KI im Zeitalter des Misstrauens

Lahmann, H.C.

ETHIK UND MILITÄR

KONTROVERSEN IN MILITÄRETHIK UND SICHERHEITSPOLITIK

AUSGABE 01/2024

KI und Autonomie in Waffen: Kriege und Konflikte außer Kontrolle?

SPECIAL

Militär, Mensch, Maschine

zebis

10 JAHRE ETHIK UND MILITÄR

KI UND AUTONOMIE IN WAFFEN: KRIEGE UND KONFLIKTE AUSSER KONTROLLE?

SPECIAL: MILITÄR, MENSCH, MASCHINE

KI IM ZEITALTER DES MISSTRAUENS

Autor: ***Henning Lahmann***

Abstract

Der Begriff „kognitive Kriegsführung" bezeichnet die böswillige Nutzung von Informationen zur Manipulation von Zielgruppen in offenen, demokratischen Gesellschaften und gilt aktuell als eine der dringendsten politischen Herausforderungen. Durch die digitale Transformation nehmen Staaten wie Russland oder China inzwischen über soziale Medien und andere digitale Kommunikationskanäle direkten Einfluss auf westliche Wählerschaften. Mit der Entwicklung bahnbrechender KI-Anwendungen zur Erzeugung und Verbreitung von Text und synthetischen Medien wird sich die Problematik in naher Zukunft voraussichtlich noch verschärfen. Inzwischen arbeitet die Forschung bereits an Konzepten für KI-gestützte „Frühwarnsysteme" für die kognitive Kriegsführung. Diese Systeme sind in der Lage, mithilfe der neuesten Lernalgorithmen Desinformationskampagnen gegnerischer Akteure zu erkennen, zu überwachen und sogar abzuwehren. Bislang hat die umfangreiche Forschungstätigkeit in den Kognitions- und Sozialwissenschaften nur spärliche Erkenntnisse über die Kausalzusammenhänge irreführender Informationen und das Ausmaß der damit verbundenen Gefährdung ergeben. Solange sich dies nicht ändert, bedrohen derlei Eingriffe jedoch kommunikationsbezogene Rechte wie die Meinungs- und Informationsfreiheit, ohne westliche Gesellschaften tatsächlich widerstandsfähiger gegenüber den Versuchen der gegnerischen Einflussnahme zu machen. Ohne eine solide Beweisgrundlage kann die fortgesetzte Versicherheitlichung und Externalisierung des Problems potenziell schädlicher Äußerungen im Internet dazu führen, dass wir genau die Rechte und Werte opfern, die solche technischen Gegenmaßnahmen eigentlich schützen sollen.

„Russland hat gerade den Ausgang der Wahlen in einem europäischen Land beeinflusst", titelt kürzlich das Magazin *Foreign Policy* aufgeregt.[1] In dem Artikel beschreibt der in Berlin lebende Journalist Paul Hockenos, wie der slowakische Präsidentschaftswahlkampf von einer Flut „antiukrainischer und prorussischer Desinformationen" beeinflusst wurde. Diese könnten letztlich „der Hebel gewesen sein (...), der das Ergebnis so dramatisch gedreht hat" und zum Sieg des prorussischen Kandidaten Peter Pellegrini führte. Betrachtet man die böswillige Einflussnahme des Kremls auf die westliche Politik über einen längeren Zeitraum hinweg, so stellen die Wahlen in der Slowakei nur den jüngsten Fall in einer Reihe von Kampagnen dar. Die Einmischung Russlands in die demokratischen Entscheidungsprozesse Europas und Nordamerikas begann spätestens 2016 mit den schockierenden Ergebnissen des Brexit-Referendums im Vereinigten Königreich und Trumps überraschendem Sieg bei den US-Präsidentschaftswahlen. Diese Ereignisse markieren den Beginn einer völlig neuen Ära hybrider Bedrohungen. Ihr Ziel ist es, das Bewusstsein der westlichen Wählerschaften zu vergiften. Bisher konzentrierte man sich auf die russischen Bemühungen, Misstrauen gegenüber jeglicher Information zu säen und die Polarisierung in den offenen westlichen Gesellschaften voranzutreiben. Doch in jüngster Zeit tritt ein weiterer potenter Akteur auf den Plan: Die Regierung in Peking hat großes Interesse, scheinbar erfolgreiche, koordinierte Informationsoperationen gegen dieselben Ziele und nach ähnlichem Muster durchführen. Wie die *New York Times* Anfang April unter Berufung auf Wissenschaftler und Regierungsbeamte berichtete, wendet China teilweise bereits dieselben Taktiken wie Russland im Vorfeld der Wahlen 2016 an, die Trump an die Macht brachten, um auf die anstehenden Präsidentschaftswahlen im November 2024 Einfluss zu nehmen. Auch in den europäischen Hauptstädten werden ähnliche Warnungen vor einer wachsenden Bedrohung durch China laut.

