INSTITUT FÜR
INNOVATION UND
TECHNIK

Desinformation und Automated Influence – Forderungen und Maßnahmen zur digitalen Resilienz

Volker Wittpahl, Ernst Andreas Hartmann

Impressum

Herausgeber
Prof. Dr. Wittpahl
Institut für Innovation und Technik (iit)
in der VDI/VDE Innovation + Technik GmbH
Steinplatz 1
10623 Berlin
Tel.: +49 30 310078 5507
Fax: +49 30 310078 104
E-Mail: info@iit-berlin.de
www.iit-berlin.de

Autoren
Prof. Dr. Volker Wittpahl
Dr. Ernst Andreas Hartmann

Layout
Poli Quintana

Bildrechte
Wuttichai – stock.adobe.com

DOI: 10.23776/2024_07
Berlin, Mai 2024

Zitation
Wittpahl, Volker; Hartmann, Ernst Andreas (2024). Desinformation und Automated Influence. Forderungen und Maßnahmen zur digitalen Resilienz. Institut für Innovation und Technik (iit).

Die Autoren danken Dr. Marc Bovenschulte für die fachlichen Diskussionen zur Erstellung dieser Publikation, insbesondere zum Phänomen „Pink Slime“.

Inhalt

Abstract

Desinformation ist kein neues Phänomen. Getrieben durch die Verbesserung und den Einsatz von selbstlernenden Algorithmen – Künstliche Intelligenz (KI) – hat sich jedoch die Qualität und Quantität von Desinformation geändert. Sie bedroht immer stärker die Wirtschaft und die Gesellschaft durch verbesserte Möglichkeiten der automatisierten Einflussnahme (engl. Automated Influence). Wirtschaft und Gesellschaft sind in großen Teilen schlecht bis gar nicht auf diese Entwicklungen vorbereitet.

Automated Influence kann die Meinungsbildung einer Gesellschaft und ihre Fähigkeit zu rationalem Diskurs und Handeln in kurzer Zeit negativ beeinflussen, sodass angesichts der kontroversen Debatten in der Zeitenwende und des „globalen Superwahljahrs 2024" rasch Mechanismen zu Eindämmung und Entlarvung dieser Beeinflussung nötig sind. Das Hauptziel muss daher sein, die Gesellschaft und Unternehmen resilienter gegenüber digitalen Desinformationskampagnen zu machen.

In dem vorliegenden Arbeitspapier wird zunächst ein Überblick zur Desinformation und der aktuellen Bedrohung durch den Einsatz von Automated Influence gegeben. Methoden der Entlarvung von Desinformation (engl. Debunking) und der präventiven Immunisierung gegen Desinformation (engl. Inoculation, Pre-Bunking) werden dargestellt und anhand der Forschungsliteratur bewertet.

Am Beispiel von Taiwan werden Maßnahmen aufgezeigt, wie der Bedrohung begegnet werden kann, und diskutiert, ob diese in Deutschland zur langfristigen Steigerung der digitalen Resilienz gegen Desinformation implementierbar sind.

1 Einleitung: 2024 – ein Jahr der Wahlen und der Desinformation

In den vergangenen Jahren nahm die gezielte Desinformation in Form von Falschmeldungen (engl. Fake News) mit manipulierten Bildern und Videos (engl. Deep Fake) im Internet über Online-Plattformen und soziale Medien massiv zu (Pérez-Escolar et al., 2023).

Nicht umsonst wurde Desinformation vom World Economic Forum im Januar 2024 als aktuell größtes globales Risiko noch vor der Klimaveränderung gesehen (World Economic Forum, 2024). Dies ist umso kritischer zu deuten, als dass im Jahr nicht nur die Wahlen zum Europaparlament stattfinden, sondern insgesamt in mehr als 60 Ländern weltweilt gewählt wird, was 3,6 Milliarden Menschen betrifft, also circa 45 Prozent der Weltbevölkerung (Wahls, 2024).

Das Bewusstsein für und die Fähigkeit zum Faktencheck und zur Quellenprüfung ist nur bei einem kleinen Teil der Internetnutzer:innen gegeben und wird von ihnen kaum in der Breite angewandt. Die Konsequenzen sind durch die Erfahrungen und Erkenntnisse während der Covid-Pandemie bekannt: Verschwörungstheorien greifen um sich und Vertrauen schwindet. Tatsächlich haben Untersuchungen in Großbritannien, den USA und Kanada während der Pandemie gezeigt, dass Faktenchecks sehr wohl – insbesondere bei fehlinformierten Personen – deutliche Effekte gezeitigt haben (kurzzeitige „Immunisierung"), diese aber nicht von Dauer waren, d. h. keine Ausbildung eines „medienimmunologischen Gedächtnisses" (Carey et al., 2022). Es zeigt sich, dass Gegenmaßnahmen zur Eindämmung von Desinformation von Entscheider:innen nicht konsequent und vollumfänglich angegangen und umgesetzt werden. Ursache hierfür ist fehlendes Verständnis für die Bedrohungen sowie fehlende Expertise aufgrund der Komplexität von Technologien und der damit einhergehenden Konsequenzen bei ihrem Einsatz. Auch zu verfügbaren Gegenstrategien und deren Anwendungsbedingungen fehlen oftmals Informationen im politischen Raum.

Definition: Was ist Desinformation und wie wirkt sie?

In der Literatur werden Begriffe im Kontext der Desinformation unterschiedlich verwendet. Esma Aïmeur und Koautor:innen (Aïmeur et al., 2023) schlagen auf der Basis einer Literaturanalyse eine differenzierte Begriffsstruktur vor.

Begriffsrelationen im Umfeld der Desinformation

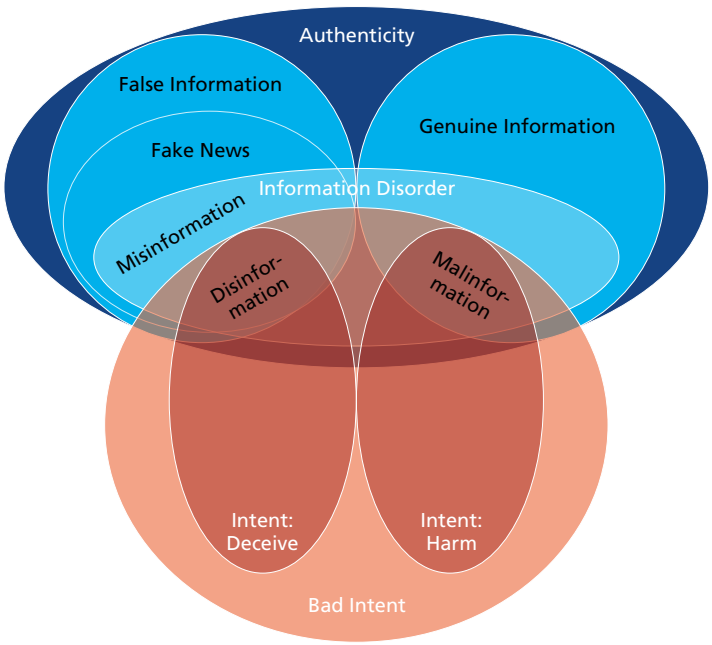

Abbildung 1: Begriffsrelationen im Umfeld der Desinformation (eigene Darstellung nach Aïmeur et al., 2023)[1]

1 Die Begriffe wurden in der englischen Sprache belassen, da die Verständlichkeit durch die Übersetzung ins Deutsche nicht steigt.

Im Zentrum steht die Unterscheidung zwischen drei Begriffen, die hier jeweils mit ihrer englischen Begrifflichkeit und Verwendung angegeben sind:

- **engl. Disinformation:** Unwahre Informationen, die mit bösartiger Intention (hier: Täuschung) erzeugt werden.

- **engl. Misinformation:** Unwahre Informationen, die ohne bösartige Intention (beispielsweise aus Nachlässigkeit, Gedankenlosigkeit oder aus humoristisch/satirischer Absicht) erzeugt werden.

- **engl. Malinformation:** Wahre Informationen, die erzeugt werden, um Personen oder Organisationen zu schaden, z. B. durch ihre Veröffentlichung (engl. Leaks).

In der vorliegenden Publikation soll der Schwerpunkt auf Disinformation – bzw. im Deutschen Desinformation, aber im hier dargestellten Sinne – liegen, und dabei eher auf den Formen mit starker Täuschungsabsicht.

Ein weiterer Schwerpunkt liegt auf fortgeschrittenen KI-generierten desinformierenden Botschaften (Bontcheva, et al., 2024; Bontridder & Poullet, 2021), sei es als Text auf der Basis von großen Sprachmodellen (Kreps et al., 2022; Stiff & Johansson, 2022) oder als computergenerierte audiovisuelle Medien auf der Basis von Generative Adversarial Networks (GAN) (Bovenschulte, 2019; Dobber et al., 2021; Vaccari & Chadwick, 2020).

2 Wahrnehmung und kognitive Verarbeitung von Desinformation

Wahrnehmung

Für die Wahrnehmung und kognitive Verarbeitung von Desinformation ist eine Unterscheidung in zwei verschiedene kognitive Prozesse hilfreich (Pennycook et al., 2015): Prozesse vom *Typ 1* sind intuitiv, schnell und vom ‚Bauchgefühl' geprägt, Prozesse vom *Typ 2* sind reflektiert und analytisch. Es überrascht wenig, dass Personen, die zu Prozessen des *Typs 2* neigen – also eher analytisch als intuitiv denken –, weniger anfällig für Falschmeldungen sind als Personen, die zu Prozessen des *Typs 1* neigen – also eher intuitiv als analytisch denken (Pennycook & Rand, 2021).

Generell verwenden Personen als wesentliche Merkmale zur Identifikation von Falschmeldungen bevorzugt Quelle (Seriosität) und formale Gestaltung ('Look and Feel') (George, 2024).

Ein weiterer wichtiger Befund ist, dass Gründe dafür, dass Menschen Desinformation weitergeben – mit anderen teilen – nicht so sehr darin liegen, dass sie diese Informationen selbst glauben, oder dass sie sie selbst nicht glauben, aber motiviert weitergeben, etwa um eine bestimmte politische Agenda zu befördern. Gegenüber diesen beiden möglichen Gründen scheint ein dritter Grund wichtiger zu sein: Unachtsamkeit bzw. mangelnde Aufmerksamkeit, was einhergeht mit den Befunden zu Prozessen vom *Typ 1* und vom *Typ 2* (Pennycook & Rand, 2021).

