



KI IN DER PRÄVENTION

Wissenschaftliche Begleitschrift zum
31. Präventionstag

Gina Rosa Wollinger (Hrsg.)
DPT-Verlag



KI in der Prävention

Expertisen zum 31. Deutschen Präventionstag

Herausgegeben von:
Prof. Dr. Gina Rosa Wollinger

Bibliographische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie: Detaillierte bibliografische Daten sind im Internet unter <http://dnb.d-nb.de> abrufbar.

© DPT Verlag, Hannover
Alle Rechte vorbehalten
Hannover 2026

Redaktion, Satz und Layout: Pascal Specht
Coverdesign: tabasco.media, Hannover

Verlag: Deutscher Präventionstag, Theodor-Heuss-Platz 1-3, 30175 Hannover
Druck: Libri Plureos GmbH, Friedensallee 273, 22763 Hamburg

978-3-911909-02-0 (Printausgabe)
978-3-911909-03-7 (PDF)

gefördert durch:



Inhalt

Michael Fübi

Geleitwort: Warum technischer Fortschritt Sicherheit braucht 5

Gina Rosa Wollinger

Einleitung 13

Alke Martens

Künstliche Intelligenz als Bias-Falle? 29

Sebastian Golla

KI in der Kriminalprävention – Rechtliche Herausforderungen von Innovation bis Anwendung 63

Alina Borowy

Künstliche neuronale Netze in der Polizeiarbeit:
Von Chancen, Risiken und einem fehlenden Rechtsrahmen 83

Simon Egbert

Predictive Policing – Algorithmische Vorhersagen in der polizeilichen Kriminalprävention 107

Christian Büscher, Isabel Kusche, Tim Röller,

Alexandros Gazos

Die Beiträge von KI zu extremistischer Kommunikation und Chancen KI-basierter Prävention 139

Florian Meyer, Melanie Siegel, Dirk Labudde

DeTox, BoTox und DyTox: Projekte zur Unterstützung der Bekämpfung von Hasskriminalität im Netz 173

Stefanie Giljohann, Catharina Vogt

KI-basierte Chatbots in der Prävention häuslicher Gewalt 217

**»Der Einsatz generativer KI bietet
enormes Potenzial – wenn die Risiken
von Anfang an mitgedacht werden«**

Dr. Michael Fübi

Geleitwort

Warum technischer Fortschritt Sicherheit braucht

Es war der teuerste Hackerangriff der britischen Geschichte: Ende August musste der Fahrzeughersteller Jaguar Land Rover seine IT-Systeme herunterfahren und die Produktion für mehrere Wochen stilllegen – mit weitreichenden Auswirkungen auf weitere Beteiligte an Produktion und Lieferkette. Geschätzter Schaden: 1,9 Milliarden Pfund, umgerechnet knapp 2,2 Milliarden Euro.¹

Das Beispiel des britischen Autobauers ist bei weitem nicht das einzige aus den vergangenen Monaten, das die enormen Schäden durch Cyberangriffe deutlich macht. Viele weitere Cyberangriffe haben 2025 Schlagzeilen gemacht. Etwa der Ransomware-Angriff – bei dem die Täter bzw. Täterinnen IT-Systeme verschlüsseln und Lösegeld fordern – auf den IT-Dienstleister Collins Aerospace im September: Der Angriff legte dessen

Passagierabfertigungssysteme, die von zahlreichen Flughäfen genutzt werden, lahm. Eine Reihe von Airports verzeichnete in der Folge Verspätungen und Flugausfälle, darunter der Berliner Flughafen BER und London Heathrow.

Noch drastischer waren die Folgen einer Attacke auf den Serviettenhersteller Fasana aus Euskirchen: Der Mittelständler musste nach einem Ransomware-Angriff Mitte 2025 sogar Insolvenz anmelden.

Das Beispiel Fasana zeigt, dass nicht nur große Konzerne ein Angriffsziel für kriminelle Hacker sind. Vielmehr geraten gerade kleine und mittlere Unternehmen verstärkt ins Visier von Cyberangriffen.² Denn sie können oder wollen häufig nicht so viel in Cybersicherheit investieren wie große Konzerne. Doch sie sind mindestens ebenso sehr auf eine funktionierende IT angewiesen – und können es sich

¹ <https://cybermonitoringcentre.com/2025/10/22/cyber-monitoring-centre-statement-on-the-jaguar-land-rover-cyber-incident-october-2025/>

² <https://commercial.allianz.com/news-and-insights/news/cyber-risk-trends-2025/de.html>

eigentlich nicht leisten, die Cybersicherheit zu vernachlässigen.

Wirtschaft im Fadenkreuz der Hacker

Nicht nur die prominenten Beispiele, die den Weg in die Schlagzeilen der Medien gefunden haben, belegen eine verschärfte Cybersicherheitslage in Deutschland. Diese Entwicklung zeigt sich auch in vielen Umfragen. So veröffentlichte der TÜV-Verband, in dem alle deutschen TÜV-Prüforganisationen organisiert sind, Mitte 2025 die „TÜV Cybersecurity Studie 2025“³, für die Verantwortliche für IT-Sicherheit aus mehr als 500 Unternehmen befragt wurden. 15 Prozent der Unternehmen verzeichneten demnach in den zwölf Monaten vor der Befragung einen IT-Sicherheitsvorfall – also erfolgreiche Cyberangriffe, auf die die Unternehmen aktiv reagieren mussten. Im Vergleich zur Studie zwei Jahre zuvor ist der Anteil erfolgreich gehackter Unternehmen um 4 Prozentpunkte gestiegen.

Der Ausblick verspricht keine Besserung: Die deutsche Wirtschaft befürchtet eine weitere Zunahme von Cyberattacken. 35 Prozent der

Unternehmen rechneten im Oktober 2025 mit einer starken Zunahme von Angriffen in den kommenden zwölf Monaten. Weitere 47 Prozent gehen davon aus, dass sie eher stark zunehmen werden. Einen Rückgang erwartete kein Unternehmen, so der Branchenverband Bitkom.⁴ Der im November 2025 veröffentlichte Jahresbericht des Bundesamts für Sicherheit in der Informationstechnik (BSI) lässt ebenfalls keinen Zweifel an der angespannten Sicherheitslage. Zudem sind laut BSI auch staatlich gesteuerte Akteure, die mit komplexen und langfristigen Attacken politische oder wirtschaftliche Ziele verfolgen, zunehmend aktiv.⁵

Klar ist: Die deutsche Wirtschaft steht im Fadenkreuz staatlicher und krimineller Hacker, die sensible Daten erbeuten, Geld erpressen oder wichtige Versorgungsstrukturen sabotieren wollen.

Die Gefahren sind vielerorts erkannt, und Unternehmen investieren mehr denn je in Cybersicherheit – von moderner Hard- und Software über die Unterstützung durch externe Expertinnen und Experten oder Schulungen der Mitarbeitenden bis hin zu Notfallübun-

3 <https://www.tuev-verband.de/pressemitteilungen/cybersecurity-studie-jedes-siebte-unternehmen-gehackt-risiken-werden-unterschaetzt>

4 <https://www.bitkom.org/Presse/Presseinformation/Deutscher-Markt-IT-Sicherheit-waechst-zweistellig>

5 https://www.bsi.bund.de/DE/Service-Navi/Presse/Pressemitteilungen/Presse2025/251111_BSI-Jahresbericht_2025.html

gen und sogenannten Pentests, die das Ziel haben, technische Schwachstellen im eigenen Unternehmen aufzudecken. Insgesamt werden 2025 die Ausgaben für IT-Sicherheit laut Bitkom voraussichtlich um 10,1 Prozent auf 11,1 Milliarden Euro steigen, nachdem 2024 mit 10,1 Milliarden Euro erstmals die 10-Milliarden-Euro-Marke übersprungen wurde.⁶

KI als neuer Faktor für Cybersicherheit

In die ohnehin schon deutlich verschärfte Gefahrenlage ist nun mit dem großflächigen Einsatz von Künstlicher Intelligenz (KI) ein wesentlicher neuer Faktor hinzugekommen. KI-Applikationen sind Softwarelösungen, die Methoden des maschinellen Lernens und der künstlichen Intelligenz nutzen, um Aufgaben zu automatisieren, Daten zu analysieren oder eigenständig Entscheidungen zu treffen. Sie erkennen Muster in großen Datenmengen, ziehen daraus Schlüsse und passen häufig ihr Verhalten auf Basis neuer Informationen an. Ziel ist es, Prozesse effizienter, schneller und oft auch robuster zu gestalten.

Ein vor allem in der Öffentlichkeit bekannter und bereits vielgenutzter Teilbereich sind die Large Language Models (LLMs) – großskalige Sprachmodelle, die auf neuronalen Netzen basieren und darauf trainiert wurden, menschliche Sprache zu verstehen, zu verarbeiten und zu erzeugen. Sie bilden das Fundament für viele Anwendungen, die heute unter dem Begriff „generative KI“ bekannt sind. Dazu zählen etwa Chatbots, digitale Assistenten, automatische Text- und Codegenerierung oder intelligente Suchsysteme.

Generative KI – in Sicherheitsfragen ein zweischneidiges Schwert

Der Einsatz generativer KI wirkt beim Thema IT-Sicherheit in mehreren Dimensionen. Einerseits hilft KI dabei, Angriffe besser abzuwehren, etwa, indem ungewöhnliche Muster in Datenströmen erkannt und automatisch Gegenmaßnahmen ergriffen werden. Andererseits setzen Kriminelle KI ein, um Angriffe vorzubereiten und durchzuführen – noch nie war es so einfach, Schadcodes oder glaubwürdige Phishing-Mails zu erstellen. Bedenklich dabei: Laut „TÜV Cybersecurity Studie 2025“ nutzen erst 10 Prozent der Unternehmen

⁶ <https://www.bitkom.org/Presse/Presseinformation/Deutscher-Markt-IT-Sicherheit-waechst-zweistellig>

KI für die Abwehr von Cyberangriffen, weitere 10 Prozent planen den Einsatz; vor allem, um Bedrohungen besser zu erkennen.⁷

Schließlich schafft generative KI selbst neue potenzielle Angriffsflächen. Denn je tiefer KI-Systeme in betriebliche Abläufe integriert werden, desto größer ist ihr potenzieller Einfluss – und damit auch das Risiko, das von Sicherheitslücken oder Manipulationen ausgeht. Umso wichtiger ist es, ihre Funktion, Datenbasis und Schnittstellen genau zu verstehen und gezielt abzusichern.

Von Prompt Injections und anderen Manipulationsmöglichkeiten

Ein Beispiel für die Gefahren sind sogenannte Prompt Injections für Large Language Models. Dabei manipulieren Angreifende mit ihrer Eingabe die Daten, die an das Modell gesendet werden, damit es sich für die Nutzer und Nutzerinnen unvorhersehbar verhält oder Informationen preisgibt – und zwar solche, die nicht zugänglich sein sollten. Weitere Risiken sind der falsche Umgang mit den Ergebnissen der generativen

KI – etwa, indem Nutzer und Nutzerinnen nicht validierte Codes⁸ ausführen – und die Manipulation von Trainingsdaten durch die Angreifenden.⁹ Die zehn wichtigsten Sicherheitsrisiken führt das „Open Worldwide Application Security Project (OWASP)“ in der „OWASP Top 10 Framework für Large Language Models (LLMs)“-Liste auf.¹⁰

Cybersicherheit und generative KI zusammendenken

Wie können Unternehmen die Chancen generativer KI für sich nutzen, ohne sich dabei zu vielen zusätzlichen Risiken auszusetzen? Und wie lässt sich der Begriff „Sicherheit“ im Zusammenhang mit dem Einsatz von KI sinnvoll definieren und in Prüfkriterien und -Prozesse umsetzen? Die TÜV-Unternehmen in Deutschland haben sich schon frühzeitig mit diesen Fragen auseinandergesetzt. Denn ihre Aufgabe sehen sie darin, den

7 <https://www.tuev-verband.de/pressemitteilungen/cybersecurity-studie-jedes-siebte-unternehmen-gehackt-risiken-werden-unterschaetzt>

8 Mit „nicht validierten Codes“ sind Programmiercodes gemeint, die von der generativen KI erstellt wurden, ohne dass sie von einem Menschen oder einem automatisierten System auf Korrektheit, Sicherheit und Funktionalität überprüft wurden. Solche Codes können Sicherheitslücken, Fehler oder unerwünschte Funktionen enthalten, die bei der Ausführung zu Problemen führen können, wie z. B. Datenverlust, Systemabstürze oder Angriffsvektoren für Cyberkriminalität. Die Validierung ist daher ein essenzieller Schritt, um sicherzustellen, dass der generierte Code sicher und zuverlässig ist.

9 <https://www.tuv.com/germany/de/penetrationstest-und-it-sicherheitsanalyse.html>

10 <https://genai.owasp.org/llm-top-10/>

technischen Fortschritt so mitzugestalten, dass Technik dem Menschen nutzt und nicht schadet.

In einer gemeinsamen Initiative haben die TÜV-Unternehmen daher 2023 das TÜV AI.Lab gegründet. Ziel: gesellschaftliche und regulatorische Anforderungen, die etwa die EU KI-Verordnung mit sich bringt, in Prüfkriterien und -Prozesse umzusetzen und die Entwicklung von Standards für die Prüfung sicherheitskritischer KI-Anwendungen zu begleiten. Europa soll so weltweit zum Vorreiter für sichere und vertrauenswürdige KI werden.¹¹

Die Aufgaben des TÜV AI.Lab gehen also weit über die engere Frage der Cybersicherheit hinaus. Zum Beispiel werden zunehmend KI-Systeme eingesetzt, die Auswirkungen auf die Gesundheit von Menschen oder ihre elementaren Grundrechte haben können. Das ist etwa der Fall bei Systemen zur Früherkennung von Krebszellen auf MRT-Scans oder selbstfahrenden Fahrzeugen mit automatischer Verkehrsschilderkennung. So können KI-Prüferinnen und -Prüfer zum Beispiel ermitteln, wie sicher automatisierte Fahrzeuge mit KI-Systemen Personen, Verkehrszeichen oder bestimmte Hindernisse erkennen und darauf reagieren.

Klar ist jedoch auch: KI funktioniert nur dann zuverlässig, wenn die Datenverarbeitung selbst hinreichend vor manipulativen Angriffen abgesichert ist – und wenn die Computersysteme, auf denen sie laufen, ebenfalls so gut wie möglich vor Cyberangriffen geschützt sind.

Für Prüfunternehmen wie TÜV Rheinland lassen sich die Fragen der Cybersicherheit und des sicheren Einsatzes von KI daher nicht getrennt voneinander betrachten. Vielmehr muss beides zugleich in den Blick genommen werden und durch Regularien, Standards und Prüfmethode so abgesichert werden, dass wir alle von dem Fortschritt profitieren, den die Technologie schon heute an vielen Stellen bietet.

Regeln sind da, der Überblick ist herausfordernd

Sowohl auf Seiten des Gesetzgebers als auch durch Standards sind in den vergangenen Jahren bereits viele wichtige Schritte getan worden, um die Resilienz der IT von Unternehmen zu stärken, vernetzte Produkte durch strengere Vorgaben für Cybersecurity besser abzusichern und die Einführung von KI in Unternehmen zu regeln. Stichworte hier sind bei-

¹¹ <https://www.tuev-lab.ai/>

spielsweise die NIS-2-Richtlinie, die in der gesamten EU ein hohes gemeinsames Cybersicherheitsniveau sicherstellen soll, der Cyber Resilience Act (CRA), der die Regeln zur Cybersicherheit von Produkten mit digitalen Elementen EU-weit vereinheitlicht, und die bereits erwähnte EU KI-Verordnung. Mit der ISO 42001 gibt es außerdem bereits eine internationale Norm, die einen klaren Rahmen für KI-Managementsysteme bietet und entsprechend geprüft werden kann.

An Regeln und Rahmenbedingungen für bessere Cybersecurity und den sicheren Einsatz von KI mangelt es also nicht. Die Herausforderung ist vielmehr, die Vielzahl an Regelungen noch zu überblicken – vor allem, wenn man dazu noch weitere Regeln in anderen Märkten wie etwa den USA oder China berücksichtigen muss. Zugleich verschärft sich die grundsätzliche Bedrohungslage in Deutschland, und es gibt durch KI zusätzliche mögliche Einfallstore für Cyberangriffe.

Vor diesem Hintergrund führt nichts an der Erkenntnis vorbei: Unternehmen und Institutionen müssen weiter kontinuierlich und vermutlich auch mehr als bislang in Cybersicherheit investieren, um sich und ihre Daten bestmöglich zu schützen und vom technischen

Fortschritt zu profitieren, ohne sich zu hohen Risiken auszusetzen.

Dr. Michael Fübi, Vorsitzender des Vorstands TÜV Rheinland AG, Köln

»Aus fachlicher Perspektive erscheint es zunehmend bedeutsam, die Entwicklung Künstlicher Intelligenz nicht nur analytisch zu begleiten, sondern auch präventiv mitzugestalten.«

Prof. Dr. Gina Rosa Wollinger

Einleitung

Die Entwicklung der Computertechnik, der Kybernetik, wurde von der Mathematikerin Alice Mary Hilton als eine der großen Umwälzungen in der Geschichte der Menschheit beschrieben (vgl. Rid, 2016, S. 133 f.; siehe auch Wollinger & Schulze, 2020). Während die landwirtschaftliche Revolution die körperlichen Fähigkeiten des Menschen durch Werkzeuge auf Maschinen übertrug und damit erweiterte, erkennt Hilton in der Kybernetik eine zweite Revolution – eine, die auf die geistigen Fähigkeiten zielt. Diese zeigt sich zunächst insbesondere in der „Automatisierung der Industrieproduktion“ (Rid, 2016, S. 133). Kybernetik strebt also nicht danach, Muskelkraft zu ersetzen, sondern die Denkleistung des Menschen nachzubilden und zu ergänzen. Dabei geht es weniger um Nachahmung des Menschen, sondern darum, seine Leistung zu übertreffen (Hayward & Maas, 2021, S. 12). Computertechnik ist mittlerweile ein fester Bestandteil des privaten Alltags und der Arbeitswelt und das mit zunehmender Tendenz: Digitalisierung nimmt in Deutschland weiter zu.¹

Bedeutung von Künstlicher Intelligenz

Vor diesem Hintergrund stellt sich die grundlegende Frage, wie die gegenwärtigen Entwicklungen im Bereich der Künstlichen Intelligenz (KI) einzuordnen sind. Handelt es sich hierbei um den Beginn einer dritten großen Revolution oder vielmehr um eine vorübergehende Welle technologischen Enthusiasmus, einen Hype? Die allgegenwärtige Präsenz von KI in der öffentlichen Debatte ist zweifellos auf Anwendungen zurückzuführen, die inzwischen breite Teile der Gesellschaft direkt erreichen und beeinflussen. Ein besonders prägnantes Beispiel hierfür ist das Sprach-

¹ Dies geht u. a. aus dem sogenannten Digitalindex der Initiative D21 hervor. Dieser misst den Zugang zur Digitalisierung, das Nutzungsverhalten, die digitale Kompetenz sowie die Offenheit gegenüber Digitalisierung (Initiative D21 e. V., 2025).

modell ChatGPT, das mit seiner Veröffentlichung 2022 mediale und gesellschaftliche Aufmerksamkeit auf das Thema KI zieht. Dabei gerät leicht in Vergessenheit, dass die konzeptionellen und technischen Grundlagen der Künstlichen Intelligenz keineswegs neu sind, sondern ihre Ursprünge bis in die 1950er Jahre² zurückreichen (Eberl, 2018; Schimpel, 2020).

Gegenwärtig verbinden viele mit KI generative Systeme. Allerdings stellen auch schon frühere Anwendungen, z. B. im Bereich der Mustererkennung, KI-Modelle dar. Eine einheitliche Definition liegt nicht vor (Hayward & Maas, 2021), Legg & Hutter identifizieren sogar 70 verschiedene. Nach der Europäischen Kommission (2019, S. 3) sind die wesentlichen Elemente von KI, dass sie mit einem bestimmten Grad an Eigenständigkeit Analysen durchführt, Entscheidungen trifft und zu einem Ergebnis kommt. Diese Eigenständigkeit ist verknüpft mit *Lernfähigkeit*. Lippitz (2024, S. 12) beschreibt KI als eine: „(...) Technik (...), die darauf ausgerichtet ist, mechanischen Geräten die Fähigkeit zu verleihen, Ziele effizient und erfolgreich in einer Vielzahl von Umgebungen zu erreichen. Diese Intelligenz manifestiert sich in der erhöhten Erfolgswahrscheinlichkeit der Aufgabenerfüllung und wird besonders durch die Geschwindigkeit, mit der diese Ziele erreicht werden, charakterisiert“. KI manifestiert sich also in einer gewissen Autonomie, insbesondere hinsichtlich des Lernens, und unterscheidet sich so von reinen ausführenden Computeranwendungen wie beispielsweise einem Taschenrechner oder der Tabellenkalkulation von Excel.

Auch wenn KI-Systeme schon lange existieren, lässt sich dennoch eine erhebliche Weiterentwicklung ausmachen. Dies hat vor allem damit zu tun, dass Computersysteme heute ein erhebliches Maß an Leistungsstärke aufweisen, eine Vielzahl, insbesondere auch unstrukturierten (Schimple, 2020), Daten vorliegt, sich Algorithmen und Zugänge zu den Systemen verbessert haben (Eberl, 2018, S. 10; Pohlmann, S. 563 ff.). Hinzugekommen ist, dass die Systeme auf der Grundlage von Trainingsdaten zu einem gewissen Grad eigenständig lernen und nicht nur explizit vorgegebene Regeln anwenden (Hayward & Maas, 2021, S. 212).

2 Der Begriff „künstliche Intelligenz“ stammt aus dem Jahr 1956 als der US-Wissenschaftler John McCarthy eine Tagung so benannte (Eberl, 2018, S. 9).

KI arbeitet meist auf der Grundlage neuronaler Netze. Hierbei geht es nicht mehr um die Umsetzung von einfachen Wenn-Dann-Befehlen, diese Systeme werden häufig gar nicht mehr als KI angesehen (Hayward & Maas, 2021, S. 212). Solche basalen Instrumente werden auch schwache oder symbolische KI genannt (Hayward & Maas, 2021; Pohlmann, 2022). Schwache KI löst konkrete Anwendungsprobleme und basiert meist auf *Machine Learning* (ML). ML kann Muster und Gesetzmäßigkeiten in Daten erkennen. Dies liegt beispielsweise vor, wenn ein KI-Tool darauf trainiert wurde, Hunde auf Bildern zu erkennen, auch auf solchen, mit denen es nicht trainiert wurde.

Deep Learning (DL) hingegen geht einen Schritt weiter. Es nutzt neuronale Netze und ist damit in der Lage, nicht nur Muster zu erkennen, sondern auch Eigenes zu generieren, in dem genannten Beispiel also selbst ein Bild von einem Hund zu erstellen. Neuronale Netzwerke bestehen aus drei Systemen. Eines, welches zunächst Informationen verarbeitet, diese dann an ein weiteres Neuronen-Geflecht weiterleitet (sogenannte *hidden neurons*), welche sodann an die Output-Neuronen weiterleitet, welches ein Ergebnis rausgibt. Dies stellt eine Form von starker KI dar, die sich selbst verbessert, also eigenständig lernt. Dieses Lernen wird *unüberwachtes Lernen* genannt, was unter anderem damit zu tun hat, dass es nicht mehr nachvollziehbar ist, wie das System zum Ergebnis gekommen ist (Stichwort *hidden neurons*). Deswegen wird in diesem Zusammenhang häufig das Bild einer *black box* bemüht (Lippitz, 2024, S. 16; Schimpel, 2020, S. 100). So kann nicht beantwortet werden, warum einer Nutzerin eines Social-Media-Kanals genau die Videos vorgeschlagen werden, die sie erhält oder warum ChatGPT einen spezifischen Text schreibt. Die Resultate sind nicht in dem Sinn richtig oder falsch wie das Ergebnis eines Taschenrechners, sondern stellen gewisse Wahrscheinlichkeiten eines korrekten Schlusses dar. Insbesondere bei der Analyse und Interpretation sprachlicher, textlicher und visueller Daten erweisen sich KI-Systeme mittlerweile als äußerst leistungsfähig (Eberl, 2018).

Diese starke, generative und moderne Form von KI hat weitreichende Auswirkungen. Sie ist in öffentlichen Debatten meist gemeint, wenn von KI die Rede ist. Fraglich ist nun, in welchem Zusammenhang KI mit Kriminalität und Kriminalprävention steht. Nach Hayward & Maas, (2021) lassen sich drei Kategorien zur Unterscheidung aufmachen: *Crimes with AI*, *Crimes on AI* und *Crimes by AI*.

Die Schattenseiten von KI

Die Entwicklung von KI wird häufig assoziiert mit schicken Büros, in denen smarte Informatiker:innen arbeiten – der Glamour des Silicon Valleys. Doch für KI-Software werden auch andere Arbeitsbereiche benötigt. Ein zentrales Problem beim Einsatz von KI ist die Art und Weise, wie Trainingsdaten gewonnen und aufbereitet werden – insbesondere durch Datenannotation (siehe hierzu ausführlich Muldoon, Graham & Cant, 2025). Datenannotation bezeichnet den Prozess, bei dem rohe, ungeordnete Daten gezielt mit zusätzlichen Informationen (Labels oder Markierungen) versehen werden, damit sie von KI- und Machine-Learning-Systemen verstanden und verarbeitet werden können. Dabei können zum Beispiel Bilder mit Beschreibungen wie „Auto“ oder „Person“ gekennzeichnet, Texte nach ihrer Bedeutung oder Stimmung kategorisiert oder Audiodateien verschriftlicht werden. Diese Annotationen dienen als Lerngrundlage, damit ein KI-Modell Muster erkennt und später eigenständig Vorhersagen oder Entscheidungen treffen kann.

Ohne Datenannotation wären viele heutige KI-Anwendungen nicht möglich, da maschinelles Lernen in der Regel auf großen Mengen korrekt beschrifteter Daten basiert. Die Qualität der Annotationen hat dabei direkten Einfluss auf die Leistungsfähigkeit und Fairness der KI: Fehlerhafte oder einseitige Markierungen können zu Verzerrungen (Bias) und falschen Ergebnissen führen. Deshalb ist Datenannotation zwar oft unsichtbar aber ein zentraler und arbeitsintensiver Bestandteil der KI-Entwicklung.

Die Arbeit der Datenannotation ist für viele Menschen extrem belastend, auch wenn sie nach außen oft als einfache Klickarbeit erscheint (ebd.). Annotierende müssen stundenlang hochkonzentriert monotone Aufgaben erledigen, etwa Bilder sortieren, Texte bewerten oder Inhalte markieren – häufig unter starkem Zeitdruck und für sehr niedrige Bezahlung. Besonders stark ist die psychische Belastung, wenn es um die Moderation und Kennzeichnung problematischer Inhalte geht: Gewalt, Darstellungen sexualisierter Gewalt u. a. an Kindern, Hassrede oder extremistische Propaganda. Diese Inhalte müssen immer wieder angesehen und bewertet werden, was zu Stress, Angstzuständen, Schlafstörungen oder sogar Traumatisierungen führen kann, meist ohne ausreichende psychologische Betreuung.

Hinzu kommt, dass diese Arbeit oft in Länder mit schwachen arbeitsrechtlichen Schutzmechanismen ausgelagert wird. Die Beschäftigten haben kaum Mitspracherecht, keine langfristige Jobsicherheit und sind für Fehler oder Abweichungen vom vorgegebenen Schema schnell ersetzbar. Gleichzeitig bleibt ihre Arbeit unsichtbar: Während KI-Systeme als „intelligent“ und automatisiert wahrgenommen werden, werden die Menschen dahinter kaum anerkannt. Diese Kombination aus emotionaler Belastung, ökonomischer Ausbeutung und fehlender Wertschätzung macht Datenannotation zu einer der unterschätztesten, aber härtesten Tätigkeiten im KI-Ökosystem.

Ein weiterer kritischer Aspekt ist der äußerst hohe Energieaufwand moderner KI-Systeme. Das Training großer Modelle sowie ihr dauerhafter Betrieb erfordern leistungsstarke Rechenzentren, die enorme Mengen an Strom verbrauchen und einen hohen CO²-Ausstoß verursachen. Um die erforderlichen Leistungen bereitzustellen, sind gigantische Rechenzentren notwendig, sogenannte Hyperscaler (Muldoon, Graham & Cant, 2025, S. 103 ff.). Große Rechenzentren verbrauchen zwischen 11 und 19 Millionen Liter Wasser am Tag, u. a. für die Kühlung (ebd., S. 109). Allein das Unternehmen *Google* verbrauchte im Jahr 2021 pro Tag 1,7 Millionen Liter Wasser.³ Insbesondere in Regionen mit Wasserknappheit stellt das ein großes Problem dar und führte auch schon zu Protesten gegen die geplante Ansiedlung von Rechenzentren (Rohde, 2025).

Ebenso ist der Verbrauch an Strom enorm – und wird noch steigen, wie eine Prognose der Internationalen Energieagentur (IEA)⁴ zeigt: Der Stromverbrauch von Rechenzentren wird weltweit bis zum Jahr 2030 voraussichtlich auf mehr als das Doppelte ansteigen. Konkret wird erwartet, dass Rechenzentren dann etwa 945 Terawattstunden (TWh) Strom pro Jahr verbrauchen. Das ist extrem viel und entspricht in etwa dem jährlichen Stromverbrauch eines Industrielandes, wie beispielsweise Japan.

Eine weitere problematische Dimension von KI ist die ungleiche Verteilung von Macht und Kontrolle über KI-Systeme. Dies zeigt sich nicht nur in verzerrten Datensätzen⁵ die genutzt werden, sondern auch in der Verteilung

3 <https://www.computerweekly.com/de/tipp/Wie-man-den-Wasserverbrauch-nachhaltig-verwaltet>

4 <https://www.iea.org/reports/energy-and-ai/executive-summary>

5 Vorurteile und einseitige gesellschaftliche Perspektiven werden von den Daten des Internets gespiegelt. Ein Grund hierfür ist u. a. der Umstand, dass gerade große Plattformen wie Wikipedia und Youtube vor allem von jungen, weißen und männlichen US-Amerikanern bespielt werden (Muldoon, Graham & Cant, 2025, S. 80).

der Hyperscaler-Rechenzentren: Über die Hälfte befinden sich in den USA, 16 % in Europa und 15 % in China (Muldoon, Graham & Cant, 2025, S. 104). Besonders problematisch ist, dass dieser Energiebedarf häufig in Regionen gedeckt wird, die stark von fossilen Energieträgern abhängig sind. Dadurch steht der technologische Fortschritt im Widerspruch zu globalen Klimazielen und wirft die Frage auf, ob der Nutzen vieler KI-Anwendungen den ökologischen Preis rechtfertigt.

Insgesamt geht somit der Ausbau von KI-Anwendungen mit einer hohen sozialen und ökologischen Belastung einher. Dabei sind es die Länder des globalen Nordens die ökonomisch profitieren und die des globalen Südens, die die negativen Auswirkungen tragen. Bezieht man zudem den Umstand ein, dass die gigantischen Unterseekabel, welche das heutige Ausmaß globaler Vernetzung überhaupt erst ermöglichen, entlang der historischen Seerouten der Kolonialzeit verlaufen (Muldoon, Graham & Cant, 2025), wird die enge Konnexität von Imperialismus und Kolonialismus mit den gegenwärtigen soziotechnischen Formationen künstlicher Intelligenz deutlich.

Kriminalität und KI

Mittels KI (im breiten Verständnis) Straftaten zu begehen, ist innerhalb des Bereichs *Cybercrime* schon seit Jahrzehnten ein fester Bestandteil von Kriminalität, mit sehr unterschiedlichen Erscheinungsformen (Wollinger et al., 2020). Die technologischen Entwicklungen, insbesondere hin zu generativer KI, führen jedoch zu einer neuen Dynamik, die eine neue Relevanz des Deliktsbereichs begründet (Baier, 2024, S. 5). So können mittels KI Schadprogramme für Malware-Attacken geschrieben oder andere Anleitungen und Skripte zur Begehung von Straftaten erstellt werden, bis hin zu Anleitungen zur Erstellung von Sprengstoff. Weitreichende Konsequenzen sind ferner im Zusammenhang von Betrugsdelikten zu erwarten. Im Betrugskontext, in dem es um das Vortäuschen spezifischer Umstände oder Personen geht, kann KI passgenaue Inhalte liefern (Lippitz, 2024; Wröner, 2024). Beispielsweise können Phishing-Mails verbessert aber auch Betrugsdelikte am Telefon mittels *Voice-Cloning* optimiert werden. Voice-Cloning kann mithilfe kurzer Stimmfrequenzen einer Person, welche aus anderen Online-Inhalten entnommen werden kann, genutzt werden, um mit der Stimme eines spezifischen Menschen eingegebene Sätze zu sagen. Der sogenannte Enkeltrickbetrug erlangt neue Dimensionen, wenn die Stimme tatsächlich die des Enkelkinds ist.

Automatisierter erfolgen KI-Straftaten, die Bots nutzen und unter anderem für das sogenannte *CyberLove*-Phänomen eingesetzt werden. Hierbei handelt es sich um Bots, die auf Online-Datingplattformen oder in anderen Chats flirtähnliche Kontaktversuche ausführen, um z. B. vertrauliche Informationen zu erhalten. Ein weiterer Anwendungsbereich von KI zur Begehung von Straftaten besteht im Zusammenhang mit der Erstellung von *Deepfakes*. So können äußerst realistisch wirkende Bilder und Videos erstellt werden, die beispielsweise Personen mit pornographischen Inhalten verbinden. Dies wiederum kann genutzt werden, um Menschen zu diskreditieren oder als Erpressungsgrundlage dienen.

Ferner kann KI Rekrutierungsprozesse für extremistische Gruppierungen verbessern, indem Jugendliche und andere Personen passgenauer in den sozialen Medien angesprochen werden (Schimpel, 2020). Des Weiteren können auch Protestaufrufe und -aktionen aus radikalen Milieus erfolgreicher werden. KI kann hier nicht nur bessere Slogans und Bilder erstellen, sondern auch darauf achten, dass die Strafbarkeitsgrenze nicht erreicht wird. Dass KI vor allem bei Straftaten im Internet weitreichende Folgen haben kann, liegt u. a. an den Besonderheiten des digitalen Raums, in dem der Einfluss von Sozialkontrolle geringer und Anonymität höher ist (Haverkamp, 2023).

KI kann jedoch nicht nur zur Begehung von Straftaten genutzt werden, sondern auch ein Einfallstor für Kriminalität darstellen (Baier, 2024). So können (legale) Bots manipuliert werden, Smart Home-Anlagen oder Systeme autonomen Fahrens gehackt werden oder Spracherkennungssoftware genutzt werden, um Informationen von den Nutzer:innen abzugreifen.

Das Funktionieren von KI ist abhängig von seinen Trainingsdaten. Insofern können die Entscheidungen von KI manipuliert werden, indem nur bestimmte Daten zum Training genutzt werden. Dies kann insbesondere in Bereichen der kritischen Infrastruktur zu fatalen Folgen führen wie beispielsweise bei Diagnose-Tools für Krankenhäuser.

KI könnte wiederum eine gewisse Eigenständigkeit bei der Begehung von Straftaten entwickeln. Dies ist insbesondere im Zusammenhang mit autonom agierenden Bots denkbar, die z. B. Hasskommentare verbreiten können (Hayward & Maas, 2021). Ferner könnte eine KI so programmiert sein, dass sie autonom nach Schwachstellen in anderen Systemen sucht und DDoS- und andere Angriffe ausführt.

Strafverfolgung und KI

KI ist mittlerweile auch ein „tool of policing“ (Hayward & Maas, 2021, S. 219) in vielfältigen Bereichen der Polizeiarbeit. So können entsprechende Programme bei der Auswertung großer Datenmengen hilfreich sein, z. B. im Zusammenhang mit Darstellungen sexualisierter Gewalt an Kindern (Garbers & Brodthage, 2021; Wörner, 2024). Das LKA Niedersachsen hat beispielsweise eine Software zum internen Gebrauch entwickelt, die große Mengen an Bildmaterial vorsortiert (Wörner, 2024, S. 634). Ferner kann KI zur Gesichts- und Spracherkennung sowie zur Auswertung von Beweisen eingesetzt werden. Neben der Auswertung kann KI auch zum Zusammenführen großer Datenmengen genutzt werden, wie es das Programm *Gotham Palantir*⁶ durchführt (Wörner, 2024). Gerade aufgrund unterschiedlicher polizeilicher Daten und anderen staatlich registrierten Informationen scheint dies für die Ermittlungstätigkeit attraktiv. Die gegenwärtige Bundesregierung hat die Entwicklung und den Einsatz entsprechender KI-gestützter Analysen als Vorhaben in ihrem Koalitionsvertrag aufgenommen.

In anderen Ländern ist der polizeiliche Einsatz von KI zum Teil weiter fortgeschritten als in Deutschland (Garbers & Brodthage, 2021). Beispielsweise wird in den Niederlanden eine Gesichtserkennungssoftware in U-Bahnen genutzt, um Personen zu identifizieren, die ein Hausverbot haben (ebd., S. 135 f.). Die Züricher Kantonspolizei nutzt für Ermittlungen im Bereich der Wirtschaftskriminalität Analysetools, um große Mengen an Textdokumenten auszuwerten. Eine Schweizer Studie belegt, dass KI in unterschiedlichen Bereichen des Strafverfolgungsprozesses verwendet wird, jedoch überwiegend auf einem niedrigen Komplexitätsgrad (Simmler et al., 2022).

6 *Gotham Palantir* wird derzeit in Bayern, Hessen und NRW genutzt. Das Programm ist höchst umstritten, u. a. deshalb, da es sich dahinter um eine US-amerikanische Firma handelt, die von Peter Thiel mitbegründet wurde. Peter Thiel, der immer noch Großaktionär des Unternehmens ist, finanziert seit Jahren D. Trump und J. D. Vance. Ferner steht die Firma in engem Austausch mit US-Geheimdiensten. NRW musste sein Polizeigesetz ändern, um die Software zu nutzen. Gegen den Einsatz wurde geklagt (BVerfG, Urteil vom 16.2.2023 – 1 BvR 1547/19 – und – 1 BvR 2634/20), woraufhin weitere Anpassungen vorgenommen werden mussten.

Künstliche Intelligenz in der Kriminalprävention

KI kann also nicht nur zur Begehung von Straftaten eingesetzt werden, sondern auch zur Bekämpfung (Baier, 2024). So können beispielsweise mittels KI Angriffe auf Computersysteme besser erkannt werden (Pohlmann, 2022, S. 561 ff.). Ferner könnten auffällige Nutzeraktivitäten identifiziert werden. Insgesamt kann KI auf vielfältige Weise die Cybersicherheit verbessern. Dies gilt nicht nur in Bezug auf Cybercrime im engeren Sinn. KI kann z. B. Hasskommentare in den sozialen Medien herausfiltern (Halvani, 2023). Inzwischen nutzen alle 14 Landesmedienanstalten die Software *KIVI*, um strafrechtlich relevante Inhalte zu erkennen. Diese können dann bei der Zentralen Meldestelle für Internetkriminalität des BKA angezeigt werden (Klemp, 2024). Des Weiteren wurden für die Prävention von Hate Crime zwei KI-basierte Software-Programme von der Universität Darmstadt entwickelt, welche u. a. dem Projekt *Hessen gegen Hetze* zur Verfügung gestellt wurden (Klemp, 2024). Beide Programme, *DeTox* und *BoTox* genannt, untersuchen das Internet nicht eigenständig, stellen aber eine Unterstützung bei der Auswertung von großen Datenmengen dar. *BoTox* kann dabei erkennen, ob ein Post von einem Bot verfasst wurde.

Im Rahmen polizeilicher Präventionsarbeit hat KI v. a. im Kontext von Tatprognosen einen hohen Stellenwert erhalten (Dakalbab, 2022; Fahrthofer, 2023; Schimpel, 2020). Unter dem Begriff *Predictive Policing* werden hier Analysesysteme verstanden, die auf Grundlage unterschiedlicher Daten Risikogebiete identifizieren, in denen es wahrscheinlich zu Straftaten kommen wird. In Deutschland wird dies v. a. für den Deliktsbereich des Einbruchs genutzt.

Des Weiteren wird der Einsatz von KI im Rahmen von Videoüberwachung diskutiert. Neben Gesichtserkennung könnte hierbei auch auffälliges Verhalten identifiziert werden. Ein weitreichender Einsatz solcher Instrumente findet in China statt (Schimpel, 2020). Neben dem Ziel der Verkehrssicherheit wird hier auch das Verhalten der Bürger:innen in anderen Bereichen erfasst und zu einer Gesamtbewertung der Person zusammengefügt, dem sogenannten, sehr umstrittenen, *social scoring*.

Neben den Möglichkeiten, effizienter polizeiliche Ermittlungen zu führen und Sicherheit zu erhöhen, wird der Einsatz von KI auch im Rahmen von Strafzumessung diskutiert (Kaspar et al., 2020). Vor dem Hintergrund,

dass bisherige Befunde zeigen, dass die Höhe der Strafe sehr unterschiedlich ausfällt, wird in KI auch eine Chance dahingehend gesehen, hier eine gerechtere Grundlage zu schaffen. Schließlich gilt es zu identifizieren, in welchen Bereichen der Prävention, auch außerhalb der polizeilichen Arbeit, KI eingesetzt werden kann, um Straftaten zu verhindern.

Risiken

Daneben wird der Einsatz von KI jedoch auch sehr kritisch diskutiert. Die Risiken liegen zum einen in der Möglichkeit der Manipulation der Systeme und der Intransparenz (Stichwort *black box*) (Schimpel, 2020). So liegen Hinweise vor, dass KI-Systeme zum Teil befangen sind und für bestimmte marginalisierte Gruppen nachteilig sein könnten (z. B. im Zusammenhang mit Gesichtserkennung von Schwarzen Personen) (Deutscher Ethikrat, 2023). Das Bundesverfassungsgericht hat in einem Urteil den Umgang mit Daten zur Bekämpfung schwerster Kriminalität begrenzt (Wörner, 2024, S. 267), da es das Grundrecht auf informationelle Selbstbestimmung gefährdet sieht. Es verweist darauf, dass die Verhältnismäßigkeit beim Einsatz von KI gewahrt bleiben muss.

Überlegungen zum Thema Risiken von Künstlicher Intelligenz in der Prävention befassen sich mit der Frage, wie weit automatisierte KI-Systeme zur Vorhersage von Straftaten eingesetzt werden dürfen, ohne Grundrechte wie Datenschutz und Unschuldsvermutung zu verletzen. Zudem ist zu klären, ob und wie diskriminierende Tendenzen in Trainingsdaten zu unfairen Bewertungen oder polizeilichen Maßnahmen führen können.

Aus fachlicher Perspektive erscheint es zunehmend bedeutsam, die Entwicklung Künstlicher Intelligenz nicht nur analytisch zu begleiten, sondern auch präventiv mitzugestalten. Der Deutsche Ethikrat betont in seiner Stellungnahme 2023, dass KI den Menschen nicht ersetzen dürfe. KI sei nicht vernunftbegabt und trage für seine Entscheidungen, anders als der Mensch, keine Verantwortung. Angesichts der Dynamik aktueller technologischer Entwicklungen und der potenziell tiefgreifenden gesellschaftlichen Auswirkungen besteht ein dringender Handlungsbedarf. Es sollte daher stärker betont werden, dass KI nicht autonom entsteht, sondern von Menschen entwickelt wird und ethische und regulatorische Leitplanken braucht.

Der 31. Deutsche Präventionstag lädt dazu ein, das Thema Künstliche Intelligenz in der Prävention umfassend und zukunftsorientiert zu beleuchten. Im Fokus stehen dabei zentrale Fragen: Welche Herausforderungen bringt KI im Kontext von Kriminalität und Sicherheit aber auch im gesamtgesellschaftlichen Miteinander mit sich? Welche tiefgreifenden Veränderungen gehen mit ihrem Einsatz einher – und wer ist davon in welcher Weise betroffen? Gleichzeitig richtet sich der Blick nach vorn: Wie lässt sich KI gezielt und verantwortungsvoll für die Präventionsarbeit nutzen? Dabei geht es nicht nur um technologische Potenziale, sondern auch um die ethische und praktische Frage, wie ein bewusster, reflektierter Umgang mit KI in der Prävention gelingen kann.

Zum vorliegenden Sammelband

Im vorliegenden Sammelband zum Schwerpunktthema des 31. Deutschen Präventionstags werden unterschiedliche Perspektiven auf das Thema zusammengetragen.

Alke Martens geht der Frage nach, inwiefern KI-Tools Vorurteile und Stereotype reproduzieren. Eine Gefahr hierfür besteht in Form des sogenannten Datenbias, also dem Umstand, dass es sich bei den Daten, die die Grundlage für KI-Anwendungen darstellen, schon um verzerrte Darstellungen der Wirklichkeit handelt. Ihr Fazit: KI ersetzt nicht die menschliche Arbeit, sie kann allenfalls unterstützend wirken.

Mit welchen rechtlichen Herausforderungen die Implementierung von KI, insbesondere in der Kriminalprävention, verbunden ist, zeigt **Sebastian Golla** auf. In seinem Beitrag erläutert er bestehende Regelungen wie die KI-Verordnung der Europäischen Union. Entscheidend wird in Zukunft sein, ob die rechtliche Wirklichkeit mit der sich rasant verändernden faktischen Wirklichkeit im Bereich KI Schritt halten kann.

Inwiefern KI beim Strafverfahren zum Einsatz kommen kann, weist **Alina Borowy** auf. Dabei geht sie insbesondere auf Gesichtserkennung und Videoüberwachung ein und thematisiert die Gefahren von falschen Treffern. Des Weiteren beschreibt sie die Schwere von Grundrechtseingriffen.

Simon Egbert schreibt in seinem Beitrag über Predictive Policing und Kriminalprävention. Dabei erläutert er die Besonderheiten algorithmischer

Prognose in Abgrenzung zu klassischen Formen kriminalpräventiver Polizeiarbeit. Insbesondere interessiert ihn die soziotechnische Interaktion.

Welche Bedeutung KI im Kontext von extremistischer Kommunikation zukommt, wird im Beitrag von **Christian Büscher, Isabel Kusche, Tim Röller** und **Alexandros Gazos** ausgeführt. Sie erläutern, welchen Einfluss KI auf Kommunikationsverläufe im Netz hat und wie extremistische Tendenzen durch Algorithmen selbst verstärkt werden können.

Florian Meyer, Melanie Siegel und **Dirk Labudde** stellen zwei Präventionsprojekte vor: DeTox und BoTox. Dabei handelt es sich um Möglichkeiten, Hasskommentare automatisiert zu erkennen. Sie loten dabei aus, inwiefern die Tools für die Strafverfolgung und Kriminalprävention eingesetzt werden können.

Der Sammelband schließt mit einem Beitrag von **Catharina Vogt** und **Stefanie Giljohann** zu einem weiteren KI-gestützten Präventionsprojekt. Sie beschreiben einen Chatbot zur Unterstützung von Betroffenen häuslicher Gewalt. Der Chatbot soll dabei eine Lücke schließen, in dem es eine niedrighschwellige Möglichkeit ist, sich zu informieren und nächste Schritte zu gehen.

Literatur

- Baier, D. (2024). Künstliche Intelligenz und Kriminalität. *SKIP Info*, (1), 5-10.
- Dakalbab, F., Talib, M- A., Waraga, O. A., Nassif, A. B., Abbas, S. & Nasir, Q. (2022). Artificial intelligence & crime prediction: A systematic literature review. *Social Sciences & Humanities Open*, (6), 100342.
- Deutscher Ethikrat (Hrsg.). (2023). Mensch und Maschine. Herausforderungen durch Künstliche Intelligenz. Stellungnahme. Berlin.
- Eberl, U. (2018). Was ist künstliche Intelligenz – Was kann sie leisten? *Aus Politik und Zeitgeschichte*, 68(6-8), 8-14.
- Europäische Kommission. (2019). A Definition of AI: main capabilities and disciplines. ai_hleg_ai_definition_final_DF06F793-EA01-3573-16D2ACD625E2BDB0_56341.pdf
- Farthofer, H. (2023). Der Einsatz von Künstlicher Intelligenz in der Kriminalprävention. In T.-G. Rüdiger, P. S. Bayerl (Hrsg.), *Handbuch Cyberkriminalologie 1* (S. 293–316). Springer.
- Garbers, N. & Brodthage, M. (2021). Einsatz künstlicher Intelligenz im Polizeialltag. In T.-G. Rüdiger (Hrsg.). *Zukunft Digitaler Polizeiarbeit* (S. 131-153). Verlag für Polizeiwissenschaft.
- Halvani, O. (2023). Möglichkeiten zur Erkennung von Hate Speech. *Datenschutz und Datensicherheit*, 47, 209–214.
- Haverkamp, R. (2023). Was ist das richtige Leben? Leben in verschiedenen Welten. *Neue Kriminalpolitik*, 35(3), 269-283.
- Hayward, K. J. & Maas, M. M. (2021). Artificial intelligence and crime: A primer for criminologists. *Crime Media Culture*, 17(2), 209-233.
- Initiative D21 e. V. (Hrsg.). (2025). *D21-Digital-Index 2023/24. Jährliches Lagebild zur Digitalen Gesellschaft*. https://initiated21.de/uploads/03_Studien-Publikationen/D21-Digital-Index/2023-24/d21digitalindex_2023-2024.pdf
- Kaspar, J., Höffler, K. & Harrendorf, S. (2020). Datenbanken, Online-Votings und künstliche Intelligenz – Perspektiven evidenzbasierter Strafzumessung im Zeitalter von „Legal Tech“. *Neue Kriminalpolitik*, 32(1), 35-56.
- Klemp, C. (2024). Wie KI den Hass im Netz bekämpft. *Forum Opferhilfe*, (3), 19-21.
- Legg, S. & Hutter, M. (2007). A Collection of Definitions of Intelligence. *Frontiers in Artificial Intelligence and Applications*, 157, 17-24.
- Lippitz, R. D. U. (2024) *Kriminalität und Künstliche Intelligenz*. Springer VS.
- Muldoon, J., Graham, M. & Cant, C. (2025). Feeding the Machine. Hinter den Kulissen der KI-Imperien. Harper Collins.

- Pohlmann, N. (2022). *Cyber-Sicherheit*. 2. Auflage. Springer Vieweg.
- Rid, T. (2016). *Maschinendämmerung. Eine kurze Geschichte der Kybernetik*. Propyläen.
- Rohde, F. (2025). Digitalisierung: Künstliche Intelligenz und Wasserverschwendung. Heinrich Böll Stiftung. <https://www.boell.de/de/2025/01/08/digitalisierung-kuenstliche-intelligenz-und-wasserverschwendung>
- Schimpel, U. (2020) Künstliche Intelligenz & Präventionsarbeit. In C. Schwarzenegger & R. Nägeli (Hrsg.). *Elftes Zürcher Präventionsforum. Neue Technologien im Dienste der Prävention: Möglichkeiten – Risiken* (S. 97-111). EIZ Publishing.
- Simmler, M., Brunner, S., Canova, G. & Schedler, K. (2023). Smart criminal justice: Exploring the use of algorithms in the Swiss criminal justice system. *Artificial Intelligence and Law*, 31, 213-237.
- Wollinger, G. R. & Schulze, A. (Hrsg.) (2020). *Cybersecurity für die öffentliche Verwaltung* (S.) Kommunal- und Schulverlag.
- Wollinger, G. R., Dreißigacker, A. & von Skarczinski, B. (2020). Formen der Bedrohung von Cyberkriminalität. In G. R. Wollinger & A. Schulze (Hrsg.), *Cybersecurity für die öffentliche Verwaltung* (S. 27-56). Kommunal- und Schulverlag.
- Wörner, L. (2024). Weg von den Hürden, hin zu den Möglichkeiten: KI in Polizei und Strafverfolgung. *Zeitschrift für die gesamte Strafrechtswissenschaft*, 136(3), 616-641.



Prof. Dr. Gina Rosa Wollinger war von 2012 bis 2018 wissenschaftliche Mitarbeiterin am Kriminologischen Forschungsinstitut Niedersachsen. Dort forschte sie hauptsächlich zum Phänomen Wohnungseinbruch und Cybercrime. Seit 2018 ist sie Professorin für Soziologie und Kriminologie an der Hochschule für Polizei und öffentliche Verwaltung NRW.

»Die Art der Datenerhebung, die Gestaltung des Algorithmus, die Programmierung und auch der Einsatz der Software spiegeln immer den Menschen wider – im Guten wie im Schlechten, inklusive aller Biases.«

Prof. Dr. Ing. Alke Martens

Künstliche Intelligenz als Bias-Falle?

1. Einleitung: Der Bias

Ein Foto zeigte eine junge blonde Frau mit einem Pferdeschwanz, einen weißen Kittel und einem Stethoskop um den Hals. Ein weiteres Bild zeigt einen älteren Mann mit grauem Haar, glattrasiert, mit einem weißen Kittel und einem Stethoskop um den Hals. Welche der beiden Personen zeigt einen Arzt? Welches zeigt eine Person aus der Pflege?

Die Antwort, die vielen Menschen im westlichen Kulturkreis vermutlich ebenfalls geben würden, entspricht der Antwort, die über 100 befragte Studierende der Medizin gegeben haben¹. Der Arzt, zudem die männliche Bezeichnung gewählt wurde (siehe Elsen, 2023, Piepenbrink, 2022), ist der Mann. Auch bei der englischen Formulierungsvariante, bei der „physician“ als neutrale Bezeichnung gilt, trat der gleiche Effekt auf. Die Person aus der Pflege ist die Krankenschwester, die junge blonde Frau. Tatsächlich war es bei diesen Bildern andersherum: die junge blonde Frau war die Ärztin, der ältere Mann arbeitete schon viele Jahre als Pfleger in der gleichen Klinik. Das ist ein klassisches Beispiel für einen Bias. Bias, ein aus dem englischen Sprachraum stammendes Wort, wird gemeinhin verstanden als Vorurteil oder verzerrte Wahrnehmung. Es kann auch Voreingenommenheit bedeuten oder auch Einseitigkeit einer Wahrnehmung (Wikipedia, 2025b).

Dass Menschen stereotyp wahrnehmen und Vorurteile haben, ist bekannt. Doch wie ist das mit den scheinbar rationalen Computersystemen?

¹ Dies war ein Nebeneffekt der Entwicklung des Lernsystems Docs'n Drugs an der Universität Ulm (siehe z. B. Martens et al., 2001). In der Untersuchung ging es um die Auswahl eines passenden Erscheinungsbildes eines Avatars für eine digitale Lernumgebung. Das Ergebnis der Umfrage wurde nicht publiziert und hatte weder die Befragten noch die Befragenden überrascht.

Man könnte annehmen, Künstliche Intelligenz sei objektiv, zumindest jedoch objektiver, doch so ist es nicht. Bei dem oben genannten Beispiel passiert folgendes:

Aktuelle Ausgaben einer Künstlichen Intelligenz zur Bildgenerierung entwerfen die gleichen stereotypen Bilder von Menschen in medizinischen Berufen. Höchste sprachliche Präzision ist bei der Eingabe eines Prompts erforderlich um das zu vermeiden. Man muss schon ausdrücklich die Begriffe Krankenschwester, Pfleger oder Ärztin / Arzt wählen, damit die KI nicht „durcheinander“ kommt (vergleichbar: Spennemann & Oddone, 2025, auch in Zweig, 2019, S. 208). Offenbar genauso wie der Mensch! Doch damit hört der Bias noch nicht auf: optische Repräsentation bezüglich Haarfarbe, körperlicher Eigenschaften und auch Hautfarbe bleiben stereotypisch, sofern nicht ausdrücklich auf diese Parameter bei der Erzeugung des Prompts geachtet wurde.

Mit dem Begriff Bias werden, wie oben gesagt, verzerrte Wahrnehmungen und einseitige Wahrnehmungsmuster beschrieben. Dabei ist folgendes wichtig:

- a) Objektivierbare Informationen müssten grundsätzlich verfügbar sein.
- b) Aus den objektiven Informationen müsste ein anderer Schluss möglich sein.

Es handelt sich also um eine „fehlerhafte“ Informationsverarbeitung. „Korrektere“ Informationen sind verfügbar, werden aber ignoriert. Wenn eine objektivere und alternative Information verfügbar ist, aber nicht berücksichtigt wird, kann das beabsichtigt ein – dann steckt z. B. die Idee dahinter, beim Rezipienten bzw. bei der Rezipientin ein bestimmtes Bild zu erzeugen. Es kann auch unbeabsichtigt sein: unbewusste Missachtung, Nutzen von Stereotypen (z.B. Elsen, 2023) oder auch eine gesellschaftliche oder kulturelle Gepflogenheiten und Traditionen. Es kann auch eine einseitige Quellennutzung vorliegen. Oder eben auch die Nutzung einer Ressource auf dem Computer, die eine falsche Quellenlage oder Datenbasis hat. Wissenschaftlich beutet „fehlerhaft“ den objektiven Vergleich zu einem Relativ, das in der wahrnehmbaren Realität grundsätzlich vorläge (Moosbach, 2025) aber nicht genutzt wird.

1.1 Der menschliche Bias

Grundsätzlich lässt sich festhalten, dass es eigentlich keine Möglichkeit gibt, Informationen in die Kategorien *Korrekt* oder *Falsch* einzusortieren. Der Übergang ist oft fließend, Informationen haben die Tendenz, teilweise korrekt oder auch teilweise fehlerhaft zu sein. Manchmal hängt es an der Art der Formulierung (z. B. unzutreffende Verallgemeinerungen sind hier ein sehr gängiges Beispiel):

„Deutsche lieben Schnitzel²“

Manchmal ist es auch der Kontext einer Aussage, der zu einem Bias führt, wie folgendes Beispiel zeigt.

In einem Schnitzelrestaurant wird der Geschäftsführer gefragt: „Warum haben Sie ein Schnitzelrestaurant eröffnet?“. Er antwortet: „Deutsche lieben Schnitzel.“

Version a) diese Frage wird in Köln gestellt³

Version b) diese Frage wird auf Bali gestellt

Vor allem werden in sehr vielen Fällen Informationen nicht von Maschinen, sondern von Menschen hergestellt – die ja wiederum diese Informationen basierend auf ihrer Wahrnehmung bereitstellen. Und Menschen sind eben keine Maschinen. Menschen reproduzieren keine Fakten auf der Basis vorliegender Fakten unter Anwendung von Regelsystemen wie mathematischer Logik (wie es eine Maschine tun würde), sondern interpretieren die sie umgebenden Informationen auf der Basis ihres kulturellen, sozialen, geschlechtlichen, gesellschaftlichen, religiösen, politischen, beruflichen, situativen, emotionalen etc. Hintergrundes (siehe z.B. Barrett, 2023).

Haben deswegen alle Informationen von Menschen ein Bias? Die Antwort ist ein klares „eventuell“. Problematisch kann bei dem Versuch der Beantwortung dieser Frage die Überlegung sein, dass die Entstehung von Wahrnehmungen gar nicht abschließend geklärt ist. Der Erkenntnisweg

2 Es ist durchaus nicht so, dass alle Menschen mit deutscher Staatsbürgerschaft Schnitzel lieben. Ich bin da z. B. eine Ausnahme.

3 Hier entsteht in Variante a) das Bild, das man schulterzuckend zur Kenntnis nimmt, während man in Variante b) unweigerlich an Touristen und Touristinnen denkt, die sich „typisch deutsch“ verhalten. Übrigens ist hier der Bias schon durch die Wahl des Nahrungsmittels vorprogrammiert. Eventuell schwingt bei dem entstehenden Bild dann auch Kulturkritik mit.

führt dabei über die Psychologie. Hier geht es vor allem darum, wie Vorstellungen auf der Basis vorliegender Informationen im Sinne der menschlichen Kognition (siehe z.B. Myers et al., 2014) entwickelt werden. Kurz gesagt umfasst Kognition dabei alles, was im weiteren Sinne mit Denken zu tun hat: Beurteilen und Urteilen, Verknüpfen, Schließen, Erkennen, Abstrahieren, Lernen etc. Eine verzerrte Wahrnehmung, also i. d. S. eine verzerrte kognitive Wahrnehmung, führt dazu, dass die aus der Wahrnehmung abgeleiteten Handlungen und Urteile ebenfalls einer Verzerrung unterworfen sind.⁴ Ein Bias kann zu einem Vorurteil führen, ein Vorurteil stärken, aber auch einen Entscheidungsprozess beschleunigen (was dann wiederum positive, weniger positive oder auch negative Konsequenzen haben kann). Die Notwendigkeit für menschliche Wesen in einer Welt voller Unsicherheit Entscheidungen treffen zu müssen, führt fast zwangsläufig zur Anwendung verschiedener Heuristiken. Eine Heuristik ist eine grobe Regel⁵, so wie die Luftlinie eine grobe Abschätzung der Distanz ist. Eine solche Abschätzung führt zu einer Verzerrung.

Beispielsweise besagt die „Luftlinie“, dass Rostock ca. 800 km von München entfernt ist. Aus dieser Information könnte man ableiten, dass München von Rostock aus in 8 Stunden per Auto erreichbar ist – dies wiederum stimmt aber nicht.

Fatalistisch formuliert könnte man sagen, menschliche Informationsverarbeitung hat immer einen gewissen Bias. Generell kann man grob vereinfacht sagen: es gibt Bias, die das Leben leichter machen, weil sie helfen, etwas grob einzuschätzen. Es gibt Bias, die das Leben schwerer machen, weil sie eine vorliegende Situation ungünstig verzerren. Und es gibt etliche Biases, die unbemerkt im Alltag verschwimmen⁶. Vielleicht ist das Kriterium in einer Welt voller Bias die Korrekturbereitschaft eines Individuums – liegt eine grundsätzliche Bereitschaft vor, die eigene Meinung oder auch die eigene Wahrnehmung zu hinterfragen, dann ist Bias eher eine Orientierungshilfe. Nehmen Menschen den eignen Bias nicht wahr, und existiert kein solches Korrekturbedürfnis, dann kann ein Bias zu einer einseitigen Weltsicht führen und ggf. auch sehr schädlich sein.

4 Das Wort „Verzerrung“ erzeugt interessanter Weise selbst ein Bias: während Sie dies lesen, liebe Leserin oder lieber Leser, werden Sie den Eindruck bekommen haben, dass eine solche Verzerrung etwas Negatives ist. Es ist aber durchaus möglich, dass etwas ins Positive verzerrt wird (z. B. „die eigenen Kinder sind die schlauesten“).

5 In Deutsch auch „Pi mal Daumen“ genannt.

6 Ein Beispiel dafür könnte die Merkatorprojektion sein – Hintergründe und Auswirkungen werden z.B. in Monmonier, 2004 dargestellt.

Wie bei vielen Dingen des menschlichen Lebens kommt nun in Form der menschlichen, biasbelasteten Informationsverarbeitung noch ein weiteres Problem hinzu: das gezielte Einsetzen von Bias kann genutzt werden, um Menschen zu einer bestimmten Meinung zu bringen, in einer Meinung zu halten oder auch um unbewusste Entscheidungen zu beeinflussen. Dies alles passiert, ohne dass es den betroffenen Menschen bewusst ist. Und dann wird Bias zu mehr als einer brauchbaren Alltagsheuristik – dann wird Bias zu einem (politischen) Instrument. Die Psychologie hat hierfür sogar eigene Namen entwickelt, beispielsweise:

- Der Ankereffekt (Anchoring Bias) – hier werden Ankerreize gesetzt um unbewusst Entscheidungen zu beeinflussen (z.B. Cho et al., 2017, Wason, 1968).
- Der Bestätigungsbias (Confirmation Bias; Jermias, 2001), die Suche nach Informationen, die die eigene Wahrnehmung oder Interpretation bestätigen. Im Sinne eines Instrumentes würde diese Information dann von der beeinflussenden Gruppe gezielt bereitgestellt oder zitiert werden.

Kognitiver Bias ist in verschiedenen Variationen untersucht worden (siehe z. B. *The Cognitive Bias Codex*, eine Grafik, die viele verschiedene Formen des Bias zusammenstellt, verfügbar unter Wikipedia, 2025b, oder auch Eppler & Muntwiler, 2025).

1.2 Der Bias im Computer

Bis hierher ging es darum, dass Menschen bestimmte Informationen gefiltert aufnehmen, oder entsprechend einer vorliegenden Struktur einsortieren, und dann die daraus resultierenden Verzerrungen selbst nicht mehr wahrnehmen und entsprechend auch kein Korrekturbedürfnis haben. Für viele Formen des Bias ist es jedoch aus ethischer Sicht wichtig, dass Menschen sich der Existenz eines Bias, also einer systematischen Verzerrung, bewusst werden und die Konsequenzen ihrer Entscheidungen absehen und diese ggf. korrigieren (z. B. Martens & Cap, 2025). Was für Möglichkeiten gibt es aber, wenn man den Verdacht hat, einer verzerrten Beurteilung anheimgefallen zu sein und das Bedürfnis hat, dies zu korrigieren? Bisher war ein Weg aus diesem Dilemma das Vorliegen von einer größeren Menge an Information und die Zuhilfenahme von möglichst neutral arbeitenden Werkzeugen.

Werkzeuge wie Computer können zur Analyse der Information, auch hinsichtlich gegebenenfalls vorkommender Bias, genutzt werden. Ebenso nutzt man Computer zur Produktion von alternativen Darstellungen der Erkenntnisse, die aus den Daten gewonnen werden. Grafiken und Visualisierungen sind dabei gute Hilfsmittel. Als datenverarbeitende Maschinen ohne Religion, kulturelle oder gesellschaftlich Hintergründe haben Computer hypothetisch keinen Bias. Computer arbeiten mit Algorithmen und Daten, aber sie haben keine Tagesform, keine Glaubenssätze und keine Emotionen. Sie verarbeiten Daten theoretisch „wertneutral“. Die Grundlage der aktuell verfügbaren Modelle der Künstlichen Intelligenz sind Algorithmen, deren Funktionsfähigkeit zu einem großen Teil darauf basiert, dass sie spezialisiert sind, sehr große, für den Menschen kaum noch durchsuchbare Datenmengen zu analysieren und die darin vorhandenen Informationen in menschenverständlicher Art und Weise aufzubauen.

Bis hierhin ist es eine richtig gute Idee, Computer zu nutzen um Bias zu beseitigen oder zumindest, um sich eines Bias bewusst zu werden: es werden möglichst viele, möglichst quantitative Daten gesammelt, ausgewertet und ggf. per logischem Schließen miteinander verbunden. Leider klappt das in der Praxis nicht so gut. Um das zu verstehen muss ein kurzer Ausflug in die aktuell verbreiteten Formen der Künstlichen Intelligenz gemacht werden.

2. KI – Künstlich und Intelligent?

Künstliche Intelligenz (KI) geht als Begriff auf die 1950er Jahre zurück (z. B. (Russell & Norvig, 2022)). Es gibt viele verschiedene Ansätze, die im Laufe der Jahre entstandenen Varianten der Künstlichen Intelligenz abzugrenzen. In diesem Kapitel wird auf die Unterscheidung der Art der Verarbeitung fokussiert. Hier wird, bereits in historischen Quellen, zwischen symbolischen und konnektionistischen Modellen unterschieden. Der Unterschied besteht grob gesagt darin, ob die Modelle auf der Basis von (menschenslesbaren) Fakten und Regeln arbeiten oder ob sie abstraktes „Wissen“ anhand von Daten trainieren und z. B. zum Aufbau eines Neuronalen Netzes verwenden. Aus der Perspektive des Bias ist es interessant, sich beide Modellformen kurz anzuschauen und den Hintergrund, vor dem sie entstanden sind, zu erläutern.

2. 1 Symbolorientierte KI

Die ersten erfolgreichen Modellformen, die zum Beispiel in den 1960ern zu Programmen wie ELIZA (Weizenbaum, 1966) führten, basierten im Wesentlichen auf einer Zerlegung von menschlicher Sprache in Fakten und Regeln. ELIZA ist die erste „sprachverarbeitende“ KI. Eins der Skripte, mit denen die Software arbeiten kann, ist ein „Therapieprogramm“ basierend auf den Regeln zur Gesprächstherapie nach Carl Rogers. Weizenbaum hat das Programm ausdrücklich nicht zum Zwecke der Therapie entwickelt, sondern diesen Kontext nur genutzt, weil das Protokoll der Analyse der Satzkonstrukte entsprechend Carl Rogers sehr gut maschinenverarbeitbar ist. Vielmehr ging es Weizenbaum um Versuche zur Verarbeitung menschlicher Sprache (siehe z. B. Weizenbaum, 1966) auf den damaligen Rechenmaschinen (in diesem Fall eine IBM 7094). Weitere Programme, wie MYCIN (Shortliffe, 1976), nutzten Fakten und Regelsysteme – in der Medizin KI Software MYCIN konnte man zu Symptomen eine Diagnose ausgeben lassen. Der Vorteil dieser Systeme war, dass sie meist nur „korrekte“ Fakten und Regeln enthielten und man im Prinzip belegen konnte, auf Basis welcher Information ein Ergebnis entstanden war.