Zu diesem besorgniserregenden Narrativ gehört auch die oft wiederholte Behauptung,

die Regierungen in Washington, Brüssel, Paris oder Berlin nähmen diesen „Informationskrieg“ trotz Moskaus und Pekings klar demonstrierter Absicht, den Informationskonflikt in den westlichen Gesellschaften weiter auszunutzen, nicht wahr und hielten sich aus falsch verstandener Sorge um die Meinungs- und Informationsfreiheit mit Gegenmaßnahmen zurück. Dabei laufe uns die Zeit davon, wenn wir unsere Demokratien erhalten wollten. Denn die Lage könne sich nur noch weiter verschlechtern, wenn solche böswilligen Akteure aus dem Ausland das volle Potenzial der künstlichen Intelligenz (KI) zu nutzen begännen, die dank der rasanten Entwicklung und Verfügbarkeit immer leistungsfähigerer großer Sprachmodelle (LLMs) die Erzeugung und Verbreitung schädlicher, irreführender Informationen entscheidend unterstützt. Was können wir angesichts dieser düsteren Aussichten tun, um unsere Demokratien vor der Geißel der kognitiven Kriegsführung zu schützen? Dieser kurze Essay will den – bereits zahlreich vorhandenen – Politikempfehlungen nicht noch weitere hinzufügen, sondern einen Schritt zurücktreten und die grundlegenden Annahmen hinterfragen, die dem Problem zugrunde liegen. Lässt sich das Rätsel des gegenwärtigen Informationschaos wirklich auf dem gerade skizzierten Wege lösen?

Manipulation durch Information

Die Idee der Beeinflussung oder sogar Manipulation der öffentlichen Meinung in einem gegnerischen Staat ist natürlich so alt wie der Krieg selbst. Militärische Feldzüge, aber auch einfach nur zwischenstaatliche Auseinandersetzungen auf politischer Ebene wurden ausnahmslos von Versuchen flankiert, die Stimmung im Ausland zu beeinflussen, um den Gegner zu schwächen oder noch zögernde potenzielle Verbündete zu überzeugen. Propaganda, ursprünglich ein eher neutraler Begriff ohne inhärent negative Konnotation, war schon immer Teil der Staatskunst und gilt daher weitgehend als akzeptable Praxis. Mit dem Aufkommen elektronischer Medien zur sofortigen und umfassenden Kommunikation über Online-Plattformen scheint sich etwas verändert zu haben. Insbesondere durch die zunehmende Verbreitung der heute vorherrschenden Social-Media-Kanäle Facebook, X (vormals Twitter) oder TikTok ist eine Entwicklung zu beobachten, die mit der rückläufigen Bedeutung der traditionellen Massenmedien für die öffentliche Meinungsbildung in demokratischen Gesellschaften Hand in Hand geht.[2]

Falsche oder irreführende Informationen, ob absichtlich von böswilligen Akteuren verbreitet oder als bloßes Nebenprodukt von Vorurteilen, Unwissenheit oder Sensationslust, durchdringen unsere digital vernetzte Gesellschaft mit einer bis dato nicht vorstellbaren Geschwindigkeit. Spätestens seit sich die britische Bevölkerung für den Austritt aus der EU entschied und die US-amerikanischen Wähler 2016 Donald Trump ins Weiße Haus brachten, gelten die jüngsten Verwerfungen der Informationslandschaft als eine der drängendsten Fragen unserer Zeit. Laut Global Risks Perception Survey des Weltwirtschaftsforums für 2023/24 steht das Thema „Falschinformationen“ ganz oben auf der Liste globaler Herausforderungen.[3] Natürlich geht das Problem weit über die Wahlmanipulation hinaus: Auf dem Höhepunkt der Covid-19-Pandemie sollen Menschen aufgrund der schnellen Verbreitung von gesundheitsbezogenen Fehlinformationen behördliche Gesundheitsmaßnahmen wie Anordnungen von Lockdowns oder Maskengebote ignoriert haben, nutzlose oder sogar schädliche Substanzen wie Hydroxychloroquin oder Bleichmittel eingenommen oder Impfungen abgelehnt haben. Als eines der jüngsten Beispiele wird die deutlich zurückgegangene Unterstützung der europäischen Gesellschaften für die Verteidigung der Ukraine gegen die russische Aggression auf eine konzertierte Kampagne zurückgeführt, die von Moskau kontrolliert und gesteuert wird.[4]

Falsche oder irreführende Informationen durchdringen unsere digital vernetzte Gesellschaft mit einer bis dato nicht vorstellbaren Geschwindigkeit

Neben den eher übergeordneten Begriffen wie „Desinformation“ oder „Computerpropaganda“ wird für die neue Realität des Informationschaos als Teil eines größeren, heraufziehenden Konflikts zwischen den westlichen Demokratien einerseits und einer wachsenden Zahl autoritärer Staaten andererseits die Bezeichnung hybride oder kognitive Kriegsführung verwendet. Ersteres benennt die vor allem von Russland, aber zunehmend auch von China angewandte, offenbar neue Strategie, traditionelle kinetische Mittel militärischer Aktivitäten mit „unkonventionellen Machtinstrumenten und Werkzeugen der Subversion (...) zu kombinieren, um die Schwachstellen eines Gegners auszunutzen und Synergieeffekte zu erzielen“[5]. Letztere wird noch genauer als nicht kinetische Formen der Kriegsführung definiert, die sich auf moderne Informations- und Kommunikationstechnologien stützen. Sie zeichnen sich dadurch aus, dass sie „ganze Bevölkerungsgruppen ins Visier nehmen (...), sich auf die Verhaltensänderung der Bevölkerung konzentrieren – und zwar durch eine Änderung ihrer Denkweisen anstatt nur durch die Einstreuung einzelner gezielter Falschinformationen zu bestimmten Themen (...) –, auf immer ausgefeiltere psychologische Manipulationstechniken zurückgreifen (...) und auf eine Destabilisierung von Institutionen, insbesondere von Regierungen abzielen, die häufig allerdings indirekt, über die Destabilisierung der Wahrheitsfindung verpflichteter Institutionen wie der Medien und Hochschulen, erfolgt.“[6]

