Global Network
on Extremism & Technology

Künstliche Intelligenz und Terrorabwehr: eine Einführung

Marie Schroeter

GNET ist ein Sonderprojekt des International Centre for the Study of Radicalisation, King's College London.

Die Autorin dieses Berichts ist Marie Schroeter, Mercator Fellow für neue Technologien in internationalen Beziehungen: Potenziale und Grenzen der künstlichen Intelligenz zur Verhinderung von gewalttätigem Extremismus im Internet

Das Global Network on Extremism and Technology (GNET) ist eine akademische Forschungsinitiative mit Unterstützung des Global Internet Forum to Counter Terrorism (GIFCT), einer unabhängigen, aber von der Wirtschaft finanzierte Initiative mit dem Ziel, die Nutzung von Technologie für terroristische Zwecke besser zu verstehen und einzudämmen. GNET wird einberufen und geleitet vom International Centre for the Study of Radicalisation (ICSR), einem akademischen Forschungszentrum innerhalb des Department of War Studies am King's College London. Die in diesem Dokument enthaltenen Ansichten und Schlussfolgerungen sind den Autoren zuzuschreiben und sollten nicht als die ausdrücklichen oder stillschweigenden Ansichten und Schlussfolgerungen von GIFCT, GNET oder ICSR verstanden werden.

Wir danken Tech Against Terrorism für die Unterstützung bei diesem Bericht.

KONTAKTANGABEN

Im Falle von Fragen oder zur Anforderung weiterer Exemplare wenden Sie sich bitte an:

ICSR
King's College London
Strand
London WC2R 2LS
United Kingdom

T. **+44 20 7848 2098**
E. **mail@gnet-research.org**

Twitter: **@GNET_research**

Wie alle anderen GNET-Publikationen kann auch dieser Bericht kostenlos von der GNET-Website unter www.gnet-research.org heruntergeladen werden.

Kurzfassung

Radikalisierung kann sowohl in der realen Welt als auch online stattfinden. Welche Rolle hierbei das Internet spielt, ist nach wie vor umstritten. Zweifellos gibt es im Netz aber radikale und extreme Communitys. Der vorliegende Bericht untersucht die Fähigkeit von Anwendungen der künstlichen Intelligenz (KI), zur Abwehr derartiger Radikalisierung beizutragen. Der Bericht stellt die Möglichkeiten und Grenzen dieser Technologie in ihren verschiedenen Formen dar und soll Entscheidungsträgern und Experten dabei helfen, unbeirrt von den Sensationsmeldungen und dem aktuellen Hype zu fundierten Entscheidungen zu gelangen. Im Wesentlichen kommt der Bericht zu folgenden Ergebnissen:

1. Entsprechend programmierte Suchmaschinen und Empfehlungssysteme können zur Gegenradikalisierung beitragen, indem sie zu moderaten Inhalten führen

Suchmaschinen und Empfehlungssysteme können maßgeblich dazu beizutragen, Online-Räume sicherer zu machen. Indem sie die Wahrscheinlichkeit verringern, auf radikalisierende Inhalte zu stoßen, tragen sie zur Prävention von gewalttätigem Extremismus bei. Suchmaschinen helfen, sich im Dschungel der Online-Informationen, die auch extremistische Inhalte einschließen, zurechtzufinden. Entsprechend programmierte Algorithmen könnten vorzugsweise auf moderate anstatt auf extremistische Inhalte verweisen. Ebenso können Empfehlungssysteme, die auf der Grundlage des Browserverlaufs das nächste Video, den nächsten Song oder den nächsten Film vorschlagen, möglicherweise extreme Standpunkte bekräftigen, indem sie bestätigende Inhalte empfehlen. Ein ausgewogenes Empfehlungssystem würde böswilligen Narrativen mit gegensätzlichen Inhalten begegnen oder Informationen über Projekte und Anlaufstellen zur Prävention und Bekämpfung von gewalttätigem Extremismus verbreiten.

2. Die Verarbeitung natürlicher Sprachen kann bei der Übersetzung von Minderheitensprachen zwecks besserer Content-Moderation helfen und auf lange Sicht die Content-Moderation von Nischen-Websites unterstützen

Die Verarbeitung natürlicher Sprachen (Natural Language Processing, NLP) bietet Potenzial für die Moderation von Online-Inhalten, insbesondere im Hinblick auf Sprachen, die nur von kleinen Gruppen von Menschen gesprochen werden. Häufig erscheint die Moderation von Inhalten in Minderheitensprachen nicht rentabel genug für die entsprechenden Investitionen. Kleinere Plattformen verfügen nicht immer über das technische Know-how oder die Ressourcen für Content-Moderationssysteme, da schon der Einsatz vorhandener Modelle einen erheblichen Zeit- und Arbeitsaufwand verlangt.

Andere vertreten eine extreme Auffassung des Rechts auf freie Meinungsäußerung und wollen aus diesem Grund ihre Nutzer nicht einschränken. Verbessertes NLP kann helfen, Inhalte in Sprachen zu übersetzen, in denen eine große Anzahl erfahrener und geschulter Moderatoren tätig ist. NLP kann außerdem ungewöhnliche semantische Muster auf Websites erkennen. Dies könnte die Erkennung kritischer Botschaften auf Plattformen unterstützen, die nicht in Content-Moderation investieren wollen oder können. Derartige Maßnahmen müssen jedoch jederzeit die Datenschutzstandards und die Menschenrechte respektieren.

3. Bei der Bekämpfung von Desinformation und manipulierten Inhalten im Internet fehlt es an automatisierten Lösungen

Bis heute gibt es keine überzeugenden automatisierten Werkzeuge, die Desinformationen und manipulierte Inhalte identifizieren und bekämpfen, die zwar schädlich, aber legal sind. Eine sehr deutlich verbesserte digitale Kompetenz der Nutzer zur Schaffung einer digitaler Souveränität scheint kurzfristig ein besserer Ansatz zu sein.

4. Übermenschliche KI wird nicht „Alarm schlagen“, wenn einzelne Personen sich im Internet radikalisieren

Eine allgemeine KI, die mit übermenschlichen Fähigkeiten den Content und das Verhalten von Einzelpersonen online überwacht und „Alarm schlägt“, wenn Indikatoren für eine Radikalisierung zusammenkommen, ist nicht realisierbar und wird aus zwei Gründen Science-Fiction bleiben. Erstens gibt es nicht genügend Daten, um einen Algorithmus mit eindeutigen Informationen über Radikalisierung zu füttern und darüber, wann ein radikalisiertes Individuum zur Gewalt greift. Solange es keine technische Innovation gibt, die zuverlässige Systeme auf einer weitaus kleineren Datenbasis ermöglicht, ist der Einsatz technischer Lösungen sinnlos, da sie ohne eine ausreichende Menge an Daten aus früheren Fällen ohnehin keine zuverlässigen Vorhersagen treffen können. Radikalisierung und Terrorismus sind – glücklicherweise – zu selten und vielfältig, um genügend Informationen für einen Algorithmus zu produzieren. Zweitens würde die Vorhersage des Verhaltens von Einzelpersonen klar zuzuordnende Daten erfordern, was in jeder Hinsicht den Schutz der Privatsphäre verletzen und potenziell zu einer Überwachung in einem noch nie dagewesenen Ausmaß führen könnte. Das beschriebene Szenario ist nicht vereinbar mit liberalen Demokratien, in denen die Persönlichkeitsrechte einen hohen Stellenwert einnehmen.

Inhalt

1 Einleitung

In der Öffentlichkeit herrscht weit verbreitet die Ansicht, künstliche Intelligenz (KI) werde alles revolutionieren und so auch die nationale Sicherheit. Inwieweit das Internet die Radikalisierung fördert, bleibt eine unbeantwortete Frage, aber die jüngsten Terroranschläge im ostdeutschen Halle, im neuseeländischen Christchurch und in der Synagoge in Poway in Kalifornien sind nur drei aktuelle Beispiele dafür, welch große Rolle der Onlinebereich heutzutage bei der Radikalisierung spielt.

Wie kann KI dazu beitragen, der Radikalisierung im Internet entgegenzuwirken? Die Expertise zu diesem Thema verteilt sich auf unterschiedliche Disziplinen, ist jedoch sowohl bei Forschenden und Fachleuten im Bereich Sicherheit und Terrorabwehr als auch bei politischen Entscheidungsträgern und technischen Sachverständigen angesiedelt, wobei diese Gruppen sich dieses Themas zunehmend gemeinsam annehmen. Die gegenwärtige Informationslandschaft macht es Entscheidungsträgern schwer, konkrete Informationen aus dem Medienrummel herauszufiltern. Dieser Bericht will die jüngsten Entwicklungen in der KI beleuchten und in den Kontext der Bemühungen zur Bekämpfung der Radikalisierung in liberalen Demokratien stellen.

Diese Publikation trägt zum Thema bei, indem sie eine Reihe von Grenzen und Möglichkeiten der KI bei der Bekämpfung von Radikalisierung im Internet aufzeigt. Das zweite Kapitel geht kurz auf die Schlüsselkonzepte und -ideen der KI ein. In einem Exkurs am Ende werden die Aspekte Datenqualität sowie Verzerrungen und Manipulationen in Datenbeständen eingehender betrachtet. Das dritte Kapitel erörtert das Potenzial und die Grenzen KI-basierter technologischer Innovationen für ein „gesundes“ Internet, das frei ist von terroristischen Inhalten, Propagandamaterial und Fake-Engagement. Hierbei wird die Annahme zugrunde gelegt, dass eine solche gesunde Online-Umgebung Radikalisierung entgegenwirken kann. Das Kapitel bewertet eine Reihe verbreiteter KI-basierter Konzepte (von Deepfakes bis hin zu Bot-Armeen, die Fake News verbreiten) und erklärt, warum Suchmaschinen, Empfehlungssysteme und insbesondere die Verarbeitung natürlicher Sprachen (Natural Language Processing, NLP) geeignet sind, auf die eine oder andere Weise zur Erreichung dieses Ziels beizutragen. Das vierte Kapitel befasst sich ausschließlich mit einer hypothetischen „generellen KI“, einem allwissenden System, das Personen, die sich radikalisieren, identifiziert und somit den Strafverfolgungsbehörden helfen kann, Verbrechen zu verhindern, bevor sie geschehen. In diesem Kapitel wird allerdings auch begründet, warum eine solche KI auf absehbare Zeit ausschließlich Science-Fiction bleiben wird. Dies leitet über zu einer Erörterung der Gründe für diese Position. Debatten zum Thema Big Data, insbesondere in Zusammenhang mit Sicherheit im herkömmlichen Sinne, müssen in liberalen Demokratien immer den Schutz der Privatsphäre berücksichtigen und priorisieren. Ein weiterer Exkurs in Kapitel vier enthält ausführlichere Informationen für interessierte Leser. Das fünfte Kapitel schließt den Bericht ab.

Der Bericht basiert auf halbstrukturierten Interviews mit Forschern, politischen Entscheidungsträgern und Beratern sowie mit Vertretern des privaten Sektors. Darüber hinaus haben Erkenntnisse aus der Schreibtischforschung und der Medienbeobachtung die Positionen dieses Berichts beeinflusst. Ich habe mit verschiedenen Stakeholdern gesprochen, um einen multidisziplinären Blickwinkel zu gewinnen, der der fragmentierten Informationslandschaft Rechnung trägt. Da Informationen über den Einsatz von maschinellem Lernen teilweise durch Geheimdienste oder Unternehmen des privaten Sektors zurückgehalten werden, gibt es jedoch klare Grenzen für die Forschung.

2 Was ist künstliche Intelligenz?

Obgleich KI heutzutage in aller Munde ist, gibt es keine weltweit einheitliche Definition. Dies ist zum Teil darauf zurückzuführen, dass KI ein sich rasant entwickelndes und populäres Forschungsgebiet ist, das ständig neue Erkenntnisse hervorbringt und die Grenzen zwischen Informatik, Statistik und Robotik verwischt. Obwohl es keinen Konsens über die Definition gibt, betrifft KI doch weite Teile unseres Lebens. Sie ist die Art von Technologie, die uns Empfehlungen für unseren nächsten Online-Einkauf gibt, Terminkalender führt und fahrerlose Autos lenkt.

Alexa, die Sprachassistentin von Amazon, zeigt die ausgefeilten Möglichkeiten automatisierter Entscheidungsfindungssysteme: Alexa kann komfortabel einen kompletten Ausgehabend organisieren, vom Kauf der Eintrittskarten für eine Vorstellung über die Tischreservierung im Restaurant und die Bestellung des Taxis bis hin zur Mitteilung Ihrer voraussichtlichen Ankunftszeit an Ihre Begleitung.[1] Allgemeiner beschreibt der Begriff KI eine Disziplin, die sich mit automatisierten und adaptiven technischen Systemen befasst. KI erfüllt Aufgaben ohne ständige Anleitung und ist in der Lage, ihre Leistung durch Lernen aus früheren Erfahrungen zu verbessern.[2]

Der Begriff „Artificial Intelligence“ (künstliche Intelligenz, KI) wurde 1956 während einer Konferenz am Dartmouth College in Hanover, New Hampshire, USA geprägt. Nach einigen Anfangserfolgen waren die Forscher optimistisch, dass computergestützte Algorithmen raschen Fortschritt herbeiführen würden. In dieser Frühphase waren sie in der Lage, Code zu schreiben, der Probleme lösen konnte; die Programme enthielten Elemente, deren Leistung sich durch Lernen verbesserte. Weil jedoch die Kapazität und Leistungsfähigkeit der Speicher und Prozessoren nicht ausreichte, folgte ein „KI-Winter“. In den 1960er Jahren wurden die Investitionen in die entsprechende Forschung eingefroren, und das Interesse ließ nach.

Der aktuelle Hype um die KI wurde durch technische Fortschritte des 21. Jahrhunderts ermöglicht, die in erster Linie durch den privaten Sektor vorangetrieben wurden. Die fallenden Preise für Massenspeicher und Software in Kombination mit der Expertise von Fachleuten und besserem Zugang zu Daten verliehen dem Bereich enormen Auftrieb. Was ist das Besondere an KI? Zunächst einmal erleichtert sie die Analyse von Massendaten; sie ist schneller und effizienter als Menschen, die vor dieselbe Aufgabe gestellt sind. Zweitens ist die entsprechende Technologie in der Lage, mit Ungewissheiten zu arbeiten und auf dieser Grundlage Vorhersagen für die Zukunft zu treffen. Wie zuverlässig diese Prognosen sind, ist dabei zunächst einmal zweitrangig. Die Stärke von Algorithmen,

1 Hao, K. (2019a), „Inside Amazon’s plan for Alexa to run your entire life“, MIT Technology Review. Abgerufen: https://www.technologyreview.com/s/614676/amazon-alexa-will-run-your-life-data-privacy/
2 Reaktor & University of Helsinki (2018), „How should we define AI?“. Abgerufen: https://course.elementsofai.com/1/1

die Vorhersagen machen können, ist diese Fähigkeit an sich. Demgegenüber ist das menschliche Gehirn nicht in der Lage, Entscheidungen auf der Grundlage großer Datenbestände, verschiedenartiger Bedingungen und Unsicherheiten zu treffen. Die Fähigkeit zu Voraussagen kann als die charakterisierende Fähigkeit von Algorithmen angesehen werden.

Der Begriff „KI“ ist an sich irreführend, denn er suggeriert, dass eine Ähnlichkeit mit der menschlichen Intelligenz oder menschlichen Lernprozessen besteht. Die neuronalen Netze des Deep Learning, einer speziellen Methode des maschinellen Lernens mit mehreren Zwischenschichten der Informationsverarbeitung, sind zwar in der Tat der Architektur des menschlichen Gehirns nachempfunden, jedoch unterscheiden sich die Fähigkeiten derartiger Systeme stark von denjenigen menschlicher Neuronen. Selbst in komplexen Fällen ist der Algorithmus in der Lage, fehlende Daten mithilfe von Prognosemodellen zu ergänzen, kann aber seinen Resultaten keine Bedeutung verleihen. Die Unterschiede zwischen menschlicher und maschineller Intelligenz werden deutlich, wenn man sich anschaut, was künstliche neuronale Netze können und was nicht. Beispielsweise kann ein Algorithmus Brustkrebs im Frühstadium zuverlässiger erkennen als der Mensch, weil er Mammographiebilder mit einer geringeren Fehlerquote analysieren kann als die Radiologen.[3] Andererseits kann der Algorithmus Emotionen der Patientin nicht verstehen bzw. deuten und angemessen reagieren. Einfühlungsvermögen (Empathie) erfordert jahrelanges Beobachten sowie emotionale Intelligenz, für die es keine Algorithmen gibt. Darüber hinaus impliziert das Wort „Intelligenz“ im KI-Begriff, dass das System in der Lage ist, originäre Gedanken hervorzubringen, was definitiv zu weit hergeholt ist. Googles KI-Programm AlphaGo kann mühelos die aussichtsreichsten Züge in dem hochkomplexen Spiel Go berechnen, was aber nicht heißt, dass AlphaGo das Spiel an sich versteht.[4] AlphaGo ist nicht in der Lage, den Kontext für solche Züge zu erklären, oder festzustellen, dass es ein Spiel spielt, geschweige denn zu begründen, warum es spielen möchte. Das Beimessen von Bedeutung und Betrachten im Zusammenhang ist eine sehr menschliche Fähigkeit; die meisten Kinder sind in der Lage zu erklären, warum Spielen ihnen Spaß macht. Demgegenüber kann das System zwar nicht erklären, warum es etwas Bestimmtes tut, kann aber dennoch für eine gegebene Situation den bestmöglichen Schritt im Hinblick auf das vorgegebene Ziel identifizieren. Es analysiert alle Optionen und entscheidet mathematisch, wie Risiken minimiert werden können und wie mit Unsicherheiten umzugehen ist.

Wenn man die Stärken und Schwächen der KI betrachten will, bieten sich zwei Kategorien an. AlphaGo lässt sich als „enge“ KI betrachten, da sie nur eine bestimmte Aufgabe erfüllt. Eine „allgemeine“ KI wäre demgegenüber in der Lage, jedwede intellektuelle Aufgabe zu erfüllen. (Eine derartige KI ist derzeit reine Science-Fiction.) Die Intelligenz eines Systems kann „schwach“ oder „stark“ sein, was einer engen bzw. allgemeinen KI entspricht. Der Begriff „enge KI“ bezieht sich auf Systeme,

3 Hao, K. (2020), „Google’s AI breast cancer screening tool is learning to generalize across countries“, MIT Technology Review. Abgerufen: https://www.technologyreview.com/f/615004/googles-ai-breast-cancer-screening-tool-is-learning-to-generalize-across-countries/

4 Gibney, E. (2017), „Self-taught AI is best yet at strategy game Go“, Nature. Abgerufen: https://www.nature.com/news/self-taught-ai-is-best-yet-at-strategy-game-go-1.22858

die vorgeben, intelligent zu sein, indem sie die gewünschten Ergebnisse hervorbringen. Dabei kann die Intelligenz oberflächlich sein und basiert oft auf falschen Strukturen: So sind unter Umständen Algorithmen, die darauf trainiert sind, Züge auf Bildern zu erkennen, nicht in der Lage, einen Zug selbst zu identifizieren. Vielmehr erkennen sie die auf Bildern von Zügen häufig vorkommenden parallelen Gleise. Allerdings konnte sich der Algorithmus in seinen neuronalen Netzen auf falsche Strukturen verlassen, weil diese das gewünschte Ergebnis lieferten.[5] Damit sind ganz offensichtlich Risiken verbunden, und es ist noch nicht vollkommen klar, welche ungewollten Folgen sich daraus ergeben können. Eine bekannte Folge ist zum Beispiel, dass Gesichtserkennungssysteme bei der Identifikation von People of Colour schwach abschneiden.[6] Eine allgemeine KI hätte einen echten Verstand, ein Bewusstsein oder eine Superintelligenz, über die populäre Medien mit ihr kommunizieren können. Noch einmal sei aber darauf hingewiesen, dass superintelligente Systeme bisher nur in Science-Fiction existieren.