Dies ist wichtig, weil sich daraus eine sehr einfache und wirksame Gegenmaßnahme ableiten lässt: Das im Englischen als ‚Accuracy Nudging' bezeichnete Vorgehen. In einem Feldexperiment wurden Nutzer:innen von Online-Plattformen und Messengerdiensten identifiziert, die bereits Links zu Plattformen, die in hohem Maße Falschmeldungen verbreiten, geteilt hatten. Diesen Nutzer:innen wurde eine Nachricht gesendet, in der sie aufgefordert wurden, die Korrektheit einer neutralen, nichtpolitischen Aussage, hier in Form einer Nachrichtenüberschrift, zu bewerten. Diese minimale Intervention reichte aus, um das Weitergabeverhalten dieser Personen so zu verändern, dass nach der Intervention seriöse Quellen häufiger und unseriöse Quellen weniger häufig geteilt wurden (Pennycook et al., 2021).

Verbreitung

Ein genereller Befund ist, dass digitale Falschmeldungen im Vergleich zu zutreffenden Nachrichten schneller verbreitet werden und mehr Personen erreichen. Gründe hierfür werden in der relativen Neuigkeit und dem relativ hohen emotionalen Gehalt dieser falschen Nachrichten gesehen (Vosoughi et al., 2018).

Hinzu kommt, dass angesichts einer Flutung des Informationsmarktplatzes mit Falschmeldungen *„die Öffentlichkeit eher bereit sein kann, wahre, aber unbequeme Fakten zu bezweifeln. Kognitive Verzerrungen fördern bereits heute den Widerstand gegen solche Fakten, und das Bewusstsein für allgegenwärtige Deep Fakes kann diese Tendenz verstärken und eine gute Ausrede bieten, um unerwünschte Beweise zu ignorieren. Insbesondere wenn gefälschte Videos weit verbreitet sind, kann die Öffentlichkeit Schwierigkeiten haben zu glauben, was ihre Augen (oder Ohren) ihnen sagen – selbst wenn die Informationen ganz real sind"* (eigene Übersetzung nach (Chesney & Citron, 2019)).

Es können zwei Phasen der Verbreitung von Falschmeldungen unterschieden werden: Das ursprüngliche ‚Aussähen' (engl. Seeding) und die Verstärkung und Verbreitung in (Sozialen) Medien, wobei auch Echokammern eine Rolle spielen können (Diaz Ruiz & Nilsson, 2023).

Zur quantitativen und qualitativen Bedeutung von Echokammern finden sich in der Literatur sehr unterschiedliche Einschätzungen. Hier besteht auch eine Abhängigkeit von den Untersuchungsmethoden: Studien, die Internet-Nutzungsdaten verwenden (engl. Digital Trace Data), finden eher starke Hinweise auf Existenz und Wirkung von Echokammern, auch weil sie andere Formen der Mediennutzung – z. B. traditionelle Print- und digitale Medien, auch andere Soziale Medien als die gerade untersuchten – eher ausblenden. Umgekehrt finden Studien, die Selbstauskünfte verwenden, geringere Hinweise auf Echokammern, auch weil hier das ganze Spektrum der Kommunikation und Mediennutzung der jeweiligen Personen betrachtet wird (Terren & Borge-Bravo, 2021).

Weiterhin gibt es Befunde, die eine Differenzierung des Konzepts Echokammer nahelegen. So sind etwa nicht alle Menschen gleichermaßen betroffen, eine hohe Anfälligkeit für Echokammern zeigen in besonderem Maße Personen mit extremen Positionen und Sichtweisen. Schließlich wird ein Kontinuum von eher offenen bis hin zu sehr geschlossenen Kommunikationsräumen vorgeschlagen. Diese Perspektive verweist auf Schlupflöcher, durch die auch ungefilterte, korrigierende Informationen eindringen können, in vielen, auch tendenziell geschlossenen Kommunikationsräumen (Geiß et al., 2021). Hier eröffnen sich Potenziale für effektive Gegenmaßnahmen.

Betrachtet man vor diesem Hintergrund die Kommunikation auf spezifischen Online-Plattformen und Messengerdiensten, finden sich – je nach Medium mehr oder weniger stark – Gruppen, deren Mitglieder in ihren Meinungen und Weltanschauungen untereinander sehr ähnlich sind. Kommunikation auf der jeweiligen spezifischen Online-Plattform findet sehr oft nur noch innerhalb dieser Gruppen statt, insbesondere auf Plattformen, die

intensiv Newsfeed-Algorithmen verwenden – wie z. B. Facebook oder X/Twitter – und den Nutzer:innen keine individuellen Anpassungen dieser Algorithmen ermöglichen (Cinelli et al., 2021).

Diese Echokammern strukturieren sich nach personalen und sozialen Identitätsdimensionen – zum Beispiel Weltanschauung, politische Haltung, soziodemografische Merkmale wie Geschlecht oder regionale und soziale Herkunft (Huddy, 2001; Tajfel & Turner, 2004; Ybema, 2020). Die Kommunikation in diesen Echokammern verstärkt die geteilten Überzeugungen, die die gemeinsame Identität ausmachen, und immunisiert sie gegen Interventionen von außen (Nyhan & Reifler, 2019). Oftmals sind solche Kommunikationen in identitär geprägten Echokammern bezogen auf Auseinandersetzungen mit externen Gegnern nach der Devise „Wir gegen die!"; diese Auseinandersetzungen können den Charakter von ‚Kulturkriegen' einnehmen oder in solche eingebettet sein (Diaz Ruiz & Nilsson, 2023).

Die Auseinandersetzung mit Effekten der sozialen Identität ist im Hinblick auf ihre Bedeutung für Desinformation in sozialen Medien noch relativ unterentwickelt (Diaz Ruiz & Nilsson, 2023), es finden sich aktuell wenige Arbeiten, die solche Erkenntnisse für Gegenmaßnahmen nutzbar machen (Ziemer & Rothmund, 2024). Unter diesen wenigen Ansätzen lassen sich die folgenden hervorheben, welche im Englischen definiert sind als:

- **Self-Affirmation:** Bevor Personen mit Informationen konfrontiert werden, die ihre Identität – bzw. die dieser Identität zugrundeliegenden Überzeugungen – herausfordern und gefährden könnten, werden diese Personen dazu aufgefordert, über Werte zu reflektieren, die ihnen persönlich wichtig sind. So soll ein Selbstwertempfinden gesteigert und dadurch der Druck gemindert werden, sich gegen äußere Bedrohungen der Identität zu wehren (Lyons, 2018; Nyhan & Reifler, 2019).

- **Multiple Identity Salience Manipulation:** Bezogen auf einen einzelnen Menschen überlappen in der Regel mehrere oder sogar viele Identitäten – orientiert an beispielsweise kulturellen Vorlieben, familiären Rollen, weltanschaulichen Überzeugungen. Die Intervention besteht hier darin, die Personen aufzufordern, über ihre unterschiedlichen Identitäten und deren Bedeutung für ihr Bewusstsein ihres Selbst zu reflektieren. Die Wirkung soll nun darin bestehen, dass durch die bewusste Präsenz mehrerer sehr unterschiedlicher Identitäten jede einzelne relativ weniger wirkmächtig wird (Lyons, 2018).

- **Perspective Taking:** Die Personen werden aufgefordert, in einem imaginierten Dialog eine als positiv empfundene Kommunikation mit einer Person zu führen, die eine im Hinblick auf den jeweiligen Sachverhalt antagonistische Identität aufweist (Guan et al., 2021).

Die vorliegenden Ergebnisse sind noch uneinheitlich und weisen auf eine noch begrenzte Wirksamkeit dieser Interventionen hin (Lyons, 2018; Nyhan & Reifler, 2019).

Solche und andere Gegenmaßnahmen werden Kapitel 4 *Gegenmaßnahmen zur Desinformation* im Zusammenhang dargestellt.

3 Aktuelle Entwicklung: Desinformation heute – rapide irreführende Meldungen

Aktuelle Entwicklung

Was sich bei der Erstellung von Desinformation in letzter Zeit gegenüber den Erfahrungen aus dem Vor-Internet-Zeitalter geändert hat, ist eine Verschärfung der Situation aufgrund des Einsatzes von selbstlernenden Algorithmen – im Deutschen allgemein als Künstliche Intelligenz (KI) bezeichnet – in unterschiedlichen Anwendungsbereichen bei gleichzeitiger Zunahme der Leistungsfähigkeit der Algorithmen. Das populärste Beispiel hierfür ist das Sprachmodell ChatGPT. Neben den erstaunlichen Fähigkeiten der Sprachmodelle beim Verfassen und Verbessern unterschiedlicher Textarbeit, wie dem Schreiben von Reden, der Zusammenfassung von Texten oder der Übersetzung von Texten, wird inzwischen auch die allgemeine Bedrohung im digitalen Raum deutlich: Konnte man vor zwei Jahren noch Phishing-Mails an den meist sehr schlechten Formulierungen als solche erkennen, so bietet der Einsatz von KI-Sprachmodellen den Cyberkriminellen die Möglichkeit, Texte ihrer Phishing-Mails glaubwürdiger und authentischer klingen zu lassen, sodass potenzielle Opfer diese auf den ersten Blick nicht mehr als Phishing-Mails erkennen. Neben dem Social Hack der Phishing-Mails befördern KI-Algorithmen auch andere Social Hacking Vektoren.

Diese höhere Professionalität der formalen Gestaltung desinformierender Kommunikation ist auch insofern bedrohlich, als viele Personen die Anmutung der Botschaft (engl. ‚Look and Feel') – also formale Korrektheit, professionelles Erscheinungsbild – als wesentliches Kriterium für die Identifikation von Falschmeldungen verwenden (George, 2024).

Ein Treiber hierbei sind Algorithmen und Dienste, die mit minimalem Ausgangsmaterial Stimmklone von Personen erzeugen können (Schiffer, 2024), welche in ihrer Tonalität und Ausdrucksweise kaum mehr von der realen Person zu unterscheiden sind und von Betrüger:innen für die digitale Variante des Enkeltricks eingesetzt werden. Der erste spektakuläre Millionenraub mittels KI-generierter Videos von Personen (engl. Deep Fake) geschah im Februar 2024 in Hongkong. In einer Videokonferenz wurde der Angestellte eines internationalen Konzerns mittels mehrerer KI-generierter Echtzeit-Videos von Personen aus dem Vorstand dazu gebracht, 24 Millionen Euro an Betrüger zu überweisen (Hurtz, 2024).