Da die technologische Seite der Gleichung in der Debatte aktuell stark betont wird, erscheinen die Inhalte der kognitiven Manipulation in gewisser Weise nachrangig

Außerhalb der rein militärischen Aspekte hat der Europäische Auswärtige Dienst (EAD) bereits versucht, die eindeutig gegnerische, strategische und externe Dimension des heutigen Informationschaos unter dem Begriff „Foreign Information Manipulation & Interference [FIMI]“ zu fassen. Der EAD definiert diese als „ein Verhaltensmuster, das Werte, Verfahren und politische Prozesse bedroht oder dazu geeignet ist, diese negativ zu beeinflussen. Es handelt sich dabei um ein absichtlich und koordiniert ausgeführtes, wesenhaft manipulatives Verhalten. Dieses kann von staatlichen oder nicht staatlichen Akteure ausgehen, einschließlich ihrer Stellvertreter innerhalb und außerhalb ihres eigenen Hoheitsgebiets.“[7] Das Gemeinsame all dieser Ansätze ist, dass sie Information als wirksam einsetzbare Waffe verstehen, die einen Gegner in einer vor der digitalen Transformation nicht vorstellbaren Weise schwächen können. Die Mittel zur Erzeugung und Verbreitung solcher manipulativen Informationen sind kein Randphänomen, sondern ein wesentliches Merkmal, ja sogar Voraussetzung für dieses Zeitalter des Informationschaos in den offenen, demokratischen Gesellschaften des Westens.

Die Anfänge KI-gestützter kognitiver Kriegsführung

Da die technologische Seite der Gleichung in der Debatte aktuell stark betont wird, erscheinen die Inhalte der kognitiven Manipulation – ob irreführende Informationen über politische Kandidaten oder Falschbehauptungen über die Gefährlichkeit eines neuartigen Virus bzw. die Einsatzmotivation und Fähigkeiten der ukrainischen Streitkräfte – in gewisser Weise nachrangig. Vor diesem Hintergrund wird nachvollziehbar, warum die Entwicklung von großen Sprachmodellen (Large Language Models, Abk. LLMs), die mittlerweile breit verfügbar sind, eine Welle neuer Besorgnis über die Zukunft des gesellschaftlichen Zusammenhalts in demokratischen Staaten ausgelöst hat. LLMs sind Algorithmen, die auf dem Prinzip des Deep Learning (einer der fortgeschrittensten Varianten des maschinellen Lernens) aufbauen. Sie werden anhand riesiger Mengen von Textdaten aus dem Internet und anderen Quellen darauf trainiert, die natürliche menschliche Sprache zu erkennen und zu interpretieren. Zunächst erlernen sie die statistischen Beziehungen zwischen Wörtern, auch als „Tokens“ bezeichnet. Danach sind sie in der Lage, auf der Grundlage einer Texteingabe einen menschlich klingenden, sinnvollen Text zu generieren, indem sie nacheinander die

wahrscheinlichste Abfolge von Wörtern oder Tokens vorausberechnen. Vor allem dank der enorm gestiegenen Rechenleistung und immer größeren Datensätze, die aus den Korpora verschiedener Online-Quellen stammen, weist die neueste Generation der LLMs – etwa ChatGPT 4 von OpenAI, Gemini von Google oder Claude von Anthropic – eine beeindruckende Bandbreite von Möglichkeiten auf, Nutzeranfragen (Prompts) mit gut formulierten, relativ komplex strukturierten Ergebnissen zu beantworten.

Da die heutigen LLMs diese Inhalte zudem in bemerkenswerter Geschwindigkeit generieren, gehen einige Experten davon aus, dass sie schon bald als immer effektivere Manipulationswerkzeuge eingesetzt werden. Dies dürfte die missliche Lage offener Gesellschaften im derzeit herrschenden informationellen Unfrieden noch weiter verschärfen. Bradley Honigberg spricht hier von einer „existenziellen Bedrohung". 2022 stellte er fest: „KI-generierte synthetische Medien und überzeugende KI-gestützte Chatbots bieten den gefährlichen Akteuren eine wachsende Anzahl wirksamer, maßgeschneiderter und schwer nachverfolgbarer Funktionen zur Nachrichtenübermittlung."[8] Neben Desinformation durch LLM-generierte Texte können andere generative KI-Tools sogenannte „Deepfakes" erstellen. Diese synthetischen (audio-)visuellen Medien lassen Prominente oder Politiker Dinge sagen oder auf eine Weise handeln, die nicht der Realität entsprechen. Die Zielgruppen sollen durch diese Manipulation dazu gebracht werden, Falschinformationen über die gezeigten Personen zu glauben und das Vertrauen in sie zu verlieren. Michal Šimečka, Spitzenkandidat der liberalen Fortschrittspartei bei den slowakischen Parlamentswahlen im September 2023, wurde bereits im Vorfeld der Wahlen Opfer dieser Taktik. Dieser Vorfall wurde – genau wie die Einmischung in die Präsidentschaftswahlen selbst – maßgeblich durch den Kreml initiiert.[9]