Die Bedeutung des Begriffs AI hat sich im Laufe der Zeit verändert. Heutzutage werden KI und maschinelles Lernen (ML) in den Medien vielfach gleichbedeutend verwendet und ebenso im vorliegenden Bericht. Generell finden zwei Bereiche des ML viel Beachtung: überwachtes und unüberwachtes Lernen. Überwachtes ML bedeutet, dass der Algorithmus Daten analysiert, nachdem er anhand einer bestimmten Menge an gekennzeichnet Daten trainiert wurde. Die Kennzeichnung kann als Unterscheidung zwischen „die Daten passen zur Bedingung“ und „die Daten passen nicht zur Bedingung“ verstanden werden. Gekennzeichnete Daten verleihen den Datenpunkten daher eine Bedeutung, was menschliches Zutun erfordert. So könnte ein Bild beispielsweise einen Apfel zeigen. Wenn dies der Fall ist, würde es als „Apfel“ gekennzeichnet. Um einen Algorithmus zu trainieren, bedarf es des Zugangs zu einer großen Menge eindeutig gekennzeichneter Daten. Mithilfe eines Testdatenbestand kann die Leistung des Algorithmus evaluiert werden. Wenn der Test als erfolgreich beurteilt wird, kann der Algorithmus neue Daten kennzeichnen. Der Vorteil des überwachten Lernens besteht darin, dass der Algorithmus die Daten genau so organisiert wie programmiert. Die Kennzeichnung von Daten von Hand ist ein arbeitsintensives und kostenintensives Unterfangen. Viele Internetnutzer haben schon selbst Daten gekennzeichnet, auf Websites die Sicherheitskontrollen wie „Ich bin kein Roboter“ beinhalten. Derartige Kontrollen können den Besucher beispielsweise auffordern, in einer Bilderauswahl all diejenigen zu kennzeichnen, auf denen Autos zu sehen sind. Dies ist der Punkt, an dem Internetnutzer Daten kennzeichnen. reCaptcha von Google ist ein solcher Validierungsservice zur Bot-Abwehr und verwendet die so gekennzeichneten Datenbestände, um ML-Algorithmen zu trainieren.[7] Zum Beispiel könnten die gekennzeichneten Daten für fahrerlose Autos nützlich sein.

5 Thesing, L. et al. (2019), „What do AI algorithms actually learn? – On false structures in deep learning“, Arxiv. Abgerufen: https://arxiv.org/abs/1906.01478
6 Simonite, T. (2019), „The best Algorithms Struggle to Recognize Black Faces Equally“, Wired. Abgerufen: https://www.wired.com/story/best-algorithms-struggle-recognize-black-faces-equally/
7 Google ReCaptcha (o. J.). Abgerufen: https://www.google.com/recaptcha/intro/v3.html

Daneben gibt es unüberwachtes ML. Hierbei muss der ML-Algorithmus in nicht gekennzeichneten, unsortierten Datenmengen Muster erkennen, ohne vorher darauf trainiert worden zu sein, in welcher Korrelation die Eingabedaten zu den Ausgabedaten stehen. Unüberwachte Algorithmen erkennen ohne weitere Anweisungen oder Klassifizierungen Muster innerhalb von Datenbeständen. Die Schwierigkeit besteht darin, dass vorher keinesfalls feststeht, ob die Muster, die der Algorithmus findet, irgendeinen Wert für die Programmierer des Algorithmus haben werden. Es ist nicht auszuschließen, dass die Ergebnisse vollkommen irrelevant sind. Nichtsdestotrotz erspart es die arbeitsintensive Kennzeichnung von Daten und nutzt die riesige Menge an ungekennzeichneten Daten, die offen zugänglich sind. Unüberwachtes Lernen kann als erster Schritt vor der Arbeit mit einem Algorithmus für überwachtes Lernen zur Anwendung kommen.

In Berichten und den Nachrichten hört man oft von „Deep Learning“. Hierbei handelt es sich um ein spezielles ML-Verfahren, bei dem zahlreiche Verarbeitungseinheiten in einem Netzwerk miteinander verbunden sind. Der Umfang dieses Netzwerks ermöglicht die Analyse komplexerer Probleme. Ein anderer Begriff in den Schlagzeilen sind „neuronale Netzwerke“. Diese sind von der Funktionsweise des menschlichen Gehirns inspirierte Verfahren, die Informationen sowohl speichern als auch verarbeiten können. Neuronale Netzwerke haben der Verwertung von Massendaten (Big Data) zum Durchbruch verholfen, da sie in der Lage sind, entsprechende Datenmengen zu verarbeiten.

Exkurs

Daten – Qualität, Verzerrung und Manipulation

Kaum eine Besprechung oder Konferenz zum Thema KI, bei der nicht jemand behauptet „Daten sind das neue Öl“. Aber stimmt das wirklich? Gewiss benötigt ML große Mengen an Daten, und deren Beschaffung ist teuer. Je besser die Algorithmen werden, desto mehr Daten brauchen sie auch. Allerdings sind nicht alle Datenpunkte gleich wertvoll, zumal auch das Gesetz vom abnehmenden Ertragszuwachs greift: Ein Algorithmus lernt mehr aus frühen Daten als aus der millionsten Wiederholung. Auch funktionieren Algorithmen besonders gut, wenn sie durch ihre Trainingsdaten auf viele Eventualitäten vorbereitet wurden – mit einem Datenbestand, die sowohl gewöhnliche als auch ungewöhnliche Elemente umfasst. „Das neue Öl“ sind Daten demnach, wenn man die Aspekte hoher Preis, hohe Nachfrage (viele neue Geschäftsmodelle, die ML verwenden, sind auf Daten angewiesen) und die derzeit noch kleine Gruppe der Besitzer dieses Rohstoffs anschaut. Während jedoch jeder Tropfen Öl zur Gesamtfördermenge beiträgt, beeinflussen die Datenpunkte den Wert von Daten insgesamt in unterschiedlichem Maße.

Algorithmen sind zudem anfällig für „Bias“, eine systematische Verzerrung der Realität, die sich aus den verwendeten Datenproben ergibt. Bias der Eingangsdaten führt unweigerlich zu Bias der Ausgangsdaten – wenn der Algorithmus ihn verstärkt, sogar in wesentlich stärkerer Ausprägung. Bias ist in aktuellen Datenbeständen mehrfach nachgewiesen worden, insbesondere in Form von Voreingenommenheit gegen Frauen und People of Colour. So musste beispielsweise Amazon sein Programm zur automatisierten Personalbeschaffung aufgeben, weil es Frauen diskriminierte. Der Algorithmus bewertete Männer besser als Frauen, weil er anhand von Bewerbungsunterlagen aus dem vergangenen Jahrzehnt trainiert worden war. Es ist hinlänglich bekannt, dass der Technologiesektor stark männlich dominiert ist, was als Verzerrung jedoch in den Auswahlentscheidungen des Algorithmus reproduziert wurde.[8]

Gute Daten ohne Bias sind die Grundvoraussetzung für Algorithmen. Nach einem Konsultationsprozess gründeten die Vereinten Nationen die Organisation Tech Against Terrorism. Diese richtete die Terrorist Content Analytics Platform (TCAP) ein, um einen solchen Datenbestand zur Nutzung durch den privaten Sektor, die Wissenschaft und die Zivilgesellschaft bereitzustellen. Ursprünglich wollte Tech Against Terrorism nur Inhalte von al-Qaida und „Islamischer Staat“ (IS) einbeziehen, gab jedoch später bekannt, dass die TCAP auch den rechtsextremen Terrorismus berücksichtigen werde. Es gibt weitere Bereiche, die Aufmerksamkeit verdienen, wie z. B. Terrorismus, der von

8 Dastin, J. (2018), „Amazon scraps secret AI recruiting tool that showed bias against women“, Reuters Technology News. Abgerufen: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

frauenfeindlichen Ideologien angeheizt wird, wie bei der Incel-Bewegung zu beobachten. Selbstverständlich müssen derartige Datenbestände den Datensicherheitsstandards genügen und auch mögliche psychische Belastungen der Begutachter berücksichtigen.

Daten können jedoch manipuliert werden. Manipulierte Datenbestände sind von außen schwer zu erkennen, insbesondere technische Laien. Dabei müssen Daten gar nicht unbedingt stark verändert werden, um einen Algorithmus zu manipulieren, wie chinesische Forscher nachgewiesen haben. Ihre Experimente brachten ein selbstfahrendes Auto dazu, auf der falschen Straßenseite zu fahren. Dies veranschaulicht, wie verwundbar diese Systeme sind.[9] Bei KI-Anwendungen, die eine Online-Radikalisierung verhindern sollen, könnte eine Manipulation die Intention ins Gegenteil verkehren. Man kann sich zum Beispiel auch ein ausgeklügeltes Content-Moderation-System vorstellen, das im Vorfeld einer Wahl automatisch Profile der Oppositionspartei aus dem Netz tilgt. Eine weitere Schwäche ergibt sich aus dem „kontradiktorischen" Training der ML-Systeme: durch zwei Systeme, die miteinander konkurrieren und sich dadurch gegenseitig in der Leistung verbessern. Zum Beispiel generiert ein ML-System gefälschte Bilder von Gesichtern, und das andere muss sie aus einer Reihe echter Bilder herausfiltern. Selbst wenn das Filtersystem immer besser wird, so entwickelt sich doch auch das System, dass die Fake-Bilder generiert, immer weiter. Wohin das führen wird, ist derzeit noch unklar.

9 Knight, W. (2019), „Military artificial intelligence can be easily and dangerously fooled", MIT Technology Review. Abgerufen: https://www.technologyreview.com/2019/10/21/132277/military-artificial-intelligence-can-be-easily-and-dangerously-fooled/

3 KI gegen Radikalisierung im Internet – der Teufel steckt im Detail

Wie können Statistik und automatisierte Entscheidungsfindung dazu betragen, Radikalisierung im Internet entgegenzuwirken? Wie immer steckt der Teufel im Detail, und der Wirbel um KI kann verwirren. Außer für technische Experten ist es daher schwer zu beurteilen, wie viel Potenzial tatsächlich vorhanden ist. Dieses Kapitel befasst sich eingehend mit den Möglichkeiten und Grenzen populärer technologischer Innovationen, die auf ML beruhen, und konzentriert sich dabei auf Aspekte, die die öffentliche Diskussion dominieren. Von Deepfakes bis hin zu automatisierter Content-Moderation, Suchmaschinen und Natural Language Processing beleuchtet dieses Kapitel die Technologien hinsichtlich ihrer Brauchbarkeit für die Gegenradikalisierung. Es gibt Überschneidungen zwischen den Elementen, was in einem sich derart schnell verändernden Umfeld mit häufigen Weiterentwicklungen unvermeidlich ist.

3.1 Beeinflussung der Online-Erfahrung – Was für Nutzer sichtbar und leicht zu finden ist

Maschinelles Lernen kann die Online-Erfahrung in hohem Maße beeinflussen, indem es beeinflusst, welche Inhalte für Internetnutzer leicht zu finden und zu betrachten sind. Die Algorithmen in ihren verschiedenen Ausprägungen haben ein großes Potenzial, der Radikalisierung entgegenzuwirken, indem sie zu einem gesünderen Online-Raum beitragen und bösartige Inhalte verhindern. Dieser Bericht befasst sich mit Suchmaschinen, Empfehlungssystemen und Content-Moderation.

Suchmaschinen helfen im Wesentlichen dabei, die Millionen von Webseiten zu durchsuchen und relevante Online-Inhalte zu finden. Wie ein Telefonbuch des 21. Jahrhunderts führen Suchmaschinen in der Masse der online verfügbaren Informationen und Daten in Richtung des Gesuchten. Maßgeblich für den Erfolg sind dabei die Algorithmen, die den Suchmaschinen zugrunde liegen. Vor zehn Jahren war die Landschaft der Suchmaschinen noch sehr viel vielfältiger, aber letztendlich machte Googles „Geheimrezept“ das Unternehmen zum Spitzenreiter, der durch die Bereitstellung relevanter Inhalte Vertrauen in seine Tools aufbaute. Heute bedient Google täglich Milliarden von Suchanfragen, darunter 15 % gänzlich neue Abfragen. Die anwenderfreundlichen Algorithmen von Google finden nicht nur die gewünschten Informationen, sondern ignorieren auch Rechtschreibfehler und schlagen automatisch das nächste Wort in der Suchleiste vor. Letzten Endes entscheidet jedoch die Programmierung des Algorithmus darüber, welche Informationen präsentiert werden. Die Zugänglichkeit von Anleitungen zum Bombenbau hat Berichten zufolge bereits direkt zu terroristischen Aktivitäten geführt, wie im Fall von Noelle Velentzas und Asia

Siddiqui. Die beiden Frauen hatten, wie ein FBI-Agent beobachtete, das al-Qaida-Magazin *Inspire*, Blog-Beiträge über selbstgemachte Sprengstoffe und *The Anarchist Cookbook* zur Herstellung von Bomben benutzt.[10] Großbritannien führte die „Operation Cupcake“ durch, so genannt, weil der Auslandsgeheimdienst MI6 und das GCHQ einen Leitfaden zur Herstellung einer Bombe in *Inspire* durch Rezepte für die „Besten Cupcakes Amerikas“ ersetzten. Zusammengefasst: Ob im Internet Anleitungen für den Bau von Bomben mit Haushaltsmitteln zu finden sind oder die Algorithmen so programmiert sind, dass extremistische Inhalte im Internet schwer aufzufinden sind, kann einen großen Unterschied machen. Zwar werden sich technisch versierte Nutzer hierdurch nicht komplett stoppen lassen, doch erhöht es zumindest die Barrieren, die zum Abruf derartiger Inhalte zu überwinden sind.

Empfehlungssysteme sind ein praktisches Hilfsmittel, um den nächsten Clip, Song, Beitrag oder Einkaufsartikel auf der Grundlage zuvor konsumierter oder gekaufter Objekte zu finden. Empfehlungssysteme können zur Entdeckung eines neuen Lieblingssongs führen oder den Kauf des passenden Bettzeugs für die neu erworbene Matratze erleichtern. Andererseits können die Algorithmen, die die nächsten Dinge zum Anschauen oder Anhören vorschlagen, auch **Filterblasen** erzeugen, die Annahmen verstärken, indem sie immer wieder ähnliches Material vorschlagen. Folglich können sie auch extremistische Haltungen bekräftigen. Die Undurchsichtigkeit der Art und Weise, wie Algorithmen in sozialen Medien sowie auf Musik-, Video- oder Film-Websites neue Elemente vorschlagen, lässt dem Verbraucher keine Wahl, nach welchen Kriterien er Empfehlungen erhalten möchte: mehr von derselben Art, gegensätzliche Ansichten oder eine bestimmte Kombination aus beidem. Wie Untersuchungen zu automatisch vorgeschlagenen Inhalten auf größeren Social-Media-Plattformen ab 2019 feststellten, trugen insbesondere die Algorithmen von YouTube zu einer Verstärkung extremistischer Ansichten bei. Sobald ein Video mit extremistischem „grenzwertigem" Inhalt angeschaut wurde, werden weitere ähnliche Inhalte empfohlen.[11] Dies ist besonders beunruhigend, wenn man bedenkt, dass YouTube die beliebteste Social-Media-Plattform bei Erwachsenen in den USA ist und es nicht unwahrscheinlich ist, dass viele Nutzer YouTube als Ihren Nachrichtenlieferanten nutzen.[12]

Mainstream-Social-Media gerieten wiederholt in die Kritik, weil sie nicht entschieden genug gegen den Missbrauch ihrer Plattformen durch Terroristen vorgegangen waren. Es wurde gefordert, dass angesichts der Bedeutung der sozialen Medien als quasi öffentlicher Raum, in dem Menschen sich begegnen, Argumente austauschen und Geschäfte tätigen, Mittelspersonen die Verantwortung für den Umgang mit derartigen Inhalten übernehmen sollten. Heute bieten jedoch unzählige **Nischendienste** eine diversifizierte Landschaft von Online-Diensten über das gesamte Spektrum der Internet-Infrastruktur an. Sie reichen von Messenger-Diensten mit einem hohen Grad an Anonymität über Nischenplattformen ohne die Möglichkeit oder

10 United States District Court, Eastern District of New York (2015), United States of America vs. Noelle Velentzas and Asia Siddiqui. Complaint and affidavit in support of arrest warrant, 2014R00196. Abgerufen: https://www.justice.gov/sites/default/files/opa/press-releases/attachments/2015/04/02/velentzas-siddiqui-complaint.pdf

11 Reed et al. (2019), „Radical Filter Bubbles“, in der 2019 GRNTT Series, an Anthology, RUSI, London.

12 Perrin, A. & Anderson, M. (2019), „Share of U.S. adults using social media, including Facebook, is mostly unchanged since 2018“, Pew Research Centre. Abgerufen: https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/

Absicht zur Überwachung von Inhalten bis hin zu Hosting-Diensten, die die Verbreitung von öffentlichen Erklärungen (Manifesten) und Live-Videos ermöglichen. In jüngster Zeit sind Websites und Dienste der Computerspielindustrie in die Kritik geraten, weil sie böswillige Nutzung zugelassen haben.[13] Die Swansea University in Wales hat in einer neueren Untersuchung gezeigt, wie das IS-Netzwerk diverse Hosting-Dienste für sein Online-Magazin Rumiyah genutzt hat. Dadurch, dass der Content auf diese Weise dezentralisiert wurde, war eine effiziente und schnelle Entfernung erschwert.[14] Weitere Untersuchungen und Metadaten dazu, wie Nischendienste von Terroristen genutzt werden, sind dringend erforderlich.

3.2 Management von User-created Content

Web 2.0 hat die Nutzungsart des Internets revolutioniert und bezeichnet die Abkehr von statischen Websites hin zu Echtzeit-Interaktionen einer großen Anzahl von Nutzern weltweit. Während diese globale Vernetzung viele internationale Communities in einem nie dagewesenen Ausmaß förderte und unterstützte, brachte sie auch neue Herausforderungen für die Radikalisierungsbekämpfung mit sich.