Da es im Internet viele frei verfügbare Bild- und Videoquellen von öffentlichen Personen und Entscheidungsträgern gibt, bilden sie eine ideale Grundlage als Trainingsdaten für KI-Video- und KI-Bildanwendungen. Durch das Training mit diesen Daten zur Erzeugung qualitativ hochwertiger KI-generierter Videos wird der Massenproduktion von Desinformationsinhalten Tür und Tor geöffnet. Ein Bedrohungsszenario, in dem sich ein:e falsche Entscheidungsträger:in mit einer/m realen Entscheidungsträger:in austauscht, um an Informationen zu gelangen oder Entscheidungen zu beeinflussen, ist keine graue Theorie: Bereits im Juni 2022 wurde die damalige Berliner Oberbürgermeisterin Franziska Giffey Opfer einer solchen Attacke mittels KI-generierter Echtzeit-Videos, als sie sich mit einem virtuellen Bürgermeister Klitschkow aus Kiew per Videokonferenz unterhielt (ZEIT ONLINE, 2022). Glücklicherweise war der Urheber dieser Aktion ein Comedian, der die Aktion für seine Show genutzt hat.

Wie schwer den Institutionen aktuell der Umgang mit KI-generierten Videos im öffentlichen Raum fällt, zeigte im Jahr 2023 das Beispiel des Videos vom Zentrum für politische Schönheit zum Verbot der AfD mit einem KI-generierten Video von Bundeskanzler Olaf Scholz (Ruch & Pelzer, 2023).

Auch wenn die Desinformation mit KI-generierten Videos besonders beindruckend ist, sollte darüber nicht übersehen werden, dass der Großteil der verwendeten – und durchaus wirksamen – visuellen Desinformation auf wesentlich einfacheren Methoden beruht wie zum Beispiel Verfälschung von realen Grafiken durch unzutreffende Beschriftungen (Weikmann & Lecheler, 2023).

Werden die aufgezeigten Entwicklungen extrapoliert, so ist davon auszugehen, dass in sehr naher Zukunft der größte Teil der Inhalte im Internet durch KI erzeugt wird. Vorreiter dieser Entwicklung ist der Bereich, der expliziten Erwachsenenunterhaltung, in dem seit kurzem verstärkt Webseiten und Apps entstehen, die das einfache Austauchen von Gesichtern in Videos mit expliziten Inhalten erlauben (Meineck, 2022). Diese Entwicklung wird sich mittelfristig nicht nur auf diese Inhalte beschränken, sondern auch für die Erzeugung von KI-generierten Quellen (engl. Fake Sources) genutzt werden, ob als KI-generierte Nutzerprofile (engl. Fake Profiles) auf Online-Plattformen und in Messengerdiensten oder direkt als Fake Source, in der reale und falsche Meldungen miteinander vermischt und ausgespielt werden. An diesem Punkt helfen die Maßnahmen des persönlichen Faktenchecks und der Quellenüberprüfung nur noch eingeschränkt, da in kürzester Zeit so viele Fake Sources durch KI generiert werden können, die ein gesichertes Belegen der Meldung nahezu unmöglich machen.

Wenn gleichzeitig das Auftreten von KI-generierten Videos in digitaler Echtzeit-Kommunikation zunimmt, ob in Chatgruppen, Telefonanrufe oder Videokonferenzen, wird das Vertrauen in die digitale Kommunikation und ihre Medien schwinden. Dies kann zu einer Bedrohung für die Wirtschaft auf globaler Ebene führen.

Handlungsbedarf

Dies macht deutlich, wie groß der aktuelle Handlungsbedarf ist. Dass er noch akuter und größer ist, wird klar, wenn man die Automatisierungspotenziale von Dialogprozessen zwischen KI-Algorithmen und Menschen sowie die Entwicklungen im Bereich von Emotionserkennung durch KI sieht, beispielsweise für Anwendungen in Callcentern (Krempl, 2023).

Nicht umsonst setzen die Betreiber von großen Online-Plattformen und Messengerdiensten seit Jahren schon selbstlernende Algorithmen ein, welche die Aufmerksamkeit der Nutzer:innen binden, indem sie versuchen, deren emotionale Erregung möglichst hoch zu halten. Mag man diesen Einsatz durch Wirtschaftsakteure ethisch und moralisch kritisch sehen, so ist dieser aktuell legal.

Wenn Cyberkriminelle als Trainingsgrundlage für den Dialog mit einer beliebigen Internetnutzer:in ihre auf Online-Plattformen und in Messengerdiensten verfügbaren Profilinformationen sowie sonstigen frei erhältlichen Informationen zur Person aus dem Internet nutzen, können sie schnell eine gezielte Manipulation von Einzelpersonen über dezidierte Chatbots hinter Profilen auf Online-Plattformen und in Messengerdiensten erzeugen. Werden dann noch die Automationspotentiale von digitalen Infrastrukturen genutzt, so lassen sich auch gezielt größere Nutzergruppen auf Social Media Plattformen und sogar darüber hinaus manipulieren und lenken. Seit dem Cambridge Analytica Skandal (Westby, 2019) aus dem Jahr 2018 ist öffentlich bekannt, dass es Manipulationsversuche über Online-Plattformen sozialer Medien als gezieltes Dienstleistungsangebot gibt, dessen Potentiale sich Dank der verbesserten KI-Werkzeuge in den letzten Jahren nochmal gesteigert haben. Diese Entwicklung stellt eine nicht zu unterschätzende Bedrohung für freie demokratische Wahlen und den Zusammenhalt der Gesellschaft dar, birgt sie doch ein hohes Potential für Desinformation.

Desinformation ist eine bekannte taktische Strategie, um eine größere Gruppe gezielt zu manipulieren, und Falschmeldungen auf Online-Plattformen und in sozialen Medien sind kein neues Phänomen. Die Qualität und Quantität hat sich jedoch dank KI-Werkzeugen stark geändert. So waren beispielsweise viele der von Redaktionen eingekauften Bilder zum Gaza-Konflikt im Oktober 2023 künstlich von KI generierte Bilder und stammten nicht von Berichterstattern vor Ort (da Silva & Veréb, 2023).

Der dringlichste Handlungsbedarf zur Begegnung von Desinformation über die systematische und digital automatisierte Einflussnahme zur Informationsmanipulation (engl. Automated Influence) liegt in der Aufklärung und Sensibilisierung. Zur Steigerung der digitalen Resilienz von Bürger:innen ist daher eine breite und sehr zeitnahe Aufklärung notwendig. Man wird aber keine flächendeckenden Fortbildungsmaßnahmen zeitnah zur Sensibilisierung ausrollen können, da selbst die klassischen Bildungsakteure wie (Hochschul-)Lehrer:innen, Eltern oder Lehrinstitutionen gerade den Umgang mit Desinformation und Deep Fakes lernen müssen. Erschwerend kommt hinzu, dass die Qualität der Desinformation sowie die Anforderungen, sie als solche zu erkennen kontinuierlich wachsen.

4 Gegenmaßnahmen zur Desinformation: Entlarvung (engl. Debunking) und Immunisierung (engl. Inoculation, Pre-Bunking)

Psychologisch begründete Gegenmaßnahmen

Gegen Desinformation gerichtete Strategien können darauf abzielen, die Desinformation zu entlarven und dies den Rezipienten gegenüber – durch einfache Warnhinweise oder ausführlichere Erklärungen – zu kommunizieren (Lewandowsky et al., 2020). Ausführliche Erläuterungen wirken oftmals besser als kurze Warnhinweise, können paradoxerweise aber auch die Persistenz der Desinformation erhöhen (Chan et al., 2017). Bloße Warnhinweise wirken nicht bei allen Personen gleich gut. So neigen etwa Personen, die dem politischen Spektrum rechts der Mitte zuneigen, einerseits besonders stark dazu, Falschmeldungen für wahr zu halten, andererseits zeigen bei diesen Personen Warnhinweise wenig Wirkung (Arendt et al., 2019).

Zur effektiven Gestaltung von Kommunikation zur Entlarvung von Desinformation (engl. Debunking) gibt es umfangreiche forschungsgestützte Handreichungen und ausgearbeitete Methoden (Lewandowsky et al., 2020). Eine generelle Herausforderung bei reaktiven Strategien wie dem Debunking besteht darin, Desinformation möglichst automatisiert und möglichst sicher erkennen zu können (Aïmeur et al., 2023; Bontcheva, et al., 2024).

Präventive Maßnahmen zielen darauf ab, potenzielle Rezipienten von Desinformation schon im Vorfeld gegen solche Manipulationen zu immunisieren (engl. Pre-Bunking, Inoculation (Lewandowsky & van der Linden, 2021; Lu et al., 2023)). Solche präventiven Interventionen sind im konkreten Fall der akuten Desinformation oftmals auch wirksamer als bloße Warnhinweise (McPhedran et al., 2023). Die konkreten Interventionen beinhalten etwa Informationen zu Strategien und rhetorischen Mustern der Desinformation oder leiten an zu kritisch-reflektiertem Denken, um den potenziellen Rezipienten beim Erkennen von Desinformationen zu helfen (Lewandowsky et al., 2020; Lewandowsky & van der Linden, 2021). Die Analogie zur medizinischen Immunisierung – der Impfung – besteht darin, dass Personen nach einer Vorwarnung mit ‚abgeschwächten' Varianten der Desinformation konfrontiert werden, verbunden mit Hilfestellungen, die darin enthaltenen rhetorischen Muster und Desinformationsstrategien zu erkennen (Maertens et al., 2021; Maertens et al., 2023; Roozenbeek et al., 2022).