Die neueste Generation von Sprachmodellen – das haben Versuche gezeigt – ist bereits in der Lage, kontextbezogene Inhalte mit schlagkräftigen Argumenten zu generieren, die sich dem Tonfall und der Sprache bestimmter Zielgruppen sowie der Aufmachung herkömmlicher Medien glaubhaft anpassen. Die Ergebnisse wirken authentisch und sind immer weniger von glaubwürdigem Journalismus zu unterscheiden. In einer kürzlich veröffentlichten Übersichtsstudie konnten drei Forscher zeigen, dass ChatGPT als LLM dazu verwendet werden kann, äußerst detaillierte und glaubwürdige Desinformationen in allen Phasen des Lebenszyklus zu führen – vom Verfassen des Prompts über die Erstellung und Verfeinerung von Falschinformationen auf der Grundlage des Prompts, die gesamte Aufmachung, die Erstellung echt wirkender, ausschließlich zur Verbreitung der Desinformationen angelegter Social Media-Accounts, die Entwicklung einer Verbreitungsstrategie sowie schließlich bis hin zur Verbreitung selbst, dem Fake Engagement (gefälschten Interaktionen), der Weiterverbreitung und, falls erforderlich, sogar der späteren Anpassung des Narrativs.[10]

Einige Experten gehen davon aus, dass Large Language Models schon bald als immer effektivere Manipulationswerkzeuge eingesetzt werden

Sollten sich diese theoretischen Forschungsergebnisse nicht nur als korrekt, sondern auch als praktisch durchführbar erweisen, bedeutet dies, dass LLMs tatsächlich sehr anfällig dafür sind, als perfekte Instrumente zum weiteren Ausbau von Desinformationskampagnen missbraucht zu werden. Es ist also vernünftigerweise davon ausgehen, dass wir uns aktuell an einem neuen Wendepunkt befinden. Maschinelles Lernen lässt sich zur Analyse riesiger Datenmengen aus der Social-Media-Kommunikation und zur Erkennung dort aufkommender Trends einsetzen. Auf dieser Grundlage können gewünschte Botschaften noch weiter verbreitet und Zielgruppen angesprochen werden, die für bestimmte manipulative Narrative potenziell empfänglicher sind. Rechnet man die sich entwickelnden Möglichkeiten zur Umsetzung von Strategien kognitiver Kriegsführung hinzu, scheint der Demokratie und unseren offenen Gesellschaften eine recht düstere Zukunft bevorzustehen.

Der Einsatz von KI zur Abwehr von Desinformationskampagnen

Wie soll darauf regiert werden? Jüngste gesetzgeberische Initiativen der Europäischen Union versuchen, der Geißel der Desinformation durch die Regulierung der Online-Plattformen Herr zu werden; sie werden flankiert von Maßnahmen zur Steigerung der Resilienz politischer Willensbildung und demokratischer Entscheidungsprozesse in der EU.[11] Der Europäische Auswärtige Dienst richtete 2015 zudem die „East Stratcom Task Force" mit der Webseite und Datenbank „EUvsDisinfo" als Tätigkeitsschwerpunkt ein. Die Task Force widmet sich ganz der Abwehr russischer Versuche, die öffentliche Meinung in Europa durch Informationsoperationen zu beeinflussen, indem sie diese aufdeckt und Gegennarrative erarbeitet. Auch die NATO betreibt seit 2014 im lettischen Riga ein Exzellenzzentrum für strategische Kommunikation. Dieses ist weniger auf die Entlarvung falscher russischer Narrative in Echtzeit ausgerichtet, befasst sich aber ebenfalls vorwiegend mit der Untersuchung und Entwicklung von Strategien, um die russischen Variante kognitiver Kriegsführung zu konterkarieren.

Angesichts der Verbreitung von KI-Technologien und des zu erwartenden Anstiegs an informationelle Bedrohungen konzentriert sich die NATO aktuell jedoch eher darauf, mithilfe des maschinellen Lernens Kampagnen der kognitiven Kriegsführung zu erkennen und zu bekämpfen.

Ein Team von Studierenden stellte in der *NATO Review* 2021 einen viel beachteten Vorschlag für ein „Überwachungs- und Warnsystem für kognitive Kriegsführung"[12] vor. Er geht davon aus, dass eine „angemessene Verteidigung zumindest das Bewusstsein erfordert, dass man zum Ziel kognitiver Kriegsführung geworden ist", sowie „die Fähigkeit zur Beobachtung und Orientierung, bevor die Entscheidungsträger sich zum Handeln entschließen können". Das anvisierte Tool soll mit Modellen auf Basis maschinellen Lernens arbeiten, die nicht nur diese Art feindlicher Aktivitäten in sozialen Netzwerken und Online-Medien durch die Identifizierung verdächtiger Verhaltensmuster aufspüren, sondern auch deren Verlauf eigenständig verfolgen und überwachen können. Zusätzlich könnte das System zum Zweck der weiteren Datenverarbeitung automatisierte, fortlaufend aktualisierte Berichte erstellen, die wiederum die Grundlage geeigneter Abwehrmaßnahmen darstellen. Die Echtzeitüberwachung – so die Idee hinter dem Konzept – ist notwendig, um angemessene Reaktionen zu entwickeln und damit eventuelle schädliche Auswirkungen solcher Kampagnen zu verhindern.