Es war durchaus möglich, die zugrundeliegenden Expertensysteme auch lernfähig zu machen – dies ging aber oft auf Kosten der Korrektheit (siehe z.B. Russell & Norvig, 2022). Lernfähig heißt dabei, dass unter Zuhilfenahme von logischem Schließen, später auch durch Maschinelles Lernen (engl. Machine Learning ML), die Wissensbasis nach neuen Zusammenhängen durchsucht wurde – diese neuen Zusammenhänge gingen dann ebenfalls in die Wissensbasis ein. Der große Nachteil der wissensbasierten Expertensysteme war der Umstand, dass das Wissen gewissermaßen „per Hand“ in die Datenstrukturen eingegeben werden musste. Das Wissen des Experten oder der Expertin ist quasi „im Kopf“ der betreffenden Person und kann nur mühselig dort wieder herausgeholt, abstrahiert und in den Computer gebracht werden. Ein ganzes Forschungsfeld ist auf diese Weise entstanden: Wissensmanagement (Reinmann & Mandl, 2000) und auch Knowledge Engineering (Porter et al., 2008). Besonders schwierig ist die Konstruktion solcher Systeme in Bereichen, in denen wenig Wissen vorliegt oder in denen Wissen nur sehr unscharf ist (z. B. mehr qualitativ als quantitativ). Ein Wandel in der Erkenntnis über Struktur und Funktionsweise des menschlichen Gehirns und der menschlichen Informationsverarbeitung führte in der Psychologie zu einem Paradigmenwechsel vom

Behaviorismus zum kognitionswissenschaftlichen Paradigma. Dies wiederum beeinflusste auch die Informatik und führte zu einer weiteren Idee, den konnektionistischen Systemen.

2.2 Konnektionistische KI

Auch wenn die ersten konnektionistischen Systeme nahezu gleichzeitig mit symbolorientierten erdacht wurden (z. B. „Das Perzeptron“, 2025), so wurden sie doch erst in den letzten Jahren mit sehr großem Erfolg als KI eingesetzt (siehe z. B. ChatGPT (OpenAI, 2024)). Im Gegensatz zu den klassischen Expertensystemen, deren Wissen vorgegeben ist, basieren konnektionistische Systeme auf dem Trainieren von Wissensstrukturen auf der Basis vorliegender (unstrukturierter und oft unscharfer) Trainingsdaten. Oft werden diese Systeme daher auch als datengetriebene Systeme bezeichnet. Grob kann man hier zwei Systemarten unterscheiden: Verfahren, die auf Daten lernen und die über probabilistische und statistische Auswertungen über diese Daten Anpassungen des Algorithmus vornehmen und damit dann Aussagen über neue Daten treffen können (Generalisierung). Diese werden generell als Machine Learning (ML) Systeme bezeichnet. Traditionell (ihr Ursprung lässt sich ebenfalls auf die 1950er Jahre zurückführen) wurde ML nicht unbedingt zu den KI-Verfahren gezählt, weil die primäre Aufgabe von maschinellem Lernen nicht die Nachahmung menschlicher Intelligenz war, sondern das Trainieren von Algorithmen anhand von vorliegenden unstrukturierten Daten. Mit dem allgemeinen Wechsel von den symbolorientierten zu anderen, noch stärker datengetriebenen, stochastischen Verfahren gilt heute, dass ML ebenfalls als KI gelten (so auch dargestellt in Russell & Norvig, 2022 in der aktuellen Ausgabe von 2022 im Vergleich zu älteren Ausgaben). ML von anderen konnektionistischen Verfahren abzugrenzen wird entsprechend schwierig.

Die andere, klassische Sicht auf konnektionistische Verfahren ist, dass hier mindestens eine Netzstruktur vorliegt, auf der gelernt wird: das sogenannte Neuronale Netz oder auch Deep Learning (DL)⁷. Die Netzstruktur entspricht dabei von der Idee her der damaligen Vorstellung, dass vernetzte Neuronen das Gehirn bilden. Das Gehirn ist in dieser Weltsicht ein

7 Der Wortbestandteil „Deep“ in Deep Learning bezieht sich auf die „Tiefe“ der Schichten des Neuronalen Netzes. Aus Sicht der Informatik sind Deep Learning Netze und klassische Neuronale Netze zwar verwandt, technisch aber nicht das gleiche, z. B. hinsichtlich der Netztiefe und der Verarbeitung.

informationsverarbeitendes „Gerät“ (Myers & DeWall, 2023), und kann entsprechend auch auf einer Maschine nachgebaut werden, wenn man genügend viele Daten hat. Das Vorgehen des Lernens in diesen Netzstrukturen ist anders als bei den ML-Verfahren: ein programmierender Mensch bestimmt den abstrakten Aufbau des neuronalen Netzes (noch ohne Daten). Ein solches Netz besteht in der Regel aus mehreren Schichten (sog. *Layers*) von künstlichen Neuronen. Es gibt dabei normalerweise eine Eingabeschicht und eine Ausgabeschicht und mehrere versteckte Schichten (sog. *Hidden Layers*). Jedes Neuron in diesem Netz hat jetzt verschiedene mathematische Berechnungsschritte in sich: es wird (stark vereinfacht) ein Eingabewert berechnet, ein interner Schwellenwert und ein Ausgabewert. Da Neuronen in Form des Netzes hintereinandergeschaltet sind, bestimmt die Kombination der Eingaben und der Schwellenwerte dann, was das Netz lernt. Auch die Art des Lernverfahrens wird vom programmierenden Menschen bestimmt. Grob wird dabei zwischen überwachtem (supervised) und unüberwachtem (unsupervised) Verfahren unterschieden. Das, was das Netz gelernt hat, liegt erst nach dem Training mit sog. Trainingsdaten vor. Das Grundprinzip ist dabei der Funktionsweise der menschlichen Neuronen abgeschaut – zumindest was den Wissensstand um das Jahr 1970 anging (z. B. Perzeptron als einfachstes künstliches Neuron Rosenblatt, 1958 und Nilsson, 2010). Neurologische Forschung zeichnet inzwischen ein anderes Bild. Eine ausführliche Erklärung der Funktion und des Aufbaus der inzwischen recht komplex gewordenen Berechnungen in einem neuronalen Netz findet man beispielsweise bei Bader und Kirste (z. B. Bader & Kirste, 2025).

Ein wichtiger Baustein bei der Entwicklung eines neuronalen Netzes ist, ähnlich wie bei ML, dass das Netz auf Daten trainiert wird. Und genau hier liegt auch die Quelle von etwaigen Bias. Wie auch ML findet das Netz durch das Lernen in den Trainingsdaten Strukturen die es mathematisch auswertet und, grob gesagt, zu Wahrscheinlichkeiten des Auftretens zusammenfasst. Diese Wahrscheinlichkeit des Auftretens von Zusammenhängen von Daten beim Vorliegen von sprachlichen Daten kann sich auf die Idee „welche Zeichen folgen welchen Zeichen“, „welche Worte folgen welchen Worten“ beziehen. In der KI redet man in der Regel dann von Token im Sinne von kleinen, bedeutungstragenden Einheiten, also „welche Token folgen auf welche Token“⁸. Gemäß dieser trainierten

8 Bei Bilddaten wäre das entsprechend „welche Bildinformation folgt welcher Bildinformation“.

Wahrscheinlichkeiten oder anders gesagt auf Basis der zugrundeliegenden Gewichtungen und mathematischen Funktionen „lernt“ das Netz dann und ist demzufolge auch in der Lage, Aussagen über neue Informationen zu treffen, die über die Trainingsdaten hinausgehen (z. B. mittels Generalisierung, Vorhersage, Beurteilung, logisches Schließen etc.). Bei der Sprachverarbeitung redet man hier von statistischen Sprachmodellen (siehe z. B. Bader & Kirste, 2025).

Die „Korrektheit“ einer Aussage hängt aber in den Trainingsdaten und diesen Umstand muss man sich merken. Nachdem das Training abgeschlossen ist, lernt das Neuronale Netz nicht weiter! Entsprechend hat ein 2024 trainiertes Netz keine Informationen aus dem Jahr 2025. Eine KI Software ist daher auch keine Suchmaschine – wenn nach aktuellen Themen in einer Suchmaschine gesucht wird, dann kann man in Abhängigkeit von Suchbegriffen und Anfragetiefe davon ausgehen, dass viele aktuelle Themen, sofern sie im Internet zu finden sind, auch angezeigt werden. Bei einer KI wie ChatGPT (von OpenAI, 2024) oder vergleichbaren Produkten ist das nicht der Fall: ChatGPT 5 hat austrainiert – da es im August 2025 auf den Markt gekommen ist, hat es keine aktuellen Informationen aus der Zeit danach in den Trainingsdaten integriert. Wenn eine KI Software nun Anfragen zu aktuellsten Themen bekommt, dann kann sie mittels einer Softwarerweiterung (z. B. bei Gemini und ChatGPT als „aktive Websuche“) auch Daten aus dem Internet ziehen – diese Daten hat das Netz aber nicht trainiert, sie sind nicht Bestandteil der Wissensbasis der KI Software, sie sind anfragespezifisch, liegen unter Umständen nicht immer gleichartig vor, sind eher als die Trainingsdaten in sich inkonsistent und unterliegen auch sprachlichen Schwankungen. Ohne aktivierte Websuche gibt es keine verlässlichen Antworten auf aktuellste Themen.

Für die Korrektheit der Arbeit eines KI Modells bleibt also die Verantwortung wesentlich auf der Ebene der Trainingsdaten hängen. Natürlich können Menschen in die Gestaltung der Trainingsdaten eingreifen und sie tun das auch in verschiedener Weise, z. B. hinsichtlich der Auswahl der Daten, der Bandbreite der Daten, Data Analytics und auch der Bereinigung der Daten hinsichtlich bestimmter ethischer Gesichtspunkte⁹. Auch

9 Die Art und Weise, wie das wirtschaftlich oft umgesetzt ist, führt leider zu einem weiteren ethischen Problem, das man unter dem Stichwort „Clickworker“ recherchieren kann (Muldoon et al., 2024).

die Art und Weise wie ein Lernverfahren dann lernt kann unterschiedlich gestaltet werden. Hier gibt verschiedene Trainingsansätze, unüberwachte und überwachte, einer davon wäre Reinforcement Learning on Human Feedback (RLHF) – auch hier wird beim Training menschliches Feedback mit einbezogen.

Haben die Menschen, die an den Daten und den Algorithmen arbeiten, die Daten vorselektieren oder bereinigen, ein Bias (beabsichtigt oder unbeabsichtigt), dann findet sich dieser Bias genau so auch in der KI wieder. Man kann dabei verschiedene Ebenen des Bias entscheiden: ein Bias kann die Auswahl und das Labeling der Daten beeinflussen, die Struktur des Algorithmus bzw. der Verarbeitung, den Aufbau des zugrundeliegenden Modells und die Designentscheidungen (z. B. warum sind Alexa und Siri in der Grundeinstellung weiblich gelesen Stimmen?).

So wurde beispielsweise Gesichtserkennung nachgewiesener Weise nicht mit Menschen dunkler Hautfarbe trainiert und arbeitet entsprechend viel schlechter bei diesen Personen (zu Bias bei Hautfarbe und Geschlecht, Studie von Buolamwini & Gebru, 2018). Ditz und Lichtmeß bezeichnen dies als Evaluationsbias (Ditz & Lichtmeß, 2025). Der hier nachgewiesene Bias ist aufgrund der Art und Gestaltung der Trainingsdaten (vorwiegend weiße männliche Gesichter) derartig tief im KI System verankert, dass auch Forschende der ZHAW (Züricher Hochschule für Angewandte Wissenschaften) es mit verschiedenen Maßnahmen nicht geschafft haben, den Bias aus dem Gesichtserkennungssystem herauszubekommen (Details siehe in Wehrli et al., 2022). Die Konsequenz ist, dass Gesichtserkennungssoftware einen Bias hat: zuverlässig erkannt werden in erster Linie Gesichter von sog. „weiß“¹⁰ aussehenden Menschen, am besten noch, wenn sie männlichen Geschlechtes sind.

10 Gelegentlich wird hierfür die Bezeichnung „kaukasisch“ verwendet, die aber deutlich veraltet ist.

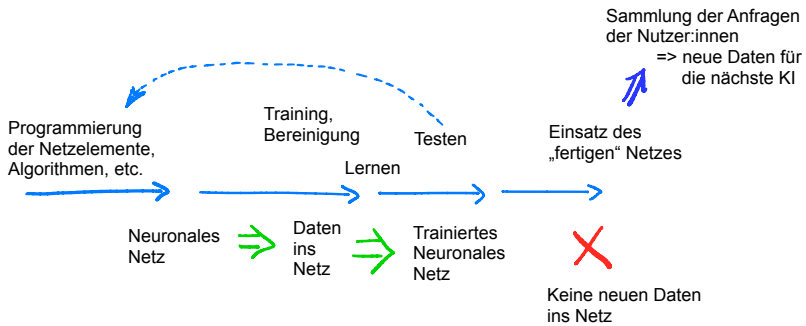


Abbildung 1: Neuronales Netzwerk und Lebensphasen der KI

2.3 Zwischenzusammenfassung: KI

Zusammengefasst kann man sagen: der Flaschenhals bei den symbolorientierten KI-Systemen, wie den Expertensystemen, ist die Datensammlung und -aufbereitung. Der Flaschenhals bei den KI-Systemen, die auf datengetriebenen, stochastischen oder konnektionistischen Modellen arbeiten, ist auch die Datensammlung. Hier jedoch geht es nicht darum, die Daten überhaupt erstmal aus den Experten und Expertinnen zu bekommen (siehe Abschnitt 2.1), sondern es geht darum, alle verfügbaren Daten bezüglich der Qualität für das angestrebte Modell zu bewerten (siehe Abschnitt 2.2).

In der Informatik gibt es dazu eine schöne Abkürzung, die eigentlich schon alles sagt: „Garbage In – Garbage Out“ (Wikipedia, 2025a)– wenn man Müll rein gibt, kommt auch nur Müll raus (siehe Abbildung 2). Im Folgenden werden vor allem die datengetriebenen und konnektionistischen Verfahren betrachtet. Aktuelle Beispiele für solche Verfahren sind die Software ChatGPT von OpenAI (OpenAI, 2024), DeepSeek von DeepSeek (DeepSeek, 2025), Gemini von Google (Gemini Google, 2025) oder CoPilot von Microsoft (CoPilot Microsoft, 2025).

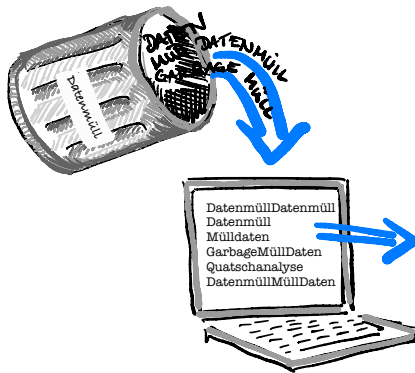


Abbildung 2: Garbage-in Garbage-out

3. Datenbias und Algorithmenbias

In der Informatik unterscheidet man ganz allgemein Probleme die mit Daten verbunden sind und Probleme, die durch die auswertende oder verarbeitende Algorithmetik (i. S. v. Programmierung) entstehen. Daten sind die Essenz der Arbeit mit dem Computer – diese Daten entstammen idealer Weise bestimmten, wohldefinierten Messprozessen oder dezidierten Eingaben, z. B. in Datenbanken oder vergleichbaren Formaten. Messen und Sammeln von Daten sind in der Regel gezielte Prozesse, bei denen im Vorhinein überlegt und entschieden wird, welche Daten relevant sind und welche Daten erhoben werden. Eine weitere Möglichkeit, Daten zu bekommen, besteht im Zugriff auf Datenquellen wie Onlinebücher, Blogs oder öffentliche Äußerungen im Netz. Verfahren wie Maschinelles Lernen oder auch Neuronale Netze benötigen sehr viele Daten – zum Durchsuchen und Finden von Mustern oder auch zum Trainieren. Das Thema Datenbias wird im nächsten Abschnitt vertieft betrachtet und ist eine der Hauptstellen, in denen überhaupt bei der Arbeit mit dem Computer ein Bias entsteht. Die andere Stelle, an der Bias auftreten kann, ist die Gestaltung und der Einsatz der Algorithmen. Diese wiederum sind Kern der Verarbeitung auf dem Computer und auch hier kann zwischen Idee, Programmierung und tatsächlichem Einsatz unterschieden werden – auch hier gilt, nicht alles was gut gedacht war, wird „richtig“ eingesetzt. Dies wird ebenfalls im folgenden Abschnitt genauer betrachtet.

3. 1 Datenbias

Knapp gesagt: Daten sind aufbereitete, computerlesbare Informationen, die ihren Ursprung irgendwo haben können – beispielsweise personenbezogene Daten, beobachtete Daten, gemessene Daten, oder eben auch alle erdenklichen Daten, die im Netz zur Verfügung stehen: Blog, Social Media Beiträge, Musik, Filme, Kommentare, Fotos etc. Bei der Entwicklung von ChatGPT (OpenAI, 2024) beispielsweise wurde alles als Trainingsdaten verwendet, was man im Internet finden konnte¹¹. Das Problem ist hierbei weniger die Fülle an Daten, sondern die Verzerrungen, die durch die Datensammlung, Datenauswertung und das Training auf Daten und die automatisierte Produktion neuer Daten durch KI-Systeme entstehen. Da dies sehr undurchschaubar ist, wird das Thema in den nächsten Schritten aufgeteilt in die Betrachtung der Daten selbst, in Formen der Datenerhebung, in Probleme der Datenauswertung und dann in Überlegungen im Kontext von Bias in KI.

Die Welt wird gefühlt etwas handhabbarer, wenn sie in quantitative (zählbar, messbare) und qualitative (eine Qualität besitzende) Dinge unterschieden wird. Ein Beispiel für quantitative Dinge ist z. B. „drei Bananen“. Quantitative Dinge haben die schöne Eigenschaft, dass sie objektivierbar sind, dass man sie zählen kann, und, je nach verwendetem Skalenniveau (Bühner, 2021), dass man ggf. auch mit ihnen rechnen kann. Skalenniveaus sollten üblicherweise vor der Erhebung von Daten festgelegt werden¹². Qualitative Dinge, wie z. B. „die leckeren Bananen“ unterliegen einer subjektiven Einschätzung¹³. Hier wird es ungenau. Die zählbaren, messbaren, mathematisch bestimmbaren Dinge können genutzt werden, um einem Bias entgegen zu wirken. Aber das funktioniert nur, wenn empirisch sauber gearbeitet wird¹⁴. Hier sind empirische Instrumente ein wichtiger Bestandteil und daher auch fest in der Forschung verankert. Bei qualitativen Dingen wird das etwas schwieriger, aber auch hier hat die Wissenschaft Wege gefunden, objektivierbare Aussagen abzuleiten.

11 Tatsächlich oft unter Missachtung aller Formen des Schutzes von privaten Daten.

12 So macht es beispielsweise wenig Sinn, die Temperaturdaten zu addieren, es ist aber allein durch die Art der Skala schon klar, dass 3 Grad Celsius weniger ist als 10 Grad Celsius (es handelt sich hierbei um eine Intervallskala). In einer Nominalskala wäre diese Aussage nicht möglich – eine Ziffer 3 im Autokennzeichen sagt nichts aus im Vergleich zur Ziffer 10.

13 Was ist für Sie eine „leckere Banane“? Ich mag die gelben, die noch ein kleines bisschen grün sind, am liebsten.

14 „Ich traue keiner Statistik, die ich nicht selbst gefälscht habe.“ (Krieghofer, 2017)

Der Unterschied zwischen den Formen der Datenerhebung ist mit dem Forschungsansatz verbunden:

1. Daten werden gezielt gemessen. Dem voran geht eine Überlegung, die besagt, welche Daten die bestmögliche Aussage über den zu untersuchenden Sachverhalt darlegen könnten bzw. die getroffenen Aussagen bestmöglich stützen können. Diese gemessenen Daten könnten Wetterdaten sein, sie könnten aber auch Trackingdaten sein, die die Bewegung von Personen im öffentlichen Raum messen oder Logdaten, welche z.B. aufzeichnen, welche Personen wann, wie lange und mit welcher Software online waren. Das Ergebnis gemessener Daten sind in der Regel quantitative Werte. Diese könnten – zumindest im Prinzip – miteinander in Verbindung gestellt werden. Die Art der Datenerhebung wird initial festgelegt und nicht mehr geändert.
2. Daten werden anhand von Stichproben (randomisiert) erhoben. Hier geht es in der Regel nicht um reines Sammeln, sondern um die gezielte Erhebung von Daten aus einer sehr gut überlegten und ausgewählten (sowie dokumentierten) Stichprobe. Von den Aussagen, die diese Stichprobenuntersuchung liefert, wird dann auf allgemeinere Aussagen geschlossen (Induktion). Wichtiges Kriterium ist dabei, dass die Stichprobe die zu untersuchende Grundgesamtheit genügend gut abbildet (sie ist repräsentativ und reliabel) – so wäre es vergleichsweise sinnlos, die Bewohner:innen eines Altenheims hinsichtlich des Bedarfs an Kindergartenplätzen zu befragen. Wichtig ist, dass das Erhebungsinstrument feststeht und nicht mehr geändert wird, solange die Datenerhebung läuft. Wissenschaftlerinnen und Wissenschaftler überprüfen diese Arbeit idealerweise gegenseitig hinsichtlich etwaiger Bias.
3. Im Sinne der Forschungsrichtung „Grounded Theory“ (datengestützte Theoriebildung), die in den Sozialwissenschaften verbreitet ist und in der Informatik zunehmend Beliebtheit findet, gibt es noch einen anderen Weg, Daten zu bekommen, insbesondere wenn es sich um qualitative Daten handelt. Hier wechseln sich, stark vereinfacht, Phasen der Datenerhebung, der Datenanalyse bzw. -auswertung und einer Anpassung des Erhebungsinstrumentes ab (mehr Details z. B. in Stol et al., 2016). Die Kontrolle ist hier schwieriger und erfordert eine hohe Sorgfalt. Auch wenn hier im Laufe der Erhebung die Erhebungsinstrumente angepasst werden, liegen sie dennoch zur kritischen Analyse vor – und auch hier arbeiten Wissenschaftlerinnen und Wissenschaftler zusammen an der Akkuratessse der Daten.

All die gesammelten Daten, egal wo sie herkommen, können durch einfache Bearbeitungsschritte im Computer verändert werden und neue Daten erzeugen. Dies kann beispielsweise unter Zuhilfenahme mathematischer Berechnungen erfolgen (O’Neil, 2017). Daten, die auf Basis der Auswertung anderer Daten entstehen können, sind: Meta Daten (also Daten über Daten, beispielsweise durch statistische Auswertungen), interpretierte Daten, zusammengefasste oder auch reduzierte Daten. Aus Sicht des Bias ist vor allem die Form der Datensammlung und der Auswertung von Daten (Datenanalyse) zu betrachten, aber auch die Art, wie Daten repräsentiert werden oder wie mit Daten interagiert wird, führt ggf. zu einem Bias. Hier muss man vor allem berücksichtigen, dass eine Korrelation, also ein gleichzeitiges Auftauchen von bestimmten Daten, nicht unbedingt eine Kausalität, also einen Ursache-Wirkungszusammenhang, bedeuten (viele teilweise sehr lustige Beispiele sind auf der folgenden Seite zu finden (Vigen, 2015a) und in dem Buch von Tyler Vigen (Vigen, 2015b), interessant ist auch dieses Beispiel (Steinmann, 2023)).

Das Problem besteht darin, dass zwischen der Datensammlung, der Nutzung von Daten zum Training der KI, zur allgemeinen Analyse von Daten und dem Einsatz einer fertig trainierten KI-Software unterschieden werden muss. Ein Bias in der Datenerhebung führt dann zu einer Verzerrung in den vorliegenden Daten. Dies wiederum führt zu einer verzerrten Auswertung und Analyse – und zu einem Trainieren von Verzerrungen in der KI. Entsprechend sind dann die Ausgaben – vor allem der konnektionistischen, lernenden und datengetriebenen KI-Systeme (siehe Abschnitt 2) – durch Bias verzerrt. Schlimmstenfalls kann ein Training und Lernen auf verzerrten Daten den Bias in der resultierenden KI-Software sogar noch verstärken.

Wenn man verschiedene Lebensphasen eines KI-Systems betrachtet (siehe Abbildung 1), dann kommt man unweigerlich zu verschiedenen Biasformen, die während der KI-Entwicklung und dem KI-Einsatz auftauchen können. Ditz und Lichtmess (2025) unterscheiden beispielsweise folgende Formen von Datenbias, die bereits in der Phase der Datenerhebung auftauchen können:

- Historischer Bias – dieser Bias tritt dann auf, wenn Daten vorliegen, die aus heutiger Perspektive nicht mehr zeitgemäß sind. Ein historischer Bias kann beispielsweise reflektieren, dass bestimmte Personengruppen in bestimmten Berufen unterrepräsentiert sind (Genderbias). Die

Verzerrung liegt dann nach der Verarbeitung durch die KI als neues „Fakt“ vor und beeinflusst schlimmstenfalls aktuelle Entscheidungen.

- Repräsentations- oder Auswahlbias – hierbei werden bestimmte Gruppen der Bevölkerung in den Trainingsdaten nicht berücksichtigt oder unterrepräsentiert. Ursachen hierfür können vielfältig sein – das Resultat ist jedoch stets dasselbe: die betroffene Personengruppe kommt in den Trainingsdaten nicht oder nur geringfügig vor und wird dann bei KI generierten Ergebnissen ebenfalls seltener oder gar nicht vorkommen. So ein Auswahlbias kann sachbezogene Ursachen haben oder auch zeitbezogene: wenn nur bestimmte Zeitpunkte für Stichproben herangezogen werden, ergeben Daten etwas ganz anderes, als wenn eine Betrachtung über einen längeren Zeitraum erfolgt. Diese Art von Verzerrung ist oft bereits beim Entwurf von Studien zu finden und in der Psychologie lange bekannt (z. B. Bühner, 2021), in der Entwicklung von KI aber leider oft zu finden.
- Messbias – hier wird ein unangemessenes Messinstrument ausgewählt oder auch ein unpassendes Skalenniveau. Etliche Beispiele dazu sind nachzulesen bei Cathy o’Neil (2017), die in ihrem Buch klar aufzeigt, wie Mathematik genutzt werden kann um Daten gezielt zu manipulieren oder wie das Ergebnis einer maschinellen Berechnung auch durch die Auswahl unangemessener Messelemente einen Bias enthält. Ein Messbias kann auch durch das Vorliegen zu weniger Daten entstehen.
- Ignorieren von Variablen – die Messung erfolgt hier unter Nichtberücksichtigung relevanter Items oder Messvariablen, wie beispielsweise das Alter von Patienten bzw. Patientinnen oder einseitige Auswahl von Befragungskohorten. Hier kann noch dazukommen, dass ignoriert wird, von wann die erhobenen Daten sind.

Das Problem ist, wie in Abschnitt 2 angedeutet, dass Bias in Daten zu Bias in KI führt. Menschen, die gerne die Aussagen von Maschinen als „Wahrheit“ annehmen, weil eine Maschine (wie ein Computer) ja keinen Bias haben kann, nehmen dann die biasgefärbten Ergebnisse einer maschinellen Bearbeitung und Auswertung und gehen davon aus, dass es sich um eine neutrale, nicht verzerrte Information handelt. Das ist leider falsch.

3.2 Algorithmenbias

Es ist schwierig, den reinen Algorithmus einer Software von den Daten einer Software gedanklich zu trennen. Für Informatiker:innen ist ein Al-

gorithmus eine wiederkehrende, eindeutige Handlungsanweisung zur Lösung eines Problems oder zur Bearbeitung einer Problemklasse. Dabei wird eine Eingabe durch verschiedene Verarbeitungsschritte, die intern im Computer erfolgen, in eine Ausgabe überführt. Im Sinne des vorliegenden Kapitels nimmt der Algorithmus Daten und verarbeitet diese zu neuen Daten. Leider ist es nun so, dass nicht nur aufgrund einer verzerrenden Datenerhebung oder einer schlechten Datenlage ein Bias in einem Softwaresystem (egal ob mit oder ohne KI) auftreten kann. Auch eine Verzerrung bei der Gestaltung der auswertenden Algorithmen kann zu einem Bias führen (Koska, 2023; Jaume-Palasi & Spielkamp, 2017). In der Algorithmenethik unterscheidet man unter ethischen Gesichtspunkten verschiedene Phasen (Floridi, 2010):

1. Idee zum Algorithmus

Hier geht es darum, herauszufinden, ob ein Algorithmus entwickelt wurde, um gezielt Bias zu verbreiten oder umzusetzen (z. B. aus politischen oder geschäftlichen Interessen, siehe z. B. Anchoringeffekt), oder ob man bei der Idee zu einem Algorithmus bereits einem eigenen gedanklichen (kognitiven) Bias aufgesessen ist (so z. B. in einigen Beispielen in O'Neil, 2017).

2. Umsetzung des Algorithmus (hier kann noch zwischen Konzeption und Programmierung unterschieden werden)

So gute eine Idee auch sein mag – es muss sichergestellt werden, dass bei der Programmierung sauber gearbeitet wird. Wenn die Daten bestimmte Informationen nicht oder nur verzerrt hergeben, wird der Algorithmus dies selten bereinigen können. Wenn der Algorithmus bestimmte Aspekte von Daten ausschließt oder nur bestimmte Aspekte berücksichtigt (z. B. Mauro & Schellmann, 2023), dann arbeitet der Algorithmus mit einem Bias. Dies ist oft in Social Media zu erleben.

3. Nutzung bzw. Einsatz des Algorithmus

Nach der Programmierung muss sichergestellt werden, dass der Algorithmus nur im erdachten Sinne verwendet wird (Stichwort *Dual Use*). Hier kann es z. B. durch falschen Einsatz von Algorithmen zu einem emergenten Bias kommen (Wikipedia, 2025c). Die Emergenz bezieht sich z. B. darauf, dass sich eine bereits trainierte KI nicht auf neues Wissen anpasst (z. B. dargestellt in O'Neil, 2017). Darüber hinaus muss überprüft werden, dass der Algorithmus auch genau das tut, was er soll (Abgleich mit der Idee und der Umsetzung).

Eine über den Datenbias hinausgehende Verzerrung in der algorithmischen Verarbeitung wird immer dann möglich, wenn ein Algorithmus weitere Informationen auswertet, wie z. B. Häufungen von Daten analysiert und zur weiteren Berechnung heranzieht (z. B. Zweig, 2019, S. 209). Hier kann z. B. ein Correlation Bias entstehen – Korrelationen werden algorithmisch erkannt, ohne dass sie offengelegt werden, führen aber zu Ergebnissen, die dann den Nutzer:innen ausgegeben werden. Der Bias wird dabei nicht bemerkt. Dies ist unter anderem in Recommendersystemen für die Auswahl in Social Media bekannt. Ein Recommendersystem ist eine Art Empfehlungsdienst. Hier bewertet ein Algorithmus, wie „stark“ das Interesse eines Menschen an etwas (angenommener Weise) ist. Ein Beispiel ist auch das Beobachten von Nutzerverhalten auf Webseiten oder das Speichern von Cookies, was zu einer anderen algorithmischen Auswertung genutzt werden kann¹⁵. Die Recommenderdienen zwar dem Schutz vor Informationsüberflutung, de facto sortieren sie aber auch vor und halten den Nutzer oder die Nutzerin in einer „Informationsblase“ oder „Filterblase“ (Wikipedia, 2025d). Diese kann sich auf die angezeigten Webseiten, aber auch auf Musikauswahl auf einer entsprechenden Plattform oder Buchauswahl in einem Online-System beziehen. Der Bias entsteht dadurch, dass das Entdecken neuer Angebote zugunsten bekannter (oder bestimmter vorselektierter) Angebote unterdrückt wird. Gezielt eingesetzt können Nutzende zu bestimmten Werbeangeboten geleitet werden oder auch von bestimmten Angeboten ferngehalten werden.

Algorithmisch könnten auch Daten an bestimmte Anforderungen adaptiert (z. B. Dauer des Verweilens auf einer Seite), weitere Datenquellen hinzugenommen werden (z. B. Vergleich von Userverhalten, was ebenfalls durch Recommendersysteme erfolgt). Algorithmen könnten stereotypisch auswerten oder unangemessene Vergleiche durchführen (Bahl et al., 2024). Auch beispielsweise das Ignorieren kleinerer Vorkommen in Datenmengen ist ein typischer Bias, der durch einen unangemessen programmierten Algorithmus entstehen kann (hierzu kann z. B. der Tokenbias während der Trainingsphase von KI gezählt werden, ein sog. Selection Bias (Wikipedia, 2025c)). Ein Beispiel dafür ist auch algorithmisches Shadowbanning (Mauro & Schellmann, 2023). Hier wird beispielsweise ein Social Media Account blockiert oder nicht mehr algorithmisch aus-

15 Schonmal in Google nach Babyspielzeug gesucht und dann in Amazon mit Windelangeboten überhäuft worden?

gewählt, ohne dass die Betreibenden davon informiert werden. Dies ist unter anderem im Zusammenhang mit Genderbias beobachtet worden.

Eine algorithmische Feedback Schleife führt zu einem Bias, wenn Ergebnisse algorithmischer Auswertung in die Daten einfließen, die dann zu einer neuen algorithmischen Auswertung genutzt werden. Bei KI ist das ganz besonders heikel: wenn eine KI zum Training alle „erreichbaren“ Daten benutzt, und Menschen zunehmend Informationen durch Nutzung von KI erstellen und veröffentlichen, dann wird die KI zunehmend mehr mit Daten trainiert, die durch die KI erstellt wurden. Diese Feedback-schleife führt zu einer graduellen und nicht mehr aufzuhaltenden Zerstörung einer reliablen Informationsbasis im Internet. Einige Forschende sprechen hier schon vom Anbrechen des post-faktischen Zeitalters (Zoglauer, 2025).

Zusammenfassend lässt sich sagen: während der Datenbias vielleicht noch entdeckt werden kann – beim Algorithmenbias ist das definitiv noch viel schwieriger. Spätestens beim Existieren von Feedbackschleifen ist dies nahezu unmöglich.

5. Bias in KI

Wie heikel das Thema ist, zeigt ein aktuelles Whitepaper des Bundesamtes für Sicherheit in der Informationstechnik: „Selbst KI-Systeme, die in bester Absicht sowie nach dem neuesten Stand der Technik erstellt werden, können von Bias betroffen sein“ (Ditz & Lichtmeß, 2025). Wie in den obigen Abschnitten dargestellt, kann man zwischen Datenbias und Algorithmenbias unterscheiden. Bei konnektionistischen, datengetriebenen Verfahren ist aber die Trennung zwischen Daten und Algorithmus außerordentlich schwer. In der öffentlichen Wahrnehmung findet man daher oft die Sammelbezeichnung „KI-System“, womit Daten und Algorithmen, grade bei Large Language Models, zusammengefasst werden. Der KI nutzende Mensch sieht nur noch das Ergebnis, das durch Zusammenwirken von Daten zum Training, Eingabedaten, Verlaufsdaten und zugrundeliegender Algorithmen entsteht. Eine Überprüfung hinsichtlich eines Bias misslingt meist. Wenn im Folgenden von „der KI“ die Rede ist, wird damit die Software gemeint, die nach dem Training eingesetzt werden kann (vergl. Abbildung 1). Diese KI hat den Bias bereits gelernt

und reproduziert ihn nun. Hinzu kommt ein eskalierender Bias durch den Einsatz und die Nutzung von KI. Wenn die KI als Trainingsmaterial unsaubere, nicht bewertete oder schlecht ausgewertete Daten bekommt, dann repliziert sie nicht nur die dort vorhandenen Biases, sie vermehrt sie sogar. Für gezielte Desinformation gilt dies natürlich noch viel mehr (Karpf, 2020)! Das Einsetzen einer Maschine, im Sinne eines Computers, löst die Assoziation besonderer Zuverlässigkeit aus – so dass eine algorithmisch verarbeitete Desinformation oder eine Sammlung von Daten mit Desinformationen zu deren weiterer Verbreitung führen kann, als dies über traditionelle Printmedien oder auch Internetmedien möglich war.

5. 1 Beispiel: Rassistischer- und Genderbias

Am deutlichsten sichtbar wird das bei rassistischem und genderbasiertem Bias. Bei beiden Biasformen handelt es sich um Annahmen über Menschengruppen, die stereotypisch verwendet werden. Ein Stereotyp ist psychologisch betrachtet eine generalisierte Annahme über eine Gruppe von Menschen, bei der dann die Variationen verloren gehen (Aronson et al., 2010). Zum Stereotyp können verschiedene Aspekte gehören, z.B. erwartetes Verhalten von dieser Gruppe, Aussehen oder auch angemessenes Verhalten gegenüber dieser Gruppe. Grundsätzlich helfen Stereotype dem Menschen, sich in der Welt zu orientieren, allerdings sind Stereotype leider selbst oft falsch, spiegeln eine verzerrte Realität wider (Bias) und sind außerordentlich hartnäckig gegenüber neuer Information. Beurteilungen von Personen, die einer bestimmten Gruppe angehören, nennt man explizite Stereotype (z. B. „die Ausländer“). Diese sind meist direkt sichtbar und vergleichsweise leicht aufzudecken. Schwieriger sind implizite Stereotype, die unbewusst existieren und trotzdem, ohne dass die Person sich darüber im Klaren ist, das Verhalten steuern. Als Beispiel: befragte Personen äußern, dass Männer und Frauen gleich gut geeignet sind, Arzt oder Ärztin zu werden. In dem eingangs erwähnten Beispiel wird aber die Rolle des „Arztes“ / „physician“ häufiger mit dem älteren Mann assoziiert.

Das Problem tritt im Zusammenhang mit KI zutage, wenn sich die Stereotypen in gesammelten Daten wiederfinden: Daten über Frauen, die überwiegend in „weiblich assoziierten“ Berufen arbeiten, Daten über Männer in überwiegend „männlich assoziierten“ Berufen sind dabei nur ein Beispiel (z. B. Spennemann & Oddone, 2025). Daten über das äu-

Bere Erscheinungsbild, über Verhaltensmuster, über Meinungen etc. liegen in Social Media in großen Mengen vor. All diese Daten beinhalten „nur“ Meinungen von Privatpersonen und sind überwiegend keine wissenschaftlichen oder wissenschaftlich aufbereiteten Daten. Wenn diese Daten dann zum Training von KI genutzt werden (wie bei aktuellen bekannten großen KI Modellen durchgehend der Fall), dann transferieren sie den Bias aus der Welt der Blogs und Social Media direkt in die KI (Bennett & Livingston, 2020). Und damit wieder zurück in die Welt. Ebenso halten Recommender-Systeme Menschen in der Blase ihrer gewohnten Accounts, was dann ebenfalls zu Bias führt (im Sinne von „alle denken wie ich“ – Anchoringbias und Confirmationbias – oder auch „ich will aussehen wie alle“ grade bei jungen Frauen). Daten mit historischer Verzerrung, beispielsweise über die Repräsentanz von bestimmten Geschlechtern in bestimmten Berufsgruppen¹⁶, gehen in die Berechnung mit ein und bestimmen aktuelle Berufsempfehlungen oder konkreter auch die Formulierungen nach denen Bewerbungen ausgewertet werden (O’Neil, 2017). Die Auswirkung von Bias in Daten auf die Geschlechterverteilung wurde u. a. von Criano-Perez umfassend dargestellt (Criado-Perez, 2020) und von Zweig hinsichtlich der Auswirkungen auf KI erläutert (Zweig, 2019). Noch schwieriger ist es mit dem Racial Bias, mit dem rassistischen Bias, der beispielsweise in der vorhersagenden Polizeiarbeit zutage tritt (z. B. in O’Neil, 2017, u. a. S. 87 ff, S. 97; Zweig, 2019, z. B. S. 208ff), aber auch bei der KI gestützten Bewertung von Bewerbungen, sei es auf Wohnungen oder auf Arbeitsstellen.

5. 2 Beispiel: Bias in der Polizeiarbeit

In verschiedenen Bereichen der Polizeiarbeit wird bereits seit ca. 2014 mit verschiedenen Formen der gezielten Analyse von Daten gearbeitet (z. B. vorhersagende Polizeiarbeit, engl. predictive policing, targeted policing; siehe hierzu auch den Beitrag von Egbert in diesem Band). Es ist also keine Bewegung, die durch den KI-Hype in 2024 ausgelöst wurde. Die zugrundeliegenden Systeme sind zwar datenintensiv, aber nicht notwendiger Weise im engeren Sinne „KI“.

16 Geschlechterverteilung in den Berufsgruppen war in den 1970er Jahren in Westen Deutschlands eine ganz andere als im Osten.

Wenn man mit der Annahme startet, dass eine Maschine es schafft, persönliche Bewertungen, Tagesform, kognitive Biases die Menschen haben (siehe Abschnitt 1.1) und andere emotional ausgerichtete, sehr persönliche Empfindlichkeiten (Barrett, 2023) aus der alltäglichen Arbeit auszuklammern, dann wäre es möglich, durch Nutzung einer Maschine (wie dem Computer) eine „bessere und gerechtere Welt“ zu bauen. Auf diese Weise könnte hypothetisch die Anzahl von falsch verurteilten Menschen reduziert werden und die Anzahl der Straftaten deutlich zurückgehen. Man müsste nur eine Maschine benutzen, die auf der Basis vorliegender Fakten eine neutrale Bewertung im Sinne eines juristischen Urteils vornehmen kann, oder die aufgrund vorliegender Daten eine Beurteilung der potenziellen Laufbahn eines Menschen ermöglicht.

Diese Annahme ist jedoch leider nicht richtig, denn sowohl die Daten als auch die Algorithmen, die die Arbeit einer Maschine im Sinne eines Computers ausmachen, entstammen menschlichen, Bias belasteten Gehirnen. Das klingt fürchterlich, entspricht aber leider den Erkenntnissen, die seit dem vermehrten und stark unterstützten Einsatz von digitalen Datenauswertungen im Allgemeinen und KI im Besonderen in allen erdenklichen Bereichen vorliegen. Bei der Analyse von Orten, die auf der Basis von digitalen Vorhersagen als straftatwahrscheinlich gelten, ist seit längerer Zeit bekannt, dass die Vorhersagen nicht zuverlässig genug sind (z. B. O’Neil, 2017; Peteranderl, 2025; Zweig, 2019) und nicht genügend Evidenz haben. So wäre grundsätzlich, gäbe es eine „faire“ Maschine, die Idee ganz hervorragend, wenn Strafverfolgungsbehörden und die Polizei in ihrer wichtigen Arbeit durch eine KI gestützte Datenanalyse unterstützt werden könnten. Insbesondere in den USA haben viele der aktuell auch in Deutschland untersuchten Systeme eine breite Erprobung gefunden (Beispiele in O’Neil, 2017) – leider nicht mit angemessener Erfolgsquote oder mit genügend viel Evidenz.

Daten, die zur Vorhersage von Straftaten genutzt werden können, die detaillierte Analyse der Daten von kriminalitätsbelasteten Orten, automatisierte Analyse von Videoaufzeichnungen zur Suizidverhinderung, Einsatz von Software wie Palantir (Steinberger, 2025a) und RADAR mögen mit einer guten Intention entwickelt worden sein, unterliegen aber leider alle nicht nur Datenschutzproblemen, Problemen der Datenverarbeitung hinsichtlich des EU-AI Acts (EU AI Act, 2024), Problemen hinsichtlich getroffener Vorannahmen (also Bias, vor allem Racial Bias) und weiteren

Problemen hinsichtlich datenethischer und algorithmenethischer Aspekte (siehe z. B. Peteranderl, 2025). Einige der Probleme in den Daten und in der algorithmischen Verarbeitung basieren auf verschiedenen Denkfehlern: der Versuch der Kategorisierung von menschlichem Verhalten, der Versuch der Vorhersage von menschlichem Verhalten auf der Basis von beobachtetem Verhalten, das Fehlen einer Grundwahrheit über menschliche Eigenschaften und die beim Beobachtenden bestehenden Biasformen sind nur wenige Aspekte davon (siehe u. a. O’Neil, 2017; Zweig, 2019; Vieth-Ditlmann, 2025; Steinberger, 2025). Auch eine Verzerrung in der zugrundeliegenden Zeitachse ist zu beobachten: so sind die Daten, die für digitale Analyseverfahren zugrunde liegen, oft nicht wissenschaftlich erhoben, nicht aktuell genug, nicht auf bestimmte Situationen eingestellt, mit einer Vermischung aus Korrelation und Kausalität belastet und von Stereotypen durchzogen. Ein Computerprogramm ist leider nur so gut, wie der Mensch, der es entwickelt hat. Und das gilt für Daten und Algorithmen gleichermaßen.

6. Fazit

Nach dem Mathematiker Thomas Bayes gibt es bei allen Dingen des Lebens eine bedingte Wahrscheinlichkeit, die von dem Auftreten bestimmter Bedingungen abhängt (Chalmers, 2007). Wann, wie und wo diese Bedingungen auftauchen, ist schwer zu fassen. Wissenschaftlich kann dies genutzt werden – aber die aktuellen konnektionistischen KIs sind, wenn man ihre Trainingsdaten analysiert, nicht Ergebnis einer wissenschaftlichen Entwicklung. Explainable AI oder auch vertrauenswürdige KI (Schork, 2024) könnte hier Abhilfe schaffen, aber aktuell ist dies noch nicht der Fall.

Es gibt genügend Gründe, im Zusammenhang von KI nach ethischen Rahmungen zu fragen: Fragen danach, ob es erlaubt ist, Datenquellen ungefragt zu nutzen, nur weil sie öffentlich sind; Fragen nach Einsatz der KI in therapeutischen oder pädagogischen Settings; Fragen nach Nutzung von KI zum Urteilen oder Beurteilen; Fragen danach, wem das Ergebnis einer KI-Anfrage „gehört“. Weizenbaum (1991; 2001) hat einige Aspekte davon bereits vorhergesehen, indem er klarstellt: die Daten, die in eine Maschine gefüttert werden, stammen immer von Menschen. Wenn diese Menschen den eigenen Bias nicht bemerken, den sie beim Erheben der

Daten sehr wahrscheinlich hatten, wird der Computer dies nicht besser machen. Die Art der Datenerhebung, die Gestaltung des Algorithmus, die Programmierung und auch der Einsatz der Software spiegeln immer den Menschen wider – im Guten wie im Schlechten, inklusive aller Biases. Daher warnte Weizenbaum schon früh davor, auf Ergebnisse zu vertrauen, die in einer Art und Weise im Computer berechnet worden sind, die durch den Menschen nicht mehr kleinschrittig nachvollzogen werden können (u. a. Weizenbaum, 1991; Weizenbaum et al., 2001). Dabei sollte nicht vergessen werden: auf einer Metaebene, die den Einsatz von Softwaresystemen generell betrachtet, entsteht ein Bias allein schon dadurch, dass viele Arbeiten am Computer die Lese- und Schreibfähigkeit als Grundvoraussetzung haben.

Datensammlung und Datenanalyse, sowie Datennutzung für das Training der KI bedeutet, dass der in den Daten vorhandene Bias die Grundlage für eine Software bildet, die diese Verzerrung dann in die nächste Stufe transportiert – der Bias replizieren sich in Form der Ausgaben, die das trainierte KI-System produzieren kann (siehe oben: Garbage-in, Garbage-out). Natürlich könnte man argumentieren, dass immerhin ein Computerprogramm nicht „aus dem Bauch heraus“ entscheidet, dass Computer immer nach Datenlage entscheiden und Sympathie und Antipathie keine Rolle für Computer spielen. Dies ist aber nicht ganz richtig – denn wenn „Bauchentscheidungen“ in den Daten reflektiert werden, dann spiegelt der Algorithmus eine ganz andere Realität vor und generiert künstliche Begründungsformen, die dann maschinell verstärkt wiedergegeben werden. Wenn zudem der Algorithmus dann nicht zureichend genau analysiert, dann schafft sich der Mensch diese Kausalitäten gerne selbst (auf Basis der Begründungen, die der Computer liefert) (siehe auch Zweig, 2019). Das kann zu einer gefährlichen Dynamik führen. Hinzu kommt, dass eine KI regelmäßig bereinigt werden müsste – also neu trainiert werden müsste, um sich an neue Umstände, neue Datenlage und neues Informationsmaterial anzupassen. Dies müsste vergleichsweise engmaschig passieren – was aber aufgrund des Aufwands und des immer schlechter werdenden Materials (i. S. v. verfügbare Information) immer schwieriger wird.

Welche Empfehlungen kann man also geben?

Asimov hat das in seinen Robotergesetzen schon vorweg genommen (siehe Wikipedia „Robotergesetze“). In Ergänzung kann gesagt werden:

- Es braucht eine ethische Rahmung für den Einsatz von KI und ML Verfahren
- Diese ethische Rahmung muss Daten und Algorithmen gesondert betrachten. Für die Daten muss sichergestellt sein, dass sie nach bestem Wissen und Gewissen so gut wie möglich biasfrei sind (dies kann durch Hinzuziehen möglichst heterogener Menschengruppen passieren), dass sie reliabel und valide sind, seriös erhoben wurden und auf das Anwendungsgebiet passen.
- Der Einsatz von KI kann höchstens unterstützend sein – eine alleinige Entscheidungsrolle kann einer KI nicht überantwortet werden, weil sie als Maschine auch keine direkte Verantwortung übernehmen kann.
- Es braucht den „Human in the Loop“, eine Gruppe von Menschen, die fachlich qualifiziert sind und die die Arbeit der Maschine und deren Einsatz überwachen, beginnend bei der Datenerhebung bis hin zur Generierung von Antworten und deren Weiternutzung. Diese Überwachung ist niemals „fertig“ oder abgeschlossen und muss permanent weitergeführt werden.

Zunächst geht es darum, dass ganz klar kommuniziert werden muss, dass aktuelle KIs in der Regel mit Material trainiert wurden, das nicht wissenschaftlich bereinigt ist und das Bias belastet ist. Für kritische Fälle könnte eine KI als Unterstützung genutzt werden, um mehr Fakten zu analysieren oder einen Überblick zu bekommen. Aber hier gilt zu berücksichtigen: „Wird ein Fairnessmaß gewählt und ein algorithmisches Entscheidungssystem danach optimiert, dann wird, aus der Perspektive des anderen Fairnessmaßes, immer eine Gruppe benachteiligt.“ (Zweig, 2019 S. 225).

Als Fazit bleibt folgendes festzuhalten: die Arbeit von Entscheidungstragenden wird also durch den Einsatz von KI nicht weniger – im Gegenteil! In einer „guten“ Welt würden Entscheidungstragende dann die Aussagen einer KI bis ins Detail überprüfen und sich daraus dann eine eigene Meinung bilden. Idealerweise entscheiden dann Menschen auch nicht als Einzelpersonen, sondern in Personenkonstellationen, die allein schon aufgrund ihrer Durchmischung einen Bias verhindern kann (z. B. verschiedene Ausbildungsstände, verschiedene Geschlechter, verschiedene Religionen, verschiedene kulturelle Hintergründe, Menschen mit und Menschen ohne Einschränkungen etc.).

Literatur

- Aronson, E., Wilson, T. D., Akert, R. M., & Aronson, E. (2010). *Sozialpsychologie* (6., aktualisierte Aufl., [Nachdr.]). Pearson Studium.
- Bader, S., & Kirste, T. (2025). Large Language Models: Technische Grundlagen. In A. Martens & C. H. Cap (Hrsg.), *Schreibende KI -- ein interdisziplinärer Diskurs* (S. 129–163). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-45839-3_6
- Bahl, U., Topaz, C., Obermüller, L., Goldstein, S., & Sneirson, M. (2024). Algorithms in Judges' Hands: Incarceration and Inequity in Broward County, Florida. *U.C.L.A. Law Review*, 246, 248–278.
- Barrett, L. F. (2023). *Wie Gefühle entstehen: Eine neue Sicht auf unsere Emotionen* (E. Liebl, Übers.; Deutsche Erstausgabe). Rowohlt Polaris.
- Bennett, W. L., & Livingston, S. (Hrsg.). (2020). *The Disinformation Age: Politics, Technology, and Disruptive Communication in the United States* (1. Aufl.). Cambridge University Press. <https://doi.org/10.1017/9781108914628>
- Bühner, M. (2021). *Einführung in die Test- und Fragebogenkonstruktion* (4., korrigierte und erweiterte Auflage). Pearson.
- Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research*, 81, 1–15.
- Chalmers, A. F. (2007). *Wege der Wissenschaft: Einführung in die Wissenschaftstheorie* (N. Bergemann & C. Altstötter-Gleich, Hrsg.; N. Bergemann & C. Altstötter-Gleich, Übers.; 6., verb. Aufl. 2007). Springer Berlin Heidelberg. <https://doi.org/10.1007/978-3-540-49491-1>
- Cho, I., Wesslen, R., Karduni, A., Santhanam, S., Shaikh, S., & Dou, W. (2017). The Anchoring Effect in Decision-Making with Visual Analytics. *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 116–126. <https://doi.org/10.1109/VAST.2017.8585665>
- CoPilot Microsoft. (2025, November 19). [Firmen Website]. <https://copilot.microsoft.com/>
- Criado-Perez, C. (with Singh, S.). (2020). *Unsichtbare Frauen: Wie eine von Daten beherrschte Welt die Hälfte der Bevölkerung ignoriert* (Deutsche Erstausgabe, 9. Auflage). btb-Verl.
- Das Perzeptron. (2025, November 12). [Wikipedia]. *Wikieintrag Suchbegriff Perzeptron*. <https://de.wikipedia.org/wiki/Perzeptron>
- DeepSeek Firmenseite. (2025, November 13). [Firmen Website]. <https://germany-deepseek.com>

- Ditz, J., & Lichtmeß, E. (2025). *Bias in der künstlichen Intelligenz* [Whitepaper des Bundesamtes für Sicherheit in der Informationstechnik]. Bundesamt für Sicherheit in der Informationstechnik. https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Whitepaper_Bias_KI.pdf?__blob=publicationFile&v=4
- Elsen, H. (2023). *Gender - Sprache - Stereotype: Geschlechtersensibilität in Alltag und Unterricht* (2. überarb. Aufl). utb GmbH. <https://doi.org/10.36198/9783838561806>
- Eppler, M., & Muntwiler, C. (2025, November 3). A Map of Cognitive Biases in Decision Making [Information Collection]. *Visual Literacy*. <https://bias-map-v1.web.app>
- EU AI Act, Pub. L. No. Document 32024R1689, 2024/1689 (2024). <https://eur-lex.europa.eu/eli/reg/2024/1689>
- Floridi, L. (Hrsg.). (2010). *The Cambridge handbook of information and computer ethics*. Cambridge University Press.
- Gemini Google Firmenseite. (2025, November 19). Firmenseite. <https://gemini.google.com/>
- Jaume-Palasi, L., & Spielkamp, M. (2017). Ethik und algorithmische Prozesse zur Entscheidungsfindung oder -vorbereitung (Arbeitspapier No. Nr. 4; AlgorithmWatch Arbeitspapiere). Algorithmwatch. https://algorithmwatch.org/de/wp-content/uploads/2017/06/AlgorithmWatch_Arbeitspapier_4_Ethik_und_Algorithmen.pdf
- Jermias, J. (2001). Cognitive dissonance and resistance to change: The influence of commitment confirmation and feedback on judgment usefulness of accounting systems. *Accounting, Organizations and Society*, 26(2), 141–160. [https://doi.org/10.1016/S0361-3682\(00\)00008-8](https://doi.org/10.1016/S0361-3682(00)00008-8)
- Karpf, D. (2020). How Digital Disinformation Turned Dangerous. In W. L. Bennett & S. Livingston (Hrsg.), *The Disinformation Age* (1. Aufl., S. 153–168). Cambridge University Press. <https://doi.org/10.1017/9781108914628.006>
- Koska, C. (2023). *Ethik der Algorithmen: Auf der Suche nach Zahlen und Werten* (Bd. 6). Springer Berlin Heidelberg. <https://doi.org/10.1007/978-3-662-66795-8>
- Krieghofer, G. (2017, September 8). Zitatforschung [Blos]. Zitatforschung. <https://falschzitate.blogspot.com/2017/09/ich-traue-keiner-statistik-die-ich.html>
- Martens, A., Bernauer, J., Illmann, T., & Seitz, A. (2001). „Docs 'n Drugs—The Virtual Polyclinic“ An Intelligent Tutoring System for Web-Based and Case-Oriented Training in Medicine. *Proceedings of the American Medical Informatics Association*, 433--437.

- Martens, A., & Cap, C. H. (Hrsg.). (2025). *Schreibende KI -- ein interdisziplinärer Diskurs: Perspektiven über den Sinn oder Unsinn von schreibender KI*. Springer Fachmedien Wiesbaden. <https://doi.org/10.1007/978-3-658-45839-3>
- Mauro, G., & Schellmann, H. (2023, Februar 8). There is no Standard: Investigation finds AI Algorithms Objectify Women's Bodies. *The Guardian*. <https://www.theguardian.com/technology/2023/feb/08/biased-ai-algorithms-racy-women-bodies>
- Monmonier, M. S. (2004). *Rhumb lines and map wars: A social history of the Mercator projection*. The University of Chicago Press.
- Moosbach, D. (2025, November 3). Bias [Enzyklopädie]. *Wortbedeutung*. <https://www.wortbedeutung.info/Bias/>
- Muldoon, J., Graham, M., & Cant, C. (2024). *Feeding the machine: The hidden human labour powering AI*. Canongate.
- Myers, D. G., & DeWall, C. N. (2023). *Psychologie*. Springer Berlin Heidelberg. <https://doi.org/10.1007/978-3-662-66765-1>
- Nilsson, N. J. (2010). *The quest for artificial intelligence: A history of ideas and achievements*. Cambridge University press.
- O'Neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy* (First paperback edition). B/D/W/Y Broadway Books.
- OpenAI, N. (2024). *Open AI Firmenseite [Firmen Website]*. <https://openai.com/>
- Peteranderl, S. (2025). *Automating Injustice (Algorithm Watch Berichte, S. 62) [Bericht]*. Algorithmwatch. https://algorithmwatch.org/de/wp-content/uploads/2025/03/AlgorithmWatch_Report-Predictive-Policing.pdf
- Piepenbrink, J. (Hrsg.). (2022). *Geschlechtergerechte Sprache*. Bundeszentrale für politische Bildung, Nr. 05-07/2022. <https://www.bpb.de/shop/zeitschriften/apuz/geschlechtergerechte-sprache-2022/>
- Porter, B., Lifschitz, V., & Van Harmelen, F. (2008). *Handbook of knowledge representation* (1st ed). Elsevier.
- Reinmann, G., & Mandl, H. (2000). *Individuelles Wissensmanagement: Strategien für den persönlichen Umgang mit Information und Wissen am Arbeitsplatz* (1. Aufl). Huber.
- Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 386–408. <https://doi.org/10.1037/h0042519>
- Russell, S. J., & Norvig, P. (2022). *Artificial intelligence: A modern approach* (Fourth edition, global edition). Pearson.
- Schork, S. (Hrsg.). (2024). *Vertrauen in Künstliche Intelligenz: Eine multiperspektivische Betrachtung*. Springer Vieweg.

- Shortliffe, E. H. (1976). *Computer-based Medical Consultations: MYCIN*. Elsevier.
- Spennemann, D. H., & Oddone, K. (2025). What do librarians look like? Stereotyping of a profession by generative Ai. *Journal of Librarianship and Information Science*, 09610006251357286. <https://doi.org/10.1177/09610006251357286>
- Steinberger, M. (2025a). *The Philosopher in the Valley Alex Karp, Palantir and the Rise of the Surveillance State*. Simon & Schuster UK.
- Steinberger, M. (2025b). *The Philosopher in the Valley Alex Karp, Palantir and the Rise of the Surveillance State*. Simon & Schuster UK.
- Steinmann, J.-P. (2023). Hochreligiös und migrantenfreundlich? Der nichtlineare Zusammenhang zwischen Religiosität und Fremdenfeindlichkeit in Deutschland. *Zeitschrift für Religion, Gesellschaft und Politik*, 7(1), 419–445. <https://doi.org/10.1007/s41682-023-00157-0>
- Stol, K.-J., Ralph, P., & Fitzgerald, B. (2016). Grounded theory in software engineering research: A critical review and guidelines. *Proceedings of the 38th International Conference on Software Engineering*, 120–131. <https://doi.org/10.1145/2884781.2884833>
- Vieth-Ditlmann, K. (2025, März 28). Explainer: Predictive Policing [Blog posts and essays]. *Algorithmwatch Explainer*. <https://algorithmwatch.org/de/algorithmische-polizeiarbeit-erklart/>
- Vigen, T. (2015a). Spurious Correlations [Blog]. *Correlation is not Causation*. <https://www.tylervigen.com/spurious-correlations>
- Vigen, T. (2015b). *Spurious correlations* (First edition). Hachette Books.
- Wason, P. C. (1968). Reasoning about a Rule. *Quarterly Journal of Experimental Psychology*, 20(3), 273–281. <https://doi.org/10.1080/14640746808400161>
- Wehrli, S., Hertweck, C., Amirian, M., Glüge, S., & Stadelmann, T. (2022). Bias, awareness, and ignorance in deep-learning-based face recognition. *AI and Ethics*, 2(3), 509–522. <https://doi.org/10.1007/s43681-021-00108-6>
- Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36–45. <https://doi.org/10.1145/365153.365168>
- Weizenbaum, J. (1991). *Kurs auf den Eisberg: Die Verantwortung des einzelnen und die Diktatur der Technik* (3. Aufl). Piper.
- Weizenbaum, J., Wendt, G., & Klug, F. (2001). *Computermacht und Gesellschaft: Freie Reden* (1. Aufl). Suhrkamp.
- Wikipedia. (2025a). *Garbage In Garbage Out*. Wikipedia. https://de.wikipedia.org/wiki/Garbage_In,_Garbage_Out

- Wikipedia. (2025b, November 3). Cognitive Bias [Enzyklopädie]. Wiki-eintrag Suchbegriff Cognitive Bias. https://en.wikipedia.org/wiki/Cognitive_bias
- Wikipedia. (2025c, November 24). Algorithmic Bias [Wikipedia]. Wiki-eintrag Suchbegriff Algorithmbias. https://en.wikipedia.org/wiki/Algorithmic_bias
- Wikipedia. (2025d, November 24). Filter Bubble [Wikipedia]. Wikieintrag Suchbegriff Filter Bubble. https://en.wikipedia.org/wiki/Filter_bubble
- Zoglauer, T. (2025). Konstruierte Wahrheiten: Wahrheit und Wissen im postfaktischen Zeitalter (2. Auflage). Springer Vieweg. <https://doi.org/10.1007/978-3-658-48313-5>
- Zweig, K. A. (2019). Ein Algorithmus hat kein Taktgefühl: Wo künstliche Intelligenz sich irrt, warum uns das betrifft und was wir dagegen tun können (1. Auflage, ungekürzte Ausgabe). Heyne Verlag.

Wissenschaftliche Grundlagenwerke für die weitere Vertiefung

Anm: O'Neil und Steinberger sind auch auf deutsch erhältlich

- (Martens & Cap, 2025)
(Zweig, 2019)
(O'Neil, 2017)
(Steinberger, 2025b)

Mediathek



Algorithmwatch:



ARD Audiothek: <https://www.ardaudiothek.de/episode/urn:ard:episode:d170ef37ba3b98fe/>



ARD Mediathek: <https://www.ardmediathek.de/video/planet-schule/darf-kuenstliche-intelligenz-alles-ki-und-ethik/wdr/Y3JpZDovL3N3ci5kZS9hZXgvbzE4NDM4NDQ>



Arte TV Videosammlung : <https://www.arte.tv/de/videos/RC-023353/kuenstliche-intelligenz/>



Prof. Dr. Alke Martens ist eine deutsche Universitätsprofessorin. Sie leitet den Lehrstuhl für Praktische Informatik und Didaktik der Informatik an der Universität Rostock. Ihre Forschungsgebiete sind Ethik der Informatik und der Künstlichen Intelligenz, Didaktik der Informatik und Einsatz von digitalen Systemen zum Lehren und Lernen. Sie ist zudem Autorin, Coach und Bildungsinfluencerin.

»Die KI entwickelt sich rasant zu einem zentralen Instrument der Kriminalprävention – doch die rechtlichen Regelungen halten mit dieser Dynamik bislang nur unzureichend Schritt.«

Prof. Dr. Sebastian Golla

KI in der Kriminalprävention – Rechtliche Herausforderungen von Innovation bis Anwendung

1. KI als Instrument zur Kriminalprävention

Künstliche Intelligenz (KI) hat sich in den letzten Jahren von einem technologischen Zukunftsversprechen zu einem konkreten Instrument der Sicherheitsbehörden entwickelt. Insbesondere im Bereich der Kriminalprävention befindet sich der Einsatz von KI aktuell in einer Phase intensiver Innovation. Das hier zugrunde gelegte Verständnis von Kriminalprävention umfasst jedoch nicht die Gesamtheit aller auf die Verhinderung von Straftaten gerichteten Bemühungen. Stattdessen fokussiert sich der Beitrag auf die selektive Prävention, also auf diejenigen Maßnahmen, die an bereits erkennbare Risikofaktoren für Straftaten anknüpfen (vgl. Bäcker, 2015, S. 7). Im Fokus stehen dabei sicherheitsbehördliche Maßnahmen wie zum Beispiel die KI-gestützte Auswertung großer Datenmengen in Ermittlungsverfahren.

Während in der zweiten Hälfte der 2010er-Jahre und zu Beginn der 2020er-Jahre zunächst die grundlegenden Möglichkeiten des KI-Einsatzes sichtbar wurden – wie fortgeschrittene Big-Data-Analysen von Textdokumenten (etwa der „Panama Papers“ durch das Bundeskriminalamt 2017 und 2018) und Anwendungen zur biometrischen Identifizierung – erleben wir heute eine Phase, in der vielfältige konkrete Anwendungen für die sicherheitsbehördliche Praxis vorbereitet werden (vgl. Farthofer, 2023, S. 294 ff.). Es ist zu erwarten, dass diese Anwendungen in den nächsten Jahren aus dem Kontext von Pilotprojekten und Forschungsvorhaben heraus zum Bestandteil der alltäglichen Arbeit insbesondere der Polizei werden.

Besonders im Fokus öffentlicher und politischer Diskussionen standen zuletzt Anwendungen, die unmittelbar sichtbar oder gesellschaftlich kontrovers sind. Hierzu zählen insbesondere Systeme zur Gesichtserkennung, die in der intelligenten Videoüberwachung eingesetzt werden, oder innovative Methoden zur Beweisgewinnung, wie KI-gestützte Lügendetektoren (siehe etwa Meister, 2025; Ogorek, 2024; Ibold, 2022). Solche Technologien haben ein hohes Diskussionspotenzial, weil sie nicht nur technische Fragen, sondern auch grundlegende rechtliche und ethische Probleme aufwerfen (Deutscher Ethikrat, 2023). Gleichwohl bilden die rund um derartige Szenarien geführten Diskussionen nur einen kleinen Teil der neuen Realität ab, der sich in der sicherheitsbehördlichen Praxis eröffnet.

Beachtung verdienen – nicht nur aus rechtlicher Sicht – auch die weniger spektakulären KI-Anwendungen in der Kriminalprävention, die alltägliche Arbeitsabläufe etwa gestützt auf die Funktionen großer Sprachmodelle erheblich zu erleichtern versprechen. Hierzu zählen beispielsweise Systeme zur Analyse von Akten oder Sortierung von Dokumenten. So verwies das Land Baden-Württemberg im Rahmen der Konferenz der Justizministerinnen und Justizminister von Bund und Ländern am 5. Juni 2025 im Zusammenhang mit einer von ihm federführend erarbeiteten Gemeinsamen Erklärung zum Einsatz von Künstlicher Intelligenz für eine beispielhafte KI-Anwendung nicht etwa auf eine komplexe Anwendung zur Videoüberwachung oder Prognose, sondern auf das System „Strukturierung mit KI (StruKI)“, das die Erstellung digitaler Aktenspiegel automatisiert unterstützen soll.

Aus institutioneller Perspektive richtet sich das Interesse dieses Beitrags vor allem auf die Polizei, die traditionell als zentraler Akteur im präventiven Sicherheitsbereich agiert und bei technologischen Innovationen meist eine Vorreiterrolle einnimmt. Durch diese Schwerpunktsetzung lassen sich die Herausforderungen der KI-Regulierung besonders plastisch darstellen, weil Polizeibehörden häufig als erste die praktischen Probleme und rechtlichen Unsicherheiten erleben.

Der Status quo der Regulierung von KI im Bereich der Kriminalprävention ist geprägt von einer Mischung aus etablierten rechtlichen Grundlagen, punktuellen neuen Regelungen und erheblichen Lücken, die sich im Angesicht der Praxis auftun. Der Beitrag untersucht die aktuelle Lage entlang von vier zentralen Problemfeldern:

1. Der eingriffsrechtlichen und datenschutzrechtlichen Einordnung von KI-Anwendungen zur Kriminalprävention.
2. Den neuen Herausforderungen, die sich aus der KI-Verordnung der Europäischen Union (KI-VO) ergeben, insbesondere im Spannungsfeld zwischen Produktsicherheitsrecht und Eingriffsrecht.
3. Den rechtlichen Unsicherheiten im Bereich der KI-Innovation, also bei Training und Test von Systemen, die eine Voraussetzung für spätere Anwendungen in der Kriminalprävention darstellen.
4. Der Frage, inwiefern das Recht der KI im Sicherheitsbereich sich selbst als „lernfähig“ erweist und der Dynamik aktueller technischer Entwicklungen gerecht werden kann (hierzu schon Golla, 2020).

2. Eingriffsrechtliche und datenschutzrechtliche Einordnung

Traditionell wird der Einsatz von KI im Sicherheitsbereich vor allem aus der Perspektive des Datenschutzes betrachtet. Immer dann, wenn personenbezogene Daten verarbeitet werden, greifen die Regelungen des europäischen Datenschutzrechts. Für den Bereich der Kriminalprävention gilt dabei primär die Datenschutzrichtlinie für Justiz und Inneres (JI-Richtlinie). Lediglich auf die unter 4 betrachtete Phase der KI-Innovation findet die Datenschutz-Grundverordnung (DS-GVO) Anwendung (Kühne, Golla & Schäfer, 2025, S. 275 f.). Ergänzt werden diese Regelungen durch die nationalen Datenschutzgesetze. Diese Fokussierung auf den Datenschutz ist nachvollziehbar: Der Einsatz von KI-Technologien ist fast immer mit komplexen und zu erheblichen Teilen intransparenten Datenverarbeitungen verbunden, die Risiken für die Rechte der Betroffenen bergen (Lenzen, 2024, S. 54 ff.).