Automatisierte Content-Moderation auf Social-Media-Plattformen soll die Verbreitung terroristischer Inhalte verhindern. 98 % der bösartigen Inhalte auf Facebook werden den jüngsten EU-Selbstbewertungsberichten zu den Verfahrensregeln gegen Desinformation bereits durch ML-Algorithmen herausgefiltert.[15] Die verbleibenden 2 % werden von Nutzern gemeldet. Twitter berichtet, dass pro Sekunde zehn Konten beanstandet werden;[16] Google, der Eigentümer von YouTube, entfernt nach eigenen Angaben 80 % der unangemessenen Videos, bevor sie überhaupt angesehen werden.[17] Zunächst klingt das alles nach erfolgreicher Moderation, und man kann auch durchaus behaupten, dass sich die Content-Moderation in den vergangenen Jahren verbessert habe. Doch noch immer gibt es zahlreiche Schlupflöcher, vor allem außerhalb der Standard-Sprachzonen. Gegenwärtig umfasst Facebooks Fact-Checking-Repertoire nur 14 der 26 offiziellen Sprachen in Europa. Durch externe Dienstleister werden mittlerweile auch 15 afrikanische Länder überwacht, aber das ist noch immer weniger als ein Drittel der Länder des Kontinents.[18] Es ist auch unklar, ob dies nur für die Amtssprachen gilt oder auch Dialekte einschließt. Das Herauslassen von Minderheitensprachen aus der Content-Moderation ist auch aus anderen bevölkerungsreichen und inhomogenen Regionen und Ländern wie Indien bekannt.[19]

13 Schlegel, L. (2020), „Points, Ratings and Raiding the Sorcerer's Dungeon: Top-Down and Bottom-Up Gamification of Radicalisation and Extremist Violence". Abgerufen: https://gnet-research.org/2020/02/17/points-rankings-raiding-the-sorcerers-dungeon-top-down-and-bottom-up-gamification-of-radicalization-and-extremist-violence/

14 Macdonald, S. et al. (2019), „Daesh, Twitter and the Social Media Ecosystem", The RUSI Journal, Bd. 164, Nr. 4, S. 60–72.

15 Facebook (2019), *Facebook report on the implementation of the Code of Practice for Disinformation*. Abgerufen: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=62681

16 Twitter (2019), *Twitter Progress Report: Code of Practice against Disinformation*. Abgerufen: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=62682

17 Google (2019), *EC EU Code of Practice on Disinformation*. Abgerufen: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=62680

18 Africa Times (2019), „Facebook expands fact-checking to 15 African nations". Abgerufen: https://africatimes.com/2019/10/10/facebook-expands-fact-checking-to-15-african-nations/

19 Perrigo, B. (2019), „Facebook Says It's Removing More Hate Speech Than Ever Before. But There's a Catch", *Time*. Abgerufen: https://time.com/5739688/facebook-hate-speech-languages/

Wie bereits erläutert, können Algorithmen den Daten keine Bedeutung zuordnen, d. h. Algorithmen können den Kontext, in dem bösartiges Verhalten stattfindet, nicht verstehen. Beispiele aus Sri Lanka zeigen, das die Algorithmen von Facebook nicht in der Lage waren, vielschichtige kulturelle Zusammenhänge zu bewerten. Vor den Bombenanschlägen in Colombo am Ostersonntag im April 2019 waren Postings durch die Maschen geschlüpft, weil die Algorithmen nicht in der Lage waren, die Komplexität der darin enthaltenen Hassreden (Hate Speech) zu verstehen. Selbst nach mehrfachen Versuchen, die Hate Speech zu melden, die eine polarisierte, antimuslimische Stimmung nährte, gelang es Facebook nicht, die Inhalte zu löschen oder mit einer angemessenen Content-Moderation zu reagieren.[20] Um die verwendete Slangsprache als Hate Speech zu einordnen zu können, hätten die an der Content-Moderation beteiligten Algorithmen in der Lage sein müssen, die Ethnizitäten der Beteiligten zu verstehen. Aber das Problem geht noch weiter: Häufig werden Sprachen, die nicht das lateinische Alphabet verwenden, aus praktischen Erwägungen in das lateinische Alphabet „übersetzt“. Einige Sprachen haben kein grammatikalisch explizites Futur; wie würde also eine Bedrohung, die ja die Zukunft impliziert, überhaupt aussehen? Wenn die automatische Filterung ausgeweitet werden soll, muss sie sich mit diesen Designfehlern auseinandersetzen.

Viele Social-Media-Unternehmen wollen verhindern, dass ihre Plattformen von böswilligen Akteuren ausgenutzt werden. Die Wirksamkeit von Algorithmen zur Erkennung von Propagandamaterial oder terroristischen Aktivitäten hängt aber auch von der Qualität und Verfügbarkeit der Daten ab, mit denen der Algorithmus trainiert wurde. Facebook räumte ein, dass ein Mangel an Trainingsdaten dafür verantwortlich war, dass es nicht gelang, Live-Streams von Schussangriffen zu identifizieren und herauszufiltern, wie im Fall des Angriffs in Christchurch, der als Live-Stream gesendet wurde. Jetzt werden Aufnahmen der Body-Cams britischer Polizeibeamter verwendet, die bei Übungen zur Terrorismusabwehr entstehen.[21]

Eine weitere ML-basierte Innovation, die zur Kontrolle von User-created Content und umfassenderer Content-Moderation beitragen könnte, ist die **Verarbeitung natürlicher Sprachen** (Natural Language Processing, NLP). Sie umfasst technische Verfahren und Werkzeuge zur Analyse und Verarbeitung von Sprache. NLP hat vielfältige Anwendungsmöglichkeiten: von Chatbots in der Kundenkommunikation über Diktiersoftware und maschinelle Übersetzung bis hin zur Unterhaltung mit Siri – NLP ist allgegenwärtig. Insbesondere im Bereich der maschinellen Übersetzung zeigt sich, welche Fortschritte die Technologie in den vergangenen Jahren gemacht hat. Früher waren brauchbare Online-Übersetzungen mehr oder weniger Glücksache, doch heute erreichen sie einen weitaus höheren Grad an Zuverlässigkeit. Dennoch ist maschinelle Übersetzung noch nicht fehlerfrei bzw. soweit, die Aufgaben von Dolmetschern zu übernehmen. In Übersetzerkreisen geriet Google Translate mit einem Screenshot in die Schlagzeilen, in dem es bei der Übersetzung von „I am smart“

20 Wijeratne, Y. (2019a), „The Social Media Block isn't helping Sri Lanka“, *Slate.* Abgerufen: https://slate.com/technology/2019/04/sri-lanka-social-media-block-disinformation.html und Wijeratne, Y. (2019b), „Big Tech is as Monolingual as Americans“, Foreign Policy. Abgerufen: https://foreignpolicy.com/2019/05/07/big-tech-is-as-monolingual-as-americans/

21 Manthorpe, R. (2019), „Police share ‚shooting' video with Facebook to help identify live-streamed attacks“, *SkyNews*. Abgerufen: https://news.sky.com/story/police-share-shooting-video-with-facebook-to-help-identify-live-streamed-attacks-11843511

(„Ich bin klug“) und „I am beautiful“ („Ich bin schön“) aus dem Englischen ins Spanische und Französische jeweils die männliche Form wählte, für den Satz „I am beautiful and not smart“ („Ich bin schön und nicht klug“) hingegen die weibliche.[22] Derartige Mängel müssen behoben werden, und die Forschung und Entwicklung sind im Gange. Ein neues Verfahren namens „Masking“, das von der chinesischen Firma Baidu entwickelt wurde, lässt Übersetzungsprogramme über eine wortweise Übersetzung hinausgehen und durch Berücksichtigung des jeweiligen Kontexts zu verlässlicheren Ergebnissen kommen.[23] Dies könnte vor dem Hintergrund der jüngsten Berichte über die rechtsextreme Boogaloo-Bewegung hilfreich sein, die im Internet angeblich kodierte Sprache verwendet, um automatisierte Löschungen auf den Social-Media-Plattformen zu umgehen.[24]

Die Technologie hat ein großes Potenzial für Content-Moderation auf Websites mit einer extremen Auslegung der Meinungsfreiheit. Bekannt sind Beispiele wie 8chan, 4chan und Gab für rechtsextreme Ideologien und zahlreiche andere Formen von Gruppendiskriminierung wie Antisemitismus, Fremdenfeindlichkeit und „White Supremacy“. Der „No-Policy"-Grundsatz kann radikalen Umgebungen Vorschub leisten, weil nach der Gesetzgebung der Vereinigten Staaten alles außer illegalen Inhalten, wie z. B. Kinderpornographie, verbreitet werden darf. Im Fall der Attentate auf die Synagoge in Poway, den Walmart-Supermarkt in El Paso und die Moscheen in Christchurch im Jahr 2019 hatten alle Täter auf 8chan gepostet, bevor sie ihre Terroranschläge verübten. Es sind detailliertere Recherchen erforderlich, aber diese letzten Postings heben sich von den üblichen Hänseleien, der Ironie und der äußerst beleidigenden Sprache auf diesen Websites ab. Sie alle beziehen sich auf andere Attentate und enthalten einen Link zu einer öffentlichen Erklärung, einem anderen Schriftstück oder einem Live-Stream. Die Schützen erwähnen, dass sie möglicherweise sterben werden. Die Postings haben einen herzlichen Ton. Es ist denkbar, dass NLP höhere Bedrohungsniveaus identifizieren könnte, wenn bestimmte Indikatoren in einem Posting erfüllt sind. Derartige Indikatoren müssten den Besonderheiten der jeweiligen Plattform angepasst werden.

Durch bessere Content-Moderation könnte NLP auch die Resilienz in Online-Communitys, die Minderheitensprachen verwenden, verbessern. Die automatisierte Content-Moderation für Minderheitensprachen liefert keine zuverlässigen Ergebnisse. Anstatt darauf zu hoffen, dass die Algorithmen bald besser funktionieren werden, obwohl der wirtschaftliche Anreiz zu gering ist, um Unternehmen zu Investitionen in derartige Verbesserungen zu bewegen, könnte die Lösung in NLP liegen, d. h. in besseren Übersetzungen in Sprachen, die von erfahrenen und geschulten Moderatoren gesprochen werden. Content-Moderation könnte andere, weniger gut abgedeckte Sprachen kontrollieren, wenn die maschinelle Übersetzung einen akzeptablen Standard hat.[25] Allerdings müssen die potenziellen Anwendungen stets die Datenschutzstandards respektieren und die Menschenrechte einhalten.

22 „Marta Ziosi“, LinkedIn, Abgerufen: https://www.linkedin.com/posts/marta-ziosi-3342007a_googletranslate-googletranslate-women-activity-6603598322009808896-MQJX
23 Baidu Research (2019), „Baidu’s Pre-training Model ERNIE Achieves New NLP Benchmark Record“. Abgerufen: http://research.baidu.com/Blog/index-view?id=128
24 Owen, T. (2020), „The Boogaloo Bois are all over Facebook“, *Vice*. Abgerufen: https://www.vice.com/en_us/article/7kpm4x/the-boogaloo-bois-are-all-over-facebook
25 Wijeratne, Y. (2019b).

3.3 Durch KI generierte Inhalte – Wie man den Spieß umdreht

Manipulierter Content hat das Potenzial, extremistisches Gedankengut in den Mainstream-Diskurs einzuschleusen und Radikalisierung zu fördern. Dies könnte zu Gewalt in der Realität führen. Politische Desinformation ist keine neue Strategie, aber die Möglichkeit, quasi auf Knopfdruck ein beispiellos großes Publikum zu erreichen, um Einfluss auf den öffentlichen Diskurs zu nehmen, stellt eine neue Herausforderung dar. Dieses Kapitel befasst sich mit Möglichkeiten zur Bekämpfung von Trollen, Bots, Fake News und Deep Fakes.

Trolle oder **Bots** sind Profile in Sozialen Medien, die bestimmte Inhalte verbreiten oder künstliches Engagement auf Social-Media-Plattformen erzeugen. „Diese Bots können so programmiert werden, dass sie Aufgaben ausführen, die normalerweise mit menschlicher Interaktion verbunden sind, wie das Folgen anderer Nutzer, das Favorisieren von Tweets oder Direct Messages (DM) an andere Benutzer. Vor allem aber können sie Content twittern und alles als Retweet posten, was von einer bestimmten Gruppe von Benutzern oder mit einem bestimmten Hashtag gepostet wurde.“[26] Das Programmieren von Bots erfordert kein großes Fachwissen und ist mithilfe von online verfügbaren Anleitungen recht einfach.[27]

In großer Anzahl eingesetzte Trolle werden auch als Troll- oder Bot-Netzwerk bzw. Troll- oder Bot-Armee bezeichnet. Manipulierte Inhalte, die massenhaft verbreitet werden, können den öffentlichen Diskurs oder die öffentliche Meinung im Sinne von Eigeninteressen beeinflussen. Ein bekanntes Beispiel ist die russische Einmischung in die amerikanischen Wahlen von 2016, als ein strategisch platziertes Fake-Engagement die Kampagne von Donald Trump unterstützte und Hillary Clinton angriff. Schätzungen zufolge sind zwischen 5 % und 15 % der Online-Profile nicht echt. (Diese Zahlen sind umstritten.)[28] Laut einer Studie von Pew Research entfallen 22 % der getwitterten Links auf die fünfhundert aktivsten Bots auf Twitter, während die fünfhundert aktivsten Menschen nur auf etwa 6 % kommen. Mittlerweile sind 66 % der Profile, die auf die populärsten Websites verlinken, Bots, und nur 34 % menschlich.[29] Die Notwendigkeit einer Regulierung von Troll-Farmen wurde unlängst durch die Nachforschungen von Katarzyna Pruszkiewicz[30] aufgezeigt, die sechs Monate lang in einem polnischen Troll-Unternehmen gearbeitet hatte. Sie und ihre Kolleginnen/Kollegen lenkten Online-Gespräche zugunsten von zahlenden Kunden, darunter auch ein öffentlich-rechtlicher TV-Sender. Es ist unklar, inwieweit sich Online-Engagement tatsächlich in Wählerstimmen niederschlägt,[31], aber es ist inakzeptabel, dass Politiker und staatliche Institutionen in Demokratien Debatten mit Geld gewinnen.

26 Symantec Security Response (2018), „How to Spot a Twitter Bot“, Symantec Blogs/Election Security. Abgerufen: https://www.symantec.com/blogs/election-security/spot-twitter-bot
27 Agarwal, A. (2017), „How to write a Twitter Bot in 5 Minutes“, Digital Inspiration. Abgerufen: https://www.labnol.org/internet/write-twitter-bot/27902/
28 Burns, J. (2018), „How many Social Media Users are Real People?“, *Gizmodo*. Abgerufen: https://gizmodo.com/how-many-social-media-users-are-real-people-1826447042
29 Wojcik, S. et al. (2017), „Bots in the Twittersphere“, Pew Research Centre. Abgerufen: https://www.pewresearch.org/internet/2018/04/09/bots-in-the-twittersphere/
30 Davies, C. (2019), „Undercover reporter reveals life in a Polish troll farm“, *The Guardian*. Abgerufen: https://www.theguardian.com/world/2019/nov/01/undercover-reporter-reveals-life-in-a-polish-troll-farm
31 Eckert, S. et al. (2019), „Die Like-Fabrik“, *Süddeutsche Zeitung*. Abgerufen: https://www.sueddeutsche.de/digital/paidlikes-gekaufte-likes-facebook-instagram-youtube-1.4728833

Mark Zuckerberg, der CEO von Facebook, nannte KI als Lösung für Content-Moderation, einschließlich der Erkennung und Entfernung von Fake-Engagement jeglicher Art. Nur automatisierte Systeme könnten die Inhalte von Millionen von Nutzern in verschiedenen Sprachen und mit unterschiedlichen kulturellen Hintergründen verarbeiten, so Zuckerberg.[32] Die Einzelheiten bleiben jedoch unklar. Zuckerberg räumte bei seiner Anhörung vor dem US-Senat 2018 auch ein, dass die KI in fünf bis zehn Jahren in der Lage sein könnte, sprachliche Nuancen aufzuspüren, aber die technischen Entwicklungen sind noch soweit.[33] Vorhandene technologische Lösungen nehmen für sich in Anspruch, dass die Erkennung von Bots möglich sei. Dabei gehen sie von der Annahme aus, dass ein Bot, der für einen bestimmten Zweck geschaffen wurde, ein einzelnes Thema oder eine sehr begrenzte Auswahl an thematischen Inhalten erstellt bzw. hiermit interagiert. Menschen würden sich im Gegensatz dazu für ein breiteres Themenspektrum interessieren. Weitere Informationen für die Analyse sind außerdem das Datum und die Uhrzeit der Profilerstellung.[34] Den Versprechen dieser Technologien stehen die Erkenntnisse des NATO-Exzellenzzentrums in Riga gegenüber. Den jüngsten Untersuchungen zufolge, die Facebook, Twitter, Instagram und YouTube umfasste, funktionieren die Identifizierung und das Löschen von vorgetäuschtem Engagement nur unzureichend.[35] Den Forschern gelang es, für nur 300,00 € 3.530 Kommentare, 25.750 Likes, 20.000 Aufrufe und 5.100 Follower zu kaufen. Den Plattformen gelang es nicht, unauthentisches Verhalten oder unauthentische Profile zu identifizieren: vier Wochen nach dem Kauf waren vier von fünf Positionen noch online. Selbst nach der Meldung eines Beispiels waren drei Wochen nach der Meldung der Websites noch 95 % der Inhalte online. Angesichts der Entschlossenheit, mit der die Akteure böswillige Inhalte über Fake-Profile oder Troll-Armeen verbreiten, müssen die Plattformen eine proaktive Strategie zur Identifizierung unechter Profile verfolgen, um zu verhindern, dass die Content-Moderation zu einem Kampf gegen Windmühlen wird.

Fake News oder **Junk News** enthalten aus der Luft gegriffene Inhalte, schlichtweg falsche Informationen oder Verschwörungstheorien, die nicht notwendigerweise illegal, in jedem Fall aber schädlich sind. Eine differenziertere Begrifflichkeit unterscheidet zwischen Desinformation und Fehlinformation, wobei erstere mit Absicht und letztere unabsichtlich verbreitet wird. Beide sind geeignet, extremistisches Gedankengut in den Mainstream-Diskurs einzuschleusen, was Radikalisierung fördern und zu realer Gewalt führen kann. Desinformation kann Teil einer politischen Strategie sein und, wenn sie effektiv verbreitet wird (eventuell über die oben beschriebenen Fake-Profile und Bot-Armeen), den öffentlichen Diskurs beeinflussen. So war beispielsweise „Pizzagate“ eine Verschwörungstheorie aus dem amerikanischen Wahlkampf 2016. Nach dem Leaken der privaten E-Mails von John Podesta, dem damaligen Wahlkampfmanager der demokratischen Präsidentschaftskandidatin Hillary Clinton, verbreiteten

32 Cao, S. (2019), „Facebook's AI Chief Explains How Algorithms Are Policing Content – And Whether It Works“ *The Observer*. Abgerufen: https://observer.com/2019/12/facebook-artificial-intelligence-chief-explain-content-moderation-policy-limitation/

33 Harwell, D. (2018), „AI will solve Facebook's most vexing problems, Mark Zuckerberg says. Just don't ask when or how“, *The Washington Post*. Abgerufen: https://www.washingtonpost.com/news/the-switch/wp/2018/04/11/ai-will-solve-facebooks-most-vexing-problems-mark-zuckerberg-says-just-dont-ask-when-or-how/

34 Gupta, S. (2017), „A Quick Guide to Identify Twitterbots Using AI“, Hackernoon. Abgerufen: https://hackernoon.com/a-quick-guide-to-identify-twitterbots-using-ai-c3dc3a7b817f

35 Bay, S. & Fredheim R. (2019), „Falling Behind: How social media companies are failing to combat inauthentic behaviour online“, NATO STRATCOM COE. Abgerufen: https://www.stratcomcoe.org/how-social-media-companies-are-failing-combat-inauthentic-behaviour-online

Gegner die Nachricht, dass man in der Flut von E-Mails einen Code finden könne, der hochrangige Führungskräfte der Demokraten mit Menschenhandel und sexuellem Kindesmissbrauch in Verbindung bringt. Vor allem Rechtsextreme verbreiteten die Theorie während des Wahlkampfes auf den Imageboards 4chan und 8chan sowie auf Reddit und Twitter. Es wurde eine Reihe von Restaurants genannt, die die „Machenschaften" der angeblichen Pädophilen ermöglicht haben sollen. Die Inhaber und Beschäftigten der Restaurants erhielten Drohungen bis hin zu Morddrohungen. Am Ende entschloss sich Edgar Maddison Welch, angestachelt von den Online-Beiträgen, eines der Restaurants aufzusuchen. Er feuerte drei Schüsse ab. Niemand wurde verletzt. In seiner Befragung nach dem Schussangriff stritt er ab, dass es sich bei den Informationen um „Fake News" handle.