Der Einsatz von KI zur Bekämpfung kognitiver Kriegsführung muss sich jedoch nicht auf die Erkennung und Überwachung beschränken. Im nächsten Schritt könnten die Möglichkeiten des maschinellen Lernens genutzt werden, um menschliche Faktenprüfer bei der schnelleren Analyse und Markierung einzelner Inhalte (Flagging) zu unterstützen.[13] Die Vorschläge gehen so weit, solche Systeme direkt in die Gegenmaßnahmen einzubeziehen und sie im Rahmen von Desinformationskampagnen verbreitete Inhalte löschen zu lassen. LLMs könnten zumindest in der Theorie sogar dazu genutzt werden, Informationen oder auch komplette Narrative zu generieren, die eine Kampagne mit Sachargumenten abwehren und Fehlinformationen direkt durch ein jeweils korrektes Gegenstück widerlegen könnten. Dieses würde dann über dieselben Kanäle verbreitet werden und sich an dieselben, im Vorfeld identifizierten Zielgruppen richten.

Diese Ideen sind größtenteils nicht wirklich neu. Die führenden Social-Media-Plattformen setzen schon seit einiger Zeit Lernalgorithmen ein, um die oben beschriebenen Aktivitäten auf ihren Seiten zu erkennen und zu unterbinden. Die Unternehmensgruppe Meta, zu der Facebook, Instagram, WhatsApp und Threads gehören, nutzt beispielsweise Systeme, die automatisch jedes „koordinierte nicht authentische Verhalten" (CIB) aufdecken. In ihren Richtlinien werden diese Aktivitäten definiert als „koordinierte Bemühungen, die öffentliche Debatte für ein strategisches Ziel zu manipulieren, wobei gefälschte Accounts im Mittelpunkt der Operation stehen"; letztere werden üblicherweise nach ihrer Erkennung gelöscht. Auch andere Plattformanbieter, wie TikTok oder YouTube, haben solche Mechanismen eingerichtet. Dar-

über hinaus werden aktuell in großem Umfang prädiktive Machine-Learning-Tools von Social-Media-Unternehmen zur algorithmischen Inhaltsmoderation auf den verschiedenen Plattformen eingesetzt. Alle Inhalte werden analysiert; dabei werden sämtliche Inhaltselemente, die gegen Geschäftsbedingungen des Anbieters verstoßen, automatisch gelöscht. Oft gehört auch Desinformation dazu – sofern sie als potenziell schädlich erkannt wird.[14]

Der Unterschied zwischen den Bemühungen einzelner digitaler Dienstleistungsunternehmen und dem oben skizzierten Ansatz, Staaten oder ein supranationales Gebilde wie die EU oder die NATO KI-gestützte Systeme zur Abwehr kognitiver Kriegsführung einsetzen zu lassen, besteht darin, dass Letztere ihre Maßnahmen per Definition auf sämtlichen Plattformen und Websites durchführen würden. Denn genau darum geht es bei den algorithmischen Tools: Sie sollen den Datenverkehr in digitalen Netzwerken im großen Stil überwachen können, im besten Fall einschließlich der Websites von Medien. Sie sollen plattformübergreifende Kampagnen erkennen und bekämpfen, die versuchen, über eine Vielzahl verschiedener Kommunikationskanäle Einfluss auf die Bevölkerung westlicher Staaten zu nehmen – bei den gewieftesten und komplexesten feindlichen Kampagnen ist dies aktuell bereits der Fall. Der kategorische Unterschied zwischen einem privatwirtschaftlichen Unternehmen, das Algorithmen zum Schutz seiner eigenen Produkte einsetzt, und staatlichen Akteuren, die diese Maßnahmen in viel größerem Maßstab umsetzen anstatt auf einem einzigen Kanal, liegt dabei auf der Hand. Dieser kategorische Unterschied besteht auch in rechtlicher, ethischer und politischer Hinsicht.

Die Mechanismen der Desinformation

Zu den größten Problemen bei der Suche nach einer Antwort auf das gegenwärtige Informationschaos zählt die Tatsache, dass praktisch alle Gegenmaßnahmen auf Annahmen über Desinformationskampagnen und deren Mechanismen fußen, die einer wissenschaftlichen Prüfung nicht standhalten. Implizite oder explizite Grundlage der Entwicklung von Gegenmaßnahmen ist offensichtlich die Prämisse, zwischen der Verbreitung irreführender Informationen und irgendeiner negativen Auswirkung bestehe ein unmittelbarer Kausalzusammenhang. Dabei wurde der entsprechende Nachweis noch gar nicht geführt. Trotz einer bereits riesigen, weiterhin rapide anwachsenden Anzahl von Studien in den Kognitions- und Sozialwissenschaften über die Mechanismen irreführender Informationen und weitere Formen der kognitiven Kriegsführung liegen bislang kaum konkrete Beweise dafür vor, dass die gegnerischen Maßnahmen irgendwelche Auswirkungen auf das Verhalten der Zielgruppen haben.[15] Wenn überhaupt, deuten die Forschungsergebnisse eher auf das Gegenteil hin.[16]