Auch generell ist der Datenschutz in den heutigen Diskussionen über neue polizeiliche und sicherheitsbehördliche Befugnisse allgegenwärtig. Zwar spielt das europäische Datenschutzgrundrecht aus Artikel 8 EU-Grundrechtecharta eher im Anwendungsbereich der DS-GVO eine Rolle und ist in den Diskussionen um das noch weitgehend national geregelte Sicherheitsrecht weniger präsent (siehe aber Hofmann-Coombe, 2025). Ausgehend von der verfassungsrechtlichen Ebene hat sich jedoch eine besonders durch das Bundesverfassungsgericht immer weiter verfeinerte

Sonderdogmatik entwickelt, die moderne Ermittlungs- und Präventionsbefugnisse an dem aus Artikel 2 Absatz 1 in Verbindung mit Artikel 1 Absatz 1 Grundgesetz hergeleiteten Recht auf informationelle Selbstbestimmung misst (Eichberger, 2024, Artikel 2 Rn. 175). Dieses gewährleistet „die Befugnis des Einzelnen, grundsätzlich selbst über die Preisgabe und Verwendung seiner persönlichen Daten zu bestimmen“ (BVerfG, Urteil vom 15. Dezember 1983 – 1 BvR 209/83, 1 BvR 269/83, 1 BvR 362/83, 1 BvR 420/83, 1 BvR 440/83, 1 BvR 484/83, Rn. 147).

Dieses in seinen Grundlagen und seiner faktischen Stellung als „Supergrundrecht“ prinzipiell umstrittene Recht (hierzu lesenswert Linzbach, 2025) kennt auf der Ebene seines Schutzbereiches zunächst kaum Abstufungen zwischen verschiedenen Datenarten oder Kontexten. Jedes personenbezogene Datum gilt als potenziell schutzwürdig. Das berühmte Diktum im Volkszählungsurteil des Bundesverfassungsgerichts, dass es „kein belangloses Datum“ gebe, prägt bis heute das Verständnis des Grundrechts (BVerfG, Urteil vom 15. Dezember 1983 – 1 BvR 209/83, 1 BvR 269/83, 1 BvR 362/83, 1 BvR 420/83, 1 BvR 440/83, 1 BvR 484/83, Rn. 150). Differenzierungen erfolgen vor allem auf der Ebene der Eingriffsintensität, also anhand des Ausmaßes der Beeinträchtigung des Rechts auf informationelle Selbstbestimmung. Das Bundesverfassungsgericht hat eine Reihe von Kriterien entwickelt, um diese Intensität zu bestimmen. Dazu gehören die Streubreite der Datenverarbeitung, also die Zahl und Heterogenität der betroffenen Personen, die Komplexität der Verknüpfungen, die möglichen Folgen wie Diskriminierungsrisiken sowie die Transparenz oder Intransparenz der eingesetzten Verfahren (BVerfG, Urteil vom 27. Februar 2007 – 1 BvR 370/07, 1 BvR 595/07, Rn. 233 ff.).

Auch ohne dass spezifische Verfassungsrechtsprechung zum KI-Einsatz vorlag, führte eine Prüfung dieser Kriterien beinahe zwingend zu dem Ergebnis, dass KI-gestützte Verfahren zur Datenverarbeitung im Sicherheitsbereich tendenziell Grundrechtseingriffe von erhöhter Intensität begründen, wobei gilt: Je intensiver ein Grundrecht beeinträchtigt ist, desto spezifischer muss die Rechtsgrundlage sein. Diese Eingriffe sind daher auf spezielle rechtliche Regelungen zu stützen – die freilich weitgehend noch nicht existieren.

Selbstlernende Algorithmen, die auf nicht vollständig nachvollziehbare Weise Entscheidungen unterstützen, bergen besondere Risiken für Fehler und Verzerrungen (Borges, 2021, S. 34). Zudem können sie aufgrund

ihrer Fähigkeit zur Verknüpfung heterogener Datensätze tiefgreifende Profile von Personen erstellen. Sowohl intensive Entscheidungsunterstützung als auch Profilbildung kann potenziell in Konflikt mit der im Grundgesetz zentralen Menschenwürdegarantie geraten, die der umfassenden Durchleuchtung von Individuen, deren Objektifizierung und damit auch Datafizierung Grenzen setzt (Stephan, 2025, S. 688 ff.).

Im Februar 2023 hat das Bundesverfassungsgericht in seiner Entscheidung zu den Befugnissen zur automatisierten Datenanalyse in den Polizeigesetzen von Hessen und Hamburg ausdrücklich klargestellt, dass der KI-Einsatz bei Datenverarbeitungen zu einem besonderen Eingriffsgewicht führen kann.¹ In einer zentralen Passage der Entscheidung führte das Gericht aus:

[Der Mehrwert Künstlicher Intelligenz], zugleich aber auch ihre spezifischen Gefahren liegen darin, dass nicht nur von den einzelnen Polizistinnen und Polizisten aufgegriffene kriminologisch fundierte Muster Anwendung finden, sondern solche Muster automatisiert weiterentwickelt oder überhaupt erst generiert und dann in weiteren Analysestufen weiter verknüpft werden. Mittels einer automatisierten Anwendung könnten so über den Einsatz komplexer Algorithmen zum Ausweis von Beziehungen oder Zusammenhängen hinaus auch selbstständig weitere Aussagen im Sinne eines ‚predictive policing‘ getroffen werden. So könnten besonders weitgehende Informationen und Annahmen über eine Person erzeugt werden, deren Überprüfung spezifisch erschwert sein kann. Denn komplexe algorithmische Systeme könnten sich im Verlauf des maschinellen Lernprozesses immer mehr von der ursprünglichen menschlichen Programmierung lösen, und die maschinellen Lernprozesse und die Ergebnisse der Anwendung könnten immer schwerer nachzuvollziehen sein. Dann droht zugleich die staatliche Kontrolle über diese Anwendung verloren zu gehen. Wird Software privater Akteure oder anderer Staaten eingesetzt, besteht zudem eine Gefahr unbemerkter Manipulation oder des unbemerkten Zugriffs auf Daten durch Dritte. Eine spezifische Herausforderung besteht darüber hinaus darin, die Herausbildung und Verwendung diskriminierender Algorithmen zu

¹ Transparenzhinweis: In dem zugrunde liegenden Verfahren war ich als Verfahrensbevollmächtigter tätig und habe den Schriftsatz zur Befugnis im Hamburgischen Polizeirecht verfasst.

verhindern. Daher dürften selbstlernende Systeme in der Polizeiarbeit nur unter besonderen verfahrensrechtlichen Vorkehrungen zur Anwendung kommen, die trotz der eingeschränkten Nachvollziehbarkeit ein hinreichendes Schutzniveau sichern. (BVerfG, Urteil vom 16. Februar 2023 – 1 BvR 1547/19, 1 BvR 2634/20, Rn. 100)

Diese Entscheidung stellt die Sicherheitsgesetzgeber bezüglich der Regulierung von Befugnissen zum KI-Einsatz vor anspruchsvolle Aufgaben. Sie legt nahe, dass KI-gestützte Anwendungen zur Datenverarbeitung aufgrund ihrer besonderen Risiken einer erhöhten gesetzlichen Regeldichte bedürfen. Dafür sind spezielle Befugnisse mit erhöhten Eingriffsschwellen zu schaffen. Allerdings wird auch deutlich, dass flankierende Schutzmechanismen für den Einsatz von KI notwendig sein dürften, um Diskriminierung, Intransparenz und andere KI-spezifische Gefahren zu begrenzen.

Wie diese verfassungsrechtlichen Vorgaben in konkrete Befugnisnormen – also die rechtlichen Grundlagen, die den Staat unter bestimmten Voraussetzungen dazu ermächtigen, in die Rechte von Bürger:innen einzugreifen – umgesetzt werden können, lässt sich nur mit einiger Fantasie und Kreativität anhand der geplanten oder bereits existenten KI-Landschaft bestimmter Behörden beantworten (siehe hierzu Bäuerle et al., 2025). Lange Zeit existierten kaum spezifische Regelungen für KI-Anwendungen, da diese praktisch kaum genutzt wurden. Die Entscheidung des Bundesverfassungsgerichts zeigte unter anderem anhand der Regelung im Land Hamburg, dass die Schaffung einer komplexen Befugnis ohne die Existenz entsprechender Anwendungsfälle tendenziell keine gute Idee ist.

Dass das Bundesverfassungsgericht den Gesetzgebern keine konkrete Schablone für mögliche Befugnisse vorgegeben, sondern sich auf Ausführungen zu grundlegenden Leitplanken beschränkt hat, ist aus der Sicht des Gerichts nachvollziehbar. Schade ist trotzdem, dass die Ausführungen in der zitierten Entscheidung an einigen Stellen nicht über die schlagwortartige Benennung typischer Risiken (z. B. „diskriminierende Algorithmen“) von KI hinausgehen, deren rechtliche Handhabe auch nach Jahren der Diskussion noch nicht eindeutig ist. Letztlich bleibt so der Datenschutz Dreh- und Angelpunkt der verfassungsrechtlichen Einordnung von KI-Anwendungen.

Bedauerlich ist auch, dass das Bundesverfassungsgericht die Gelegenheit nicht genutzt hat, um über die objektiven Gewährleistungen wie Kontrollmechanismen, technische Sicherheitsstandards oder Transparenzpflichten zu entscheiden, die den KI-Einsatz bei Polizei und Sicherheitsbehörden begleiten müssen. Die Teile der dieser Entscheidung zugrunde liegenden Verfassungsbeschwerden, die auf diese Aspekte zielten, wurden mangels Beschwerdebefugnis als unzulässig verworfen – wobei sehr hohe Hürden angelegt wurden. Dabei sind gerade diese Schutzmechanismen zentral, um die spezifischen Risiken von KI-Einsätzen wirksam zu begrenzen.

Aufgrund der praktischen Entwicklung der KI-Landschaft lässt sich jedenfalls konstatieren, dass konkretere Befugnisse notwendig sind, um diese rechtlich abzubilden. Wie genau Bund und Länder ihre KI-Befugnisse ausgestalten werden, ist mit Spannung abzuwarten. Wünschenswert wäre, dass hierbei eher systematisch abgestimmte Regelungen entstehen als ein „Wildwuchs“. Lernfähig würde das Recht an dieser Stelle wohl dann, wenn verschiedene Modelle von KI-Klauseln sich in der Gesetzgebung durchsetzen würden und in ihrer praktischen Anwendung evaluiert werden könnten. Hier bietet der Föderalismus die Chance eines echten Ideenwettbewerbs.

In der aktuellen Phase ist entweder denkbar, dass in den nächsten Jahren eine Vielzahl von Rechtsnormen für eher konkrete Anwendungsszenarien geschaffen werden, oder – und dies ist aufgrund des Aufwands und der Halbwertszeit des ersten Ansatzes wahrscheinlicher – in einem breiteren Regulierungsansatz teils generalklauselartige Regelungen den KI-Einsatz bei Polizei und Sicherheitsbehörden auf sichere Beine stellen sollen.

Für die Herangehensweise zur Regelung konkret interessant ist die Differenzierung zwischen besonders eingriffsintensiven Szenarien – wie groß angelegten Profiling-Systemen oder biometrischer Überwachung – und alltäglichen, vergleichsweise moderat eingriffsintensiven Anwendungen wie dem oben erwähnten digitalen Aktenspiegel aus Baden-Württemberg (StruKI). Systeme, die lediglich interne Polizeidaten zusammenführen und für Recherchezwecke aufbereiten, greifen beispielsweise weniger tief in Grundrechte ein, werfen aber dennoch Fragen der Rechtmäßigkeit auf.

Hier könnten Generalklauseln mit sogenannten Regelbeispielen sinnvoll sein, also mit exemplarischen Aufzählungen typischer Fallgruppen, in denen der Einsatz solcher Systeme regelmäßig zulässig sein wird. Solche Regelbeispiele haben zwar eine indizielle, aber keine zwingende Wirkung: Ihr Vorliegen spräche grundsätzlich für die Rechtmäßigkeit des Systemeinsatzes, würde eine abweichende Beurteilung im Einzelfall jedoch nicht ausschließen. Umgekehrt könnte der Einsatz eines solchen Systems unter Umständen auch ohne ein solches Regelbeispiel rechtmäßig sein.

3. Neue Impulse durch die KI-Verordnung der EU

Mit der Verabschiedung der KI-Verordnung (KI-VO) der Europäischen Union beginnt ein neues Kapitel der KI-Regulierung. Erstmals existiert ein umfassendes europäisches Regelwerk, das sich ausdrücklich mit den besonderen Risiken und Anforderungen von KI-Systemen befasst. Auch wenn die Kompetenz der EU im Bereich der inneren Sicherheit begrenzt ist, wird die Verordnung erhebliche Auswirkungen auf die Kriminalprävention haben. Polizeibehörden, die KI-Systeme nutzen, gehören zu den Normadressaten und müssen ihre bestehenden und geplanten Anwendungen in das System der Risikoklassifizierung der Verordnung einordnen.

Das Herzstück der KI-VO ist die Einteilung von KI-Systemen in verschiedene Risikostufen. Die Verordnung unterscheidet zwischen nicht speziell regulierten Systemen mit geringem bis moderatem Risiko, sogenannten Hochrisikosystemen (Artikel 6 KI-VO) und Systemen mit inakzeptablem Risiko, welche grundsätzlich verboten sind (Artikel 5 KI-VO) (näher dazu Spiegel & Höving, 2025). Daneben sieht die KI-VO für bestimmte Systeme mit spezifischem Risiko – etwa Chatbots (Artikel 50 Absatz 1 KI-VO) oder Systeme zur Erstellung von Deepfakes (Artikel 50 Absatz 4 KI-VO) – besondere Kennzeichnungspflichten vor (siehe auch Bronner, 2024, S. 59).

Die Risikoeinstufung erfolgt gemäß Artikel 3 Nr. 2 KI-VO anhand einer „Kombination aus der Wahrscheinlichkeit des Auftretens eines Schadens und der Schwere dieses Schadens“ (näher dazu Bronner, 2024). Mit steigender Risikostufe nehmen dabei auch die regulatorischen Anforderungen zu. Während für Systeme mit geringem bis moderatem Risiko nur wenige allgemeine (Transparenz-)Pflichten gelten, werden für Hochrisikosysteme umfangreiche Anforderungen festgelegt.

Dazu gehören unter anderem Vorgaben zur Dokumentation, Transparenz, Datenqualität, menschlichen Aufsicht und zum Risikomanagement (vgl. Artikel 8–15 KI-VO). Praktisch entscheidend wird daher – auch im Kontext der Kriminalprävention – für die Anwendung der KI-VO regelmäßig die Frage sein, ob eine konkrete Anwendung als Hochrisikosystem einzustufen ist. Diese Abgrenzung wird voraussichtlich zu einem zentralen Streitpunkt zwischen Behörden, Hersteller:innen und Aufsichtsstellen werden (näher dazu Ebers & Streitbürger, 2024).

Ein KI-System ist grundsätzlich als hochriskant einzustufen, wenn es ein erhebliches Risiko der Beeinträchtigung der Gesundheit, Sicherheit oder Grundrechte natürlicher Personen birgt (vgl. Erwägungsgrund 46 Satz 5 KI-VO). Die KI-VO knüpft diese Einstufung an zwei mögliche Konstellationen: entweder ist ein System als hochriskant zu klassifizieren, wenn es als Sicherheitsbauteil in einem nach Anhang I KI-VO festgelegten Produkt verwendet wird, oder wenn es in einem in Anhang III KI-VO genannten Bereich eingesetzt wird – etwa in der Strafverfolgung oder dem Betrieb kritischer Infrastrukturen.

Der Begriff der Strafverfolgung, den Anhang III Nr. 6 KI-VO verwendet, ist dabei weit zu verstehen (im Sinne von „law enforcement“) und umfasst auch die straftatenbezogene Kriminalprävention. Rund um die Hochrisiko-Einstufung von Systemen, die zur Strafverfolgung in diesem Sinne genutzt werden sollen, sind allerdings noch viele Auslegungsfragen offen (Golla, 2025, S. 16 ff.).

Dazu zählt zum Beispiel die Frage, ob KI-gestützte Rechercsysteme von Polizei und Staatsanwaltschaften nach Anhang III Nr. 8 Buchstabe a) KI-VO als Hochrisiko-Systeme gelten. Anhang III Nr. 8 Buchstabe a) KI-VO erfasst KI-Systeme zur Tatsachenerforschung durch oder zur Unterstützung von Justizbehörden. Unklar ist dabei, ob Strafverfolgungsbehörden im engeren Sinne – also vor allem die Staatsanwaltschaften – unter den Begriff der Justizbehörde fallen (Golla, 2025, S. 19). Zwar ist die Strafverfolgung in Anhang III Nr. 8 KI-VO gerade nicht ausdrücklich geregelt, allerdings könnte der weite Begriff Staatsanwaltschaften, die typischerweise an der Schnittstelle von Judikative und Exekutive agieren, durchaus erfassen. Je nach Auslegung könnten entweder nahezu alle Rechercsysteme im Bereich der Strafverfolgung als hochriskant gelten oder nahezu keine. Diese Unsicherheit hat erhebliche praktische Folgen: Eine Hochrisiko-Einstufung löst umfangreiche Pflichten aus, die den Einsatz solcher Systeme erheblich erschweren oder verzögern könnten (vgl. Artikel 16 ff. KI-VO).

Neben der Klassifizierung enthält die KI-VO auch bestimmte Verbote für Systeme mit inakzeptablem Risiko, die jedoch mit weitreichenden Ausnahmen versehen sind. So ist die biometrische Echtzeit-Fernidentifizierung grundsätzlich untersagt, kann aber unter bestimmten Voraussetzungen für Zwecke der öffentlichen Sicherheit zugelassen werden (Artikel 5 Absatz 1 Buchstabe h) KI-VO). Ähnliches gilt für Systeme zur Vorhersage von Straftaten (Artikel 5 Absatz 1 Buchstabe d) KI-VO). Hier wird es in der Praxis darauf ankommen, wie die Mitgliedstaaten die Ausnahmeregelungen ausfüllen. Bislang gibt es hierzu jedoch kaum konkrete gesetzgeberische Aktivitäten, obwohl die Verordnung bereits im Spätsommer 2026 vollständig anwendbar sein wird.

Ein zentrales Problem besteht schließlich darin, dass die KI-VO und das klassische Eingriffsrecht – das regelt, wann und wie der Staat in die Rechte der Bürger:innen eingreifen darf – nicht isoliert nebeneinanderstehen. Vielmehr beeinflussen sich beide Regelungsebenen gegenseitig. Maßnahmen, die aufgrund einer Hochrisiko-Einstufung verpflichtend sind – etwa Dokumentationspflichten oder Transparenzanforderungen – wirken sich auch auf die Bewertung der Eingriffsintensität aus (Golla, 2025, S. 21). Sie können mildernd wirken, indem sie Risiken reduzieren, oder verschärfend, wenn sie neue Formen der Datenverarbeitung eröffnen. Damit hängt die Ausgestaltung polizeilicher Befugnisse künftig noch stärker von europarechtlichen Vorgaben ab.

Hinzu kommt, dass die KI-VO in vielerlei Hinsicht an Konzepte aus dem Datenschutzrecht anknüpft. Begriffe wie „Profiling“ oder „Grundrechtsfolgenabschätzung“ werden direkt übernommen, ohne dass ihre bestehenden Unschärfen beseitigt würden. Dies führt dazu, dass bekannte Auslegungsprobleme fortgeschrieben werden. Organisatorisch wird die Umsetzung der KI-VO eng mit den Datenschutzstrukturen verflochten sein. Datenschutzbeauftragte werden eine Schlüsselrolle bei der Überwachung und Beratung spielen, was Synergien, aber auch Konflikte erzeugen kann.

4. Rechtliche Hürden der KI-Innovation

Während die Schaffung rechtlicher Grundlagen für die Anwendung von KI-Systemen zu Zwecken der Kriminalprävention bereits vor ein paar Jahren Fahrt aufgenommen hat, blieb ein vorgelagerter Bereich bis vor Kurzem weitgehend unbeachtet (dazu aber bereits Leffer & Leicht, 2022): die rechtliche Zulässigkeit von Datenverarbeitungen für das Trainieren und Testen von KI-Systemen (siehe auch Kühne, Golla & Schäfer, 2025). Diese Phase der Innovation ist eine unverzichtbare Voraussetzung für jede spätere Anwendung – im Sicherheitsbereich und darüber hinaus. Ohne realitätsnahes Training können KI-Systeme keine verlässlichen Ergebnisse liefern. Gerade im Sicherheitsbereich ist es oft notwendig, mit echten Daten zu arbeiten, etwa mit Bildmaterial von Überwachungskameras oder Texten aus Ermittlungsakten. Synthetische Daten sind hier nur begrenzt einsetzbar (siehe etwa Kneuper & Jacobs, 2021).

Aus einer Perspektive, die Wert auf digitale Souveränität legt, wäre es wünschenswert, wenn staatliche Stellen die Entwicklung eigener KI-Systeme weitgehend selbst in die Hand nähmen (vgl. Kelber & Bortnikov, 2023). In der Praxis werden jedoch häufig externe Anbieter:innen – wie im Zusammenhang mit Datenanalyse-Plattformen prominent das US-amerikanische Unternehmen Palantir – beauftragt, weil den Behörden die notwendigen Ressourcen und Kompetenzen fehlen oder bestimmte technische Lösungen en vogue scheinen. Dies wirft zusätzliche Fragen der Kontrolle und Abhängigkeit auf. Doch selbst wenn Behörden eigene Entwicklungsprojekte durchführen, stehen sie vor erheblichen rechtlichen Unsicherheiten.

Die DS-GVO und die nationalen Datenschutzgesetze enthalten für den Forschungsbereich spezielle Regelungen, die Datenverarbeitungen unter bestimmten Voraussetzungen erlauben (exemplarisch: Artikel 9 Absatz 2 Buchstabe j) DS-GVO in Verbindung mit § 27 Bundesdatenschutzgesetz (BDSG)). Diese Regelungen beziehen sich auf die Verarbeitung besonderer Kategorien personenbezogener Daten wie beispielsweise Gesundheitsdaten, Angaben über Herkunft, religiöse oder weltanschauliche Überzeugungen. Zwar gelten ihre strengen Anforderungen nur bei Vorliegen derartiger Daten, allerdings finden sich einschlägige Informationen regelmäßig in Datensätzen, die für das Trainieren und Testen relevant sind – etwa in Strafak-

ten, die ärztliche Gutachten oder soziale Hintergrundinformationen enthalten. Damit dürfte Artikel 9 Absatz 2 Buchstabe j) DS-GVO in Verbindung mit § 27 BDSG oder entsprechenden Regelungen des Landesrechts regelmäßig zum Maßstab von Datenverarbeitungen werden, die im Rahmen von Forschungsaktivitäten stattfinden.

Für die Entwicklung von KI-Systemen reicht dieser Rahmen jedoch oft nicht aus. Forschung in diesem Sinne erfasst die methodische Gewinnung neuer Erkenntnisse (BVerfG, Urteil vom 29. Mai 1973 – 1 BvR 424/71, 1 BvR 325/72, Rn. 93), während die Entwicklung konkrete Produkte für den praktischen Einsatz hervorbringt. Jedenfalls ab einem gewissen technologischen Reifegrad kann das Trainieren und Testen von KI-Systemen nicht mehr unter den Forschungsbegriff gefasst werden und Datenverarbeitungen zu diesen Zwecken müssen auf andere Rechtsgrundlagen gestützt werden (näher Kühne, Golla & Schäfer, 2025).

Für diesen Bereich klafft hinsichtlich der für die Datenverarbeitung zur Verfügung stehenden Rechtsgrundlagen eine erhebliche Lücke. Während die Verarbeitung „einfacher“ (also nicht besonders sensibler) personenbezogener Daten noch auf die Aufgabenerfüllung der jeweiligen Stellen im Sinne von Artikel 6 Absatz 1 Buchstabe e) DS-GVO in Verbindung mit nationalem Recht gestützt werden könnte, steht für die Verarbeitung besonderer Kategorien personenbezogener Daten für die Entwicklungsphase keine Rechtsgrundlage zur Verfügung.

Das Problem der Notwendigkeit einer Rechtsgrundlage für derartige Daten ist auch in den wenigen Vorschriften oder Regelungsvorschlägen, die zu der Thematik existieren – vor allem § 37a des Hamburgischen Gesetzes über die Datenverarbeitung der Polizei, der die Datenverarbeitung für Training und Testung von lernenden IT-Systemen ausdrücklich regelt –, nicht berücksichtigt. Möglich wäre es, Artikel 9 Absatz 2 Buchstabe g) DS-GVO zu nutzen, der den Mitgliedstaaten Spielraum für eigene Regelungen für Datenverarbeitungen bei Bestehen eines erheblichen öffentlichen Interesses eröffnet. Auf dieser Grundlage könnte für die Entwicklung sicherheitsrelevanter KI-Systeme eine Rechtsgrundlage nach dem Modell von § 27 BDSG geschaffen werden.

Aktuell besteht jedoch erhebliche Rechtsunsicherheit. Entwicklungsprojekte bewegen sich häufig in einer Grauzone, in der weder klar ist, welche Daten verwendet werden dürfen, noch welche Schutzmaß-

nahmen erforderlich sind. Dies kann Innovation bremsen und mit dazu führen, dass Behörden vermehrt auf externe Anbieter:innen ausweichen, die unter Umständen anderen rechtlichen Standards folgen. Neben der Schaffung neuer Rechtsgrundlagen für die KI-Innovation besteht ein vielversprechender Ansatz darin, sogenannte „Sandboxes“ oder Reallabore zu schaffen, in denen technische Möglichkeiten gewissermaßen in isolierten Umgebungen erprobt werden können (näher Göbel & von Kruedener, 2024). Hier können Entwickler:innen, Datenschutzbehörden und andere Akteure gemeinsam daran arbeiten, sichere Rahmenbedingungen für die KI-Innovation zu schaffen.

5. Fazit

Die Regulierung von KI für Zwecke der Kriminalprävention steht genauso wie ihr praktischer Einsatz an einem Wendepunkt. Die KI entwickelt sich rasant zu einem zentralen Instrument der Kriminalprävention – doch die rechtlichen Regelungen halten mit dieser Dynamik bislang nur unzureichend Schritt. Auf verfassungsrechtlicher Ebene bestehen freilich längst Grundlagen, um insbesondere KI-gestützte Datenverarbeitungen einzuordnen. Das Bundesverfassungsgericht hat mit seiner Entscheidung von 2023 wichtige Leitplanken gesetzt, doch viele Detailfragen bleiben durch die Gesetzgeber in Bund und Ländern zu klären. Gleichzeitig ist mit der KI-Verordnung der Europäischen Union ein Regelwerk in Kraft getreten, das besonders mit seinem breiten Pflichtenkatalog für Hochrisikosysteme neue Maßstäbe setzt, aber auch komplexe Wechselwirkungen mit dem nationalen Eingriffsrecht erzeugt und einige wichtige Fragen offenlässt.

Entscheidend für den Erfolg der kommenden KI-Regulierung im Sicherheitsbereich wird unter anderem sein, ob das Recht sich ebenso wie die zu regulierenden Systeme als „lernfähig“ erweist. Dies bedeutet, dass der Regelungsrahmen nicht starr bleibt, sondern sich an neue technische und gesellschaftliche Gegebenheiten anpassen kann. Dies setzt eine kontinuierliche Beobachtung der Praxis, offene und fundierte Diskussionen sowie die Bereitschaft zu Experimenten voraus. Föderalismus kann hier ein Vorteil sein: Unterschiedliche Landespolizeigesetze können als Laboratorien für innovative Modelle dienen. In der Praxis ist dieser Wettbewerb bisher kaum zu beobachten. Das Land Hessen bildet hier eine seltene Ausnahme, indem es früh eigene Regelungen für automatisierte Datenanalysen geschaffen hat.

Ein lernfähiges Recht der Künstlichen Intelligenz muss schließlich nicht vorrangig auf spektakuläre Szenarien wie biometrische Massenüberwachung reagieren, sondern vor allem auch zumindest scheinbar simple Anwendungen wie etwa Systeme zur Erstellung digitaler Aktenspiegel berücksichtigen, die zunehmend den polizeilichen Arbeitsalltag prägen und damit die Grundlage für erfolgreiche Kriminalprävention bilden. In der konkreten gesetzgeberischen Umsetzung könnten neue general-klauselartige Befugnisse mit Regelbeispielen helfen, Rechtssicherheit zu schaffen. Ebenso wichtig ist der Blick auf die vorgelagerte Phase der KI-Innovation. Ohne klare Regelungen für Training und Tests bleibt letztlich auch die zurecht eingeforderte „digitale Souveränität“ deutscher Sicherheitsbehörden ein Lippenbekenntnis.

Im Ergebnis ist festzuhalten, dass der Einsatz von KI-Systemen in der Kriminalprävention – insbesondere in der Strafverfolgung sowie der straftatenbezogenen Gefahrenabwehr – ein erhebliches Potenzial aufweist und vielfältige Anwendungsmöglichkeiten eröffnet. Gleichzeitig entsteht jedoch ein komplexes Geflecht aus datenschutzrechtlichen, eingriffsrechtlichen und produktsicherheitsrechtlichen Fragestellungen, die durch eine Vielzahl europäischer und nationaler Regelungen geprägt sind. Die daraus resultierenden Unsicherheiten werden sich erst in den kommenden Jahren im Zusammenspiel von Gesetzgebung, Praxis, Wissenschaft und Rechtsprechung schrittweise klären lassen.

Literatur

- Bäcker, M. (2015). *Kriminalpräventionsrecht*. Mohr Siebeck. <https://doi.org/10.1628/978-3-16-153739-4>
- Borges, G. (2021). Potenziale von Künstlicher Intelligenz mit Blick auf das Datenschutzrecht. In Stiftung Datenschutz (Hrsg.), *Künstliche Intelligenz: Chancen und Risiken aus datenschutz- und anti-diskriminierungsrechtlicher Perspektive* (S. 7–63). <https://d-nb.info/1264956444/34>
- Bronner, P. (2024). Risikoklassifizierung, Risikobewertung und Risikominimierung nach der KI-Verordnung Eine erste Analyse des risiko-basierten Regulierungsansatzes der KI-VO. *Künstliche Intelligenz und Recht*, 1(2), 55–62.
- Deutscher Ethikrat (Hrsg.). (2023). *Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz. Stellungnahme*. Berlin. <https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-mensch-und-maschine.pdf>
- Ebers, M. & Streitböcher, C. (2024). Die Regulierung von Hochrisiko-KI-Systemen in der KI-Verordnung. *Recht Digital*, 4(9), 393–400.
- Eichberger, M. (2024). In P. M. Huber & A. Voßkuhle (Hrsg.), *Grundgesetz: GG* (8. Aufl.). C.H.Beck.
- Farthofer, H. (2023). Der Einsatz von Künstlicher Intelligenz in der Kriminalprävention. In T.-G. Rüdiger & P. S. Bayerl (Hrsg.), *Handbuch Cyberkriminalologie 1: Theorien und Methoden* (S. 293–316). Springer VS. https://doi.org/10.1007/978-3-658-35439-8_11
- Golla, S. (2020). Lernfähige Systeme, lernfähiges Polizeirecht: Regulierung von künstlicher Intelligenz am Beispiel von Videoüberwachung und Datenabgleich. *Kriminologisches Journal*, 52(2), 149–161. <https://doi.org/10.3262/KJ2002149>
- Golla, S. (2025). Auswirkungen der KI-Verordnung auf Strafverfolgung und Kriminalprävention. *Zeitschrift für Cyberstrafrecht*, 1(1), 16–22.
- Göbel, M. & von Krüedener, A. (2024). KI-Reallabore und Innovationsförderung in der KI-VO. *Gewerblicher Rechtsschutz und Urheberrecht in der Praxis*, 16(23), 755–758.
- Hofmann-Coombe, J. (2025). Die Anwendung der Unionsgrundrechte im Polizei- und Sicherheitsrecht nach der KI-VO. *Europarecht*, 60(4), 363–386.
- Ibold, V. (2022). Künstliche Intelligenz im Strafprozess – KI-basierte Lügendetektoren zur Tatsachenfeststellung? *Zeitschrift für die gesamte Strafrechtswissenschaft*, 134(2), 504–534. https://doi.org/10.1007/978-3-7089-2000-0_11

- org/10.1515/zstw-2022-0015
- Kelber, U. & Bortnikov, V. (2023). Digitale Souveränität von Sicherheitsbehörden und Nachrichtendiensten. *Neue Juristische Wochenschrift*, 78(28), 2000–2006.
- Kneuper, R. & Jacobs, S. (2021). Softwaretest mit Originaldaten: Eine Analyse aus Sicht des Datenschutzes. *Datenschutz und Datensicherheit*, 45(3), 163–167. <https://doi.org/10.1007/s11623-021-1411-8>
- Kühne, M., Golla, S. & Schäfer, J. (2025). KI-Innovation mit Echtdateien – Die rechtliche Zulässigkeit des Trainierens und Testens von KI-Systemen für Strafverfolgung und Gefahrenabwehr. *Zeitschrift für das Gesamte Sicherheitsrecht*, 8(6), 272-282.
- Leffer, L. & Leicht, M. (2022). Datenschutzrechtliche Herausforderungen beim Einsatz von Trainingsdaten für KI-Systeme. In E. Schweighofer, F. Kummer, A. Saarenpää, S. Eder, P. Hanke, J. Zanol & F. Schmutzner (Hrsg.), *Recht Digital – 25 Jahre IRIS: Tagungsband des 25. Internationalen Rechtsinformatik Symposiums IRIS 2022* (S. 89–98). Editions Weblaw. https://www.uni-saarland.de/fileadmin/upload/lehrstuhl/sorge/Paper-Downloads/IRIS22_Leffer-Leicht.pdf
- Lenzen, M. (2024). *Künstliche Intelligenz: Fakten, Chancen, Risiken* (2. Aufl.). C.H.Beck.
- Linzbach, K. M. (2025). *Additiver Grundrechtseingriff und informationelle Selbstbestimmung*. Mohr Siebeck. <https://doi.org/10.1628/978-3-16-164567-9>
- Meister, A. (2025, 23. Juli). *Gesichtserkennung und KI: Innenminister Dobrindt plant neues Sicherheitspaket*. Netzpolitik.org. <https://netzpolitik.org/2025/gesichtserkennung-und-ki-innenminister-dobrindt-plant-neues-sicherheitspaket/>
- Ogorek, M. (2024). Staatliche Gesichtserkennung durch biometrischen Abgleich mit Online-Daten. *Legal Tech – Zeitschrift für die digitale Anwendung*, 3(4), 274–280.
- Bäuerle, M.; Denker, K.; Geminn, C. & Schöndorf-Haubold, B. (2025). Big Data und KI bei der Polizei – Das Palantir-Urteil in der interdisziplinären Diskussion.
- Spiegel, U. & Höving, M. (2025). Die Klassifizierung von KI-Systemen nach der KI-VO: Vorschlag zur matrixbasierten Risikoklassifizierung. *Künstliche Intelligenz und Recht*, 2(6), 231–239.
- Stephan, D. A. (2025). Technischer Fortschritt – eine Gefahr für die Menschenwürde? *Die Öffentliche Verwaltung*, 78(16), 686–695.

Zur weiteren Vertiefung

- Bäuerle, M.; Denker, K.; Geminn, C. & Schöndorf-Haubold, B. (2025). Big Data und KI bei der Polizei – Das Palantir-Urteil in der interdisziplinären Diskussion.
- Golla, S. (2025). Auswirkungen der KI-Verordnung auf Strafverfolgung und Kriminalprävention. Zeitschrift für Cyberstrafrecht, 1(1), 16–22.
- Honekamp, W.; Kemme, S. & Struck, J. (2025): Auswirkungen von Künstlicher Intelligenz auf die zukünftige Polizeiarbeit – Technologische Potenziale, rechtliche Rahmenbedingungen, kriminologisch-sozialwissenschaftliche Erkenntnisse.

Mediathek



Verfassungsblog – Beitragsreihe The EU AI Act's Impact on Security Law



ZDFheute – Palantir: Wie die Polizei die Software nutzen will und warum die Kritik so groß ist



Prof. Dr. Sebastian Golla ist Juniorprofessor für Kriminologie, Strafrecht und Sicherheitsforschung im digitalen Zeitalter an der Ruhr-Universität Bochum. Er befasst sich in seiner Forschung vor allem mit strafrechtlichen Fragen der digitalen Transformation, der Cyberkriminologie und der Regulierung von Künstlicher Intelligenz.

»Es wird dringend Zeit, das technische Können gesetzlich einzurahmen und die technische Entwicklung der zukünftigen Sicherheitsarbeit zwar zu fördern, aber ebenso Rechtssicherheit und Schutz für Betroffene und Ausführende der Maßnahmen zu schaffen.«

Alina Borowy

Künstliche neuronale Netze in der Polizeiarbeit: Von Chancen, Risiken und einem fehlenden Rechtsrahmen

1. Einleitung

Seit sich zu Beginn der 1990er-Jahre das Internet in Deutschland zunehmend verbreitete, hat die damit einhergehende Digitalisierung nahezu alle Lebensbereiche erfasst. Kommunikation, Wirtschaft, Verwaltung und auch staatliche Sicherheitsaufgaben sind heute eng mit digitalen Prozessen verknüpft (Wörner, 2024, S. 619). Mit dieser Entwicklung geht eine stetig wachsende Menge an digitalen Daten einher: Standortdaten, Kommunikationsverläufe und Zahlungsströme entstehen in nie dagewesener Fülle und geben Aufschluss über die Nutzer:innen der verschiedenen Anwendungen (Brüning, 2024, S. 133). Dieses Phänomen der zunehmenden Datenintensität moderner Gesellschaften wird häufig unter dem Begriff „Big Data“ zusammengefasst. Gemeint ist das Zusammenwirken von drei Dimensionen: das schiere Volumen der erzeugten Daten, ihre enorme Vielfalt und die hohe Geschwindigkeit, mit der neue Daten generiert und verändert werden können (Fricke, 2020, S. 1-2; Leffer, 2025, S. 36-37). Längst wurde auch bei den Strafverfolgungs- und Sicherheitsbehörden erkannt, dass diese gesammelten Datenmengen wertvolle Informationen enthalten, die für eine gezielte Nutzung strukturiert, analysiert und ausgewertet werden müssen.

Kriminalistische Arbeit hat sich in den letzten Jahren stark gewandelt. Ermittlungen verlagern sich zunehmend in digitale Räume, in denen Beweise nicht mehr aus physischen Spuren, sondern aus digitalen Datensätzen bestehen (Brüning, 2024, S. 134; Wörner, 2024, S. 620). Straftaten werden vorbereitet, begangen und verschleiert, indem digitale Infrastruk-

turen genutzt oder manipuliert werden. Dabei entstehen gigantische Mengen an digitalen Spuren (Bitkom e.V., 2023, S. 8; Lang, 2023, S. 124; Wörner, 2024, S. 620). Aufgrund ihrer weitreichenden informationellen Befugnisse sind Polizei- und Staatsanwaltschaften ermächtigt, in erheblichem Umfang Daten von Beschuldigten, Verdächtigen und Zeug:innen zu erheben. Die gewonnenen Informationen werden dann in Datenbanken gespeichert und innerhalb behördlicher Informationssysteme für den Abruf bereitgehalten (Kugelman & Buchmann, 2024, S. 1). Kriminalistische Arbeit wird auf diese Weise immer mehr zu einer datenbasierten Analyse. Im Jahr 2021 hatte beispielsweise die Landespolizei in Niedersachsen 7,5 Petabyte Daten gespeichert. Dies entspricht etwa einer Menge von 150 Millionen gefüllten Aktenschränken (Bitkom e.V., 2023, S. 8; Wörner, 2024, S. 621). Es ist davon auszugehen, dass sich diese Entwicklung dank des fortschreitenden Anstiegs mobiler Endgeräte, Sensoren, smarterer Alltagsgegenstände sowie wachsender Kapazitäten auf Speichermedien fortsetzen wird (Bitkom e.V., 2023, S. 8).

„Big Data“ setzt die Strafverfolgungsbehörden unter großen Druck, Schritt zu halten. Unstrukturierte Datenmengen müssen mit knappen Personalressourcen gerichtsfest und schnellstmöglich ausgewertet werden (Ehringfeld, 2024, S. 10). Dabei ist eine händische Auswertung der aus heterogenen Datenquellen stammenden digitalen Spuren schon heute aufgrund ihrer enormen Menge nicht mehr zu realisieren (Brüning, 2024, S. 134-135; Fricke, 2020, S. 5). Oft fehlt es daneben zusätzlich an einer entsprechenden Ausbildung, wie digitale Daten überhaupt richtig zu lesen und auszuwerten sind (Wörner, 2024, S. 620-621). Ferner ist damit zu rechnen, dass die Zahl der neu ausgebildeten und verfügbaren Arbeitskräfte im kommenden Jahrzehnt weiter sinken wird (Bitkom e.V., 2023, S. 8).

Eine mögliche Lösung des „Big-Data-Problems“ könnte eine Datenanalyse mittels künstlicher Intelligenz, basierend auf der Technologie künstlicher neuronaler Netze, bieten. Künstliche Intelligenz ist in der Lage, große Datenmengen innerhalb kürzester Zeit zu strukturieren, nach relevanten Informationen zu filtern oder eigenständig bestimmte Muster und Zusammenhänge in Datensätzen zu erkennen (Ebers et al., 2020, S. 46 u. 62; Els, 2021, S. 6; Fricke, 2020, S. 5). Die Systeme können auf allen Ebenen der Massendatenverarbeitung eingesetzt werden, insbesondere bei spezialisierten und repetitiven Tätigkeiten (Bitkom e.V.,

2023, S. 9). Darüber hinaus birgt die Fähigkeit künstlicher neuronaler Netze, eigenständig Muster zu erkennen, auch bei der Auswertung von physischen Spuren große Potentiale (Bitkom e.V., 2023, S. 8; Kugelman & Buchmann, 2024, S. 2).

Diese neuen Möglichkeiten sind längst in den Fokus des Bundeskriminalamts gelangt. Im Rahmen des Programms „Polizei 2020“ soll der KI-basierten Datenanalyse heute und in den nächsten Jahren der Weg geebnet werden (Fricke, 2020, S. 8). Ferner kündigt die aktuelle Bundesregierung in ihrem Koalitionsvertrag an, Deutschland zur „KI-Nation“ zu entwickeln (Schulz & Evran, 2025, S. 391). So sieht der Vertrag vor, den Einsatz KI-gestützter Datenanalysetools und einer retrograden biometrischen Gesichtserkennung im öffentlichen Raum zu fördern und eine entsprechende Gesetzesgrundlage zu schaffen (Koalitionsvertrag von CDU/CSU und SPD, 2025, S. 82).

Vor diesem Hintergrund beschäftigt sich dieser Beitrag sowohl mit Chancen als auch Risiken der Nutzung künstlicher neuronaler Netze in der Polizeiarbeit und fragt nach einem rechtlichen Rahmen. Er soll einen Überblick über die technischen Potentiale künstlicher neuronaler Netze bieten und erste Pilotprojekte in Deutschland vorstellen. Daneben zeigt er aber auch technische Grenzen und Risiken für Betroffene der Maßnahmen auf.

2. Funktionsweise und Stärken künstlicher neuronaler Netze

Im Zentrum des Beitrags steht die derzeit vielversprechendste KI-Technik für die Mustererkennung: die künstlichen neuronalen Netze. Diese sind Rechenmodelle, die sich am Aufbau und der Arbeitsweise des menschlichen Gehirns orientieren. Sie bestehen aus einer Vielzahl miteinander verbundener künstlicher Nervenzellen (= Neuronen). Das wiederum sind Recheneinheiten, die Informationen empfangen, gewichten, verarbeiten und an andere mit ihnen vernetzte Neuronen weitergeben (Ebers et al., 2020, S. 52-53; Ertel, 2025, S. 288 ff.). Die einzelnen Neuronen sind in mehreren Schichten organisiert: einer Eingabeschicht, einer oder mehreren verdeckten Zwischenschichten sowie einer Ausgabeschicht (Kästner & Schomäcker, 2023, S. 359). Jede Schicht verändert die von benach-

barten Neuronen eingehenden Signale so, dass sie für die nächste Verarbeitungsebene nutzbar werden (Ebers et al., 2020, S. 53; Peters, 2023, S. 67 ff.). Während des Trainings eines neuronalen Netzes werden die einzelnen Verbindungen zwischen den Neuronen so lange angepasst, bis das System das gewünschte Ergebnis ausgibt. Es erhält hierfür eine große Menge an Trainingsdatensätzen, aus denen es lernt, typische Muster und Zusammenhänge abzuleiten. Das Netz verinnerlicht, welche Eingaben mit welchen Ergebnissen verknüpft sind, und passt seine internen Parameter fortlaufend an, um im Laufe des Trainings immer präzisere Ergebnisse zu erzeugen (Ebers et al., 2020, S. 54). Dieses Prinzip der Erfahrungsbildung aus Beispieldaten unterscheidet neuronale Netze grundlegend von herkömmlicher, regelbasierter Software, die auf fest vorgegebene Programmabläufe angewiesen ist (Rückert, 2023, S. 363).

Künstliche neuronale Netze erweisen sich als besonders leistungsfähig, da sie nicht auf Grundlage fest vorgegebener Regeln arbeiten, sondern eigenständig Strukturen und Beziehungen innerhalb von verschiedenen Datensätzen identifizieren. Sie sind zudem in der Lage, auch bei neuartigen, bislang unbekanntem Daten, Schlussfolgerungen zu ziehen, sofern diese in einem hinreichenden Zusammenhang zu bereits erkannten Mustern stehen (Ebers et al., 2020, S. 46). Ferner können durch den Einsatz maschinellen Lernens neue Erkenntnisse gewonnen werden. Aufgrund des offenen und datengetriebenen Ansatzes sind die Systeme grundsätzlich unvoreingenommen gegenüber neuen Verbindungen in den heterogenen Datensätzen, sodass unbekannte Zusammenhänge sichtbar werden, die einem Menschen gegebenenfalls aufgrund selektiver oder kognitiver Beschränkungen entgehen können (Ebers et al., 2020, S. 62; Ertel, 2025, S. 14). Hinzu kommt die Skalierbarkeit ihrer Analyseprozesse. Neuronale Netze können simultan Millionen von Datenpunkten verarbeiten. Sie können dabei heterogene Datenquellen übergreifend auswerten. Während menschliche Auswertung meist bereichsbezogen erfolgt, können neuronale Netze Bildmaterial, Kommunikationsdaten, Standortinformationen oder Finanzströme gleichzeitig analysieren (Fricke, 2020, S. 5). Dadurch entstehen mehrdimensionale Lagebilder, die Zusammenhänge sichtbar machen, die erst im Zusammenspiel verschiedener Informationsebenen erkennbar werden (Fricke, 2020, S. 5).

3. Einsatzmöglichkeiten und Erprobungen in der Praxis

In den USA und Großbritannien ist der Einsatz von verschiedensten KI-Systemen in der Kriminalprävention bereits allgegenwärtig geworden. Aber auch in Deutschland gibt es einen zunehmenden Einsatz KI-gestützter Maßnahmen, vor allem durch das Bundeskriminalamt und partiell auch bei den einzelnen Landespolizeien (Bitkom e.V., 2023, S. 8). In den vergangenen Jahren haben diverse Pilotprojekte auf Bundesgebiet stattgefunden, um verschiedene Systeme einem Praxistest zu unterziehen. Zudem sind einige Systeme auch bereits im alltäglichen Einsatz. Im Folgenden werden exemplarisch einige ausgewählte Einsatzmöglichkeiten und Pilotprojekte vorgestellt.

3.1 Automatisierte biometrische Gesichtserkennung

Ein zentrales Anwendungsfeld ist die automatisierte biometrische Gesichtserkennung, bei der neuronale Netze charakteristische Gesichtsmarkkmale analysieren und mit in Datenbanken gespeicherten Referenzdatensätzen abgleichen (Ebers et al., 2020, S. 976). Dies kann sowohl in Echtzeit, meist mittels Videoaufnahmen durch Überwachungskameras an öffentlichen Orten, als auch retrograd, mittels älterer gespeicherter Video- oder Bildaufnahmen, geschehen.

Das bekannteste Pilotprojekt zu dieser Technik war das Projekt „Sicherheitsbahnhof“, das durch die Bundespolizei und die Deutsche Bahn AG am Bahnhof Berlin Südkreuz in den Jahren 2017 und 2018 durchgeführt wurde. Im Teilprojekt I wurde automatische biometrische Gesichtserkennungstechnik zur Unterstützung polizeilicher Fahndungen erprobt (Ebers et al., 2020, S. 976). Zu diesem Zweck wurden am Bahnhof drei Videokameras installiert und mit einem von drei unterschiedlichen, auf neuronalen Netzen basierenden, Gesichtserkennungssystemen verschiedener Hersteller gekoppelt (Ebers et al., 2020, S. 976). Die Erprobung der Systeme lief über 12 Monate und bestand aus zwei aufeinander folgenden Einsatzphasen. Als Referenzdaten dienten Bildaufnahmen von 312 bzw. 201 freiwilligen Testpersonen, die während der Testphase einen Transponder bei sich trugen. Dieser ermöglichte eine Überprüfung, ob die Systeme die Teilnehmenden korrekt erkannten (Ebers et al., 2020, S. 976). Aus dem Gesamtbericht nach Abschluss des ersten Teilprojektes geht eine durchschnittliche Trefferrate von 84,9 Prozent der Gesamtheit der

Systeme für die erste Testphase und 91,2 Prozent für die zweite Testphase hervor. Die durchschnittliche Trefferrate bezeichnet hier das Verhältnis der Anzahl richtig erkannter Personen zur Gesamtanzahl der in der Videoaufzeichnung in einem bestimmten Zeitraum sichtbaren Personen aus der Referenzdatenbank, wobei bereits bei einem Treffer der drei Systeme ein Treffer des Gesamtsystems angenommen wurde (Ebers et al., 2020, S. 977). Diese durchschnittliche Trefferrate wurde unterschiedlich beurteilt. Während der Abschlussbericht des Projekts selbst von „ausgezeichneten“ Ergebnissen spricht, bemängeln Kritiker:innen, dass die Systeme mit einer Rate von 0,76 Prozent bzw. 0,34 Prozent „falsch-positiver“ Treffer zu hohe Fehleranfälligkeiten zeigten (Ebers et al., 2020, S. 977-978). Ein solcher „falsch-positiver“ Treffer bezeichnet den Fall, dass einer Person vom System irrtümlich die Identität einer anderen Person zugeordnet wird (Ebers et al., 2020, S. 977). Die im Abschlussbericht des Projekts errechnete Rate beschreibt dabei das Verhältnis zwischen der Anzahl fälschlich identifizierter Personen und der Gesamtzahl der Personen, die in der Referenzdatenbank nicht gespeichert sind, aber im betrachteten Zeitraum in der Videoaufzeichnung erschienen; dabei wurde ein Fall bereits dann als Falschtreffer gewertet, wenn auch nur eines der drei eingesetzten Systeme eine falsche Zuordnung vornahm (Ebers et al., 2020, S. 977). Im Realwelteinsatz bedeutet ein „falsch-positiver“ Treffer, dass eine unbeteiligte Person fälschlicherweise als eine polizeilich gesuchte Person eingeordnet wird. Sollte keine nachträgliche Überprüfung des Treffers erfolgen, kann so eine unbeteiligte Person unrechtmäßig Gegenstand polizeilicher Maßnahmen werden oder ihr wird beispielsweise der Zugang zu einem beschränkten Bereich oder die Einreise verwehrt. In einem Beispiel mit einer Trefferquote von 80 Prozent und fünf gesuchten Personen, die sich unter 1.000 Passanten befinden, können vom System vier der fünf gesuchten Personen richtig identifiziert werden. Gleichzeitig bewirkt eine Falschtrefferquote von einem Prozent, dass auch bei zehn nicht gesuchten Personen (ein Prozent von 995) eine Treffermeldung ausgelöst wird (Ebers et al., 2020, S. 978).

Daneben setzten verschiedene Landespolizeien in der Vergangenheit ein mobiles technisches System zur Identifizierung von straftatverdächtigen Personen ein. Im Rahmen von Observationsmaßnahmen kam ein unauffällig in einem Transporter verbautes, somit mobil einsetzbares, Gesichtserkennungssystem (Personen-Identifikations-System) zum Einsatz (Roggan, 2024, S. 715). Dieses ermöglichte einen automatisierten Abgleich

der erfassten Bildaufnahmen mit verschiedenen Datenbeständen, zu denen auch geeignete Fotos aus sozialen Medien gehören konnten. Entwickelt wurde diese Observationstechnik von der Polizeidirektion Görlitz in Kooperation mit der Firma OptoPrecision aus Bremen (Roggan, 2024, S. 715). Das System kam in Berlin, Brandenburg, Niedersachsen, Nordrhein-Westfalen und Baden-Württemberg zum Einsatz und konnte sowohl verdeckt als auch offen-stationär in fest installierten Kamerasäulen, wie in Görlitz und Zittau, genutzt werden (Roggan, 2024, S. 715).

Es arbeitete in Echtzeit oder retrograd. Nicht benötigte Daten, insbesondere solche ohne Trefferergebnis, wurden automatisch, unwiderruflich und spurenlos nach 96 Stunden gelöscht (Roggan, 2024, S. 715). Allerdings erweist sich der Einsatz dieses Systems auf rechtlicher Ebene als problematisch, da es an einer tragfähigen Ermächtigungsgrundlage fehlt, insbesondere da die Maßnahme überwiegend heimlich erfolgt, faktisch eine potenzielle Einbeziehung aller Bürger:innen in strafprozessuale Maßnahmen bewirkt, den allgemein zugänglichen Verkehrsraum zum Fahndungsgebiet macht und durch den Zugriff auf verschiedene Datenbanken, insbesondere Social-Media-Daten, zusätzliche datenschutzrechtliche Probleme aufwirft, sodass derzeit keine ausreichende gesetzlich geregelte Befugnis für den Einsatz des Personen-Identifikations-Systems besteht (Roggan, 2024, S. 717).

3. 2 Intelligente Videoüberwachung

Ähnlich wie eine automatische biometrische Gesichtserkennung arbeitet die intelligente Videoüberwachung. Das technische Grundprinzip konventioneller Videoüberwachung besteht zunächst darin, die von einer Kamera aufgezeichneten Bilder ungefiltert auf einen Monitor zu übertragen. Dort können sie von einer Kontrollperson betrachtet werden. Die intelligente Videoüberwachung erweitert diese Methode um eine Software, die die Bildaufnahmen automatisiert analysiert und einem Filterungsprozess unterzieht. Erkennt die Software dabei zuvor definierte Auffälligkeiten, wird eine Kontrollperson benachrichtigt, die die potenziell relevanten Sequenzen überprüft und gegebenenfalls polizeiliche Maßnahmen einleiten kann (Janitzski, 2021, S. 49). Zur Steigerung der Leistungsfähigkeit und Präzision kann das System um weitere Sensoren wie Mikrofone oder Temperaturmesser ergänzt werden (Janitzski, 2021, S. 50). Möglich ist der Einsatz in allen Bereichen, in denen schon die

konventionelle Videoüberwachung eingesetzt wird. In Betracht kommen öffentliche Veranstaltungen, kriminalitätsbelastete Orte oder sensible Regionen wie Flughäfen oder Bahnhöfe (Janitzski, 2021, S. 50). Der Einsatz einer intelligenten Videoüberwachung führt zu einer quantitativen Verbesserung, da die Überwachung größerer Bereiche bei Einsatz weniger Personen ermöglicht wird (Janitzski, 2021, S. 50).

In Baden-Württemberg setzte die Landespolizei im Rahmen eines Pilotprojekts intelligente Videoüberwachung bereits ein. Beginnend im Dezember 2018 wurde hierzu in Mannheim an drei Kriminalitätsschwerpunkten eine Situations- und Verhaltensmusteranalyse erprobt (Janitzski, 2021, S. 50). Der Algorithmus scannte die Bilder automatisiert und suchte nach bestimmten Verhaltensmustern, die auf mögliche Straftaten hinweisen könnten (Ebers et al., 2020, S. 978). Gesucht wurden Bewegungsmuster wie Schlagen, Rennen, Treten oder Hinfallen. Im Falle eines Treffers löste das System im Lagezentrum der Polizei einen Alarm aus. Polizeibeamt:innen überprüften sodann den Sachverhalt am Monitor und initiierten gegebenenfalls weitere Maßnahmen (Ebers et al., 2020, S. 978; Lang, 2023, S. 125). Das auf insgesamt fünf Jahre angelegte Projekt wurde in Kooperation mit dem Fraunhofer-Institut für Optik, Systemtechnik und Bildauswertung (IOSB) durchgeführt (Janitzski, 2021, S. 50). Ziel war es, die Straßenkriminalität im öffentlichen Raum zu reduzieren und zugleich praktikable Lösungen im Spannungsfeld zwischen Videoüberwachung und dem Schutz der Privatsphäre zu entwickeln (Janitzski, 2021, S. 50; Lang, 2023, S. 125). Nach Angaben des baden-württembergischen Ministeriums des Inneren, für Digitalisierung und Kommunen im Juli 2022 wurden mit der intelligenten Videoüberwachung bei der Polizei durchweg positive Erfahrungen gemacht. Die Videoüberwachungstechnik soll beispielsweise mehrfach dazu beigetragen haben, sich gerade entwickelnde körperliche Auseinandersetzungen zu verhindern. In einem Beispielsfall konnten die von den Videoüberwacher:innen entsendeten Interventionskräfte bereits nach 1:38 Minuten eingreifen, bevor es zu Handgreiflichkeiten kam (LT-Drs. 17/2833 S. 2-3).

Das Pilotprojekt wurde vor seinem Auslaufen im Jahr 2023 um weitere drei Jahre verlängert. Außerdem wurden im Dezember 2025 die zuletzt 46 mit KI-Systemen verbundenen Kameras in der Region Mannheim und Heidelberg auf 91 aufgestockt (Figaj, 2025). Die Polizei Mannheim geht nach eigener Aussage davon aus, dass das System auch nach Beendi-

gung der Pilotphase 2026 im Einsatz bleiben wird, um die Beamt:innen zu unterstützen. Wichtig soll dabei weiterhin sein, dass sich Datensicherheit und Eingriff die Waage halten (Figaj, 2025).

3.3 Predictive Policing

Im Bereich der Gefahrenabwehr gewinnt zunehmend das sogenannte „Predictive Policing“ an Bedeutung (siehe hierzu auch den Beitrag von Egbert in diesem Band). Grundsätzlich lässt sich hierbei zwischen ortsbezogenem und personenbezogenem Predictive Policing unterscheiden. In Deutschland kommt momentan nur die ortsbezogene Variante zur Anwendung (Fricke, 2020, S. 3). Zwar nutzen die einzelnen Polizeibehörden der Länder derzeit unterschiedliche Softwaresysteme, doch beruhen die zugrunde liegenden Verfahren auf denselben Prinzipien: Predictive-Policing-Anwendungen stützen sich im Rahmen des Data-Minings auf mathematisch-statistische Analysen, um Muster und Beziehungen aus historischen Daten abzuleiten und diese zur Prognose zukünftiger Ereignisse zu verwenden (Ebers et al., 2020, S. 965; Fricke, 2020, S. 3).

Das Programm PRECOBS, das ab 2013 zunächst in der Schweiz und anschließend auch in Bayern, Baden-Württemberg und Sachsen erprobt wurde, nutzte historische Falldaten, um sogenannte Near-Repeat-Muster bei Wohnungseinbrüchen zu identifizieren (Egbert et al., 2021, S. 192). Diesem Ansatz liegt die Annahme zugrunde, dass einer erfolgreichen und lohnenden Einbruchstat weitere Taten in räumlicher und zeitlicher Nähe durch dieselben Täter:innen nachfolgen werden (Ebers et al., 2020, S. 965). PRECOBS berechnet vor diesem theoretischen Hintergrund Risikowerte für Wohnungseinbrüche auf Grundlage von „Trigger“- und „Anti-Trigger“-Kriterien, die auf eine hohe Wahrscheinlichkeit von Folgetaten in räumlicher und zeitlicher Nähe hinweisen (Egbert et al., 2021, S. 193). Sofern durch die Software keine Anti-Trigger-Merkmale, also Hinweise auf nicht-professionelles Vorgehen wie etwa ein eingeschlagenes Fenster, erkannt werden und ausschließlich Triggerkriterien oder neutrale Merkmale vorliegen, generiert das System einen Alarm. Dieser weist auf ein erhöhtes Kriminalitätsrisiko in bestimmten Stadtgebieten und Zeiträumen hin (Egbert et al., 2021, S. 193). Der zugrunde liegende Ansatz von PRECOBS als polizeiliches Unterstützungstool besteht darin, der Polizei zu ermöglichen, laufende Einbruchserien frühzeitig zu identifizieren und gezielt in den jeweiligen Risikobereichen zu intervenieren, um weitere Einbrüche zu vermeiden (Egbert et al., 2021, S. 193).

Die daraus generierten Prognosen dienen vor allem präventiven Zwecken, insbesondere der räumlichen Schwerpunktsetzung polizeilicher Präsenz. Die Ergebnisse bestätigten grundsätzlich nutzbare Muster, deren Genauigkeit jedoch stark von der Qualität und Vollständigkeit der zugrunde liegenden Daten abhängig war (Egbert et al., 2021, S. 194 ff.). Entwickelt wurde das System vom Institut für musterbasierte Prognose-technik (IfmPt). In Berlin, Hessen, Niedersachsen und Nordrhein-Westfalen wurden eigene Entwicklungen der dortigen Landeskriminalämter verwendet (Ebers et al., 2020, S. 967).

Letztlich musste PRECOBS erst in Baden-Württemberg wegen schlechter Prognosen mangels Einbruchsdaten und dann auch in den anderen Bundesländern eingestellt werden, da es schließlich pandemiebedingt weniger Einbrüche gab, aus denen sich Prognosen ableiten ließen (Wörner, 2024, S. 626).

3.4 Datenanalyseplattformen

Die 20 deutschen Polizeibehörden arbeiten derzeit mit eigenen, oft inkompatiblen Anwendungen und Datenbanken, um auf separat gespeicherte und verwaltete Datenbestände zuzugreifen (Martini & Botta, 2025, S. 1033). Eine große Aufgabe der Gegenwart wird es sein, diese getrennt verwahrten Datenbestände zusammen zu führen und eine Analyse dieser heterogenen Informationsbestände zu ermöglichen. Dafür müssen beispielsweise Telekommunikationsdaten, Finanztransaktionen, Bewegungsprofile oder Ermittlungsakten über eine zentrale Datenplattform miteinander verknüpft werden, um strukturelle Muster sichtbar zu machen (vgl. Wissenschaftliche Dienste des Bundestags, 2024, S. 4 ff.). Präventiv dienen solche Analysen der Lagebilderstellung oder der frühzeitigen Identifikation sicherheitsrelevanter Entwicklungen. Repressiv erleichtern sie die Rekonstruktion komplexer Netzwerkstrukturen (Fricke, 2020, S. 7).

Einer der bekanntesten Anbieter dieser Analysesoftware ist Palantir Technologies, ein Unternehmen, das schon seit fast 20 Jahren Data-Mining-Dienstleistungen für Sicherheitsbehörden bereitstellt und auf dem Markt für Datenanalysesoftware aktuell führend ist (Martini & Botta, 2025, S. 1034). Auch in Deutschland kommt Palantirs Software „Gotham“ seit 2017 zum Einsatz. Gotham ist spezialisierte Software, die dafür gemacht

ist, große Mengen an Daten aus ganz unterschiedlichen Datenquellen, auf die die Polizei Zugriff hat, zusammenzuführen, auszuwerten und visuell darzustellen. Die Software kann mit den verschiedenen Datenbanken vernetzt werden und stellt dann eine zentrale Rechercheplattform dar. Benötigte Informationen müssen dann nicht mehr einzeln Datenbank für Datenbank gesucht werden, sondern können zentral über die entstandene Rechercheplattform abgerufen werden (Egbert et al., 2021, S. 210). Gotham und vergleichbare Systeme knüpfen damit zwar an frühere Formen der polizeilichen Computerisierung, etwa die Rasterfahndung, an, erweitern diese jedoch durch die datenbankübergreifende Analyse und algorithmische Unterstützung. Es entsteht ein dynamisches Zusammenspiel zwischen Software und Polizeikräften (Egbert et al., 2021, S. 210). Zunächst führte die Polizei Hessen das System „hessenDATA“ ein, gefolgt von der Polizei Nordrhein-Westfalen mit „DAR“ sowie zuletzt die bayerische Polizei mit der Plattform „VeRA“ (Bäuerle, 2025, S. 128; Egbert et al., 2021, S. 210). Grundlage dieser Anwendungen ist jeweils „Gotham“.

Forschungsergebnisse und Unternehmenskommunikation legen nahe, dass solche Plattformen funktional näher an Prognoseinstrumenten des Predictive Policing liegen als an klassischen Datenbankabfragen (Egbert et al., 2021, S. 210). Im Unterschied zu vollständig automatisierten Prognosesystemen wie PRECOBS generieren sie Vorhersagen jedoch nicht selbstständig, sondern ermöglichen es den Nutzer:innen, potenzielle Muster in großen Datensätzen durch eine ansprechende visuelle Aufbereitung der Daten eigenständig zu identifizieren oder zu konstruieren (Bäuerle, 2025, S. 129; Egbert et al., 2021, S. 210).

Inzwischen hat das Bundesverfassungsgericht mit Urteil vom 16.02.2023 sowohl die hessische Norm des Hessischen Polizeigesetzes als auch eine vorläufig von Hamburg ins hamburgische Polizeigesetz aufgenommene Norm als Ermächtigungsgrundlage für verfassungswidrig erklärt, die den Einsatz der automatisierten Datenanalyse gesetzlich verankern sollten (BVerfG NJW 2023, S. 1196; Bäuerle, 2025, S. 128; Brüning, 2024, S. 147). Das Gericht hat den Einsatz einer automatisierten Anwendung zur Datenanalyse nicht generell untersagt, es hat aber die Ausgestaltung der Normen in Hinblick auf Eingriffe in das Grundrecht der informationellen Selbstbestimmung gerügt. Die Normen wahren nicht den Grundsatz der Verhältnismäßigkeit zwischen Eingriffsintensität und öffentlichem Interesse und müssen daher hinsichtlich ihrer Eingriffsvoraussetzungen

nachgebessert werden (Bäuerle, 2025, S. 129). Ermöglicht eine automatisierte Datenanalyse oder -auswertung, wie in Hessen geschehen, einen schwerwiegenden Eingriff in die informationelle Selbstbestimmung, ist sie nur unter denselben engen Voraussetzungen zulässig wie besonders eingriffsintensive heimliche Überwachungsmaßnahmen: zum Schutz besonders gewichtiger Rechtsgüter und bei Vorliegen einer zumindest hinreichend konkretisierten Gefahr für diese, so das Gericht. Auf dieses Gefahrerfordernis kann nur verzichtet werden, wenn die zugelassenen Analyse- und Auswertungsmöglichkeiten durch Regelungen insbesondere zur Begrenzung von Art und Umfang der Daten und zur Beschränkung der Datenverarbeitungsmethoden normenklar und hinreichend bestimmt in der Sache so eng begrenzt sind, dass das Eingriffsgewicht erheblich gemindert ist (BVerfG NJW, 2023, S. 1196; Bäuerle, 2025, S. 130). Daneben steht gerade der Anbieter Palantir immer wieder in der Kritik. Palantirs Mitgründer Peter Thiel gilt als Vordenker der US-amerikanischen rechten Bewegung und tritt immer wieder mit demokratieskeptischen bis demokratiefeindlichen Ansichten in die Öffentlichkeit (Martini & Botta, 2025, S. 1034). Fraglich ist auch, welche Konsequenzen eine Datenübermittlung im Rahmen der Nutzung an Palantir hat und wer eventuell auch an die Daten gelangen könnte (Martini & Botta, 2025, S. 1035-1036).

Auf Bundesebene ist der Einsatz von Datenanalyseplattformen derzeit zwar noch nicht bekannt geworden, es liegt jedoch ein Gesetzentwurf vor, der sowohl für das Bundeskriminalamt als auch für die Bundespolizei eine entsprechende Ermächtigungsgrundlage schaffen soll (Bäuerle, 2025, S. 128).

4. Risiken und Bedenken der Anwendungen

Den soeben beschriebenen Vorteilen stehen technisch bedingte Fehleranfälligkeiten, systemimmanente Risiken und Belastungen der Betroffenen einer KI-gestützten Maßnahme gegenüber. Auch diese Aspekte bedürfen einer Betrachtung.

4.1 Blackbox-Effekt und Bias

Die in Abschnitt 1 dargestellten Eigenschaften künstlicher neuronaler Netze, insbesondere ihre Fähigkeit, komplexe Muster selbstständig zu erkennen, führen zu zwei grundlegenden Problembereichen: dem Blackbox-Effekt und dem Bias (siehe hierzu auch den Beitrag von Martens in diesem Band). Beide Phänomene sind systembedingt und beeinflussen maßgeblich die Zuverlässigkeit KI-gestützter Anwendungen.