Auch wenn Inhalte nicht illegal sind, können Plattformbetreiber Verstöße gegen selbst festgelegte Community-Standards ahnden. Andererseits kann das Herausfiltern von Fake News dem Geschäftsmodell von Social-Media-Unternehmen zuwiderlaufen. Immerhin erhöhen polarisierende, spektakuläre Inhalte das Engagement der Nutzer und lassen sie mehr Zeit auf den Websites verbringen, wodurch die Unternehmen mehr Kundendaten sammeln können. Diese Daten bilden das Rückgrat ihres Finanzmodells, weil zielgerichtete Werbung hiervon profitiert und höhere Erträge erwirtschaftet. Dennoch gibt es Bemühungen von verschiedenen Akteuren, gegen Fake News vorzugehen. Forscher der kanadischen University of Waterloo haben ein KI-basiertes Tool entwickelt, das die Überprüfung von Fakten in einem bislang ungeahnten Ausmaß unterstützen kann. Durch den Abgleich von Aussagen eines Artikels mit anderen Quellen kann das System angeben, ob es sich wahrscheinlich um Fake News handelt oder nicht. Den Forschern zufolge liegt das System in neun von zehn Fällen richtig.[36] Dies könnte im Kampf gegen Fake News ein wichtiger Schritt nach vorn sein.

Das privatwirtschaftliche Projekt Newsguard ist ein Negativbeispiel. Newsguard ist ein Add-on für Webbrowser, das in Form von Bewertungen einen Anhaltspunkt für die Glaubwürdigkeit von Medien liefert. In den sozialen Medien zeigt Newsguard ein kleines Label, um die Vertrauenswürdigkeit von Informationen anzuzeigen. Es handelt sich um eine unkomfortable Lösung: Der Nutzer muss von sich aus das Add-on herunterladen, das dann noch nicht einmal bei der Bewertung konkreter Artikel hilft, sondern nur eine allgemeine Bewertung des Mediums abgibt. Im Fall des beschriebenen Pizzagate-Skandals, der über private Accounts verbreitet wurde, hätte es rein gar nichts ausgerichtet. Breitbart, ein Medium, das rassistische („White Supremacy") und rechtsextreme Ideologien verbreitet, oder der russische Propagandakanal RT werden beide mit einem insgesamt grünen Label bewertet. Dann aber weist Newsguard im Text unter seinem Label darauf hin, dass diese Sites „erhebliche Defizite" hätten. Durch Breitbart geriet auch Facebook in die Kritik: Facebook startete ein „News"-Angebot mit Beiträgen aus verifizierten Medien. Diese Medien wurden in Zusammenarbeit mit Journalisten ausgewählt und halten sich an die internen Richtlinien von Facebook gegen Hassreden und Clickbaiting-Content. Dass auch Breitbart aufgenommen wurde, löste Proteste aus. Bislang verteidigt Facebook diese Entscheidung mit

36 Grant, M. (2019), „New tool uses AI to flag fake news for media fact-checkers", *Waterloo News*. Abgerufen: https://uwaterloo.ca/news/news/new-tool-uses-ai-flag-fake-news-media-fact-checkers

dem Argument der Meinungsfreiheit. Unterdessen gab Twitter im Zuge der britischen Parlamentswahlen 2019 ein Verbot jeglicher politischer Werbung bekannt.[37]

Generell entspricht das Verbot extremistischer Inhalte der Hypothese, dass das Internet zu einem gesünderen Raum werde, wenn die Wahrscheinlichkeit abnimmt, dort potenziell radikalisierende Inhalte anzutreffen. Nichtsdestotrotz kann das Verbot immer nur ein Teil der Antwort auf schädliche Online-Inhalte sein, da es die Ursachen der geäußerten Meinungen außer Acht lässt.

Deepfakes sind eine extreme Form von synthetischen und manipulierten Daten und geben der Redewendung „jemandem Worte in den Mund legen" eine ganz neue Dimension. Die neuesten technologischen Entwicklungen ermöglichen es, Videos mit der Mimik und Stimme realer Menschen zu erstellen. So können Videoinhalte produziert werden, die sehr authentisch aussehen und klingen, obwohl die betreffende Person die wiedergegebenen Worte nie gesagt hat. Ein indischer Politiker nutzte unlängst im Wahlkampf ein Deepfake-Video, um ein gemischtsprachiges Publikum zu erreichen, was nicht nur positive Reaktionen hervorrief.[38] Auch Organisationen, die vor Deepfakes warnen, haben schon Deepfakes verbreitet, wie zum Beispiel das Video, in dem Boris Johnson und sein Gegner Jeremy Corbyn für die Wahl in Großbritannien im Dezember 2019 jeweils eine Wahlempfehlung für den Gegner abgeben. Am weitesten verbreitet sind gefälschte Daten in der Pornographie, was Frauen zu den Hauptopfern dieser neuen Technologie macht. Sie macht es möglich, dass Frauen ohne ihr Wissen oder ihre Zustimmung als Akteurinnen in Pornovideos erscheinen. Besonders hinterhältig erscheint diese Technologie in Kombination mit so genannten **De-Identification**-Tools. Diese Werkzeuge verändern Bilder und Videos auf eine Weise, die es den Algorithmen unmöglich macht, die neue, leicht veränderte Version als das Originalgesicht zu identifizieren, und sie würden es als ein ganz neues Gesicht kategorisieren. Die Nutzer würden jedoch auch in der neuen, geänderten Version das ursprüngliche Gesicht wiedererkennen. Dies könnte ein rasches und zuverlässiges Löschen erschweren. Das von der Branche betriebene Global Internet Forum to Counter Terrorism (GIFCT) hat ein „Hash-Sharing"-Konsortium gegründet – eine Datenbank mit digitalen „Fingerabdrücken" bösartiger Inhalte, so genannter Hashes.[39] Durch die Zusammenarbeit zwischen den Unternehmen will das Forum die Wirksamkeit steigern.[40] Es ist unklar, ob die Datenbank dem systematischen Einsatz von „De-Identifikation"-Software standhalten könnte, zumal extremistische Inhalte strategisch über eine Reihe von Akteuren und über verschiedene Plattformen verbreitet werden.

37 Die Unterscheidung zwischen politischer und thematischer Werbung bleibt aber umstritten, und es gibt keine allgemein akzeptierten Definitionen. Die sich daraus ergebenden Schwierigkeiten führten zu Forderungen, alle Werbebeiträge nach einem einheitlichen, strengen Standard zu behandeln, um die Transparenz zu erhöhen und ihre Wirkung untersuchen zu können. Weitere Informationen: Frederik J. Zuiderveen Borgesius et al. (2018): Online Political Microtargeting: Promises and Threats for Democracy. Utrecht Law Review, 14 (1). 82–96. Abgerufen: https://www.ivir.nl/publicaties/download/UtrechtLawReview.pdf; und Universal Advertising Transparency by default (2020). Abgerufen: https://epd.eu/wp-content/uploads/2020/09/joint-call-for-universal-ads-transparency.pdf

38 Christopher, N. (2020), „We've just seen the First Use of Deepfakes in an Indian Election Campaign", *Vice*. Abgerufen: https://www.vice.com/en_in/article/jgedjb/the-first-use-of-deepfakes-in-indian-election-by-bjp

39 GIFCT, „Joint Tech Innovation". Abgerufen: https://www.gifct.org/joint-tech-innovation/

40 Llansó, E. (2019), „Platforms want centralised Censorship. That should scare you", *Wired*. Abgerufen: https://www.wired.com/story/platforms-centralized-censorship/; und Windwehr, S. und York, Jillian (2020), „One Database to rule them all: The invisible Content Cartel that undermines the freedom of expression online", EFF. Abgerufen: https://www.eff.org/deeplinks/2020/08/one-database-rule-them-all-invisible-content-cartel-undermines-freedom-1.

Realistische Täuschungsmanöver erfordern Expertenwissen – vor allem, wenn die Ergebnisse darauf abzielen, Menschen zu täuschen. Das notwendige technische Wissen verhindert immer noch eine schnelle Ausweitung der Technologie, vor allem wenn es andere, weniger anspruchsvolle Methoden gibt, um dasselbe Ziel zu erreichen. Wie Hwang berichtet, besteht auch die Gefahr der Bloßstellung durch De-Identification-Tools. Verbotsstrategien und das Risiko öffentlicher Bloßstellung kann die Attraktivität von Deepfakes zur Beeinflussung von Kampagnen schmälern.[41] Twitter verwendete seine neuen Labels für manipulierten Content zum ersten Mal für Inhalte, die vom Verantwortlichen für Soziale Medien im Weißen Haus erstellt worden waren.[42] Die Regeln besagen, dass Twitter manipulierte Videos oder Fotos kennzeichnet, aber nicht entfernt, es sei denn, der Content gefährdet die physische Sicherheit einer Person.

Desinformation und künstliches Engagement in den Griff zu bekommen, erfordert eine höhere digitale Kompetenz der Nutzer. Wie Untersuchungen über die Verbreitung von Fake News auf Twitter zeigen, verbreiten sich Unwahrheiten schneller und weiter als die Wahrheit, und der Grund hierfür ist menschliche Interaktion. Bots tragen zur Viralität bei, verursachen aber nicht die weite Verbreitung von Unwahrheiten. Die Forscher machten die emotionale Reaktion und den relativen Neuigkeitscharakter des Inhalts für die Verbreitung verantwortlich.[43] Das Ergebnis zeigt deutlich, dass es keine Alternative zu einer angemessenen Bildung gibt, die Nutzer befähigt, sich mit höherer Resilienz im Internet zu bewegen.

41 Hwang, T. (2020), „Deepfakes – A grounded threat assessment", Centre for Security and Emerging Technology.
42 Dent, S. (2020), „Twitter labels video retweeted by Trump as ‚manipulated data'", Engadget Online. Abgerufen: https://www.engadget.com/2020/03/09/twitter-labels-trump-retweet-manipulated-media/
43 Vosoughi, S. et al. (2018), „The spread of true and false news online", *Science* Bd. 359, Nr. 6380, S.1146–51. Abgerufen: https://science.sciencemag.org/content/359/6380/1146

4 Radikalisierung vorhersagen, bevor sie stattfindet – Allgemeine KI für die Strafverfolgung

Man kann sich leicht den Ort vorzustellen, wo sich das Wunder vollzieht; borgen wir uns zum Beispiel die Darstellung aus dem Science-Fiction-Klassiker *Minority Report*: ein großer blauer Touchscreen, der die Ergebnisse einer superintelligenten Maschine anzeigt, die die Strafverfolgung unterstützen soll. Das angezeigte Ergebnis basiert auf verfügbaren Daten über Einzelpersonen sowie auf dem Online-Verhalten in Echtzeit. Die Warnglocken des Systems schrillen, sobald es kritisch zunehmende Risikofaktoren erkennt, die auf ein gefährliches Maß an Radikalisierung hindeuten. Je nach beobachtetem Verhalten schickt das System dann entsprechende Einsatzteams an den Standort. Dank der Vorhersagekraft des starken KI-basierten Systems ist die Polizei in der Lage einzugreifen, bevor etwas passiert. Dieses Szenario mag verlockend erscheinen, auch wenn es übertrieben ist und mehr mit Science-Fiction zu tun hat als mit der Realität. Dennoch gibt es im Spannungsfeld zwischen neuen Technologien und Sicherheit Interesse an einer solchen Entwicklung. Dieses Kapitel befasst sich ausschließlich mit dem Mythos einer superintelligenten, allgemeinen KI, die Inhalte und das Verhalten von Einzelpersonen online überwacht, um einer Radikalisierung entgegenzuwirken.

Projekte der vorhersagenden Polizeiarbeit (Predictive Policing) untersuchen, wie KI die Strafverfolgungsbehörden bei ihrer Arbeit unterstützen kann. Bei derartigen Projekten geht es um ML-Anwendungen, die auf der Grundlage statistischer Korrelationen Prognosen über zukünftige Verbrechen erstellen, um der Strafverfolgung zuzuarbeiten.[44] Die Wirksamkeit derartiger Systeme ist jedoch heftig umstritten. Die Polizei im britischen Kent gab die Verwendung der amerikanischen Software PredPol für die vorausschauende Polizeiarbeit wieder auf, nachdem deren Nutzen nicht überzeugen konnte.[45] Die Bürgerrechtsorganisation Big Brother Watch berichtete über eine Massenüberwachung von Minderheiten und ein Wiedererstarken der Voreingenommenheit gegenüber Zielbezirken der Verbrechensprävention, weil vermehrte Patrouillen in Gebieten mit historisch höherer Kriminalitätsrate automatisch zu mehr Anzeigen führen und eine systematische Verzerrung schaffen würden.[46] Die vermehrte Anwendung von Indikatoren für die Kriminalitätsprognose setzt interpretatives Denken und nichtkausale Formen des Risikodenkens voraus. Es zeichnet sich eine stärkere

44 Moses, B. L. und Chan, J. (2018), „Algorithmic prediction in policing: assumptions, evaluation, and accountability“, *Policing and Society*, Bd. 28, Nr. 7, S. 806–22.
45 Big Brother Watch Submission to the Centre for Data Ethics and Innovations (2019), „Bias in Algorithmic Decision Making (Crime and Justice)“, Big Brother Watch. Abgerufen: https://bigbrotherwatch.org.uk/wp-content/uploads/2019/06/Big-Brother-Watch-submission-to-the-Centre-for-Data-Ethics-and-Innovation-Bias-in-Algorithmic-Decision-Making-Crime-and-Justice-June-2019.pdf
46 Ebd.

Berücksichtigung des Kontexts bei der Risikoanalyse ab, was als eine Abkehr von der Profilerstellung (Profiling) angesehen werden kann, die in vielen Gesellschaften als ungerecht und diskriminierend empfunden wird.[47] Ein auf Abstammung oder Zugehörigkeit zu einer Volksgruppe beruhendes Profiling belastet die Beziehung zwischen Strafverfolgung und pluralistischen Gesellschaften.[48]

Der indikatorengestützte Ansatz könnte hingegen für Onlinestrategien gegen Radikalisierung attraktiv erscheinen. Eine Liste von Indikatoren auf der Grundlage des Online-Verhaltens sowie der konsumierten Inhalte könnte die Suche nach Personen unterstützen, die dabei sind, sich zu radikalisieren. Eine Übertragung aktueller Systeme für Predictive Policing auf die Arbeit zur Radikalisierungsbekämpfung ist jedoch schwer vorstellbar. Hierfür gibt es drei Hauptgründe.

Die erste Schwierigkeit besteht darin, dass es an Klarheit bzw. einem präzisen Verständnis von Radikalisierungsprozessen mangelt.[49] Unklar ist auch, an genau welchem Punkt ein radikalisiertes Individuum zu einer Straftat übergeht, was ein Eingreifen rechtfertigen würde. Radikalisierung und Terrorismus sind glücklicherweise nicht häufig genug, um eine zuverlässige Datengrundlage zu liefern. Radikalisierung ist ein hochkomplexer und individueller Prozess, und obwohl Forscher bestimmte Elemente identifiziert haben, die in Radikalisierungsprozessen immer wieder auftauchen,[50] gibt es nicht genügend Informationen, um einen Algorithmus zu trainieren. Die derzeitigen KI-Systeme benötigen aber enorme Datenmengen, um eine Vorhersagekraft zu entwickeln. Solange kein technologischer Durchbruch dafür sorgt, dass die KI-Technologie mit deutlich weniger Daten auskommt, ist dieser Ansatz nicht sehr vielversprechend. Derzeit gibt es keine Anzeichen dafür, dass sich das ändern wird.

Aktuelle Predictive-Policing-Systeme beruhen auf gruppenbezogenen Annahmen in Kombination mit Orts- und Zeitdaten zu kriminalitätsgefährdeten Gebieten. Das bedeutet, die Algorithmen werden mit einer Mischung aus Open-Source-Informationen, Regierungsdaten und Daten aus der Privatwirtschaft gefüttert, damit sie zu fundierten Vorhersagen kommen. Die zugrundeliegenden Annahmen gehen dabei von einem hohen Maß an wirtschaftlichen und rationalen Erwägungen aus. Nehmen wir Einbruchsdelikte als Beispiel: Der Einbrecher, der einmal zu einer bestimmten Zeit und an einem bestimmten Ort erfolgreich war, wird wahrscheinlich die nächste Straftat an einem ähnlichen Ort und zu einer ähnlichen Zeit begehen, um seinen Erfolg zu wiederholen. Die Grundannahme ist, dass Kriminelle ihr Risiko minimieren und den Erfolg so weit wie möglich maximieren wollen. Diese Annahmen lassen sich aber nicht ohne Weiteres auf Radikalisierung und Terrorismus übertragen. Das soll nicht heißen, dass es bei der Entscheidung für terroristische Aktivitäten keine rationale Argumentation gibt, sie läuft aber nicht in der gleichen Weise ab wie bei Eigentumsdelikten. So war das „Sterben für den richtigen Zweck“ ein expliziter Pull-Faktor der IS-Propagandakampagne. Mit dem Argument, man sterbe nur einmal

47 Monaghan, J. und Molnar, A. (2016), „Radicalisation theories, policing practices, and ‚the future of terrorism?‘“, *Critical Studies on Terrorism*, Bd. 9, Nr. 3, S. 393–413.
48 Open Society Foundations (2019), „Ethnic Profiling: What it is and Why it must end“, Abgerufen: https://www.opensocietyfoundations.org/explainers/ethnic-profiling-what-it-and-why-it-must-end
49 siehe Monaghan und Molnar.
50 siehe Neumann.

– warum dann nicht als Märtyrer, wurde Gefolgschaft für das „Kalifat“ angeworben.[51]

Der dritte Grund sind die Einschränkungen in liberalen Demokratien, die sich aus der Überzeugung ergeben, dass der Einzelne durch den Staat und vor dem Staat zu schützen ist. Ein Gedankenexperiment: Was wäre nötig, um das Verhalten von Einzelpersonen voraussagen zu können? Ein Algorithmus, der individuelles Verhalten prognostizieren kann, bräuchte weitaus differenziertere Daten als die verfügbaren gruppenbezogenen Informationen, nämlich nicht anonymisierte Daten über das Verhalten von Einzelpersonen. Dabei würde gelten: Je mehr solcher Daten vorliegen, desto zuverlässiger sind die Vorhersagen. Dies würde jedoch eine Überwachung des individuellen Verhaltens in einem noch nie dagewesenen Ausmaß erfordern: die Live-Beobachtung einer ganzen Gesellschaft wäre dazu notwendig. Dies ist mit den bestehenden Persönlichkeitsrechten nicht vereinbar und in einer freien Gesellschaft aus ethischen und moralischen Gründen auch nicht wünschenswert. Sie würde die Grundrechte beschneiden, wie unter anderem die Rede-, Presse-, Vereinigungs- und Versammlungsfreiheit sowie das Fernmeldegeheimnis.[52] Dies wäre zweifelsohne ein dystopisches Szenario.