Durchaus erwiesen ist jedoch, dass gegnerische Akteure, vor allem Russland und China, versucht haben und auch weiterhin versuchen, die öffentliche Meinung in westlichen Ländern durch den Einsatz kognitiver Kriegsführung zu beeinflussen. Doch wie bereits mit Verweis auf den Wahlerfolg Donald Trumps im Jahr 2016 angemerkt, ist „der Nachweis anhaltender Bemühungen nicht gleichzusetzen mit dem Nachweis von Auswirkungen oder Prävalenz“.[17] Dies liegt nicht daran, dass die empirische Forschung in diesem Bereich übermäßig komplex wäre, sondern eher daran, dass die gängigen Annahmen darüber, welche Informationen das Verhalten der Zielgruppen tatsächlich beeinflussen, womöglich größtenteils falsch sind. Anstatt von leichtgläubigen Rezipienten auszugehen, die sich unter dem Einfluss irreführender Informationen falsche Überzeugungen zu eigen machen und bereitwillig danach handeln, trifft Hannah Arendts Beobachtung auch heute noch im Großen und Ganzen zu: „Massen werden so wenig durch Tatsachen überzeugt, daß selbst erlogene Tatsachen keinen Eindruck auf sie machen. Auf sie wirkt

Die gängigen Annahmen darüber, welche Informationen das Verhalten der Zielgruppen tatsächlich beeinflussen, sind womöglich größtenteils falsch

nur die Konsequenz und Stimmigkeit frei erfundener Systeme, die sie miteinzuschließen versprechen."[18] Tatsächlich bestätigt die kognitionswissenschaftliche Forschung, dass der Empfänger einer Information wahrscheinlich nur dann beeinflusst werden kann, wenn eine neue Information – unabhängig von ihrem Wahrheitsgehalt – an Überzeugungen anknüpft, die bereits vorher Teil seines Glaubenssystems waren.[19]

Entscheidend ist, dass sich dieses Bild mit dem zunehmenden Einsatz großer Sprachmodelle zur Erstellung und Verbreitung falscher und irreführender Informationen wahrscheinlich nicht ändern wird. Selbst wenn der Einsatz solcher auf maschinellem Lernen basierenden Algorithmen zu mehr, besseren und stärker personalisierten falschen und irreführenden Informationen führt, konnte kürzlich gezeigt werden, dass „die vorhandenen Forschungsergebnisse höchstens geringfügige Auswirkungen der generativen KI auf die Desinformationslandschaft nahelegen". Die immer öfter heraufbeschworenen Gefahren für die Informationsökosysteme in demokratischen Staaten erscheinen daher als „aufgebauscht"[20]. Gängige Medienberichte wie das eingangs zitierte Beispiel, dem zufolge der Kreml mit einer konzertierten Beeinflussungskampagne erneut den Ausgang einer Wahl in einem westlichen Land bestimmt habe, sind also bestenfalls irreführend; schlimmstenfalls befeuern sie einen gefährlich Alarmismus.

Die immer öfter heraufbeschworenen Gefahren für die Informationsökosysteme in demokratischen Staaten erscheinen als aufgebauscht

Überreaktionen und ihre Gefahren

Es ist wichtig, dass wir die Mechanismen der Informationsmanipulation richtig verstehen. Denn jeder Eingriff auf staatlicher Ebene bedeutet einen erheblichen Eingriff in unsere Grundrechte, besonders wenn er mithilfe von KI erfolgt. Zum einen hat ein algorithmisches Frühwarnsystem für Kampagnen kognitiver Kriegsführung offensichtliche Auswirkungen auf Privatsphäre und Datenschutz. Denn ein solches Modell kann nur funktionieren, wenn es sowohl zu Trainingszwecken als auch während des tatsächlichen Einsatzes mit großen Mengen von Eingabedaten gefüttert wird. Da das hauptsächliche Ziel darin besteht, ungewöhnliche Aktivitäten in digitalen Netzen aufzudecken, umfassen diese Daten automatisch auch große Mengen personenbezogener Informationen über die Nutzer von Social-Media-Plattformen und anderen Websites.

Noch wichtiger zu erwähnen ist vielleicht, dass die Befugnis eines Staates oder einer anderen maßgeblichen Stelle, zwischen „guten" und „schlechten" Online-Inhalten zu unterscheiden, sich sehr nachteilig auf die individuellen Kommunikationsrechte wie das Recht auf freie Meinungsäußerung und das Informationsrecht auswirken könnte. In den vergangenen Jahren hat sich bereits gezeigt, welche Folgen es hat, wenn Staaten entscheiden, welche Art von Online-Inhalten als „Desinformation" gelten und vom Netz genommen werden sollten. Das Problem erstreckt sich dabei nicht nur auf die Beweisführung und Fragen einer Definition, die rechtsstaatlichen Standards entspricht. Vor allem nach Beginn der mit Covid-19 einhergehenden globalen Gesundheitskrise nutzten einige autoritär gesinnte Regierungen die aufkommende Debatte zur gesundheitsbezogenen Desinformation, um die freie Meinungsäußerung zu unterbinden.[21] Diese Bedenken sollten auch dann ernst genommen werden, wenn sich die algorithmischen Systeme auf die Erkennung und Überwachung möglicher gegnerischer Kampagnen beschränken. Denn sobald der Algorithmus eine solche erkennt, wird der Druck groß sein, dagegen vorzugehen, indem damit zusam-