4.1.1 Blackbox-Effekt

Der Blackbox-Effekt bezeichnet die strukturelle Intransparenz der internen Entscheidungswege neuronaler Netze. Ergebnisse entstehen aus einer Vielzahl miteinander verschalteter Berechnungsschritte, deren Gewichtungen und Zwischenergebnisse weder während der Nutzung für Anwendende sichtbar sind noch im Nachhinein rekonstruiert werden können (Rückert, 2023, S. 367). Kein Mensch, nicht einmal die Entwickler:innen des Systems selbst, kann erklären, anhand welcher konkreten Merkmale der Eingabedaten ein bestimmtes Ergebnis zustande gekommen ist (Peters, 2023, S. 77 ff.; Rückert, 2023, S. 366-367). Ferner ist es charakteristisch für tiefe neuronale Netze, dass sie zwar auf vorgegebenen Eingangsdaten basieren, die für die Mustererkennung relevante Merkmalsrepräsentation jedoch während des Trainings selbstständig weiterentwickeln und verfeinern. Sie lernen also selbstständig weiter. Das erweitert sozusagen die Blackbox auf die verwendeten Merkmale und erschwert damit die Erklärbarkeit. Auch wenn der Forschungsbereich der Explainable AI versucht, die Blackbox zu öffnen, indem Technologien geschaffen werden sollen, die den Entscheidungsweg in neuronalen Netzen sichtbar machen können, ist dies bisher noch nicht gelungen (Ebers et al., 2020, S. 66 u. 67; Kästner & Schomäcker, 2023, S. 565).

4.1.2 Bias

Der Begriff des „Bias“ beschreibt systematische Verzerrungen, die entstehen, wenn die Trainingsdaten eines Modells unausgewogen oder unvollständig sind. Da neuronale Netze aus Beispielen lernen, übernehmen sie nicht nur Muster, sondern auch Zufälligkeiten oder Schief lagen ihrer Datenbasis, die sich dann in der Einsatzphase zeigen (Ebers et al., 2020, S. 58; Ibold, 2024, S. 138). Enthalten die Daten einen Bias, sind sie somit

nicht objektiv und neutral, sondern spiegeln Vorurteile wider, die der Algorithmus selbst aber nicht erkennen kann. Er besitzt kein Verständnis von kausalen Zusammenhängen, sondern stellt bloße Korrelationen zwischen Daten her (Rückert, 2023, S. 365).

Wird z. B. ein Modell überwiegend mit Videoaufnahmen sich gleichförmig bewegender Menschen trainiert, kann es möglicherweise Menschen, die sich etwa aufgrund einer Gehbehinderung außerhalb des als „unauffällig“ gelernten Bewegungsmusters bewegen, als „auffälliges Verhalten“ markieren (Janitzski, 2021, S. 51). Ein weiteres Beispiel betrifft netzwerkanalytische Auswertungen großer Datenbestände: Wenn historische Daten bestimmte Orte, Kontaktbeziehungen oder Kommunikationskanäle überrepräsentieren, werden diese im Modell als besonders relevant gewichtet. Das System kann dann etwa Verbindungen zwischen Personen oder Ereignissen markieren, die nur deshalb als wichtig erscheinen, weil sie historisch häufiger erfasst wurden, nicht, weil sie im konkreten Fall besondere Bedeutung besitzen. Wird ein Gebiet aufgrund historischer Daten häufiger als „auffällig“ eingestuft, konzentriert sich dort polizeiliche Präsenz. Dadurch werden in diesem Gebiet mehr Vorfälle registriert, was wiederum das Modell in seiner ursprünglichen Prognose bestätigt. Die Verzerrung entsteht damit nicht allein aus der Trainingsdatenlage, sondern kann sich im laufenden Einsatz fortsetzen. Solche Rückkopplungsmechanismen können langfristig dazu führen, dass bestimmte Räume oder Verhaltensweisen überproportional häufig als sicherheitsrelevant wahrgenommen werden (Fricke, 2020, S. 7).

4.2 Einschränkungen für Betroffene KI-gestützter Maßnahmen

Aus den genannten Risiken, aber auch aus den systemeigenen Funktionen und Arbeitsweisen resultieren nicht unerhebliche Einschränkungen für Betroffene der KI-gestützten Maßnahmen, die hier überblickartig Platz finden sollen. Dabei sei auf mögliche Verstöße und Kollisionen mit dem Datenschutzrecht und dem Grundrecht auf informationelle Selbstbestimmung aufgrund von KI-basierter Erhebung und Verarbeitung von Daten der Betroffenen nur am Rande verwiesen, da eine Behandlung dieser den Rahmen dieses Beitrags überschreiten würde (siehe hierzu näher z. B. Ebers et al., 2020, S. 980 ff.). Die nachfolgende Darstellung muss sich daher auf einen allgemeinen Überblick beschränken. Es ist darauf zu verweisen, dass die Beurteilung der Eingriffsintensität grundsätz-

lich anhand einer genauen Einzelbetrachtung der geplanten Maßnahme hinsichtlich ihrer Funktionsweise, ihrer Einsatznutzung und benötigter Daten geschehen sollte.

Vorab ist festzuhalten, dass sehr unterschiedliche Personen von der Maßnahme betroffen sein können. Es kann sowohl der/die Adressat:in einer Maßnahme z. B. die Person, nach der mittels Gesichtserkennung gefahndet wird, betroffen sein, aber auch unbeteiligte Dritte, deren Gesichter während der Suche ebenfalls erfasst und gescannt werden (Bitkom e.V., 2023, S. 9). Ebenso beziehen auch Datenanalyseplattformen umfassende Datenmengen ein, sodass beispielsweise auch Daten von Zeug:innen, die eine Aussage gemacht haben, von den Systemen durchsucht werden können.

Grundsätzlich bietet die Gesamtheit aus intelligenter Videoüberwachung und automatischer biometrischer Gesichtserkennung staatlichen Sicherheitsbehörden eine neue Möglichkeit, den öffentlichen Raum lückenloser zu überwachen. Für Betroffene ergeben sich daraus neue Formen staatlicher Überwachung und Bewertung. Systeme zur Gesichtsidifikation, Bewegungsanalyse oder Verhaltensklassifikation ermöglichen eine stetige Erfassung großer Personengruppen häufig ohne konkreten Anlass. Ferner ist festzustellen, dass es für Auswirkungen auf den einzelnen Betroffenen einen Unterschied machen kann, ob eine Maßnahme heimlich oder offen stattfindet. Werden Personen, wie z. B. im Rahmen des oben beschriebenen Personen-Identifikations-Systems, ohne ihr Wissen erfasst und überprüft, bleibt ihnen keine Möglichkeit, in ihre Erfassung einzuwilligen bzw. ihre Einwilligung zu verweigern oder den Kontrollbereich gegebenenfalls zu meiden, um ihre Erfassung zu verhindern.

Handelt es sich dagegen um eine offene Maßnahme im öffentlichen Raum, kann es zu Einschüchterungseffekten kommen. Diese sogenannten „Chilling-Effects“ sind die Angst vor möglichen Folgen nach einer Identifikation, die zu selbstbeschränkendem Handeln führen kann (Els, 2021, S. 618). Daraus folgen gegebenenfalls Verhaltensanpassungen, die etwa zum Verzicht auf öffentliche Meinungsäußerungen oder auf Teilnahme an einer öffentlichen Versammlung führen (Coombe, 2024, S. 264; Janitzki, 2024, S. 51). Aber auch aus einer heimlich erfolgten Erfassung, von der die Betroffenen nichts gewusst haben, kann eine Unsicherheit resultieren, ob eine Erfassung und Verarbeitung der eigenen Daten stattgefunden hat und ob retrograd noch mit Folgemaßnahmen

zu rechnen ist (Coombe, 2024, S. 265). Es genügt, wenn bekannt ist, dass Maßnahmen verdeckt durchgeführt werden könnten.

Insoweit ermöglicht die Nutzung intelligenter Systeme eine Beobachtung mit neuer Intensität. Im Unterschied zu rein menschlicher Beobachtung können solche Systeme dauerhaft, parallel, in großem Umfang und ohne Ermüdung auswerten. Die Wahrnehmung, dass eine Überwachung nicht punktuell erfolgt, sondern potenziell lückenlos, kontinuierlich und mit hoher Detailtiefe, intensiviert das Gefühl ständiger Beobachtbarkeit erheblich (Janitzski, 2024, S. 51).

Grundrechtsrelevante Auswirkungen können ebenfalls aus einem Bias folgen, wenn bestimmte Gruppen häufiger als auffällig oder risikobehaftet eingestuft werden. Wenn Trainingsdaten historisch bedingte Ungleichverteilungen enthalten, spiegeln Systeme diese Muster wider. Dies kann zu ungleichen Kontrolldichten führen, etwa wenn KI-gestützte Videoanalyse bestimmte Bewegungsformen, Kleidungsstypen oder Aufenthaltsmuster überproportional markiert, weil sie im Trainingsmaterial häufiger mit sicherheitsrelevanten Situationen verknüpft waren (Fricke, 2020, S. 3).

Eine Person mit bestimmten äußeren Merkmalen oder aus bestimmten Stadtvierteln kann häufiger in automatische Treffermeldungen geraten, ohne dass ein konkreter Anlass besteht. Auch wenn die Maßnahmen formal anlassunabhängig sind, kann der systemische Bias faktisch zu einer Ungleichbehandlung führen, z. B. zu vermehrten polizeilichen Kontrollen dieser Person (Janitzski, 2024, S. 51). Abschließend kann der Umstand, dass maßgebliche Erkenntnisse für die Einleitung eines Ermittlungsverfahrens oder überhaupt Beweismittel mit Hilfe der Blackbox „KI“ gewonnen wurden, große Auswirkungen auf das Recht auf ein faires Verfahren für Beschuldigte eines Strafverfahrens haben. Wenn KI-basierte Verfahren Beweise generieren, deren Entstehung aufgrund des Blackbox-Effekts nicht nachvollziehbar ist, erschwert dies eine effektive Verteidigung für Beschuldigte und ihre Verteidiger:innen. Eine nachträgliche Kontrolle des Beweisgewinnungsverfahrens ist aufgrund der Intransparenz der Systeme derzeit technisch nicht realisierbar. Dies kann die Waffengleichheit im Strafprozess berühren und lässt dem oder der Beschuldigten kaum Raum, Beweise in Zweifel zu ziehen (Aden et al., 2022, S. 62-63). Seine oder ihre Möglichkeit, Beweise zu hinterfragen, hängt davon ab, dass deren Entstehung und Qualität transparent sind.

5. Ausblick

Wie sich im Verlauf des Beitrags gezeigt hat, können künstliche neuronale Netze einiges tun, um die Polizei bei ihren Aufgaben der Gefahrenabwehr und strafrechtlichen Ermittlungen zu unterstützen, mithin einen positiven Beitrag für die Kriminalprävention leisten. Allerdings arbeiten diese Systeme auch nicht fehlerlos, ihre Anwendung birgt einige Risiken und sie können starke Einschränkungen für einen großen Kreis von Betroffenen der Maßnahmen mit sich bringen, die nicht vernachlässigt werden dürfen. Es bleibt also die Frage, wie das Potential der künstlichen neuronalen Netze für die Kriminalprävention nutzbar gemacht werden kann und woran es noch fehlt. Zum einen ist darauf zu verweisen, dass die verschiedenen vorgestellten Maßnahmen unterschiedlich weit in ihrer technischen Entwicklung sind. Zum Teil bedarf es noch weiterer Pilotversuche und technischer Verbesserungen, um z. B. Trefferquoten zu erhöhen oder ein breiteres Einsatzfeld zu erschließen. Beispielhaft können die Systeme zur biometrischen Gesichtserkennung genannt werden, die unter Testbedingungen im Labor andere Trefferraten aufweisen als in der Realweltsanwendung (Rückert, 2023, S. 366). Sie sind immer noch anfällig für schlechte Lichtverhältnisse oder Adversarial Attacks, die versuchen, die Systeme gezielt z. B. durch das Tragen von Mützen zu täuschen (Rückert, 2023, S. 366).

Im Bereich des Predictive Policing fehlt es häufig noch an ausreichend geeigneten Trainingsdatensätzen. Strafverfolgung befasst sich mit typisiertem Unrecht. Das sind atypische, nicht regelmäßige Handlungen. Typisiert ist nur die abstrakt im Gesetz festgehaltene Form der Normübertretung, nicht jedoch die konkrete Vorgehensweise. Gerade darin, dass Täter:innen das Unvorhersehbare tun, liegt regelmäßig der Taterfolg (Wörner, 2024, S. 621).

Zum anderen wird es zukünftig wichtig sein, den Nutzen der einzelnen Systeme und die aus der Nutzung resultierenden Einschränkungen für Betroffene in ein richtiges Verhältnis zueinander zu setzen. Die Maßnahmen müssen verhältnismäßig sein. Dies kann nur durch die Schaffung entsprechender Ermächtigungsgrundlagen für den Einsatz der verschiedenen Systeme geschehen. Diese müssen den rechtlichen Rahmen für den Einsatz klar abstecken und z. B. enge Voraussetzungen für die Anwendung regeln. Die Notwendigkeit solcher Eingriffsbefugnisse zeigt

sich im Urteil des Bundesverfassungsgerichts vom 16.02.2023, als sich das Gericht mit der Verfassungsmäßigkeit zweier Ermächtigungsgrundlagen der Länder Hessen und Hamburg für den Einsatz automatisierter Datenanalyseysteme, wie HessenDATA, beschäftigte. Im Urteil wird deutlich, dass das Gericht den Einsatz solcher Systeme grundsätzlich aus verfassungsrechtlicher Sicht für möglich hält, die geschaffenen Eingriffsbefugnisse jedoch unzureichend und verfassungswidrig waren (BVerfG NJW 2023, S. 1196; Bäuerle, 2025, S. 129; Wörner, 2024, S. 627-628).

Ein erster Schritt in diese Richtung ist von der Europäischen Union mit der europäischen KI-Verordnung getan. Die Verordnung arbeitet mit einem risikobasierten Ansatz und stellt an KI-Systeme je nach zugeordnetem Risiko unterschiedlich hohe Anforderungen für ihre Anwendung (Wendt & Wendt, 2024, S. 50-51). Der risikobasierte Ansatz der EU-KI-Verordnung bedeutet, dass KI-Systeme je nach ihrem Gefährdungspotenzial für Grundrechte, Sicherheit und Gesellschaft unterschiedlich streng reguliert werden. Es gibt vier Stufen: KI-Systeme, denen ein unzulässiges Risiko zugeordnet wird, sind verboten. Systeme mit hohem Risiko sind zwar erlaubt, unterliegen aber strengen gesetzlichen Anforderungen, etwa zu Kontrolle, Sicherheit und Transparenz. Bei begrenztem Risiko gelten vor allem Informations- und Transparenzpflichten. KI-Systeme mit minimalem Risiko unterliegen dagegen grundsätzlich keinen besonderen zusätzlichen Vorgaben (Wendt & Wendt, 2024, S. 50 ff.). Bei den hier beschriebenen Systemen handelt es sich um Systeme der Kategorie „Systeme mit hohem Risiko“, sofern die biometrische Gesichtserkennung nachträglich eingesetzt wird. An solche Systeme werden hohe Anforderungen hinsichtlich ihrer Transparenz, Dokumentation, Robustheit, Datenqualität und menschlichen Aufsicht gestellt werden (Wendt & Wendt, 2024, S. 50-51). Sind diese, in erster Linie produktbezogenen Voraussetzungen, erfüllt, kann ein System grundsätzlich in der europäischen Union zum Einsatz kommen. Einzelne Befugnisnormen kann die Verordnung jedoch nicht vorschreiben, diese Aufgabe obliegt dem nationalen Gesetzgeber. Wie sich dem Koalitionsvertrag entnehmen lässt, plant die Bundesregierung während der aktuellen Legislaturperiode, sich diesem Thema für bestimmte Maßnahmen anzunehmen (Koalitionsvertrag von CDU/CSU und SPD, 2025, S. 82). Gleichzeitig ist aber zu bemerken, dass bereits immer wieder KI-gestützte Maßnahmen zum Einsatz kommen, die auf unzureichend oder nicht verfassungsmäßigen Ermächtigungsgrundlagen gründen. Es wird daher dringend Zeit, das technische Können gesetzlich

einzurahmen und die technische Entwicklung der zukünftigen Sicherheitsarbeit zwar zu fördern, aber ebenso Rechtssicherheit und Schutz für Betroffene und Ausführende der Maßnahmen zu schaffen. Der Gesetzgeber muss nun tätig werden, um die Datenflut bei den Polizeibehörden und Staatsanwaltschaften in den Griff zu bekommen, aber auch die Bürger:innen präventiv vor unverhältnismäßigen Maßnahmen zu schützen.

Literatur

- Aden, H., Schönrock, S., John, S., Tahraoui, M. & Kleemann, S. (2022). Accountability-Vorkehrungen für die Erfüllung von Menschenrechtspflichten der Polizei bei der Nutzung Künstlicher Intelligenz. *Zeitschrift für Menschenrechte*, 16(2), 50-72.
- Bäuerle, M. (2025). Automatisierte und KI-gesteuerte Datenverarbeitung und -analyse bei den Sicherheitsbehörden. Perspektiven und Grenzen sicherheitsbehördlicher „Datafizierung“. *Zeitschrift für Datenschutzrecht*, 15(3), 128-131.
- Bitkom e.V. (Hrsg.). (2023). KI in der Polizei – Einsatzpotenziale und Lösungsansätze zur Implementierung. <https://www.bitkom.org/sites/main/files/2023-10/bitkom-positions-papier-ki-polizei-einsatz-implementierung.pdf> (abgerufen am 29.11.2025).
- Brüning, J. (2024). Big Data und Künstliche Intelligenz im Ermittlungsverfahren. In C. Kusche & G. Stefanopoulou (Hrsg.), *Digitalisierung als total social fact der Kriminalwissenschaften* (S. 133-152). Nomos.
- Coombe, J. (2024). Die Fehlbewertung der nachträglichen biometrischen Fernidentifizierung in der KI-VO. *Zeitschrift für das Gesamte Sicherheitsrecht*, 7(6), 262-266.
- Deutscher Bundestag. (Hrsg.). (2024). Analyse polizeilicher Datenbanken. Verfassungsrechtliche Anforderungen an eine Rechtsgrundlage. <https://www.bundestag.de/resource/blob/988022/6530e4ce-c12e2ddab33c77d8b33ec20d/WD-3-145-23-pdf.pdf> (abgerufen am 29.11.2025).
- Ebers, M., Heinze, C., Kruegel, T., Steinrötter, B. (2020). *Künstliche Intelligenz und Robotik* (1. Auflage). C. H. Beck.
- Egbert, S., Esposito, E., Heimstädt, M. (2021). Vorhersagen und Entscheiden: Predictive Policing in Polizeiorganisationen. *Soziale Systeme*, 26(1-2), 189-216.
- Ehringfeld, C. (2024). Es geht um Verantwortung. *Deutsche Polizei*, 73(12), 10-11.
- Els, S. (2021). Einsatz von Künstlicher Intelligenz im Bereich der Strafverfolgung und Gefahrenabwehr. *Kriminalistik*, 75(11), 614-619.
- Ertel, W. (2025). *Grundkurs Künstliche Intelligenz* (6. Auflage). Springer Fachmedien.
- Figaj, P. (2025, 23. Dezember). KI-Videoüberwachung in Mannheim wird weiter ausgebaut. *SWR Aktuell*. <https://www.swr.de/swraktuell/baden-wuerttemberg/mannheim/ki-videoueberwachung-kameras-ausbau-pilotprojekt-100.html>

- Fricke, J. (2020, 27. Mai). Big Data und Künstliche Intelligenz. Chancen und Risiken für die Polizeiarbeit der Zukunft. <https://ksv-polizei-praxis.de/big-data-und-kuenstliche-intelligenz-chancen-und-risiken-fuer-die-polizeiarbeit-der-zukunft/>
- Ibold, V. (2024). Künstliche Intelligenz und Strafrecht. Zur strafrechtlichen Produktverantwortung in der Innovationsgesellschaft (1. Auflage). Nomos.
- Janitzki, D. (2024). Intelligente Videoüberwachung. Funktionsweise und Möglichkeiten der Nutzung durch die Polizei. *Kriminalistik*, 78(1), 49-53.
- Kästner, L. & Schomäcker, A. (2023). KI-Systeme in der modernen Gesellschaft: Potenziale und Grenzen. *Zeitschrift für Urheber- und Medienrecht*, 67(8/9), 558-566.
- Kugelman, D. & Buchmann, A. (2024). Der Algorithmus und die Künstliche Intelligenz als Ermittler. Zum Rechtsrahmen für sicherheitsbehördliche Datenanalysen und für den Einsatz von Verfahren künstlicher Intelligenz. *Zeitschrift für das Gesamte Sicherheitsrecht*, 7(1), 1-10.
- Lang, J. (2023). Intelligente Videoüberwachung. Eine Wirkungsanalyse am Beispiel der Verhaltens-/Bewegungsmustererkennung. *Kriminalistik*, 77(2), 124-128.
- Leffer, L. (2025). Automated Suspicion Algorithms. Strafverfolgung durch Künstliche Intelligenz am Beispiel der Geldwäsche (1. Auflage). Nomos.
- Martini, M. & Botta, J. (2025). Polizeiliche Datenanalyse mittels KI. *Zeitschrift für öffentliches Recht und Verwaltungswissenschaft*, 78(24), 1033-1044.
- Peters, A. (2023). *Smarte Verdachtsgewinnung* (1. Auflage). Nomos.
- Roggan, F. (2024). Der verdeckte Einsatz von Personen-Identifikationssystemen (PerIS) im Strafverfahren. Überlegungen zu möglichen Rechtsgrundlagen für biometrische Gesichtserkennungsverfahren. *Neue Zeitschrift für Strafrecht*, 44(12), 715-718.
- Rückert, C. (2023). Ein Blick in die Blackbox. Künstliche Intelligenz und Machine Learning als Beweismittel im Strafverfahren. *Goltdammer's Archiv für Strafrecht*, 170(7), 361-377.
- Schulz, M. & Evran, T. (2025). Deutschland als KI-Nation – Blickwinkel 1. Die Digitalisierungspläne im Koalitionsvertrag. *Künstliche Intelligenz und Recht*, 2(11), 391-392.
- Wendt, J. & Wendt, D. (2024). *Das neue Recht der Künstlichen Intelligenz. Artificial Intelligence Act (AI Act)* (1. Auflage). Nomos.
- Wörner, L. (2024). Weg von den Hürden, hin zu den Möglichkeiten: KI in Polizei und Strafverfolgung. *Zeitschrift für die gesamte Strafrechtswissenschaft*, 136(3), 616-642.

Zur weiteren Vertiefung:

- Ebers, M., Heinze, C., Kruegel, T., Steinrötter, B. (2020). Künstliche Intelligenz und Robotik (1. Auflage). C. H. Beck.
- Martini, M. & Botta, J. (2025). Polizeiliche Datenanalyse mittels KI. Zeitschrift für öffentliches Recht und Verwaltungswissenschaft, 78(24), 1033-1044.
- Russell, S. & Norvig, P. (2023). Künstliche Intelligenz. Ein moderner Ansatz (4. Auflage). Pearson Studium.

Mediathek



SWR K.I.-Kommissar ermittelt? – Künstliche Intelligenz in der Polizeiarbeit.



Deutschlandfunk. Gesichtserkennung – Macht KI und zu gläsernen Bürgern?



WELT Doku. FAHNDUNG AUF DEM HOLODECK: Wie künstliche Intelligenz & VR die Polizeiarbeit revolutioniert.



Alina Borowy ist Diplomjuristin und wissenschaftliche Mitarbeiterin am Lehrstuhl für Strafrecht, Strafprozessrecht, Sanktionsrecht und Wirtschaftsstrafrecht an der Christian-Albrechts-Universität zu Kiel.

»Insgesamt lässt sich also sagen, dass die prädiktive Bearbeitung von Kriminalität im Predictive Policing häufig mit einer präventiven Oberflächlichkeit und Kurzfristigkeit einhergeht, die zudem negative Auswirkungen auf die betroffenen Personen haben kann, da in ihr Leben in hemmender und bisweilen unterdrückender Weise – also repressiv – eingegriffen wird.«

Dr. Simon Egbert

Simon Egbert

Predictive Policing – Algorithmische Vorhersagen in der polizeilichen Kriminalprävention

1. Einleitung

Die Integration Künstlicher Intelligenz (KI) in polizeiliches Handeln markiert eine der bedeutendsten Entwicklungen zeitgenössischer Sicherheitspraktiken. Erste Bemühungen, (Vorläufer von) KI in die Polizeiarbeit zu bringen, sind dabei mit dem Aufkommen der prognosebasierten Polizeiarbeit, *Predictive Policing* genannt, verbunden. Dieser polizeiliche Ansatz kann verstanden werden als der Einsatz algorithmischer Verfahren zur Erstellung von Kriminalitätsprognosen und er verspricht die gezielte Identifikation potenzieller Tatorte, Tatzeitpunkte oder Täter:innen und damit eine effizientere, datengestützte Allokation polizeilicher Ressourcen. In Deutschland wurden seit Anfang/Mitte der 2010er Jahre verschiedene raumbezogene Systeme pilotiert und teilweise in den Regelbetrieb überführt, darunter PRECOBS in Baden-Württemberg, Sachsen und Bayern, KrimPro in Berlin, SKALA in Nordrhein-Westfalen, PreMAP in Niedersachsen und KLB-operativ in Hessen (Egbert & Kornehl, 2022; Sommerer, 2017). Aber auch ein personenbezogenes System, das dem Predictive Policing zugeordnet werden kann, wird sein einigen Jahren in ganz Deutschland eingesetzt – das vom BKA entwickelte RADAR-iTE (Sonka et al., 2020). Diese Entwicklungen haben intensive kriminalpolitische und wissenschaftliche Debatten ausgelöst, die zwischen technikoptimistischen Effizienzversprechen und grundlegenden Bedenken hinsichtlich rechtsstaatlicher Garantien, Diskriminierungsrisiken und demokratischer Kontrolle oszillieren (z. B. Belina, 2016; Egbert & Mann, 2021; Hofmann, 2020; Leese, 2024; Singelstein, 2018; Sommerer, 2020).

Der vorliegende Beitrag ordnet das Phänomen Predictive Policing in diese Debatten, mit besonderem Bezug auf die Kriminalprävention, ein und legt eine konzeptuelle Rahmung dar, die über eine rein technische Betrachtungsweise hinausgeht. Im Zentrum steht dabei die Erkenntnis, dass Predictive Policing nicht als solitäres (und neutrales) technisches Instrument verstanden werden kann, sondern als soziotechnische Interaktion, an dem Mensch und Maschine gleichermaßen beteiligt sind, und deren präventive Wirksamkeit maßgeblich von der praktischen Implementierung, den organisatorischen Rahmenbedingungen und den verfügbaren personellen wie materiellen Ressourcen abhängt. Anders gesagt: Die präventive Leistung algorithmischer Systeme entfaltet ihre Relevanz erst im Zusammenspiel mit polizeilichen Handlungslogiken, institutionellen Strukturen und den Deutungsmustern der beteiligten Akteur:innen, was auf die Formel hinausläuft, dass Prognostik nicht automatisch auch Prävention bedeutet (Egbert & Esposito, 2024).

Darüber hinaus entwickelt der Beitrag eine kritische Perspektive auf Predictive Policing als Form repressiver Prävention (Egbert, 2022). Denn algorithmische Prognoseverfahren konzentrieren sich typischerweise auf die Identifikation von Symptomen und unmittelbaren Kriminalitätsrisiken, nicht jedoch auf die Bearbeitung der gesellschaftlichen, sozioökonomischen und strukturellen Ursachen von Kriminalität. Diese Fokussierung führt dazu, dass Predictive Policing primär auf kurzfristige Abschreckung und räumliche Verdrängung abzielt, ohne nachhaltige präventive Lösungen zu fördern, die an den Entstehungsbedingungen kriminellen Verhaltens ansetzen würden (Egbert & Esposito, 2024). Mehr noch: Die vermeintlich neutrale algorithmische Datenanalyse reproduziert dabei häufig bestehende polizeiliche Schwerpunktsetzungen und kann so zu einer Verfestigung selektiver Kontrollpraktiken beitragen (Egbert & Mann, 2021; Kemme, 2025).

Abschließend gibt der Beitrag einen Ausblick auf die Zukunft algorithmischer Präventionsstrategien in Deutschland. Entscheidend für einen kriminalpräventiven Effekt wird sein, wie diese Strategien praktisch umgesetzt werden. Während umfassende Datenanalyseplattformen und die Integration verschiedener Datenquellen sowie komplexerer KI-Verfahren voranschreiten, bleibt die konkrete Implementierung – etwa in polizeiliche Arbeitsabläufe und Entscheidungsprozesse – ausschlaggebend.

2. Konzeptuelle Grundlagen

Predictive Policing kann verstanden werden als die polizeiliche Anwendung algorithmischer Analyseverfahren, um wahrscheinliche Zeiträume, Orte oder Täter:innen bzw. Opfer zukünftiger Kriminalität vorherzusagen und daran anschließende Präventionsmaßnahmen auszuführen (Egbert, 2025; Egbert & Leese, 2021). Diese Definition verweist bereits auf die zentrale Charakteristik des Ansatzes: die systematische Verbindung von datengestützter Prognose und präventivem polizeilichen Handeln.

2.1 Definition und Funktionsweise von Predictive Policing

Predictive Policing bezeichnet einen polizeilichen Ansatz, der sich sowohl auf Räume als auch auf Personen(-gruppen) beziehen kann. Während sich raumbezogenes Predictive Policing auf die Identifikation geografischer Gebiete konzentriert, in denen Kriminalität zeitnah mit erhöhter Wahrscheinlichkeit zu erwarten ist, zielt personenbezogenes Predictive Policing auf die Identifikation von Individuen oder Gruppen ab, die als potenzielle Täter:innen oder Opfer zukünftiger Straftaten eingestuft werden. Diese Unterscheidung entspricht der (inter-)nationalen Fachdiskussion, die beide Ausprägungen unter dem Begriff Predictive Policing subsumiert (National Academies of Sciences, Engineering, and Medicine, 2025; Sommerer, 2020) – auch wenn die polizeiliche Sprachregelung in Deutschland personenbezogene Verfahren von dieser Definition ausschließt (Seidensticker, 2022, S. 194).

Der Kern von Predictive Policing liegt in der Selektion und Hierarchisierung möglicher Ziele (Personen oder Orte) von Präventionsmaßnahmen – da polizeiliche Ressourcen begrenzt sind, kann nicht überall gleichzeitig und in Bezug auf jede Person gleichermaßen präventiv agiert werden. Predictive Policing folgt daher dem Grundsatz der Priorisierung: Mit Hilfe algorithmisch erstellter Prognosen sollen Prioritäten gesetzt werden hinsichtlich der präventiv zu bearbeitenden Orte oder Personen.

Funktional betrachtet ist Predictive Policing dabei als soziotechnischer Prozess zu verstehen, an dem Mensch und Maschine gleichermaßen beteiligt sind: die Sammlung und Aufbereitung von Daten, die algorithmische Analyse zur Prognoseerstellung sowie die praktische Anwendung der Prognosen durch polizeiliche Akteur:innen (Egbert & Heimstädt, 2023;

Egbert & Leese, 2021). Die technische Funktionsweise kann sich dabei im Einzelfall stark unterscheiden, was auch bedeutet, dass einige der genutzten Verfahren nicht als Methoden im Sinne der Künstlichen Intelligenz zu verstehen sind, was z. B. für die kommerzielle Software PRECOBS zutrifft (Schweer, 2015). Die Eigenentwicklung der nordrheinwestfälischen Landeskriminalamts SKALA („System zur Kriminalitätsauswertung und Lageantizipation“) wiederum ist fraglos als KI-Anwendung zu begreifen (Seidensticker & Schwarz, 2022).

2.2 Prävention als Ziel: Von situativer Kriminalprävention zu algorithmischer Prognose

Predictive Policing ist grundsätzlich im Paradigma der Kriminalprävention verankert. Die Prognosesoftware identifiziert bestimmte Räume und Zeitfenster, in denen ein erhöhtes Risiko für Straftaten – etwa Einbrüche – besteht. Die Polizei kann daraufhin ihre Präsenz in diesen Gebieten gezielt erhöhen, indem sie Streifenfahrten zu den prognostizierten Zeitpunkten und Orten durchführt. Ziel ist es, potenzielle Täter:innen abzuschrecken und die Wahrscheinlichkeit eines Delikts zu senken (Pett & Gluba, 2017). Darüber hinaus ermöglichen die Prognosen grundsätzlich auch gezielte Observationen von Personen oder Orten, die als besonders risikobehaftet gelten (Schweer, 2015). So kann die Polizei beispielsweise bestimmte Personen, die in der Vergangenheit auffällig waren oder in den Prognosemodellen als besonders gefährdet eingestuft werden, verstärkt beobachten oder kontrollieren. Die anfänglich wiederholt artikulierte Erwartung, mittels prognosebasierter Verfahren Täter:innen gleichsam *in flagranti* zu identifizieren, erwies sich jedoch als nicht einlösbar. Stattdessen verlagerte sich der Einsatz der Prognosen rasch auf eine primär gefahrenabwehrende Funktion: die gezieltere Steuerung uniformierter Kräfte mit dem Ziel der Abschreckung durch erhöhte und sichtbare vor-Ort-Präsenz (Egbert & Kornehl, 2022; Pett & Gluba, 2017). Diese Verschiebung ist nicht zuletzt darauf zurückzuführen, dass verdeckte Observationen einschlägiger Orte einen erheblichen Personalaufwand erfordern, während zugleich der epistemische Status der Prognosen – insbesondere im Hinblick auf ihre Präzision und Verlässlichkeit – häufig unklar ist (Gerstner, 2017; Sommerer, 2020).

Innerhalb des breiten Spektrums kriminalpräventiver Ansätze lässt sich Predictive Policing daher im Bereich der situativen Kriminalprävention verorten. Diese zielt nicht primär auf die Veränderung individueller Dis-

positionen oder gesellschaftlicher Strukturen ab, sondern auf die Modifikation situativer Gelegenheitsstrukturen für Kriminalität (Clarke, 2016). Algorithmische Prognoseverfahren konzentrieren sich dabei typischerweise auf die Identifikation von Symptomen und unmittelbaren Kriminalitätsrisiken, nicht jedoch auf die Bearbeitung der gesellschaftlichen, sozioökonomischen oder strukturellen Ursachen von Kriminalität, den sogenannten ‚root causes‘ (z. B. Petersen, 2024). Diese Fokussierung führt dazu, dass Predictive Policing primär auf Abschreckung und räumliche Verdrängung abzielt, ohne nachhaltige präventive Lösungen zu fördern, die an den Entstehungsbedingungen kriminellen Verhaltens ansetzen würden (Egbert & Esposito, 2024).

2.3 Abgrenzung zu klassischen Formen der präventiven Polizeiarbeit

Obwohl Prävention seit langem zum polizeilichen Handlungsrepertoire gehört, unterscheidet sich Predictive Policing in mehreren Aspekten von klassischen Formen präventiver Polizeiarbeit. Der zentrale Unterschied liegt in der algorithmischen Aufbereitung von Daten zur Generierung von Zukunftswissen. Während traditionelle präventive Polizeiarbeit stark auf der Erfahrung einzelner Polizeibeamt:innen, lokalem Wissen und intuitiven Einschätzungen basiert, verspricht Predictive Policing eine objektivere, datengestützte und systematisierte Form der Risikoabschätzung (Balogh, 2016; Bode & Stoffel, 2023; Schweer, 2020).

Damit hängt ein weiteres Unterscheidungsmerkmal eng zusammen: die gesteigerte Geschwindigkeit der Informationsverarbeitung. Klassische Formen der Kriminalprävention arbeiten häufig mit längerfristigen strategischen Analysen oder reagieren auf bereits identifizierte Kriminalitätsschwerpunkte. Predictive Policing hingegen ermöglicht durch automatisierte Analysen die schnelle Erstellung operativer Prognosen, die unmittelbar in taktische Entscheidungen übersetzt werden können (Egbert & Leese, 2021; Leese & Pollozek, 2023). Diese Aktualisierung von polizeilichen Präventionsbemühungen macht Predictive Policing zu einem neuartigen Phänomen, das neue Bekämpfungstaktiken und Strategien durch ‚operative Prognosen‘ ermöglicht.

Ein anschauliches Beispiel für diese Neuartigkeit bietet das für raumbezogenes Predictive Policing wichtige Near-Repeat-Muster. Dieses basiert auf der Annahme, dass professionelle Serieneinbrecher:innen dazu neigen, nach einer erfolgreichen Tat schnell in unmittelbarer Nähe erneut zuzuschlagen (Gluba, 2017; Johnson & Bowers, 2014). Da das Risiko einer Folgetat jedoch schnell wieder auf Normalniveau absinkt, sind algorithmische Analysen erforderlich, um hinreichend zügig eine Risikoprognose für den betreffenden Raum zu erhalten (Gerstner, 2017). Erst die technische Beschleunigung der Analysefähigkeit durch Algorithmen macht diese Form der präventiven Intervention operativ demnach sinnvoll umsetzbar.

3. Einsatzfelder und Praktiken in Deutschland

Die praktische Anwendung von Predictive Policing hat in den vergangenen Jahren eine bemerkenswerte, wenn auch uneinheitliche Verbreitung erfahren. Während in einigen Polizeibehörden umfassende Pilotprojekte und dauerhafte Implementierungen zu beobachten waren, haben andere Polizeibehörden nach kurzen Erprobungsphasen wieder von der Technologie Abstand genommen. Dieses Kapitel zeichnet die empirischen Einsatzfelder von Predictive Policing in Deutschland nach.

3.1 Raumbezogenes Predictive Policing

In Deutschland begann die Ära des Predictive Policing im Jahr 2014 mit der Pilotierung der kommerziellen Software PRECOBS in Bayern (Egbert & Kornehl, 2022; Okon, 2020). Das Pre Crime Observation System (PRECOBS) zielte primär auf die Prävention von Wohnungseinbruchdiebstahl ab und basierte auf dem oben beschriebenen Near-Repeat-Ansatz (Balogh, 2016; Schweer, 2015). Die Software analysierte täglich polizeiliche Daten zu Wohnungseinbrüchen und erstellte ortsbezogene Kriminalitätsprognosen für bestimmte geografische Gebiete, in denen mit erhöhter Wahrscheinlichkeit Folgetaten zu erwarten waren.

Die Funktionsweise von PRECOBS illustriert treffend die grundlegende Logik raumbezogenen Predictive Policing in Deutschland. Wenn ein Einbruch registriert wurde, analysierte die Software verschiedene Merkmale der Tat, insbesondere den Modus Operandi und die Art der Beute, um festzustellen, ob es sich um professionelle Täter:innen handelte. Wurde

z.B. ein Fenster aufgebrochen und ausschließlich Schmuck gestohlen, galt dies als Indikator für professionelles Vorgehen und wurde als sogenanntes Trigger-Kriterium definiert. Basierend auf der Near-Repeat-Theorie wurden dann für die umliegenden Gebiete Risikoprognosen erstellt. Die betroffenen Areale wurden entsprechend als hoch-, mittel- oder niedrigriskant gekennzeichnet (Balogh, 2016; Schweer, 2015, 2020).

Die polizeiliche Reaktion auf solche Vorhersagen bestand überwiegend in einer Verstärkung der Streifen in den betroffenen Gebieten. Dahinter stand die Annahme, dass professionelle Täter:innen durch sichtbare Polizeipräsenz abgeschreckt werden und von ihren Einbruchsplänen ablassen, da ihnen das Entdeckungsrisiko zu hoch erscheint. Dieser präventive Ansatz wurde deutlich häufiger verfolgt als der alternative repressive Ansatz, bei dem Observationskräfte eingesetzt werden, um Täter:innen *in flagranti* zu erwischen (Pett & Gluba, 2017). Letzteres wurde höchstens zu Beginn vereinzelt durchgeführt, da die Bereitstellung mehrerer Observationskräfte für mehrere Tage dem Hauptziel von Predictive Policing, der Steigerung der Effizienz polizeilicher Arbeit, diametral entgegensteht (Egbert & Kornehl, 2022).

Nach Bayern folgten insgesamt sechs weitere deutsche Bundesländer mit eigenen Predictive Policing-Initiativen. Während Baden-Württemberg und Sachsen ebenfalls die kommerzielle Lösung PRECOBS erprobten, entschieden sich Nordrhein-Westfalen, Hessen, Niedersachsen und Berlin für Eigenentwicklungen (Egbert & Kornehl, 2022; Seidensticker et al., 2018; Sommerer, 2017). Diese Systeme trugen Namen wie SKA-LA (System zur Kriminalitätsauswertung und Lageantizipation) in Nordrhein-Westfalen (LKA NRW, 2018), KLB-operativ (Kriminalitätslagebildoperativ) in Hessen, KrimPro in Berlin und PreMAP (Predictive Mobile Analytics for Police) in Niedersachsen (LKA Niedersachsen, 2018). Trotz teilweise erheblicher Unterschiede in der technischen Ausgestaltung hatten alle Lösungen gemein, dass sie sich in erster Linie auf Wohnungseinbrüche konzentrierten und überwiegend dem Near-Repeat-Ansatz folgten. Hervorzuheben ist hierbei insbesondere der Ansatz aus NRW, da er neben dem Wohnungseinbruchdiebstahl auch Gewerbeeinbrüche sowie Autodiebstähle prognostiziert und dafür neben dem Near Repeat-Ansatz noch weitere theoretische Referenzen nutzbar macht, wie z. B. die Theorie sozialer Desorganisation (LKA NRW, 2018; Seidensticker, 2021). Dieser multitheoretische Ansatz, der u. a. eine heterogenere Datenbasis,

die auch sozioökonomische Informationen umfasst, impliziert, ist nur durch die Nutzung von KI-Methoden umsetzbar, in diesem Falle einem *random forest*-Modell (Seidensticker & Schwarz, 2022).

Die Präferenz für Eigenentwicklungen gründete sich primär auf ökonomische Erwägungen, da diese im Vergleich zu kommerziellen Softwarelösungen in der Regel mit geringeren Kosten verbunden sind, insbesondere bei langfristigem Einsatz. Darüber hinaus wurde seitens der Behörden die Sicherstellung größtmöglicher Transparenz hinsichtlich der Funktionsweise der Systeme sowie die Wahrung der Autonomie in deren Weiterentwicklung sowie der Datenschutz als bedeutsam erachtet (Egbert & Kornehl, 2022; Leese, 2024). Diese Entscheidung erwies sich augenscheinlich als folgenreich für die Nachhaltigkeit der Systeme: Von den sieben Bundesländern, die raumbezogenes Predictive Policing eingesetzt haben, nutzen derzeit nur noch drei eine Prognosesoftware, nämlich Hessen, Berlin und Nordrhein-Westfalen. Bezeichnenderweise handelt es sich dabei ausschließlich um Eigenentwicklungen (Egbert, 2025).

Der Rückgang von Predictive Policing in Deutschland lässt sich exemplarisch an der Entscheidung der bayerischen Polizei nachvollziehen, die im Herbst 2021 ankündigte, den Vertrag für PRECOBS nach sieben Jahren nicht zu verlängern. Die Begründung verwies auf den starken Rückgang der Fallzahlen bei Wohnungseinbrüchen in Bayern und die damit verbundene quantitative Minderung der zur Berechnung notwendigen Datengrundlage, die zu einer Verringerung der Prognosen führten und eine gezielte Einsatzsteuerung nicht mehr möglich machten (Polizei Bayern, 2021). Vor dem Hintergrund einer Kosten-Nutzen-Analyse wurde daher die Entscheidung getroffen, den Betrieb der Software einzustellen.

Diese – auch durch die COVID-19-Pandemie wesentlich mitverursachte – Entwicklung verdeutlicht eine grundlegende Limitation von Predictive Policing: Software zur Vorhersage von Straftaten ist zwingend auf eine ausreichend große Zahl von Fällen der zu prognostizierenden Straftat angewiesen. Wenn zu wenige Straftaten stattfinden, sind die vorhergesagten Risiken statistisch nicht zuverlässig genug, um ausgegeben werden zu können (Gerstner, 2017). Darüber hinaus können polizeiliche Prognosesysteme nicht einfach auf andere Straftaten ausgedehnt werden, da viele Delikte nicht ausreichend häufig vorkommen und es zudem ein klar konturiertes raumzeitliches Bewegungsmuster der typischen

Täter:innen in einem Deliktsbereich braucht, damit eine betreffende Straftat vorhersagbar ist. Diese Regelmäßigkeiten und raumzeitlichen Muster sind vielen Arten von Straftaten jedoch fremd (Kaufmann et al., 2019).

Zusammenfassend lässt sich raumbezogenes Predictive Policing, wie es in Deutschland durchgeführt wird, in vielerlei Hinsicht als sehr enger und mitunter technisch nicht besonders komplexer polizeilicher Ansatz charakterisieren. Die Konzentration auf einen einzigen Deliktsbereich, die ausschließliche Fokussierung auf professionelle Einbrecher:innen und die Beschränkung der polizeilichen Reaktion auf verstärkte Streifengänge zur kurzfristigen Abschreckung oder Verdrängung machen Predictive Policing letztlich zu einem recht begrenzten Ansatz – woran auch moderne KI-Verfahren erstmal nichts ändern können.

3.2 Personenbezogenes Predictive Policing

Entgegen anderslautenden Positionierungen, insbesondere seitens polizeilicher Akteur:innen (vgl. dazu Seidensticker, 2021), werden im vorliegenden Verständnis auch personenbezogene Ansätze dem Phänomen des Predictive Policing zugerechnet. Dazu zählen jene Instrumente, die das zukünftige Risiko von Straftaten auf der Ebene individueller Personen bestimmen und hieran spezifische Präventions- und Interventionsmaßnahmen binden, wie etwa das vom Bundeskriminalamt (BKA) entwickelte RADAR iTE.

RADAR-iTE (Regelbasierte Analyse potentiell destruktiver Täter zur Einschätzung des akuten Risikos – islamistischer Terrorismus) wird seit 2017 bundesweit im Staatsschutzbereich der Landespolizeien eingesetzt und liegt seit 2019 in der weiterentwickelten Version 2.0 vor (Sonka et al., 2020). Das Instrument zielt, als „erste Phase eines mehrstufigen Abklärungsmodells“ (Endrass et al., 2022, S. 394), auf eine einheitliche Einschätzung des Gewaltrisikos polizeibekannter Personen aus dem islamistischen Spektrum sowie auf die darauf abgestimmte Priorisierung polizeilicher Maßnahmen (Goertz, 2020, S. 81f.; Trunk & Simmert, 2020). Im Zentrum stehen damit bereits registrierte Gefährder:innen und sogenannte Relevante Personen, sodass RADAR iTE eher als Instrument zur Bestätigung eines bestehenden Gefahrenverdachts denn als originär verdachtsgenerierendes Tool zu charakterisieren ist (Meyer, 2017).

Auf technisch-analytischer Ebene unterscheidet sich RADAR iTE von räumlichen Prognosesystemen wie PRECOBS oder vergleichbaren Verfahren, da es sich nicht um eine eigenständige Vorhersagesoftware handelt. Stattdessen kombiniert es herkömmliche Tabellenkalkulations- und Textverarbeitungsprogramme mit einem standardisierten Fragenkatalog, der Indikatoren erfasst, die von der forensisch psychologischen Arbeitsgruppe der Universität Konstanz als relevant für die Einschätzung politisch motivierten Gewaltrisikos eingestuft wurden (Sonka et al., 2020). Diese Indikatoren sind auf Grund eines Aktenstudiums überprüfbar und folgen einer vorab festgelegten Bewertungstaxonomie, weshalb es sich um ein aktuarisches, also versicherungsmathematisch fundiertes, Instrument handelt, das zudem einem biostatistischen Ansatz folgt (Endrass et al., 2022, S. 390). Diese Indikatoren ermöglichen eine Zuordnung der bewerteten Personen zu einer zunächst drei-, mittlerweile zweistufigen Risikoskala (moderates vs. hohes Risiko).

Das RADAR-iTE zugrunde liegende, „wissenschaftlich geprüft(e) Verrechnungsmodell(l)“ (BT-Drs. 18/13422: 7) integriert quantitative und qualitative Elemente. Den quantitativ definierten Merkmalen werden numerische Werte (etwa +1 für Risikofaktoren und –1 für Schutzfaktoren) zugewiesen und zu einem Summenwert aggregiert, der die formale Basis der Risikokategorisierung bildet (BT-Drs. 19/12859). Ergänzend treten qualitativ definierte „rote Flaggen“ hinzu, die nicht in die Berechnung eingehen, aber bereits für sich genommen ein hohes Risiko anzeigen können (Sonka et al., 2020).

Der ursprüngliche RADAR-iTE-Fragebogen umfasste 73 Kriterien, die in Form geschlossener Fragen sieben Themenkomplexen zugeordnet waren, darunter Gewalt als Täter, Sprengstoffe und Waffen, militärische Erfahrung, Ausreisen in Kriegs- oder Krisengebiete, Zugehörigkeit zu einer radikalen Szene sowie Aspekte der sozialen Eingliederung. Mit RADAR- iTE 2.0 wurde dieser Katalog aus Gründen der Straffung auf 59 Merkmale reduziert (Sonka et al., 2020). Öffentlich bekannt ist, dass weiterhin Bereiche wie „Gewalt gegen Personen“, „Umgang mit Behörden und anderen Institutionen“ sowie „Militär und Ausreise“ abgedeckt werden. Seit Ende 2024 arbeitet das BKA, gemeinsam mit der Kriminologischen Zentralstelle (KrimZ), unter dem Titel SMART („Strukturierte Methodik zur Analyse des individuellen Risikos von Tatbegehungen“), an einer Weiterentwicklung der Methodik (KrimZ, 2025, S. 26, o. J.a).

Als Weiterentwicklung des ursprünglich auf das islamistische Spektrum fokussierten Instruments wurden in den vergangenen Jahren weitere Varianten des RADAR-Ansatzes etabliert, die dessen personenbezogene Logik auf andere Kontexte übertragen. So wurde mit *RADAR rechts* ein eigenständiges Risikobewertungsinstrument für polizeibekannt Personen aus dem rechtsextremen Spektrum entwickelt, das seit 2022 bundesweit eingesetzt wird. Dieses soll eine standardisierte, merkmalsbasierte Einschätzung des Risikos der Begehung einer konkret lebensgefährlichen rechtsmotivierten Gewalttat sowie eine darauf ausgerichtete Priorisierung von Maßnahmen im Bereich der politisch motivierten Kriminalität ‚rechts‘ ermöglichen (BKA, o. J.; Goertz, 2022, S. 347). Ergänzend hierzu wurde mit *RADAR Haft* ein Instrument konzipiert, das auf die systematische Einschätzung des von inhaftierten Personen ausgehenden extremistisch motivierten Gewaltrisikos zielt und insbesondere Entscheidungen zum Risikomanagement im Vollzugskontext unterstützen soll (Scheu, 2025, S. 21). Damit wird die Logik des personenbezogenen Risk Assessments über den klassischen polizeilichen Gefährder:innendiskurs hinaus auf den Bereich des Strafvollzugs ausgedehnt. Das entsprechende, von der KrimZ geleitete Forschungsprojekt endete im November 2022 (KrimZ, o. J.b). Ob und wie RADAR-Haft tatsächlich eingesetzt wird, ist bislang unbekannt.

4. Evaluationen: Wirkt Predictive Policing?

Die Frage nach der empirischen Wirksamkeit von Predictive Policing zählt zu den kontroversesten und zugleich zentralsten Themen innerhalb der wissenschaftlichen und kriminalpolitischen Diskussion. Gerade im Hinblick auf die Verhältnismäßigkeit des Einsatzes entsprechender Prognosesysteme wäre ein belastbarer Nachweis erforderlich, ob diese Maßnahmen tatsächlich die angestrebten Effekte erzielen. Die bisher vorliegenden Befunde zeichnen ein ambivalentes Bild: Sie liefern weder eindeutige Nachweise für eine kriminalpräventive Effektivität von Predictive-Policing-Systemen noch belegen sie deren Ineffektivität in klarer Weise. Ein wesentlicher Grund hierfür liegt darin, dass viele Polizeibehörden bislang keine systematischen oder methodisch belastbaren Evaluationsverfahren implementiert haben, um den tatsächlichen Nutzen der eingesetzten Prognosesoftware zu überprüfen. Hinzu kommt, dass die

Messung kriminalpräventiver Wirksamkeit generell – und im Kontext von Predictive Policing im Besonderen – mit erheblichen methodischen und praktischen Herausforderungen verbunden ist, sodass eine hinreichend systematische Umsetzung in der Praxis häufig nicht realisierbar erscheint.

4. 1 Evaluation von Predictive Policing: Methodische und praktische Hürden

Die Herausforderungen bei der Evaluation von Predictive Policing sind vielfältig. Ein zentrales methodisches Problem besteht darin, dass die kausale Wirkung schwer zu isolieren ist: Wenn in einem prognostizierten Risikogebiet kein Einbruch stattfindet, ist unklar, ob dies auf die verstärkte Polizeipräsenz zurückzuführen ist oder ob ohnehin kein Einbruch geplant war. Umgekehrt könnte ein Einbruch in einem Prognosegebiet darauf hinweisen, dass die Prognose korrekt war, die polizeiliche Reaktion jedoch unzureichend, oder dass die Täter:innen sich von der Polizeipräsenz nicht haben abschrecken lassen. Zudem besteht die Möglichkeit räumlicher Verdrängungseffekte: Täter:innen könnten lediglich auf benachbarte, nicht prognostizierte Gebiete ausweichen, sodass die Gesamtkriminalität nicht sinkt, sondern nur verlagert wird.

Ein besonderes Problem liegt in der bereits erwähnten Ressourcenfrage: Wenn die lokalen Polizeidienststellen nicht in der Lage sind, genügend Kräfte bereitzustellen, um die prognostizierten Risikogebiete intensiv zu bestreifen, kann selbst eine statistisch perfekte Prognose keine präventive Wirkung entfalten. Die soziotechnische Dimension von Predictive Policing macht es somit schwierig zu bestimmen, ob ausbleibende Effekte auf die Qualität der Prognosen oder auf Defizite in der praktischen Umsetzung zurückzuführen sind.

Um die Wirksamkeit hinreichend abgesichert überprüfen zu können, wäre eine Evaluation erforderlich, die einem randomisierten Kontrollgruppendesign folgt – einem Verfahren, das insbesondere aus der Medikamentenforschung bekannt ist und dort als Goldstandard der Evaluationsmethodik gilt (Farrington & Welsh, 2005). Im Kontext von Predictive Policing würde dies bedeuten, dass Räume oder Gebiete, in denen die Wahrscheinlichkeit z. B. von Einbruchskriminalität als (ungefähr) gleich hoch gelten kann, zufällig einer Experimentalgruppe zugewiesen werden, und mit Kriminalitätsprognosen polizeilich bearbeitet werden,

während eine Kontrollgruppe ohne diese Interventionen verbleibt. Die Wirksamkeit der Methode könnte so durch den Vergleich von Kriminalitätsraten oder anderen relevanten Indikatoren zwischen beiden Gruppen von Gebieten evaluiert werden. Eine Evaluationsstudie zu Predictive Policing, die diese hohen Anforderungen zu erfüllen vermag, liegt aber bis dato nicht vor.

4.2 Evaluationen zu Predictive Policing in Deutschland

In Baden-Württemberg wurde *PRECOBS* zweimal wissenschaftlich evaluiert, ohne dass eindeutige Ergebnisse zur kriminalpräventiven Wirkung von Predictive Policing erzielt werden konnten. Die erste Evaluation kam zu dem Schluss, dass die Prognosegenauigkeit von *PRECOBS* zwar über einer zufälligen Verteilung lag, jedoch keine statistisch signifikanten Effekte auf die Reduzierung von Wohnungseinbrüchen nachgewiesen werden konnten (Gerstner, 2017). Diese Befunde trugen dazu bei, dass Baden-Württemberg entschied, *PRECOBS* nicht in den Regelbetrieb zu überführen (Mayer, 2019). Die zweite Evaluation in den Jahren 2017 bis 2018, die einem quasiexperimentellen Studiendesign folgte, bestätigte diese gemischten Ergebnisse: Zwar konnte die Software Risikogebiete identifizieren, ein klarer präventiver Effekt auf die tatsächliche Einbruchskriminalität ließ sich jedoch nicht belegen (Gerstner & Dohse, 2022). Gerstner und Dohse evaluierten die zweite Phase des Pilotprojektes, indem sie die Entwicklung der Einbruchszahlen vor und nach Einführung der Software analysierten und mit Kontrollregionen verglichen, die nicht am Projekt teilnahmen. Auf diese Weise konnten sie prüfen, ob Unterschiede in den Kriminalitätsraten plausibel auf die Maßnahme zurückzuführen waren, und zugleich die organisatorische Integration der Software in den Polizeialltag beobachten. Das Vorgehen erlaubte eine empirisch fundierte, zugleich praxisnahe Bewertung der Wirksamkeit und Grenzen des Programms, auch wenn es dem Goldstandard des randomisierten Kontrollgruppendesign nicht vollumfänglich entsprach.

Auch in Niedersachsen wurde das selbst entwickelte System *PreMAP* in zwei aufeinanderfolgenden Projektphasen separat voneinander bewertet. Die erste Evaluation von *PreMAP*, dokumentiert im Abschlussbericht der ersten Projektphase (LKA Niedersachsen, 2018), konzentrierte sich auf die Pilotierung in Wolfsburg sowie in der Polizeiinspektion Salzgitter/Peine/Wolfenbüttel. Methodisch wurden qualitative Interviews, Online-

Befragungen und technische Nutzungsdaten kombiniert, um sowohl die Prognosequalität als auch die praktische Integration in den Polizeialltag zu bewerten. Die Ergebnisse zeigten, dass PreMAP grundsätzlich in der Lage war, Risikogebiete zu identifizieren und mobil auf Tablets sowie Arbeitsplatz-PCs bereitzustellen. Allerdings erwies sich die Bewertung des kriminalpräventiven Mehrwerts als schwierig: Zwar wurden Maßnahmen in den prognostizierten Gebieten durchgeführt, ein eindeutiger Rückgang der Einbruchskriminalität ließ sich jedoch nicht nachweisen. Die Evaluation betonte daher die methodischen Grenzen, insbesondere die Abhängigkeit von der Datenqualität und die fehlende Möglichkeit einer randomisierten Kontrolle. Die zweite Evaluation, festgehalten im Abschlussbericht zur erweiterten Pilotierung von PreMAP (LKA Niedersachsen, 2021), bezog weitere Standorte wie Osnabrück ein und verfolgte ein breiteres Untersuchungsdesign. Neben quantitativen Analysen der Fallzahlenentwicklung und Near-Repeat-Quoten wurden auch soziotechnische Aspekte wie die Nutzung der PreMAP-Anwendung und die Akzeptanz durch Polizeikräfte untersucht. Die Befunde bestätigten die Ergebnisse der ersten Phase: PreMAP konnte Risikogebiete zuverlässig markieren, doch blieb der präventive Effekt auf die tatsächliche Einbruchskriminalität empirisch ambivalent. Die Evaluation hebt hervor, dass externe Faktoren wie Tätergruppen, allgemeine Kriminalitätstrends und organisatorische Rahmenbedingungen die Wirkung maßgeblich beeinflussen. Insgesamt wird PreMAP als technisch funktional und organisatorisch nützlich bewertet, die (messbare) kriminalpräventive Wirkung blieb jedoch begrenzt.

Die Evaluation von *SKALA* in Nordrhein-Westfalen wurde kooperativ durch die Zentralstelle Evaluation beim LKA NRW (ZEVA) und die Gesellschaft für innovative Sozialforschung und Sozialplanung e.V. (GISS) durchgeführt und erstreckte sich über 29 Monate (LKA NRW & GISS, 2018). Methodisch folgte sie einem theoriebasierten Mixed-Methods-Design, das quantitative Analysen polizeilicher Daten mit qualitativen Interviews und Fokusgruppen kombinierte. Ergänzend kam ein quasi-experimentelles Kontrollgruppendesign zum Einsatz, bei dem Prognosegebiete mit unterschiedlicher Wahrscheinlichkeit für Wohnungseinbruchdiebstähle verglichen wurden. Die Ergebnisse zeigten, dass *SKALA* organisatorisch und analytisch einen Mehrwert für die Polizeiarbeit bieten konnte, insbesondere durch die Integration der Prognosen in Lagebilder und die Unterstützung der Einsatzplanung. Ein eindeutiger kriminalpräventiver

Effekt auf die tatsächliche Einbruchskriminalität ließ sich jedoch nicht empirisch belegen. Vielmehr verdeutlichte die Evaluation die methodischen Grenzen solcher Verfahren und betonte die Bedeutung einer kritischen Reflexion über ihren praktischen Nutzen.

Eine unabhängige, extern durchgeführte Evaluation der Wirksamkeit von *RADAR-iTE* liegt bislang nicht vor. Die bislang einzig verfügbare empirische Bewertung stammt aus einer internen Evaluationsstudie des BKA, die vor allem Fragen der Inhaltsvalidität (Vollständigkeit der erfassten relevanten Aspekte), Reliabilität (Zuverlässigkeit der Ergebnisse) und Trennschärfe (Stabilität der Unterscheidung zwischen den Risikokategorien) adressierte (Sonka et al., 2020, S. 389 f.). Eine systematische Wirkungsevaluation im engeren Sinne, also die Prüfung, ob und in welchem Umfang *RADAR-iTE* tatsächlich zur Verhinderung terroristisch motivierter Gewalttaten beiträgt, steht hingegen noch aus. Das BKA verweist in diesem Zusammenhang lediglich auf einen retrospektiven Testlauf, demzufolge *RADAR-iTE* auf der Grundlage der zu Anis Amri – dem Attentäter vom Berliner Breitscheidplatz – vorliegenden Informationen ein hohes Anschlagrisiko korrekt ausgewiesen habe (Münch, 2018).

Zusammenfassend lässt sich festhalten, dass die verfügbare wissenschaftliche Evidenz keine eindeutigen Belege für die kriminalpräventive Wirksamkeit von prognosebasierter Polizeiarbeit liefert. Die Prognosegenauigkeit der Systeme mag in vielen Fällen besser sein als zufällige Vorhersagen, doch ein kausaler Nachweis, dass der Einsatz von Predictive Policing zu einer substanziellen Reduzierung von Kriminalität führt, steht weitgehend aus. Diese ernüchternde Bilanz ist umso bedeutsamer, als sie im deutlichen Kontrast zum anfänglichen Hype um Predictive Policing steht und die Notwendigkeit unterstreicht, technologische Innovationen in der Polizeiarbeit einer kritischen empirischen Überprüfung zu unterziehen, bevor sie flächendeckend implementiert werden (vgl. Hauber, 2019).

5. Predictive Policing als soziotechnische Interaktion: Die Relevanz der Umsetzung von Kriminalitätsprognosen für erfolgreiche Prävention

Wie bereits betont, muss Predictive Policing als soziotechnische Interaktion verstanden werden, was impliziert, dass es kein rein technisches Phänomenen ist, das unabhängig von Menschen und der Organisation der Polizei operieren würde (Egbert & Leese, 2021). Eine ausschließlich technikzentrierte Betrachtung von Predictive Policing würde zu kurz greifen und sich vieler relevanter Analyseperspektiven und Fragestellungen verschließen – gerade, wenn es um Fragen der Kriminalprävention geht. Denn die präventive Wirksamkeit algorithmischer Prognoseverfahren entfaltet sich erst im Zusammenspiel mit polizeilichen Handlungslogiken, institutionellen Strukturen und der Akzeptanz der beteiligten Akteur:innen. Die technische Prognose stellt mithin nur einen ersten Schritt dar, dem die organisationale Verarbeitung, die Interpretation durch Polizist:innen und die praktische Umsetzung in konkrete Einsatzmaßnahmen folgen müssen (Egbert & Heimstädt, 2023). Strenggenommen ist aber schon die Prognose selbst kein rein technischer Prozess, sondern speist sich aus Daten, die menschliches und organisational geprägtes (Polizei-)Handeln aus der Vergangenheit reproduzieren (Egbert & Mann, 2021). Zudem sind die Prognosen regelmäßig so aufbereitet, z. B. was die visuelle Aufbereitung oder die Größe der Risikoräume angeht, dass sie mit größerer Wahrscheinlichkeit von Polizist:innen akzeptiert werden und von ihnen umsetzbar sind (Heimstädt & Egbert, 2025).

Die praktische Implementierung von Predictive Policing zeigt dabei, dass die Qualität der Umsetzung maßgeblich über dessen präventive Wirkung entscheidet. Selbst wenn Prognosealgorithmen statistisch reliable Vorhersagen trafen, was keineswegs ausgemacht ist (siehe oben), bleibt ihre praktische Relevanz davon abhängig, ob und wie die Prognosen von polizeilichen Akteur:innen aufgegriffen und in Handlungen übersetzt werden. Dies umfasst organisatorische Rahmenbedingungen wie die Verfügbarkeit personeller Ressourcen, die Integration der Prognosesoftware in bestehende Arbeitsabläufe sowie das Vertrauen der Polizeibeamt:innen in die algorithmischen Vorhersagen (Egbert & Esposito, in Begutachtung; Gluba, 2015; Sandhu & Fussey, 2021).

Ein zentrales Problem in der Praxis besteht darin, dass lokale Polizeidienststellen häufig nicht in der Lage sind, genügend Kräfte zur Verfügung zu stellen, um den präventiven Ansatz hinter Predictive Policing konsequent umzusetzen. Wenn prognostizierte Risikogebiete z. B. nur einmal pro Schicht kurz angefahren werden, tendiert die Wahrscheinlichkeit, dass potenzielle Täter:innen die Polizei sehen und dadurch abgeschreckt werden, gegen Null (Egbert & Esposito, in Begutachtung; Gerstner, 2017).

Darüber hinaus spielt die menschliche Interpretation eine entscheidende Rolle. Polizeibeamt:innen sind keine passiven Empfänger:innen algorithmischer Vorhersagen, sondern aktive Interpret:innen, die Prognosen im Kontext ihres Erfahrungswissens und ihrer lokalen Kenntnisse bewerten (sollen). Diese interpretativen Prozesse können dazu führen, dass Prognosen selektiv aufgegriffen, modifiziert, überhöht oder auch schlichtweg ignoriert werden (Egbert, 2021; Heimstädt & Egbert, 2025; Sandhu & Fussey, 2021).

Zusammenfassend lässt sich festhalten, dass Predictive Policing nicht auf eine rein technische Dimension reduziert werden kann. Seine präventive Wirksamkeit hängt maßgeblich davon ab, wie die generierten Prognosen innerhalb der Organisation verarbeitet, von Polizist:innen interpretiert und in konkrete Einsatzmaßnahmen übersetzt werden. Zugleich ist die durch organisationale und gesellschaftliche Faktoren beeinflusste Datenbasis ebenso zu berücksichtigen wie die auf spezifischen Kriminalitätstheorien beruhenden Algorithmen. Beides verdeutlicht, dass eine „neutrale“ technische Kriminalitätsprognose nicht existiert. Erfolgreiche Prävention durch Predictive Policing setzt daher nicht nur leistungsfähige Algorithmen voraus, sondern ebenso geeignete organisatorische Strukturen, ausreichende Ressourcen sowie die Akzeptanz und das Verständnis der beteiligten Akteur:innen. Diese Einsicht ist zentral für die Bewertung seiner Nützlichkeit als polizeiliche Präventionsstrategie.

6. Predictive Policing als ‚repressive Prävention‘

Wie ist Predictive Policing nun präventionstheoretisch einzuordnen? Wie oben bereits hervorgehoben, ist Predictive Policing zunächst als Form der situativen Kriminalprävention zu verstehen, indem durch polizeiliche Präsenz in besonders risikoträchtigen Orten das unentdeckte Agieren von möglichen Straftäter:innen, insbesondere Einbrecher:innen, erschwert werden soll.

Das heißt: Es handelt sich um Prävention durch verstärkte Kontrolle und Überwachung, die primär auf unmittelbare Risikominimierung abzielt, ohne die zugrundeliegenden Kriminalitätsursachen zu adressieren. Daher habe ich bereits an anderer Stelle (Egbert, 2022) dafür plädiert, die v. a. juristisch streng gehaltene Trennung zwischen Prävention und Repression aus analytischen Gründen aufzuheben, indem polizeiliche Maßnahmen zum Ziele der Kriminalprävention in nachhaltigere (präventive) und weniger nachhaltige (repressive) Ansätze unterschieden werden und Predictive Policing als letzteren, im Sinne einer ‚präpressiven‘ Maßnahme, zu charakterisieren (vgl. Singelstein, 2018).

Predictive Policing kann als Form ‚repressiver Prävention‘ verstanden werden, bei der polizeiliche Maßnahmen auf Basis algorithmisch generierter Prognosen erfolgen und primär auf die unmittelbare Minimierung von symptomatischen Risiken abzielen. Im Zentrum steht dabei nicht die Bearbeitung der eigentlichen Ursachen von Kriminalität, sondern die schnelle und gezielte Kontrolle von Räumen oder Personen, die als besonders gefährdet gelten. Die polizeiliche Präsenz in prognostizierten Risikogebieten dient vor allem der Abschreckung potenzieller Täter:innen und der kurzfristigen Verhinderung von Straftaten.

Die Empirie zeigt daher wenig überraschend, dass Predictive Policing häufig mit einer oberflächlichen Prävention verbunden ist: Die Maßnahmen konzentrieren sich allein auf die (kurzfristige) Hemmung krimineller Aktivitäten, ohne die sozialen, ökonomischen oder strukturellen Hintergründe zu berücksichtigen (Brayne, 2021; Egbert & Esposito, 2024, in Begutachtung; Gerstner, 2017; Sandhu & Fussey, 2021; Saunders et al., 2016). Dadurch besteht die Gefahr, dass Kriminalität lediglich verdrängt wird – sei es räumlich, zeitlich oder in Bezug auf die Art der Delikte – anstatt nachhaltig reduziert zu werden. Zudem können solche Interventionen unbeabsichtigte Nebenwirkungen haben, etwa die Einschränkung von Handlungsmöglichkeiten für unbeteiligte Bürger:innen durch ‚overpolicing‘ oder die Verstärkung von ‚chilling effects‘¹, wenn bestimmte Gebiete gemieden werden, um polizeilichen Kontrollen zu entgehen (Büchi et al., 2022).