51 Kingsley, P. (2014), „Who is behind Isis's terrifying online propaganda operation?“, *The Guardian*. Abgerufen: https://www.theguardian.com/world/2014/jun/23/who-behind-isis-propaganda-operation-iraq

52 Ganor, B. (2019), „Artifical or Human: A New Era of Counterterrorism Intelligence?“, *Studies in Conflict and Terrorism*.

Exkurs

Inhärent demokratische KI

Algorithmen funktionieren mit Daten, und ihr Datenhunger ist unersättlich. Zunächst müssen sie mit riesigen Datenbeständen trainiert und dann mit noch mehr Daten getestet werden, um sich schließlich ständig durch noch mehr Daten hindurchzuarbeiten. So drängen sich Themen wie Privatsphäre und Datenschutz unmittelbar auf, insbesondere in Kombination mit KI und Fragen nach der Sicherheit.

Anstatt zu versuchen, Systeme für automatisierte Entscheidungsfindung so zu regulieren, dass sie den Standards demokratischer Gesellschaften genügen, sollten demokratische Werte von vornherein in die Technologie einfließen.
Die technologischen Entwicklungen sollten „inhärent privat“ sein, d. h. Anwenderdaten nach den höchsten Datenschutzstandards behandeln, sofern sich der Anwender nicht von sich aus zur Preisgabe der Informationen bereiterklärt hat. Dies hat Auswirkungen auf das Kuratieren von Inhalten auf Social-Media-Plattformen und die massenhafte Erfassung von Daten zu persönlichen Verhaltensweisen. Darüber hinaus müssen die Systeme transparente Ergebnisse oder Erklärungen liefern, die es menschlichen Bedienern ermöglichen, die Ergebnisse des Algorithmus zu bewerten und über deren Glaubwürdigkeit zu entscheiden. Dies wäre eine klare Alternative zur derzeitigen „Blackbox-KI“, deren Ergebnisse sich der Erklärung entziehen. Höhere Transparenz beispielsweise bei Empfehlungssystemen oder der Kuratierung von Content würde darüber hinaus öffentliches Interesse und Forschung ermöglichen und so zu einem besseren Verständnis von Radikalisierung im Internet führen. Rechenschaftspflicht in Entscheidungsprozessen ist nur mit einer transparenten und vertrauenswürdigen KI zu erreichen. Audits der Anwendungen automatisierter Entscheidungsfindung würden eine rechtmäßige Nutzung gewährleisten und Anreize für eine faire und demokratische KI schaffen.

5 Schlussfolgerungen

Das Ziel dieses Berichts war es, zu erörtern, wie KI-fähige Technologien zur Bekämpfung von Radikalisierung im Internet beitragen können.

KI eröffnet neue Möglichkeiten, Massendaten zu analysieren und Zukunftsprognosen zu treffen. Es gibt Raum für eine eingeschränkte Anwendung der Technologie, um die Prävention von gewalttätigem Extremismus zu unterstützen und das Risiko zu verringern, online auf radikalisierende Inhalte zu stoßen. Besonders vielversprechende KI-basierte Werkzeuge sind Suchmaschinen, Empfehlungssysteme und die Verarbeitung natürlicher Sprachen (Natural Language Processing, NLP). NLP bietet Potenzial für eine verbesserte Content-Moderation im Internet, insbesondere für Sprachen, die nur von kleinen Gruppen von Menschen gesprochen werden. Für größere Plattformen sind die zu erwartenden finanziellen Erträge einer Investition in die Content-Moderation von Minderheitensprachen nicht groß genug – insbesondere wenn hierfür zusätzliches Personal beschäftigt werden muss. Kleinere Plattformen verfügen nicht immer über das technische Know-how oder die Ressourcen für Content-Moderationssysteme, da schon der Einsatz vorhandener Modelle einen erheblichen Zeit- und Arbeitsaufwand verlangt. Wieder andere Plattformen legen das Recht auf Meinungsfreiheit sehr weit aus und argumentieren, dass sie ihre Nutzer in keiner Weise einschränken wollen. Eine verbesserte NLP kann helfen, Inhalte in Sprachen zu übertragen, für die geschulte Moderatoren zur Verfügung stehen. Darüber hinaus kann sie auffällige semantische Muster auf Websites bei Anbietern erkennen, die nicht in Content-Moderation investieren wollen. Derartige Maßnahmen müssen jedoch stets die Datenschutzstandards und die Menschenrechte respektieren.

Content-Moderation für große Social-Media-Plattformen ist und bleibt eine Herausforderung. Die riesige Anzahl von Sprachen in Kombination mit einer bunten Palette an kulturellen Kontexten ist noch immer zu komplex als dass Algorithmen bösartige Inhalte herausfiltern könnten. Es bedarf eines breiten öffentlichen Diskurses über Material in der „Grauzone“, d. h. Inhalte, die zwar legal, aber schädlich sind. Gesellschaften müssen gemeinsam zu einem Konsens kommen, wo die Meinungsfreiheit im Internet an ihre Grenzen stößt. Diese Entscheidung darf nicht allein den Privatunternehmen überlassen werden. Auch ist die KI-Technologie noch nicht weit genug entwickelt und ungeeignet, um künstliches Engagement im Internet, wie Trolle, Bots und Fake News zu bekämpfen. Derartige Mechanismen werden immer noch unzureichend erkannt. Die Nutzer müssen deshalb befähigt und unterwiesen werden, sich im Umgang mit dem Internet verantwortungsbewusst zu verhalten und digitale Souveränität zu erreichen. Die Gestaltung der Plattformen muss ein transparentes „Notice-and-Action“-System zulassen. Dabei darf die Last jedoch nicht allein auf den Schultern der Kunden oder Nutzer liegen. Das Streben nach einem sichereren Internet muss durch politische Maßnahmen flankiert werden, die der Verbreitung bösartiger oder gefälschter Online-Inhalte entgegenwirken und Geschäftsmodelle verbieten, die schädlichem Content Vorschub leisten, weil dieser

das Online-Engagement und damit die Werbeeinnahmen erhöht. Die anspruchsvolle Aufgabe für die Plattformbetreiber wird es sein, nicht zu viele Inhalte aus dem Netz zu nehmen und die Meinungsfreiheit nicht zu stark einzuschränken, aber gleichzeitig angemessene Maßnahmen zur Verhinderung bösartiger Inhalte zu ergreifen.

Wie sich gezeigt hat, ist eine allgemeine KI, ein System mit Superintelligenz, ganz offensichtlich keine Option, um die Online-Radikalisierung von Personen vorherzusagen. Hierfür gibt es zwei Gründe. Der erste ist technischer Natur: Beim derzeitigen Entwicklungsstand der KI-Technologie benötigen die Algorithmen riesige Datenmengen, um brauchbare Vorhersagen für die Zukunft treffen zu können. Glücklicherweise sind Radikalisierung und Terrorismus zu selten, als dass sie genügend Daten für eine allgemeine KI liefern könnten, die eine mögliche Radikalisierung konkreter Personen im Internet vorhersagen könnte. Die Quote der falsch-positiven und falsch-negativen Ergebnisse wäre inakzeptabel. Investitionen in Humanressourcen wären nutzbringender. Der zweite Grund ist der Schutz der Privatsphäre: Systeme, die das Online-Verhalten von Personen in Echtzeit beobachten, die Daten speichern und analysieren, würden den Datenschutzgrundsätzen liberaler Demokratien zuwiderlaufen. Sie könnten potenziell zu einer Überwachungsgesellschaft führen.

Langfristig muss die Anwendung KI-basierter Technologien klaren Standards folgen. Diese Standards müssen die Nutzer vor unfairer automatisierter Entscheidungsfindung schützen, z. B. infolge verzerrter Datengrundlagen oder diskriminierender Content-Kuratierung aufgrund von Geschlecht, sexueller Orientierung, Religion oder anderen durch die Menschenrechte geschützten Merkmalen. Die Ergebniserreichung der Algorithmen muss transparent sein, um die auf den algorithmischen Berechnungen beruhenden Entscheidungen nachvollziehbar zu machen. Der vorgezeichnete Weg für die KI muss eine Entwicklung sein, die das Recht auf Privatsphäre respektiert, frei von Diskriminierung und durch den Betreiber nachvollziehbar.

Die politische Landschaft

Dieser Abschnitt wurde von Armida van Rij und Lucy Thomas, beide Research Associates am Policy Institute des King's College London, verfasst. Er bietet einen Überblick über den politischen Kontext des Berichtsthemas.

Einleitung

Die Verhinderung von Gewaltverherrlichung und Terrorismus, die Verbreitung von Desinformationen und anderen Formen extremistischer Inhalte im Internet sind allgegenwärtige Herausforderungen, vor denen politische Akteure und Technologieplattformen auf der ganzen Welt stehen. *Künstliche Intelligenz und Terrorabwehr: eine Einführung*, ein Bericht für das Global Internet Forum to Counter Terrorism (GIFCT), bietet einen umfassenden Überblick über die Chancen und Risiken der künstlichen Intelligenz (KI) in Bezug auf die Terrorabwehr im Netz.

Nationale und internationale politische Entscheidungsträger sowie Technologieunternehmen stehen unter wachsendem Druck, extremistische Inhalte schneller zu moderieren und zu tilgen. Diese Notwendigkeit ergibt sich unter anderem daraus, dass es durch bösartige Inhalte im Internet bereits zu Vorfällen gekommen ist, die in ihrer Häufigkeit und Tragweite tragische Ausmaße haben, wie das Attentat auf die Moscheen im neuseeländischen Christchurch im März 2019, der Schussangriff auf eine Kirche in Charleston 2015 in den USA und der Angriff auf eine Moschee im Kanadischen Québec im Jahr 2017. In der Folge unternahmen die Technologiebranche sowie die nationalen und internationalen politischen Entscheidungsträger Anstrengungen, um bestimmte extremistische Inhalte zu entfernen, die von der Terrormiliz „Islamischer Staat“ (IS) und gewalttätigen dschihadistischen Gruppen produziert wurden, sowie rassistische (insbesondere „White Supremacy“), frauenfeindliche, antisemitische und islamfeindliche Inhalte.

Künstliche Intelligenz und Terrorabwehr: eine Einführung beschreibt vorhandene KI-Technologien, die zur Unterstützung, Beschleunigung und präzisen Moderation und Entfernung von Online-Inhalten entwickelt wurden. Sie reichen von Werkzeugen, die in der Sprache „geschult“ werden, um schädliche Inhalte zu erkennen und zu kennzeichnen, bis hin zu Technologien, die Deepfake-Videos aufspüren, und den Prozessen des maschinellen Lernens (ML) zur Erstellung von Erkennungsalgorithmen. Der Bericht nennt auch diverse Herausforderungen und Hindernisse für den wirksamen Einsatz dieser Technologien. Zunächst können Empfehlungssysteme, die auf genau derselben KI-Technologie basieren, Nutzer gerade in eine Abwärtsspirale mit zunehmend schädlichem Content führen. Zweitens führt der Umstand, dass sich die Content-Moderation auf eine Auswahl der europäischen Sprachen konzentriert, dazu, dass Inhalte in selteneren Sprachen zu wenig beachtet und zu wenig moderiert werden. Drittens gibt es derzeit keine wirksamen Maßnahmen, um der Desinformation im Internet entgegenzuwirken, weder aus technischer noch aus ethischer Sicht. Zu guter Letzt wäre ein allgemeines

KI-System zur Echtzeitüberwachung des Geschehens im Internet auf ein unmögliches Maß an nicht anonymisierten Daten angewiesen, was das Recht auf den Schutz der Privatsphäre und das Recht auf freie Meinungsäußerung gefährden würde.

In diesem Bericht untersuchen wir den Umgang von neun wichtigen nationalen und supranationalen politischen Akteuren mit diesen Chancen und Risiken: Kanada, Frankreich, Japan, Ghana, Neuseeland, Großbritannien, USA, die Europäische Kommission und das Counter-Terrorism Executive Directorate der Vereinten Nationen. Wir präsentieren einen fallweisen Überblick über die entsprechenden Bemühungen und geben abschließend politische Empfehlungen ab.

Künstliche Intelligenz und Terrorabwehr: Umgang mit den Herausforderungen und Beurteilung neuer Entwicklungen

Kanada

Die kanadische Regierung hat eine robuste Strategie zur Terrorismusabwehr entwickelt, und ihre Bemühungen und -Initiativen zur Terrorabwehr im Internet sind Teil einer umfassenderen, ganzheitlichen Politik zur Terrorismusbekämpfung. Wie bei vielen Regierungen wurden die Investitionen und die Aufmerksamkeit für die Bekämpfung von gewalttätigem Extremismus im Internet leider erst durch reale Katastrophen motiviert.

Ende Januar 2017 eröffnete der gebürtige Québecer Alexandre Bissonnette das Feuer auf das Islamische Kulturzentrum in Québec, wobei sechs Menschen getötet und zahlreiche weitere verletzt wurden. Nachfolgende Untersuchungen ergaben, dass Bissonnette vor den Schüssen in rechtsextremen und rassistischen Online-Kreisen aktiv gewesen war und regelmäßig Twitter-Accounts von Verschwörungstheoretikern, weißen Nationalisten und rechtsextremen Online-Persönlichkeiten wie Ben Shapiro und Alex Jones von InfoWars verfolgt hatte.[53]

Anders als die Täter vieler anderer online motivierter Terroranschläge postete Bissonnette jedoch weder ein Manifest noch eine Absichtserklärung.[54] Trotzdem ist die zunehmende Verbreitung terroristischer Manifeste ein um sich greifender Trend, zu dessen Bekämpfung KI einen Beitrag leisten könnte. Rechtsextreme Manifeste beziehen sich häufig aufeinander, z. B. in Form von Bewunderung für Angriffe aus der früheren oder jüngeren Vergangenheit oder das Zitieren von Memes oder Internet-Abkürzungen. KI könnte dabei helfen, Uploads schädlicher Inhalte, wie beispielsweise rechtsextremer

53 Riga, A. (17. April 2018), „Quebec Mosque Killer Confided He Wished He Had Shot More People, Court Told", *Montreal Gazette*. Abgerufen: https://montrealgazette.com/news/local-news/quebec-mosque-shooter-alexandre-bissonnette-trawled-trumps-twitter-feed/. Siehe auch: Mahrouse, G. (2018), „Minimizing and denying racial violence: Insights from the Quebec Mosque shooting", *Canadian Journal of Women and the Law*, Bd. 30, Nr. 3, S. 471–93.

54 So haben z. B. die Täter Robert Bowers (Attentat in der Synagoge in Pittsburgh 2018), Dylann Roof (Anschlag in einer Kirche in Charleston 2015), Brenton Tarrant (Terroranschlag auf zwei Moscheen in Christchurch 2019), Patrick Crusius (Anschlag in El Paso 2019), Anders Breivik (Massaker von Utøya 2011) und viele andere kurz vor ihren Taten Erklärungen auf verschiedenen Online-Plattformen veröffentlicht. Siehe: Ware, J. (2020), *Testament to Murder: The Violent Far-Right's Increasing Use of Terrorist Manifestos* – Policy Brief, International Centre for Counter-Terrorism – The Hague. Abgerufen: https://icct.nl/publication/testament-to-murder-the-violent-far-rights-increasing-use-of-terrorist-manifestos/

Manifeste, zu erkennen, um eine Intervention zu ermöglichen, bevor Angriffe in der realen Welt stattfinden.

Wie in der National Strategy on Countering Radicalization to Violence[55] dargelegt, reagiert Kanada auf gewalttätigen Extremismus im Internet mit einer dreigleisigen Strategie: die Entwicklung von Botschaften gegen Extremismus in der Zivilgesellschaft (engl. „counter-messaging"), die Unterstützung der Forschung zur Terrorabwehr und die Partnerschaft mit internationalen Initiativen und Technologieunternehmen. Im Rahmen der dritten Strategiekomponente hat Kanada in den Anknüpfungspunkt von KI und Terrorabwehr investiert.

Insbesondere hat Kanada 2019 Tech Against Terrorism, eine internationale, von der UNO geförderte Initiative, die mit der globalen Technologiebranche zusammenarbeitet, mit der Entwicklung der Terrorist Content Analytics Platform (TCAP)[56] beauftragt, einer Datenbank, die verifiziertes terroristisches Material und Inhalte aus bestehenden Datenbeständen und offenen Quellen enthält. Man hofft, dass die TCAP als Echtzeit-Warndienst für terroristische und gewalttätige extremistische Inhalte auf kleineren Internet-Plattformen fungieren kann: verifizierte bösartige Inhalte auf diesen Plattformen werden dann rasch der Content-Moderation zugeführt und ggf. gelöscht. Mittel- und langfristig soll die TCAP als historisches Archiv für eine Verbesserung der quantitativen und qualitativen wissenschaftlichen Analyse fungieren.[57]

Speziell im KI-Bereich ist eines der erklärten Ziele der TCAP die Unterstützung eines Ökosystems algorithmischer Content-Klassifikatoren.[58] Wie der GNET-Einführungsbericht, *Künstliche Intelligenz und Terrorabwehr*, zeigt, „benötigen die Algorithmen riesige Datenmengen, um brauchbare Vorhersagen für die Zukunft treffen zu können".[59] Mechanismen der automatisierten Content-Moderation, die auf maschinellem Lernen (ML) und der Verarbeitung natürlicher Sprachen (Natural Language Processing, NLP) basieren, stützen sich auf die Analyse von Massendaten (Big Data). Sie dienen dazu, KI dahingehend zu trainieren, dass sie Daten als das erkennen, wofür sie stehen, und die Elemente von IS-Propagandavideos (Logos, Flaggen usw.) identifizieren, damit zukünftige Videos mit gleichen oder ähnlichen Elementen gefunden und markiert werden können. Die TCAP ist als erste einheitliche Plattform für terroristische Inhalte im Internet eine überaus wichtige Daten- und Informationsressource für Programmierer, die ML-Algorithmen zur Identifizierung und Klassifizierung terroristischen Materials entwickeln.

Indem die TCAP verifizierte terroristische Inhalte von verschiedenen Internet-Plattformen als historisches Archiv zur Verfügung stellt, könnte sie einen bedeutenden technologischen Fortschritt bei der Bekämpfung des gewalttätigen Extremismus im Internet bewirken.

---

55 „National Strategy on Countering Radicalization to Violence", Public Safety Canada. Abgerufen: https://www.publicsafety.gc.ca/cnt/rsrcs/pblctns/ntnl-strtg-cntrng-rdclztn-vlnc/index-en.aspx#s7

56 Die Terrorist Content Analytics Platform (TCAP) wurde auch im Abschnitt „Die politische Landschaft" des GNET-Berichts „Hass decodieren: Klassifizierung terroristischer Inhalte mittels experimenteller Textanalyse" beschrieben. Abgerufen: https://gnet-research.org/wp-content/uploads/2020/09/GNET-Report-Decoding-Hate-Using-Experimental-Text-Analysis-to-Classify-Terrorist-Content_GERMAN.pdf

57 „Press Release: Tech Against Terrorism Participates in UN General Assembly Week in New York", Tech Against Terrorism. Abgerufen: https://www.techagainstterrorism.org/2019/10/08/press-release-tech-against-terrorism-participates-in-un-general-assembly-week-in-new-york/

58 Ebd.

59 Schroeter, M. (2020), „Künstliche Intelligenz und Terrorabwehr: eine Einführung", Global Network on Extremism and Technology, S. 28.