Der Autor

***Henning Lahmann** ist Assistant Professor am Center for Law and Digital Technologies der juristischen Fakultät der Universität Leiden. Seine Forschungstätigkeit konzentriert sich auf die Schnittstelle zwischen Technologie und Völkerrecht. 2021/22 war Henning Lahmann über ein Forschungsstipendium des Deutschen Akademischen Austauschdienstes als Hauser Post-Doctoral Global Fellow an der NYU School of Law tätig. Er promovierte im Völkerrecht an der Universität Potsdam.*

menhängende Accounts oder Informationen gelöscht werden – sei es manuell oder mithilfe eines weiteren Algorithmus, der automatisch gegen solche Inhalte vorgeht. Bei der letztgenannten Variante ist der Eingriff in die Kommunikationsrechte natürlich noch gravierender. Die gegnerischen Akteure in Moskau und Peking mögen alles in ihrer Macht Stehende tun, um die Bürger in den westlichen Demokratien zu beeinflussen; dennoch muss mit Nachdruck betont werden, dass solche Versuche weder das Recht auf freie Meinungsäußerung noch die Informationsfreiheit außer Kraft setzen. Eine funktionierende liberale Demokratie zeichnet sich vielmehr dadurch aus, dass die Bürger selbst entscheiden können, wo sie sich informieren und sogar welchen Darstellungen der Realität sie Glauben schenken wollen. Deshalb steht den Regierungen im Allgemeinen kein Recht zu, paternalistisch in den Prozess der politischen Willensbildung einzugreifen. Auf die kollektive Ebene bezogen, gefährdet ein KI-gestützter informationeller Interventionismus letztlich das Selbstbestimmungsrecht eines Volkes.

Dieser Vorbehalt verweist auf eine übergeordnete Überlegung. Der im Westen geführte politische Diskurs rund um die Begriffe „Desinformation“, „kognitive Kriegsführung“ bzw. „ausländische Informationsmanipulation und Einmischung“ setzt implizit voraus, das gegenwärtige „Zeitalter des Misstrauens“ stelle in erster Linie ein Sicherheitsproblem und eine Bedrohung aus dem Ausland dar. Doch genau die Versicherheitlichung und Externalisierung des Problems hat erst dazu geführt, dass überhaupt in militärischen Kategorien nach Lösungen des Problems gesucht wurde. Mehrere Forschende konnten zeigen, dass die gefährlichsten falschen politischen Narrative vorwiegend im eigenen Land entstehen und verbreitet werden. Externe Akteure nutzen diese lediglich aus und verstärken sie.[22] Wenn dies zutrifft, muss eine tatsächliche Lösung in erster Linie die zunehmende Polarisierung in unserer eigenen Gesellschaft angehen, anstatt zu versuchen, die böswilligen Machenschaften externer Akteure zu durchkreuzen. „Frühwarnsysteme“ brauchen wir schließlich mit Sicherheit zur Erkennung anfliegender Raketen oder anderer bewaffneter gegnerischer Aktivitäten. Ob wir sie auch für den Umgang mit potenziell problematischen verbalen Äußerungen brauchen, ist weit weniger klar.

Die digitale Transformation, insbesondere der Aufstieg der sozialen Medien, geht mit einer zunehmenden Fragmentierung der Informationsökosysteme in den offenen Gesellschaften der westlichen Demokratien einher. Diese Entwicklung versuchen gegnerische Staaten wie Russland und China für ihre eigenen geopolitischen Ziele zu nutzen. Die Entwicklung und der erwartete allgegenwärtige Einsatz großer Sprachmodelle könnte zu einer weiteren Verschärfung der Situation beitragen, auch wenn sich die tatsächlichen Auswirkungen der LLMs letztendlich als weniger gravierend erweisen könnten als gemeinhin befürchtet. Wir sollten vor allem bei der Formulierung politischer Antworten auf das wahrgenommene Informationschaos vorsichtig sein, damit sich der Einsatz von Algorithmen nicht als Versuch erweist, den Teufel mit dem Beelzebub auszutreiben. Vielmehr sind die politischen Entscheidungsträger gut beraten, sich auf langfristige Strategien zu konzentrieren; diese mögen technologisch weniger attraktiv erscheinen, haben dafür aber den Vorteil, Einschränkungen genau derjenigen Rechte und Werte zu vermeiden, die mit dem Kampf gegen das Informationschaos geschützt werden sollen.

Wir sollten vor allem bei der Formulierung politischer Antworten auf das wahrgenommene Informationschaos vorsichtig sein