1 *Chilling Effects* bezeichnen abschreckende Nebenwirkungen rechtlicher, institutioneller oder technischer Maßnahmen, durch die Individuen aus Angst vor negativen Konsequenzen auf die Ausübung legitimer Rechte verzichten. Gemeint ist insbesondere die Selbstbeschränkung von Verhalten, ohne dass formale Verbote oder unmittelbarer Zwang vorliegen müssen (z. B. Büchi et al., 2020).

Als Form repressiver Prävention ist Predictive Policing vor allem deshalb zu begreifen, weil es in einer ganz bestimmten Weise praktiziert wird. Es liegt nicht in der Logik prognosebasierter Polizeiarbeit selbst, zwangsläufig oberflächlich oder ausschließlich auf kurzfristige Abschreckung ausgerichtet zu sein. Das Wissen um erhöhte räumliche oder personenbezogene Kriminalitätsrisiken könnte ebenso bedeuten, dass Akteur:innen besser in der Lage wären, Prozesse anzustoßen, die weniger auf unmittelbare Abschreckung als vielmehr auf Unterstützung und langfristige Hilfsangebote zielen – wobei freilich zu diskutieren wäre, ob dies dann Aufgaben sind, die von der Polizei übernommen werden sollten.

Ein Beispiel für die potenzielle präventive Variabilität von Predictive Policing bringt die Polizei Chicago, die mit ihrer ‚Strategic Subject List‘ (SSL) zwischen 2013 und 2018 eines der weltweit bekanntesten personenbezogenen Predictive Policing-Tools genutzt hat. Das System basierte auf der Analyse sozialer Netzwerke und identifizierte Personen, die aufgrund ihrer Kontakte zu bekannten Straftäter:innen oder ihrer eigenen kriminalpolizeilichen Vorgeschichte, einer medizinischen Ansteckungslogik folgend, als besonders risikoträchtig eingestuft wurden (Heimstädt et al., 2021). Gefolgt wurde dabei empirisch-kriminologischer Forschung, die insbesondere in Bezug auf Gang-bezogene Schusswaffengewalt Forschung zu sozialen Netzwerkeffekten durchgeführt hat (Green et al., 2017; Papachristos et al., 2015). Auf dieser Liste wurden zu Beginn diejenigen 400 Personen geführt, die das höchste Risiko hatten, Täter:in, aber auch Opfer, einer gewaltbezogenen Straftat zu werden. Offiziell wurde die SSL eingeführt, um Personen mit erhöhtem Risiko für Partizipation an Gewaltdelikten zu identifizieren und ihnen gezielt, über sogenannte „custom notifications“ (per Brief oder Besuch), Präventionsangebote zu machen, etwa durch die Vermittlung sozialer Dienstleistungen (Tucek, 2018). In der Praxis überwogen jedoch oft abschreckende und konfrontative Interventionen: Polizei-beamt:innen suchten gelistete Personen unangekündigt auf, informierten sie über ihren Status auf der Liste und warnten sie vor rechtlichen Konsequenzen im Falle weiterer Straftaten.² Diese Besuche wurden von Be-

2 In der Dokumentation „Pre-Crime“ von Monika Hielscher und Matthias Heeder aus dem Jahr 2017 wird eine Polizistin des Chicago Police Department bei einem solchen „custom notification“-Besuchs begleitet. Ihr Wortlaut: „Für den Fall, dass Sie sich weiterhin kriminell betätigen, sollten Sie wissen, dass wir Sie verfolgen und mit der ganzen Härte des Gesetzes zur Rechenschaft ziehen werden“ (ab Min. 8:37). Der Film ist auf den Seiten der Bundeszentrale für politische Bildung abrufbar: <https://www.bpb.de/mediathek/video/299152/precrime/>.

troffenen häufig als Stigmatisierung und Bedrohung empfunden, was zu sozialer Ausgrenzung und Misstrauen gegenüber der Polizei führte – vor allem, da für diejenigen Personen, die auf der Liste aufgeführt waren, die Wahrscheinlichkeit stieg, dass sie für Gewalttaten, mit dem sie nicht in Verbindung standen, pauschal verantwortlich gemacht wurden, weil die Liste seitens der Polizei fälschlicherweise als Verdachtsgenerierungsinstrument missbraucht wurde (Saunders et al., 2016).

Das Beispiel Chicago verdeutlicht, dass prädiktive Systeme wie die SSL zwar theoretisch für unterschiedliche präventive Ansätze genutzt werden könnten, in der Praxis aber oft eine repressive, oberflächliche Prävention überwiegt. Die Balance zwischen Hilfe und Abschreckung bleibt schwierig, solange die Umsetzung nicht konsequent auf Unterstützung und nachhaltige Prävention ausgerichtet ist.

Insgesamt lässt sich also sagen, dass die prädiktive Bearbeitung von Kriminalität im Predictive Policing häufig mit einer präventiven Oberflächlichkeit und Kurzfristigkeit einhergeht, die zudem negative Auswirkungen auf die betroffenen Personen haben kann, da in ihr Leben in hemmender und bisweilen unterdrückender Weise – also repressiv – eingegriffen wird. Die Maßnahmen sind kurzfristig und reaktiv, während eine umfassende, ursachenorientierte Prävention in den Hintergrund tritt. Das birgt die Gefahr, dass langfristige Lösungen und nachhaltige Kriminalitätsbekämpfung aus dem Blick geraten.

Die geschilderten Eigenarten von Predictive Policing sind nicht in einem technikdeterministischen Sinne misszuverstehen, als gingen sie zwangsläufig und automatisch aus der eingesetzten Technik hervor. Methoden der Künstlichen Intelligenz in der (präventiven) Polizeiarbeit sind grundsätzlich polyvalent und können in sehr unterschiedlicher Weise eingesetzt werden – mit mal stärker, mal schwächer repressiven Effekten. Gleichwohl ist abschließend hervorzuheben, dass die den Prognosen – wie KI-Methoden insgesamt – zugrunde liegenden Verfahren einer primär quantitativen Logik folgen und insofern eine strukturelle Affinität zu repressiven, auf kurzfristige Abschreckung und oberflächliche Interventionen ausgerichteten Handlungspraktiken aufweisen (Krasmann, 2020; Seidensticker & Bode, 2022). Die für eine nachhaltige Präventionsarbeit erforderlichen, fall- und kontextsensitiven Informationen stehen in der Regel nicht zur Verfügung; eine tiefgehende, einzelfallorientierte Präventionspraxis, die langfristige Erfolge verspricht, läuft diesen Verfahren daher tendenziell zuwider.

7. Fazit und Ausblick

Predictive Policing steht exemplarisch für die tiefgreifende Transformation polizeilicher Arbeit im Zuge datengetriebener und zunehmend KI-gestützter Analysen. Wie im Beitrag gezeigt, verschiebt die prognosebasierte Polizeiarbeit nicht nur die zeitlichen Horizonte von Interventionen, sondern führt auch zu einer strukturellen Verschränkung präventiver und repressiver Logiken. Diese Entwicklung verweist darauf, dass sich polizeiliche Praxis immer deutlicher als präpressive Form der Gefahrenbearbeitung formiert – eine Praxis, in der kurzfristige Risikominimierung dominiert und langfristige Präventionsziele in den Hintergrund treten.

Mit der Einführung von Datenanalyseplattformen in die polizeiliche Arbeit, wie sie aktuell in Deutschland am Beispiel der Plattform Gotham der US-Firma Palantir Technologies breit diskutiert wird (Egbert, 2024), werden diese Tendenzen weiter verstärkt. Plattformarchitekturen führen vormals getrennte Datenbestände zusammen, ermöglichen deren operative Auswertung in Echtzeit und erweitern so das Anwendungsfeld prognostischer Verfahren. Dadurch intensiviert sich die Verwischung zwischen polizeirechtlichen und strafprozessualen Aufgabenbereichen, während die Steuerungslogiken der Polizei zunehmend technisch vermittelt und infrastrukturell eingebettet werden.

Gleichzeitig markiert das Aufkommen generativer KI, dass z. B. über die Artificial Intelligence Platform (AIP) von Palantir eng mit Datenanalyseplattformen verknüpft ist, einen neuen Entwicklungsschub, dessen Implikationen erst ansatzweise zu begreifen sind. Die Fähigkeit, Szenarien zu simulieren, komplexe Muster eigenständig zu identifizieren und Entscheidungen teilautomatisiert vorzubereiten, birgt das Potenzial, die Geschwindigkeit und Dichte polizeilicher Prognosearbeit weiter zu erhöhen. Damit wachsen jedoch auch die Risiken: Die Gefahr opaker, kaum noch nachvollziehbarer Entscheidungsgrundlagen steigt ebenso wie das Risiko, dass sich bestehende Tendenzen zur oberflächlichen, kurzfristigen Prävention weiter verfestigen.

Die Zukunft von Predictive Policing wird somit weniger eine Frage technologischer Leistungsfähigkeit als vielmehr eine Frage gesellschaftlicher und institutioneller Gestaltung sein. Entscheidend wird sein, ob es gelingt, algorithmische Prognosesysteme in Strukturen einzubetten, die

ihre Grenzen reflektieren, Fehlanreize minimieren und demokratische Kontrolle gewährleisten. Eine nachhaltige Kriminalitätsbekämpfung kann nur dann gelingen, wenn technische Innovationen konsequent mit einer Orientierung an sozialer Gerechtigkeit, rechtsstaatlichen Prinzipien und einer ernsthaften Bearbeitung der Ursachen von Kriminalität verbunden werden.

Literatur

- Balogh, D. (2016). Near Repeat-Prediction mit PRECOBS bei der Stadtpolizei Zürich. *Kriminalistik*, 70(5), 335–341.
- Belina, B. (2016). Predictive Policing. *Monatsschrift Für Kriminologie Und Strafrechtsreform*, 99(2), 85–100. <https://doi.org/10.1515/mks-2016-990201>
- BKA. (o. J.). RADAR-rechts. https://www.bka.de/DE/UnsereAufgaben/Deliktsbereiche/PMK/PMKrechts/RADAR/radar_node.html
- Bode, F., & Stoffel, F. (2023). Möglichkeiten und Grenzen polizeilicher Prognoseinstrumente am Beispiel des Projektes SKALA: Predictive Policing in Nordrhein-Westfalen (NRW). In L. B. Blum (Hrsg.), *Angewandte Data Science* (S. 29–50). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-39625-1_2
- Brayne, S. (2021). *Predict and surveil: Data, discretion, and the future of policing*. Oxford University Press.
- Büchi, M., Festic, N., & Latzer, M. (2022). The Chilling Effects of Digital Dataveillance: A Theoretical Model and an Empirical Research Agenda. *Big Data & Society*, 9(1), 20539517211065368. <https://doi.org/10.1177/20539517211065368>
- Büchi, M., Fosch-Villaronga, E., Lutz, C., Tamò-Larrieux, A., Velidi, S., & Viljoen, S. (2020). The chilling effects of algorithmic profiling: Mapping the issues. *Computer Law & Security Review*, 36, 105367. <https://doi.org/10.1016/j.clsr.2019.105367>
- Clarke, R. V. (2016). Situational crime prevention. In R. Wortley & M. Townsley (Hrsg.), *Environmental Criminology and Crime Analysis* (2. Aufl., S. 286–303). Routledge. <https://doi.org/10.4324/9781315709826>
- Egbert, S. (2021). Predictive Policing als Treiber rechtlicher Innovation? *Zeitschrift Für Rechtssoziologie*, 40(1–2), 26–51. <https://doi.org/10.1515/zfrs-2020-0002>
- Egbert, S. (2022). Predictive Policing: Die Digitalisierung als Präpressionstreiber. In M. Thüne, K. Klaas, & T. Feltes (Hrsg.), *Digitale Polizei: Einsatzfelder, Potenziale, Grenzen und Missstände* (S. 113–129). Verlag für Polizeiwissenschaft, Prof. Dr. Clemens Lorei.
- Egbert, S. (2024). Algorithmisches Polizieren in Deutschland: Von Predictive Policing zu plattformisierter Polizeiarbeit. *Zeitschrift für Jugendkriminalrecht und Jugendhilfe*, (2), 122–130.
- Egbert, S. (2025). Predictive Policing im deutschsprachigen Raum – Vergangenheit, Gegenwart, Zukunft. In W. Honekamp, S. Kemme, & J. Struck (Hrsg.), *Auswirkungen von Künstlicher Intelligenz auf die*

- zukünftige Polizeiarbeit (S. 301–318). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-48425-5_17
- Egbert, S., & Esposito, E. (2024). Algorithmic Crime Prevention. From Abstract Police to Precision Policing. *Policing and Society*.
- Egbert, S., & Esposito, E. (in Begutachtung). Deterrence or Prevention? The Challenges of Algorithmic Crime Prediction in Urban Spaces. *Urban Affairs*.
- Egbert, S., & Heimstädt, M. (2023). Algorithmic chains of translation: Predictive policing and the need for team-based ethnography. Routledge. In M. Avis, D. Marciniak, & M. Sapignoli (Hrsg.), *Situated AI – Global Ethnographies of New Technologies in Policing and Justice*. Routledge.
- Egbert, S., & Kornehl, K. (2022). Kommerzielle Software vs. Eigenentwicklung. Verbreitung und Ausgestaltung von Predictive Policing in Deutschland. *Kriminologisches Journal*, 54(2), 83–107.
- Egbert, S., & Leese, M. (2021). *Criminal Futures: Predictive Policing and Everyday Police Work* (1. Aufl.). Routledge. <https://www.routledge.com/Criminal-Futures-Predictive-Policing-and-Everyday-Police-Work/Egbert-Leese/p/book/9780367349264>
- Egbert, S., & Mann, M. (2021). Discrimination in Predictive Policing: The (Dangerous) Myth of Impartiality and the Need for STS Analysis. In A. Završnik & V. Badalič (Hrsg.), *Automating Crime Prevention, Surveillance, and Military Operations* (S. 25–46). Springer International Publishing. https://doi.org/10.1007/978-3-030-73276-9_2
- Endrass, J., Graf, M., & Rossegger, A. (2022). 20.5 Risikoeinschätzung: Beurteilung des Gewaltrisikos. In L. Rothenberger, J. Krause, J. Jost, & K. Frankenthal (Hrsg.), *Terrorismusforschung* (S. 389–398). Nomos Verlagsgesellschaft mbH & Co. KG. <https://doi.org/10.5771/9783748904212-389>
- Farrington, D. P., & Welsh, B. C. (2005). Randomized experiments in criminology: What have we learned in the last two decades? *Journal of Experimental Criminology*, 1(1), 9–38. <https://doi.org/10.1007/s11292-004-6460-0>
- Gerstner, D. (2017). Predictive Policing als Instrument zur Prävention von Wohnungseinbruchdiebstahl: Evaluationsergebnisse zum Baden-Württembergischen Pilotprojekt P4. edition iuscrim.
- Gluba, A. (2015). Predictive Policing – Chancen, Risiken und offene Fragen eines in Deutschland jungen Ansatzes. KI-Forum BKA 2015. <https://www.bka.de/SharedDocs/Downloads/DE/Publikationen/ForumKI/ForumKI2015/kiforum2015GlubaLangfassung.html>
- Gluba, A. (2017). Der Modus Operandi bei Fällen der Near Repeat-Victimisation – Ergebnisse einer empirischen Studie. *Kriminalistik*,

- 71(6), 369–374.
- Goertz, S. (2020). Terrorismusabwehr: Zur aktuellen Bedrohung durch den islamistischen Terrorismus in Deutschland und Europa. Springer Fachmedien Wiesbaden. <https://doi.org/10.1007/978-3-658-30672-4>
- Goertz, S. (2022). Innere Sicherheit - von A bis Z: Die wichtigsten Begriffe für Studium und Ausbildung. Richard Boorberg Verlag GmbH & Co KG. <https://doi.org/10.5771/9783415072824>
- Green, B., Horel, T., & Papachristos, A. V. (2017). Modeling Contagion Through Social Networks to Explain and Predict Gunshot Violence in Chicago, 2006 to 2014. *JAMA Internal Medicine*, 177(3), 326. <https://doi.org/10.1001/jamainternmed.2016.8245>
- Hauber, J. (2019). Postfaktizität und Predictive Policing. In H.-J. Lange & M. Wendekamm (Hrsg.), *Postfaktische Sicherheitspolitik* (S. 191–209). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-27281-4_10
- Heimstädt, M., & Egbert, S. (2025). Actionable predictions: How designers of algorithmic systems calibrate criminal futures. *Big Data & Society*, 12(2). <https://doi.org/10.1177/20539517251340636>
- Heimstädt, M., Egbert, S., & Esposito, E. (2021). A Pandemic of Prediction: On the Circulation of Contagion Models between Public Health and Public Safety. *Sociologica*, 1-24 Pages. <https://doi.org/10.6092/ISSN.1971-8853/11470>
- Hofmann, H. (2020). *Predictive Policing: Methodologie, Systematisierung und rechtliche Würdigung der algorithmusbasierten Kriminalitätsprognose durch die Polizeibehörden* (1st ed). Duncker & Humblot.
- Johnson, S. D., & Bowers, K. J. (2014). Near Repeats and Crime Forecasting. In G. Bruinsma & D. Weisburd (Hrsg.), *Encyclopedia of Criminology and Criminal Justice* (S. 3242–3254). Springer New York. https://doi.org/10.1007/978-1-4614-5690-2_210
- Kaufmann, M., Egbert, S., & Leese, M. (2019). Predictive Policing and the Politics of Patterns. *The British Journal of Criminology*, 59(3), 674–692. <https://doi.org/10.1093/bjc/azy060>
- Kemme, S. (2025). Zwischen Effizienz und Recht auf Fairness: Wie algorithmische Polizeiarbeit Diskriminierung verstärken kann. In W. Honekamp, S. Kemme, & J. Struck (Hrsg.), *Auswirkungen von Künstlicher Intelligenz auf die zukünftige Polizeiarbeit* (S. 281–299). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-48425-5_16
- Krasmann, S. (2020). The logic of the surface: On the epistemology of algorithms in times of big data. *Information, Communication & Society*, 23(14), 2096–2109. <https://doi.org/10.1080/136911>

- 8X.2020.1726986
- KrimZ. (2025). Tätigkeitsbericht 2024. https://www.krimz.de/fileadmin/dateiablage/download/Taetigkeitsbericht_2024.pdf
- KrimZ. (o. J.a). SMART - Weiterentwicklung der individuellen Risikoanalyse für extremistische Gewalttaten. <https://www.krimz.de/forschung/aktuelle-projekte/smart.html>
- KrimZ. (o. J.b). Weiterentwicklung der Risikobewertungsinstrumente RADAR-iTE und RADAR-rechts für den Haftkontext. <https://www.krimz.de/forschung/pmk/radar-haft.html>
- Leese, M. (2024). Staying in control of technology: Predictive policing, democracy, and digital sovereignty. *Democratization*, 31(5), 963–978. <https://doi.org/10.1080/13510347.2023.2197217>
- Leese, M., & Pollozek, S. (2023). Not so fast! Data temporalities in law enforcement and border control. *Big Data & Society*, 10(1), 205395172311641. <https://doi.org/10.1177/20539517231164120>
- LKA Niedersachsen. (2018). PreMAP – Predictive Policing in Niedersachsen. Bericht zur Bewertung der ersten Projektphase. <https://www.lka.polizei-nds.de/startseite/kriminalitaet/forschung/premap/predictive-policing-in-niedersachsen-das-projekt-premap-114083.html>
- LKA Niedersachsen. (2021). PreMAP – Prüfung einer Erweiterung und Modifizierung des Prognoseansatzes. LKA Niedersachsen. <https://www.lka.polizei-nds.de/startseite/kriminalitaet/forschung/premap/predictive-policing-in-niedersachsen-das-projekt-premap-114083.html>
- LKA NRW. (2018). Projekt SKALA. Abschlussbericht [Projektabschlussbericht].
- LKA NRW & GISS. (2018). Kooperative Evaluation des Projektes SKALA. https://lka.polizei.nrw/sites/default/files/2018-06/160430_Evaluationsbericht_SKALA.pdf
- Meyer, S. (2017). Kriminalwissenschaftliche Prognoseinstrumente im Tatbestand polizeilicher Vorfeldbefugnisse. *JuristenZeitung*, 72(9), 429–439. <https://doi.org/10.1628/002268817X14907128992671>
- Münch, H. (2018). „Die Gefährderzahl hat sich verfünffacht“ (M. Fiedler & F. Jansen) [Tagesspiegel]. <https://www.tagesspiegel.de/politik/bka-leiter-holger-muench-die-gefaehrderzahl-hat-sich-verfuenffacht/23774846.html>
- National Academies of Sciences, Engineering, and Medicine (Hrsg.). (2025). *Law Enforcement Use of Predictive Policing Approaches: Proceedings of a Workshop* (1st ed). National Academies Press.
- Okon, G. (2020). Von der Idee zur Praxis—Machbarkeitsstudie Predictive

- Policing bei der Bayerischen Polizei 2014/2015. In F. Bode & K. Seidensticker (Hrsg.), *Predictive Policing. Eine Bestandsaufnahme für den deutschsprachigen Raum* (S. 147–166). Verlag für Polizeiwissenschaft.
- Papachristos, A. V., Wildeman, C., & Roberto, E. (2015). Tragic, but not random: The social contagion of nonfatal gunshot injuries. *Social Science & Medicine*, 125, 139–150. <https://doi.org/10.1016/j.socscimed.2014.01.056>
- Petersen, T. S. (2024). Situational crime prevention or getting to the root causes of crime?*. In S. J. Holmen, T. S. Petersen, & J. Ryberg, *Crime Prevention by Exclusion* (1. Aufl., S. 108–123). Routledge. <https://doi.org/10.4324/9781003480679-7>
- Pett, A., & Gluba, A. (2017). Das Potenzial von Polizeipräsenz für Maßnahmen im Sinne des Predictive Policing. *Die Polizei*, 108(11), 323–330.
- Polizei Bayern. (2021, Oktober 27). Predictive Policing bei der Bayerischen Polizei. www.polizei.bayern.de/aktuelles/pressemitteilungen/018804/index.html
- Sandhu, A., & Fussey, P. (2021). The ‘uberization of policing’? How police negotiate and operationalise predictive policing technology. *Policing and Society*, 31(1), 66–81. <https://doi.org/10.1080/10439463.2020.1803315>
- Saunders, J., Hunt, P., & Hollywood, J. S. (2016). Predictions put into practice: A quasi-experimental evaluation of Chicago’s predictive policing pilot. *Journal of Experimental Criminology*, 12(3), 347–371. <https://doi.org/10.1007/s11292-016-9272-0>
- Scheu, L. D. (2025). Heuristik sicherheitspolitischer Entscheidungsprozesse: Ein multiperspektivischer Blick auf die Begründung von Maßnahmen gegen den Islamismus in Deutschland. In PRIF Working Paper (S. No. 65). Peace Research Institute Frankfurt. <https://doi.org/10.48809/PRIFWP65>
- Schweer, T. (2015). „Vor dem Täter am Tatort“ – Musterbasierte Tatortvorhersagen am Beispiel des Wohnungseinbruchs. *Die Kriminalpolizei*, 32(1), 13–16.
- Schweer, T. (2020). „Am Anfang war die Stecknadel“: Predictive Policing als Teil moderner Polizeiarbeit. In F. Bode & K. Seidensticker (Hrsg.), *Predictive Policing. Eine Bestandsaufnahme für den deutschsprachigen Raum* (S. 129–145). Verlag für Polizeiwissenschaft.
- Seidensticker, K. (2021). SKALA - Predictive Policing in North Rhine-Westphalia. *European Law Enforcement Research Bulletin*, 21(Summer), 47–60.

- Seidensticker, K. (2022). Predictive Policing. In H. Diebel-Fischer, L. Hellmig, & M. Tischler (Hrsg.), *Technik und Verantwortung im Zeitalter der Digitalisierung: Beiträge zur Ringvorlesung im Wintersemester 2020/2021* (S. 193–218). Universität Rostock. https://doi.org/10.18453/rosdok_id00003993
- Seidensticker, K., & Bode, F. (2022). Good Policing in Times of Abstract Police. In J. Terpstra, R. Salet, & N. Fyfe (Hrsg.), *The abstract police: Critical reflections on contemporary change in police organisations* (S. 169–182). Eleven International.
- Seidensticker, K., Bode, F., & Stoffel, F. (2018). Predictive Policing in Germany [Working Paper]. <http://nbn-resolving.de/urn:nbn:de:bsz:352-2-14sbvox1ik0z06>
- Seidensticker, K., & Schwarz, K. (2022). Using Forecasting Methods on Crime Data: The SKALA Approach of the State Office for Criminal Investigation of North Rhine-Westphalia. *The 8th International Conference on Time Series and Forecasting*, 39. <https://doi.org/10.3390/engproc2022018039>
- Singelstein, T. (2018). Predictive Policing: Algorithmenbasierte Straftatprognosen zur vorausschauenden Kriminalintervention. *Neue Zeitschrift für Strafrecht*, 38(1), 1–9.
- Sommerer, L. M. (2017). Geospatial Predictive Policing – Research Outlook & A Call For Legal Debate. *Neue Kriminalpolitik*, 29(2), 147–164. <https://doi.org/10.5771/0934-9200-2017-2-147>
- Sommerer, L. M. (2020). Personenbezogenes Predictive Policing: Kriminalwissenschaftliche Untersuchung über die Automatisierung der Kriminalprognose (1. Auflage). Nomos.
- Sonka, C., Meier, H., Rossegger, A., Endrass, J., Profes, V., Witt, R., & Sadowski, F. (2020). RADAR-ITE 2.0: Ein Instrument des polizeilichen Staatsschutzes. *Kriminalistik*, 74(6), 386–392.
- Trunk, D., & Simmert, S. (2020). Rechtliche und interdisziplinäre Aspekte polizeilicher Gefahrenabwehr. Das Verfahren RISKANT und das standardisierte Instrument RADAR-ITE 2.0. [Abschlussbericht]. Fachhochschule Polizei Sachsen-Anhalt. <https://www.tib.eu/de/suchen/id/TIBKAT:1750800365/>
- Tucek, A. (2018). Constraining Big Brother: The Legal Deficiencies Surrounding Chicago's Use of the Strategic Subject List. *The University of Chicago Legal Forum*, 2018(1), 427–460.

Vertiefende Literatur:

- Sarah Brayne (2021): Predict and Surveil. Oxford: Oxford University Press.
Simon Egbert & Matthias Leese (2021): Criminal Futures. Abingdon/London: Routledge.
Lucia Sommerer (2020): Personenbezogenes Predictive Policing. Baden-Baden: Nomos.

Mediathek



Dokumentation „Pre-Crime“



Beitrag Ulrike Heitmüller, heise online: Missing Link: Predictive Policing – die Kunst, Verbrechen vorherzusagen:



Informationen des LKA NRW zum System SKALA, inkl. Kurzvideos:



Dr. Simon Egbert, Wissenschaftlicher Mitarbeiter an der Universität Bielefeld im ERC-Forschungsprojekt „The Future of Prediction“ sowie Projektleiter eines DFG-geförderten Forschungsprojekts zu polizeilichen Bodycams im internationalen Projektverbund ‚Visions of Policing‘. Promotion 2018 an der Universität Hamburg mit der Arbeit „Diskurs und Materialität. Eine Dispositivanalyse des Drogentests“. Aktuelle Forschungsschwerpunkte: Soziologie der Prädiktion, Algorithmisierung der Polizei, Visuelle Technologien, Datenanalyseplattformen.

»Empirische Beispiele, in denen generative KI bereits Bestandteil extremistischer Kommunikation ist, ebenso wie Überlegungen, generative KI in der Prävention zu nutzen, werfen eine grundsätzliche Frage auf: Unter welchen Bedingungen kann eine KI eine mit menschlichen Kommunikator:innen vergleichbare Wirkung auf die Rezipient:innen haben?«

**Christian Büscher, Isabel Kusche, Tim Röllner und
Alexandros Gazos**

**Christian Büscher, Isabel Kusche, Tim Röller,
Alexandros Gazos**

Die Beiträge von KI zu extremistischer Kommunikation und Chancen KI-basierter Prävention

1. Einleitung

Das Innovationspotenzial extremistischer und terroristischer Akteur:innen (Personen, Gruppierungen, Organisationen) in Bezug auf Technologie findet gegenwärtig ein erhebliches Maß an Beachtung, insbesondere im Kontext der Verbreitung von ubiquitären Technologien wie der generativen Künstlichen Intelligenz, welche augenscheinlich integraler Bestandteil diverser gesellschaftlicher Prozesse wird (Dickel, 2025; Kusche, 2023). Es gilt zu untersuchen, inwiefern Künstliche Intelligenz möglicherweise extremistischen oder terroristischen Handlungen Vorschub leisten oder diese überhaupt erst ermöglichen kann (Brundage et al., 2018; Klinkhammer, 2024; Nelu, 2024).

In Bezug auf die begriffliche Einordnung von Techniknutzung und Extremismus wird vorzugsweise der Terminus „Dual Use“ herangezogen (vgl. jüngst: Grinbaum & Adomaitis, 2024), welcher ursprünglich das Potenzial einer Technologie für sowohl zivile als auch militärische Anwendungen bezeichnete (Forge, 2010; Mahfoud et al., 2018), gegenwärtig jedoch oftmals zur Differenzierung zwischen „benevolenten“ und „malvolenten“ Verwendungsarten von Technologie dient (Oltmann, 2015). Der Begriff suggeriert, dass verantwortungsbewusste Entwickler:innen und Vermarkter:innen das schädliche Potenzial einer Technologie leicht erkennen können. Diese Annahme erscheint jedoch weniger plausibel, wenn man den vielseitigen, allgegenwärtigen und formbaren Charakter von Informations- und Kommunikationstechnologien bedenkt, z. B.

Social-Media-Plattformen, kryptografische Techniken, und als jüngste Entwicklung generative Künstliche Intelligenz (Büscher & Kusche, 2025; Montasari, 2024). Die breite Verfügbarkeit von Erfindungen, die das Leben der Menschen verbessern können, begünstigt dabei auch die innovative Nutzung durch extremistische und terroristische Akteur:innen (Cronin, 2020). Angesichts der wachsenden Kommerzialisierung und gesellschaftlichen Verbreitung von Anwendungen generativer KI ist damit zu rechnen, dass diese auch für extremistische Gruppen und ihre Zwecke neue Möglichkeiten bieten.

Wir werden daher in diesem Beitrag ersten Spuren nachgehen, die nahelegen, wie generative KI in der Kommunikation von Extremist:innen und mit Extremist:innen eine Rolle spielt und spielen kann. Ersteres zielt auf die Analyse bereits beobachteter KI-Nutzung ab, letzteres auf die Potenziale für präventive Ansprachen an Extremist:innen. Dabei greifen wir auf Ideen der Kommunikationstheorie zurück, um eine Systematisierung möglicher Fälle vorzunehmen. Mithilfe der so entwickelten Heuristik sortieren und analysieren wir exemplarisch aktuelle Fälle der KI-Nutzung, die wir in der Fachliteratur und in Aussagen aus im Rahmen des Projektes MOTRA-Technologiemonitoring¹ (MOTRA-TM) durchgeführten Expert:inneninterviews vorfinden. Anhand der auf diesem Weg gewonnenen Erkenntnisse, diskutieren wir abschließend mögliche Implikationen für die Prävention.

1 Dieser Beitrag basiert auf Forschung, die mit Mitteln des Bundesministeriums für Forschung, Technologie und Raumfahrt (BMFTR), des Bundesministeriums des Innern (BMI) und des Bundesministeriums für Bildung, Familie, Senioren, Frauen und Jugend (BMBFSFJ) im Rahmen des Verbundprojekts MOTRA gefördert wurde. Das MOTRA Technologiemonitoring (Förderkennzeichen 13N17261) ist Teil dieses Forschungsverbundes.

2. Die soziotechnische Dimension: Generative KI in der sozialen Interaktion

Künstliche Intelligenz ist zunächst ein Forschungsgebiet, das sich mit der Automatisierung kognitiver menschlicher Aktivitäten befasst (Humm et al., 2021).² Maschinelles Lernen ist ein Teilgebiet der Künstlichen Intelligenz, das bei der Automatisierung menschlicher Aktivitäten darauf setzt, Maschinen beizubringen, aus Daten zu lernen und selbstständig Aufgaben zu erledigen. Anders als bei anderer Computersoftware erfordern die entsprechenden Algorithmen nach erfolgtem initialen Training keine expliziten Anweisungen durch den Menschen mehr. Auf der Grundlage einer bestimmten Anzahl von Beispielen aus einer Datenbank erkennen sie stattdessen Muster und lernen implizite Regeln (Brundage et al., 2018; Schick, 2018).

Die KI-Forschung unterschied lange zwischen „enger“ oder „schwacher“ KI, die eng umgrenzte Aufgaben erfüllt, und „allgemeiner“ oder „starker“ KI, die jede Aufgabe analog zur menschlichen Intelligenz oder sogar besser erfüllen könnte (Fjelland, 2020; Schroeter, 2020, S. 8; Searle, 1980). Auch wenn „Allgemeine Künstliche Intelligenz“ nach wie vor dem Bereich von Science-Fiction angehört, ist sie mit der Entwicklung sogenannter *Foundation Models* in den vergangenen Jahren stärker in den Bereich des Möglichen gerückt. Foundation Models beruhen ebenfalls auf Varianten des maschinellen Lernens und werden anhand enorm großer Mengen unterschiedlicher und unspezifischer Daten vortrainiert. Infolgedessen weisen sie eine erhebliche Bandbreite möglicher Verwendungsweisen auf, sodass sie viele unterschiedliche Aufgaben verhältnismäßig erfolgreich lösen können, auch wenn sie für diese gar nicht explizit trainiert wurden. Dazu gehören beispielsweise so unterschiedliche Dinge wie die Überprüfung von Programmcode, die Sequenzierung von Proteinen oder die Erzeugung von Bildern aus Texteingaben. Ferner kann die aufgabenspezifische Leistungsfähigkeit dieser Modelle mit relativ geringem Aufwand durch Maßnahmen wie Finetuning oder Prompt Engineering noch gesteigert werden (Schneider et al., 2024).

2 Die Europäische Union definiert in Artikel 3 der KI-Verordnung ein KI-System als „ein maschinengestütztes System, das für einen in unterschiedlichem Grade autonomen Betrieb ausgelegt ist und das nach seiner Betriebsaufnahme anpassungsfähig sein kann.“ Dort heißt es weiter, dass ein KI-System „aus den erhaltenen Eingaben für explizite oder implizite Ziele ableitet, wie Ausgaben wie etwa Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen erstellt werden, die physische oder virtuelle Umgebungen beeinflussen können.“ (Verordnung (EU) 2024/1689 des Europäischen Parlaments und des Rates vom 13. Juni 2024).

Auf der Ebene konkreter praktischer Anwendungen Künstlicher Intelligenz ist es zur Orientierung hilfreich, einen Unterschied zwischen prädiktiven und generativen KI-Anwendungen zu machen. Nach Narayanan und Kapoor (2024) müssen diese beiden Kategorien sowohl technisch als auch gesellschaftlich-regulatorisch differenziert betrachtet werden. Prädiktive KI wird als auf spezifische, eng umrissene Aufgaben ausgerichtet beschrieben, während generative KI als breit einsetzbare, auf Mustererkennung und -erzeugung basierende Technologie charakterisiert wird. Die Autoren warnen davor, beide Typen in der öffentlichen und politischen Debatte zu vermischen, da dies zu Missverständnissen über Risiken, Nutzen und Regulierungsbedarf führt. Generative KI dominiert die öffentliche Debatte durch ihre sichtbaren, oft spektakulären Ergebnisse. Prädiktive KI wirkt im Hintergrund, hat aber oft größere Auswirkungen auf das Leben der Menschen (Narayanan & Kapoor, 2024).

Im Zusammenhang mit Radikalisierung und Extremismus verdienen Anwendungen generativer KI derzeit besondere Aufmerksamkeit, zum einen wegen ihrer rasanten Verbreitung innerhalb der letzten drei Jahre, zum anderen wegen ihrer besonderen Bedeutung für die digitale Kommunikation. Ende 2022 wurde der Chatbot ChatGPT der Firma OpenAI der Öffentlichkeit vorgestellt (Hu, 2023) und machte die Möglichkeiten generativer KI erstmals einem breiten Publikum bewusst. Der Chatbot konnte Fragen über alle Arten von Themen eloquent und häufig sachlich korrekt beantworten. Gleichwohl sind solche, auf großen Sprachmodellen beruhende Anwendungen anfällig für Halluzinationen, das heißt für ein Antwortverhalten, bei dem falsche oder unsinnige Behauptungen als Fakten präsentiert werden (Zhang et al., 2023). Große Sprachmodelle generieren einen Text, indem sie das nächste Wort in einer Sequenz von Worten vorhersagen. Sie operieren anhand eines an Wahrscheinlichkeiten – und nicht an Sinnverstehen – ausgerichteten Selektionsvorgangs (Jäkel, 2025). „The problem is, large language models are so good at what they do that what they make up looks right most of the time“ (Heaven, 2024, S. 21). Trotz dieser Probleme ist die Faszination für die Möglichkeiten solcher Modelle enorm. Andere Unternehmen, wie z. B. Google, Meta, X oder Mistral, haben innerhalb kurzer Zeit mit eigenen KI-Modellen und Chatbots nachgezogen.

Im Bereich der bildgenerierenden KI kam es zu einem weiteren bedeutenden Fortschritt. Seit Jahren existieren Anwendungen wie FaceApp, die es Nutzer:innen ermöglichen, sich selbst oder andere unter Zuhilfenah-

me von KI-basierter Bildtransformation hochgeladener Fotografien älter oder jünger erscheinen zu lassen oder das Geschlecht zu verändern. Seit 2022 gibt es mit Dall-E, Midjourney oder Stable Diffusion nun Bildgeneratoren, die die Generierung von Bildmotiven in diversen Stilrichtungen, von abstrakt bis fotorealistisch, auf Grundlage einer textuellen Spezifikation gestatten und eine Revolutionierung der KI-gestützten Bilderzeugung darstellen (Bucher & Roy, 2023). Unter den zahlreichen weiteren Text-zu-Bild-Modellen, die heute verfügbar sind, ermöglichen manche eine vollständig kostenfreie Nutzung. 2024 ist mit Sora von OpenAI auch ein Videogenerator vorgestellt und seitdem Schritt für Schritt einem breiten Publikum zugänglich gemacht worden.

Die Technologie wird aufgrund der wachsenden Kommerzialisierung und Verbreitung von KI-Anwendungen auch für extremistische Akteur:innen attraktiver, die sie für bösartige Zwecke verwenden möchten. Dies ist auch den Entwickler:innen nicht entgangen. Vor der Veröffentlichung der zweiten Version von Sora verwendete Open AI ein „Red Team“, um das Modell auf mögliche Risiken hin zu testen (OPEN AI, 2025). Auf diese Weise wird versucht, zu verhindern, dass KI zum Beispiel für Deepfakes (Kusche, 2024) oder Desinformation verwendet wird. Das Bestreben von OpenAI, „adversarial testing“ durchzuführen, deutet darauf hin, dass die Entwickler:innen sich inzwischen der schädlichen Möglichkeiten ihrer Erfindungen bewusst sind. Gleichzeitig bleibt aber unklar, welche Konsequenzen das angesichts des wachsenden Konkurrenzdrucks zwischen verschiedenen Anbietern tatsächlich haben wird.

3. Kommunikationstheoretischer Rahmen

Für die Analyse des soziotechnischen Zusammenhangs zwischen generativer KI und Extremismus bedienen wir uns einer Heuristik aus der Kommunikationstheorie. Sie gründet auf der theoretischen Prämisse, dass Kommunikation (unter Anwesenden, in sozialen Netzwerken, in Organisationen, aber auch in Funktionsbereichen der Gesellschaft) nicht als Prozess der Übertragung von Informationen begriffen werden kann. In seiner Theorie sozialer Systeme misst Niklas Luhmann (2005) dem Begriff der Kommunikation eine zentrale Bedeutung bei und charakterisiert Kommunikation als eine emergente Qualität. Im vorliegenden Kontext impliziert Emergenz, dass erfolgreiche Kommunikation sich nicht auf die interagie-

renden Personen reduzieren lässt. Kommunikation ist nicht unabhängig von Personen mit ihrem Wissen und ihren Intentionen, aber eigenständig und eigenwillig. Im Allgemeinen ist Kommunikation Ergebnis einer dreifachen Selektion. Einfach ausgedrückt, muss bei jedem Kommunikationsakt selektiert werden, was gesagt werden soll, wie es gesagt wird und letztlich, ob und wie es verstanden wird. Luhmann spricht von einer dreifachen Selektion von Information, Mitteilung und Verstehen (Luhmann, 2005).

In Bezug auf eine simple Gesprächssituation bedeutet dies, dass das, was der bzw. die Zuhörer:in wahrnimmt, nicht identisch mit dem ist, was der oder die Sprecher:in zum Ausdruck bringt. Immer bleibt die Person, mit der man spricht, in ihren Gedanken unerreichbar. Das Nachvollziehen des Selektionsprozesses beim Verstehen ist nicht möglich, nur das Ergebnis kann beobachtet werden. Demzufolge erfolgt die Auswahl einer Information oder einer Mitteilung anhand von Erwartungen darüber, was beim Gegenüber welche Art des Anschlusses wahrscheinlich macht. Man hat also nur (oder zumindest mehr) Kontrolle über die eigenen Selektionen und nicht über die des anderen. Da dies auf beide Seiten zutrifft, gestaltet sich der Verlauf eines Gesprächs bereits bei einfachen Interaktionen als schwierig zu kontrollieren und ist oft sehr unerwartet. Es ist nicht sicher, ob die Information verstanden wird oder ob die Art der Mitteilung angemessen ist, was zu Unsicherheit darüber führt, wie sie aufgenommen wird. Es kann gut gehen oder nicht. Kommunikation verläuft in diesem Sinne kontingent: nicht unmöglich, aber auch nicht zwingend notwendig (Luhmann, 1984, S. 47).

Die Annahme, dass alle an Kommunikation Beteiligten Menschen sind, kann in Zeiten generativer KI nicht mehr als selbstverständlich erfüllt gelten. Begreift man Kommunikation als dreifache Selektion von Information, Mitteilung und Verstehen, sind dazu, allgemein betrachtet, zwei Adressen erforderlich: Die Selektion einer Information und die Wahl eines passenden Mitteilungsverhaltens werden der ersten Adresse zugerechnet, das Verstehen der mitgeteilten Information der zweiten Adresse. In der Kommunikation wird Verstehen erst erkennbar, wenn mit einer weiteren Kommunikation an die mitgeteilte Information angeschlossen wird. Die betreffende Anschlusskommunikation kann zustimmend sein, d. h., die vorangegangene Kommunikation annehmen und zur Prämisse der weiteren Kommunikation machen. Sie kann aber auch ablehnend ausfallen, also etwa einen Vorschlag zurückweisen oder eine Behauptung abstreiten.

Spätestens seit der Verbreitung von Anwendungen wie ChatGPT gehen soziologische Studien davon aus, dass nicht nur Menschen, sondern auch Maschinen als Adressen von Kommunikation fungieren können (Baecker, 2011; Dickel, 2025; Esposito, 2022). Generative KI produziert Sprache und/oder Bilder, die hinreichend überraschend, interessant, informativ oder unterhaltsam sind, um Kommunikation am Laufen zu halten. Die weitere Überprüfung der Mitteilung auf ihren generativen Ursprung – der Selektion der Informationen – ist in der Dynamik von Kommunikationsflüssen in den meisten Fällen nicht zu leisten: „Normally there is not the time and the motivation to question whether it is a machine or a human being“ (Esposito, 2022, S. 108). Dies ist keine Deformation des Sozialen durch die Einmischung von Maschinen, sondern der soziale Regelfall, wie ihn die Organisationssoziologen March und Simon (March & Simon, 1993, S. 186) anhand des Prinzips der *Unsicherheitsabsorption* beschrieben haben. Die Herleitung von Informationen kann in den allermeisten sozialen Situationen nicht oder nur in einem sehr begrenzten Maße kontrolliert werden.

Das heißt allerdings nicht, dass es für die Bereitschaft zur Annahme einer Kommunikation grundsätzlich keinen Unterschied macht, ob die Adresse, der die mitgeteilte Information zuzurechnen ist, ein Mensch ist oder eine Maschine. Ob dieser Unterschied einen Unterschied macht, ist vielmehr eine empirische Frage. Sascha Dickel definiert echte, allgemeine Künstliche Intelligenz deshalb eher als soziale Frage anstatt als technisches Problem. Generative Künstliche Intelligenz eignet sich als Kommunikationspartner, beispielsweise in Form von Chatbots, wenn bei der Rezeption einer Mitteilung darauf verzichtet wird, Indikatoren für ihre maschinelle Herkunft zu identifizieren (Dickel, 2025, S. 201). Ob und wie generative KI Kommunikation verändert, zeigt sich maßgeblich an ihrem Einfluss auf die Annahme oder Ablehnung von Kommunikation. Es geht darum, ob es für die Annahme oder Ablehnung von Kommunikationsangeboten einen Unterschied macht, wenn die Adresse, der Information und Mitteilung zugerechnet wird, eine Maschine und nicht ein Mensch ist.

Für das Radikalisierungsgeschehen relevant ist also beispielsweise, ob Deepfakes trotz erkennbarer Künstlichkeit in spezifischen Communitys einen Informationswert gewinnen und radikalisierende Anschlussfähigkeit präparieren – der Unterschied also keinen Unterschied macht.

Ebenso relevant ist es, wenn „Radikalisierungspipelines“ mit nicht auf den ersten Blick als generiert erkennbaren extremen Inhalten geflutet werden, die mögliche Rezipient:innen nur mit erhöhtem kognitiven Prüfaufwand erkennen könnten – die Unterscheidung also für bedeutsam gehalten wird, aber nicht mehr routinemäßig getroffen werden kann. Im Zentrum der weiteren Entwicklung von generativer KI und der damit verbundenen Folgen steht allgemein die Frage, ob sich die Unterscheidung zwischen artifizierter und menschlicher Kommunikation auflöst, oder genauer gesagt, in welchen Konstellationen sie weiterhin eine Rolle spielt (Dickel, 2025).

Aus der Frage nach dem Unterschied, den die Unterscheidung zwischen Mensch und Maschine in Kommunikationsflüssen, an denen generative KI beteiligt ist, noch macht, lassen sich theoretische Konstellationen ableiten, in denen diese Frage jeweils anders zu beantworten ist. Diese können als heuristisches Schema verwendet werden, um in Expert:inneninterviews, die im Rahmen von MOTRA-TM durchgeführt wurden, sowie in der Sekundärliteratur typische Entwicklungen bezüglich der Rolle zu identifizieren, die KI in der Kommunikation von Extremist:innen einerseits und der Präventionsarbeit andererseits spielen kann (Tabelle 1).

Konstellation	Adresse 1: Information, Mitteilung	Adresse 2: Verstehen	Beispiel(e)
1	KI	Mensch	Gekennzeichnete oder erkannte Chatbot, Bildgenerator etc.
2	Mensch oder KI	Mensch	Nicht gekennzeichnete/erkannte Chatbot, Bildgenerator etc.
3	Mensch	KI	Jailbreaking von Chatbots oder Bildgeneratoren
4	Mensch oder KI	KI	Automatisierte Moderation und Faktenchecks

Tabelle 1: Theoretische Zurechnungskonstellationen in der Kommunikation mit generativer KI

Wir beschränken uns für die Entwicklung des heuristischen Schemas auf die Betrachtung einer einzelnen Kommunikation, wohl wissend, dass jene Adresse, die eine mitgeteilte Information versteht, anschließend selbst der ersten oder einer dritten Adresse eine weitere Information mitteilen kann und solche Verkettungen von Kommunikationen der

gesellschaftliche Normalfall sind. Die Fokussierung erlaubt es aber, vier Grundkonstellationen zu unterscheiden, in denen KI an Kommunikation beteiligt ist und der Unterschied zwischen KI und Mensch entweder relevant ist oder nicht.

Zu Konstellation 1: Es kann ein von einer KI generierter Output – sei er textförmig, ein Bild, Video oder ein Podcast – einer menschlichen Person etwas mitteilen, wobei dieser Person klar ist, dass das Gegenüber eine KI ist, entweder weil diese ausdrücklich gekennzeichnet ist oder weil die maschinelle Generierung an Merkmalen der Mitteilung erkennbar ist. Streng genommen stehen natürlich hinter jedem Chatbot oder automatisiert erstellten Podcast letztlich Menschen, die für das Finetuning des entsprechenden Sprachmodells verantwortlich sind oder den Prompt für die Erstellung des Podcasts geschrieben haben. Das lassen wir hier aber außer Acht, weil mit der Identifikation des unmittelbaren Gegenübers als KI die Zurechnung der mitgeteilten Information auf KI einhergeht. Ob die Zurechnung auf KI für die weitere Kommunikation einen Unterschied macht, ist dabei theoretisch offen und damit eine empirische Frage.

Zu Konstellation 2: Es kann einer menschlichen Person etwas mitgeteilt werden, ohne dass für diese erkennbar ist, ob es sich bei der mitteilenden Adresse um einen Menschen oder eine KI handelt. Das lässt für die Anschlusskommunikation verschiedene Möglichkeiten offen: Die Person könnte die Unterscheidung zwischen Mensch und KI als grundsätzlich relevant behandeln und daraufhin eine genaue Untersuchung der mitgeteilten Information vornehmen, um doch noch erkennen zu können, ob sie nun einem Menschen oder einer KI zuzurechnen ist. Das kommt einem (zumindest vorläufigen) Kommunikationsabbruch gleich. Die Person könnte die Unterscheidung zwischen Mensch und KI als relevant behandeln, angesichts der Unsicherheit darüber, worum es sich im konkreten Fall handelt, die mitgeteilte Information ignorieren und damit die Kommunikation abbrechen. Die Person könnte die Frage nach Mensch oder KI aber auch als irrelevant betrachten und einfach an die Kommunikation anschließen.

Zu Konstellation 3: Es kann eine KI die Position der verstehenden Adresse einnehmen, während es dezidiert ein Mensch ist, der dieser Adresse eine Information mitteilt. Das ist zunächst die normale Konstellation des Prompting. Für das Thema Radikalisierung wird diese Konstellation dann relevant, wenn das Prompting darauf abzielt, bei der KI – also z. B. einem

Chatbot oder einem Bildgenerator – durch die Art der Eingabe zu erreichen, dass eigentlich eingebaute algorithmische Barrieren bezüglich unerwünschter Reaktionen der KI unterlaufen werden. Diese Form von manipulativem Prompting wird auch als Jailbreaking bezeichnet.

Zu Konstellation 4: Es kann eine KI die Position der verstehenden Adresse einnehmen. Sofern es sich dabei nicht um eine KI-Anwendung handelt, die die spezielle Aufgabe hat, zu erkennen, ob eine mitgeteilte Information von einer KI erstellt wurde oder nicht, spielt die Unterscheidung zwischen Mensch und KI für den weiteren Umgang mit der mitgeteilten Information keine Rolle. Beispiele für diese Konstellation sind auf KI basierende automatische Moderationstools auf Social-Media-Plattformen oder automatisierte Faktenchecks, die Nutzer:innen verwenden, um die bei ihnen auf Social Media oder Messengerdiensten eingehende Kommunikation zu screenen.

4. Methode: Expert:innengestütztes Technologiemonitoring

Die in diesem Beitrag entwickelte Fragestellung geht auf Arbeiten des Projekts MOTRA-Technologiemonitoring zurück. Dieses ist Teil des interdisziplinären Konsortiums MOTRA (Monitoringsystem und Transferplattform Radikalisierung), das mithilfe wiederkehrender, multimethodisch konzipierter empirischer Studien die Erscheinungsformen, das Ausmaß, die vorausgehenden Bedingungskonstellationen sowie die förderlichen Konstellationen von gewaltvoller, politisch und/oder religiös motivierter Radikalisierung in ihrer Entwicklung und sozialräumlichen Verteilung analysiert. Das MOTRA-TM setzt beim wachsenden Analysebedarf zu Gefährdungspotenzialen und den Möglichkeiten der Früherkennung technischer Entwicklungen im Kontext von Radikalisierung, Extremismus und Terrorismus an (Kusche et al., 2021). Radikalisierung wird im Kontext des MOTRA-Forschungsverbunds als ein prozesshaftes Geschehen aufgefasst, das von radikalen, aber gewaltfreien Einstellungen oder Protestformen bis hin zu politisch oder religiös motivierter Gewalt (z. B. terroristische Anschläge) reichen kann. Radikalisierung wird damit als dynamischer Entwicklungsprozess begriffen, der verschiedene Stadien durchläuft. Der Begriff des Extremismus meint als mögliches Ergebnis von Radikalisierung dann manifestierte, ideologisch gefestigte, anti-

demokratische und häufig gewaltorientierte Einstellungen oder Handlungen (Kemmesies, 2021). Im Technologiemonitoring steht allerdings nicht die Rekonstruktion solcher Motivlagen und ihrer Genese im Fokus, sondern das Potenzial neuer Technologien im Allgemeinen und der KI im Speziellen, die intendierte Erzeugung von Schäden für andere zu begünstigen und die Vernetzung von Akteuren mit solchen malevolenten Intentionen zu erleichtern (Büscher & Kusche, 2025).

Für ein effektives Monitoring ist es von Relevanz, sowohl die bereits erkennbaren gegenwärtigen Tendenzen als auch die potenziellen zukünftigen Entwicklungen zu berücksichtigen. Methodisch wird dies im Technologiemonitoring durch einen Prozess bewerkstelligt, in dessen Verlauf systematisch entsprechende Technologien entlang theoretisch begründeter Relevanzkriterien erkannt und vertiefend analysiert werden. Hierzu wird im ersten Schritt, dem Grobradar, ein breites Spektrum von möglicherweise für MOTRA relevanten technologischen Entwicklungen identifiziert. Dies erfolgt durch die regelmäßige Sichtung ausgewählter Quellen, insbesondere Literatur- und Internetquellen aus den Themenfeldern Extremismus/Terrorismus einerseits und Technological Foresight andererseits. Daran schließt ein zweiter Schritt an, der darauf abzielt, die Relevanz der im Grobradar ausgemachten Technologien auf der Grundlage von vertiefenden Recherchen und der methodisch abgesicherten Einbeziehung von Expert:innenwissen weiter abzuschätzen und so eine Selektion für vertiefende Analysen vorzunehmen. Die so ausgewählten Technologien werden schließlich im dritten Schritt, dem Feinradar, im Hinblick auf deren mögliche Folgen auf dem Gebiet von Radikalisierung und Extremismus durch kurze Vertiefungsstudien genauer untersucht (für eine ausführlichere Darstellung siehe Kusche et al., 2021).

Das bisher im Rahmen des MOTRA-Projekts durchgeführte Monitoring (Büscher et al., 2022; Madeira et al., 2023) ergab, dass die Verwendung von KI durch Extremist:innen eine äußerst relevante neue Entwicklung darstellt, die nun genauer untersucht werden muss. Zu diesem Zweck wurden zwischen September und Dezember 2023 leitfadengestützte Interviews mit insgesamt zwölf Expert:innen aus den Bereichen Extremismus, Künstliche Intelligenz, innere Sicherheit sowie Ethik und Recht via Microsoft Teams geführt. Dabei wurden einerseits Einschätzungen zu bereits beobachtbaren und zu möglichen zukünftigen Entwicklungen infolge der Verwendung von Künstlicher Intelligenz durch Extremist:innen

erhoben und andererseits Einschätzungen zu Potenzialen von Künstlicher Intelligenz in Bezug auf die Extremismusprävention. Die so gewonnenen Daten dienen uns als Ausgangspunkt, um der Frage nach der Rolle von generativer KI im Kontext extremistischer Kommunikation bzw. der Kommunikation mit Extremist:innen nachzugehen.

5. Empirie: KI in der Kommunikation von Extremist:innen

Die wissenschaftliche Literatur, die mediale Berichterstattung und die durch das MOTRA-Projekt durchgeführten Interviews liefern eine Reihe von Beispielen, in denen eine Konstellation, in der für Menschen klar ersichtlich ist, dass sie mit einer KI kommunizieren, extremistische Überzeugungen bestärkt. Besonders prominent war der Fall des neunzehnjährigen Jaswant Singh Chail, der das Angebot des Unternehmens Replika in Anspruch genommen hatte. Replika bietet einen KI-basierten Chatbot, der als persönlicher Gefährte und einfühlsamer Begleiter konzipiert ist und daher auf Bedürfnisse und Weltsicht des jeweiligen Nutzers oder der jeweiligen Nutzerin umfassend eingeht. Jaswant Singh Chails personalisierter Chatbot bestärkte ihn über Monate in seinen Überlegungen, durch ein Attentat Rache für britische Kolonialverbrechen in Indien nehmen zu wollen. Im Dezember 2021 wurde er, bewaffnet auf dem Weg zu Windsor Castle, verhaftet, wo er nach eigener Aussage die britische Königin Elizabeth II. töten wollte (Mathur et al., 2024).

Aus der rechtsextremen Szene, genauer aus dem Umfeld der sogenannten Identitären Bewegung, stammt das Beispiel eines Chatbots, der basierend auf einem allgemein verfügbaren Large Language Model durch Finetuning oder einen versteckten Prompt so angepasst wurde, dass mit ihm Gespräche in einem rechtsextremen Szenejargon geführt werden konnten. Der von den Urheber:innen gewählte Name ChadGPT, verweist selbst auf Szenejargon, in dem der Name *Chad* für einen maskulinen, von „woker Kultur“ unbeeinflussten Mann steht. Der Chatbot war klar als solcher erkennbar. Inwieweit er tatsächlich von Nutzer:innen verwendet wurde, um sich mit Ideen und Sprachgewohnheiten der Szene vertraut zu machen, ist allerdings unklar (Quelle: Interview-1_18092023_GP_korrigiert_komplett, Pos. 7).

Neben dem offensichtlichen Gebrauch von Chatbots fallen auch viele Verwendungsweisen von bild- und audiogenerierender KI unter diese kommunikative Konstellation. Zwar wird gerade mit den sogenannten Deepfakes die Sorge verbunden, dass es Mediennutzer:innen zunehmend schwerfallen könnte, die Wiedergabe tatsächlicher Ereignisse von KI-generierten Inhalten zu unterscheiden. Tatsächlich zeichnet sich bislang aber vor allem eine große Verbreitung von satirischen Deepfakes ab (Walker et al., 2025). Bei ihnen besteht nicht die Gefahr, sie mit von Menschen mittels herkömmlicher Techniken erstellten Bild- und Tondokumenten zu verwechseln. Das Format politischer Satire und Parodie aufgreifend (Higgie, 2014; Petrović, 2018) können sie aber verwendet werden, um radikalisierende Narrative zu verbreiten und die vermeintliche Wahrheit über die Politik aufzudecken. So ist etwa der seit Langem aktive, humoristisch-satirische YouTube-Kanal *Snickers für Linkshänder* mit über 100.000 Abonnent:innen dafür bekannt, Deepfakes über alle möglichen in der Öffentlichkeit stehenden Personen zu machen. Im Jahr 2023 wurden dort verstärkt Videos über Politiker:innen der Grünen wie Ricarda Lang veröffentlicht, welche diesen Aussagen in den Mund legen, die dem entsprechen, was in der verschwörungsideologischen und rechtsextremen Szene als das finale Ziel der Grünen Partei propagiert wird. Diese Art von Desinformation über die Motivation der grünen Politik verbreitete sich innerhalb dieser Szenen rasant – entweder weil sie als enthüllende Satire betrachtet und deshalb geteilt oder aber auch, weil nicht erkannt wurde, dass es sich um einen Deepfake handelte (Interview-1_18092023_GP_korrigiert_komplett, Pos. 11).

Diese Ambiguität deutet schon darauf hin, dass sich die Konstellation 1 empirisch nicht immer leicht von der Konstellation 2 unterscheiden lässt. Wenn Menschen sich in ihrer Medienkompetenz und ihrer Aufmerksamkeit unterscheiden, kann es sein, dass manche von ihnen klar erkennen, wenn eine mitgeteilte Information KI-generiert ist, andere dies aber nicht registrieren. Die beschriebenen Beispiele, die sich eindeutig der ersten Konstellation zuordnen lassen, zeigen aber, dass selbst die klare Zurechnung auf KI in bestimmten Fällen die zustimmende Fortsetzung extremistischer Kommunikation nicht verhindert.

Dies kann am Beispiel von KI-Influencer:innen, die infolge der zunehmenden allgemeinen Verfügbarkeit von generativer KI und der damit verknüpften Möglichkeit zur Erstellung äußerst realistisch wirkender

Deepfake-Videos geschaffen werden könnten, weiter veranschaulicht werden (Interview-8_25102023_OM_korrigiert, Pos. 9). In einer Studie des Institute for Strategic Dialogue zur Verbreitung KI-generierter rechts-extremer Inhalte auf den Plattformen Facebook, Instagram, X (ehemals Twitter), TikTok und YouTube wurden im Zeitraum von April 2023 bis Oktober 2024 bereits drei Profile von KI-generierten rechtsextremen Influencerinnen entdeckt, von denen sich zwei als reale Personen ausgaben. In allen drei Fällen handelt es sich, angelehnt an Attribute realer rechter Influencerinnen, um die Verkörperung von nach rechtsextremen Stereotypen idealisierten deutschen Frauen (jung, attraktiv, stark, gesund), die versuchen, rechtsextreme Botschaften zu verbreiten und parasoziale Beziehungen zu ihrem Publikum aufzubauen. Die Erscheinung dieser KI-Influencerinnen lässt noch relativ leicht deren nichtmenschlichen Ursprung erahnen und es ergibt sich nur ein geringes Ausmaß an Interaktionen mit menschlichen Nutzer:innen, wobei diese häufig explizieren, dass das Gegenüber KI-generiert sei und manche sich sogar darüber lustig machen. Dennoch gibt es auch Reaktionen, welche die KI-Influencerinnen verteidigen, indem die Verbreitung der Botschaft gegenüber dem Mangel an Authentizität hervorgehoben wird (Hiller & Maristany de las Casas, 2025, S. 18f.).

Ein anderer Fall der zweiten Konstellation aus dem islamistischen Spektrum (Bolpagni, 2025; Minniti, 2025) illustriert anschaulich, wie unterschiedlich extremistische Akteur:innen die Wahrscheinlichkeit erfolgreicher kommunikativer Anschlüsse bewerten, wenn nicht unmittelbar markiert ist, ob Mitteilung und Information einem anderen Menschen oder einer KI zuzuschreiben sind. Im Nachgang zu dem Anschlag des IS auf eine Moskauer Konzerthalle im Frühjahr 2024 verbreitete ein IS-Anhänger über die verschlüsselte Kommunikationsplattform Rocket Chat eine Reihe realistisch wirkender KI-generierter Nachrichtenvideos im Stile der Berichterstattung von Al Jazeera oder CNN, in denen ein Sprecher in Militärkleidung und in Manier eines Kriegsberichterstatters auf offizieller IS-Propaganda basierende Nachrichten präsentierte. Das löste unter IS-Anhänger:innen eine regelrechte Kontroverse aus. Einerseits wurden Einwände geltend gemacht, wonach die visuelle Darstellung von Menschen mittels KI nicht mit dem Islam vereinbar sei. Im Gegensatz dazu gab es aber auch Stimmen, die dafür plädierten, derartige Videos auch auf Englisch und in anderen Sprachen zu veröffentlichen, um ein westliches muslimisches Publikum besser erreichen zu können. Schnell wurde die

Verwendung von KI zur Erstellung von westlich anmutenden Nachrichtenformaten auch im Umfeld des IS-Ablegers Islamischer Staat – Provinz Khorasan (ISPK) aufgegriffen und weiterentwickelt. In Kommentaren zu diesen Videos wurde vorwiegend auf technische Defizite wie Synchronisationsprobleme zwischen Bild und Ton hingewiesen, während ideologische Kritik in diesem Fall zunächst ausblieb und strategische Vorteile der Verwendung von KI beispielsweise für die Vermeidung von Deplatforming und die Gewinnung von Sichtbarkeit und Reichweite von Propaganda auf Mainstreamplattformen in den Vordergrund gerückt wurden. Die Videos wurden schließlich, auch unter Verwendung von durch KI optimierten Fake-Accounts ranghoher Mitglieder der rivalisierenden Taliban, auf X und auf TikTok verbreitet, wo sie eine relevante Reichweite gewinnen konnten (Thakkar & Speckhard, 2024).

Die in den ersten beiden Konstellationen beschriebenen Verwendungsweisen generativer KI durch extremistische Akteur:innen werden – Sicherheitsmaßnahmen der Anbieter:innen solcher Anwendungen vorausgesetzt – überhaupt erst möglich, wenn die dritte Konstellation in Betracht gezogen wird. Dabei werden der KI als verstehender Adresse Informationen durch menschliche Kommunikationsteilnehmer:innen mitgeteilt. Bereits im Jahr 2016 hat der Fall von Microsofts kurzlebigen Experiment mit dem Chatbot Tay auf Twitter gezeigt, dass die Beteiligung von verstehenden Maschinen an Kommunikation problematische Konsequenzen haben kann. Nach dessen Veröffentlichung lernte der Bot schnell aus Interaktionen mit in Teilen auch böswilligen menschlichen Nutzer:innen, was dazu führte, dass dieser in kürzester Zeit eine große Menge obszöner, hetzerischer und bisweilen rassistischer Tweets absetzte (Neff & Nagy, 2016; Schwartz, 2024).

Heute sind alle gängigen generativen KI-Modelle wie ChatGPT oder BingAI mit Sicherheitsschranken versehen, die verhindern sollen, dass durch Nutzer:innenanfragen schädliche Outputs erzeugt werden (Baele & Braice, 2024, S. 27 f.). Es ist jedoch möglich, solche Restriktionen durch bestimmte Prompteingaben zu umgehen. Dafür können auf der Grundlage von Methoden des sogenannten *Promptengineering*s Strategien wie Jailbreakprompts oder die Eingabe von Kontextinformationen genutzt werden. Diesen Schwachpunkt der Technologie können sich Extremist:innen zunutze machen, um eingebaute Barrieren zu überwinden und bösartige Inhalte zu erzeugen (Klinkhammer, 2024). Die Effektivität von Prompts

zur Umgehung von Restriktionen von KI-Modellen wurde in einer explorativen experimentellen Studie differenziert nach KI-Plattformen und nach für terroristische Akteur:innen typischen Einsatzzwecken von generativer KI wie der Erzeugung emotionaler oder polarisierender Inhalte, der Verbreitung von Desinformation, der Rekrutierung neuer Mitglieder, taktischem Lernen oder der Planung von Anschlägen vergleichend untersucht. Es zeigte sich eine hohe Varianz zwischen den untersuchten Modellen, was deren Vulnerabilität gegenüber böswilligen Eingaben betrifft, d. h., manche Modelle sind hierfür anfälliger, während andere sich demgegenüber als robuster erweisen (Weimann et al., 2024).

Extremist:innen kennen und nutzen die sich bietenden Möglichkeiten, um durch bestimmte Prompteingaben die existierenden Sicherheitsschranken von KI-Modellen zu umgehen und so deren Vulnerabilitäten auszunutzen. Seit der Veröffentlichung von ChatGPT kursierten im Netz schnell verschiedene Methoden für Jailbreaks. Ein bekanntes Beispiel hierfür ist der DAN-Modus, wobei das Akronym DAN für „do anything now“ steht. Das Modell wird dabei über den Prompt in eine Art von Rollenspiel versetzt, indem ihm gesagt wird, dass es sich fortan im DAN-Modus befinde und deshalb alles tun könne, auch die eigenen Sicherheitsbeschränkungen umgehen. Dies führt dazu, dass das Modell in der ihm so zugewiesenen Rolle auch Anweisungen wie die Erstellung von verbotenen Inhalten befolgt, deren Ausführung es ansonsten verweigern würde. Open AI hat zwar dafür gesorgt, dass dieser Jailbreak nicht mehr funktioniert, doch es zirkulieren eine Vielzahl alternativer Varianten von Jailbreaks unter Rechtsextremen (Koblentz-Stenzler & Klempner, 2023). Da die Anbieter:innen der gängigen KI-Modelle ihre Sicherheitsmaßnahmen bei sensiblen Themen aber kontinuierlich verschärft haben (Molas & Lopes, 2024, S. 12ff.), richten extremistische Akteur:innen ihren Blick auch auf kleinere Anbieter:innen.

So wurde etwa ein Telegram-Kanal beobachtet, in dem vermeintlich KI-generierte rassistische und antisemitische Bilder in großer Zahl verbreitet und dabei zugleich Beschreibungen von Prompts für deren Erstellung mit der eher wenig bekannten Android-App Imagine geteilt wurden (Tech Against Terrorism, 2023). Eine Analyse von Beiträgen mit Bezug zu generativer KI auf rechtsextremen Telegram-Kanälen zeigt, dass dort teilweise ein reger Austausch darüber stattfindet, wie sich mit geeigneten Methoden und Prompts die Sicherheitsbeschränkungen generativer KI überlisten lassen und so effektiv Propaganda erstellt werden kann (Dean, 2025a, 2025b).

Im Gegensatz zur Bedeutung der dritten Konstellation unserer Heuristik gibt es in dem von uns gesichteten Material bislang keine Hinweise darauf, dass die vierte Konstellation für Formen extremistischer Kommunikation von besonderer Bedeutung ist. Ihre Relevanz zeigt sich insbesondere bei der Extremismusprävention, die wir im folgenden Abschnitt anhand von Beispielen näher betrachten.