Die Regierung Kanadas hat als Co-Sponsor der Plattform gezeigt, wie gezielte und kluge Investitionen in sektorübergreifende Initiativen Gelegenheit für die Zusammenarbeit von Wissenschaft, Industrie und Zivilgesellschaft schaffen können, um KI im Hinblick auf die Terrorabwehr voranzubringen.

Europäische Kommission

In ihrem KI-Weißbuch vom Februar 2020 merkt die Europäische Kommission an, dass „mithilfe von KI-Tools EU-Bürger u. U. besser vor Verbrechen und terroristischen Anschlägen geschützt werden können".[60] Die EU will bei der Nutzung von KI ein duales Konzept verfolgen: ausgerichtet auf Regulierung und Finanzierung. Insbesondere geht es ihr darum, eine „vertrauenswürdige KI" zu ermöglichen, indem sie einen soliden Rechtsrahmen zum Schutz der europäischen Bürgerinnen und Bürger einrichtet und „zur Schaffung eines reibungslos funktionierenden Binnenmarkts" für die Weiterentwicklung der KI beiträgt.[61] „Vertrauenswürdige KI bedeutet in diesem Fall: technisch solide und präzise Systeme".[62] Die EU beabsichtigt außerdem, die Investitionen in KI bis 2030 auf mindestens 20 Mrd. € pro Jahr zu erhöhen.[63]

Die Europäische Kommission ernannte 2019 eine High-Level Expert Group on Artificial Intelligence (AI HLEG). Diese Gruppe hat sieben Voraussetzungen festgelegt, die eine vertrauenswürdige KI gewährleisten sollen. Diese sieben Voraussetzungen sind: Vorrang menschlichen Handelns und menschlicher Aufsicht; Robustheit und Sicherheit; Privatsphäre und Datenqualitätsmanagement; Transparenz; Vielfalt, Nichtdiskriminierung und Fairness; gesellschaftliches und ökologisches Wohlergehen sowie Rechenschaftspflicht.[64] Vor diesem Hintergrund verlangt das Weißbuch der Kommission ein Ökosystem für Vertrauen, um den Schutz der Grundrechte zu gewährleisten.[65]

Die Reaktionen der großen Technologieunternehmen auf das KI-Weißbuch waren uneinheitlich. Google forderte die EU auf, bestehende Regelungen und rechtliche Rahmenbedingungen auszuschöpfen, anstatt neue rechtliche Rahmenbedingungen zu schaffen, an die sich Technologieunternehmen halten müssen. Parallel dazu müssen sich Google, Facebook und andere Technologieplattformen auf das „Gesetz für digitale Dienste" (Digital Services Act) vorbereiten, das noch in diesem Jahr erwartet wird und darauf abzielt, „das Online-Ökosystem in vielen Bereichen zu regulieren, darunter ... anstößige Inhalte".[66] Des Weiteren wird erwartet, dass die EU ihrem KI-Weißbuch im weiteren Verlauf des Jahres 2020 Rechtsvorschriften zu KI und Sicherheit, Haftung, Grundrechten und Umgang mit Daten folgen lassen wird.[67]

---

60 Europäische Kommission, (19. Februar 2020), „Weißbuch zur Künstlichen Intelligenz – Ein europäisches Konzept für Exzellenz und Vertrauen", S. 2. Abgerufen: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_de.pdf
61 Ebd., S. 10.
62 Ebd., S. 20.
63 Französische Regierung, Ministerium für Europa und Äußeres, „Transparency and accountability: The challenges of artificial intelligence". Abgerufen: https://www.diplomatie.gouv.fr/en/french-foreign-policy/digital-diplomacy/transparency-and-accountability-the-challenges-of-artificial-intelligence/
64 Ebd.
65 Ebd.
66 Stolton, S. (23. Juni 2020), „Platform clamp down on hate speech in run up to Digital Services Act", *EURACTIF*. Abgerufen: https://www.euractiv.com/section/digital/news/platforms-clamp-down-on-hate-speech-in-run-up-to-digital-services-act/
67 Kayali, L., Heikkila, M. und Delcker, J. (19. Februar 2020), „Europe's digital vision, explained", *Politico*. Abgerufen: https://www.politico.eu/article/europes-digital-vision-explained/

Frankreich

In Frankreich gibt es mehrere Schlüsselakteure, deren Aufgabenbereich die KI umfasst. Der Koordinator einer nationalen Strategie für künstliche Intelligenz ist mit der Analyse und Entwicklung von Vorschlägen für Änderungen im Zusammenhang mit digitalen Innovationen im Sicherheitsbereich beauftragt.[68] Innerhalb des Verteidigungsministeriums gibt es ebenfalls eine Koordinierungsstelle für künstliche Intelligenz im Verteidigungsbereich, die Teil der Agentur für Verteidigungsinnovation ist.

Frankreich hat seinen Rechtsrahmen angepasst, um eine sichere und effizientere Nutzung von KI-fähigen Technologien zum Schutz der französischen Bevölkerung zu ermöglichen. Was die Politik betrifft, so hat Frankreich im März 2018 seine KI-Strategie veröffentlicht. Ihre Hauptziele sind: die Verbesserung des Bildungs- und Ausbildungsumfelds für KI, um die besten KI-Talente zu fördern und anzuziehen, die Einführung einer Open-Data-Richtlinie für die Implementierung von KI-Anwendungen und die gemeinsame Nutzung von Ressourcen sowie die Entwicklung eines ethischen Rahmens für eine transparente und faire Nutzung von KI-Anwendungen.[69] In Übereinstimmung mit der EU-Richtlinie über die Sicherheit von Netzen und Informationssystemen hat Frankreich ein eigenes Gesetz zur Cybersicherheit formuliert.[70] Frankreich systematisiert derzeit seine Überlegungen zur Nutzung KI-fähiger Technologien für militärische Zwecke.[71]

Frankreich hat eine Reihe von Initiativen zur KI gestartet. Auf der G7-Multistakeholder-Konferenz zur künstlichen Intelligenz im Jahr 2018 kündigten Frankreich und Kanada die Einrichtung eines internationalen Gremiums für künstliche Intelligenz an, das die verantwortungsvolle Einführung der KI unterstützen soll.[72] Darüber hinaus waren sie gemeinsam federführend bei der Gründung der neuen Initiative Global Partnership on Artifical Intelligence (GPAI), der sich weitere Länder angeschlossen haben. Die Initiative soll eine verantwortungsvolle Weiterentwicklung und Nutzung der KI unter Berücksichtigung von Menschenrechten, Inklusion, Vielfalt, Innovation und Wirtschaftswachstum lenken.[73] Konkret wird sie dazu dienen, die Kluft zwischen Theorie und Praxis der KI zu überbrücken, indem sie die Forschung zu KI-bezogenen Aktivitäten unterstützt. In Frankreich wird die GPAI durch ein Kompetenzzentrum unterstützt, einer Schwesterinstitution des GPAI-Kompetenzzentrums im kanadischen Montreal. Die GPAI wird außerdem von der OECD unterstützt.

Auf der Grundlage der KI-Strategie könnte Frankreich die Einführung des Gesetzes für die „Digitale Republik" ins Auge fassen, das dazu dienen würde, „öffentliche Daten zu öffnen, den Schutz der Nutzerrechte und den Datenschutz zu stärken und sicherzustellen, dass die Chancen der Digitalisierung allen zugute kommen".[74]

---

68 Französische Regierung, Büro des Premierministers, (13. Juli 2018), „Action plan against terrorism", S. 20. Abgerufen: http://www.sgdsn.gouv.fr/uploads/2018/10/20181004-plan-d-action-contre-le-terrorisme-anglais.pdf

69 Europäische Kommission, *France AI strategy report*. Abgerufen: https://ec.europa.eu/knowledge4policy/ai-watch/france-ai-strategy-report_en

70 Französische Regierung, Nationale Agentur für Cybersicherheit, *Directive network and information system security (NIS)*. Abgerufen: https://www.ssi.gouv.fr/entreprise/reglementation/directive-nis/

71 Pannier, A. und Schmitt, O. (2019), „To fight another day: France between the fight against terrorism and future warfare", *International Affairs* Bd. 95, Nr. 4. Abgerufen: https://academic.oup.com/ia/article/95/4/897/5492774

72 Kanadische Regierung, Büro des Premierministers (6. Dezember 2018), *Mandate for the International Panel of Artificial Intelligence*. Abgerufen: https://pm.gc.ca/en/news/backgrounders/2018/12/06/mandate-international-panel-artificial-intelligence

73 Französische Regierung, Ministerium für Europa und Äußeres (15. Juni 2020), *Launch of the Global Partnership on Artificial Intelligence by 15 founding members*. Abgerufen: https://www.diplomatie.gouv.fr/en/french-foreign-policy/digital-diplomacy/news/article/launch-of-the-global-partnership-on-artificial-intelligence-by-15-founding

74 Europäische Kommission, *France AI strategy report*.

Ghana

Die Anstrengungen zur Abwehr von gewalttätigem Extremismus im Internet sind begrenzt, weil die politisch motivierte Gewalt im Land, im Gegensatz zu den Nachbarstaaten Nigeria und Tschad, nicht durch terroristische Aktivitäten angeheizt wurde.[75] Die Global Terrorism Database, eine Datenbank der weltweiten Terroranschläge seit 1970, listet in Ghana in 50 Jahren nur 21 Vorfälle mit 23 Todesopfern auf.[76]

Was die Abschaltung des Internets durch die Regierung oder die Ausnutzung sozialer Medien zur Unterdrückung politisch Andersdenkender betrifft, ist Ghana nicht von den gleichen Problemen betroffen wie einige der Nachbarländer.[77] Deren Regierungen haben aus der Kolonialzeit stammende Gesetze, die historisch zur Beschneidung von Freiheiten verwendet wurden, dazu ausgenutzt, um „viele ... Versuche zu legitimieren, außergesetzliche Forderungen an den privaten Sektor zu stellen".[78] Der Bericht „Ranking Digital Rights" 2019 zeigt, dass Social-Media-Plattformen und Internetdienstanbieter auf außergesetzliche Forderungen mit Abschaltdrohungen der Regierungen reagieren mussten, was den Verdacht übermäßiger Überwachung und Zensur aufkommen lässt.[79]

Auch wenn die ghanaische Regierung bisher keine derartigen illegalen Forderungen gestellt hat, haben zivilgesellschaftliche Gruppen und Journalisten Besorgnis über die Zukunft geäußert.[80] Vor den Wahlen 2016 hatte der ghanaische Polizeichef eine mögliche Abschaltung der sozialen Medien angekündigt.[81] Obwohl sich der Präsident derartigen Plänen widersetzte, wächst in Ghana die Sorge um die digitalen Rechte.

Die liberalen Gesetze zur Meinungsfreiheit in Ghana lassen digitale Räume offen für Missbrauch wie Hassreden und Cyberbullying (insbesondere gegen Frauen gerichtet).[82] Die Forderungen nach einer strengeren Regulierung von Social-Media-Plattformen nehmen daher zu. Ein Experte der Freedom of Expression Media Foundation für Westafrika gibt zu bedenken: „Wenn es keine Regulierung gibt, werden andere Gesetze herangezogen werden, um Menschen in unter Umständen überzogener Weise zu verfolgen", ähnlich wie bei den oben erwähnten Regierungsforderungen.

Die staatlichen Regulierungen sozialer Medien müssen jedoch ein Gleichgewicht zwischen dem Schutz der Nutzer vor Schaden und dem Schutz der freien Meinungsäußerung finden. Eine prominente zivilgesellschaftliche Gruppe, die gegen Internetabschaltungen

75 Dank an Tomiwa Ilori, Forscher an der Expression, Information and Digital Rights Unit am Centre for Human Rights der Universität Pretoria, für diese Informationen per E-Mail.
76 Global Terrorism Database, START. Abgerufen: https://www.start.umd.edu/gtd/
77 Ilori, T. (2020), „Content Moderation Is Particularly Hard in African Countries", Information Society Project at Yale Law School. Abgerufen: https://law.yale.edu/isp/initiatives/wikimedia-initiative-intermediaries-and-information/wiii-blog/moderate-globally-impact-locally-content-moderation-particularly-hard-african-countries
78 Ilori, T. (2020), „Stemming digital colonialism through reform of cybercrime laws in Africa", Information Society Project at Yale Law School. Abgerufen:https://law.yale.edu/isp/initiatives/wikimedia-initiative-intermediaries-and-information/wiii-blog/stemming-digital-colonialism-through-reform-cybercrime-laws-africa
79 Ranking Digital Rights, „2019 RDR Corporate Accountability Index". Abgerufen: https://rankingdigitalrights.org/index2019/assets/static/download/RDRindex2019report.pdf
80 Majama, K. (2019), „Africa in urgent need of a homegrown online rights strategy", Association for Progressive Communications. Abgerufen: https://www.apc.org/en/news/africa-urgent-need-homegrown-online-rights-strategy
81 Olukotun, D. (16. August 2019), „President of Ghana says no to internet shutdowns during coming elections", *AccessNow*. Abgerufen: https://www.accessnow.org/president-ghana-says-no-internet-shutdown-elections-social-media/
82 Endert, J. (2018), „Digital backlash threatens media freedom in Ghana", DW Akademie. Abgerufen: https://www.dw.com/en/digital-backlash-threatens-media-freedom-in-ghana/a-46602904

kämpft, hat vor einer staatlichen Regulierung der sozialen Medien gewarnt: „Sobald man der Regierung erlaubt, das Internet zu regulieren – und da gibt es Beispiele aus anderen Ländern – wird diese Regierung vorgeben, wie man das Internet zu nutzen hat.“[83]

Es ist derzeit nicht klar, ob es in Ghana Pläne zur Entwicklung von KI-basierten Instrumenten zur Unterstützung der Regulierung von Online-Inhalten gibt. Nichtsdestotrotz haben in der Region Bedrohungen der Meinungsfreiheit durch Gesetze aus der Kolonialzeit gezeigt, dass das Land bei allen technischen Werkzeugen oder gesetzgeberischen Bemühungen zur Überwachung des Internets auf schädliche Inhalte den digitalen Rechten seiner Bürgerinnen und Bürger oberste Priorität einräumen muss. In einem begrüßenswerten Schritt hat Ghana 2019 ein Gesetz über das Recht auf Information verabschiedet, das den Zugang zu Informationen im Besitz öffentlicher Institutionen garantiert.[84] Der Gesetzentwurf signalisiert, dass die ghanaische Regierung mit digitalen Rechten transparent und verantwortungsbewusst umgehen will. Alle zukünftigen Entwicklungen im Bereich Content-Moderation sollten diesen Verpflichtungen und Standards folgen, um den Schutz personenbezogener Daten und das Recht auf freie Meinungsäußerung zu gewährleisten.

Japan

Japan kanalisiert seine Aktivitäten zur Terrorabwehr größtenteils über den Verband Südostasiatischer Nationen (Association of Southeast Asian Nations, ASEAN).[85] Bereits 2004 gaben die ASEAN-Mitgliedsländer in Partnerschaft mit Japan eine Reihe von Erklärungen zur Zusammenarbeit bei der Bekämpfung des internationalen Terrorismus ab. Über die politische Absichtserklärung hinaus verpflichtete die Erklärung die Unterzeichner zur „Verhütung, Unterbindung und Bekämpfung des internationalen Terrorismus durch Informationsaustausch, gemeinsame Nutzung nachrichtendienstlicher Erkenntnisse und Aufbau von Kapazitäten“ und schuf damit ein Vorbild für die regionale Zusammenarbeit bei der Bekämpfung von gewalttätigem Extremismus und Terrorismus.[86]

Im Jahr 2015 bekräftigte Japan sein Bekenntnis zur multinationalen Zusammenarbeit in Südostasien bei der Abwehr von gewalttätigem Extremismus und Terrorismus und zur Zusammenarbeit bei der Umsetzung des ASEAN-Aktionsplans zur Verhinderung und Bekämpfung der Zunahme von Radikalisierung und gewalttätigem Extremismus (2018–2025).[87] Der Aktionsplan setzt auf die Partnerschaft „mit der Wirtschaft und dem Technologiesektor bei der Förderung von Moderation und der Intensivierung des Dialogs zur Prävention von Radikalisierung und gewalttätigem Extremismus“ und stärkt die „strategische Kommunikation“ zur Verhinderung

83 Ebd.
84 Yahya Jafru, M. (26. März 2019), „Right to information – RTI bill passed into law“, *Graphic Online*. Abgerufen: https://www.graphic.com.gh/news/politics/ghana-news-rti-bill-passed.html
85 „Japan: Extremism & Counter Extremism“, Counter-Extremism Project. Abgerufen: https://www.counterextremism.com/countries/japan
86 „ASEAN-Japan Joint Declaration for Cooperation to Combat International Terrorism“, ASEAN. Abgerufen: https://asean.org/?static_post=asean-japan-joint-declaration-for-cooperation-to-combat-international-terrorism-2
87 „Chairman's Statement of the 18th ASEAN-Japan Summit, Kuala Lumpur, November 22 2015“, ASEAN. Abgerufen: https://www.asean.org/wp-content/uploads/2015/12/6.-Chairmans-Statement-of-the-18th-ASEAN-Japan-Summit-Final-Final.pdf; „Japan's cooperation with ASEAN 2025 (Counter-Terrorism)“, Mission of Japan to ASEAN. Abgerufen: https://www.asean.emb-japan.go.jp/asean2025/jpasean-ps03.html

des Missbrauchs sozialer Medien durch gewalttätige Extremisten und Terroristen.[88]

Die Olympischen Spiele und Paralympics 2020 (wegen der Coronavirus-Krise auf 2021 verschoben) werden in Tokio stattfinden. Die Ausrichtung der Olympischen Spiele wird traditionell als „Test" der Sicherheitsfähigkeiten einer Nation angesehen. Insofern bieten die Spiele auch Gelegenheit, Innovationen in den Bereichen KI, Security und Strafverfolgung zu erproben.[89]

In einem entsprechenden Pilotprojekt vor den Spielen im Jahr 2018 kündigte die Polizei der Präfektur Kanagawa die Einführung eines Predictive-Policing-Systems an, um Verbrechen und Angriffe mittels eines Deep-ML-Algorithmus vorherzusagen.[90] Inzwischen haben führende Technologieunternehmen die Bereitstellung umfangreicher Systeme für Gesichtserkennung, biometrische Authentifizierung und Verhaltenserkennung bei den Spielen sowie in Häfen und an Flughäfen bestätigt.[91] Diese Systeme werden in der Lage sein, Gesichter auf bestimmte Emotionen zu scannen und die Identität anhand von Gesichtszügen und personenbezogenen Informationen zu kontrollieren.

Es steht noch unklar, ob sich diese KI-Sicherheitsfunktionen auch auf Social Media und Online-Aktivitäten erstrecken werden. Die in Kanagawa erprobten Systeme hätte Berichten zufolge auch die Überwachung von Social-Media-Inhalten zur Verbrechensabwehr beinhalten können. Dies ließe sich dahingehend interpretieren, dass die japanischen Strafverfolgungsbehörden die KI-gestützte Überwachung sozialer Medien zur Bekämpfung bösartiger oder potenziell gefährlicher Inhalte im Internet begrüßen. Ein solcher Schritt wäre allerdings riskant angesichts der Empörung im Jahr 2017, als Kritiker dem umstrittenen Antiterrorgesetz der Regierung eine Gefährdung der bürgerlichen Freiheiten anlasteten.[92] Japan wird dem Schutz der Rechte seiner Bürgerinnen und Bürger auf Privatsphäre und Meinungsfreiheit bei der Entwicklung von KI-Technologien zur Bekämpfung von gewalttätigem Extremismus im Internet und in der Realität ernsthafte Aufmerksamkeit widmen müssen.