1 Hockenos, Paul (2024): Russia Just Helped Swing a European Election. Foreign Policy. https://foreignpolicy.com/2024/04/17/slovakia-president-pellegrini-russia-election-interference-disinformation/ (Stand aller Internetquellen: 30. April 2024).
2 Siehe z. B. Newman, Nic (2023): Young People Are Abandoning News Websites – New Research Reveals Scale of Challenge to Media. The Conversation. https://theconversation.com/young-people-are-abandoning-news-websites-new-research-reveals-scale-of-challenge-to-media-207659; Lipka, Michael and Shearer, Elisa (2023): Audiences Are Declining for Traditional News Media in the U.S. – With Some Exceptions. Pew Research Center. https://www.pewresearch.org/short-reads/2023/11/28/audiences-are-declining-for-traditional-news-media-in-the-us-with-some-exceptions/.
3 World Economic Forum (2024): Global Risks Reports 2024, S. 18. https://www3.weforum.org/docs/WEF_The_Global_Risks_Report_2024.pdf.
4 Digital Forensics Research Lab (2024): Undermining Ukraine: How Russia Widened Its Global Information War in 2023. https://www.atlanticcouncil.org/in-depth-research-reports/report/undermining-ukraine-how-russia-widened-its-global-information-war-in-2023/.
5 Bilal, Arsalan (2021): Hybrid Warfare – New Threats, Complexity, and 'Trust' as the Antidote. NATO Review. https://www.nato.int/docu/review/articles/2021/11/30/hybrid-warfare-new-threats-complexity-and-trust-as-the-antidote/index.html. (Übersetzung aus dem Englischen.)
6 Miller, Seumas (2023): Cognitive Warfare: An Ethical Analysis. In: Ethics and Information Technology, Bd. 25, S. 1. (Übersetzung aus dem Englischen.)
7 Europäischer Auswärtiger Dienst (2021): Tackling Disinformation, Foreign Information Manipulation & Interference. https://www.eeas.europa.eu/eeas/tackling-disinformation-foreign-information-manipulation-interference_en. (Übersetzung aus dem Englischen.)
8 Honigberg, Bradley (2022): The Existential Threat of AI-Enhanced Disinformation Operations. Just Security. https://www.justsecurity.org/82246/the-existential-threat-of-ai-enhanced-disinformation-operations/. (Übersetzung aus dem Englischen.)
9 Meaker, Morgan (2023): Slovakia's Election Deepfakes Show AI Is a Danger to Democracy. Wired. https://www.wired.com/story/slovakias-election-deepfakes-show-ai-is-a-danger-to-democracy/.
10 Barman, Dipto, Guo, Ziyi und Conlan, Owen (2024): The Dark Side of Language Models: Exploring the Potential of LLMs in Multimedia Disinformation Generation and Dissemination. In: Machine Learning with Applications 16, Art. 100545. https://www.sciencedirect.com/science/article/pii/S2666827024000215.
11 Europäische Kommission (2024): Leitlinien für Anbieter von VLOPs und VLOSEs zur Minderung systemischer Risiken für Wahlprozesse. https://digital-strategy.ec.europa.eu/de/library/guidelines-providers-vlops-and-vloses-mitigation-systemic-risks-electoral-processes.
12 Johns Hopkins University und Imperial College London (2021): Countering Cognitive Warfare: Awareness and Resilience. NATO Review. https://www.nato.int/docu/review/articles/2021/05/20/countering-cognitive-warfare-awareness-and-resilience/index.html. (Übersetzung aus dem Englischen.)
13 Bateman, Jon und Jackson, Dean (2024): Countering Disinformation Effectively: An Evidence-Based Policy Guide. Carnegie Endowment for International Peace, S. 87.
14 Gorwa, Robert, Binns, Reuben und Katzenbach, Christian (2020): Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance. In: Big Data & Society 7 (1). https://journals.sagepub.com/doi/epub/10.1177/2053951719897945.
15 Siehe u. a. Bateman, Jon et al. (2021): Measuring the Effects of Influence Operations: Key Findings and Gaps from Empirical Research. Carnegie Endowment for International Peace; Maschmeyer, Lennart et al. (2023): Donetsk Don't Tell – 'Hybrid War' in Ukraine and the Limits of Social Media Influence Operations. In: Journal of Information Technology & Politics. https://www.tandfonline.com/doi/epdf/10.1080/19331681.2023.2211969?needAccess=true.
16 Mercier, Hugo und Altay, Sacha (2022): Do Cultural Misbeliefs Cause Costly Behavior? In: Joseph Sommer et al. (Hg.): The Cognitive Science of Belief: A Multidisciplinary Approach. Cambridge, S. 193-208.
17 Benkler, Yochai, Faris, Robert und Roberts, Hal (2018): Network Propaganda. Oxford, S. 254. (Übersetzung aus dem Englischen.)
18 Arendt, Hannah (1975): Elemente und Ursprünge totaler Herrschaft. Band 3: Totalitarismus. Frankfurt a. M., Berlin, Wien, S. 93 f.
19 Jowett, Garth S. und O'Donnell, Victoria (2012): Propaganda and Persuasion. 5. Auflage. Los Angeles et al., S. 34.
20 Simon, Felix M., Altay, Sacha und Mercier, Hugo (2023): Misinformation Reloaded? Fears About the Impact of Generative AI on Misinformation Are Overblown. In: Harvard Kennedy School Misinformation Review 4, S. 3. https://misinforeview.hks.harvard.edu/wp-content/uploads/2023/10/simon_generative_AI_fears_20231018.pdf. (Übersetzung aus dem Englischen.)
21 Siehe Kaye, David (2020): Disease Pandemics and the Freedom of Opinion and Expression: Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression. Vereinte Nationen. https://documents.un.org/doc/undoc/gen/g20/097/82/pdf/g2009782.pdf?token=vxfXRfLUEqLTf8Syoi&fe=true.
22 Benkler, Yochai, Faris, Robert und Roberts, Hal (2018), siehe Endnote 16.