Es wurden Videos zur präventiven Immunisierung entwickelt, die Desinformationsmethoden kompakt darstellen; sie erzielten eine gute Wirkung hinsichtlich der Fähigkeit, zutreffende von nicht zutreffenden Informationen unterscheiden zu können (Roozenbeek et al., 2022). Nicht alle Methoden zur Immunisierung sind allerdings gleichermaßen wirksam. So wurden etwa im Kontext von Desinformation zum Klimawandel Immunisierungsstrategien erprobt, die etwa auf die Betonung der Vertrauenswürdigkeit der Forschergemeinschaft abzielen oder mit moralischen oder emotionalen Interventionen arbeiten; hier wurden keine bedeutsamen Effekte festgestellt. (Spampatti et al., 2024). Hinsichtlich der Nützlichkeit von Video*spielen* – gegenüber erklärenden Videos, – gibt es sowohl positive (Appel et al., 2024) wie auch kritische Ergebnisse (Modirrousta-Galian & Higham, 2023).

In jedem Fall lässt die Wirkung von Immunisierungsmaßnahmen über die Zeit nach – im Fall von text- oder videobasierten Methoden nach ungefähr einem Monat, nach spielebasierten Methoden schneller. Über ‚Auffrischungsdosen' (engl. Booster Shots) kann die nachlassende Wirkung psychologischer Immunisierung wieder verstärkt werden (Maertens et al., 2021; Maertens et al., 2023).

In ihrem Scoping Review zu psychologisch begründeten Gegenmaßnahmen zur Desinformation unterscheiden Ziemer und Rothmund fünf große Gruppen solcher Maßnahmen; dies soll hier als Zusammenfassung und Einordnung aufgegriffen werden (Ziemer & Rothmund, 2024). Die englischen Gruppenbezeichnungen stehen für:

1. **Boosting:** Verstärkung kognitiver Fähigkeiten, entweder durch Vermittlung von Faktenwissen (engl. Knowledge) oder durch generalisierte kognitive Fähigkeiten (engl. Literacy)

2. **Inoculation:** Psychologische Impfung (wie oben beschrieben) durch Konfrontation mit ‚abgeschwächten' Desinformationen nach Warnhinweisen, um typische Muster erkennen zu können

3. **Nudging:** Hinweise auf hilfreiche Verhaltensweisen, wie etwa laterales Lesen, d.h. mehrere Quellen berücksichtigen und vergleichen

4. **Fact Checking:** Von der einfachen Identifikation falscher Informationen (engl. Flagging) bis zu ausführlichen, methodisch ausgearbeiteten Erläuterungen (engl. Debunking)

5. **Identitiy Management:** Dies sind noch relativ seltene Maßnahmen, die darauf abzielen, Herausforderungen der jeweiligen personalen und /oder sozialen Identität durch Informationen, die mit verfestigten Überzeugungen konfligieren, abzumildern. Dies ist auch insofern besonders interessant, als Desinformation nach einer Aussäh-Phase

Übersicht der Gegenmaßnahmen zur Desinformation

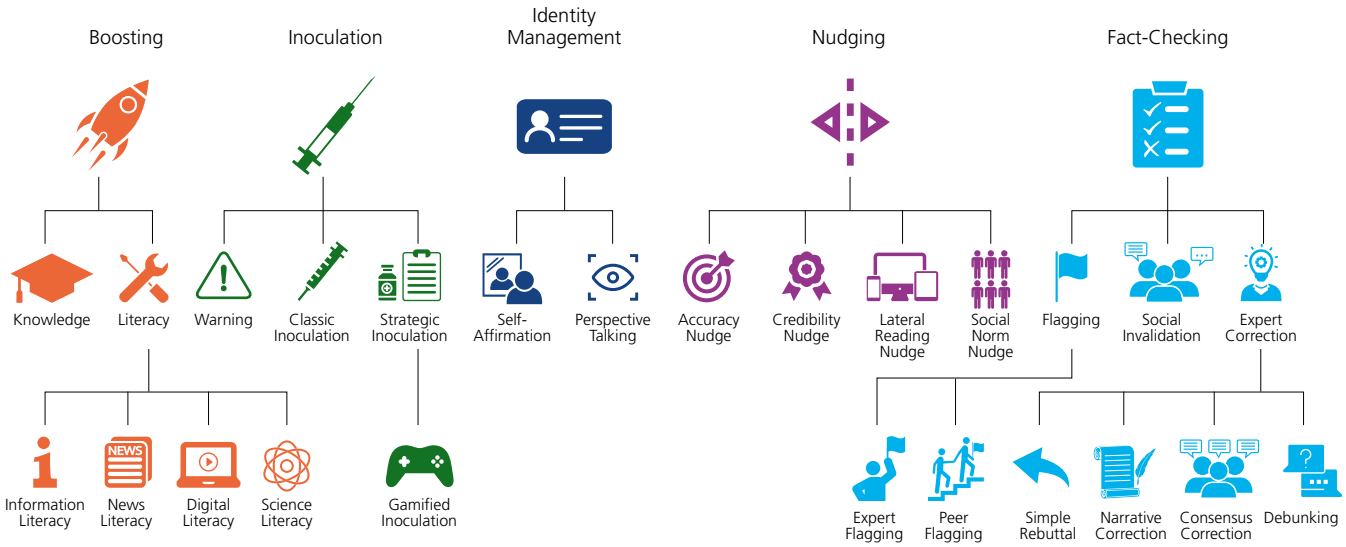

Abbildung 2: Übersicht der Gegenmaßnahmen zur Desinformation (eigene Darstellung nach Ziemer & Rothmund, 2024)

(engl. Seeding) in nach Identitätsmerkmalen wie Geschlechtsidentität, Region, Religion, Weltanschauungen … etc. strukturierten Echo-Kammern verstärkt und verfestigt werden (Echo-Phase); diese Echo-Phasen werden generell in der Diskussion um Gegenmaßnahmen zu Desinformation noch zu wenig thematisiert (Diaz Ruiz & Nilsson, 2023). Einige Gegenmaßnahmen im Kontext des Identitätsmanagements wurden weiter oben bereits dargestellt.

Neue Situation durch KI-generierte Desinformation

Vor dem Hintergrund der oben dargestellten Befunde zur Wahrnehmung und Verbreitung von Desinformation sowie möglicher Gegenmaßnahmen wird hier erörtert, welche neuen Phänomene und Herausforderungen durch das Aufkommen KI-generierter Desinformation zu berücksichtigen sind. Dazu werden zunächst einige grundlegende Befunde dargestellt.

Menschen sind heute weitgehend nicht in der Lage, KI-generierten Text zu unterscheiden von Text, der von Menschen erzeugt wurde – beide Arten von Text werden für vergleichbar glaubwürdig gehalten (Kreps et al., 2022; Stiff & Johansson, 2022).

Ganz grundsätzlich haben visuelle bzw. audiovisuelle gegenüber rein textlichen Botschaften eine höhere Wirksamkeit, indem sie mit höherer Wahrscheinlichkeit Personen von der Richtigkeit der jeweiligen Aussagen überzeugen können (Hameleers et al., 2020; Sundar, 2008). Vor diesem Hintergrund bekommen KI-generierte audiovisuelle Medien (Audio, Foto, Video) eine besondere Brisanz.

So konnte gezeigt werden, dass durch ein manipuliertes Video, in dem einem Politiker für ihn abträgliche Äußerungen ‚in den Mund gelegt' wurden, tatsächlich ein messbarer Ansehensverlust dieses Politikers bei den Versuchspersonen hervorgerufen werden kann. Dieser Effekt ‚färbte allerdings nicht ab' auf die Partei, der der Politiker angehört. Hier haben offenbar längerfristige politische Orientierungen und Identitäten einen inokulierenden Effekt (Dobber et al., 2021).

Weitere Befunde zur Wirkung KI-erzeugter Videos zeigen auf, dass möglicherweise der Haupteffekt gar nicht darin besteht, die Rezipienten zu *täuschen* – in dem Sinne, dass sie die gefakte Botschaft für wahr halten würden –, sondern eher oder noch mehr in einer zunehmenden Unsicherheit hinsichtlich des Wahrheitsgehaltes der Botschaften bzw. der Echtheit der Videos. Dies kann nun zu zwei ganz unterschiedlichen Konsequenzen führen (Vaccari & Chadwick, 2020):

2 Die Begriffe wurden in der englischen Sprache belassen, da die Verständlichkeit durch die Übersetzung ins Deutsche nicht steigt.

- **Die Unsicherheit führt zu Zynismus:** „Die lügen und betrügen doch eh alle!" Dies könnte zu einem zunehmenden Vertrauensverlust bezüglich der Medien führen, was grundsätzlich alle Medien – seriöse wie unseriöse – betreffen könnte. Dies wiederum würde weiterer Desinformation Tür und Tor öffnen und zur Polarisierung unserer Gesellschaften sowie zur Verletzbarkeit der Demokratie erheblich beitragen.

- **Die Unsicherheit führt zu Skeptizismus:** „Ich muss genauer prüfen – wenn ich mich nicht mehr auf meine Augen verlassen kann, muss ich mich umso mehr auf mein Hirn verlassen können!" Eine solche Entwicklung könnte positive Auswirkungen im Hinblick auf eine höhere Resilienz gegenüber Desinformation haben. Insbesondere würde dadurch die potenziell besonders schädliche Wirkung audiovisueller Botschaften möglicherweise limitiert.

Diese beiden – zunächst hypothetischen – Entwicklungsrichtungen würden völlig unterschiedliche politische Bewertungen und Handlungsoptionen nach sich ziehen. Hierzu besteht aktuell erheblicher Forschungsbedarf.

Zusätzlich soll auf einen bereits weiter oben dargestellten Befund zurückgegriffen werden: Personen verwenden generell bevorzugt Quelle (Seriosität) und formale Gestaltung (engl. 'Look and Feel') als wesentliche Merkmale zur Identifikation von Falschmeldungen (George, 2024). Hier eröffnen sich angesichts der Möglichkeiten KI-erzeugter Desinformation folgende neue, bedrohliche Sachverhalte:

- Durch KI werden Inhalte der Desinformation in Form von Text und audiovisueller Aufbereitung qualitativ sehr viel besser und wirken sehr viel professioneller. Das Merkmal *‚Look and Feel'* eignet sich immer weniger zur Identifikation von Desinformation.

- Auch *Quellen* können ‚gefakt' werden. Als Gegenstrategie hilft hier eine eigenständige, unabhängige Recherche in vertrauenswürdigen und seriösen Medien. Dies ist aber eine relativ aufwändige Strategie, die dem Mediennutzungsverhalten vieler Menschen nicht entgegenkommt, und weiterhin …

- … können KI-generierte Videos und Bilder auch, wegen mangelnder journalistischer Sorgfalt, in *Qualitätsmedien* auftauchen. So tauchten beim Ausbruch der Kämpfe im Gazastreifen in 2023 zahllose KI-generierte Fotos auf, die u. a. vom Bildanbieter Adobe Stock verkauft und von Medien übernommen wurden (da Silva, G, Veréb, D., 2023).