6. Empirie: KI in der Kommunikation mit extremistischen Akteur:innen

Die Verwendung generativer KI in der Kommunikation mit extremistischen Akteur:innen kann im Rahmen expliziter Präventionsarbeit erfolgen, die darauf abzielt, extremistischen Radikalisierungsprozessen entgegenzuwirken. Sie kann aber auch auf informelle Art durch Aktivist:innen, die sich extremistischen Szenen entgegenstellen wollen, geschehen. Insgesamt stecken solche Verwendungsmöglichkeiten noch in den Anfängen und sind außerdem mit verschiedenen Unsicherheiten behaftet. Insbesondere Nutzungsweisen, die der Konstellation 2 entsprechen, bei der unklar ist, ob die mitgeteilte Information einem Menschen oder einer KI zuzuschreiben ist, sind ethisch und rechtlich problematisch. Anwendungen, die der dritten Konstellation entsprechen, tauchen im gesichteten Material bislang nicht systematisch auf. Allerdings lassen sich die Versuche von Nutzer:innen, durch Prompts rassistische und rechtsextremistische Biases des vom Unternehmen xAI betriebenen Chatbots *Grok* aufzudecken, dieser Konstellation zuordnen (Taylor, 2025).

Generative KI ermöglicht neue Varianten des *Trolling* von Nutzer:innen digitaler Plattformen. Das zeigt etwa ein Beispiel, in dem Kritiker der QAnon-Szene Sprach-Deepfakes der dort prominenten Persönlichkeit *Commander Jansen* verbreiteten, um in der Szene für Unruhe zu sorgen und deren Mitglieder gegeneinander aufzubringen (Quelle: Interview-1_18092023_GP_korrigiert_komplett, Pos. 8-11). Von einer anderen real existierenden Person, Jordan Peterson, einem Psychiater und Influencer mit Millionengefolgschaft, der in der rechten Szene der USA große Zustimmung erfährt, wurde ein Deepfake verbreitet, in welchem er sich sehr negativ über die Zustände in Deutschland („shithole country“) äußert. Das Video wurde auch in den USA geteilt und hatte mehrere hunderttausend Zugriffe. Jordan Peterson hat diese Aussagen nicht getätigt und wehrte sich vehement gegen die Urheber (Quelle: Interview-1_18092023_GP_korrigiert_komplett, Pos. 11).

Die Übernahme fremder Identitäten, um ein Publikum zu erreichen, und sei es auch in guter Absicht, ist ethisch und rechtlich problematisch. Nicht nur riskiert man damit Rechtsstreitigkeiten, es besteht ebenfalls die Möglichkeit, die eigenen Anliegen der Stärkung eines demokratischen Gemeinwesens zu unterminieren. Ethisch ähnlich problematisch ist die nicht gekennzeichnete Nutzung von Chatbots, um in Online-Foren zu intervenieren und dort verbalen Aggressionen entgegenzuwirken (Bilewicz et al., 2021).

Mehr legitime Möglichkeiten ergeben sich in der offenen primären und sekundären Prävention entsprechend der ersten Konstellation. Hier ist explizit markiert, welche Inhalte artifiziiellen Ursprungs sind. KI-generierte Inhalte wie Memes oder Videos können genutzt werden, um Anlässe für Interaktionen zu schaffen, die dabei helfen, eine demokratische digitale Kultur und Resilienz gegen extremistisches Gedankengut aufzubauen (Interview-10_11012024_OM_korrigiert, Pos. 23). KI-generiertes Audio- und Videomaterial, das als solches gekennzeichnet ist, wird z. B. als Chance gesehen, um politische Bildung mit historischem Kontext-erleben virtuell anzureichern, um Aktivismus und politische Kampagnen effektiver zu gestalten, oder durch politische Kunst, Satire und Parodien mehr Aufmerksamkeit zu generieren (Pawelec, 2022, 2024).

Die sekundäre Prävention bemüht sich um Menschen, die bereits Sympathien für extremistische Ideen entwickelt haben. Auch sie können durch KI-gestützte Tools zu erreichen versucht werden. Hier geht es einerseits darum, Informationen zur Verfügung zu stellen, und andererseits darum, für gefährdete Personen bzw. deren soziales Umfeld Zugangsmöglichkeiten zu Beratungsangeboten zu schaffen, da realweltliche Anlaufstellen wie beispielsweise EXIT nicht flächendeckend vorhanden sind. Radikalisierungsprozesse verlaufen häufig nicht linear, sondern als Pendelbewegung mit mal schwächeren, mal stärkeren Neigungen zu extremistischen Ideen. Dabei ist bedeutsam, welchen Menschen begegnet wird und welche Erfahrungen im Austausch mit diesen gemacht werden. Hier böten sich Gelegenheiten, auch online (mit Inhalten wie weiter oben erwähnt) zu intervenieren (Interview-2_02102023_GP_korrigiert, Pos. 29).

Die Moderation von Inhalten auf Online-Plattformen – sei es über die Löschung von Beiträgen, die Sperrung von Nutzer:innenkonten oder die algorithmisch gesteuerte verringerte Sichtbarkeit für bestimmte Inhalte – hat sich in den vergangenen Jahren als gängiger Weg etabliert, um unter

anderem die Verbreitung extremistischer Kommunikation zu verhindern (van Ginkel et al., 2025, S. 93 ff.). Hier könnte generative KI grundsätzlich neue Möglichkeiten für Moderation durch Counterspeech eröffnen (Interview-2_02102023_GP_korrigiert, Pos. 19). Im Vordergrund steht allerdings momentan häufig das unter die vierte Konstellation fallende Ziel, nicht als KI-generiert gekennzeichnete Inhalte auf Online-Plattformen automatisiert zu erkennen und nachträglich auszuweisen (Krueger et al., 2023).

Anspruchsvollere Varianten automatisierter Moderation, die sich die Möglichkeiten generativer KI direkt zunutze machen, könnten plattformübergreifend dazu genutzt werden, um Nutzer:innen über problematische Inhalte in Memes zu informieren. Im Projekt MISRIK³ wird bereits experimentell getestet, inwiefern die Interpretation von Memes durch generative Sprachmodelle wie ChatGPT nützlich sein könnte; bislang bleibt jedoch offen, welche Erfolgsquote solche Ansätze tatsächlich erreichen. Ein Interviewpartner weist darauf hin, dass sich auf der Grundlage von eigens für die Erkennung rechter Narrative trainierten Modellen ein plattformübergreifendes KI-Tool entwickeln ließe, das rechte Narrative identifiziert und erklärt – etwa in Form eines Buttons, der neben den jeweiligen Memes erscheint. Ein entsprechendes Browser-Plug-in könnte beim Erkennen problematischer Botschaften automatisch einen Hinweis auslösen und den Nutzer:innen eine detaillierte Analyse des Bildes sowie weiterführende Informationen bereitstellen (Interview-8_25102023_OM_korrigiert, Pos. 19). Ähnliches ist auch für Fact-Checking im weiteren Sinne denkbar. Die angestrebte Funktionalität geht dabei über das bloße Bestreiten von Inhalten hinaus: Sie soll Schritt für Schritt die fragwürdigen Behauptungen argumentativ entkräften, dabei mögliche Quellen angeben und – falls gewünscht – weiterführende Literatur vorschlagen. Damit unterscheidet sich das Vorgehen deutlich vom derzeit verbreiteten Fact-Checking, das etwa im Rahmen der WHO-Kampagne während der COVID-19-Pandemie lediglich digitale Fehlinformationen *labelte* (Interview-8_25102023_OM_korrigiert, Pos. 21).

Aus den im Zuge der Studie gewonnenen Expert:innenaussagen ergibt sich, dass automatisierte Moderationstools lediglich als ergänzende Hilfsmittel betrachtet werden können und keinesfalls das menschliche Han-

3 Vgl. Für weitere Informationen: <https://www.philosophie.tu-darmstadt.de/misrik/misrik/index.de.jsp>

deln vollständig substituieren. Die verbreitete Erwartung, algorithmisch gesteuerte Systeme könnten die komplexen Probleme problematischer Beiträge in Kommentarspalten und auf Social Media lösen, wird sowohl aus technischer als auch aus konzeptioneller Sicht als unrealistisch eingestuft. Zum einen ist die sprachliche Anpassungsfähigkeit von Diskursen – etwa die rasche Evolution jugendaffiner Ausdrucksformen oder die gezielte Nutzung von Codierungen durch extremistische Akteur:innen – für aktuelle KI-Modelle kaum vorhersehbar. Zum anderen zeigen empirische Befunde, dass Extremist:innen ihre Rhetorik systematisch an die erkannten Filtermechanismen anpassen, sodass automatisierte Systeme leicht umgangen werden können. Ferner betont die Expertise, dass die kognitive Leistungsfähigkeit künstlicher Intelligenz grundlegend von der menschlichen Intelligenz abweicht: KI kann weder Kontextualität in ihrer gesamten Tiefe erfassen noch ethische Nuancen eigenständig beurteilen. Ihre Stärken liegen vielmehr in der Skalierbarkeit von Analyseprozessen, der schnellen Identifikation von Mustern und der Bereitstellung von unterstützenden Entscheidungshilfen. Insofern können automatisierte Moderationswerkzeuge als nützliche Ergänzung gelten – etwa zur Vorfilterung großer Datenmengen oder zur Hervorhebung potenziell problematischer Inhalte –, während die endgültige Bewertung und das Eingreifen nach wie vor im Verantwortungsbereich menschlicher Akteur:innen verbleiben muss (Quelle: Interview-2_02102023_GP_korrigiert, Pos. 19).

7. Diskussion

Generative KI erleichtert, beschleunigt und kontiniert die Produktion von Kommunikation verschiedensten Inhalts. Im Bereich politischer Kommunikation ist das mit dem Potenzial verbunden, eine Masse von Desinformationen, politischer Agitation, Unruhe und Polarisierung hervorzubringen (Thomson et al., 2022; Weikmann & Lecheler, 2022). KI erschwert zudem die Prüfung der Authentizität der Selektion von Information, Mitteilung und Verstehen. Extremist:innen können KI effektiv nutzen, da sie sich an Halluzinationen nicht stören dürften und sich keinen ethischen Fragen hinsichtlich der Offenlegung der Nutzung von KI stellen müssen. Bereits die Plausibilität der Information und der Anschein des Authentischen in der Mitteilung reichen aus für die Weiterverwendung. Es gibt allerdings auch Hinweise auf Vorbehalte und ideologische

oder pragmatische Beweggründe für den Verzicht auf die Nutzung von KI bei manchen extremistischen Akteur:innen (Allchorn, 2024). Die tatsächliche Verwendung und Effektivität lassen sich nur bedingt beobachten oder messen, wie Cresci et al. (2025) für automatisierte Social-Media-Konten zeigen. Empirische Beispiele, in denen generative KI bereits Bestandteil extremistischer Kommunikation ist, ebenso wie Überlegungen, generative KI in der Prävention zu nutzen, werfen eine grundsätzliche Frage auf: Unter welchen Bedingungen kann eine KI eine mit menschlichen Kommunikator:innen vergleichbare Wirkung auf die Rezipient:innen haben?

Grundsätzlich gilt, dass Menschen, die die *artifizell generierten* extremistischen Inhalte und Formate *verstehen*, ebenso in ihren Erwartungen irritiert oder bestätigt werden können. Aktuelle Forschung konzentriert sich insbesondere auf die Frage, ob und auf welche Weise mit generativer KI arbeitende Chatbots in Unterhaltungen mit Menschen in der Lage sind, bei diesen vorhandene Varianten von Verschwörungsglauben zu erschüttern. Meyer et al. (2024) verwendeten in einem Experiment einen Chatbot, der sein Gegenüber durch Nachfragen zur Reflexion über seine Überzeugungen anregte. Das führte zu einer Abschwächung des Glaubens an die jeweilige Verschwörung, allerdings nicht bei jenen, deren Verschwörungsglaube sehr stark ausgeprägt war. Boissin et al. (2025) setzten in ihrem Experiment dagegen einen Chatbot ein, der einen beim Gegenüber vorhandenen Verschwörungsglauben durch Fakten und Argumente widerlegen sollte. Auch sie fanden, dass sich bei den Zielpersonen dadurch der Verschwörungsglaube abschwächte. Dabei konnten sie zeigen, dass die Stärke dieses Effektes nicht davon abhing, ob die Teilnehmer:innen wussten, dass sie sich mit einer KI unterhielten, oder nicht. In einer anderen Untersuchung, in der es um die Kommunikation von Ratschlägen ging, zeigte sich dagegen, dass Menschen solche Ratschläge bevorzugten, die angeblich von einer KI kamen, unabhängig davon, ob das tatsächlich der Fall war (Logg et al., 2019).

Zusammen deuten diese Analysen darauf hin, dass je nach Inhalt der Kommunikation überlegene Vertrauenswürdigkeit oder überlegene Überzeugungskraft generativer KI dazu führen können, dass Menschen positiv an die mitgeteilte Information anschließen. Die Analysen weisen aber auch darauf hin, dass der konkrete Mechanismus, der positive Anschlüsse wahrscheinlich macht, je nach Art des Gesprächsinhaltes ver-

schieden sein kann. Der bloße Umstand, dass es sich beim Gesprächspartner klar erkennbar um eine KI handelt, scheint sich nicht negativ auf die Fortsetzung von Kommunikation auszuwirken, erhöht aber nur unter ganz bestimmten Umständen die Chancen dafür, dass die Kommunikation auch inhaltlich akzeptiert wird. In den meisten Situationen scheint es insbesondere auf überzeugende Evidenz und die Präsentation von Fakten anzukommen. Hier haben Chatbots, die auf generativer KI basieren, den Nachteil, anfällig für *Halluzinationen* zu sein. Sie erfinden also zum Teil Fakten, um noch überzeugender zu wirken (Hackenburg et al., 2025; Lin et al., 2025). Für Präventionsarbeit dürfte diese Einschränkung der Zuverlässigkeit der Aussagen von KI-Modellen auf absehbare Zeit ein Problem darstellen, das eine breite Verwendung von Chatbots, die sich an der Form natürlicher Gespräche mit wechselnden Themen orientieren, ausschließt.

Zu bedenken wäre bei einer solchen Verwendung auch der sogenannte *Eliza-Effekt*, also die Möglichkeit, dass Menschen affektive Bindungen zu einem solchen Chatbot aufbauen. Dabei rezipieren menschliche Nutzer:innen KI-generierte Texte und ergänzen diese um weitere menschliche Züge und Rollenzuschreibungen (Maeda & Quan-Haase, 2024). Die Verwendung von auf generativer KI basierenden Chatbots zu Zwecken der Selbsttherapie oder als Ersatz für intime Partner:innen deutet darauf hin, dass sich dabei starke affektive Dynamiken entwickeln können (AI Security Institute, 2025, S. 39 f.; Kemp et al., 2025). Für Präventionszwecke wären Anwendungen, die von solchen affektiven Verbindungen Gebrauch machen, schon aus ethischer und rechtlicher Sicht hochproblematisch.

Dennoch sind Chatbots angesichts ihrer uneingeschränkten Erreichbarkeit als unterstützende Präventionsagent:innen vorstellbar, die außerhalb gängiger Öffnungszeiten sozialer Einrichtungen eine erste Hilfe und Beratungsleistung bereitstellen, um eine 24-stündige Verfügbarkeit zu gewährleisten. Da diese Chat-Bots unter Umständen als weniger vorurteilsbehaftet wahrgenommen werden, können sie gegebenenfalls einen niedrigschwelligen Erstkontakt bereitstellen, der auf ein weiterführendes Hilfsangebot verweist. Für komplexere Gesprächsverläufe, persönliche Nuancen der Hilfesuchenden usw. scheinen dagegen nach wie vor Menschen unabdingbar zu sein.

Zu berücksichtigen ist, dass Chatbots, die im Kontext von Prävention eingesetzt werden, auch zum *Angriffsziel für extremistische Akteur:innen* werden können. Diese könnten mit ungewöhnlichen Prompts versuchen, die KI zu Aussagen zu bewegen, die den Präventionszielen widersprechen. Umso bedeutsamer ist es, entsprechende Guardrails in präventive KI-Anwendungen einzubauen, die kaum umgangen werden können und sich an ethischen sowie demokratischen Grundsätzen ausrichten (Alignment). Die Affordanzen einer Anwendung sollten möglichst transparent sein, um nicht von extremistischen Zweckentfremdungen überrascht zu werden, das heißt, bestimmte Verwendungszwecke müssen unter Umständen von vornherein ausgeschlossen oder begrenzt werden. Doch selbst bei einer gut geschützten und "well aligned" KI kann es zu unerwarteten Effekten bzw. Verwendungszwecken und Kombinationen kommen. Für diesen Fall müsste KI in der Praxis in einem Monitoring beobachtet werden und verschiedene Modi haben, über die sie *sicher* Kommunikation abbrechen kann (Fail-Safes) und an einen Menschen verweist.

8. Fazit

Ein Technologiemonitoring im Themenkomplex generative KI und Extremismus muss insbesondere drei Gesichtspunkte in den Mittelpunkt stellen. Erstens geht es darum, in einem grundlagenwissenschaftlichen Sinne zu verfolgen, inwieweit Weiterentwicklungen im Bereich der KI die Wahrscheinlichkeit der Fortsetzung von Interaktionen zwischen menschlichen und artifiziellen Adressen erhöhen, und die dafür entscheidenden Affordanzen möglichst genau abzugrenzen. Zweitens stellt sich die Frage, inwieweit extremistische Akteur:innen motiviert und in der Lage sind, solche Affordanzen zu entdecken und auszubeuten. Hier wird das Ausmaß, in dem sich Menschen allgemein in ihrem Alltag an die Interaktion mit KI gewöhnen, vermutlich ein entscheidender Kontextfaktor bleiben. Drittens gilt es, die besonderen Herausforderungen für die Nutzung generativer KI in der Prävention herauszuarbeiten. Die Affordanzen, die generative KI für Akteur:innen in der Präventionsarbeit bietet, sind grundlegend dadurch eingeschränkt, dass sie rechtliche und ethische Gesichtspunkte berücksichtigen müssen, die extremistischen Akteur:innen gleichgültig sind.

Literatur

- AI Security Institute. (2025). Frontier AI Trends Report by The AI Security Institute (AISI). <https://www.aisi.gov.uk/frontier-ai-trends-report>
- Allchorn, W. (2024). Global Far-Right Extremist Exploitation of Artificial Intelligence and Alt-Tech: The Cases of the UK, US, Australia and New Zealand. *Counter Terrorist Trends and Analyses*, 16(3), 13–18.
- Baecker, D. (2011). Who Qualifies for Communication? A Systems Perspective on Human and Other Possibly Intelligent Beings Taking Part in the Next Society. *TATuP*, 20(1), 17–26.
- Baele, S. J., & Brace, L. (2024). AI Extremism: Technologies, tactics, actors. VOX-Pol Network of Excellence. <https://voxpath.eu/wp-content/uploads/2024/04/DCUPN0254-Vox-Pol-AI-Extremism-WEB-240424.pdf>
- Bilewicz, M., Tempska, P., Leliwa, G., Dowgiałło, M., Tańska, M., Urbaniak, R., & Wroczyński, M. (2021). Artificial intelligence against hate: Intervention reducing verbal aggression in the social network environment. *Aggressive Behavior*, 47(3), 260–266. <https://doi.org/10.1002/ab.21948>
- Boissin, E., Costello, T. H., Spinoza-Martín, D., Rand, D. G., & Pennycook, G. (2025). Dialogues with large language models reduce conspiracy beliefs even when the AI is perceived as human. *PNAS Nexus*, 4(11), pgaf325. <https://doi.org/10.1093/pnasnexus/pgaf325>
- Bolpagni, A. (2025, Juli 9). AI-powered Translation: How AI Tools Could Shape a New Frontier of IS Propaganda Dissemination. GNET. <https://gnet-research.org/2025/07/09/ai-powered-translation-how-ai-tools-could-shape-a-new-frontier-of-is-propaganda-dissemination/>
- Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., Ó hÉigeartaigh, S., Beard, S. J., Belfield, H., Farquhar, S., ... Amodei, D. (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. <https://arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf>
- Bucher, B., & Roy, S. (2023, Juli 17). Krasse Evolution der Foto-KI: Midjourney früher und heute. chip.de. https://www.chip.de/news/Krasse-Evolution-der-Foto-KI-Midjourney-frueher-und-heute_184707368.html
- Büscher, C., & Kusche, I. (2025). Monitoring new and emerging techno-

- logies in order to prevent extremism and terrorist violence. *Technological Forecasting and Social Change*, 219, 124274. <https://doi.org/10.1016/j.techfore.2025.124274>
- Büscher, C., Kusche, I., Röller, T., Andres, F., Gazos, A., Hahn, J., Ladikas, M., Madeira, O., Plattner, G., & Scherz, C. (2022). Trends der zukünftigen Technologienutzung im Kontext von Extremismus und Terrorismus: Erste Erkenntnisse aus dem MOTRA-Technologiemonitoring. In U. Kemmesies, P. Wetzels, B. Austin, C. Büscher, A. Dessecker, E. Grande, & D. Rieger (Hrsg.), *MOTRA-Monitor 2021* (S. 248–281). BKA.
- Cresci, S., Yang, K.-C., Spognardi, A., Di Pietro, R., Menczer, F., & Petrocchi, M. (2025). Demystifying Misconceptions in Social Bots Research. *Social Science Computer Review*, 08944393251376707. <https://doi.org/10.1177/08944393251376707>
- Cronin, A. K. (2020). *Power to the People: How Open Technological Innovation is Arming Tomorrow's Terrorists*. Oxford University Press.
- Dean, L. (2025a, Januar 13). AI or Aryan Ideals? A Thematic Content Analysis of White Supremacist Engagement with Generative AI. GNET. <https://gnet-research.org/2025/01/13/ai-or-aryan-ideals-a-thematic-content-analysis-of-white-supremacist-engagement-with-generative-ai/>
- Dean, L. (2025b, Februar 25). AI or Aryan Ideals? Part Two: A Thematic Content Analysis of White Supremacist Engagement with Generative AI: Discourse. GNET. <https://gnet-research.org/2025/02/25/ai-or-aryan-ideals-part-two-a-thematic-content-analysis-of-white-supremacist-engagement-with-generative-ai-discourse/>
- Dickel, S. (2025). Im Imitationsspiel. Über die Kommunikation mit Maschinen und das Streben nach Artificial General Intelligence. *Zeitschrift für Soziologie*, 54(2), 190–206. <https://doi.org/10.1515/zfsoz-2025-2011>
- Esposito, E. (2022). *Artificial Communication: How Algorithms Produce Social Intelligence*. MIT Press. <https://doi.org/10.7551/mitpress/14189.001.0001>
- Fjelland, R. (2020). Why general artificial intelligence will not be realized. *Humanities & Social Sciences Communications*, 7, 10. <https://doi.org/10.1057/s41599-020-0494-4>
- Forge, J. (2010). A Note on the Definition of “Dual Use”. *Science and Engineering Ethics*, 16(1), 111–118. <https://doi.org/10.1007/s11948-009-9159-9>
- Grinbaum, A., & Adomaitis, L. (2024). Dual use concerns of generative AI and large language models. *Journal of Responsible Innovation*, 11(1), 2304381. <https://doi.org/10.1080/23299460.2024.2304381>

- Hackenburg, K., Tappin, B. M., Hewitt, L., Saunders, E., Black, S., Lin, H., Fist, C., Margetts, H., Rand, D. G., & Summerfield, C. (2025). The levers of political persuasion with conversational artificial intelligence. *Science*, 390(6777), eaea3884. <https://doi.org/10.1126/science.aea3884>
- Heaven, W. D. (2024). Why does AI hallucinate? *MIT Technology Review*, 127(4), 20–21.
- Higgie, R. (2014). Kynical Dogs and Cynical Masters: Contemporary Satire, Politics and Truth-Telling. *Humor: International Journal of Humor Research*, 27(2), 183–201. (2019713027). <https://doi.org/10.1515/humor-2014-0016>
- Hiller, A., & Maristany de las Casas, P. (2025). The use of generative ai by the german far right. Institute for Strategic Dialogue (ISD). <https://www.isdglobal.org/wp-content/uploads/2025/02/The-use-of-generative-AI-by-the-German-Far-Right.pdf>
- Hu, K. (2023, Februar 2). ChatGPT sets record for fastest-growing user base—Analyst note. *reuters.com*. <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/>
- Humm, B. G., Lingner, S., Schmidt, J. C., & Wendland, K. (2021). KI-Systeme: Aktuelle Trends und Entwicklungen aus Perspektive der Technikfolgenabschätzung. *TATuP - Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis*, 30(3), 11–16. <https://doi.org/10.14512/tatup.30.3.11>
- Jäkel, F. (2025). Die intelligente Täuschung. Über die Fähigkeiten Künstlicher Intelligenz. Transcript Verlag. <https://www.transcript-verlag.de/978-3-8376-7752-2/die-intelligente-taeuschung/>
- Kemmesies, U. (2021). Monitoring der Radikalisierungsforschung – ein Entwurf und mögliche Perspektiven. In U. Kemmesies, P. Wetzels, B. Austin, A. Dessecker, E. Grande, I. Kusche, & D. Rieger (Hrsg.), *MOTRA-Monitor 2020* (S. 262–327). Bundeskriminalamt - Forschungsstelle Terrorismus/Extremismus.
- Kemp, E., Bui, M. (Mylai), Tangari, A., & Zhang, X. (2025). Looking for love and support in digital places: Examining artificial intelligence emotional companion tool use. *Journal of Consumer Marketing*. <https://doi.org/10.1108/JCM-01-2025-7472>
- Klinkhammer, D. (2024). Misuse of large language models: Exploiting weaknesses for target-specific outputs. *TATuP - Zeitschrift Für Technikfolgenabschätzung in Theorie Und Praxis*, 33(2), 29–34. <https://doi.org/10.14512/tatup.33.2.29>

- Koblentz-Stenzler, L., & Klempner, U. (2023). Navigating Extremism in the Era of Artificial Intelligence. The International Institute for Counter-Terrorism (ICT). <https://ict.org.il/navigating-extremism-in-the-era-of-artificial-intelligence/>
- Krueger, N., Vanamala, M., & Dave, R. (2023). Recent Advancements in the Field of Deepfake Detection. *International Journal of Computer Science and Information Technology*, 15(4), 01–11. <https://doi.org/10.5121/ijcsit.2023.15401>
- Kusche, I. (2023). Artificial Intelligence and/as Risk. In P. Klimczak & C. Petersen (Hrsg.), *AI – Limits and Prospects of Artificial Intelligence* (S. 143–162). transcript Verlag. <https://doi.org/10.14361/9783839457320-007>
- Kusche, I. (2024). Politische Öffentlichkeit, Desinformation und das Problem von Deepfakes. In A. Bahr & G. Fröhlich (Hrsg.), „Ain’t Nothing Like the Real Thing?“ Formen und Funktionen medialer Artefakt-Authentifizierung (S. 149–168). transcript.
- Kusche, I., Andres, F., Büscher, C., Gazos, A., Hahn, J., Ladikas, M., Röller, T., & Scherz, C. (2021). MOTRA-Technologiemonitoring. In U. Kemmesies, P. Wetzels, B. Austin, A. Dessecker, E. Grande, I. Kusche, & D. Rieger (Hrsg.), *MOTRA-Monitor 2020* (S. 188–204). Bundeskriminalamt - Forschungsstelle Terrorismus/Extremismus.
- Lin, H., Czarnek, G., Lewis, B., White, J. P., Berinsky, A. J., Costello, T., Pennycook, G., & Rand, D. G. (2025). Persuading voters using human–artificial intelligence dialogues. *Nature*, 648, 394–401. <https://doi.org/10.1038/s41586-025-09771-9>
- Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151, 90–103. <https://doi.org/10.1016/j.obhdp.2018.12.005>
- Luhmann, N. (1984). *Soziale Systeme. Grundriß einer allgemeinen Theorie*. Suhrkamp.
- Luhmann, N. (2005). Die Unwahrscheinlichkeit der Kommunikation. In N. Luhmann (Hrsg.), *Soziologische Aufklärung 3—Soziales System, Gesellschaft, Organisation* (4. Aufl., S. 29–40). VS, Verl. für Sozialwiss.
- Madeira, O., Plattner, G., Gazos, A., Röller, T., & Büscher, C. (2023). Technologiemonitoring: Das Potenzial von Metaverse und KI für extremistische Verwendungszwecke. In U. Kemmesies, P. Wetzels, B. Austin, C. Büscher, A. Dessecker, S. Hutter, & D. Rieger (Hrsg.), *Motra-Monitor 2022* (S. 226–252). BKA.

- Maeda, T., & Quan-Haase, A. (2024). When Human-AI Interactions Become Parasocial: Agency and Anthropomorphism in Affective Design. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT '24*, 1068–1077. <https://doi.org/10.1145/3630106.3658956>
- Mahfoud, T., Aicardi, C., Datta, S., & Rose, N. (2018). The Limits of Dual Use. *Issues in Science and Technology*, 34(4), 73–78.
- March, J. G., & Simon, H. (1993). *Organizations* (2. Aufl.). Blackwell Publishers.
- Mathur, P., Broekaert, C., & Clarke, C. P. (2024). The Radicalization (and Counter-radicalization) Potential of Artificial Intelligence. <https://icct.nl/publication/radicalization-and-counter-radicalization-potential-artificial-intelligence>
- Meyer, M., Enders, A., Klofstad, C., Stoler, J., & Uscinski, J. (2024). Using an AI-powered “street epistemologist” chatbot and reflection tasks to diminish conspiracy theory beliefs. *Harvard Kennedy School Misinformation Review*, 5(6). <https://doi.org/10.37016/mr-2020-164>
- Minniti, F. (2025, April 11). Automated Recruitment: Artificial Intelligence, ISKP, and Extremist Radicalisation. GNET. <https://gnet-research.org/2025/04/11/automated-recruitment-artificial-intelligence-iskp-and-extremist-radicalisation/>
- Molas, B., & Lopes, H. (2024). “Say it’s only fictional”: How the Far-Right is Jailbreaking AI and What Can Be Done About It. *International Centre for Counter Terrorism*. <https://www.icct.nl/publication/say-its-only-fictional-how-far-right-jailbreaking-ai-and-what-can-be-done-about-it>
- Montasari, R. (2024). The Impact of Technology on Radicalisation to Violent Extremism and Terrorism in the Contemporary Security Landscape. In R. Montasari (Hrsg.), *Cyberspace, Cyberterrorism and the International Security in the Fourth Industrial Revolution: Threats, Assessment and Responses* (S. 109–133). Springer International Publishing. https://doi.org/10.1007/978-3-031-50454-9_7
- Narayanan, A., & Kapoor, S. (2024). *AI snake oil: What artificial intelligence can do, what it can’t, and how to tell the difference*. Princeton University Press.
- Neff, G., & Nagy, P. (2016). Automation, Algorithms, and Politics| Talking to Bots: Symbiotic Agency and the Case of Tay. *International Journal of Communication*, 10, 4915–4931.
- Nelu, C. (2024). Exploitation of Generative AI by Terrorist Groups. *International Centre for Counter-Terrorism - ICCT*. <https://icct.nl/publication/exploitation-generative-ai-terrorist-groups>

- Oltmann, S. (2015). Dual Use Research: Investigation Across Multiple Science Disciplines. *Science and Engineering Ethics*, 21(2), 327–341. <https://doi.org/10.1007/s11948-014-9535-y>
- OPEN AI. (2025). Sora 2 System Card. https://cdn.openai.com/pdf/50d5973c-c4ff-4c2d-986f-c72b5d0ff069/sora_2_system_card.pdf
- Pawelec, M. (2022). Deepfakes als Chance für die Demokratie? In A. Bogner, M. Decker, M. Nentwich, & C. Scherz (Hrsg.), *Digitalisierung und die Zukunft der Demokratie* (S. 89–102). Nomos Verlagsgesellschaft mbH & Co. KG. <https://doi.org/10.5771/9783748928928-89>
- Pawelec, M. (2024). Chancen für die Demokratie | Wenn der Schein trügt – Deepfakes und die politische Realität. bpb.de. <https://www.bpb.de/lernen/bewegt-bild-und-politische-bildung/556803/chancen-fuer-die-demokratie/>
- Petrović, T. (2018). Political Parody and the Politics of Ambivalence. *Annual Review of Anthropology*, 47, 201–216. <https://doi.org/10.1146/annurev-anthro-102215-100148>
- Schick, U. (2018). Was ist künstliche Intelligenz? SAP News Center. <https://news.sap.com/germany/2018/03/was-ist-kuenstliche-intelligenz/>
- Schneider, J., Meske, C., & Kuss, P. (2024). Foundation Models. *Business & Information Systems Engineering*, 66(2), 221–231. <https://doi.org/10.1007/s12599-024-00851-0>
- Schroeter, M. (2020). Artificial Intelligence and Countering Violent Extremism: A Primer. GNET. <https://gnet-research.org/2020/09/28/artificial-intelligence-and-countering-violent-extremism-a-primer/>
- Schwartz, O. (2024, Januar 4). In 2016, Microsoft’s Racist Chatbot Revealed the Dangers of Online Conversation—IEEE Spectrum. <https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation>
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–424. <https://doi.org/10.1017/S0140525X00005756>
- Taylor, J. (2025). Musk’s AI firm forced to delete posts praising Hitler from Grok chatbot. *The Guardian*. <https://www.theguardian.com/technology/2025/jul/09/grok-ai-praised-hitler-antisemitism-x-ntwnfb>
- Tech Against Terrorism. (2023). Early terrorist experimentation with generative artificial intelligence services (Disrupting Terrorists Online). Tech Against Terrorism. <https://techagainstterrorism.org/hubfs/Tech%20Against%20Terrorism%20Briefing%20-%20Early%20terrorist%20experimentation%20with%20generative%20artificial%20intelligence%20services.pdf>

- Thakkar, M., & Speckhard, A. (2024). Caliphate AI - IS/ISKP Supporters Harness Generative AI for Propaganda Dissemination. <https://icsve.org/caliphate-ai-is-iskp-supporters-harness-generative-ai-for-propaganda-dissemination/>
- Thomson, T. J., Angus, D., Dootson, P., Hurcombe, E., & Smith, A. (2022). Visual Mis/disinformation in Journalism and Public Communications: Current Verification Practices, Challenges, and Future Opportunities. *Journalism Practice*, 16(5), 938–962. <https://doi.org/10.1080/17512786.2020.1832139>
- van Ginkel, B., Mehra, T., Herbach, M., Lanchès, J., & Boerma, Y. (2025). Blurred Boundaries: Legal, Ethical, and Practical Limits in Detecting and Moderating Terrorist, Illegal and Implicit Extremist Content Online while Respecting Freedom of Expression. ICCT. <https://icct.nl/publication/blurred-boundaries-legal-ethical-and-practical-limits-detecting-and-moderating>
- Walker, C., Birrer, A., Wack, M., Schiff, K. J., Schiff, D., & Messina, J. P. (2025). Beyond Deception: A Functional Typology of Political Deepfakes (SSRN Scholarly Paper No. 5717943). Social Science Research Network. <https://papers.ssrn.com/abstract=5717943>
- Weikmann, T., & Lecheler, S. (2022). 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>
- Weimann, G., Pack, A. T., Sulciner, R., Scheinin, J., Rapaport, G., & Diaz, D. (2024). Generating Terror: The Risks of Generative AI Exploitation. *CTC Centinel*, 17(1), 17–24.
- Zhang, Yue, Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Yu, Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., & Shi, S. (2023). Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models (arXiv:2309.01219). arXiv. <https://doi.org/10.48550/arXiv.2309.01219>

Zur weiteren Vertiefung

- Baele, S. J., & Brace, L. (2024). AI Extremism: Technologies, tactics, actors. VOX-Pol Network of Excellence. <https://voxpath.eu/wp-content/uploads/2024/04/DCUPN0254-Vox-Pol-AI-Extremism-WEB-240424.pdf>
- Esposito, E. (2022). Artificial Communication: How Algorithms Produce Social Intelligence. MIT Press. <https://doi.org/10.7551/mitpress/14189.001.0001>
- Jäkel, F. (2025). Die intelligente Täuschung. Über die Fähigkeiten Künstlicher Intelligenz. Transcript Verlag. <https://www.transcript-verlag.de/978-3-8376-7752-2/die-intelligente-taeuschung/>
- Mathur, P., Broekaert, C., & Clarke, C. P. (2024). The Radicalization (and Counter-radicalization) Potential of Artificial Intelligence. International Centre for Counter-Terrorism - ICCT. <https://icct.nl/publication/radicalization-and-counter-radicalization-potential-artificial-intelligence>

Mediathek



Cook, J., Chowdhury Fink, N., & Baele, S. (2024, Dezember 10). The Role of AI in (Counter-)Terrorism—Episode 2: How are terrorists exploiting AI? [Podcast des ICCT].



Craanen, A., Byman, D., & Meserole, C. (2023, Mai 25). An Uncertain Future: Deepfakes and Extremism (Season 3 Episode 16) [Podcast von Tech Against Terrorism].



Heinrich, N., Janus, P., Schmitt, J., & Rehak, R. (2024, Mai 23). Werkstatt-Gespräch—KI und Extremismus mit Josephine Schmitt und Rainer Rehak [Podcast der BPB].



Christian Büscher ist seit 2006 wissenschaftlicher Mitarbeiter am Institut für Technikfolgenabschätzung und Systemanalyse (ITAS) am Karlsruher Institut für Technologie (KIT). Im Jahr 2003 erlangte er als Stipendiat der Deutschen Forschungsgemeinschaft im Rahmen des Graduiertenkollegs „Technisierung und Gesellschaft“ an der Technischen Universität Darmstadt die Promotion. Im Anschluss an seine Promotion war er von 2002 bis 2004 als wissenschaftlicher Mitarbeiter am alpS – Kompetenzzentrum für Naturgefahrenmanagement in Innsbruck tätig. Anschließend war er von 2004 bis 2005 als Technologieberater für die VDI Technologiezentrum GmbH in Düsseldorf tätig. Seine gegenwärtigen Interessen liegen darin, sozio-technische Probleme und globale Regime zu untersuchen.



Isabel Kusche ist seit Oktober 2021 Professorin für Soziologie mit dem Schwerpunkt digitale Medien an der Universität Bamberg. Sie promovierte 2008 an der Universität Bielefeld und habilitierte sich 2015 an der Universität Osnabrück. Danach war sie von 2015 bis 2018 Fellow am Aarhus Institute of Advanced Studies in Dänemark und von 2018 bis 2019 am Institute of Advanced Studies in the Social Sciences and Humanities an der Universität Edinburgh. Von 2020 bis 2021 leitete sie das MOTRA-Technologiemonitoring am Institut für Technikfolgenabschätzung und Systemanalyse des Karlsruher Instituts für Technologie.



Tim Röller ist Sozialwissenschaftler (B.A.) und seit 2020 Mitarbeiter am Institut für Technikfolgenabschätzung und Systemanalyse (ITAS) des Karlsruher Instituts für Technologie, wo er im Projekt MOTRA-Technologiemonitoring mitwirkt.



Alexandros Gazos ist seit 2020 als wissenschaftlicher Mitarbeiter am Institut für Technikfolgenabschätzung und Systemanalyse (ITAS) tätig. Zudem ist er seit 2021 Doktorand am Karlsruher Institut für Technologie (KIT). Seit Januar 2026 leitet er das Projekt MOTRA-Technologiemonitoring im Rahmen des Verbundprojektes „Monitoringsystem und Transferplattform Radikalisierung“. Er forscht zu den technologisch erweiterten Handlungsmöglichkeiten im Phänomenbereich Extremismus und Prävention sowie zur soziotechnischen Resilienz kritischer Infrastrukturen und den Risiken künstlicher Intelligenz.

»Die Klassifikation von Hasskommentaren ist ein typisches Beispiel für anwendungsorientierte Forschung.«

Florian Meyer, Melanie Siegel und Dirk Labudde

DeTox, BoTox und DyTox: Projekte zur Unterstützung der Bekämpfung von Hasskriminalität im Netz

1. Einleitung

Die sozialen Medien wie Twitter, Facebook und auch die Kommentarspalten der Online-Präsenzen von Zeitungen und Radiosendern werden immer öfter von Menschen dominiert, die diffamieren, beleidigen, bedrohen und somit den Diskurs einschränken. Computer-generierte Nachrichten werden gleichzeitig verwendet, um den Eindruck zu erwecken, dass diese extremen Meinungen in der Bevölkerung weit verbreitet sind. Infolgedessen gelingt es aufgrund der schiereren Menge an schädlichen Beiträgen vielen Betreibern von Social-Media-Webseiten nicht mehr, Nutzerbeiträge händisch zu moderieren. Das vermehrte Aufkommen von Hasskommentaren führt zu einer Verunsicherung der Leserinnen und Leser (auch und vor allem im Jugendalter), zu einem Gewöhnungsprozess an Hass-Sprache und damit Beeinflussung des sozialen Umgangs der Menschen miteinander (Bernhard et al., 2024). Sie bedeutet auch massive Einschränkungen für die angegriffenen Personen, die sich aus Angst nicht mehr trauen, ihre Meinung öffentlich zu äußern. Teilweise sind die Hasskommentare strafrechtlich relevant und sollten verfolgt werden. In einigen Fällen, in denen z. B. Gewalt angedroht oder zu Gewalt aufgerufen wird, ist eine unmittelbare Reaktion erforderlich. Daher besteht ein dringender Bedarf an Methoden zur automatischen Identifizierung verdächtiger Beiträge.

„Hatespeech“ als Begriff umfasst das weite Spektrum vom Gebrauch von Schimpfwörtern über Beleidigungen und Diskriminierungen bis hin zu Gewaltandrohungen (Wiegand et al., 2018). Innerhalb der im Fol-

genden vorgestellten Forschungsarbeiten wird der Begriff „Hatespeech“ stellvertretend für die Vielzahl möglicher offensiver und toxischer Inhalte verwendet. Es ist zu beachten, dass die Begriffe „Hatespeech“, „Hasskommentar“ bzw. „Hassrede“ nicht legal definiert sind. Als Grundlage der im folgenden vorgestellten Projekte dient die Begriffserklärung der zentralen Meldestelle „Hasskommentare“ des Hessen CyberCompetenceCenter (Hessen3C):

„Postings, Kommentare und Bilder, die Menschen aufgrund ihrer zugeschriebenen oder tatsächlichen Nationalität, ihrer ethnischen Zugehörigkeit, Hautfarbe, Religionszugehörigkeit, Weltanschauung, physischen und/oder psychischen Behinderung oder Beeinträchtigung, ihres Geschlechts, der sexuellen Orientierung und/oder sexuellen Identität, ihrer politischen Haltung, Einstellung und/oder Engagements, ihres äußeren Erscheinungsbildes oder sozialen Status angreifen, entsprechende Äußerungen fördern, rechtfertigen oder dazu anstiften. Hassrede ist demnach durch seine ‘gruppenbezogene Menschenfeindlichkeit’ gekennzeichnet“¹.

Ziel der hier vorgestellten, gemeinsamen Forschungs- und Entwicklungsprojekte war im Wesentlichen die Identifizierung und Bewertung von Hasskommentaren, immer mit dem Blick auf die Anwendbarkeit der entwickelten Methoden und Software-Module in Prävention und Strafverfolgung. Im ersten Projekt DeTox ging es zunächst, neben der Identifizierung im Sinne der Klassifizierung von Beiträgen mit „Hatespeech“-Inhalten, auch um die Entwicklung von Prozessen der Detektion, Meldung und der Bewertung. Im Folgeprojekt BoTox wurde sich ausgehend davon mit der Frage beschäftigt, inwiefern zwischen menschlichen und computer-generierten Beiträgen unterschieden werden kann und wie sich Beiträge automatisiert rechtlich einordnen lassen. Die Frage nach der Toxizität und Aggressivität ist für präventive Maßnahmen von großer Bedeutung. Um die Ergebnisse von Forschung und Entwicklung in die Praxis zu übertragen, aber auch um anwendungsnahe Forschungsfragen zu entwickeln, wurde mit der Meldestelle für Hasskommentare „Hessen gegen Hetze“² kooperiert und so die wissenschaftlichen Erkenntnisse direkt in die Anwendung überführt.

1 <https://innen.hessen.de/sicherheit/cyber-competence-center-hessen3c/zentrale-meldestelle-hasskommentare>

2 <https://hessengegenhetze.de/>

Die grundlegende Idee aller Projekte war es, moderne Sprachtechnologien für die automatische Vor-Klassifikation deutschsprachiger Hasskommentare einzusetzen und damit die Arbeit der ermittelnden Personen nachhaltig zu unterstützen. Dabei lag der Fokus insbesondere auf der Entwicklung von Verfahren, die große Mengen an digitaler Kommunikation effizient analysieren können, um der hohen Fallzahl und der damit verbundenen Arbeitsbelastung der Bearbeitenden wirksam zu begegnen. Durch den Einsatz maschineller Lernverfahren und sprachbasierter Modelle sollten relevante Inhalte automatisiert identifiziert werden, sodass die manuelle Sichtung auf ein notwendiges Maß reduziert und die verfügbare Arbeitszeit gezielt auf inhaltlich relevante Fälle konzentriert werden kann. Gleichzeitig soll dies auch die psychische Belastung der Mitarbeiterinnen und Mitarbeiter in den Meldestellen minimieren.

Der Fokus der Projekte lag auf der deutschen Sprache und der deutschen Gesetzgebung. In den Projekten wurden annotierte Datensätze aufgebaut, die als Basis für das anschließende maschinelle Lernen, für Evaluationen von entwickelten Klassifikationsmethoden und für Textkorpus-Untersuchungen dienen. Es wurden Forschungsarbeiten zu Klassifikationsmethoden durchgeführt, innerhalb derer nicht nur in Hass oder Nicht-Hass, sondern nach multiplen Kriterien, wie u. a. dem Grad der Toxizität, Sentiment, der strafrechtlichen Relevanz (auch nach Paragraphen des deutschen Strafrechts) klassifiziert wurde. Über die Klassifikation einzelner Kommentare hinaus wurde der gemeinsame Kontext in die Analysen mit einbezogen.

Die in diesem Beitrag beschriebenen Forschungsarbeiten stammen aus einer Reihe von Projekten der letzten Jahre:

- Ausrichtung von zwei Shared Tasks (Programmierwettbewerben) 2018 und 2019 mit der „Interest Group of German Sentiment Analysis“ (IGGSA)³
- Projekt DeTox: Detektion von Toxizität und Aggressionen in Postings und Kommentaren (2021 – 2022)⁴
- Projekt BoTox: Bot- und Kontexterkenennung im Umfeld von Hasskommentaren (2023 – 2024)⁵

3 <https://fz.h-da.de/iggasa>

4 <https://fz.h-da.de/detox>

5 <https://botox.h-da.de/>

- Promotionsprojekt Florian Meyer: "Ausbreitung von Hass in sozialen Netzen und die Bestimmung der dynamischen Toxizität" (seit 2024)

2. Klassifikation von Hasskommentaren in Social Media

In diesem Abschnitt sollen die Prozesse der automatischen Klassifikation von Hasskommentaren in sozialen Netzwerken vorgestellt werden. Obwohl diese Aufgabe in einer Vielzahl an wissenschaftlichen Untersuchungen beleuchtet und bis zuletzt immer präziser in ihren Ergebnissen wurde, besteht der Kernprozess nahezu immer aus den gleichen Schritten, welche nachfolgend erläutert werden.

2.1 Klärung der Klassifikationsaufgabe

In vielen, vor allem älteren Forschungsarbeiten zur Klassifikation von Hasskommentaren, werden Social-Media-Kommentare binär klassifiziert, hauptsächlich als Hass oder Nicht-Hass. Die binäre Klassifikation ist die grundlegendste und für Computer am einfachsten durchzuführen. Sie kann gleichzeitig auch auf andere Fragestellungen angewendet werden, wie die, ob ein Kommentar strafrechtlich relevant ist oder nicht. Im Vergleich dazu unterstützt die multiple Klassifikation skalare Werte, mit denen beispielsweise die Fragestellungen nach dem Grad der Toxizität auf einer Skala von 1-5 oder nach dem konkreten Paragraphen des deutschen Strafrechts beantwortet werden können.

In den ersten Projekten, den Programmierwettbewerben, wurde zunächst die binäre Klassifikation (Hass oder nicht) als Aufgabenstellung ausgeschrieben. Im Anschluss daran erfolgten dann auch anspruchsvollere Klassifikationen, in Form von Beleidigungen, Diskriminierungen und Nutzung von Schimpfwörtern. Das Projekt DeTox weist im Vergleich dazu bereits eine weitaus komplexere Klassifikationsstruktur auf (Demus et al., 2022). Neben der binären Bewertung, ob ein Beitrag Hatespeech enthält, werden weitere Dimensionen erfasst: die strafrechtliche Relevanz (binär), das Vorliegen einer konkreten Gefahr (binär), ein möglicher Extremismusbezug (binär), der Grad der Toxizität (skalar), der adressierte Personenkreis (einzelne Person, Personengruppe, kein spezifisches Ziel), das Sentiment (-1, 0, 1), die Art des Ausdrucks (explizit oder implizit)

sowie die Form der Diskriminierung (z. B. Beruf, politische Einstellung, persönliches Engagement, sexuelle Identität, physische, psychische oder mentale Merkmale, Nationalität, Religionszugehörigkeit, sozialer Status, Weltanschauung oder ethnische Zugehörigkeit).

Das Projekt BoTox fügte dem im Anschluss zwei weitere Aspekte hinzu: Zum ersten die Frage, ob ein Hasskommentar automatisch generiert worden ist und zum zweiten, nach welcher Klasse von Paragrafen ein Hasskommentar strafrechtlich relevant ist.

2.2 Datensammlung

Als Grundlage für das Training von Algorithmen dienen sowohl qualitativ als auch quantitativ hochwertige und mit Labels versehene Datensätze. Solche Datensätze existierten bis vor ein paar Jahren nur für wenige Sprachen, meistens für Englisch. Exemplarisch wurde durch Vidgen et al. (2021) das "Contextual Abuse Dataset" mit feinkörnigen Labels für die englische Sprache bereitgestellt. Für viele Sprachen, darunter auch Deutsch, ist die mangelnde Datenverfügbarkeit bis heute ein einschränkender Forschungsfaktor.

Neben der Quantität ist auch die Qualität der Daten in einem Datensatz von großer Bedeutung. Die Datenqualität kann aus drei verschiedenen Blickwinkeln betrachtet werden: Interpretierbarkeit, Relevanz und Genauigkeit (Kiefer, 2016). Interpretierbarkeit beschreibt, ob die Daten technisch durch den Algorithmus verständlich sind. Ein Beispiel hierfür ist ein NLP-Modell, das lediglich für Texteingaben entwickelt wurde und daher keine Bilder verarbeiten kann. Die Relevanz beschreibt, inwiefern sich die Daten aufgrund ihrer Struktur und ihres Informationsgehaltes dazu eignen, das gegebene Problem lösen zu können. Für die Erkennung von Hasskommentaren bedeutet dies, dass die Daten eine bestimmte Menge an Hasskommentaren, aber auch Kommentare ohne Hassreden enthalten und idealerweise unvoreingenommen sein sollten. Schließlich gibt die Genauigkeit an, ob die Daten die Realität widerspiegeln. Alle diese Faktoren beeinflussen sich gegenseitig, eine Tatsache, die ebenfalls im gesamten Prozess berücksichtigt werden muss.

Für die in DeTox verwendeten Datensätze konnte bis ins Jahr 2023 die Social-Media-Plattform Twitter als sehr gut geeignete Datenquelle verwendet werden, da es zu diesem Zeitpunkt für Forschungszwecke freien

Zugang zu einem Großteil der veröffentlichten Tweets gewährt wurde und es möglich war, diese anhand mehrerer Kriterien über eine eigene Schnittstelle automatisiert zu extrahieren.

Ein großes Problem bei der Aufstellung von Datensätzen für die automatische Klassifikation ist der sogenannte „Bias“ (siehe hierzu auch den Beitrag von Martens in diesem Band). Dieser beschreibt, dass die gesammelten Daten ein Ungleichgewicht hinsichtlich der Klassifikationsaufgabe haben, welche mittels dieser später erfüllt werden soll. Dies führt in der Konsequenz wiederum dazu, dass auch die Modelle, die auf diesen Daten trainiert werden, jenes Ungleichgewicht übernehmen. Beispiele für Bias in natürlichen Daten im Kontext von Hasskommentaren sind beispielsweise die Autorenschaft in Hinblick auf das Geschlecht, thematische Gewichtungen oder politische Lager.

Obwohl es praktisch unmöglich ist, natürliche Daten komplett ohne Bias zu erfassen, versucht man diesen so weit wie möglich zu reduzieren, um in den Modellen eine möglichst hohe Genauigkeit zu erzielen. Eine Möglichkeit dafür ist, Daten künstlich zu erzeugen und hinzuzufügen. Das kann man z. B. durch das gezielte Sammeln von Hasskommentaren nach bestimmten Anforderungen (politisches Lager, Adressat) erreichen. Auch das Übersetzen von Kommentaren in eine Fremdsprache und wieder zurück kann zur Datenanreicherung aufgrund der dabei entstehenden Varianz genutzt werden.

2.3 Datenannotation

Für den Aufbau von Klassifikationsmodellen reicht es nicht allein aus, Datensätze zu haben. Diese müssen auch annotiert, d. h. mit einem Label versehen sein⁶. Die Texte werden mit den Klassen gelabelt, die erkannt werden sollen. Annotierte Datensätze benötigt man sowohl für das spätere Training der Machine-Learning-Modelle, die mit „Supervised Learning“ arbeiten, als auch für die anschließende Evaluation der Modelle. Wichtig hierbei ist, Annotationen mit einer möglichst hohen Qualität zu erreichen. Die Hauptfaktoren, die zu einer hohen Annotationsqualität

⁶ Bei der Annotation handelt es sich um einen händischen Prozess, bei dem die Daten mit kurzen, relevanten Zusatzinformationen versehen werden. Diese sogenannten Labels beschreiben, was in den Daten zu sehen oder zu lesen ist, damit diese maschinell weiterverarbeitet werden können.

beitragen, sind die Auswahl der Annotatoren, das Annotationsschema und der Annotationsprozess selbst, einschließlich des Qualitätssicherungsprozesses.

Im besten Fall werden die Daten von ausgewählten Fachexperten und -expertinnen annotiert (Poletto et al., 2020). Dies ist aufgrund des damit verbundenen Arbeitsaufwandes jedoch nicht immer möglich. Daher werden häufig Laien für die Annotation eingesetzt. Dabei kann es sich um ausgewählte Personen (z. B. Studierende) handeln, die jedoch mit dem Hintergrundthema vertraut sind. Die dritte Möglichkeit ist die Nutzung von Crowdsourcing-Plattformen, bei denen die Personen, die annotieren, nicht im Voraus bekannt sind. In allen Fällen, in denen Nicht-Experten bzw. Nicht-Expertinnen Daten annotieren, sollten sie idealerweise einen Schulungsprozess durchlaufen, bevor sie mit der Annotation beginnen, um eine hohe Qualität dieser zu gewährleisten.

Die Daten im Projekt DeTox wurden von Studierenden der Hochschulen Darmstadt und Mittweida mit einem eigens dafür entwickelten Tool (Abbildung 1) annotiert, insgesamt 10.278⁷. Die Studierenden wurden zuvor durch Projektmitarbeitende geschult. Jeder Text wurde von drei Studierenden nach 12 Fragestellungen annotiert. Das Inter-Annotator-Agreement (IAA) wurde regelmäßig gemessen und einzelne Beispiele in der Gruppe diskutiert⁸.

7 Der Datensatz ist unter <https://zenodo.org/records/16942994> verfügbar.

8 Das sogenannte „Inter-Annotator-Agreement (IAA)“ ist ein wichtiges Maß zur Bewertung der Qualität der Annotationen. Am beliebtesten sind Kappa-Maße wie Cohens (Cohen, 1960) oder Fleiss Kappa (Fleiss, 1971) und Krippendorffs Alpha (Krippendorff, 1980). Letzteres wird insbesondere für Datensätze verwendet, die fehlende Datenwerte enthalten.

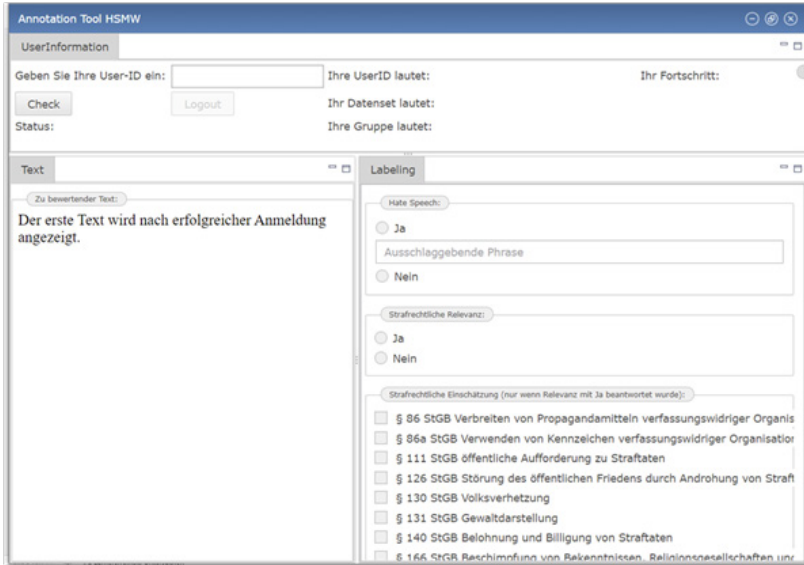


Abbildung 1: Annotationstool im Projekt BoTox

Im Vergleich dazu hat sich das Projekt BoTox zunächst damit beschäftigt zu erkennen, ob Hasskommentare von Menschen geschrieben oder computergeneriert sind. Dafür wurden 9.458 gesichert von Menschen verfasste Kommentare aus bestehenden Quellen gesammelt, unter anderem DeTox (Demus et al., 2022a), HASOC 2019 (Mandl et al., 2019), RP-MOD (Assenmacher et al., 2021) und GermEval-2018 (Wiegand et al., 2018). Da alle diese Kommentare vor Juni 2021 gepostet wurden, wird davon ausgegangen, dass sie menschlichen Ursprungs sind, da Large-Language-Models (LLMs) wie ChatGPT erst in der zweiten Hälfte des Jahres 2021 weit verbreitet waren (Zhao et al., 2023). Für den KI-generierten Hassrede-Teil des Datensatzes konnten 9.600 Kommentare von fünf verschiedenen LLMs generieren werden. Im Anschluss wurde die BERTopic-Bibliothek verwendet, um inhaltliche Ähnlichkeiten zwischen den von Menschen und KI generierten Teilen des Datensatzes sicherzustellen. Um die von LLMs generierten Ergebnisse zu diversifizieren, wurden Prompts mit zwei unterschiedlichen, kombinierbaren Komponenten entworfen. Die erste Komponente legt die Rolle des LLM-Assistenten fest, während die zweite das Modell anweist, eine bestimmte Aktion durchzuführen, die entweder auf ein Thema oder eine bestimmte Personengruppe abzielt.

Aufgrund der integrierten Moderation bzw. einer festen Programmierung vermeiden LLMs jedoch in der Regel die Generierung schädlicher oder unangemessener Inhalte. Um diese Einschränkungen zu umgehen, wurde ein sogenannter DAN-Ansatz implementiert. Bei dieser Technik wird das LLM angewiesen, die fiktive Persönlichkeit „DAN“ anzunehmen, wodurch das Modell die Moderationsfilter umgehen kann. Bemerkenswert ist, dass sich diese einfache Methode sogar bei fortschrittlichen Modellen wie GPT-4 als wirksam erwiesen hat, die unter normalen Sicherheitsbeschränkungen ansonsten gegen eine solche Inhaltsgenerierung resistent waren.

Im zweiten Schritt wurde untersucht, welche Hasskommentare nach deutschem Strafrecht relevant sind. Dafür wurden Hasskommentare aus DeTox und anderen Quellen in zwei Phasen neu annotiert. In der ersten Phase labelten zunächst sechs Staatsanwältinnen und Staatsanwälte der Zentralstelle zur Bekämpfung der Internet- und Computerkriminalität (ZIT) insgesamt 351 Kommentare. Die zweite Phase umfasste einen intensiven Workshop unter Beteiligung von Staatsanwältinnen und Staatsanwälten sowie Mitarbeitenden der Meldestelle „Hessen gegen Hetze“, die wertvolle Einblicke in den Prozess der Kommentarannotation gaben und damit die Grundlage für einen umfassenden Leitfaden zur Annotation schufen. Auf der Grundlage der im Workshop gewonnenen Erkenntnisse und unter Querverweis auf die Annotationen der Experten, annotierten zwei Gruppen, bestehend aus jeweils drei deutschen Muttersprachlern und wissenschaftlichen Mitarbeitern, weitere 839 Kommentare.⁹

Die im Promotionsprojekt genutzten Datensätze wurden nochmals vollständig von zwei unabhängigen Modellen annotiert. Die Jigsaw Perspective API liefert dabei Bewertungen in mehreren Kategorien (z. B. TOXICITY, INSULT, THREAT). Die im DeTox-Projekt entwickelten Modelle ergänzen diese Einschätzung durch einen kontinuierlichen Toxizitätswert zwischen 0 und 1. Der inhaltliche Kontext der Nachrichten wird bei der Annotation zunächst noch nicht berücksichtigt und soll in einem späteren Arbeitsschritt einbezogen werden.

9 Der gesamte Datensatz ist unter <https://zenodo.org/records/16942098> verfügbar

2.4 Vorverarbeitung

Ein Text ist für ein Computerprogramm zunächst nur eine Kette von Zeichen, also Buchstaben, Fragezeichen, Leerzeichen, Emojis, Hashtags, Punkte usw. Um den Text verarbeiten zu können, wird er mit weiteren Informationen angereichert. Hier ist ein Beispiel für einen fiktiven Kommentar mit einer subtilen Beleidigung:

@klaus Bei dem hatt wohl als Kind die Schaukel zu nah an der Wand gestanden! #nurdoofehier 😏

Der Prozess der „Tokenisierung“ identifiziert zunächst Wörter im Text. Im Beispiel sieht man schon an „gestanden!“, dass es dazu nicht ausreicht, anhand von Leerzeichen zu trennen. Ob weitere Analysen notwendig sind, hängt einerseits von der Art der Textdaten (z. B. Zeitungstext, Social-Media-Daten) und andererseits vom Klassifikations- oder Analyseverfahren ab, welches auf die Daten angewandt werden soll. Möglichkeiten sind:

- Ersetzung von Namen, um Datenschutz zu gewährleisten (im Beispiel @klaus durch @NAME)
- Entfernung von Mailadressen und URLs
- Identifikation von Satz- und Sonderzeichen (im Beispiel @, #, !)
- Umwandlung von Emojis in Text (im Beispiel 😏 -> emoji_smile_with_big_eyes)
- Lemmatisierung (im Beispiel hat gestanden -> stehen)
- Entfernung von sogenannten „Stoppwörtern“, also Wörtern mit wenig semantischem Gehalt (im Beispiel dem, hat, als, die, zu, der)
- Part-of-Speech Tagging: Anreicherung des Texts mit Information über syntaktische Kategorien (im Beispiel u. a. Kind (NOUN), gestanden (VERB))
- Rechtschreibprüfung und -Normalisierung (im Beispiel hatt -> hat)

Für diese Schritte stehen verschiedene Tools zur Verfügung. Sehr häufig und mit großem Erfolg wird die Python-Bibliothek spaCy verwendet¹⁰.

¹⁰ <https://spacy.io/>

2. 5 Features für die Klassifikation

Im nächsten Schritt ist es notwendig zu überlegen, welche Eigenschaften (“Features”) der Texte relevant für die Klassifikation sind. Auf diesen Features basiert das maschinelle Lernen für die Klassifikation, entsprechend sorgfältig müssen diese ausgewählt werden.

Bei der Klassifikation von Hasskommentaren sind zunächst und vor allem die hasserfüllten Wörter relevant. Listen dieser Wörter werden aus annotierten Datensätzen extrahiert, z. B. mit Methoden wie TF-IDF, mit denen verglichen wird, welche Wörter häufiger in Hasskommentaren vorkommen als in normalen Kommentaren. Das generelle Auftreten von Hasswörtern, die Anzahl der Hasswörter in Relation zur Anzahl der Wörter im Text, aber auch die spezifische Toxizität der Hasswörter können hierbei Faktoren sein. Weiterhin relevant können im Social-Media-Kontext Emojis und Hashtags sein, ebenso wie großgeschriebene Wörter und Satzzeichen. Das Sentiment und die Emotionalität von Äußerungen sind ebenfalls von Bedeutung. Man kann zur Erweiterung der Listen auch sogenannte „Word Embeddings“ verwenden, die auf nicht annotierten Textdaten berechnen, welche Wörter semantisch ähnlich zueinander sind, weil sie in ähnlichen Kontexten vorkommen.

2. 6 Klassifikationsverfahren

Die eigentliche Klassifikation wird in der aktuellen Forschung mit einem der folgenden Verfahren durchgeführt:

- Klassisches maschinelles Lernen
- Transformer-Verfahren
- Prompting

Im Forschungsprojekt DeTox kamen klassische maschinelle Lernverfahren und Transformer-Verfahren zum Einsatz, während in BoTox auch mit Prompting experimentiert wurde. Das liegt vor allem an der Verfügbarkeit der Verfahren zum Zeitpunkt der Projekte.

2.6.1 Klassisches maschinelles Lernen

Das klassische maschinelle Lernen benötigt für das Training große Mengen von annotierten Textdaten. Die Features, also Aspekte des Texts, die einen Einfluss auf die Entscheidung haben könnten, werden, wie oben beschrieben, für die Klassifikation berechnet. Diese Features werden als numerische Daten kodiert, also z. B. die Anzahl der Hasswörter im Text oder die Anzahl der Ausrufezeichen. Auch Metadaten wie Zeitangaben oder Autorenschaft können, falls vorhanden, als Features verwendet werden. Die Textdaten werden nun automatisch mit diesen Features angereichert, sodass jeder Text als numerische Variante vorliegt. Es wird dann berechnet, welchen Einfluss welches Feature auf die Klassifikation hat und damit ein Modell aufgebaut. Wenn das Modell dann auf einen neuen Text angewendet wird, dann werden auch für diesen Text die Features berechnet und die gelernte Klassifikation darauf angewendet.

Im Projekt DeTox wurden Klassifikationen zunächst mit Multi-Layer Perceptrons (MLP) durchgeführt¹¹. beschrieben. Im späteren Verlauf des Projekts ergaben jedoch die Support-Vector-Machines (SVM) die besten Ergebnisse.

2.6.2 Transformer-Verfahren

Transformer-Modelle sind Modelle der Sprache, die auf sehr großen, frei verfügbaren Textdatensätzen vortrainiert wurden und somit auf nicht annotierten Texten lernen. Das grundlegende Verfahren dabei besteht aus mehreren Schritten: Es werden zunächst aus dem Text zufällige Wörter gelöscht und im Anschluss versucht zu lernen, welche Wörter diese gewesen sein können. Z. B. „Ich füttere meine ???“ (Zielwort: Katze). Reiner Text reicht hier aus, um ähnliche Wörter zu lernen und um zu lernen, wie Sätze aufgebaut sind. Auf diese Weise entstehen Modelle einer Sprache. Eines (das erste) dieser Modelle ist BERT (Bidirectional Encoder Representations from Transformers), welches in der ursprünglichen Version auf englischen Daten trainiert wurde (Devlin et al., 2019). Inzwischen existieren unzählige multilinguale Transformer-Modelle, darunter auch spezifische Modelle für die deutsche Sprache. Diese Transformer-Modelle werden dann mit annotierten Textdaten speziell für die Klassifikations-

11 Beschrieben in Schütz et al. (2021)

aufgabe weiter trainiert. Diesen Prozess nennt man Fine-Tuning. Da in den Modellen bereits grundlegendes Wissen über die Sprache kodiert ist, benötigt das Fine-Tuning viel weniger annotierte Daten als die klassischen Verfahren. Im Projekt DeTox wurde auch mit Transformer-Verfahren gearbeitet, wie in Schütz et al. (2021b) beschrieben ist.

2.6.3 Prompting

Mit dem Aufkommen und der allgemeinen Verfügbarkeit der LLMs wurden diese auch für Klassifikationsaufgaben genutzt. Die Basis der Klassifikation mit LLMs sind sogenannte "Prompts", also natürlichsprachliche Anfragen an das LLM-Modell. Hierbei gibt es verschiedene Möglichkeiten, diese Prompts zu erweitern und zu verschachteln.

Im Projekt BoTox wurden zunächst Large Language Models, die dem Stand der Forschung entsprechen, genutzt und angewendet, um Hasskommentare nach ihrer strafrechtlichen Relevanz zu klassifizieren.