Neuseeland

Die Lenkung der Bekämpfung von gewalttätigem Extremismus im Internet erfordert in Neuseeland die Koordination zahlreicher Stellen und Gremien. Dazu gehören das Cabinet External Relations and Security Committee, die Polizei, der Geheimdienst und Behörden für Sicherheitskommunikation sowie die Behörden

88 „2018 ASEAN Plan of Action to Prevent and Counter the Rise of Radicalisation and Violent Extremism (2018–2025), adopted in Myanmar, October 31 2018", ASEAN. Abgerufen: https://cil.nus.edu.sg/wp-content/uploads/2019/10/2018%20ASEAN%20Plan%20of%20Action%20to%20Prevent%20and%20Counter%20the%20Rise%20of%20Radicalisation%20and%20Violent%20Extremism%20(2018-2025).pdf

89 Siehe zum Beispiel: Soria, V. (2011), „Beyond London 2012: The Quest for a Security Legacy," *The RUSI Journal*, Bd. 156, Nr. 2, S. 36–43.

90 „Kanagawa police to launch AI-based predictive policy system before Olympics", *Japan Times* (29. Januar 2018). Abgerufen [Paywall]: https://www.japantimes.co.jp/news/2018/01/29/national/crime-legal/kanagawa-police-launch-ai-based-predictive-policing-system-olympics/

91 Japanische Regierung (2019), „All is Ready for a Safe and Secure Tokyo Games". Abgerufen: https://www.japan.go.jp/tomodachi/2019/autumn-winter2019/tokyo2020.html; „NEC Becomes a Gold Partner for the Tokyo 2020 Olympic and Paralympic Games", NEC Corporation (2015). Abgerufen: https://www.nec.com/en/press/201502/global_20150219_01.html

92 „Japan passes controversial anti-terror conspiracy law", *BBC* (15. Juni 2017). Abgerufen: https://www.bbc.co.uk/news/world-asia-40283730

für auswärtige Angelegenheiten, Handel, Verteidigung, Verkehr, Innovation und Entwicklung. Neuseelands Gesamtstrategie ist in dem im Februar 2020 veröffentlichten Strategieplan zur Terrorismusbekämpfung beschrieben.[93]

Nach dem Attentat in den Moscheen in Christchurch im März 2019 brachten die Regierungen Neuseelands und Frankreichs unter dem Christchurch Call to Eliminate Terrorist and Violence Extremist Content Online („Christchurch-Appell") eine Koalition von Staatsoberhäuptern mit Social-Media- und Technologieunternehmen zusammen.[94] Der Aufruf verpflichtet die unterstützenden Länder, Gesetze durchzusetzen, die die Verbreitung terroristischer und gewalttätiger extremistischer Inhalte im Internet verbieten und gleichzeitig die internationalen Menschenrechtsgesetze einschließlich des Rechts auf freie Meinungsäußerung respektieren. Die Länder arbeiten auch daran, Rahmenbedingungen, Kapazitätsaufbau und Sensibilisierungsmaßnahmen zu unterstützen, um der Nutzung von Online-Diensten zur Verbreitung terroristischer und gewalttätiger extremistischer Inhalte entgegenzuwirken.

Der Christchurch-Appell verpflichtet auch Unternehmen wie Amazon, Facebook, Google, Twitter und YouTube zu mehr Rechenschaftspflicht und Transparenz in der Branche. Die Unternehmen müssen ihre Community-Standards und Nutzungsbedingungen durchsetzen, indem sie Maßnahmen zur Content-Moderation und Entfernung von Inhalten Priorität einräumen und Inhalte in Echtzeit zur Überprüfung und Bewertung identifizieren. Gemeinsam entwickeln die Länder und Unternehmen mit der Zivilgesellschaft Maßnahmen, um von der Community ausgehende Aktivitäten zu fördern und so in die Prozesse der Online-Radikalisierung einzugreifen.

Nach dem Attentat vom März 2019 wurde eine Untersuchungskommission (Royal Commission of Inquiry) eingesetzt, um die Reaktion der Behörden auf die Schussangriffe zu bewerten und zu ermitteln, welche zusätzlichen Maßnahmen ergriffen werden können, um künftige Angriffe zu verhindern.[95] Der Bericht der Kommission wird die aktuelle Strategie zur Terrorabwehr und die zukünftige Ausrichtung in Neuseeland beleuchten und wertvolle Erkenntnisse darüber liefern, inwieweit KI Teil dieser zukünftigen Strategie sein soll. Aufgrund der Coronavirus-Krise verzögert sich die Veröffentlichung des Berichts bis zum Winter 2020.

Die neuseeländische Regierung verpflichtet sich auch bei der Anwendung von Algorithmen für die Regierungsführung strengeren Standards für Transparenz und Rechenschaftspflicht. Wie *Künstliche Intelligenz und Terrorabwehr: eine Einführung* beschreibt, kann die Anwendung von Algorithmen bereits bestehende Verzerrungen verstärken.[96] Im Juli 2020 veröffentlichte die Regierung die Algorithm Charter for Aotearoa New Zealand, eine umfassende Überprüfung

93 Neuseeländische Regierung, Officials' Committee for Domestic and External Security Coordination, Counter-Terrorism Coordination Committee (Februar 2020), „Countering terrorism and violent extremism national strategy overview". https://dpmc.govt.nz/sites/default/files/2020-02/2019-20 CT Strategy-all-final.pdf

94 Siehe https://www.christchurchcall.com/

95 The Royal Commission of Inquiry into the Attack on Christchurch mosques. Siehe: https://christchurchattack.royalcommission.nz/

96 Siehe auch Babuta, A. und Oswald, M. (2019), „Briefing Paper: Data Analytics and Algorithmic Bias in Policing," RUSI. Abgerufen: https://www.gov.uk/government/publications/report-commissioned-by-cdei-calls-for-measures-to-address-bias-in-police-use-of-data-analytics; Benjamin, R. (2019), *Race After Technology: Abolitionist Tools for the New Jim Code* (Polity); Benjamin, R., „A New Jim Code?", Berkman Klein Center for Internet & Society at Harvard University. Abgerufen: https://cyber.harvard.edu/events/new-jim-code

des staatlichen Einsatzes von Algorithmen in Bereichen von Verkehr bis Justiz, und eine Verpflichtung zu mehr Transparenz, Einbindung von Interessengruppen, Stakeholder Engagement, Maßnahmen zum Schutz der Privatsphäre und zur menschlichen Aufsicht über den Einsatz von Algorithmen.[97] Die Charta – weltweit die erste ihrer Art – ist zum Zeitpunkt dieser Niederschrift von fünfundzwanzig staatlichen Stellen unterzeichnet.

Wesentlich aber ist, dass die für Terrorabwehr im Internet zuständigen Stellen und Behörden noch fehlen. Insofern bleibt es unklar, inwieweit die politischen Entscheidungsträger in Neuseeland die Weiterentwicklung von KI und algorithmischen Werkzeugen zur Bekämpfung bösartiger Inhalte im Internet erwägen und an welche Standards diese gebunden sein sollen. Die Charta signalisiert einen Schritt in eine positive Richtung, und die Anwendung derartiger Standards auf KI-basierte Maßnahmen zur Terrorabwehr wäre eine begrüßenswerte Entwicklung in der Politik.

Vereinigtes Königreich

Im Februar 2018 kündigte Großbritannien die Entwicklung eines auf maschinellem Lernen (ML) basierenden algorithmischen Werkzeugs an, um terroristische Inhalte des IS im Internet aufzuspüren. Die Software wurde auf die Erkennung und Kennzeichnung audiovisueller Elemente in IS-Propagandainhalten trainiert – Flaggen, Logos, Formatierungen, Strukturen und Tonspuren. Große Technologieplattformen wie YouTube und Facebook haben im Laufe der Jahre stark in die Entwicklung eigener automatisierter Tools zur Moderation von Inhalten investiert. Das Tool wurde so konzipiert, dass es plattformunabhängig ist und daher als Open Source durch kleinere Internet- und Social-Media-Plattformen wie Vimeo genutzt werden kann.

Trotz des vielversprechenden Ansatzes ist das Instrument in seiner Wirksamkeit jedoch stark begrenzt und wird uneinheitlich aufgenommen. Erstens reichen, wie unser Kollege Charlie Winter betonte, die Online-Inhalte des IS von Videos bis hin zu Fotos, schriftlichen Abfassungen und Radiobeiträgen. Wenn auch die Bekämpfung von Video-Inhalten ein positiver Schritt ist, „wird sie [das Problem] bestenfalls etwas entschärfen, aber sie ist noch weit von einer Lösung entfernt“.[98] Zweitens gab das britische Innenministerium (Home Office) das Tool in Auftrag, um die explizitesten und schockierendsten IS-Videos zu erkennen. Das Softwareentwicklungsunternehmen, das das Tool entwickelt hat, erklärte, dass es „weniger um den Umfang als vielmehr darum gegangen sei, für wie effektiv es [das Home Office] die Lösung in Bezug auf bestimmte Arten von Videos gehalten habe“.[99] In vielen akademischen Studien wurden jedoch die weitreichenden Auswirkungen gerade der „weicheren“ propagandistischen Inhalte und ihres radikalisierenden Potenzials über lange Zeiträume beschrieben.[100] Die gleich dreifache Einschränkung der KI – IS, Video-Content und extreme Inhalte – steht der technischen

97 „Algorithm charter for Aotearoa New Zealand“, data.govt.nz. Abgerufen: https://data.govt.nz/use-data/data-ethics/government-algorithm-transparency-and-accountability/algorithm-charter
98 Temperton, J. (13. Februar 2018), „ISIS could easily dodge the UK's AI-powered propaganda blockade“, *Wired*. Abgerufen: https://www.wired.co.uk/article/isis-propaganda-home-office-algorithm-asi
99 Ebd.
100 „Hashtag Terror: How ISIS Manipulates Social Media“, Anti-Defamation League (21. August 2014). Abgerufen: https://www.adl.org/education/resources/reports/isis-islamic-state-social-media

Wirksamkeit des Tools im Wege. Das Tool wurde kleineren Technologieunternehmen kostenlos zur Verfügung gestellt. Allerdings plante mit Stand April 2019 noch kein Unternehmen eine Einführung des Tools.[101]

Das Vorhaben der britischen Regierung, KI zur Bekämpfung von gewalttätigem Extremismus im Internet einzusetzen, macht außerdem potenzielle Interessenkonflikte in der Zusammenarbeit zwischen Regierungen und Industrie deutlich. Das im April 2019 veröffentlichte Online Harms White Paper der Regierung begründet ausführlich, warum eine stärkere nationale Regulierung der sozialen Medien notwendig sei.[102] Dieser neue Rechtsrahmen erlegt Social-Media- und Technologieunternehmen eine neue gesetzliche Sorgfaltspflicht gegenüber ihren Nutzern auf, die über die britische Medienaufsichtsbehörde (Office of Communications, Ofcom) durchsetzbar ist. Bei Nichteinhaltung des rechtlichen Rahmens und Verstößen gegen die gesetzliche Sorgfaltspflicht verhängt Ofcom finanzielle und technische Strafen über die Plattformen – Websites könnten auf ISP-Ebene gesperrt und mit Bußgeldern von bis zu 4 % ihres weltweiten Umsatzes belegt werden.[103] Bei der Ankündigung des algorithmischen Werkzeugs im Februar 2018 signalisierte die damalige Innenministerin Amber Rudd, dass Unternehmen möglicherweise gesetzlich zu dessen Verwendung verpflichtet werden könnten.

An und für sich geben derartige regulatorische Schritte keinen Anlass zu Bedenken. Das Datenanalyse- und Softwareentwicklungsunternehmen, das das Tool entwickelt hat, ehemals ASI Data Science (jetzt unter dem Namen Faculty bekannt), war jedoch mit der Datenmodellierung in den Vote-Leave- und Leave.EU-Kampagnen beauftragt worden und als solches in den Cambridge-Analytica-Skandal verwickelt.[104] Darüber hinaus hat das Unternehmen bis Mai 2020 innerhalb von achtzehn Monaten mindestens sieben öffentlich finanzierte Regierungsaufträge erhalten und unterhält beachtenswerte persönliche und geschäftliche Verbindungen zu Dominic Cummings, dem Chefberater des Premierministers.[105]

Diese Tatsachen geben Anlass zu Bedenken hinsichtlich eines möglichen Interessenkonflikts. Das Vertrauen der Öffentlichkeit und der Wirtschaft in das Instrument wird untergraben, wenn der entsprechende Entwicklungsauftrag an eine Firma geht, die enge Verbindungen zu zentralen Regierungskreisen unterhält und in einen öffentlichen Skandal verwickelt ist. Wenn man sich noch dazu für eine gesetzlich verpflichtende Nutzung des Instruments durch die Social-Media-Plattformen einsetzt, die dem Unternehmen laufende Umsätze bescheren würde, erscheinen die regulatorischen Bestrebungen in einem zweifelhaften Licht.

---

101 Murgia, M. und Bond, D. (6. April 2019), „Businesses show no appetite for anti-terror AI tool", *Financial Times*. Abgerufen [Paywall]: https://www.ft.com/content/fda2d218-56fb-11e9-91f9-b6515a54c5b1

102 Britische Regierung (April 2019), „Online Harms White Paper". Abgerufen: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf

103 Crawford, A. (29. Juni 2020), „Online Harms bill: Warning over ‚unacceptable' delay", *BBC*. Abgerufen: https://www.bbc.co.uk/news/technology-53222665

104 Cadwalladr, C. (7. Mai 2017), „The great British Brexit robbery: how our democracy was hijacked", *The Guardian*. Abgerufen: https://www.theguardian.com/technology/2017/may/07/the-great-british-brexit-robbery-hijacked-democracy

105 Evans, R. und Pegg, D. (4. Mai 2020), „Vote Leave AI firm wins seven government contracts in 18 months", *The Guardian*. Abgerufen: https://www.theguardian.com/world/2020/may/04/vote-leave-ai-firm-wins-seven-government-contracts-in-18-months; Pegg, D., Evans, R. und Lewis, P. (12. Juli 2020), „Revealed: Dominic Cummings firm paid Vote Leave's AI firm £260,000", *The Guardian*. Abgerufen: https://www.theguardian.com/politics/2020/jul/12/revealed-dominic-cummings-firm-paid-vote-leaves-ai-firm-260000; Pegg, D. und Evans, R. (2. Juni 2020), „AI firm that worked with Vote Leave given new coronavirus contract", *The Guardian*. Abgerufen: https://www.theguardian.com/technology/2020/jun/02/ai-firm-that-worked-with-vote-leave-wins-new-coronavirus-contract

Eine von der Regierung unabhängige Entwicklungsfirma hätte ein technisch besser wirksames Werkzeug entwickeln können, das umfassende Anforderungen erfüllt (z. B. mehr als nur eine Teilmenge von IS-Videoinhalten erkennt) und sowohl das Vertrauen in das Werkzeug als auch seine Akzeptanz erhöht. Transparenz und Rechenschaftspflicht sind „nicht bloß Schlagworte, die mit Lippenbekenntnissen abgehakt werden können: sie sind maßgeblich für den Erfolg der Problemlösungsansätze“ zur Bekämpfung von gewalttätigem Extremismus im Internet mittels KI.[106] Das Vereinigte Königreich hat eine wichtige politische Chance vertan, modernste KI-Technologie zur Moderation schädlicher Online-Inhalte zu entwickeln und bereitzustellen, indem es der Vertrauenswürdigkeit des Werkzeugs geschadet und sein technisches Potenzial beschnitten hat.

Counter-Terrorism Committee Executive Directorate der Vereinten Nationen

Das Counter-Terrorism Committee Executive Directorate der Vereinten Nationen (UN CTED) wurde vom Sicherheitsrat der Vereinten Nationen mit der Resolution 1535 (2004) eingerichtet, um als Expertengremium das Counter-Terrorism Committee des Sicherheitsrats zu unterstützen.[107] Sein anfängliches Ziel bestand darin, die Implementierung von Resolutionen des Sicherheitsrats zur Terrorismusbekämpfung durch die UN-Mitgliedstaaten zu bewerten und diese Bemühungen im Wege eines Dialogs zu unterstützen. Das UN CTED steht in enger Zusammenarbeit mit dem Sicherheitsrat sowie den großen Technologieunternehmen und Organisationen der Zivilgesellschaft im GIFCT.

Gegenwärtig gibt es mehrere Resolutionen des Rats der Vereinten Nationen zum Missbrauch des Internets für terroristische Zwecke, und das UN CTED bemüht sich um mehr Kohärenz und Rationalisierung der Schnittmenge zwischen den Resolutionen des UN-Sicherheitsrats und den Aufgaben der IT-Seite. Resolution 2129 (2013) des Sicherheitsrats verweist auf die sich weiterentwickelnde Verflechtung von Terrorismus und IT sowie auf die Nutzung von Technologien wie dem Internet zur Begehung und Förderung von Terrorakten, indem sie die Anstiftung, Anwerbung, Geldbeschaffung oder Planung solcher terroristischer Handlungen ermöglichen.[108] Diese Resolution bekräftigt zudem das Mandat des UN CTED. Die Resolutionen 2354 (2017), 2395 (2017) und 2396 (2017) rufen die Mitgliedstaaten der Vereinten Nationen auf, durch Kooperation untereinander sowie mit dem privaten Sektor und der Zivilgesellschaft zu verhindern, dass IT und das Internet von Terrororganisationen missbraucht werden.[109] Die Resolution 1373 des Sicherheitsrats fordert die Mitgliedstaaten der Vereinten Nationen auf, den „Austausch operativer Informationen“ über den Einsatz von IT durch terroristische Organisationen weiterzuentwickeln und zu beschleunigen und die Rekrutierung von Terroristen zu stoppen.[110]

106 „Tackling the Information Crisis: A Policy Framework for Media System Resilience“, The Report of the LSE Commission on Truth Trust & Democracy, S. 32. Abgerufen: https://www.lse.ac.uk/media-and-communications/assets/documents/research/T3-Report-Tackling-the-Information-Crisis-v6.pdf

107 Chowdhury Fink, N. (2012), „Meeting the challenge: A guide to United Nations counterterrorism activities“, *International Peace Institute*, S. 45. https://www.ipinst.org/wp-content/uploads/publications/ebook_guide_to_un_counterterrorism.pdf

108 UN, Security Council Counter-terrorism Committee, (14. September 2018), „Public-private efforts to address terrorist content online: A year of progress – what's next?“. Abgerufen: https://www.un.org/sc/ctc/news/event/public-private-efforts-address-terrorist-content-online-year-progress-whats-next/; Global Initiative Against Transnational Organised Crime, *Responding to terrorist use of the internet* (21. Januar 2019). Abgerufen: https://globalinitiative.net/terrorist_use_internet/

109 Vereinte Nationen, Security Council Counter-Terrorism Committee, 2018.

110 Global Initiative Against Transnational Organised Crime, 2019.

Das High-Level Panel on Digital Cooperation des Generalsekretärs der Vereinten Nationen sucht nach Lösungen zur Minderung von KI-Risiken.[111] Die Empfehlung 3C des Gremiums lautet:

> *„Wir sind der Überzeugung, dass autonome intelligente Systeme so gestaltet werden müssen, dass ihre Entscheidungen nachvollzogen werden können und der Mensch für ihre Nutzung verantwortlich ist. Audits und Zertifizierungssysteme müssen die Konformität der KI-Systeme mit technischen und ethischen Standards überwachen, die im Rahmen von Multi-Stakeholder- und multilateralen Ansätzen aufgestellt werden müssen. Entscheidungen über Leben und Tod dürfen nicht an Maschinen delegiert werden. Wir rufen zu einer verstärkten digitalen Zusammenarbeit mit verschiedenen Interessengruppen auf, um die Gestaltung und Anwendung dieser Standards und Prinzipien wie Transparenz und Unvoreingenommenheit in autonomen intelligenten Systemen in verschiedenen sozialen Umfeldern zu durchdenken.“*[112]

Einer der Schwerpunkte liegt auf dem Schutz der Menschenrechte im digitalen Zeitalter.[113]

USA

Die Terrorabwehrstrategie der Vereinigten Staaten identifiziert die Abwehr im Internet als einen Schwerpunktbereich und verpflichtet sich zur Zusammenarbeit mit der Wirtschaft und IT-Branche, um die Anwerbung von Terroristen, Geldbeschaffung und Radikalisierungsprozesse im Internet zu bekämpfen. Was länderübergreifende Initiativen betrifft, so arbeiten die USA mit Initiativen wie Tech Against Terrorism und dem Global Counterterrorism Forum zusammen, das in Partnerschaft mit anderen Unterzeichnern, der Zivilgesellschaft und dem Technologiesektor Konzepte zur mittel- und langfristigen Bekämpfung von gewalttätigem Extremismus im Internet entwickelt.