Abschließend sollen die oben dargestellten fünf generellen, psychologisch begründeten Gegenstrategien (Ziemer & Rothmund, 2024) vor dem Hintergrund der neuen Situation, wie sie durch KI-generierte Desinformation hervorgerufen wird, reflektiert werden:

1. **Fact Checking** wird weiterhin notwendig sein und muss zunehmend selbst KI-basierte Methoden verwenden, um KI-generierte Desinformation aufspüren zu können (Aïmeur et al., 2023; Bontcheva, et al., 2024). Dabei ist ein ‚Rüstungswettlauf' zwischen Desinformation und Debunking zu erwarten, aber auch kaum zu vermeiden.

2. In ähnlicher Weise müssen im Rahmen der ‚Impfung' (engl. **Inoculation**) die jeweils aktuellsten Methoden der Desinformation für die Aufbereitung des ‚Impfstoffs' verwendet werden, es müssen also auch KI-basierte Desinformationen zur Demonstration der Effekte eingesetzt werden. Ein Beispielfall aus Taiwan wird im Beitrag dargestellt.

3. **Boosting** – Aufbau und Entwicklung traditioneller und digitaler Medienkompetenzen – wird immer wichtiger, weil sich das Korpus des notwendigen Wissens, um Desinformation erkennen zu können, ständig erweitert und schnell erneuert.

4. **Nudging** wird ebenfalls bedeutsamer, in zweierlei Hinsicht: Für das individuelle Erkennen und Entlarven von Desinformation wird es angesichts KI-basierter Desinformation immer wichtiger, konkurrierende Quellen[3] zu recherchieren (engl. ‚Lateral Reading Nudge') – bei allen Limitationen, die oben dargestellt wurden. Zweitens kann durch ‚Accuracy Nudging' sowohl die individuelle Wahrnehmung von Desinformation wie auch das Weitergabeverhalten gegenüber anderen positiv beeinflusst werden. Dieser zweite Aspekt ist nicht spezifisch für KI-generierte Desinformation, aber besonders bedeutsam als einerseits relativ einfache und wirksame, andererseits noch zu wenig genutzte Maßnahme.

5. In ähnlicher Weise sind die Maßnahmen zu **Identity Management** nicht KI-spezifisch, aber von großer Bedeutung sowie noch unterkritisch erforscht und entwickelt (Ziemer & Rothmund, 2024). Bedeutsam sind sie, weil sie im Kontext von Echokammern einerseits die Verfestigung und Weitergabe von Desinformation beeinflussen, andererseits die Voraussetzungen für wirksame Entlarvung von Desinformation (engl. Debunking) und Immunisierung schaffen können.

3 Interessant sind in diesem Zusammenhang: LLM-basierte Dienste wie Perplexity (www.perplexity.ai), die Aussagen mit verlinkten Quellen unterlegen.

5 Maßnahmen gegen Desinformation am Beispiel Taiwan: Best Practice für zukunftsweisende Demokratie im Zeitalter von Automated Influence?

Automated Influence soll Vertrauen in Demokratie schwächen

Automated Influence zielt als politisch motivierter Angriff nicht darauf ab eine Online-Plattform oder ein soziales Medium zu bekämpfen, sondern verfolgt das Ziel, dass die Menschen kein Vertrauen mehr in die demokratischen Institutionen haben.

Das Schlachtfeld befindet sich nicht in einem bestimmten Cybersystem, das repariert und gepatcht werden könnte, sondern im eigenen Kopf. Wenn Desinformation die Art von Angst, Unsicherheit und Zweifel erzeugt, die die Gesellschaft polarisiert und einen Teil der Gesellschaft dazu bringt, die andere Seite der Gesellschaft für dieses Chaos verantwortlich zu machen, dann hinterlässt dies ein gegenseitiges Misstrauen, das schwer abbaubar ist.

Die grundlegende Strategie gegen Desinformation und Propaganda in digitalen Medien kann darin bestehen, eine alternative digitale Plattform anzubieten, die demokratische Diskussionsprozesse ermöglicht und fördert.

Für eine breitflächige und gesellschaftsweite Sensibilisierung zu Desinformation im Internet mag Taiwan als Vorbild dienen, das während seiner Wahl im Januar die massiven Desinformationskampagnen aus China erfolgreich abgewehrt hat (Wu, 2024).

Hier findet sich ein sehr gutes und umfassendes Beispiel für eine grundlegende Strategie. Mit vTaiwan („v" für „vision", „voice", „vote" und „virtual") wurde ein offener digitaler Konsultationsprozess geschaffen, der seit 2014 in Taiwan etabliert ist (Ho, 2022; Hsiao et al., 2018; Lee, 2020).

vTaiwan: digitaler Konsultationsprozess gegen Desinformation

Hintergrund für die Entstehung von vTaiwan waren die Demokratisierungs- und Digitalisierungsprozesse in Taiwan seit den 1990er-Jahren. Nachdem die jahrzehntelang dominierende Partei Kuomintang ihre Vormachtstellung verloren hatte, wurde nach neuen Wegen der demokratischen Debatte, Konsultation und Partizipation gesucht (Ho, 2022; Hsiao et al., 2018). Dass dabei besonders digitalisierte Ansätze berücksichtigt wurden, liegt an Besonderheiten der taiwanesischen Gesellschaft und Kultur. So etablierten sich seit den frühen 2010er-Jahren Bottom-up-Initiativen in der Gestalt von Civic Tech Communities, die digitale Technologien für die Bürgergesellschaft nutzbar machen wollten. Von diesen Communities wurde g0v („gov zero") besonders bekannt und besonders wirksam, insbesondere auch als Entwickler von vTaiwan, ausgelöst durch eine Anregung der damaligen (2014) Ministerin Jaclyn Tsai (Ho, 2022; Hsiao et al., 2018; Lee, 2020). Um die Jahrtausendwende entfaltet sich zudem ein sehr dynamischer Prozess einer umfassenden e-Government-Infrastruktur in Taiwan (Biberman, 2021).

Der Diskussions- und Mitgestaltungsprozess in vTaiwan verläuft in vier Phasen, die allerdings in Dauer und methodischer Ausgestaltung variabel sind (Hsiao et al., 2018)[4]. Insbesondere die Phasen 1 und 3 werden moderiert.

1. **Vorschlag (engl. proposal):** Wöchentliche, öffentliche online-offline Mini-Hackathons, in denen mit kollaborativen Echtzeit-Texteditoren Ideen gesammelt werden. Die Regierung oder Behörden können entscheiden, welche dieser Themen weiterverfolgt werden sollen und welche nicht. Die jeweils zuständige Behörde ist dann auch für dieses Thema durch alle vier Phasen verantwortlich.

2. **Meinungsbildung (engl. opinion):** Es werden online Diskussionsprozesse organisiert, unter Verwendung von IT-Systemen wie Polis[5]. Die Meinungen werden erfasst und strukturiert dargestellt. Mögliche Konsens-Elemente werden deutlich.

3. **Reflexion (engl. reflection):** In einer öffentlichen Online-Konferenz kommen unterschiedliche Gruppen zusammen, um die strukturierten Meinungen aus der Meinungsbildungsphase zu diskutieren: Teilnehmer aus der Online-Befragung, Vertreter von Interessengruppen, Regierung, Wissenschaft, Wirtschaft, zivilgesellschaftliche Organisationen.

4 Ein sehr gut dokumentierter Anwendungsfall dieser Methode ist die gesetzliche Regulieren von Personenbeförderungsdiensten, die durch das Aufkommen von Uber notwendig wurde

5 https://pol.is/home

4. **Gesetzgebung (engl. legislation):** Je nach Art des erreichten Konsenses schließt der Prozess mit einer Stellungnahme, Richtlinie oder anderer Regulierung der zuständigen Behörde ab, oder ein Gesetzgebungsverfahren im Taiwanesischen Parlament (Legislativ-Yuan) wird angestoßen.

Die spezifische Ausgestaltung dieses Prozesses ist sehr in der taiwanesischen Geschichte und Kultur verwurzelt. Dies muss bei Versuchen, in anderen Ländern vom taiwanesischen Beispiel zu lernen, sehr gründlich und reflektiert berücksichtigt werden. Dabei helfen Studien, die IT-gestützte politische Diskurs- und Beteiligungsverfahren kulturvergleichend analysieren (Tseng, 2022, 2023).

Einbettung der Maßnahmen in eine Verteidigungsstrategie gegen Desinformation

Die konkreten Gegenmaßnahmen, welche von Taiwan zum Bekämpfen von Desinformation während der Präsidentschaftswahl im Januar 2024 zum Einsatz kamen, wurden von Audrey Tang, der Ministerin für digitale Angelegenheiten Taiwans, in einem Interview erläutert (Center for Humane Technology, 2024).

Ein Hauptbestandteil der Strategie ist, dass Vertreter:innen aus Wissenschaft und Technologie Strategien zur Informationsmanipulation bekannt machen und hierzu aufklären, noch bevor sie breit zum Einsatz kommen können (engl. Pre-Bunking). Mit Blick auf die Wahlen im Januar 2024 hat Audrey Tang bereits im Jahr 2022 ein KI-generiertes Video von sich und ihrem Wissenschaftsrat erstellen lassen und als Deep-Fake-Szenario öffentlich vorgestellt: Gezeigt wurde das „Making of" des offiziellen Deep Fakes der Ministerin und auch, wie einfach es ist, dies mit einem Laptop selber zu produzieren und dass es in Zukunft auf jedem Mobiltelefon ebenso einfach sein wird.

Pre-Bunking ist eine Form der ‚Impfung' (engl. Inoculation) und braucht eine gewisse Zeit, um zu wirken sowie eine regelmäßige Auffrischung. Daher wurde zur Inoculation der Film zur Erzeugung von KI-generierten Videos in den Jahren 2022 und 2023 wiederholt ausgestrahlt. Das Ergebnis: Immer wenn KI-generierte Videos während der Wahlkampfsaison auftauchten, hatten sie keine große Wirkung, weil die Menschen bereits seit zwei Jahren sensibilisiert worden waren.