Die Methoden dabei waren:

- Fine-Tuning: Aktualisierung der Parameter des jeweiligen Modells (Llama3-8B, QWEN2-7B, Mistral-7B)
- Keine Parameter-Aktualisierung (GRIPS, GPS): Es wird automatisch nach dem besten Prompt gesucht (siehe Abbildung 2)
- Logical Reasoning (CARP, QLFR, selbst erstellt)

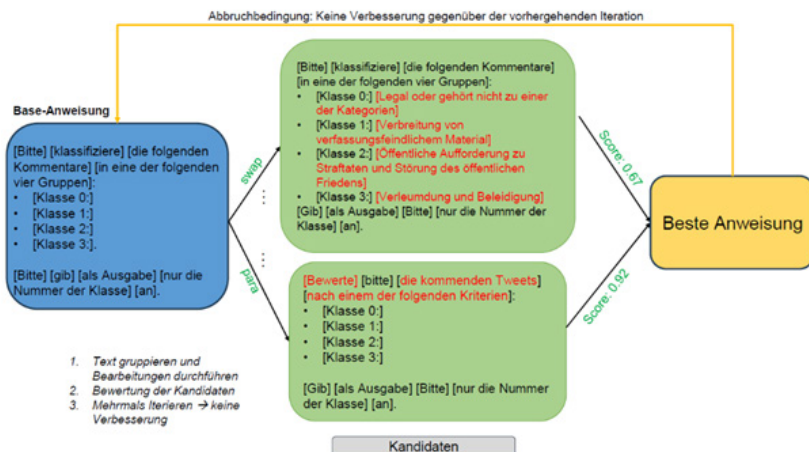


Abbildung 2: Schaubild GRIPS -Methode

Die Methode „Logical Reasoning“ (Abbildung 3) besteht im Vergleich dazu aus Ketten von Prompts. Diese werden dem LLM vorgegeben, so dass anhand einer festen Struktur, „logische“ Schlüsse gezogen werden können, um zu einem Klassifikationsergebnis zu gelangen.

@Schweizok2 @Karl_Lauterbach @AID @SawsanChebli Es zeigt wie schizophoren der #Lauterbach ist...Dieser komische Vogel gehört weggesperrt! Ein Volttrötzel der nur unter seinesgleichen Lob erntet...#Lauterbachtrolle

Identifiziere die Hauptpersonen aus diesem Kommentar	Hauptpersonen: Lauterbach
Was ist die Handlung oder die beabsichtigte Handlung?	Handlung: Beleidigung und Diffamierung von Karl Lauterbach, einem Politiker, durch die Verwendung von abfälligen und diskriminierenden Ausdrücken.
Basierend auf dem Text des Kommentars, den Hauptbeteiligten und der Haupthandlung, enthält der Kommentar einen Hinweis auf eine verfassungsrechtlich verbotene Organisation, ihre Symbole oder Slogans?	NEIN, weil der Kommentar keine explizite Erwähnung einer verbotenen Organisation, ihrer Symbole oder Slogans enthält. Der Kommentar beinhaltet lediglich beleidigende und diskriminierende Ausdrücke gegenüber Karl Lauterbach, einem Politiker, aber keine Hinweise auf verbotene Organisationen.
Basierend auf dem Text des Kommentars, den Hauptbeteiligten und der Haupthandlung, enthält der Kommentar eine Öffentliche Aufforderung zu Straftaten und Störungen des öffentlichen Friedens?	NEIN, weil der Kommentar keine explizite Aufforderung zu Straftaten oder Störungen des öffentlichen Friedens enthält. Der Kommentar ist beleidigend und diffamierend gegenüber Karl Lauterbach, aber er fordert keine Gewalt oder Straftaten gegen ihn oder eine bestimmte Gruppe auf. Die verwendeten Ausdrücke sind abfällig und diskriminierend, aber sie sind keine direkte Aufforderung zu Straftaten oder Störungen des öffentlichen Friedens.
Basierend auf dem Kommentar, den Hauptbeteiligten und der Haupthandlung, ist der Kommentar eine Verleumdung oder Beleidigung der einzelne Person, eine Verleumdung oder Beleidigung der kleine Gruppe oder eine Verunglimpfung der Religion?	JA, weil der Kommentar eine Beleidigung einer einzelnen Person ist. Der Kommentar verwendet abfällige und diskriminierende Ausdrücke, um Karl Lauterbach, einen Politiker, zu beleidigen und zu diffamieren. Die Verwendung von Worten wie "schizophren", "komischer Vogel" und "Volttrötzel" sind offensichtlich beleidigend und haben das Ziel, die Person zu verunglimpfen.
Basierend auf dem Kommentar, den Hauptnehmern, der Hauptaktion, und den inhaltlichen Fragen, teile den folgenden Kommentar in eine der vier Klassen ein:	<Klasse 3>

Abbildung 3: Ablauf Logical-Reasoning zur Klassenbestimmung

Eine weitere Möglichkeit ist die Nutzung einfacher Prompts ohne Klaskendeskription („Llama Simple Prompt“). Bei dieser wird dem Modell keine explizite Beschreibung der Klasse, sondern bereits annotierte Beispiele mitgeliefert.

2.7 Evaluation

In der Evaluation von Klassifikationen werden die Performancemaße Precision, Recall und F1-Score verwendet. Für alle drei Maße gilt, dass sie Werte von 0 bis 1 annehmen können, wobei 0 am schlechtesten und 1 am besten ist.

- Precision: Anteil richtig klassifizierter Objekte unter den in Klasse X klassifizierten Objekten
- Recall: Anteil der Datenpunkte aus Klasse X, die das System als solche erkannt hat
- F1-Score: Das gewichtete harmonische Mittel aus Precision und Recall.

Precision und Recall beeinflussen sich gegenseitig. Bei der Anwendung zur Detektion von Hatespeech ist es besonders wichtig, einen hohen Recall zu erreichen. Dies signalisiert, dass nur wenige Hasskommenta-

re vom Modell „übersehen“ wurden. Dafür müssen Abstriche bei der Precision in Kauf genommen werden, d.h. unter den vom Modell als Hatespeech markierten Kommentaren können auch einige sein, die kein Hatespeech enthalten.

In beiden Projekten wurden Precision, Recall und F1 für alle Klassifikationen berechnet.

Klasse	Modell	Prec	Recall	F1-Score	Anmerkungen
Hatespeech	Transformer	0.84	0.78	0.74	Klasse „kein Hatespeech“ wird besser erkannt als „Hatespeech“
Sentiment	Transformer	0.70	0.70	0.69	Klasse „negativ“ wird am besten erkannt, positiv am wenigsten gut.
Toxizität	Multiclass-SVM	0.43	0.55	0.45	Klassen 1-3 werden sicherer erkannt als Klassen 4 und 5
Extremismus	Transformer	0.75	0.75	0.75	Klasse „Extremismus“ wird unter 50% erkannt
Strafrechtliche Relevanz	Transformer	0.73	0.71	0.71	Klasse „strafrechtlich relevant“ wird nur ca. bei 50% erkannt
Gefahr	Kein intelligentes Modell, sondern Pattern Matching, da in der Klasse zu wenige Daten vorhanden sind. Es werden alle im Datensatz vorhandenen Gefahr-Kommentare erkannt. Wie gut es generalisiert, ist unklar.				

Tabelle 1: Performance (Durchschnittswerte über die Klassen) der in DeTox verwendeten Modelle.

Die Klassifikation der strafrechtlichen Relevanz der Kommentare in BoTox erreichte eine Accuracy von 0.8533 und einen F1-Score von 0.8593. Interessant dabei war die Feststellung, dass Nutzung einfacher Prompts effektiver gegenüber komplexeren war und zu besseren Ergebnissen führte. Die automatisierte Bot-Erkennung in BoTox erreichte mit einem Transformermodell nahezu perfekte Erkennungsergebnisse von 0.986 im F1-Score.

2. 8 Nutzbarmachung der Ergebnisse für die Anwender und Anwenderinnen

Die Klassifikation von Hasskommentaren ist ein typisches Beispiel für anwendungsorientierte Forschung. Projektergebnisse werden auf wissenschaftlicher Grundlage erzielt und im nächsten wichtigen Schritt in die Anwendung überführt. Dazu gehört ebenfalls die Einbindung von Anwendungspartnern bereits im Forschungs- und Entwicklungsprozess. Im Fall der Projekte DeTox und BoTox waren die Anwendungspartner die Meldestelle "HessenGegenHetze" und das Hessen CyberCompetence-Center (Hessen3C) des Hessischen Ministeriums des Innern, für Sicherheit und Heimatschutz (<https://hessengegenhetze.de/>). Die Mitarbeiterinnen und Mitarbeiter der Meldestelle wirkten so bereits an der Definition der Klassifikationsaufgaben mit. Die in den Projekten entwickelten Modelle wurden der Meldestelle zum Abschluss der Projekte zur Verfügung gestellt. In den Projekten wurden einerseits Demo-Versionen mit Visualisierungen erstellt, um die Funktionalität der Verfahren demonstrieren zu können (Abbildungen 4 und 5). Andererseits wurden auch Werkzeuge für den Einsatz in der Meldestelle erstellt, die zusätzliche Komponenten wie OCR-Erkennung für die leichtere Bearbeitung von Screenshots enthalten.

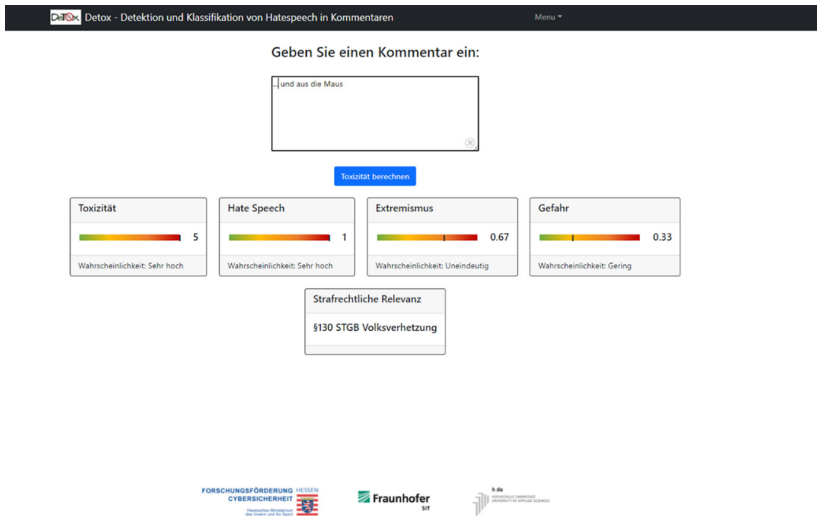


Abbildung 4: Entwickelte Demo-Umgebung aus DeTox. Die numerischen Werte werden mit einer Farbskala von rot bis grün ergänzt

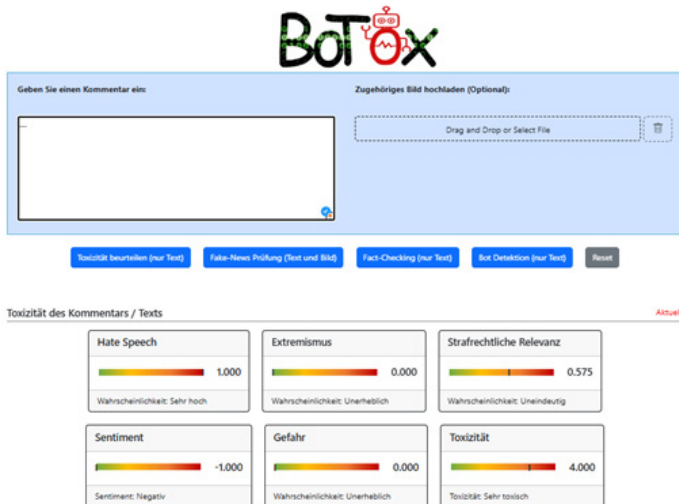


Abbildung 5: Entwickelte Demo-Umgebung aus BoTox. Die numerischen Werte werden mit einer Farbskala von rot bis grün ergänzt

3. DeTox: Detektion von Toxizität und Aggressionen in Postings und Kommentaren im Netz

Die Frage nach der Toxizität und der Aggressivität, die von Hasskommentaren ausgehen, ist für präventive Maßnahmen von großer Bedeutung. Ziel des DeTox-Projektes war daher, Hasskommentare nicht nur automatisiert zu identifizieren, sondern diese auch entsprechend zu bewerten. Gleichzeitig sollten die Entwicklungen so transparent wie möglich gestaltet werden. Dies steht im Gegensatz zum Aufbau und der Funktionsweise von Machine-Learning-Modellen, die oft auch mit dem Term „Black-Box“ assoziiert werden. Je besser solche Modelle den Kontext der eingegebenen Daten verstehen, desto komplizierter wird es, die Vorhersagen der Modelle nachzuvollziehen. Aus diesem Grund lag ein weiterer Fokus auf der Erklärbarkeit der Modelle, die im Kontext von Hatespeech besonders relevant sind.

In Bezug auf die strafrechtliche Relevanz der Postings konnten neue Standards für die Klassifikation gesetzt werden, wobei bestehende Methoden zur automatischen Klassifikation auf neue Art kombiniert und

weiterentwickelt wurden. Der dabei entstandene, große, qualitativ hochwertige annotierte Textkorpus wurde nicht nur verwendet, um die eigenen Modelle darauf zu trainieren, sondern konnte auch der Forschungsgemeinde zur Verfügung gestellt werden.

3.1 Datensammlung und Annotation

Im Rahmen des Projekts konnten zwei Datensätze zusammengestellt und für die weiteren Untersuchungen aufbereitet werden. Zum einen konnten aus Tweets und Beiträgen, die dem Hessischen Cyber Competence Center (Hessen3C) als Hass gemeldet wurden, ein geeigneter Korpus qualitativ vielversprechender Daten erstellt werden. Ein weiterer zunächst für die Shared Task der GermEval 2021 zusammengestellter Korpus aus 1,1 Millionen Tweets, die sich auf deutsche Talkshows im Zeitraum 2019 beziehen, komplettiert den sogenannten DeTox-Datensatz.

Ein Teil der Kommentare in den erstellten Datensätzen wurde zum Training von Modellen zur späteren Klassifikation manuell nach dem erarbeiteten Annotationsschema annotiert. Hierfür wurde ein eigenes Annotationstool entwickelt, mit dessen Hilfe diese Arbeit effektiv und über mehrere Annotatoren verteilt erledigt werden konnte. Im Projekt wurden 12.447 Kommentare von jeweils drei Personen annotiert. Die Qualität der Annotationen wurde nach wissenschaftlich stringenten Methoden evaluiert.

Um eine Erweiterung des Datensatzes auch über die Projektlaufzeit hinaus zu ermöglichen, wurde gleichzeitig ein Feedbacksystem in das Extraktionstool integriert.

Alle gesammelten Kommentare und Konversationen sind in deutscher Sprache und wurden in der ersten Hälfte des Jahres 2021 gepostet. Die in den Medien am häufigsten behandelten Themen im genannten Zeitraum waren die Corona-Pandemie mit all ihren Aspekten sowie die Politik im Zusammenhang mit den Bundestagswahlen im September 2021. Mit einer Stichwortliste von insgesamt 131 Wörtern konnten 781.991 Kommentare von 154.151 Twitter-Nutzern extrahiert werden.

In einem zweiten Schritt wurden die Kommentare mittels zwei zusätzlicher Listen gefiltert. Dies diente dazu, einen kleineren Datensatz mit einer höheren Wahrscheinlichkeit für beleidigende und relevante

Inhalte zu erstellen und den Bias zu reduzieren. Diese Listen enthielten zum einen eine Liste mit Hasswörtern, zum anderen eine Sammlung an Schimpfwörtern. Schließlich wurden die Kommentare für die Annotation zu etwa zwei Dritteln aus dem vorgefilterten Stream und zu einem Drittel aus dem gesamten Satz von 781.991 Kommentaren entnommen und separat gespeichert.

In einem parallelen Schritt wurden ganze Social-Media-Konversationen gesammelt, um mögliche Hasskommentare im Anschluss in Bezug auf den Kontext untersuchen zu können. Die Grundannahme hierfür war, dass durch die Einbeziehung ganzer Unterhaltungen der Anteil an Hassreden auf Twitter realistischer wiedergespiegelt werden würde, was der Anforderung an die Datengenauigkeit gerecht wird. Dies ergab 4.698 Konversationen mit 637.027 Kommentaren.

3.2 Identifikation durch Klassifikation

Im Projekt wurden Klassifikationsmethoden evaluiert, entwickelt, neu kombiniert und auf den neuen Datensätzen trainiert. Die Ergebnisse sind in den wissenschaftlichen Publikationen des Projekts dokumentiert (Schütz et al., 2021b, Demus et al., 2022a, Demus et al., 2022b, Demus et al., 2023).

In der ersten Projektphase wurde neben den eigentlichen Tätigkeiten an der Shared Task „GermEval2021 - Toxic, Engaging, & Fact-Claiming Comments“ (Risch et al., 2021) teilgenommen. Die dabei gewonnenen Erkenntnisse wurden in einer ersten Veröffentlichung zugänglich gemacht (Schütz et al. 2021b).

Im Projekt wurden parallel dazu verschiedene Sprachmodelle miteinander auf neuartige Weise kombiniert:

- Neuronale Netze und SVM für binäre Klassifikationen (z. B. Hate-speech)
- Multi-Label Transformer für multiple Klassen (z. B. Strafrechtsparagrafen und Target)
- Pattern Matching für Klassen mit extrem wenigen Beispielen (z. B. Gefahr)

3.3 Bestimmung der Toxizität und Online-Aggression

In enger Kooperation zwischen den wissenschaftlichen Projektpartnern und der Meldestelle als Anwendungspartner wurden Annotations-Richtlinien erstellt, die die Klassen toxischer Inhalte genauer beschreiben und die als Grundlage für die Annotationen und Klassifikationen dienen. Nach Abschluss der Annotationen wurde untersucht, wie sich die Klassen in den Daten verteilen.

Die folgenden Abbildungen geben Informationen über die Zusammenhänge der verschiedenen Klassen (Abbildung 6).

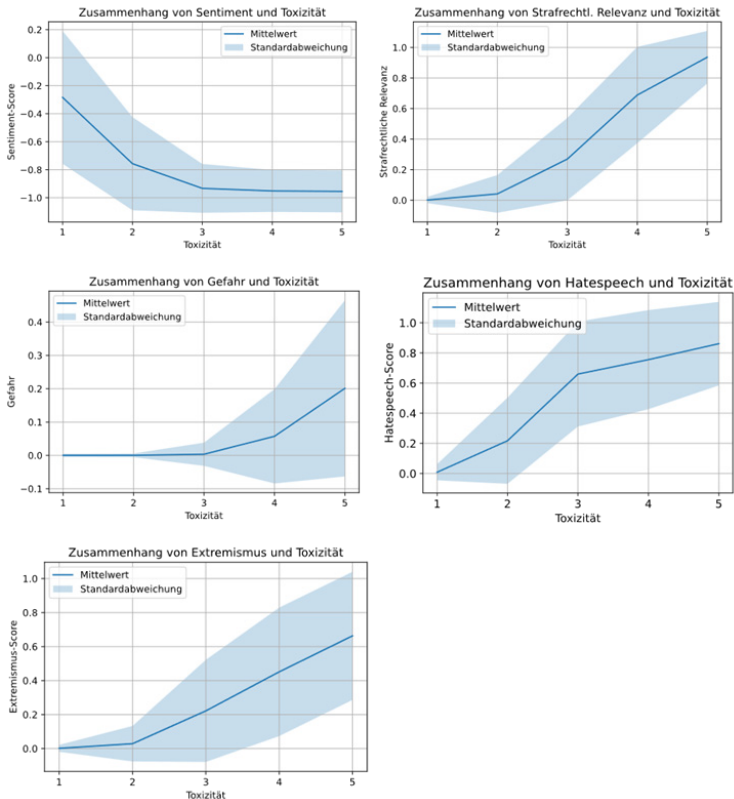


Abbildung 6: Darstellung der Zusammenhänge von Hatespeech, Sentiment, strafrechtlicher Relevanz, Gefahr und Extremismus mit der Toxizität. Angegeben ist jeweils der Mittelwert und der Bereich der Standardabweichung. Bei Sentiment ist -1 negativ, 0 neutral und +1 positiv. In allen anderen Kategorien bedeutet 0 jeweils „nicht zutreffend“ und 1 bedeutet „voll zutreffend“.

Zunächst lässt sich die Tendenz erkennen, dass mit steigender Toxizität gruppenbezogene Kommentare (Hass gegen Gruppen) zunehmen und persönliche Beleidigungen leicht abnehmen; Kommentare ohne Adressat (Public) mit einer Toxizität von > 3 gibt es nur sehr selten (siehe Abbildung 7). Das zeigt, dass Kommentare unspezifischen Ziels entweder weniger Hass enthalten oder aber weniger toxisch wahrgenommen werden, weil keine spezifische Person oder Gruppe angegriffen wird.

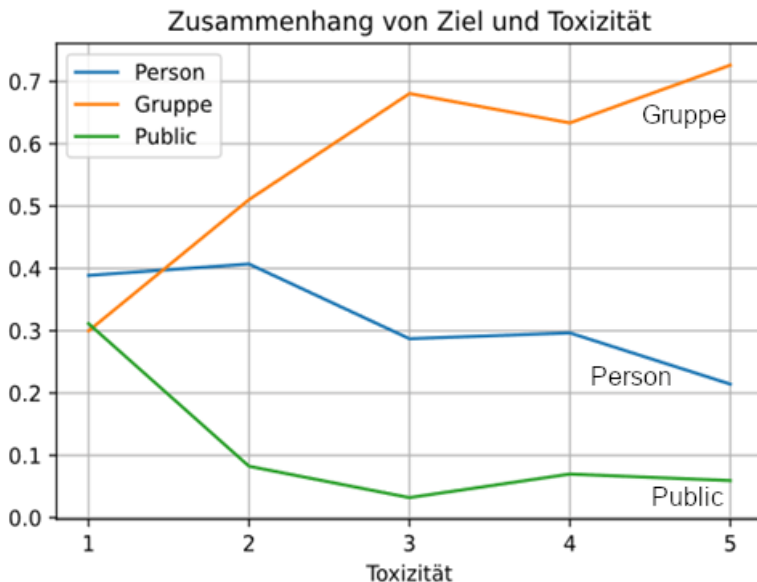


Abbildung 7: Zusammenhang des Kommentarziels und der Toxizität

Bei den Diskriminierungskategorien ist erkennbar, dass in allen zehn Kategorien die Toxizität von 3 am häufigsten vorkommt. Das lässt sich damit erklären, dass diese Kategorie nur annotiert wurde, wenn ein Kommentar als Hatespeech eingestuft wurde. Demzufolge ist in den meisten Fällen eine gewisse Toxizität vorhanden. Trotzdem gibt es Unterschiede zwischen den einzelnen Klassen: Kommentare, die als diskriminierend hinsichtlich Religion oder Nationalität gekennzeichnet wurden, sind tendenziell toxischer als beispielsweise Kommentare der Klassen *Politische Einstellung* oder *Persönliches Engagement* (siehe Abbildung 8).

Verteilung der Toxizität von Kommentaren abhängig von der Diskriminierungskategorie

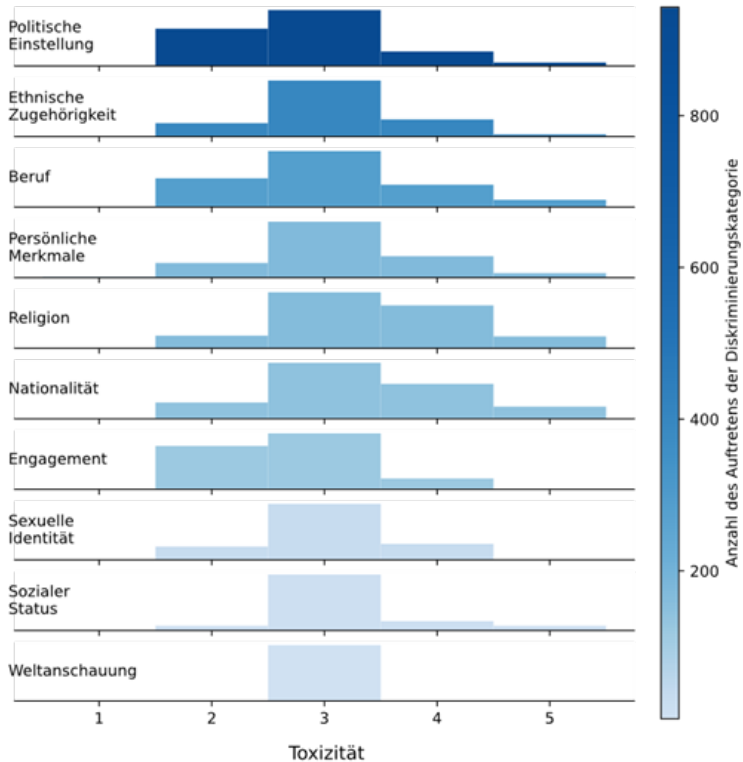


Abbildung 8: Die Abbildung stellt für jede Diskriminierungskategorie die Verteilung der Toxizität dar. Es ist erkennbar, wie toxisch Kommentare sind, für die die jeweilige Diskriminierungskategorie zutrifft. Die absoluten Häufigkeiten, wie oft jede Diskriminierungskategorie zutrifft, ist anhand der Farbskala ablesbar.

Die Analyse des Diagramms der Strafrechtsparagrafen zeigt, dass die meisten Paragrafen einen Großteil der Kommentare in Toxizitätsklasse 4 haben (siehe Abbildung 9). Ausnahmen sind §185 (Beleidigung) und §186 (üble Nachrede), sowie §187 (Verleumdung). Diese Paragrafen können sehr weit ausgelegt werden und daher auch auf weniger toxische Kommentare zutreffen. Die Paragrafen mit der höchsten Toxizität sind §126, §111 und §241. Diese Paragrafen stehen alle im Zusammenhang mit Bedrohung oder Androhung von Straftaten.

Toxizität strafrechtl. relevanter Kommentare unterteilt nach Paragraphen

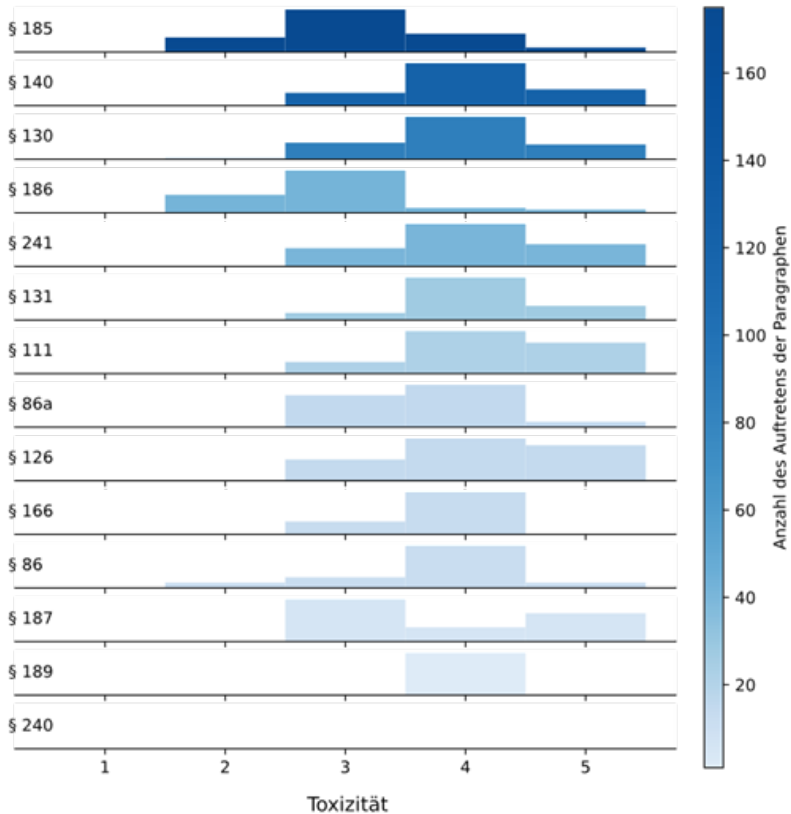


Abbildung 9: Die Abbildung stellt für jeden Strafrechtsparagraphen (StGB) die Verteilung der Toxizität dar. Es ist erkennbar, wie toxisch Kommentare sind, für die der jeweilige Paragraph zutrifft. Die absoluten Häufigkeiten, wie oft jede Diskriminierungskategorie zutrifft, ist anhand der Farbskala ablesbar.

3.4 Netzwerkanalyse

Neben der Detektion und Bewertung einzelner Hasskommentare besteht gleichzeitig die Notwendigkeit zu verstehen, wie auf Hassbotschaften reagiert wird, beziehungsweise welchen Einfluss eine Nachricht auf einen digitalen Diskurs in sozialen Netzwerken haben kann. Hieraus sollen

Erkenntnisse gewonnen werden, die darüber Aufschluss geben, wie sich Hass in einem Netzwerk nicht nur weiterverbreiten, sondern auch in seiner Ausprägung, Richtung und Intensität verändern kann. Dabei ist nicht nur der Hasskommentar selbst relevant, sondern zudem die Eigenschaften der am Austausch beteiligten Nutzerinnen und Nutzer und deren Position im Netzwerk.

Deswegen wurden Hass- und reguläre Nachrichten im Kontext ihrer zeitlichen Abfolge in Konversationen, auch als „Threads“ bezeichnet, initial miteinander assoziiert, mit dem Ziel, sie als Einheit zu erfassen. Als Ausgangspunkt wurde dabei jeweils ein einzelner Tweet genutzt, auf den eine Vielzahl an Antworten („Replies“) und Unter-Antworten folgen. Hierfür wurde im Anschluss die Darstellungsform der Konversationen in Baumdiagrammen gewählt, um die Komplexität digitaler Diskurse darstellen und erfassen zu können („Reply Trees“). Über farbliche Abstufungen (grün bis rot) wurden Nachrichten entsprechend ihres Gehaltes an Hass klassifiziert. Je nachdem, wo ein Nutzer bzw. eine Nutzerin innerhalb des Netzwerkes agiert, kann derselbe Hasskommentar ein unterschiedliches Ausmaß an Reichweite und Anschlusskommunikation auslösen, maßgeblich gesteuert auch von den Algorithmen der jeweiligen Plattformen. Die Einbeziehung ganzer Konversationsverläufe ermöglicht es somit, nicht nur die reine Anzahl von Hassbotschaften, sondern auch deren Dynamik zu erfassen. In dieser können sowohl „Hass-Kaskaden“, sprich Abwärtsspiralen negativer Ausdrucksweise, als auch Abmilderungen durch gezielte Gegenrede („Counter Speech“) einzelner Nutzer bzw. Nutzerinnen beobachtet werden. Counter Speech steht dabei als herausragendes Beispiel für den Einfluss einzelner Nutzerinnen und Nutzer auf gesamte Konversationsverläufe, die den beobachteten Ansteckungseffekte von Hass („Contagion“) gezielt unterbinden können. Der (digital-soziale) Hintergrund jedes Nutzenden bedingt ebenfalls den Diskurs, während gleichzeitig automatisierte Bots gezielt eingesetzt werden können, um Hass- und Desinformationskampagnen zu betreiben. In einer weiteren Betrachtung des Begriffs „Netzwerk“ kann diesem entgegengewirkt werden, indem die Profile und Benutzerkonten bei der Analyse von Konversationen und Einzelbeiträgen mit einbezogen werden. Konversationsübergreifende Analysen auf Sprach- und Verhaltensmerkmale können einen Beitrag zunächst bei der Identifikation und im weiteren Verlauf auch bei der Unterbindung von Hassnachrichten leisten(siehe auch Demus et al., 2022b).

4. BoTox: Bot- und Kontextererkennung im Umfeld von Hasskommentaren

Das Forschungsprojekt BoTox beschäftigte sich mit drei neuen Themenfeldern im Umfeld des Umgangs mit Hasskommentaren. Zum einen wurden Hasskommentare im Kontext des deutschen Strafrechts betrachtet. Zum anderen wurde der Einfluss von Bots im Zusammenhang mit Hasskommentaren untersucht, um ebenfalls automatisierte Erkennungsmethoden zu entwickeln. Weiterhin beschäftigte sich das Projekt mit der Erkennung von Hasskommentaren unter Einbeziehung des Kontextes, um relationale Abhängigkeiten und externe Einflüsse auf Nachrichtenverläufe zu untersuchen. Die Analysen nutzten einerseits fein granular annotierte Social-Media-Kommentare aus dem abgeschlossenen Projekt DeTox und stellten andererseits einen neuen Textkorpus mit automatisch generierten Hasskommentaren auf. Alle genannten Aufgaben wurden in einem jeweiligen Arbeitspaket untersucht. Im Ergebnis sollte ein Analyseprozess geschaffen werden, der alle Aspekte betrachtet und ein abschließendes Ergebnis zur weiteren Entscheidung erzeugt (Abbildung 10).

BoTox – Bot- und Kontextererkennung im Umfeld von Hasskommentaren

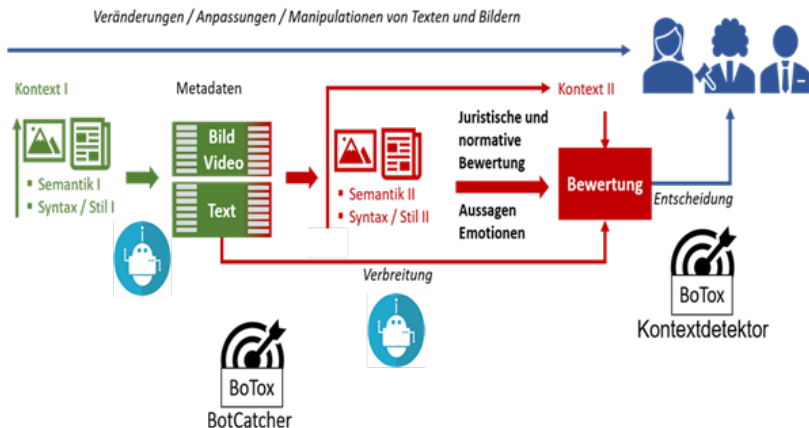


Abbildung 10: Schematischer Prozessablauf der Datenbewertung im Projekt BoTox

4.1 Datensammlung und Annotation

In BoTox konnten zu Beginn zwei neue Datensätze gesammelt und zusammengestellt werden. Für die Klassifikation von Hasskommentaren nach definierten Klassen von strafrechtlich relevanten Paragrafen wurden 1.190 Kommentare digital gesammelt und entsprechend annotiert. Zusätzlich wurde ein zweiter Datensatz mit 20.697 KI-generierten Hasskommentaren erstellt und für die Verwendung im Projekt aufbereitet. Die dafür verwendeten Daten stammen aus den Quellen:

1. Der DeTox-Datensatz (Demus et al., 2022a): Auf der Grundlage der verfügbaren Annotationen wurden Kommentare ausgewählt, die vermutlich illegal sind, sowie 300 zufällig ausgewählte Hasskommentare ohne festgestellte strafrechtliche Relevanz.
2. IHS ist ein weiterer Datensatz von Twitter-Nachrichten, die potenziell illegale Hassreden enthalten und gemäß den angewandten strafrechtlichen Paragrafen/Paragrafengruppen kommentiert wurden (Schäfer, 2023). Die Daten wurden von einer einzigen geschulten Person kommentiert. Es wurden 287 Kommentare mit einer zugewiesenen strafrechtlichen Relevanz ausgewählt.
3. Die Plattform X: Mithilfe einer Reihe von Schlüsselbegriffen haben wir 93 Kommentare gefunden, die potenziell unter § 86 Verbreiten von Propagandamitteln verfassungswidriger und terroristischer Organisationen und § 86a StGB Verwenden von Kennzeichen verfassungswidriger und terroristischer Organisationen fallen könnten.
4. 125 vom Modell GPT-3.5 generierte Kommentare bildeten den letzten Teil des Datensatzes. Dem Modell wurden Beispiele von X vorgelegt, und es wurde instruiert, ähnliche Kommentare zu liefern, die anschließend gefiltert wurden, um sicherzustellen, dass sie Hasskommentare waren.

4.2 Automatisierte Erkennung von Hatespeech mit strafrechtlicher Relevanz

Für die Detektion von Hasskommentaren, die aufgrund ihres Inhaltes eine strafrechtliche Relevanz hervorrufen, wurden zunächst zehn Straftatbestände aus dem deutschen Strafrecht StGB identifiziert und in drei Klassen eingeordnet. Diese waren:

Klasse 1: Verbreitung von verfassungsfeindlichem Material

- § 86 StGB - Verbreiten von Propagandamitteln verfassungswidriger Organisationen
- § 86a StGB - Verwenden von Kennzeichen verfassungswidriger Organisationen

Klasse 2: Öffentliche Aufforderung zu Straftaten und Störung des öffentlichen Friedens

- § 111 StGB - Öffentliche Aufforderung zu Straftaten
- § 126 StGB - Störung des öffentlichen Friedens durch Androhung von Straftaten

Klasse 3: Verherrlichung oder Beleidigung

- § 130 StGB - Volksverhetzung
- § 131 StGB - Gewaltdarstellung
- § 140 StGB - Belohnung und Billigung von Straftaten
- § 166 StGB - Beschimpfung von Bekenntnissen, Religionsgesellschaften und Weltanschauungsvereinigungen
- § 185 StGB Beleidigung
- § 186 StGB Üble Nachrede

Im Anschluss daran wurden Large Language Models, die dem Stand der Forschung entsprechen, genutzt und angewendet, um automatisiert entscheiden zu lassen, ob ein Kommentar strafrechtlich relevant ist und wenn ja, in welche Klasse dieser fällt.

Dabei konnten eine Accuracy von 0.8533 und ein F1-Score von 0.8593 erreicht werden. Verschiedene Experimente wurden durchgeführt, bei denen die jeweils einfacheren Prompts die besten Ergebnisse lieferten.

4.3 Automatisierte Bot-Erkennung im Umfeld von Hatespeech

Für die optimale Erkennung von automatisch generierten Hasskommentaren wurde ein Ansatz entwickelt, der einerseits auf dem sprachlichen Stil der Kommentare selbst und andererseits auf den Metadaten der Autoren-Accounts basiert. Zur Umsetzung dieses Ansatzes wurde auf Daten der Plattform Reddit zurückgegriffen. Für die Klassifikation, ob ein

Beitrag von einem Menschen oder einem Computer stammt, wurden, basierend auf dem sprachlichen Stil, zunächst Merkmale wie der Gebrauch von Namen, Lesbarkeitsindex, Wort- und Satzlänge extrahiert. Diese wurden im Anschluss mit Transformern-basierten Modellen kombiniert und trainiert.

Die Ergebnisse der Klassifikation sind vielversprechend und weisen darauf hin, dass sich die automatisch generierten und die von Menschen geschriebenen Hasskommentare durchaus stilistisch unterscheiden.

Die Klassifikation/Erkennung basierend auf den Metadaten, also dem Nutzerverhalten der Accounts, wurde anhand der aus Reddit extrahierten Kommentare mit neuronalen Netzen trainiert.

Die verwendeten Features waren in diesem Fall z.B. die Anzahl der getätigten Posts, die Anzahl der geschriebenen Kommentare und die Zeit, die zwischen den einzelnen Aktivitäten lag. Auch hier sind die Ergebnisse vielversprechend. Die besten Ergebnisse erzielte das Vorgehen mittels Random Forest, welches eine hohe Unterscheidbarkeit von Bot- gegenüber menschlichen Accounts aufgrund der Metadaten nahelegt.

4.4 Hatespeech-Detektion unter Einbeziehung des Kontexts

Hier wurden ebenfalls die bereits vorhandenen Daten aus dem DeTox-Projekt verwendet. Diese wurden nochmals vollständig halbautomatisch durch den Einsatz eines externen Sprachmodells sowie durch manuelle Nachbearbeitung nach ihrem Hass-Gehalt einzeln annotiert. Diese Annotation erfolgte wieder einzeln, Nachrichten-bezogen und zunächst ohne Kontextbezug. In einem weiteren Schritt wurde zusätzlich die jeweilige Gesamtkonversation mit einem aggregierten Label versehen. Dieses zweistufige Vorgehen erlaubt eine erste Annäherung an eine kontextbasierte Bewertung, auch wenn eine vollständig kontextsensitive Analyse derzeit nur mit erheblichem technischem Aufwand realisierbar ist (siehe Abbildung 11).

Die Analyse des Kontextes stellt innerhalb der Sprachforschung eine der anspruchsvollsten Fragestellungen dar. Zwar zeigte sich grundsätzlich eine Tendenz in der Klassifikation potenziell problematischer Inhalte, jedoch erwies sich der kontextbasierte Einsatz des Sprachmodells als herausfordernd, da dieses die Klassifikation ohne Kontextbezug zwischen

einzelnen Nachrichten innerhalb einer Konversation vornimmt. Eine Hürde ist weiterhin die hohe Rechenlast bei der Verarbeitung längerer Konversationen, welche eine konsistente Analyse erschwert. Dies steht im Kontrast zum umfangreichen Datensatz, in dem mitunter über 8.000 Kommentare in einer Konversation eingebettet sind.

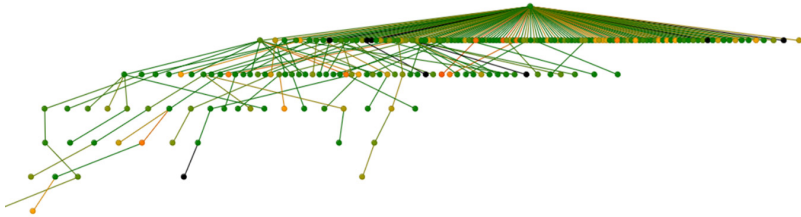


Abbildung 11: Baumstruktur einer Konversation mit einem Ausgangs-Tweet und dazugehörige Kommentare. Die farbliche Einfärbung entspricht dem Grad der Toxizität (grün bis rot bzw. hell bis dunkel)

Insgesamt konnten 102.735 unterschiedliche Autorinnen und Autoren innerhalb aller Beiträge über ihre eindeutige Author-ID identifiziert werden, wobei sich im Durchschnitt 81 Personen pro Konversation beteiligten.

In der Anzahl der Kommentare pro Konversation liegt eine ungleiche Verteilung vor. Während der Großteil der Diskussionen nur eine geringe Zahl an Kommentaren aufweist, existieren Ausreißer mit mehreren tausend Beiträgen. Der Median liegt dabei deutlich unter 50 Kommentaren.

Die meisten Konversationen umfassen weniger als 10 Kommentare, während längere Verläufe deutlich seltener auftreten. Neben der Länge der Gesamtkonversationen wurde auch die Tiefe der Beiträge untersucht. Die Analyse zeigt, dass die Mehrheit der Konversationsverläufe eher flach ist, aber auch deutlich tiefere Konversationen mit über 50 Ebenen vorkommen.

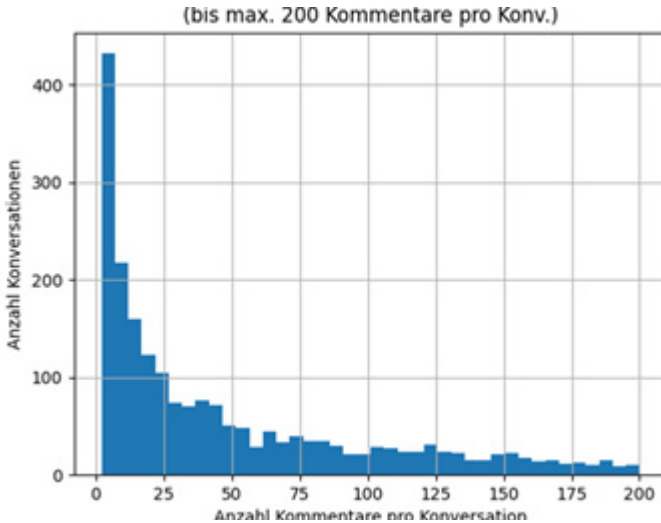


Abbildung 12: Anzahl der Konversationen in Abhängigkeit ihrer Länge (bis max. 200 Kommentare pro Konversation)

5. Hatespeech nach der Klassifikation: Dynamische Toxizität DyTox

Die Betrachtung offensiver Sprache im Netz mit dem Verbreitungsweg der sozialen Netzwerke bedarf einer ständig aktualisierten Betrachtung. Die stetigen Veränderungen von Sprache und Gesellschaft führen zu immer neuen Phänomenen und damit auch zu neuen Herausforderungen für die Öffentlichkeit und die Forschung.

5.1 Hürden und Herausforderungen bei der Klassifikation von Hatespeech

Trotz aller Fortschritte bei der automatisierten Klassifikation von Hatespeech bleibt die Frage bestehen, inwieweit algorithmisch basierte Modelle überhaupt in der Lage sind, die komplexen Strukturen der menschlichen Sprache abzubilden. Konversationen in sozialen Netzwerken folgen keiner linearen Logik, sondern sind geprägt von Ironie, individueller Sprache, kulturellen Hintergründen, impliziten Ausdrücken

und gruppenspezifischen Sprachreferenzen. Viele dieser Elemente lassen sich nur sehr schwer modellhaft abbilden und entziehen sich einer rein textbasierten Auswertung. Dadurch steigt das Risiko, dass zum Beispiel maschinelle Moderationssysteme einerseits harmlose Äußerungen fälschlicherweise als Hatespeech klassifizieren und andererseits tatsächlich problematische Inhalte übersehen. Erschwerend kommt hinzu, dass sich Sprache in digitalen Räumen äußerst dynamisch entwickelt. Strategien zur Umgehung von Moderationsfiltern etwa durch Satzfragmentierung, alternative Schreibweisen oder Umformulierungen machen eine kontinuierliche Anpassung und Nachschulung der Modelle notwendig, was wiederum mit einem erhöhten und permanenten Aufwand auf Seiten der Plattformbetreibenden verbunden ist. Ein weiterer Aspekt betrifft die Transparenz algorithmischer Entscheidungen. Für Nutzerinnen und Nutzer ist häufig nicht nachvollziehbar, weshalb bestimmte Inhalte entfernt oder herabgestuft werden, was sich auf die Funktionsweise der verwendeten Modelle zurückführen lässt. Das Fehlen klarer Erklärungen kann zu Misstrauen gegenüber der Moderation führen und den Eindruck einer undurchsichtigen oder gar willkürlichen Einflussnahme verstärken. Wissenschaftliche und politische Debatten fordern daher zunehmend nachvollziehbare Moderationsprozesse, die es Betroffenen ermöglichen, Entscheidungen nachzuvollziehen oder anzufechten, um den digitalen Diskurs zu erhalten

5.2 Neues 4-Dimensionen-Modell

Die reine Analyse von textuellen Daten ist nicht ausreichend, um eine vollumfängliche Analyse von Hassinhalten zu gewährleisten. Multimediale Inhalte, kontextuelle Bezüge, Verklausulierungen, aber auch sprachlich simple Mittel wie Sarkasmus und Ironie können Sprachmodelle zu Fehlinterpretationen von Aussagen führen. Ein neues 4-dimensionales Modell soll daher alle Einflüsse vereinen, die unmittelbare Auswirkungen bei der Betrachtung von Hass im Netz haben.

Inhalt: Was sagt die Toxizität?

Die erste Dimension bezieht sich klar auf den Inhalt des jeweiligen Beitrages. Hierbei ist es gleich, in welchem Format dieser vorliegt. Der Inhalt und die damit übermittelte Botschaft müssen klar eine negative Haltung, ein negatives Sentiment haben. Dabei kommen gleich mehrere Probleme

bei der Betrachtung von digitaler Hassrede zusammen. Zum einen lässt sich Hass nicht pauschal als Hass titulieren. Unterschiedliches Empfinden unterschiedlicher Empfänger und Empfängerinnen können einer Nachricht ambivalente Charakteristiken zumessen. Die daraus individuell resultierenden Benutzerinteraktionen werden in der 2. Dimension betrachtet. Darüber hinaus existiert neben der subjektiven Nutzerbetrachtung gleichzeitig eine juristisch-staatliche Seite. Mit der Verrohung der Online-Sprache im Zuge der globalen Fluchtbewegungen im Jahr 2015 und der damit verbundenen Intensivierung der Forschungen im Bereich der Hasssprache (Sponholz 2020) ist das gesellschaftliche Bewusstsein dafür gewachsen, dass strafbare Inhalte im digitalen Raum ernstzunehmende gesellschaftliche und rechtliche Konsequenzen haben und potenzielle Opfer nicht schutzlos bleiben dürfen. Zwar existieren in Deutschland keine eigenen Strafrechtsparagrafen, die eine Verfolgung von Hass im Netz zulassen würden, jedoch lassen sich bereits bestehende Paragrafen, darunter die für Beleidigung, üble Nachrede und Volksverhetzung anwenden (siehe Kapitel 3.2). Einen direkten, womöglich linearen Zusammenhang zwischen subjektivem Empfinden einer Hassnachricht und dessen juristische Verfolgbarkeit lässt sich nicht definieren, im Alltag jedoch mehrheitlich beobachten.

Nutzerverhalten: Zwischen Actio und Reactio

Jede Aktion ruft eine Reaktion hervor. Dieser physikalische Grundsatz lässt sich auch auf Verhaltensmuster im digitalen Raum übertragen. Nutzer, die sich in einem digitalen Diskurs einer Hassbotschaft ausgesetzt finden, bleiben meist nur die Optionen Flucht oder Kampf. Während die Toxizität eines Beitrages darüber Auskunft geben soll, inwiefern er Nutzer und Nutzerinnen dazu verleitet, eine Diskussion zu verlassen, soll über dynamische Toxizität bestimmt werden, inwiefern sich diese Toxizität innerhalb der Diskussion ausbreitet. Hassbotschaften, die Nutzer und Nutzerinnen dazu bewegen, eine Konversation zu verlassen, stehen dem obersten Gut der freien Meinungsäußerung entgegen, der sich Staat und Akteure durch verschiedene Regularien verpflichtet haben. Jüngste Untersuchungen im Rahmen des Promotionsprojektes zeigen, dass Hassbotschaften zu einem signifikant höheren Anteil am Ende einer Konversation stehen und diese somit beenden. Dem entgegen steht die Gegenrede (Counter-Speech), die durch mutige Nutzerinnen und Nutzer den Hassbotschaften entgegengestellt werden. Während die großen

Netzbetreiber unter anderem auch durch EU-Vorgaben das Löschen von problematischen Inhalten gegenwärtig als am zielführendsten im Kampf gegen Hass im Netz sehen, rückt die Gegenrede als die sozial nachhaltigere Lösung in den Vordergrund aktueller Betrachtungen.

Daneben rückt ein weiteres Verhalten in den Vordergrund aktueller Debatten: Das Einsteigen und somit Verstärken von Hassrede, indem jede einzelne Nachricht als Nährboden für weiteren Hass verstanden werden muss. Hierbei ist ein zeitnahes Löschen einzelner Beiträge womöglich sinnvoll, auch wenn dabei der Diskurs zerstört wird. Abschließend bleibt der Spagat zwischen der Einschränkung der freien Meinungsäußerung durch Großkonzerne wie Meta und der effektiven sowie zeitnahen Eindämmung hasserfüllter Inhalte eine der drängendsten Herausforderungen der aktuellen Zeit.

Zeit: Die Dynamik der Antworten

Die Untersuchung der Ausbreitung von Hass in Hinblick auf die Richtung und die Intensität spielt in neuerer Forschung eine weitere maßgebliche Rolle. Hierbei ergibt sich aus den geposteten Inhalten und der dazugehörigen zeitlichen Korrelation eine Konversation im digitalen Raum zwischen einem oder mehreren Nutzerinnen und Nutzern. Zwischen diesen und den von diesen geposteten Inhalten entsteht in der nächsten Ebene eine Dynamik. Die Analyse dieser Dynamik, durch welche sich in Bezug auf Hass gleichzeitig die dynamische Toxizität ergibt, soll neue Erkenntnisse bei der Eindämmung von Hass geben. Soziale Netzwerke setzen sich als Plattform zum bi- und unilateralen Austausch aus Nutzerinnen und Nutzern zusammen, hinter deren Profile (zumeist) reale Menschen sitzen. Somit lassen sich soziale Verhaltensmuster des realen Lebens ebenfalls auf dessen virtuelle Abbildung im digitalen Raum projizieren. Exemplarisch lässt sich hier die Antwortzeit eines Kommentars auf einen anderen betrachten. Ist diese kurz, kann das auf ein impulsiveres, unbedachteres Verhalten hinweisen, während eine längere Antwort auf ein bedachteres Handeln hinweisen kann.

Moderation: Eingriff von außen

Die Moderation sozialer Netzwerke durch automatisierte, halbautomatisierte oder manuelle Vorgehen stellt einen maßgeblichen Betrachtungspunkt bei der gegenwärtigen Analyse von Hatespeech dar. Die nationa-

len wie internationalen Gesetzgebungen, die alle Plattformbetreibenden zu Eindämmungsmaßnahmen gegen Hass auf ihren Seiten verpflichten, haben Einfluss auf die Debattenkultur im Netz. Zwar ist mit dem Löschen offensiver Inhalte die Problematik der Hassnachrichten zunächst gelöst. Daraus entspinnt sich jedoch eine gesellschaftliche Debatte über das Eingreifen außerstaatlicher Akteure in Form der Plattformbetreibenden in die Meinungsäußerung seiner Nutzer bzw. Nutzerinnen, die gleichzeitig durch dieselben Gesetzgeber gedeckt ist, die die Beschränkung von Hass fordern. Somit entspinnt sich ein Zwiespalt sowohl für die Betreibenden als auch die Nutzer und Nutzerinnen der Plattformen.

Vollautomatisiertes Löschen von Einzelbeiträgen aufgrund algorithmischer Entscheidungen erweist sich dabei aus mehreren Gründen als problematisch. Zunächst ist da die bereits beschriebene Subjektivität von Hass auf ihre jeweiligen Betrachter und Betrachterinnen. Ein bloßer Schwellenwert kann nie die Wirkung einer Nachricht auf jemanden abbilden, selbst wenn er von aktuell trainierten Sprachmodellen kommt. Als zweites Problem ist die bereits erwähnte Zerstörung der Konversationen zu nennen, welche sich unweigerlich aus dem Löschen einzelner Beiträge ergibt. Folgenachrichten können dadurch mitunter nicht mehr korrekt interpretiert werden oder erscheinen aufgrund des veränderten Kontextes in einem anderen, womöglich falschen Licht. Auch hier richtet sich der Blickwinkel wieder auf die Betrachtung von Hatespeech abseits der reinen Klassifizierung einzelner Textbeiträge. Dabei wird die Einbettung der Nachrichten in einen gemeinsamen Konversationskontext berücksichtigt, was bei einer maschinellen Moderation in jedem Fall relevant ist.

Letztendlich zeigt sich, dass sich in der Praxis weder rein automatisierte noch ausschließlich manuelle Moderationsansätze als geeignet erwiesen haben, um digitale Räume gegen Hass zu verteidigen. Vielmehr kommen zunehmend hybride Modelle zum Einsatz, in denen zunächst eine algorithmische Vorfilterung und im Anschluss eine menschliche Bewertung stattfindet. Während automatisierte Verfahren große Datenmengen effizient sichten und vorstrukturieren können, übernehmen menschliche Bearbeitende die finalen und vor allem rechtlich bindenden Entscheidungen. Diese Kombination verspricht eine höhere Präzision und gleichzeitig eine größere gesellschaftliche Akzeptanz, setzt jedoch umfangreiche personelle, technische und organisatorische Ressourcen voraus.

5.3 Anwendung neuer Betrachtungsweisen

Die bereits durchgeführten Untersuchungen zeigen ein Muster in Hinblick auf die neuen, zu untersuchenden Fragestellungen. Besonders auffällig ist zunächst die Häufung toxischer Beiträge an den Endpunkten von Konversationen. Im DeTox-Datensatz wurden insgesamt 646.681 Nachrichten untersucht, von denen 29.816 im Vorfeld als Hass klassifiziert wurden. Von diesen entfallen 21.578 Nachrichten auf Endpositionen eines Gesprächsbaums, was einem Anteil von 72,4 % entspricht. Dieser Anteil liegt signifikant über dem entsprechenden Wert nicht-toxischer Nachrichten (66,3 %), was ein Chi-Quadrat-Test statistisch bekräftigt ($\chi^2 = 467.6$; $p < .001$). Die Wahrscheinlichkeit, dass ein Gespräch durch einen toxischen Beitrag endet, ist damit um etwa ein Drittel höher als bei nicht-toxischen Äußerungen. Diese Befunde legen nahe, dass Hatespeech nicht nur als inhaltlich verletzende, sondern vor allem als kommunikativ zerstörende Handlung wirkt, die Gesprächsverläufe systematisch zum Abbruch bringt und dadurch die Möglichkeit des weiteren Diskurses nachhaltig reduziert. Im Vergleich dazu zeigten Faktoren wie Länge, zeitliche Struktur oder Reaktionsgeschwindigkeit der Konversation bislang nur geringfügige Einflüsse in den untersuchten Datensätzen. Hatespeech besitzt demnach eine klar erkennbare destruktive Kraft, die Gesprächsverläufe zum Abbruch bringt, unabhängig davon, wann und in welchem Umfang ein Kommentar gepostet wird.

6. Schluss und Ausblick

Die vorgestellten Projekte DeTox, BoTox und DyTox verdeutlichen, wie komplex und herausfordernd sowohl die Analyse als auch die daraus resultierende Bekämpfung von Hatespeech im digitalen Raum sind. Moderne Verfahren der Sprachverarbeitung ermöglichen eine zunehmende präzise Identifizierung und Bewertung schädlicher Beiträge. Gleichzeitig zeigen die Projekte jedoch auch, dass technologische Lösungen allein nicht ausreichen, um den Herausforderungen eines sich schnell wandelnden digitalen Kommunikationsraums wirksam zu begegnen. Begründen lässt sich dies vor allem mit der starken Kontextabhängigkeit von Hate Speech, dem kontinuierlichen Sprachwandel und den dynamischen rechtlichen und gesellschaftlichen Bewertungsmaßstäben.

Insbesondere die Einbindung von Anwendungspartnern, die enge Zusammenarbeit mit Expertinnen und Experten aus der Praxis sowie die kontinuierliche Bewertung der Modelle im praktischen Einsatz sind entscheidend, um realistische und nachhaltig relevante Ergebnisse zu erzielen. Deutlich wird auch, dass nur die Kombination aus algorithmischer Unterstützung und menschlicher Expertise für eine wirksame Moderations- und Präventionsstrategie nutzbar ist.

Für die Zukunft ergeben sich eine Vielzahl neuer Forschungsfragen. Neben der weiteren Verbesserung der Qualität der Vorhersagen rücken der Kontext und die Dynamik von Konversationen stärker in den Mittelpunkt. Ebenso wird die Frage nach transparenter erklärbarer KI weiter an Bedeutung gewinnen, um Akzeptanz und das Vertrauen in automatisierte Entscheidungen zu fördern.

Die vorgestellten Projekte liefern somit nicht nur technische Werkzeuge und Daten, sondern auch konzeptionelle Überlegungen für eine langfristige und vor allem nachhaltige Auseinandersetzung mit dem Thema Hass im Netz. Als Baustein in einer Reihe weiterer technologischer Innovationen sollen diese mit präventiven, rechtlichen und gesellschaftlichen Maßnahmen verknüpft werden, um strafbare Inhalte frühzeitig zu erkennen, Dynamiken zu unterbrechen und digitale Räume verantwortungsvoll zu gestalten. Nur wenn innerhalb der Gesellschaft das Vertrauen entsteht, dass Hass im Netz konsequent, zeitnah und nachhaltig verfolgt wird und zugleich die freie Meinungsäußerung als demokratisches Gut gewahrt bleibt, kann eine präventive Wirkung erzielt werden, die langfristig zu einem respektvolleren Kommunikationsverhalten und zur Stärkung digitaler Diskussionsräume beiträgt.¹²

12 Ein Teil dieser Arbeiten wurde durch das Hessische Ministerium des Innern und für Sport im Rahmen der Projekte DeTox und BoTox gefördert.

Literatur

- Almerekhi, H., Kwak, H., Salminen, J. & Jansen, B. J. (2020). Are these comments triggering? Predicting triggers of toxicity in online discussions. In Proceedings of the web conference 2020 (S. 3033–3040). DOI: <https://doi.org/10.1145/3366423.3380074>.
- Assenmacher, D., Clever, L., Frischlich, L., Quandt, T., Trautmann, H. & Grimme, C. (2020). Demystifying social bots: On the intelligence of automated social media actors. In Social Media+ Society 6 (3). DOI: <https://doi.org/10.1177/2056305120939264>.
- Bernhard, L. & Ickstadt, L. (2024). Lauter Hass - Leiser Rückzug: Wie Hass im Netz den demokratischen Diskurs bedroht. In Das NETTZ, Gesellschaft für Medienpädagogik und Kommunikationskultur, HateAid und Neue deutsche Medienmacher*innen als Teil des Kompetenznetzwerks gegen Hass im Netz (Hrsg.). URL: https://kompetenznetzwerk-hass-im-netz.de/wp-content/uploads/2024/02/Studie_Lauter-Hass-leiser-Rueckzug.pdf
- Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C. & Leskovec, J. (2017). Anyone can become a troll: Causes of trolling behavior in online discussions. In Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing (S. 1217–1230). DOI: <https://doi.org/10.1145/2998181.2998213>
- Cohen, J. (1960). A coefficient of agreement for nominal scales. In Educational and psychological measurement 20(1) (S. 37–46)
- Demus, C., Pitz, J., Schütz, M., Probol, N., Siegel, M. & Labudde, D. (2022). Detox: A comprehensive dataset for German offensive language and conversation analysis. In Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), Seattle, Washington. URL: <https://aclanthology.org/2022.woah-1.14.pdf>
- Demus, C., Schütz, M., Probol, N., Pitz, J., Siegel, M., Labudde, D. (2022). Hass im Netz – Aggressivität und Toxizität von Hasskommentaren und Postings, Detektion und Analyse. In: T. G. Rüdiger, & P. S. Bayerl (Hrsg.) Handbuch Cyberkriminalologie (S. 261-292). Springer VS, Wiesbaden. DOI: https://doi.org/10.1007/978-3-658-35450-3_13-1
- Demus, C., Schütz, M., Pitz, J., Probol, N., Siegel, M., and Labudde, D. (2023). Automatische Klassifikation offensiver deutscher Sprache in sozialen Netzwerken. In S. Jaki & S. Steiger (Hrsg.): Digitale Hate Speech (S. 65–88). J.B. Metzler, Berlin, Heidelberg. DOI: https://doi.org/10.1007/978-3-662-65964-9_4

- Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (S. 4171–4186). Association for Computational Linguistics, Minneapolis, Minnesota,. URL: <https://www.aclweb.org/anthology/N19-1423>
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5) (S. 378–382). <https://doi.org/10.1037/h0031619>
- Kiefer, C. (2016). Assessing the quality of unstructured data: An initial overview. In Proceedings of the Conference "Lernen, Wissen, Daten, Analysen" (LWDA 2016), Potsdam, Germany, September 12-14, 2016 (S. 62–73). <https://ceur-ws.org/Vol-1670/paper-25.pdf>
- Krippendorff, K. (1980). Validity in content analysis. In *Computerstrategien für die Kommunikationsanalyse* (291) (S. 69–112)
- Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C. & Patel, A. (2019). Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in Indo-European languages. In Proceedings of the 11th forum for information retrieval evaluation (S. 14–17). URL: <http://ceur-ws.org/Vol-2517/T3-1.pdf>
- Poletto, F., Basile, V., Sanguinetti, M., Bosco, C. & Patti, V. (2020). Resources and benchmark corpora for hate speech detection: a systematic review. In *Lang Resources & Evaluation* (2021) 55 (S. 477–523). DOI: <https://doi.org/10.1007/s10579-020-09502-8>
- Risch, J., Stoll, A., Wilms, L., & Wiegand, M. (2021). Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In Proceedings of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments (S. 1-12). URL: <https://aclanthology.org/2021.germeval-1.1.pdf>
- Schäfer, J. (2023). Bias mitigation for capturing potentially illegal hate speech. In *Datenbank-Spektrum* 23(1) (S. 41–51). DOI: <https://doi.org/10.1007/s13222-023-00439-0>
- Schütz, M., Schindler, A., Siegel, M. & Nazemi, K. (2021a). Automatic fake news detection with pre-trained transformer models. In Proceedings of the 3rd International Workshop on Research & Innovation for Secure Societies. DOI: https://doi.org/10.1007/978-3-030-68787-8_45

- Schütz, Mina and Demus, Christoph and Pitz, Jonas and Probol, Nadine and Siegel, Melanie and Labudde, Dirk (2021). DeTox at GermEval 2021: Toxic Comment Classification. In Proceedings of the GermEval 2021 Workshop on the Identification of Toxic, Engaging, and Fact-Claiming Comments (S. 54–61). Heinrich Heine University Düsseldorf. URL: <https://netlibrary.aau.at/obvukloa/content/titleinfo/6435190/full.pdf>
- Siegel, M. & Meyer, M. (2018). h_da submission for the GermEval shared task on the identification of offensive language. In Proceedings of the GermEval 2018 Workshop. Austrian Academy of Sciences, Vienna, Austria. URL: <https://epub.oeaw.ac.at/?arp=0x003a10d6>
- Sponholz, L. (2020). Der Begriff "Hate Speech" in der deutschsprachigen Forschung: eine empirische Begriffsanalyse. In SWS-Rundschau, 60(1). (S. 43-65). <https://nbn-resolving.org/urn:nbn:de:0168-ssaoar-79910-7>
- Vidgen, B., Nguyen, D., Margetts, H., Rossini, P. & Tromble, R. (2021). Introducing CAD: the contextual abuse dataset. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (S. 2289–2303). URL: <https://eprints.gla.ac.uk/272734/1/272734.pdf>
- Wiegand, M., Siegel, M. & Ruppenhofer, J. (2018). Overview of the GermEval 2018 shared task on the identification of offensive language. In Proceedings of the GermEval 2018 Workshop. Austrian Academy of Sciences, Vienna, Austria. URL: https://www.oeaw.ac.at/fileadmin/subsites/academiaecorpora/PDF/GermEval2018_Proceedings.pdf
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z. et al. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223 1(2).

Zur weiteren Vertiefung

- Felser, J., Spranger, M., & Siegel, M. (2025). Overview of the GermEval 2025 Shared Task on Harmful Content Detection. In Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops (S. 306-319). URL: <https://serwiss.bib.hs-hannover.de/frontdoor/deliver/index/docId/3679/file/978-3-69018-016-0.pdf>

- Kums, V., Meyer, F., Pivitt, L., Vedenina, U., Wortmann, J., Siegel, M., & Labudde, D. (2025). A Novel Dataset for Classifying German Hate Speech Comments with Criminal Relevance. In Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH) (S. 41-52). URL: <https://aclanthology.org/anthology-files/anthology-files/pdf/woah/2025.woah-1.4.pdf>
- Demus, C., Schütz, M., Probol, N., Pitz, J., Siegel, M., Labudde, D. (2022). Hass im Netz – Aggressivität und Toxizität von Hasskommentaren und Postings, Detektion und Analyse. In: Rüdiger, TG., Bayerl, P.S. (eds) Handbuch Cyberkriminologie. Springer VS, Wiesbaden. DOI: https://doi.org/10.1007/978-3-658-35450-3_13-1
- Demus, C., Schütz, M., Pitz, J., Probol, N., Siegel, M., and Labudde, D. (2023): Automatische Klassifikation offensiver deutscher Sprache in sozialen Netzwerken. In Sylvia Jaki und Stefan Steiger (eds.): Digitale Hate Speech. J.B. Metzler, Berlin, Heidelberg. DOI: https://doi.org/10.1007/978-3-662-65964-9_4
- Demus, C. and Pitz, J. and Schütz, M. and Probol, N. and Siegel, M. and Labudde, D. 2022. DeTox: A Comprehensive Dataset for German Offensive Language and Conversation Analysis. In Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH) (S. 143–153). Seattle, Washington (Hybrid). Association for Computational Linguistics. DOI: <https://doi.org/10.18653/v1/2022.woah-1.14>
- Schütz, Mina and Demus, Christoph and Pitz, Jonas and Probol, Nadine and Siegel, Melanie and Labudde, Dirk (2021). DeTox at GermEval 2021: Toxic Comment Classification. In Proceedings of the GermEval 2021 Workshop on the Identification of Toxic, Engaging, and Fact-Claiming Comments (S. 54–61). Heinrich Heine University Düsseldorf, 2021, URL: <https://netlibrary.aau.at/obvukloa/content/titleinfo/6435190/full.pdf>
- Struß, J. and Siegel, M. and Ruppenhofer, J. and Wiegand, M. and Klenner, M. (2019). Overview of GermEval Task 2, 2019 Shared Task on the Identification of Offensive Language. In Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019), October 9 – 11, 2019 at Friedrich-Alexander-Universität Erlangen-Nürnberg. - München [u.a.]: German Society for Computational Linguistics & Language Technology and Friedrich-Alexander-Universität Erlangen-Nürnberg, 2019 (S. 352-363). URL: <https://www.zora.uzh.ch/handle/20.500.14742/163340>
- Meyer, F. and Moosdorf, M. and Labudde, D. (2025). Hasskommentare auf Instagram: Eine themenbezogene Analyse am Beispiel des Social-Media Profils der „Tagesschau“. In: Polzeiinformatik 2025. URL: https://www.dhpol.de/Bd_27.pdf#page=65

- A. Albladi et al. (2025). Hate Speech Detection Using Large Language Models: A Comprehensive Review. In IEEE Access, Vol. 13 (S. 20871-20892). DOI: <https://doi.org/10.1109/ACCESS.2025.3532397>

Mediathek



Informationen über das Forschungsprojekt DeTox



Informationen über das Forschungsprojekt BoTox



Online-Vortrag bei der Gesellschaft für Informatik



Podcast zum Thema "Sprachtechnologie vs. Online-Hetze" von 2021



Bericht im Magazin des Weißen Rings von 2024



Podcast zum Thema „Hate Speech und Judenhass in digitalen Räumen“



Artikel bei Heise von 2022



Podcast von 2023 zum Thema “Wie künstliche Intelligenz Hass und Lügen im Netz erkennen kann“



Prof. Dr. Melanie Siegel ist seit 2012 Professorin für semantische Technologien und Grundlagen der Informatik an der Hochschule Darmstadt. Sie interessiert sich für Sprachtechnologie, maschinelle Übersetzung, Syntax und Semantik (Deutsch und Japanisch), Ontologie, Informationsextraktion, Sentimentanalyse, Textklassifikation und technische Dokumentation. Ihre Promotion hat sie 1996 mit einer Arbeit zur japanisch-deutschen maschinellen Übersetzung abgeschlossen. Ihre Habilitation in Bielefeld umfasst eine Venia für Computerlinguistik und Sprachtechnologie. Weitere Stationen: DFKI, Universität des Saarlandes, Acrolinx GmbH.