Was die innenpolitische Gesetzgebung in den USA angeht, so lösten Berichte über russische Einmischung und Medienmanipulation bei den Präsidentschaftswahlen 2016 Forderungen nach einer Regulierung der sozialen Medien und Technologieplattformen aus. Im selben Zeitraum sind die Social-Media-Unternehmen in Bezug auf die Nutzerzahlen und zugehörigen Dienstleistungen und Produkte weiter gewachsen. Ende 2019 hielten der Bankenausschuss des US-Senats und der Energie- und Handelsausschuss des Kongresses Anhörungen zu dem von Facebook vorgeschlagenen Kryptowährungsdienst Libra ab. Die Anhörungen boten den Gesetzgebern in den USA die Gelegenheit, Facebook-Führungskräfte zu Manipulation und Missbrauch der Plattform[114] zu befragen und eine „Big-Tech“-Regulierung als geeignete legislative Option ins Spiel zu bringen.[115]

---

111 Französische Regierung, Ministerium für Europa und Äußeres, „Transparency and accountability: The challenges of artificial intelligence“.

112 Vereinte Nationen (16. Dezember 2019), High-level Panel Follow-up Roundtable 3C – Artificial Intelligence – Meeting note. Abgerufen: https://www.un.org/en/pdfs/HLP%20Followup%20Roundtable%203C%20Artificial%20Intelligence%20-%201st%20Session%20Summary.pdf

113 Vereinte Nationen, Secretary-General’s High-level Panel on Digital Cooperation. Abgerufen: https://www.un.org/en/digital-cooperation-panel/

114 US House of Representatives Committee on Energy and Commerce, „Facebook: Transparency and Use of Consumer Data,“ Niederschrift vom 11. April 2018, S. 33. Abgerufen: https://docs.house.gov/meetings/IF/IF00/20180411/108090/HHRG-115-IF00-Transcript-20180411.pdf

115 Molla, R. und Stewart, E. (2019), „How 2020 Democrats think about breaking up Big Tech“, *Vox*. Abgerufen: https://www.vox.com/policy-and-politics/2019/12/3/20965447/tech-2020-candidate-policies-break-up-big-tech

Während der Kongress gesetzgeberische Maßnahmen erwägt, hat die Geheimdienst-Community der USA den Einsatz von KI zur Bekämpfung von gewalttätigem Extremismus im Internet vorangetrieben. Im Frühjahr und Sommer 2019 wurden die USA von einer Serie von Massenmorden erschüttert, deren Täter eine umfangreiche Vorgeschichte im Umfeld von gewalttätigem Extremismus im Internet gemeinsam hatten. Der Täter beispielsweise, der Ende April 2019 in der Synagoge im kalifornischen Poway auf Anwesende schoss, hatte kurz vor dem Angriff eine Erklärung auf 8chan gepostet. Hierin bezieht er sich auf andere durch Aufmerksamkeit im Internet mitmotivierte Schusswaffenangriffe, wie im Fall der Moscheen in Christchurch und in der Synagoge in Pittsburgh, sowie auf typische Online-Persönlichkeiten und Quellen des rechtsextremen und weiß-nationalistischen Umfelds.

In diesem Zusammenhang startete das Federal Bureau of Investigation (FBI) eine Ausschreibung, die private Auftragnehmer zur Entwicklung einer Technologie aufrief, die dem FBI „nahezu in Echtzeit Zugang zu einem vollständigen Spektrum von Social-Media-Plattformen“ verschaffen und es ermöglichen sollte, “eine stetig wachsende Vielfalt von Bedrohungen der nationalen Interessen der USA aufzudecken, zu unterbinden und zu untersuchen“.[116] Ein ähnliche Ausschreibung wurde im Januar 2020 veröffentlicht.[117] Als im Juni 2020 die #BlackLivesMatter-Proteste das gesamte Land überzogen, verlängerte das FBI seine Verträge mit Dataminr, einem Unternehmen zur Überwachung und Analyse sozialer Medien, und Venntel, einem Unternehmen für Ortungsdaten.[118]

Diese Technologien und der Datenzugriff wären eine parallele zu einem allgemeinen KI-System, wie in Abschnitt vier von *Künstliche Intelligenz und Terrorabwehr: eine Einführung* dargelegt, ein vorausschauendes System, das es den Strafverfolgungsbehörden ermöglicht, auf der Grundlage eines Warnmechanismus einzugreifen. Derartige Instrumente würden eine beachtliche ethische Bedrohung für die Persönlichkeitsrechte der Nutzer darstellen, da die Echtzeitüberwachung des individuellen Verhaltens für die Strafverfolgung auf nicht anonymisierten Daten beruhen würde. Die Erfassung von Personen zuordenbaren Daten würde das Recht auf persönliche Sicherheit, Schutz der Identität und Meinungsfreiheit untergraben.

Der Ansatz, nach dem die USA künstliche Intelligenz zur Bekämpfung von gewalttätigem Extremismus im Internet einsetzen wollen, zeigt die rechtlichen und ethischen Herausforderungen, die mit der Verfolgung und Moderation von Online-Material verbunden sind. Wie Marie Schroeter schreibt, wäre dies „zweifelsohne ein dystopisches Szenario“.[119]

116 US Government Federal Acquisitions Service, „Contract Opportunity: Social Media Alerting Subscription.“ Abgerufen: https://beta.sam.gov/opp/b6de554012cf4ab9ab795f52c638467c/view

117 US Government Federal Acquisitions Service, „Request for Proposal – FBI Social Media Alerting.“ Abgerufen: https://beta.sam.gov/opp/2b3003e9b0b34b639687786e8420013b/view

118 US Government Federal Acquisitions Service, „Contract Information – Dataminr. Inc.“ Abgerufen: https://beta.sam.gov/entity/962138942?keywords=Dataminr&sort=-relevance&index=&is_active=true&page=1&status=active; Fang, L. (24. Juni 2020), „FBI Expands Ability to Collect Cellphone Location Data, Monitor Social Media, Recent Contracts Show“, *The Intercept*. Abgerufen: https://theintercept.com/2020/06/24/fbi-surveillance-social-media-cellphone-dataminr-venntel/

119 Schroeter, M. (2020), „Künstliche Intelligenz und Terrorabwehr: eine Einführung“, Global Network on Extremism and Technology, S. 25.

Empfehlungen zur Politik

Bestehende Initiativen und Aktionen, wie die oben beschriebenen, liefern Erkenntnisse und Empfehlungen für politische Entscheidungsträger in aller Welt. Auf der Grundlage unserer Ergebnisse geben wir die folgenden politischen Empfehlungen ab:

Empfehlung 1: Einrichtung einer unabhängigen Regulierungsbehörde auf transnationaler Ebene zur Aufsicht über die nationalen Bemühungen zur Bekämpfung des gewalttätigen Extremismus im Internet mithilfe künstlicher Intelligenz

Gesetze der Regierungen, die Strafen für Social-Media-Unternehmen vorsehen, die es versäumen, gegen schädliche Inhalte vorzugehen,[120] können sehr wirksam sein,[121] bergen jedoch die Gefahr, dass das Recht der Bürgerinnen und Bürger auf freie Meinungsäußerung eingeschränkt wird, da die Furcht vor Strafen zu übertrieben weit gehender Entfernung von Inhalten führen kann. Wie oben beschrieben, besteht im Falle Großbritanniens, Japans und der USA auch ein Risiko, dass die Bemühungen um Content-Moderation und die Gesetzgebung in rechtliche und ethische Konflikte im Zusammenhang mit Datenschutz, Vertrauenswürdigkeit und Rechenschaftspflicht geraten.

Die Selbstregulierung der sozialen Medien, wonach die Unternehmen ihre eigenen Standards, Kodizes und Richtlinien zur Entfernung bösartiger Inhalte im Netz erstellen und durchsetzen, kann funktionieren, aber die Anwendbarkeit der Standards kann uneinheitlich und intransparent sein.[122] Viele große Unternehmen veröffentlichen zwar High-Level-Daten zur Content-Moderation, aber hierzu besteht keine Verpflichtung.[123]

Um in den Bereichen Rechenschaftspflicht, Transparenz und Ethik die Einhaltung von Standards zu gewährleisten, sollte eine gemeinsame Regulierung durch Regierung, Zivilgesellschaft und IT-Branche etabliert werden, die von einem unabhängigen, transnationalen Gremium beaufsichtigt wird. Ein unabhängiges Gremium, das sich globalen Standards zum Schutz der Privatsphäre verpflichtet hat,[124], das über Durchsetzungsmechanismen verfügt und von den Regierungen unabhängig ist, würde die beschriebenen Konflikte abmildern.

Eine gemeinsame Regulierung durch Regierung, Social-Media-Plattformen und Zivilgesellschaft würde sicherstellen, dass im Vordergrund der Regulierungsbemühungen die Interessen der Nutzer stehen. Eine Regierungsgesetzgebung, die das Regulierungsgremium

120 So sieht beispielsweise das deutsche Netzdurchsetzungsgesetz (NetzDG) von 2017 Strafen von bis zu 50 Mio. € für Social-Media-Plattformen vor, die illegale Inhalte nicht innerhalb von vierundzwanzig Stunden entfernen. Siehe: http://wp.ceps.eu/wp-content/uploads/2018/11/RR%20No2018-09_Germany's%20NetzDG.pdf

121 Elhai, W. (2020), „Regulating Digital Harm Across Borders: Exploring a Content Platform Commission“, *SMSociety'20: International Conference on Social Media and Society*, https://doi.org/10.1145/3400806.3400832, S. 223–4.

122 See Matsakis, L. (2. März 2018), „YouTube Doesn't Know Where Its Own Line Is“, *Wired*. Abgerufen: https://www.wired.com/story/youtube-content-moderation-inconsistent/

123 Ebd.

124 Wie z. B. die Global Network Initiative. Siehe https://globalnetworkinitiative.org/

autorisiert, schützt sich vor wechselnden politischen Interessen und gewährleistet, dass ihre Durchsetzungsmechanismen funktionieren. Indem die Plattformen verpflichtet werden, sich an die Mechanismen zu halten, werden die Moderationsbemühungen gleichmäßiger und gerechter angewendet. Schutz der Privatsphäre, Redefreiheit und Rechenschaftspflicht sollten die Grundlage der Werte dieses Gremiums bilden und dessen Lenkungsform bestimmen.

Empfehlung 2: Einbeziehung von Maßnahmen zur Bekämpfung algorithmischer Verzerrungen direkt bei der Software-Entwicklung

Die Grundsätze und die Praxis der Inhaltsmoderation und Kuratierung durch Online-Plattformen beinhalten in den seltensten Fälle Input aus der Öffentlichkeit oder Rechenschaftspflichten.[125] Algorithmen, die in der Technologiebranche oft „hinter verschlossenen Türen" entwickelt werden und großen Einfluss auf die Online-Erfahrung von Milliarden von Nutzern haben, können extrem einseitig sein. Algorithmen „lernen, indem sie mit bestimmten Bildern gefüttert werden, die oft von technischen Fachleuten ausgewählt werden", einer Personengruppe, in der männliche Weiße überrepräsentiert sind.[126] Entsprechende Verzerrungen haben schon zu gravierenden Problemen in der realen Welt geführt, so eine Software, die dunkelhäutige Angeklagte mit einer größeren Wahrscheinlichkeit einer Wiederholungstäterschaft assoziiert,[127] und die Google-Photos-App, die Fotos eines dunkelhäutigen Nutzers versehentlich als Fotos von Gorillas kategorisiert hat.[128]

Im Zusammenhang mit der Terrorabwehr bedeutet eine entsprechende algorithmische Verzerrung, dass nicht westliche extremistische Inhalte unterschätzt und untermoderiert werden. Da die größten Technologieunternehmen der Welt ihren Sitz im Westen haben, ist davon auszugehen, dass „den Fachleuten und Führungskräften, die für die Entwicklung von Technologieprodukten verantwortlich sind, die Katalysatoren von Gewalt und Diskriminierung in anderen Kulturen als der eigenen nicht vertraut sind".[129] Die Lage in Myanmar zeigt auf drastische Weise, was eine nicht adäquate Content-Moderation und Tilgung nicht westlicher bösartige Inhalte bedeutet. Online-Hassreden in burmesischer Sprache gegen die Volksgruppe der Rohingya haben zu weit verbreiteter Gewalt angestachelt. Doch die Eskalation des rassistischen Hasses gegen die Minderheit blieb weitgehend ungebremst, weil Facebook lediglich zwei Burmesisch sprechende Content-Moderatoren beschäftigte.[130]

125 „Tackling the Information Crisis: A Policy Framework for Media System Resilience", The Report of the LSE Commission on Truth Trust & Democracy, S. 18. Abgerufen: https://www.lse.ac.uk/media-and-communications/assets/documents/research/T3-Report-Tackling-the-Information-Crisis-v6.pdf

126 Crawford, K. (25. Juni 2016), „Artificial Intelligence's White Guy Problem", *New York Times. Abgerufen [Paywall]:* https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html

127 Angwin, J. et al. (23. Mai 2016), „Machine Bias", *ProPublica.* Abgerufen: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

128 Nieva, R. (1. Juli 2015), „Google apologizes for algorithms mistakenly calling black people ‚gorillas'", *CNET.* Abgerufen: https://www.cnet.com/news/google-apologizes-for-algorithm-mistakenly-calling-black-people-gorillas/

129 Elhai, S. 221.

130 Stecklow, S. (2018), „Special Report: Why Facebook is losing the war on hate speech in Myanmar", *Reuters.* Abgerufen: https://www.reuters.com/article/us-myanmar-facebook-hate-specialreport/special-report-why-facebook-is-losing-the-war-on-hate-speech-in-myanmar-idUSKBN1L01JY

Um derartigen Ungleichheiten zu begegnen, können Technologieplattformen in der Zivilgesellschaft und akademischen Welt vorhandene Expertise nutzen, um sie in die Softwareentwicklungsphase einfließen zu lassen. Die Unternehmen sollten umfangreiche und regelmäßige Audits zur Verwendung von Algorithmen durchführen. Die Ergebnisse dieser Prüfungen sollten der Öffentlichkeit zugänglich gemacht werden, um der Rechenschaftspflicht, der Transparenz und der Vertrauenswürdigkeit in den Augen der Öffentlichkeit Genüge zu tun.[131]

Der Ausbau der Fähigkeiten zur Inhaltsmoderation – in sprachlicher und geografischer Hinsicht bei Mitarbeiterteams sowie bei der Entwicklung nicht westlicher KI-Tools – ist für die Social-Media- und Technologieunternehmen eine kostspielige Investition. Allerdings können Bestrebungen im Hinblick auf diese dringend benötigten Erweiterungen innerhalb der Technologiebranche die Chance bieten, sich als Branchenführer für Content-Moderation und Beseitigungsstrategien zu positionieren.

Empfehlung 3: Finanzierung von Publikationen, die die Technologie, Herausforderungen und Chancen der künstlichen Intelligenz in klarer und zugänglicher Sprache darstellen, durch nationale und multinationale Akteure und Initiativen

Wie *Künstliche Intelligenz und Terrorabwehr: eine Einführung* feststellt, gibt es rund um das Thema KI eine erhebliche Menge an Hype und Unsachlichkeit. Selbst viele politische Entscheidungsträger haben keine korrekte Vorstellung davon, was KI ist und was sie leisten kann. Darüber hinaus „zeigten parlamentarische Diskussionen und Ausschussanhörungen im Vereinigten Königreich nach dem Cambridge-Analytica-Skandal 2018, im US-Kongress und im Europäischen Parlament unter hochrangigen Parlamentariern und politischen Entscheidungsträgern ein erschreckend niedriges Niveau der Medienkompetenz und des Medienverständnisses“.[132]

Unzureichende Kenntnisse der digitalen und medialen Umgebung sowie der Möglichkeiten und Grenzen der KI kann zu politischen Fehlentscheidungen führen. Weiterbildung der politischen Entscheidungsträger in diesem Bereich ist vonnöten, damit sie fundierte, ausgewogene und gut informierte politische Entscheidungen treffen können.

Nationale und multinationale Initiativen sollten die regelmäßige Erstellung und Veröffentlichung von Orientierungshilfen zum Wesen der KI im Auge haben und welche Herausforderungen und Chancen sie für die Politik birgt. Zivilgesellschaftliche und akademische Experten für schädliche Inhalte im Internet und deren Folgen in bestimmten Kontexten sollten außerdem reaktive Leitfäden für neu auftretende und künftige Bedrohungen erstellen. So würden beispielsweise Experten, die mit einer bevorstehenden umstrittenen

131 Turner Lee, N. (2019), „Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms“, Brookings Institute. Abgerufen: https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

132 „Tackling the Information Crisis: A Policy Framework for Media System Resilience“, The Report of the LSE Commission on Truth Trust & Democracy, S.38. Abgerufen: https://www.lse.ac.uk/media-and-communications/assets/documents/research/T3-Report-Tackling-the-Information-Crisis-v6.pdf

Wahl in einem nicht westlichen Kontext vertraut sind, für politische Entscheidungsträger ein Briefing zum Kontext und den Katalysatoren für schädliche Inhalte und über die möglichen Konsequenzen in der realen Welt erstellen.

Derartige Orientierungshilfen und Briefings müssen in klarer und verständlicher Sprache verfasst sein die den Fachjargon oder sensationslüsternen Hype rund um das Thema KI durchbricht. Auf diese Weise könnten politische Entscheidungsträger und Laien gleichermaßen zum öffentlichen Diskurs über künstliche Intelligenz und Terrorabwehr beitragen.

Global Network
on Extremism & Technology

KONTAKTANGABEN

Im Falle von Fragen oder zur Anforderung weiterer Exemplare wenden Sie sich bitte an:

ICSR
King's College London
Strand
London WC2R 2LS
United Kingdom

T. **+44 20 7848 2098**
E. **mail@gnet-research.org**

Twitter: **@GNET_research**

Wie alle anderen GNET-Publikationen kann auch dieser Bericht kostenlos von der GNET-Website unter www.gnet-research.org heruntergeladen werden.