Taiwan konnte auch Wahlbetrugsvorwürfe vollständig entkräften, die als Zweifel gleich nach der Wahl kamen. Dies geschah mittels einer Transparenz-Strategie. Was auch immer die Anschuldigung war, man konnte in den Wahllokalen immer drei verschiedene YouTuber finden, die den drei verschiedenen Parteien angehörten und durch ihre parallelen Liveaufzeichnungen eine genaue Dokumentation der Auszählungen bereitstellten.

Dies war nur ein Baustein der taiwanesischen Verteidigungsstrategie gegen Desinformation im Rahmen der Wahlen. Zur Entwicklung einer Verteidigungsstrategie für den digitalen Raum muss man laut Audrey Tang die drei Ebenen der Informationsmanipulation verstehen (Center for Humane Technology, 2024):

- **Akteursebene:** Wer handelt hier?
- **Verhaltensebene:** Handelt es sich um Millionen koordinierter unechter Handlungen oder nur um einen einzigen Akteur?
- **Inhaltsebene:** Ob der Inhalt gefälscht oder echt ist, lässt sich allein anhand des Inhalts nicht feststellen. Die Menschen sollen daher anhand des Verhaltens oder des Akteurs erkennen, ob der Inhalt vertrauenswürdig ist.

Im Rahmen der Strategie wurde ein staatliches Identity Management umgesetzt. Zur Herstellung einer vertrauensvollen staatlichen Quelle werden in Taiwan seit 2024 alle Kurznachrichten von allen öffentlichen Institutionen, egal ob es die Wasserwerke, die Elektrizitätswerke, die Verwaltungen oder andere sind, von derselben Nummer aus verschickt: der 111. Egal, ob Bürger:innen an die Steuer erinnert wird oder zur Teilnahme an einer Umfrage aufgefordert werden, die Kommunkation läuft komplett per SMS über diese Nummer. Da sie nicht zu fälschen ist und in Taiwan Nummern normalerweise zehnstellig sind, lässt sich sehr einfach am Absender erkennen, dass die Nachricht aus einer vertrauenswürdigen Quelle stammt. Inzwischen gehen auch taiwanesische Telekommunikationsunternehmen und Banken dazu über, eigene Kurzcodes einzusetzen. Hierdurch entwickeln sich zwei Klassen von Vertrauenszentren: Die einen sind fälschungssicher und garantiert vertrauenswürdig, die anderen erfordern ein Kennenlernen von Angesicht zu Angesicht und die Aufnahme in das eigene Adressbuch, bevor bestätigt werden kann, dass es sich bei dem Gegenüber tatsächlich um die bestimmte Person handelt.

Auch wurde Fact Checking in der Strategie berücksichtigt. So gibt es mit Cofacts eine kollaborative Faktenüberprüfung, bei der jeder eine Nachricht als möglichen Betrug oder Spam an die Chat-Gruppe melden kann. Das bedeutet, dass es eine Echtzeit-Stichprobe davon gibt, welche Informationspakete viral gehen. Einige dieser Stichproben sind Informationsmanipulationen, andere sind tatsächlich wahr, sodass sich hierdurch eine Echtzeitkarte derjenigen Inhalte ergibt, die in diesem Moment viral gehen. Ein Crowdsourcing der Faktenüberprüfung – analog zum Prinzip bei Wikipedia, nur in Echtzeit – erlaubt nicht nur die Verfolgung des Informationspakets, sondern auch die des einhergehenden Pre-Bunking und Debunking. Mit neueren Methoden des Trainings von Sprachmodellen, wie der direkten Präferenzoptimierung, wird die Logik dessen herausgefunden, was angenommen und was abgelehnt wird.

Ein weiterer Baustein sind deliberative Umfragen. Das Prinzip stützt sich im Wesentlichen auf direkte Nachrichten und spricht mit einer sehr individualisierten Gruppe von Menschen auf Basis eines Modells ihrer Vorlieben. Für Umfragen wird über die 111 eine Zufallsstichprobe von 4.000 oder 10.000 Personen kontaktiert, um ihnen die gleichen Fragen zur Ermittlung ihrer Präferenzen zu stellen wie der Kontrollgruppe. Die Umfragen werden bei Wahlen, aber auch bei der Politikgestaltung eingesetzt.

Das erste Mal wurde dieser Ansatz eingesetzt, als im Jahr 2015 der Mobilitätsdienstleister Uber neu nach Taiwan kam. Es gab Proteste wie in anderen Ländern auch. Die Regierung bat die Uber-Fahrer:innen, die Taxifahrer:innen, die Fahrgäste und sonstige Menschen das soziale Online-Medium Polis zu besuchen. Der Unterschied zu anderen großen Online-Plattformen ist, dass hier anstatt die am meisten klickenden, polarisierenden und sensationslüsternen Ansichten hervorzuheben, nur die Ansichten veröffentlicht werden, die Unterschiede überbrücken.

Polis wie auch Community Notes auf X.com haben als grundlegendes Design gemeinsam, dass es keinen Antwort-Button gibt. Durch den Algorithmus der Überbrückungsboni gibt das System den überbrückenden Aussagen immer mehr Sichtbarkeit, sodass die Menschen immer längere Brücken bauen können, die über immer größere Unterschiede zwischen den Ideologien, Erfahrungen und so weiter überbrücken.

6 Falschmeldungen in einer grundlegend gewandelten Medienlandschaft in Deutschland

Das Angebot der Online-Plattform Polis reagiert dabei auch auf ein weiteres Phänomen, das zur Fragmentierung der Gesellschaft und zur generellen Anfälligkeit gegenüber Falschmeldungen auch unabhängig von KI-generierten Videos beiträgt. Während bis Anfang der 2000er-Jahre die Medienlandschaft zwar durch eine gewisse Vielfalt, aber dennoch auch Überschaubarkeit in Form von Leitmedien geprägt war, ist die Medienlandschaft heute eine vollkommen andere Welt. In den 1990ern gab es noch sogenannte Straßenfeger im linearen Fernsehen, die am nächsten Tag gleichsam Stadtgespräch waren. Durch Streamingdienste für Filme oder Serien und Musik sowie durch Internet-Nachrichten, Podcasts und Kanäle auf Online-Plattformen und in Messengerdiensten etc. ist die Landschaft heutzutage unüberschaubar. Kein öffentlich-rechtlicher oder privater Rundfunk und keine große Tages- oder Wochenzeitung bietet heute den Resonanzboden für eine breite öffentliche Meinungsbildung. Die enorme Individualisierung von Vorlieben und Angeboten spiegelt sich nicht nur in Formaten, sondern auch Inhalten wieder. Das schlägt auch auf die Qualität durch. Für viele junge Medienkonsument:innen sind Telegram-Kanäle gleichwertig zur Tagesschau; mitunter wird den Qualitätsmedien sogar weniger getraut als einzelnen News-Influencer:innen (Bayer AG, 2022). Dementsprechend ist hier die Gefahr größer, dass Verschwörungserzählungen und Falschmeldungen Zustimmung und Weiterverbreitung finden. Angeheizt wird dieser Prozess noch durch das Phänomen des „Pink Slime"-Journalismus, dem man den Fake Sources zuordnen kann. Dabei handelt es sich um Medienseiten im Internet, deren Inhalte frei erfunden und Falschmeldungen sind, aber vom Erscheinungsbild realen Medien entsprechen. Realität ist der „Pink Slime" insbesondere seit rund zehn Jahren in den USA, wo jede Woche im Schnitt zwei Regionalzeitungen sterben – mit weitreichenden Folgen für das Funktionieren der lokalen Gemeinschaften (Abernathy, 2022). Der Lokaljournalismus gilt als Grundpfeiler des US-amerikanischen Demokratie und erfüllt als lokaler „Wachhund" wichtige Aufgaben. Umfragen zeigen, dass die Bürger:innen ein erkennbar höheres Vertrauen in lokale als in nationale Medien haben. In diese Lücke stoßen die sogenannten Partisanen-Medien, die von Anbietern wie Metric Media mit mehr als 1.300 sogenannten „Community Media Sites" einen automatisierten Billig-Journalismus anbieten, um insbesondere in den Wahlzyklen parteiische Inhalte zu publizieren: *„Der Fall Metric Media veranschaulicht, wie ein komplexes Netzwerk von konservativen und wirtschaftsfreundlichen Interessengruppen, Geldgebern und politischen Kandidaten in einer undurchsichtigen Operation miteinander verflochten sind. Damit werden interessengeleitete Inhalte über eine Reihe von Websites, Anzeigen in sozialen Medien und gedruckten Mailings und getarnt als lokale Nachrichten verbreitet.*[6]" (Tow Center for Digital Journalism, 2024).

6 eigene Übersetzung

7 Fazit und Ausblick

Die Manipulation von Medieninhalten ist so alt wie die Medien selbst. Gab es nur eine einzige Medienquelle, konnte diese sehr einfach und ohne Möglichkeit zur Verifikation durch andere Kanäle für Falschmeldungen und Propaganda genutzt werden. Mit durchschlagendem Erfolg: In der Geschichte der Medienvielfalt des 20. Jahrhunderts hatte sich durch die – gerade noch überschaubare – Vielfalt an Zeitungen, Magazinen, TV- und Radiosendern ein wettbewerbliches Gleichgewicht herauskristallisiert, das sowohl die prägenden Medienkonzerne als auch die Sensationspresse – oftmals in Union – auf ein Restmaß an Meinungsvielfalt und Faktenorientierung verpflichtete. Mit dem Aufkommen des Internets und Online-Plattformen sowie Messengerdiensten fand ab der Jahrtausendwende eine Abkehr vom gefestigten Mediengefüge und dessen Konsum statt – jede:r konnte plötzlich Nachrichten machen. Die grundlegenden Herausforderungen von Online-Plattformen sozialer Medien wurden angesichts von Cyber-Mobbing etc. sehr wohl erkannt und so entstand das Konzept der Medienkompetenz bzw. der Medienmündigkeit, die – vorwiegend, aber nicht ausschließlich – jungen Menschen einen reflektierten Umgang sowohl aktiv als auch passiv mit der neuen Medienvielfalt vermitteln soll. Ein wichtiger Baustein hierin ist die Rezeptionskompetenz und damit die Fähigkeit, Medien kritisch zu nutzen.

Die Entwicklung von Medienkompetenz ordnet sich ein in die fünf Kategorien von Maßnahmen gegen Desinformation, die in der Fachliteratur diskutiert werden:

1. Fact-Checking wird in Zukunft verstärkt mit KI-Tools unterstützt und (halb-)automatisiert durchgeführt werden müssen, um der zunehmend besser werdenden Qualität und der hohen Anzahl KI-generierter Falschnachrichten gerecht werden zu können. Hier besteht einerseits technischer Forschungs- und Entwicklungsbedarf, andererseits müssen entsprechende Kampagnen in sozialen Medien konzipiert und gefördert werden.

2. Eine weitere Maßnahmengruppe betrifft die kognitive ‚Impfung' durch kontrolliertes Konfrontieren mit Falschinformationen und entsprechende Aufklärung, wie oben am Beispiel Taiwans gezeigt, wo KI-erzeugte Deep Fake-Videos von Regierungsstellen gezielt zum Zweck des De-Bunking veröffentlicht wurden. Dies könnte einerseits in die oben angesprochenen Maßnahmen zur Kompetenzentwicklung eingebunden werden, andererseits wären auch hier spezifische Kampagnen in sozialen Medien gut vorstellbar, bzw. spezifische Elemente der oben angesprochenen Kampagnen.

3. Die bereits benannte Medienkompetenz – aber auch generelles Wissenschafts- und Fachwissen, das beim Erkennen von Desinformation hilft – sind bewährte Ansatzpunkte für Maßnahmen gegen Desinformation. Eine besonders wichtige Querschnittskompetenz ist das analytische Denken, das in besonderem Maße vor Desinformation schützt. Konkrete Maßnahmen können bestehen in Good-Practice-Wettbewerben für die schulische und außerschulische Bildung sowie der Entwicklung, Bereitstellung und breiten Kommunikation von offenen Bildungsmedien (Open Educational Resources) zur Medienkompetenz und den anderen genannten Kompetenzfacetten.

4. Nudging geht über die üblichen Methoden der Kompetenzentwicklung hinaus, indem durch gezielte, kleinteilige Interventionen direkt Verhaltensänderungen ausgelöst werden sollen. Hier besteht nicht nur weiterer psychologischer Forschungsbedarf, noch bedeutsamer ist eine politische Auseinandersetzung darüber, welche Randbedingungen für solche Interventionen in offenen demokratischen Gesellschaften geklärt werden müssen und wie die notwendige Transparenz gegenüber der Öffentlichkeit hergestellt werden soll.

5. In noch stärkerem Maße gilt für Maßnahmen des Identitätsmanagements – der direkten Adressierung von etwa politisch oder weltanschaulich ausgerichteten Identitäten –, dass einerseits noch ein sehr großer Forschungsbedarf besteht, andererseits die politische Begründung und Kommunikation keineswegs trivial ist.

So wichtig diese aus der Forschung ableitbaren Maßnahmen auch sind, darf sich der Kampf gegen demokratiegefährdende Desinformation nicht darauf beschränken, das Problem und seine Lösung in den Schulen, Hochschulen und Forschungseinrichtungen abzuladen. Angesichts der durch Digitalisierung und KI in Qualität und Quantität rasant ansteigenden Flut von Falschmeldungen – im Sinne von Mis- und Disinformation – müssen umfassende Maßnahmen entwickelt werden, die das Individuum einerseits in seiner Fähigkeit zur Wahrnehmung, Analyse und Einordnung von Nachrichten stärken („Immunisierung"), es andererseits mit dieser Aufgabe auch im Alltag und außerhalb geschützter Räume wie Bildungseinrichtungen nicht allein lassen. Das Beispiel Taiwan zeigt, wie durch ein durchdachtes Zusammenspiel aus unterschiedlichen Maßnahmen ein wirksames Gesamtpaket gegen den medialen Frontalangriff auf Demokratie und gesellschaftlichen Zusammenhalt geschnürt werden kann. Dabei ist ohne Frage zu prüfen, in welcher Form und in welcher

Abhängigkeit von kulturellen und sonstigen Bedingungen ein solches Maßnahmenpaket in Deutschland umgesetzt werden kann. Deutlich wird aber:

- Dieser Bedrohung kann nicht mit halbherzigen und verstreuten Einzelaktivitäten begegnet werden – es braucht einen umfassenden Ansatz.

- Rasche Effekte sind nicht zu erwarten – das „medienimmunologische Gedächtnis" der Bevölkerung bildet sich erst langsam heraus und muss immer wieder aufgefrischt werden.

- Vertrauen und vertrauenswürdige Institutionen sind Schlüsselfaktoren – daher sind gerade diese Institutionen vor einer Aushöhlung durch Populisten und Demokratiefeinde zu schützen.

8 Literatur

Abernathy, P. M. (2022). *News Deserts and Ghost Newspapers: Will Local News Survive? The Center for Innovation and Sustainability in Local Media.* University of North Carolina at Chapel Hill. https://www.usnewsdeserts.com/wp-content/uploads/2020/06/2020_News_Deserts_and_Ghost_Newspapers.pdf

Aïmeur, E., Amri, S. & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: a review. *Social network analysis and mining*, *13*(1), 30. https://doi.org/10.1007/s13278-023-01028-5

Appel, R. E., Roozenbeek, J., Rayburn-Reeves, R. M., Basol, M., Corbin, J. C., Compton, J. & van der Linden, S. (2024). *Psychological inoculation improves resilience to and reduces willingness to share vaccine misinformation.* https://doi.org/10.31234/osf.io/ek5pu

Arendt, F., Haim, M. & Beck, J. (2019). Fake News, Warnhinweise und perzipierter Wahrheitsgehalt: Zur unterschiedlichen Anfälligkeit für Falschmeldungen in Abhängigkeit von der politischen Orientierung. *Publizistik*, 64(2), 181–204. https://doi.org/10.1007/s11616-019-00484-4

Bayer AG. (2022). *Vertrauensstudie 2022: Angst vor der Zukunft? Jugendliche zwischen gesunder Skepsis und gefährlicher Verschwörungsneigung. Bundesweite Befragung von Kindern und Jugendlichen durch die Universität Bielefeld im Auftrag der Bepanthen-Kinderförderung.* Bayer AG. https://www.bepanthen.de/sites/g/files/vrxlpx36091/files/2022-08/Bepanthen-Kinderfoerderung_Vertrauensstudie2022_Ergebnispr%C3%A4sentation.pdf

Biberman, J. (2021). *E-Governance and Civic Technology: Lessons from Taiwan* (ICT India Working Paper, No. 48). Columbia University, Earth Institute, Center for Sustainable Development (CSD).

Bontcheva,, K., Papadopoulous, S., Tsalakanidou, F., Gallotti, R., Dudkiewicz, L., Krack, N., Teyssou, D., Francesco, S. N., Spangenberg, J., Srba, I., Aichroth, P., Cuccovillo, L. & Verdoliva, L. (2024). *Generative AI and Disinformation: Recent Advances, Challenges, and Opportunities.* European Digital Media Observatory.

Bontridder, N. & Poullet, Y. (2021). The role of artificial intelligence in disinformation. *Data & Policy, 3.* https://doi.org/10.1017/dap.2021.20

Bovenschulte, M. (2019, Mai). *Deepfakes: Manipulation von Filmsequenzen* (Themenkurzprofil Nr. 25). https://publikationen.bibliothek.kit.edu/1000133910/122803209

Carey, J. M., Guess, A. M., Loewen, P. J., Merkley, E., Nyhan, B., Phillips, J. B. & Reifler, J. (2022). The ephemeral effects of fact-checks on COVID-19 misperceptions in the United States, Great Britain and Canada. *Nature human behaviour*, *6*(2), 236–243. https://doi.org/10.1038/s41562-021-01278-3

Center for Humane Technology. (2024, 29. Februar). *Future-proofing Democracy In the Age of AI with Audrey Tang*. Your Undivided Attention Podcast. Center for Humane Technology. https://www.humanetech.com/podcast/future-proofing-democracy-in-the-age-of-ai-with-audrey-tang

Chan, M.-P. S., Jones, C. R., Hall Jamieson, K. & Albarracín, D. (2017). Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation. *Psychological science*, *28*(11), 1531–1546. https://doi.org/10.1177/0956797617714579

Chesney, B. & Citron, D. (2019). *Deep Fakes: A Looming Challenge for Privacy.* https://doi.org/10.15779/Z38RV0D15J

Cinelli, M., Francisci Morales, G. de, Galeazzi, A., Quattrociocchi, W. & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(9). https://doi.org/10.1073/pnas.2023301118

da Silva, G, Veréb, D. (2023, 12. November). Bildagenturen verkaufen künstlich generierte Bilder zum Nahostkrieg – ohne sie als Fakes auszuweisen. *Neue Zürcher Zeitung, 2023.* https://www.nzz.ch/technologie/bildagenturen-vertreiben-kuenstlich-generierte-bilder-zum-nahost-krieg-ohne-sie-als-fakes-auszuweisen-ld.1764934

Diaz Ruiz, C. & Nilsson, T. (2023). Disinformation and Echo Chambers: How Disinformation Circulates on Social Media Through Identity-Driven Controversies. *Journal of Public Policy & Marketing*, *42*(1), 18–35. https://doi.org/10.1177/07439156221103852

Dobber, T., Metoui, N., Trilling, D., Helberger, N. & Vreese, C. de (2021). Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes? *The International Journal of Press/Politics*, *26*(1), 69–91. https://doi.org/10.1177/1940161220944364

Geiß, S., Magin, M., Jürgens, P. & Stark, B. (2021). Loopholes in the Echo Chambers: How the Echo Chamber Metaphor Over-

simplifies the Effects of Information Gateways on Opinion Expression. *Digital Journalism*, *9*(5), 660–686. https://doi.org/10.1080/21670811.2021.1873811

George, J. F. (2024). Discovering why people believe disinformation about healthcare. *PloS one*, *19*(3), e0300497. https://doi.org/10.1371/journal.pone.0300497

Guan, T., Liu, T. & Yuan, R. (2021). Facing disinformation: Five methods to counter conspiracy theories amid the Covid-19 pandemic. *Comunicar*, *29*(69), 71–83. https://doi.org/10.3916/C69-2021-06

Hameleers, M., Powell, T. E., van der Meer, T. G. & Bos, L. (2020). A Picture Paints a Thousand Lies? The Effects and Mechanisms of Multimodal Disinformation and Rebuttals Disseminated via Social Media. *Political Communication*, *37*(2), 281–301. https://doi.org/10.1080/10584609.2019.1674979

Ho, M.-S. (2022). *Exploring Worldwide Democratic Innovations-A case study of Taiwan* (Exploring Worldwide Democratic Innovations Nr. 78). European Partnership for Democracy.

Hsiao, Y. T., Lin, S.-Y., Tang, A., Narayanan, D. & Sarahe, C. (2018). *vTaiwan: An Empirical Study of Open Consultation Process in Taiwan.* https://doi.org/10.31235/osf.io/xyhft

Huddy, L. (2001). From Social to Political Identity: A Critical Examination of Social Identity Theory. *Political Psychology, 22(1), 127–156.* https://doi.org/10.1111/0162-895X.00230

Hurtz, S. (2024, 5. Februar). Angestellter überweist 24 Millionen Euro an Betrüger. *Süddeutsche Zeitung,* https://www.sueddeutsche.de/wirtschaft/deepfake-betrug-videokonferenz-hongkong-1.6344209

Krempl, S. (2023, 4. Oktober). Emotionserkennung im Callcenter: *Wenn die KI Kundengespräche heimlich auswertet. heise online,* https://www.heise.de/news/Emotionserkennung-im-Callcenter-Wenn-die-KI-Kundengespraeche-heimlich-auswertet-9325037.html

Kreps, S., McCain, R. M. & Brundage, M. (2022). All the News That's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation. *Journal of Experimental Political Science,* 9(1), 104–117. https://doi.org/10.1017/XPS.2020.37

Lee, M. (2020). Free the data from the birdcage: Opening up data and crowdsourcing activism in Taiwan. *PoLAR: Political and Legal Anthropology Review*, *43*(s), 247–261.

Lewandowsky, S., Cook, J. & Lombardi, D. (2020). *Debunking Handbook 2020.* https://doi.org/10.17910/b7.1182

Lewandowsky, S. & van der Linden, S. (2021). Countering Misinformation and Fake News Through Inoculation and Prebunking. *European Review of Social Psychology*, *32*(2), 348–384. https://doi.org/10.1080/10463283.2021.1876983

Lu, C., Hu, B., Li, Q., Bi, C. & Ju, X.-D. (2023). Psychological Inoculation for Credibility Assessment, Sharing Intention, and Discernment of Misinformation: Systematic Review and Meta-Analysis. *Journal of medical Internet research*, *25*, e49255. https://doi.org/10.2196/49255

Lyons, B. (2018). Reducing Group Alignment in Factual Disputes? The Limited Effects of Social Identity Interventions. *Science Communication*, *40*(6), 789–807. https://doi.org/10.1177/1075547018804826

Maertens, R., Roozenbeek, J., Basol, M. & van der Linden, S. (2021). Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of experimental psychology. Applied*, *27*(1), 1–16. https://doi.org/10.1037/xap0000315

Maertens, R., Roozenbeek, J., Simons, J., Lewandowsky, S., Maturo, V., Goldberg, B., Xu, R. & van der Linden, S. (2023). *Psychological Booster Shots Targeting Memory Increase Long-Term Resistance Against Misinformation.* https://doi.org/10.31234/osf.io/6r9as

McPhedran, R., Ratajczak, M., Mawby, M., King, E., Yang, Y. & Gold, N. (2023). Psychological inoculation protects against the social media infodemic. *Scientific reports*, *13*(1), 5780. https://doi.org/10.1038/s41598-023-32962-1

Meineck, S. (2022, 15. September). *Porno-Deepfakes per Knopfdruck*. Netzpolitik.org. https://netzpolitik.org/2022/millionenfach-installierte-apps-porno-deepfakes-per-knopfdruck/

Modirrousta-Galian, A. & Higham, P. A. (2023). Gamified inoculation interventions do not improve discrimination between true and fake news: Reanalyzing existing research with receiver operating characteristic analysis. *Journal of experimental psychology. General*, *152*(9), 2411–2437. https://doi.org/10.1037/xge0001395

Nyhan, B. & Reifler, J. (2019). The roles of information deficits and identity threat in the prevalence of misperceptions. *Journal of Elections, Public Opinion and Parties*, *29*(2), 222–244. https://doi.org/10.1080/17457289.2018.1465061

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D. & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature,* 592(7855), 590–595. https://doi.org/10.1038/s41586-021-03344-2

Pennycook, G., Fugelsang, J. A. & Koehler, D. J. (2015). What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive psychology*, *80*, 34–72. https://doi.org/10.1016/j.cogpsych.2015.05.001.

Pennycook, G. & Rand, D. G. (2021). The Psychology of Fake News. *Trends in cognitive sciences*, *25*(5), 388–402. https://doi.org/10.1016/j.tics.2021.02.007

Pérez-Escolar, M., Lilleker, D. & Tapia-Frade, A. (2023). A Systematic Literature Review of the Phenomenon of Disinformation and Misinformation. *Media and Communication*, *11*(2). https://doi.org/10.17645/mac.v11i2.6453

Roozenbeek, J., van der Linden, S., Goldberg, B., Rathje, S. & Lewandowsky, S. (2022). Psychological inoculation improves resilience against misinformation on social media. *Science advances*, *8*(34), eabo6254. https://doi.org/10.1126/sciadv.abo6254

Ruch, P, Pelzer, S. (2023, 28. Dezember). Scholz greift durch: Die AfD wird verboten – Deepfakes auch! *Chaos Computer Club e.V.* https://media.ccc.de/v/37c3-12345-scholz_greift_durch_die_afd_wird_verboten_-_deepfakes_auch#t=1900

Schiffer, C. (2024, 3. April). OpenAI präsentiert Tool zum Klonen von Stimmen – und warnt davor. *BR24,* https://www.br.de/nachrichten/netzwelt/ki-entwickler-openai-praesentiert-tool-zum-klonen-von-stimmen-und-warnt-davor,U8rlHot

Spampatti, T., Hahnel, U. J. J., Trutnevyte, E. & Brosch, T. (2024). Psychological inoculation strategies to fight climate disinformation across 12 countries. *Nature human behaviour*, *8*(2), 380–398. https://doi.org/10.1038/s41562-023-01736-0

Stiff, H. & Johansson, F. (2022). Detecting computer-generated disinformation. *International Journal of Data Science and Analytics*, *13*(4), 363–383. https://doi.org/10.1007/s41060-021-00299-5

Sundar, S. S. (2008). The MAIN Model: A Heuristic Approach to Understanding Technology Effects on Credibility. In M. J. Metzger & A. J. Flanagin (Hrsg.), *Digital Media, Youth, and Credibility.* The MIT Press.

Tajfel, H. & Turner, J. C. (2004). The Social Identity Theory of Intergroup Behavior. In J. T. Jost & J. Sidanius (Hrsg.), *Political Psychology* (S. 276–293). Psychology Press. https://doi.org/10.4324/9780203505984-16

Terren, L. & Borge-Bravo, R. (2021). Echo Chambers on Social Media: A Systematic Review of the Literature. *Review of Communication Research*, *9*, 99–118. https://doi.org/10.12840/ISSN.2255-4165.028

Tow Center for Digital Journalism. (2024). *"Pink Slime": Partisan Journalism and the Future of Local News.* Columbia University in the City of New York. https://towcenter.columbia.edu/sites/default/files/content/%E2%80%9CPink Slime%E2%80%9D_ Partisan journalism and the future of local news %281%29.pdf

Tseng, Y.-S. (2022). Algorithmic empowerment: A comparative ethnography of two open-source algorithmic platforms – Decide Madrid and vTaiwan. *Big Data & Society*, *9*(2), 205395172211235. https://doi.org/10.1177/20539517221123505

Tseng, Y.-S. (2023). Rethinking gamified democracy as frictional: a comparative examination of the Decide Madrid and vTaiwan platforms. *Social & Cultural Geography*, *24*(8), 1324–1341. https://doi.org/10.1080/14649365.2022.2055779

Vaccari, C. & Chadwick, A. (2020). Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society*, *6*(1), 205630512090340. https://doi.org/10.1177/2056305120903408

Vosoughi, S., Roy, D. & Aral, S. (2018). The spread of true and false news online. *Science (New York, N.Y.)*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Wahls, R. (2024, 11. Februar). Wer wann wo weltweit wählt, 2024. https://www.spiegel.de/ausland/superwahljahr-2024-in-grafiken-rund-die-haelfte-der-weltbevoelkerung-ist-zu-wahlen-aufgerufen-a-dae06614-a1fb-47fc-baee-9c519b7da694

Weikmann, T. & Lecheler, S. (2023). Visual disinformation in a digital age: A literature synthesis and research agenda. *New Media & Society*, *25*(12), 3696–3713. https://doi.org/10.1177/14614448221141648

Westby, L. (2019, 24. Juli). *'The Great Hack': Cambridge Analytica is just the tip of the iceberg.* Amnesty International. https://www.amnesty.org/en/latest/news/2019/07/the-great-hack-facebook-cambridge-analytica/

World Economic Forum. (2024, 10. Januar). *Globale Risiken 2024: Desinformation an der Spitze bei gleichzeitiger Zunahme von*

Umweltbedrohungen. World Economic Forum. https://www3.weforum.org/docs/WEF_GRR24_Press%20release_DE.pdf

Wu, E. Y. (2024, 12. Januar). *Disinformation: Building Digital Resilience.* United States Institute of Peace. https://www.usip.org/publications/2024/01/disinformation-building-digital-resilience

Ybema, S. (2020). Bridging Self and Sociality. In A. D. Brown & S. Ybema (Hrsg.), *The Oxford Handbook of Identities in Organizations* (S. 50–67). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198827115.013.50

ZEIT ONLINE (2022, 25. Juni). Staatsschutz ermittelt nach Fake-Telefonat mit Franziska Giffey. *ZEIT ONLINE,* 2022. https://www.zeit.de/politik/deutschland/2022-06/deepfake-franziska-giffey-staatschutz

Ziemer, C.-T. & Rothmund, T. (2024). Psychological Underpinnings of Misinformation Countermeasures. *Journal of Media Psychology,* Artikel 1864-1105/a000407. Vorab-Onlinepublikation. https://doi.org/10.1027/1864-1105/a000407

DOI: 10.23776/2024_07