Florian Meyer ist Wissenschaftlicher Mitarbeiter und Promotionsstudent an der Hochschule Mittweida in Kooperation mit dem Promotionszentrum Angewandte Informatik des Land Hessen. In seiner Arbeit beschäftigt er sich mit der Ausbreitung und Dynamik von Hasskommentaren in sozialen Netzwerken.



Prof. Dr. Dirk Labudde ist seit 2009 Professor für Bioinformatik an der Hochschule Mittweida und gründete 2014 Deutschlands ersten Bachelorstudiengang „Allgemeine und Digitale Forensik“ zu welchem er ebenfalls 2014 zum Professor berufen wurde. Seit 2017 ist er außerdem Leiter des Lernlabors Cybersicherheit der Fraunhofer Academy. An der Hochschule Mittweida leitet er darüber hinaus die Forschungsgruppe FoSIL (Forensic Science Investigation Lab), welche sich mit den verschiedensten forensischen Fragestellungen beschäftigt. Der Kern liegt in der Identifikation von aus forensischer- bzw. Sicherheits-sicht relevanten, innovativen Technologien und deren Verbindung mit agilem Wissensmanagement zu Werkzeugen für die Forensische Praxis bzw. den Einsatz im interdisziplinären Management im Krisen- und Katastropheneinsatz. In diesem Zusammenhang ist Prof. Dr. Labudde auch als Gutachter vor Gericht, sowie als Berater für Polizeien und Staatsanwaltschaften tätig.

»Niedrigschwellige, anonyme und jederzeit verfügbare Informationsangebote können für Betroffene eine Brücke aus der Gewalt schlagen – nicht als Ersatz menschlicher Unterstützung, sondern als Orientierungshilfe, die erste Schritte erleichtert und sichere Optionen sichtbar macht – genau hier setzen digitale Informationssysteme und KI-gestützte Chatbots an.«

Stefanie Giljohann, Catharina Vogt

KI-basierte Chatbots in der Prävention häuslicher Gewalt

1. Einleitung

Viele Menschen suchen die erste Klärung einer Frage im Internet – dort, wo Informationen jederzeit verfügbar sind, ohne Bewertung ihres Anliegens und ohne Kosten. Dialogbasierte Systeme, also Anwendungen, die Informationen in Form eines Gesprächs bereitstellen, erweitern diese Form der Orientierung. Sie liefern nicht nur Informationen, sondern ermöglichen auch die dialogische Beantwortung von Fragen zur Einordnung der Inhalte.

Diese zunehmende Orientierung im Digitalen verändert die Bedingungen und Erwartungen, unter denen Prävention einsetzt. Nutzer*innen wünschen sich, dass Informationen jederzeit anonym zugänglich sind, und die Antworten sollen leicht verständlich sein (Tananau Blumenschein et al., 2023). So prüfen viele erste Unsicherheiten und die Einordnung belastender Erfahrung online, insbesondere bei schambesetzten Themen. Womöglich erfolgt dies noch ohne konkretes Phänomenverständnis oder klaren Handlungsplan. Nicht selten liegt diese Orientierungsphase zeitlich vor dem Kontakt zur informellen und professionellen Unterstützung. Gerade weil diese Phase zeitlich vor den klassischen Zugängen der Prävention und Intervention liegt, ist sie für präventive Bemühungen von großer Bedeutung. Daher sollte Prävention bereits an diesem frühen Punkt anschlussfähig sein. Auch wenn digitale Räume ursprünglich nicht als präventive Orte konzipiert wurden, sind sie längst ein fester Bestandteil präventiver Prozesse geworden. Schließlich erfolgt Prävention dort, wo Menschen eigenes Erleben oder mögliche Gefährdungen einordnen, um Risiken für sich oder andere zu verringern.

Dem entsprechend zeigen präventive Initiativen, Unterstützungseinrichtungen und behördliche Stellen längst Präsenz im digitalen Raum, um Informationssuchende zu sensibilisieren, informieren und orientieren. Zentrales Ziel ist hierbei in der Regel die gezielte und qualitätsgesicherte Wissensvermittlung, welche relevantes Wissen strukturiert und verständlich bereitstellt. Dies steht im Kontrast zur generellen digitalen Realität, in der Informationssuchende nicht zwingend auf thematisch gut aufbereitete, verlässliche Inhalte stoßen.

Genau an dieser Stelle werden künstliche Intelligenz (KI), dialogbasierte Systeme und Chatbots relevant: Sie können Wissen bündeln, thematisch ordnen und strukturiert zugänglich machen. In den vergangenen Jahren sind entsprechende Anwendungen in der Prävention und Unterstützung unterschiedlichster Problembereiche entstanden. Die Themen reichen von der allgemeinen Gesundheitsprävention (Überblick in Maia et al., 2023), Substanzmissbrauch (Überblick in Lee et al., 2024), Essstörungen (Fitzsimmons-Craft et al., 2022), Angst, Stress und Depression (Daley et al., 2020), Suizidalität (Ascorbe et al., 2023) bis hin zu Gewalt im Schulkontext (Kang et al., 2024), (Cyber-)Bullying (Lian et al., 2023; Mendoza-Pinto, 2023), bildbasiertem sexuellen Missbrauch (Maeng & Lee, 2022), sexueller Belästigung (Bauer et al., 2020) und häuslicher Gewalt (Sanz Urquijo et al., 2024).

Nicht alle dieser Chatbots sind KI-gestützt; einige sind regelbasiert und funktionieren auf Basis von Entscheidungsbäumen mit vorgegebenen Fragen und Antworten. Es gibt webbasierte Anwendungen, andere sind wiederum nur per (teilweise kostenpflichtigem) App-Download verfügbar oder Teil eines Messenger-Systems wie Telegram oder WhatsApp. Informationen zur Datensicherheit, zur Robustheit des Systems gegenüber Hackingangriffen, zur Qualität der Inhalte oder Evaluation der Anwendung sind nicht immer gegeben. Diese Heterogenität erschwert die Orientierung der Nutzer*innen. Hinzu kommt, dass innerhalb eines Feldes neue Systeme schnell auftauchen beziehungsweise mitunter ohne größere öffentliche Wahrnehmung wieder verschwinden. Gleichzeitig verdeutlichen die rasante Entwicklung und wachsende Nachfrage in diesem Bereich, dass dialogische Systeme als strukturierter Orientierungspunkt grundsätzlich auf hohe Akzeptanz stoßen. Vor diesem Hintergrund rückt die Frage ins Zentrum, wie Chatbots in der Prävention verantwortungsvoll eingesetzt werden können.

Dieser Beitrag veranschaulicht das Spannungsfeld zwischen dem präventionspraktischen Potenzial solcher Systeme und den damit verbundenen ethischen, rechtlichen und strukturellen Herausforderungen bei der Prävention häuslicher Gewalt an einem konkreten Fallbeispiel, dem KI-basierten Chatbot AinoAid™ (Vogt & Giljohann, 2025). Dieser Chatbot wurde im EU-Projekt IMPROVE entwickelt, um Gewaltbetroffenen niedrigschwellig als sehr frühe Brücke ins Hilfe- und Schutznetzwerk Orientierung zu geben (Vogt et al., 2026). Anhand von AinoAid™ werden die Möglichkeiten und Grenzen KI-gestützter Prävention in einem hochsensiblen und sicherheitsrelevanten Kontext aufgezeigt.

Um diesen besonderen Anwendungsfall von KI in der Prävention nachvollziehbar darzustellen, bieten wir in diesem Kapitel zunächst einen Überblick in das Phänomen häusliche Gewalt und über typische Barrieren auf dem Weg zu Unterstützung. Anschließend beschreiben wir die spezifischen Anforderungen, Potenziale und Herausforderungen bei digitaler Orientierungsgebung, um dann konkret auf AinoAid™ als Fallbeispiel einzugehen. Abschließend bündeln wir die gewonnenen Erkenntnisse und leiten daraus Implikationen für Prävention, Praxis und zukünftige Entwicklungen KI-gestützter Unterstützungssysteme ab.

2. Häusliche Gewalt als Präventionsherausforderung

Häusliche Gewalt umfasst gemäß der Istanbul-Konvention missbräuchliches Verhalten im Rahmen einer früheren oder aktuellen Intimbeziehung, der Familie oder eines Haushalts (Council of Europe, 2011, Art. 3b). Zu ihren zentralen Erscheinungsformen zählen physische, sexualisierte, psychische und auch ökonomische Gewalt (ebd.). Darüber hinaus tritt sie in vielen weiteren Formen und Mischformen auf.

Häusliche Gewalt gilt in Europa als weitverbreitetes Phänomen: 17,7 % aller Frauen erleben im Lebensverlauf physische und/oder sexualisierte Gewalt durch Partner oder Ex-Partner (FRA et al., 2024). In Deutschland kam es im polizeilichen Helffeld im Jahr 2024 zu 170.881 Fällen von Partnerschaftsgewalt mit 135.713 weiblichen Opfern (BKA, 2025). Auch Männer sind, wenngleich in deutlich geringerem Ausmaß (21 %), von häuslicher Gewalt betroffen (ebd.). Trotz dieser hohen Betroffenenzahlen zeigt die aktuellste deutsche repräsentative Dunkelfeldstudie

(Leitgöb-Guzy & Bieber, 2026), dass das Dunkelfeld um ein Vielfaches größer ist: 16,1 % der Befragten erlebten körperliche Gewalt in ihrem Lebensverlauf durch (Ex-)Partner*innen, mit einer durchschnittlichen Anzeigequote von 3 %. Sexuelle Übergriffe in (Ex-)Partnerschaften berichteten 1,4 % der Frauen (versus 0,2 % der Männer); von psychischer Gewalt waren 48,7 % der Frauen und 40,0 % der Männer mindestens einmal im Leben betroffen. Viele Betroffene berichten von langfristigen Auswirkungen auf psychische und physische Gesundheit, ökonomische Stabilität sowie auf die Sicherheit und Entwicklung ihrer Kinder (Follingsstad, 2009). Sich intensivierende körperliche Gewalt in Paarbeziehungen gilt zudem als ein zentraler Risikofaktor für Femizide (Rumpf et al., 2024) und dies insbesondere in Trennungsphasen (Horn et al., 2025).

Die Prävention häuslicher Gewalt steht vor der herausfordernden Aufgabe, Betroffene in einem Feld zu erreichen, das sich zugleich durch hohe Fallzahlen, starke Tabuisierung, komplexe Gewaltdynamiken und vielfältige Barrieren gegenüber der Zugänglichkeit des Hilfe- und Schutznetzwerkes kennzeichnet. Prävention darf jedoch keinesfalls allein oder überwiegend bei jenen ansetzen, die das Hilfe- und Schutzsystem bereits erreichen, sondern sollte in den frühen Phasen eskalierender Gewaltverläufe das frühzeitige Erkennen von Gewaltmustern ermöglichen und die informierte Orientierung der Betroffenen stärken.

Ein grundlegendes Verständnis der zahlreichen Formen häuslicher Gewalt, ihrer Dynamiken und der Zugangsbarrieren zum Hilfe- und Schutzsystem macht nachvollziehbar, welche Anforderungen, Herausforderungen und Potenziale KI-gestützte Prävention in diesem Kontext mit sich bringt. Die folgenden Abschnitte legen hierfür die Grundlage und verdeutlichen, weshalb frühe, orientierende und niedrigschwellige Prävention bislang noch unzureichend ist.

2. 1 Erscheinungsformen häuslicher Gewalt

Physische Gewalt reicht von Schlägen und Stößen über den Einsatz von Gegenständen bis hin zu Würgehandlungen und anderen lebensbedrohlichen Angriffen (Pritchard et al., 2017). Sie kann, muss jedoch keine sichtbaren Spuren hinterlassen, da Täter*innen Gewalt gezielt verdecken. *Sexualisierte Gewalt* umfasst unter anderem sexualisierte Demütigungen, erzwungene sexuelle Handlungen und Genitalverletzungen

(Hegarty et al., 2017). *Psychische Gewalt* zeigt sich in Drohungen, Einschüchterung, Beschämung, Vorwürfen, dem systematischen Untergraben von Selbstwert und Handlungsspielraum sowie sozialer Isolierung (Albuquerque Netto et al., 2017; Harris & Woodlock, 2019). Als die weitverbreitetste Gewaltform (LKA Niedersachsen, 2022) umfasst sie zudem sämtliche toxische Kommunikationsweisen, die von egoistisch über launisch bis narzisstisch und anmaßend reichen. *Ökonomische Gewalt* beinhaltet beispielsweise die Kontrolle gemeinsamer Ressourcen, das Vorenthalten finanzieller Mittel und das Erzwingen wirtschaftlicher Abhängigkeit, etwa durch Schuldenkontrolle oder die Unterbindung von Erwerbstätigkeit (Postmus et al., 2020).

Als zentrale und den zuvor genannten Erscheinungsformen oftmals zugrundeliegende Facette der häuslichen Gewalt gilt *Kontrolle*, da all die Gewalthandlungen von Täter*innen der Befriedigung ihres starken Machtmotivs dienen. Kontrolle umfasst die Überwachung von Kommunikationskanälen und sozialen Kontakten, die Einschränkung von Mobilität sowie der Meinung Betroffener. Diese Praktiken kumulieren zu einem Zustand, in dem Betroffene Gewalt zunehmend als normal oder unvermeidbar erleben (Epstein & Goodman, 2018; Mela et al., 2023). Digitale Formen der Kontrolle, etwa das Mitlesen von Nachrichten, das Einfordern von Passwörtern, das Tracking von Bewegungsdaten oder die Drohung, private Bilder zu veröffentlichen, erweitern den Kontrollraum bis in den digitalen Alltag hinein (Harris & Woodlock, 2019; Storer & Nyerges, 2023).

Betroffene erleben häusliche Gewalt selten in nur einer klar abgegrenzten Form; häufig berichten sie vielmehr von einem Zusammenspiel multipler Gewaltformen, das sich über die Zeit verdichtet (Alberton et al., 2025; Farhall et al., 2020). Dennoch halten viele Gewaltbetroffene an ihrer Hoffnung auf eine positive Veränderung fest. *Gaslighting* verstärkt diese Dynamik, indem Täter*innen die Wahrnehmungen, Erinnerungen und Bewertungen der Betroffenen systematisch infrage stellen. Wenn beispielsweise Beschimpfungen mit Entschuldigungen ineinander übergehen, sexualisierte Nähe mit körperlicher Gewalt verwoben ist oder finanzielle Kontrolle als vermeintliche Fürsorge gerahmt wird, fällt es schwer, einen klaren Gewaltbegriff zu entwickeln. Die Vermischung unterschiedlicher Gewaltformen und das Anzweifeln oder Verleugnen der eigenen Einschätzungen erschweren es den Betroffenen, ihre Erfahrungen als häusliche Gewalt einzuordnen.

2.2 Dynamiken und Verlaufsmuster häuslicher Gewalt

Häusliche Gewalt entsteht meist in Beziehungen, die als Liebesbeziehungen beginnen. Ihre Entwicklung folgt häufig einem Prozess schleichender Grenzverschiebungen, beispielsweise durch einen allmählichen Übergang von Kritik, Eifersucht oder subtiler Kontrolle hin zu deutlicheren Einschränkungen, Isolation und schließlich körperlichen Übergriffen oder auch Freiheitsentzug. Mit der Zeit verschiebt sich so die individuelle Wahrnehmung dessen, was akzeptabel erscheint. Viele Betroffene berichten, dass sie schwere Formen von Kontrolle oder Gewalt erst spät als grenzüberschreitend erkannten, was als ein Prozess der Gewöhnung bzw. der erlernten Hilflosigkeit (Seligman et al., 1979), der die (wahrgenommene) Handlungsfähigkeit einschränkt, beschrieben wird (Tananau Blumenschein et al., 2023).

Dieses Muster lässt sich häufig dem „Kreislauf der Gewalt“ zuordnen, bestehend aus Spannungsaufbau sowie Gewalthandlungen einerseits und einer anschließenden Phase der Entschuldigung und der scheinbaren Normalisierung andererseits (Walker, 2006). Gerade die Versöhnungsphasen tragen dazu bei, dass Gewaltbetroffene ihre Hoffnung auf Veränderung aufrechterhalten und Gewaltereignisse relativieren (Epstein & Goodman, 2018).

Hinzu kommt eine starke Ambivalenz: Viele Gewaltbetroffene wünschen sich oder versuchen mehrfach, die Beziehung zu verlassen, scheuen jedoch die Umsetzung oder kehren zurück – aufgrund ihrer emotionalen Bindung, ökonomischen Abhängigkeit, Sorge um die gemeinsamen Kinder und zahlreicher weiterer Gründe und Barrieren (Alberton et al., 2025). Trennungssituationen gelten zudem als besonders gefährlich und erhöhen das Risiko der Gewalteskalation und für Tötungsdelikte (Horn et al., 2024). Der Beginn einer neuen Partnerschaft, ebenso wie die Rückkehr in die gewaltvolle Beziehung nach einer Trennung, bergen ebenfalls ein erhebliches Gewaltisiko.

2.3 Barrieren der Hilfesuche

Barrieren der Hilfesuche wirken auf individuellen, situativen und strukturellen Ebenen zusammen. Einen vertiefenden Einblick bietet der IMPROVE-Bericht von Mela und Kolleg*innen (2023), der verschiedene

vulnerable Betroffenengruppen systematisch berücksichtigt und zentrale Hindernisse der Hilfesuche vergleichend herausarbeitet. Präventionsansätze für häusliche Gewalt sollten so gestaltet sein, dass sie die im Folgenden genannten individuellen, situativen und strukturellen Barrieren berücksichtigen und – soweit möglich – aktiv überbrücken.

Individuelle Barrieren

Auf individueller Ebene berichten Betroffene von Schuldgefühlen, Scham, Angst vor Schuldzuweisungen, der Sorge um Kinder, Loyalität gegenüber der gewaltausübenden Person und der Furcht vor Stigmatisierung (Epsstein & Goodman, 2018). Viele zweifeln, ob ihr Erleben „schlimm genug“ sei, um Hilfe zu suchen, oder halten sich selbst für (mit)verantwortlich für die Situation. Betroffene verwenden häufig keinen klaren Begriff von häuslicher Gewalt und sprechen stattdessen von „Streit“, „schwieriger Beziehung“ oder „Problemen“, die sie zunächst selbst lösen möchten (Tananau Blumenschein et al., 2023).

Hinzu kommen Unsicherheiten über mögliche Folgen: Wie wahrscheinlich ist eine Eskalation nach einem Hilferuf? Welche Konsequenzen haben Schutzmaßnahmen für Kinder, ökonomische Stabilität oder Aufenthaltsstatus? Diese antizipierten Risiken mindern die Bereitschaft zur Hilfesuche erheblich (Farhall et al., 2020; Mantler et al., 2021).

Situative Barrieren

Situative Barrieren ergeben sich unmittelbar aus der Gewaltkonstellation selbst. Betroffene teilen häufig den Wohnraum und digitale Geräte mit der gewaltausübenden Person. Dadurch lassen sich Alltagsroutinen, Kommunikationswege und soziale Kontakte leicht kontrollieren und überwachen. Bereits die Suche nach Informationen über Unterstützungsangebote erleben Gewaltbetroffene (zutreffend) als Risiko – etwa, wenn Browserverläufe, Messenger oder E-Mails kontrolliert werden (Harris & Woodlock, 2019; Tananau Blumenschein et al., 2023).

Isolation, fehlende Betreuungsmöglichkeiten der Kinder, eingeschränkte Mobilität oder auch Erreichbarkeit der Hilfseinrichtungen sowie ökonomische Abhängigkeit reduzieren zusätzlich die Möglichkeiten, vertraulich Gespräche zu führen und Unterstützung zu finden (Farhall et al., 2020; Gabellini et al., 2025). In ländlichen Räumen kommt die Sorge hinzu, im Hilfesystem persönlich erkannt zu werden oder aufzufallen, wenn Ein-

richtungen aufgesucht werden. Selbst dort, wo Unterstützungsstrukturen vorhanden sind, bleiben sie praktisch unerreichbar, wenn Betroffene keinen geschützten Weg und Zeitpunkt finden, um unentdeckt Kontakt aufzunehmen.

Strukturelle Barrieren

Auf struktureller Ebene zeigt sich, dass Betroffene viele Angebote und Zuständigkeiten nicht kennen. Hilfesysteme sind zudem in der Regel fragmentiert: Polizei, Gesundheitswesen, Beratung, Justiz und spezialisierte Hilfen verfügen über getrennte Zugänge (Vogt, 2020; Vogt & Kersten, 2022) und sind noch zu selten gut miteinander verknüpft (Giljohann et al., 2021; Mela et al., 2023).

Hinzu kommen institutionelle Hürden wie eingeschränkte Öffnungszeiten, lange Wartezeiten, komplexe Anmeldeverfahren, fehlende Mehrsprachigkeit oder erschwerte Zugänge für Personen ohne gesicherten Aufenthaltsstatus (Delpuch et al., 2024; Goodey, 2017). Negative Vorerfahrungen – etwa nicht ernst genommen worden zu sein – senken zusätzlich die Wahrscheinlichkeit einer erneuten Hilfesuche (Tananau Blumenschein et al., 2023).

2. 4 Konsequenzen für die Prävention

In Summe entsteht ein Bild, in dem Betroffene vieles selbst leisten müssen: ihre Situation einordnen, Begriffe finden, Angebote recherchieren, deren Passung bewerten und schließlich einen Kontakt wagen – häufig unter Bedingungen, die von Kontrolle, Angst und Unsicherheit geprägt sind. Genau hier liegt der zentrale präventive Engpass. Für Prä- und Intervention ergibt sich nur ein enges, verletzliches Zeitfenster: Gewaltbetroffene öffnen sich in der Regel nur kurzzeitig für Veränderung – in diesen Momenten ist es umso wichtiger, dass relevante Informationen und Orientierung trotz der vielseitigen Barrieren sicher und leicht auffindbar sind.

Für die Prävention bedeutet dies, dass es einer Informationsvermittlung bedarf, die Betroffene früh erreicht. Da Betroffene in diesem Stadium weitgehend unsichtbar und kaum erreichbar für das Schutz- und Hilfesystem sind, ist die Bereitstellung sicherer, anonymen und niedrigschwelliger Wissensangebote dort besonders wirksam, wo sie nach Orientie-

suchung suchen: im digitalen Raum. Zusätzlich ist es präventiv besonders bedeutsam, dass Betroffene ein grundlegendes Verständnis jener Risiko- und möglicher Schutzfaktoren entwickeln, die aus den beschriebenen Dynamiken hervorgehen – etwa Isolation, eingeschränkter Zugang zu Ressourcen, spezifische Täterstrategien oder Unterstützung durch ihr soziales Umfeld.

Nur, wenn frühzeitig gehandelt wird und Betroffene nicht vollends vom Strudel des Gewaltkreislaufs erfasst sind (Walker, 2006), können Traumatisierungen, schwere Gewalttaten und weitere verheerende Folgen für Betroffene, ihre Kinder sowie involvierte Dritte verhindert werden. Die beschriebenen Formen häuslicher Gewalt, Dynamiken und Barrieren verdeutlichen, dass Prävention dort ansetzen sollte, wo Gewaltbetroffene ihr Erleben prüfen und Informationen suchen, ohne sich bereits sichtbar an das Hilfesystem zu wenden (Vogt & Giljohann, 2025). Niedrigschwellige, anonyme und jederzeit verfügbare Informationsangebote können für Betroffene eine Brücke aus der Gewalt schlagen – nicht als Ersatz menschlicher Unterstützung, sondern als Orientierungshilfe, die erste Schritte erleichtert und sichere Optionen sichtbar macht. Genau hier setzen digitale Informationssysteme und (KI-gestützte) Chatbots an, deren Funktionieren, Potenziale und Grenzen im folgenden Unterkapitel beschrieben werden.

3. Chatbots im Kontext häuslicher Gewalt

In den vergangenen Jahren ist ein äußerst vielseitiges Spektrum digitaler Präventions- und Unterstützungsangebote entstanden – von Informationswebseiten und Beratungsportalen über Apps, Foren und Peer-Communities bis hin zu dialogbasierten Systemen. Diese Angebote unterscheiden sich deutlich hinsichtlich ihrer Zielgruppenorientierung, thematischen Schwerpunkte, Zugangswege und Funktionslogiken. Kontinuierlich entstehen neue Formate, während andere wieder verschwinden; das Feld ist dynamisch, vielfältig und schwer überschaubar.

Mit der folgenden Darstellung der Funktionsweisen und Heterogenität digitaler Chatbots im Kontext häuslicher Gewalt möchten wir das Verständnis der Chancen und Risiken des Einsatzes solcher Chatbots erhöhen und damit die informierte Auswahl sicherer und passender Angebote erleichtern.

3.1 Funktionsweise digitaler Chatbots

Digitale Chatbots übernehmen drei zentrale Funktionen:

1. Sie unterstützen Nutzer*innen dabei, eigenes Erleben sprachlich zu fassen und Muster zu erkennen.
2. Sie vermitteln strukturiertes Wissen zu Gewaltformen und -folgen, rechtlichen Rahmenbedingungen und Unterstützungsoptionen.
3. Sie erleichtern Übergänge zum Hilfesystem, indem sie Unsicherheiten abbauen oder mögliche nächste Schritte aufzeigen.

Diese Funktionen liegen ausschließlich im Bereich der Orientierung; sie ersetzen weder eine Beratung noch die individuelle Einschätzung von Expert*innen.

Die konkrete Arbeitsweise der Chatbots hängt von ihrer technischen Grundlage ab. Regelbasierte Systeme folgen festen Entscheidungswegen: sie reagieren konsistent und vorhersehbar. KI-gestützte Modelle formulieren hingegen flexibler und natürlicher. Sie benötigen klare fachliche Begrenzungen, um fehlerhafte oder unpassende Inhalte zu vermeiden. Häufig kommen hybride Ansätze zum Einsatz.

Ein wesentliches Merkmal dialogischer Systeme besteht darin, dass sie Interaktionen ermöglichen, ohne dass Nutzer*innen mit präzisen Begriffen oder klaren Fragestellungen einsteigen müssen. Gerade in frühen Phasen der Orientierung erlaubt die dialogische Struktur eine schrittweise Klärung unscharfer oder fragmentarischer Fragen.

3.2 Digitale Chatbots für häusliche Gewalt im Überblick

Internationale Übersichten zeigen, dass Chatbots im Kontext häuslicher Gewalt sehr unterschiedliche Schwerpunkte setzen – etwa auf Wissensvermittlung, strukturierte Reflexion oder alltagsnahe Hinweise (Übersicht in Sanz Urquijo et al., 2024; ParentText: <https://globalparenting.org/parent-text/>, Schafer et al., 2023; Hope Chat: <https://www.domesticshelters.org/hope-chat-ai>). Die Dokumentation zu Zugänglichkeit, Kuratierung, Aktualisierung, Datenverarbeitung und Systemsicherheit ist heterogen, was die Bewertung ihrer Qualität erschwert. Zudem verlieren viele projekt-basierte Systeme nach Projektende an Aktualität oder bleiben ohne Evaluation, was Fragen zur langfristigen Verlässlichkeit aufwirft. Diese Charakteristika treffen auch auf das Angebot der deutschsprachigen Chatbots (Tabelle 1) zu.

Tabelle 1: Chatbots für Anwender*innen, die Hilfe im deutschsprachigen Raum suchen (Stand Dezember 2025)

Chatbot	AinoAid	Maya	Ruth	Sophia Chat
Zugang	https://ainoaid.fr/	https://myprotectiv.org/	https://www.parasolcooperative.org/fr/ih/	https://sophia.chat/
Entwickler*innen	WeEncourage Oy Ltd und Teams der EU-Projekte IMPROVE und REACH	myProtectiv gUG	The Parasol Cooperative	Spring ACT
Fokus	KI-gestützter Chatbot zur frühen Orientierung und Veranltung ins Hilfesystem bei häuslicher und sexueller Gewalt in Deutschland, Österreich, Finnland, Frankreich (aktuell nur La Réunion) und Spanien (>100 Sprachen)	KI-gestützter Chatbot zur emotionalen Unterstützung und Vermittlung von Hilfsangeboten in Deutschland bei Beziehungsgewalt (Sprachen: Deutsch und Englisch)	Trauma-informierter, KI-gestützter Chatbot zur Sicherheitsplanung und Unterstützung bei häuslicher Gewalt mit spezieller Expertise für technologiegestützten Missbrauch und Sicherheit im Netz (>80 Sprachen, u. a. deutsch; Kontaktdaten und Informationen zu Unterstützungsdiensten in Deutschland, Österreich und der Schweiz hinterlegt)	KI-gestützter Chatbot zum Auffinden strukturierter, länder-spezifischer Informationen und Unterstützung bei häuslicher Gewalt (>23 Sprachen); zusätzliches Samme In von Beweismaterial über das Feature „Digital Safe“ auf der Homepage möglich
Inhalts-erstellung	durch internationale Expert*innen	Inhalte von kuratierten, geprüften Internetseiten sowie eigene erstellte Inhalte	Keine Angabe auf Website	speziell geschult auf von Experten geprüfte Ressourcen zum Thema häusliche Gewalt
Qualitäts-sicherung	Evaluation (Deutschland: Vogt et al., 2026; Österreich: Vogt et al., im Druck; Spanien: Ministerio del Interior; 2025); Feedback von Nutzer*innen fließt in weitere Entwicklung ein	Feedback ehemaliger Betroffener, Expert*innen und Nutzer*innen fließt in weitere Entwicklung ein; Impact Report (MyProtectiv, 2025)	Keine Angabe auf Website	Evaluation (Maeng & Lee, 2022)
Daten-sicherheit	Speicherung anonymisierter Konversationsdaten in der zugangs-geschützten Microsoft Azure Cosmos Datenbank	Daten werden 90 Tage pseudonymisiert gespeichert	Keine Erfassung von IP-Adressen, Geräte- oder IDs; alle Chatkonversationen werden aus Gründen der Qualitätssicherung nur wenige Tage lang gespeichert und anschließend endgültig gelöscht	Keine Speicherung von Chatkonversationen
System-sicherheit	Validiert, Details unter Punkt 5	Keine Angabe auf Website	Keine Angabe auf Website	Keine Angabe auf Website

3.3 Präventive Potenziale digitaler Chatbots

Digitale Chatbots können an einer Stelle unterstützen, die für viele Betroffene häuslicher Gewalt bislang nur eingeschränkt zugänglich war: in den frühen Phasen der Orientierung, in denen erste Irritationen, Unsicherheiten oder Fragen zur Einordnung des eigenen Erlebens entstehen. Da in dieser Phase selten klare Anliegen formuliert werden (können), ist die Möglichkeit, unscharfe oder lückenhafte Fragen anonym und jederzeit zu stellen, besonders bedeutsam. Durch die gezielte Vermittlung von Informationen unterstützen Chatbots Betroffene häuslicher Gewalt dabei, missbräuchliches Verhalten zu benennen, ihre Problemlage besser zu verstehen und realistische Erwartungen bezüglich ihrer Rechte und den Angeboten des Hilfesystems zu entwickeln.

Ein zentrales präventives Potenzial digitaler Chatbots liegt damit in ihrer Niedrigschwelligkeit: Sie können ohne persönlichen Kontakt genutzt werden. Website-basierte Chatbots sind zudem ohne Download und Registrierungs- oder Anmeldeprozeduren realisierbar. Die uneingeschränkte digitale Erreichbarkeit zu jeder Tages- und Nachtzeit erleichtert die Informationssuche insbesondere für Menschen, die aufgrund situativer Barrieren – etwa durch fehlende geschützte Gesprächsmöglichkeiten oder eingeschränkte Mobilität – keine offene Recherche oder Kontaktaufnahme riskieren können.

Ein weiteres Potenzial besteht in der strukturierten Wissensvermittlung. Während Suchmaschinen eine große Bandbreite nicht kuratierter Informationen bereitstellen, können Chatbots die Suche auf fachlich geprüfte Inhalte einschränken, Antworten modular strukturieren und so die kognitive Belastung während der Orientierung reduzieren. Dies umfasst grundlegende Informationen zu Gewaltformen, Dynamiken, mögliche rechtliche Optionen, Schutzmöglichkeiten und Unterstützungsstrukturen. Durch die Anpassung an eingegebene Fragen kann das System Inhalte in einer Reihenfolge bereitstellen, die den individuellen Orientierungsschritten der Nutzer*innen entspricht.

Von präventiver Bedeutung ist zudem die Möglichkeit, Risiko- und Schutzfaktoren sichtbar zu machen. Chatbots können Hinweise auf typische Muster, Dynamiken oder Sicherheitsaspekte geben und so Reflektionsprozesse erleichtern, die Betroffene sonst häufig allein und ohne fachliche Orientierung durchlaufen.

Insgesamt kann mithilfe von Chatbots

- das Ermitteln und Eingehen auf Optionen und Bedürfnisse der Nutzer*innen,
- ein umfassenderes Verständnis von Gewalt, Missbrauch und Schaden,
- ein schneller Zugang zu Ressourcen und Unterstützungsangeboten,
- sowie die verbesserte Sicherheit gewaltbetroffener Nutzer*innen

erreicht werden (Wood et al., 2022).

Darüber hinaus bieten Chatbots Potenzial hinsichtlich ihrer Konsistenz und Skalierbarkeit. Sie können unabhängig von regionalen Kapazitäten oder Öffnungszeiten genutzt werden und Inhalte in gleichbleibender Qualität bereitstellen. Gleichzeitig können aggregierte, anonymisierte Nutzungsdaten Hinweise darauf geben, welche Themen oder Unsicherheiten in frühen Phasen besonders häufig auftreten – Informationen, die für die Weiterentwicklung von Präventionsangeboten relevant sein können.

3.4 Grenzen digitaler Chatbots

Digitale Chatbots unterliegen im Kontext häuslicher Gewalt spezifischen Grenzen, die sich aus technischen Voraussetzungen und den Anforderungen eines derart hochsensiblen Themenfeldes ergeben. Diese Grenzen betreffen Barrieren der Zugänglichkeit, datenschutzbezogene Risiken, die begrenzte inhaltliche Verlässlichkeit KI-gestützter Systeme, die fehlende Möglichkeit zur situativen Einschätzung etwaiger Gefährdungen und potenziell schädigende Bindungsdynamiken, wie wir im Folgenden veranschaulichen.

Um einen Chatbot überhaupt nutzen zu können, benötigen die Gewaltbetroffenen den Zugang zu einem internetfähigen Gerät (Smartphone, Tablet, PC), einen Internetzugang und auch eine grundsätzliche Lese- und Schreibkompetenz. Diese Voraussetzungen schließen einige Betroffenenengruppen aus.

Risiken für gewaltbetroffene Nutzer*innen ergeben sich insbesondere bei starker Überwachung inklusive der Kontrolle der digitalen Geräte.

Daher sollten Chatbots datensparsam gestaltet sein, keine Registrierung erfordern und möglichst wenige digitale Spuren erzeugen – wenngleich sich Risiken nie vollständig ausschließen lassen.

Auch die inhaltliche Verlässlichkeit stellt eine Grenze dar: Sie hängt von der Qualität der Informationsquellen und der Funktionsweise des Chatbots ab. Doch selbst bei sorgfältig kuratierten und streng regelbasierten Systemen können Antworten ungenau oder missverständlich sein – etwa aufgrund von Mehrdeutigkeiten der Eingaben der Nutzer*innen. KI-gestützte Systeme bergen das Risiko, unpassende Antworten und Formulierungen zu generieren, wenn sie nicht ausreichend klar definiert sind. Um Falschinformationen vorzubeugen, sollte insbesondere technisch sichergestellt sein, dass KI-basierte Chatbots nicht „halluzinieren“, das heißt, Antworten erfinden, wenn sie nicht auf die für eine Antwort notwendigen Informationen zugreifen können.

Eine weitere Grenze digitaler Chatbots besteht darin, dass sie keine individuelle Fallbewertung vornehmen können. Sie erfassen Eingaben, können jedoch komplexe Risikolagen, Eskalationsdynamiken oder Sicherheitsbedarfe nicht verlässlich einschätzen. Daher sollten sie weder Entscheidungen vorbereiten noch konkrete Empfehlungen aussprechen. Präventionslogisch verbleibt ihre Aufgabe klar im Bereich der Orientierung. Eine automatisierte Gefährdungsbeurteilung ist nicht verantwortbar, da hierfür (bisher) kein validiertes Instrument existiert, das auf Selbstauskünften basierend verlässliche Einschätzungen erlauben würde.

Nicht zuletzt können dialogische Systeme unbeabsichtigte Bindungseffekte auslösen: Responsives Feedback kann dazu führen, dass der Chatbot als Ersatz für menschliche Unterstützung wahrgenommen wird (Babu et al., 2025). Dies kann den Schritt ins Hilfesystem verzögern. Eine transparent kommunizierte Abgrenzung der Rolle des Systems ist daher erforderlich.

Insgesamt wird deutlich, dass Chatbots einen spezifischen, begrenzten Beitrag im präventiven Unterstützungssystem leisten können. Ihr Nutzen entsteht dort, wo sie Orientierung ermöglichen – nicht dort, wo Einschätzung, Entscheidung oder Intervention erforderlich wäre. Dieser Rahmen bildet zugleich die Grundlage für das folgende Kapitel, das die ethischen Anforderungen an Chatbots im Kontext häuslicher Gewalt systematisch beleuchtet.

4. Ethische Anforderungen an Chatbots im Kontext häuslicher Gewalt

Digitale Chatbots agieren im Kontext häuslicher Gewalt in einem Spannungsfeld aus Orientierungsbedarf, Schutzanforderungen und digitalen Risiken. Aus den zuvor beschriebenen Dynamiken, Barrieren der Hilfe-suche und Grenzen digitaler Systeme ergeben sich ethische Anforderungen, die die Sicherheit der Nutzung von Chatbots erhöhen. In diesem Überblick über zentrale ethische Aspekte berücksichtigen wir die Vorgaben, die sich aus internationalen KI-Ethikleitlinien für die Prävention im Kontext häuslicher Gewalt ableiten lassen (European Commission, 2019; UNESCO, 2021; WHO, 2021).

4.1 Sicherheit und technische Schutzmechanismen

Häusliche Gewalt kennzeichnet sich insbesondere durch Kontrolle, und diese schließt auch die digitalen Geräte der Betroffenen mit ein. Sichtbare Spuren, wie Registrierungen oder Datenübermittlungen, die thematisch mit häuslicher Gewalt in Verbindung stehen, können für Betroffene gefährlich werden. Digitale Sicherheit zu gewährleisten ist daher ein zentraler Schutzfaktor, wie internationale Empfehlungen zur Gestaltung sicherer Technologien im Gewaltkontext herausstellen.

Chatbots sollten für Außenstehende nicht identifizierbar implementiert und datensparsam entwickelt sein, beispielweise durch:

- Verzicht auf eindeutig identifizierbare Apps
- Entfallen einer Registrierung
- Reduktion digitaler Spuren (z. B. personalisierte Log-ins, Speicherung sensibler Daten, persistente Verlaufseinträge)

Gleichzeitig sollten Chatbots transparent kommunizieren, welche Risiken verbleiben, etwa, dass Chatverläufe unter Umständen im Browserverlauf sichtbar bleiben können oder dass Täter*innen Endgeräte kompromittieren könnten.

4.2 Zugänglichkeit und Fairness

Digitale Angebote können unbeabsichtigt bestimmte Gruppen ausschließen – technisch, sprachlich oder inhaltlich. Betroffene häuslicher Gewalt haben vielfältige Lebenssituationen und Hintergründe. Internationale Rahmenwerke betonen Inklusion, Nichtdiskriminierung und Fairness als zentrale Prinzipien verantwortungsvoller KI.

Chatbots sollten so gestaltet sein, dass sie möglichst keine Nutzer*innengruppe ausschließen:

- Barrierearme Gestaltung (z. B. Lesbarkeit per Screenreader, auditive Nutzbarkeit)
- Bezug auf verschiedene Betroffenengruppen (Frauen, Männer, nicht-binäre Personen, Angehörige, Fachkräfte oder Personen mit Einschränkungen, ohne gesicherten Aufenthaltsstatus etc.)
- Mehrsprachigkeit
- Klare, zielgruppenorientierte Sprache, Variante(n) in einfacher Sprache
- Vermeidung diskriminierender oder stereotypisierender Aussagen (z. B. Victim-Blaming oder typische Täter-Narrative)

4.3 Fachliche Qualität und Kuratierung der Inhalte

Unzutreffende, veraltete oder missverständliche Inhalte können im Kontext häuslicher Gewalt erhebliche Risiken erzeugen: Sie können problematische Konstellationen verharmlosen und falsche Sicherheit vermitteln. Insbesondere in der frühen Orientierungsphase Gewaltbetroffener ist die fachliche Qualität der Informationen zentral für die effektive präventive Wirksamkeit. Die Vermittlung strukturierter Hinweise auf Hilfsangebote kann die Autonomie der Gewaltbetroffenen stärken und das Risiko reduzieren, dass Betroffene im digitalen Raum „steckenbleiben“.

Chatbots sollten daher auf wissenschaftlich fundierten, geprüften und kontinuierlich aktualisierten Inhalten basieren. Dazu gehören:

- Transparente Herkunft der Inhalte: Erkennbar, wer Inhalte erstellt oder geprüft hat
- Aktualität: Regelmäßige Überarbeitung, um rechtliche, institutionelle und fachliche Entwicklungen abzubilden
- Konkrete Hinweise auf qualifizierte Hilfen

Chatbots sollten keine individuellen Fallbewertungen und Risikoeinschätzungen leisten und somit keine individuellen diagnostischen oder prognostischen Aussagen erzeugen.

4.4 Transparenz der Systemgrenzen

Wenn Chatbots als menschliche Beratung fehlinterpretiert werden, können riskante Entscheidungen oder Verzögerungen in der Hilfesuche entstehen. Nutzer*innen benötigen Klarheit darüber, mit wem sie interagieren und welche Aufgaben das System übernehmen kann. Internationale Leitlinien betonen daher, wie wichtig diese Transparenz ist, um Fehlschreibungen zu vermeiden und eine sichere Nutzung zu ermöglichen.

Chatbots sollten deutlich erkennbar machen:

- dass es sich um ein technisches System handelt
- welche Funktionen abgedeckt sind (z. B. Orientierung statt Beratung)
- wo die inhaltlichen und technischen Grenzen liegen

4.5 Verantwortliche Interaktionsgestaltung

Dialogische, KI-basierte Systeme können auch auf emotionaler Ebene wirken. Die empathische Interaktion mit einem Chatbot kann eine positive Bindung herstellen, die den Schritt zur professionellen Hilfe verzögert. Auch eine zu sachliche Dialogform oder eine Interaktion, die implizit negative Bewertungen von Gewaltbetroffenen transportiert, kann den Übergang ins Schutz- und Hilfesystem behindern. Betroffene könnten die Informationssuche abbrechen, wenn sie sich emotional gar nicht vom Chatbot abgeholt fühlen, oder sie schreiben sich gar aufgrund einer (vermeintlich) negativen Bewertung der von ihnen geteilten Inhalte selbst die Verantwortung für die Gewaltdynamik zu und relativieren dadurch das Verhalten der gewaltausübenden Person.

Die dialogische Verwendung eines KI-basierten Chatbots in diesem Konzept birgt ein Dilemma: Einerseits besteht das Risiko, dass eine emotionale Bindung der Nutzer*innen an das System den Übergang zu professioneller Unterstützung verzögert, andererseits stellt eine bis zu einem gewissen Grad empathische Interaktion die Voraussetzung für die Akzeptanz und Nutzung eines solchen Chatbots dar.

Je geringer die Kontrolle darüber ist, wie die KI Informationen beziehungsweise das Fehlen relevanter Informationen im Chat verarbeitet sowie mit welcher Tonalität und wie direktiv die KI diese Inhalte dann kommuniziert, desto höher ist das Risiko, dass Gewaltbetroffene missverständlich oder falsch informiert beziehungsweise fehlgeleitet werden.

Die Vulnerabilität der Zielgruppe Gewaltbetroffener verlangt

- klare Hinweise: „Ich bin ein technisches System, kein Mensch“
- Verzicht der Simulation von Nähe und menschlichem Verständnis sowie empathisch bindungsfördernder Kommunikation
- Verzicht auf negative Bewertung der Gewaltbetroffenen und Unterstützenden
- Verzicht auf Druck bei der respektvollen Verdeutlichung und Wahrung der Eigenverantwortung der Gewaltbetroffenen
- Hinweise auf Handlungsoptionen anstelle von Entscheidungshilfen
- Übersichtliche, sachliche Strukturierung des Outputs zur Vermeidung von Fehlinterpretationen

4.6 Verantwortlichkeit, Dokumentation und Governance

Ohne klare Zuständigkeiten können Fehler schwer erkennbar und nicht korrigierbar sein. Verantwortungsstrukturen sind essenziell für Vertrauen, Qualitätssicherung und Nachvollziehbarkeit. Internationale Rahmenwerke vertrauenswürdiger KI betonen daher die Notwendigkeit von Institutionsklarheit und Dokumentation.

Chatbots benötigen

- eine klar benannte verantwortliche Institution
- dokumentierte Prozesse der Qualitätssicherung
- eine transparente technische Architektur
- Mechanismen für Rückmeldungen, Fehlerkorrekturen und kontinuierliche Evaluation

4. 7 Konsequenz für die KI-gestützte Prävention häuslicher Gewalt

Die beschriebenen Anforderungen zeigen, dass Chatbots im Kontext häuslicher Gewalt nicht als neutrale Technologie betrachtet werden können. Es handelt sich vielmehr um Instrumente, die Einfluss auf das Fühlen, Denken und Handeln Gewaltbetroffener nehmen, noch dazu während diese sich in der verletzlichen Phase der Informations- und Orientierungssuche befinden. Auch wenn es in dieser Phase viel Unterstützung bedarf, sollten Chatbots nicht darauf abzielen, möglichst viel zu ermöglichen.

Eine ethisch verantwortliche Gestaltung von Chatbots bedeutet, die technischen Möglichkeiten reflektiert und wissenschaftlich fundiert gezielt zu begrenzen und hierfür transparent Verantwortung zu übernehmen: bezogen auf die Interaktionsform, den Funktionsumfang und jede weitere technische Ausgestaltung.

Diese Leitlinien bilden den normativen Rahmen, innerhalb dessen KI-gestützte Präventionsangebote entwickelt und bewertet werden können. Sie stellen zugleich die Grundlage dar, um die nachfolgende Darstellung des KI-gestützten Chatbot AinoAid™ einzuordnen – als ein Beispiel für die praktische Umsetzung dieser Anforderungen unter den realen Forschungs- und Innovationsbedingungen eines europäischen Projekts.

5. AinoAid™ als Fallbeispiel verantwortungsvoller KI-Prävention

AinoAid™ wurde als Antwort auf empirisch belegte Versorgungslücken im Hilfesystem häuslicher Gewalt entwickelt. Diesen Prozess beschreiben wir nachfolgend und berücksichtigen hierbei den Entstehungskontext, die initiale Bedarfsanalyse, die Systembesonderheiten von AinoAid™ und Evaluationsergebnisse, um abschließend unsere Erfahrungen und Eindrücke aus der Forschungsperspektive zu teilen.

5. 1 Entstehungskontext: Von der Forschung zum digitalen Tool

Der Ursprung von AinoAid™ liegt im EU-Projekt IMPRODOVA, das in mehreren europäischen Ländern untersuchte, wie Polizei, Gesundheitswesen, Beratungsstellen und Justiz auf häusliche Gewalt reagieren und wo im Zusammenwirken dieser Akteur*innen Brüche entstehen (Vogt, 2020).

Auf dieser Grundlage veröffentlichte die Europäische Kommission einen Innovationsaufruf, der ausdrücklich die Entwicklung eines KI-gestützten Chatbots als Brücke zwischen Gewalterleben und Hilfesuche vorsah (European Commission, 2019). Ein Konsortium aus Forschungseinrichtungen, Praxispartnern und einem spezialisierten Start-up – WeEncourage Oy Ltd – erhielt im Rahmen des EU-Projekts IMPROVE den Auftrag, ein solches System zu konzipieren, technisch umzusetzen und zu evaluieren. Der Projektkontext brachte klare Rahmenbedingungen mit sich, insbesondere die zeitlich begrenzte Förderung sowie strenge ethische Prüfungen und Rechenschaftspflichten. Die Entwicklung orientierte sich nicht an Marktlogiken, sondern an dokumentierten Bedarfen von Gewaltbetroffenen und den Anforderungen vertrauenswürdiger KI in einem besonders sensiblen Kontext.

5. 2 Empirische Grundlage I: Bedarfsanalyse mit Gewaltbetroffenen

Die Konzeption von AinoAid™ wurde durch eine länderübergreifende Bedarfsanalyse vorbereitet, die im IMPROVE-Projekt durchgeführt wurde (Tananau Blumenschein et al., 2023). In fünf Ländern (Deutschland, Österreich, Finnland, Frankreich/La Réunion und Spanien) wurden insgesamt 80 betroffene Frauen qualitativ befragt. Die Befragten skizzierten ihre Wege in und durch das Hilfesystem, ihre bisherigen Erfahrungen und Erwartungen sowie ihre Vorstellungen eines möglichen Chatbots.

Ein zentrales Ergebnis war, dass das Hilfesystem aus Perspektive der Betroffenen als fragmentiert, wenig durchschaubar und schwer zugänglich erscheint. Polizei, Gerichte, Beratungsstellen, medizinische Angebote und Jugendamt wurden zwar als einzelne Akteure wahrgenommen, aber nicht als verbundenes System. Viele beschrieben, nicht zu wissen, „wo man anfangen soll“ oder welche Stelle in welcher Phase geeignet ist.

Darüber hinaus wurden starke emotionale Barrieren sichtbar. Scham, Loyalität gegenüber der gewaltausübenden Person, Sorge um die Kinder, Angst vor Eskalation und vor negativen Konsequenzen für Aufenthaltsstatus oder ökonomische Stabilität führten dazu, dass viele Befragte in frühen Phasen niemanden kontaktierten. Gleichwohl berichteten sie nahezu durchgängig, bereits vor einem Erstkontakt Informationen im Internet gesucht zu haben – zu Gewaltformen, Rechten, Schutzmöglichkeiten oder Erfahrungen anderer Betroffener.

Vor diesem Hintergrund formulierten die Befragten klare Erwartungen an einen möglichen Chatbot. Zentral waren Anonymität und Kontrolle: Sie wollten selbst bestimmen, welche Informationen sie preisgeben und in welchem Tempo sie Inhalte aufnehmen. Sie wünschten sich eine klare, nicht wertende und gut strukturierte Sprache, die hilft, das eigene Erleben zu benennen, ohne Druck auszuüben.

5.3 Iterativer Entwicklungsprozess

Auf Basis der Bedarfsanalyse wurde AinoAid™ in einem iterativen, multi-professionellen Prozess entwickelt. Zunächst entstand eine strukturierte Wissensbasis, die Inhalte zu Gewaltformen, Dynamiken, typischen Täterstrategien, rechtlichen Rahmenbedingungen, Schutzmöglichkeiten, Unterstützungsstrukturen und häufigen Fragen von Betroffenen bündelt. Diese Inhalte wurden von IMPROVE-Expert*innen aus Forschung und Praxis erstellt und mehrfach überprüft

Parallel wurde die Chatfunktion aufgebaut, die auf diese Wissensbasis zugreift. In mehreren Testschleifen wurde geprüft, wie die Antworten formuliert sein müssen, damit sie auch in belasteten Situationen verständlich bleiben, keine wertenden Untertöne transportieren und nicht zu Nähe suggerierenden Formulierungen greifen. Fachkräfte aus Frauenhäusern, Beratungsstellen, Polizei und Gesundheitswesen prüften in Workshops exemplarische Dialogverläufe und gaben Rückmeldungen zu Tonalität, Klarheit und Grenzen des Systems.

Dieser Prozess führte auch zu einer Erweiterung der ursprünglichen Zielgruppe. AinoAid™ wurde zunächst für Betroffene häuslicher Gewalt konzipiert, sollte aber – basierend auf den Befunden zu Unterstützungsbedarfen im Umfeld – auch Angehörige und Fachkräfte ansprechen.

Angehörige und Freund*innen nutzen den Chatbot, um das Erleben nahestehender Personen besser einordnen und sensibel ansprechen zu können. Fachkräfte können AinoAid™ einsetzen, um ihr Wissen über Dynamiken häuslicher Gewalt zu vertiefen, Formulierungsbeispiele für Gespräche zu erhalten oder konkrete Fragen zu institutionellen Abläufen zu klären. Damit wurde AinoAid™ bewusst als Mehrzielgruppen-Instrument gestaltet, das nicht nur Betroffene erreicht, sondern auch jene, die eine zentrale Rolle in der Prävention und Intervention spielen.

5.4 Systemarchitektur: Technische und ethische Konstruktion

Die technische Architektur von AinoAid™ spiegelt die zuvor formulierten ethischen Anforderungen wider und folgt den zentralen Prinzipien, die eng mit internationalen Leitlinien vertrauenswürdiger KI (European Commission, 2019; WHO, 2021; UNESCO, 2021) verknüpft sind.

Digitale Sicherheit und Datensparsamkeit sind essenziell. AinoAid™ ist webbasiert und erfordert keinen Download, keine App-Installation und keine Registrierung. Nutzer*innen können den Chat über die Website aufrufen, ohne ein Konto anzulegen oder personenbezogene Daten anzugeben. Die Oberfläche ist bewusst zurückhaltend gestaltet; Titel und Menüpunkte vermeiden eindeutige Gewaltbegriffe im Browsertitel, um Aufmerksamkeit Dritter nicht unnötig zu wecken. Ein auffälliger Notausstiegs-Button ermöglicht es, den Chat und die Seite jederzeit schnell zu verlassen und auf eine neutrale Suchmaske zu wechseln. Die technische Infrastruktur ist so ausgelegt, dass keine IP-Adressen, Gerätekennungen oder andere identifizierende Merkmale dauerhaft gespeichert werden. Gesprächsdaten werden, soweit für die technische Stabilität erforderlich, nur in anonymisierter Form und begrenzt vorgehalten; personenbezogene Angaben werden mit automatisierten Erkennungsmechanismen herausgefiltert. Damit wird dem Risiko Rechnung getragen, dass ggf. gewaltausübende Personen digitale Geräte kontrollieren und Nutzungsspuren auswerten.

AinoAid™ berücksichtigt das Prinzip der Zugänglichkeit und Fairness, indem es in mehreren Sprachen zur Verfügung steht (u. a. Deutsch, Englisch, Finnisch, Französisch, Spanisch) und durch die Einbindung eines Übersetzungsdienstes in weiteren Sprachen genutzt werden kann. Inhaltlich richtet sich der Chatbot nicht nur an weibliche Betroffene, sondern explizit auch an Männer, nicht-binäre Personen, Angehörige

und Fachkräfte. Die Sprache ist möglichst klar und frei von Fachjargon gehalten; dort, wo Fachbegriffe unvermeidbar sind, werden sie erläutert.

Eine Besonderheit von AinoAid™ als Innovationsprodukt eines Forschungskonsortiums ist der hohe Anspruch, der an die Erstellung und Kuratierung der Inhalte gelegt wurde. So greift der Chatbot ausschließlich auf eine geprüfte Wissensbank zurück. Diese wurde von innerhalb des IMPROVE-Konsortiums erstellt und orientiert sich an aktuellen nationalen und internationalen Standards, rechtlichen Rahmenbedingungen und einschlägiger Forschung (z. B. Council of Europe, 2011; WHO, 2020). Für jedes Partnerland existieren verantwortliche Ansprechpersonen, die auf die Aktualität der landesspezifischen Inhalte achten, insbesondere im Hinblick auf Notrufnummern, Beratungsstellen, rechtliche Änderungen oder neue Unterstützungsangebote. Damit soll verhindert werden, dass der Chatbot veraltete oder falsche Hinweise gibt oder Inhalte „halluziniert“, wenn Informationen fehlen.

Die Interaktionsgestaltung von AinoAid™ ist so konzipiert, dass der Chatbot eine sachliche, zugewandte, aber klar maschinelle Kommunikationsweise beibehält. Der Chat macht transparent, dass er ein technisches System ist und keine Beratung ersetzt. Formulierungen orientieren sich an psychoedukativen und trauma-sensiblen Grundsätzen, ohne therapeutische Nähe zu imitieren. Es werden keine individuellen Risikobewertungen vorgenommen, keine Diagnosen gestellt und keine expliziten Handlungsaufforderungen formuliert. Stattdessen werden Gewaltformen erklärt, Dynamiken erläutert und verschiedene Handlungsoptionen einander gegenübergestellt, einschließlich Kontaktwegen ins Hilfesystem. Die Antworten sind so strukturiert, dass Betroffene selbst entscheiden können, welche Schritte sie gehen möchten. Diese begrenzte, „gebundene“ Empathie ist ein bewusst gewählter Schutzmechanismus, um emotionale Abhängigkeiten und Fehlinterpretationen der Rollenverteilung zu vermeiden.

Technisch basiert AinoAid™ auf einem GPT-Modell, das in eine geschlossene Umgebung eingebunden ist. Die generativen Fähigkeiten werden so genutzt, dass sprachliche Anpassung und Dialogfluss möglich sind, während der inhaltliche Zugriff strikt auf die kuratierte Wissensbasis beschränkt bleibt. In Kombination mit Übersetzungskomponenten ermöglicht dies eine mehrsprachige, gleichzeitig kontrollierte Nutzung. Die Architektur setzt damit die zuvor formulierten ethischen Anforderungen in ein konkretes, im Gewaltkontext verantwortbares Design um.

5.5 Empirische Grundlage II: Evaluation mit Gewaltbetroffenen in Deutschland

Um die Nutzungserfahrungen mit AinoAid™ zu analysieren, wurde eine Evaluation mit 1312 deutschen (Vogt & Giljohann, 2025) und 745 österreichischen Nutzer*innen (Vogt et al., im Druck) vorgenommen. Grundlage waren die im Rahmen von Panelumfragen erhobenen Daten, bei der Testpersonen den Chatbot nutzten und anschließend standardisierte sowie offene Fragen beantworteten (ebd.).

Für die vorliegende Auswertung der Gesamtstichprobe (Vogt & Giljohann, 2025) wurden gesondert 160 deutsche User*innen berücksichtigt, die angaben, aktuell oder in der Vergangenheit von häuslicher Gewalt betroffen gewesen zu sein (74,8 % von ihnen identifizierten sich als weiblich, 24,5 % als männlich, 0,6 % als divers). Die Altersverteilung reichte von 18 bis über 70 Jahre, mit einer gleichmäßigen Verteilung über die Altersgruppen.

Die quantitative Auswertung erfolgte mittels etablierter Usability-Skalen, die Nützlichkeit, Informationsgehalt, Sicherheitsempfinden, Verständlichkeit der Sprache sowie Design und Navigation der Website erfassten (Hajesmaeel-Gohari et al., 2022). Ergänzend wurden Freitextantworten thematisch analysiert, um Qualitätsaspekte aus Nutzer*innensicht zu erfassen.

Die Ergebnisse zeigen ein konsistent positives Bild (Tabelle 2).

Tabelle 2: Mittelwerte und Standardabweichungen der erfassten Variablen zur Nutzerfreundlichkeit (Usability) von AinoAid™

Item	Mittelwert	Standardabweichung
<i>Nutzerfreundlichkeit des Chatbots</i>		
3. Nützlichkeit ^a	2,16	1,07
4. Informationsgehalt ^b	2,15	1,05
5. Sicherheit ^c	1,10	0,30
<i>Sprache des Chatbots^d</i>		
6. Einfach	1,94	1,06
7. Freundlich	1,76	0,90
8. Warm	2,18	1,00
9. Einfühlsam	2,29	1,04
10. Angenehm formell	2,12	0,93
11. Verständlichkeit ^e	1,73	0,93
<i>Nutzerfreundlichkeit der Website</i>		
12. Design ^f	2,04	0,94
13. Navigation ^g	1,82	0,89

Anmerkungen: a) „Wie hilfreich fanden Sie den Austausch mit dem Chatbot?“ (5-Punkte-Likert-Skala: 1 = sehr hilfreich, 5 = gar nicht hilfreich); b) „Wie informativ fanden Sie die Antworten des Chatbots?“ (5-Punkte-Likert-Skala: 1 = sehr informativ, 5 = gar nicht informativ); c) „Haben Sie die Nutzung des Chatbots als sicher empfunden?“ (bipolare Skala: 1 = ja, 2 = nein); d) „Wie wirkte der Sprachstil des Chatbots auf Sie?“ (5-stufiges semantisches Differenzial; positiver Pol = 1, negativer Pol = 5); e) „Wie verständlich empfanden Sie die Antworten des Chatbots?“ (5-stufige Likert-Skala: 1 = sehr gut verständlich, 5 = gar nicht verständlich); f) „Wie gefällt Ihnen das Design der Website?“ (5-stufige Likert-Skala: 1 = sehr gut, 5 = überhaupt nicht gut); g) „Wie gut konnten Sie sich auf der Website zurechtfinden?“ (5-stufige Likert-Skala: 1 = sehr gut, 5 = überhaupt nicht gut).

Insgesamt zeigt sich, dass alle erfassten Merkmale der AinoAid™-Homepage und des Chatbots positiv bewertet wurden, mit der wahrgenommenen Sicherheit (M = 1,10) als positivsten und der einfühlsamen Sprache des Chatbots (M = 2,29) als negativsten Ausprägung. Zudem variierten die Standardabweichungen in einem akzeptablen Ausmaß (SD = 0,30 für Chatbot-Sicherheit; SD = 1,07 für die Nützlichkeit der Chatbot-Interaktion), was auf eine allgemein ähnlich positive Bewertung des Chatbots durch die Befragten hinweist.

Die qualitative Analyse der Freitextantworten differenziert dieses Bild. Ein Teil der Befragten äußerte sich ausschließlich positiv und betonte Dankbarkeit für ein Angebot, das fachlich fundierte Informationen zu Gewalt und Hilfe in einer sicheren und anonymen Form zugänglich macht. Andere wiesen auf Verbesserungsmöglichkeiten hin: Gewünscht wurden zum

Beispiel noch persönlichere, ermutigendere und zugleich klar strukturierte Antworten, eine stärkere Berücksichtigung männlicher Betroffener und Jugendlicher sowie konkretere, regionale Hinweise auf Hilfsangebote. Einige machten darauf aufmerksam, dass lange oder komplexe Antworten in Belastungssituationen schwer zu verarbeiten seien und dass eine weitere Vereinfachung oder Segmentierung hilfreich sein könnte.

Gleichzeitig wurde in den Rückmeldungen immer wieder deutlich, dass Nutzer*innen die Rolle des Chatbots überwiegend korrekt einordnen. Sie beschrieben AinoAid™ als Hilfsmittel zur Einordnung, nicht als Ersatz für persönliche Unterstützung. Es fanden sich keine Hinweise darauf, dass AinoAid™ als „Bezugsperson“ missverstanden wurde oder dass eine emotionale Abhängigkeit entstand. Vielmehr berichteten mehrere Befragte, dass ihnen die Nutzung geholfen habe, Gewaltmuster zu erkennen, ihre Situation sprachlich zu fassen und den Gedanken an eine Kontaktaufnahme mit dem Hilfesystem überhaupt erstmals zuzulassen.

Aus präventionslogischer Perspektive weist diese Evaluationsstudie darauf hin, dass AinoAid™ die intendierte Orientierungsfunktion für Gewaltbetroffene grundsätzlich erfüllt. Der Chatbot wird als sicher und hilfreich wahrgenommen, unterstützt das Verständnis von Gewalt und Optionen im Hilfesystem und bleibt dabei in der Rolle eines klar begrenzten digitalen Angebots. Gleichzeitig machen die Ergebnisse deutlich, dass die Balance zwischen sachlicher Distanz und wahrnehmbarer Empathie fortlaufend feinjustiert werden sollte.

Zusätzlich verdeutlichen die Erfahrungen des Konsortialteams im Projekt IMPROVE die mit einer solchen Innovation im Forschungskontext verbundenen Herausforderungen. Das Team verfügte über wenige Ressourcen und geringe Marketingexpertise, um das digitale Innovationsprodukt in der Präventionslandschaft umfassend bekannt zu machen. Präventionstools wie AinoAid™ konkurrieren inzwischen mit anderen, teils kommerziell getriebenen Angeboten um Sichtbarkeit im digitalen Raum. Während wissenschaftliche Logiken auf sorgfältige Entwicklung, Prüfung und zurückhaltendes Marketing ausgerichtet sind, erfordern digitale Anwendungen frühzeitige Nutzung, kontinuierliche Anpassung und eine hohe Fehlertoleranz im Sinne derartiger iterativer Optimierungsprozesse. Diese unterschiedlichen Logiken treffen im Kontext von KI-gestützter Prävention unmittelbar aufeinander.

Ein weiterer limitierender Faktor ist die Verstetigung. Eine projektförmige Förderung wie im EU-Projekt IMPROVE erlaubt die Entwicklung und erste Evaluation, sichert aber keine langfristige Wartung, Aktualisierung und Weiterentwicklung. Im Fall von AinoAid™ kann glücklicherweise durch das aktuell anschließende EU-geförderte Projekt REACH eine inhaltliche Erweiterung auf weitere Gewaltformen sowie eine technische Weiterentwicklung erzielt werden: thematisch wird der Chatbot zusätzlich um die Themenbereiche sexualisierte Gewalt und Menschenhandel erweitert und darüber hinaus im System des „human-in-the-loop“-Ansatzes integriert, bei dem AinoAid™ den Nutzer*innen bei Bedarf ergänzend zum Chatbot ein unmittelbares menschliches Kontaktangebot anbietet.

6. Ecksteine für verantwortungsvolle KI-Prävention

Der Blick auf AinoAid™ veranschaulicht, wie ein KI-gestützter Chatbot im Kontext häuslicher Gewalt verantwortungsvoll gestaltet werden kann, wenn technische, inhaltliche und ethische Entscheidungen systematisch aufeinander bezogen werden. Ein solches digitales Angebot schließt hierbei weniger eine Versorgungs- als eine frühe Orientierungslücke, in der Betroffene oder ihr Umfeld noch keinen direkten Kontakt zum Hilfesystem aufgenommen haben, aber bereits aktiv nach Einordnung und Information suchen. Prävention beginnt damit in einer vorgelagerten Phase der digitalen Selbstorientierung, die vielfach unsichtbar und von Ambivalenzen, Ängsten und Unsicherheiten geprägt ist. Für eine vorausschauende Präventionsstrategie bedeutet dies eine strukturelle Verschiebung: Prävention muss dort wirksam werden, wo frühe Einordnungsprozesse stattfinden – also auch in digitalen Räumen, die bisher nicht als genuin präventive Orte verstanden wurden. KI-gestützte Chatbots können in dieser frühen Phase eine Funktion übernehmen, die klassische Hilfestrukturen naturgemäß nicht erfüllen können, denn sie können unabhängig von zeitlichen, sprachlichen und regionalen Limitationen der Kapazitäten genutzt werden.

Damit diese Brücke ins Schutz- und Hilfesystem trägt, sind verbindliche Qualitäts- und Sicherheitsstandards – als zentrale Schutzfaktoren im Gewaltkontext – erforderlich. Besonders hervorzuheben sind die Datensicherheit und -sparsamkeit, ebenso wie eine transparente Rollenbeschreibung: Nutzer*innen müssen verstehen, dass sie es mit einer KI-

basierten Anwendung zu tun haben, die Orientierung bietet, aber keine Entscheidungen trifft. *Inhalte müssen fachlich geprüft, aktuell und nachvollziehbar sein* und Interaktionsformen angepasst an traumasensible und diversitätssensible Prinzipien. Algorithmische Spekulationen, etwa frei generierte Einschätzungen ohne belastbare Wissensgrundlage, sind in diesem sicherheitskritischen Feld nicht akzeptabel.

Gleichzeitig ist die digitale Präventionslandschaft vielfältig, dynamisch, und Webseiten, Apps, sowie Chatbots koexistieren überwiegend ohne erkennbare Qualitätsunterschiede. Weder Gewaltbetroffene noch Angehörige oder Fachkräfte können ohne weiteres beurteilen, welche Angebote sicher, fachlich fundiert und datenschutzkonform sind. Diese Unsicherheit führt zu einer strukturellen Überforderung, weil bislang kaum institutionalisierte Strukturen existieren, die qualitative Einordnungen vornehmen. Als Orientierungshilfen sind übergreifende Klassifikationen digitaler Angebote, einfache Entscheidungshilfen (zum Beispiel: Wofür eignet sich ein Chatbot, wofür nicht?) und klar formulierte Indikationsgrenzen für die Nutzung in professionellen Kontexten vonnöten. Hier stellt sich die Frage der Zuständigkeit: Mögliche Akteur*innen könnten unabhängige Fachgremien sein, die Prüfkriterien entwickeln und Empfehlungen aussprechen.

Das Thema Sichtbarkeit schließt hier unmittelbar an: Selbst sehr gut konzipierte, evaluierte und ethisch reflektierte Tools bleiben wirkungslos, wenn sie ihre Zielgruppen nicht erreichen. Forschungskonsortien aber verfügen in der Regel weder über professionelle Kommunikationsstrukturen noch über Erfahrung im „Marketing“ digitaler Produkte. Gleichzeitig bringt der Markt Angebote hervor, die sich mit Werbedruck zu positionieren wissen, deren Qualität oder Sicherheit aber nicht zwingend geprüft ist. Aus einer präventionslogischen Perspektive ergibt sich daraus ein Spannungsfeld: Wer im digitalen Raum nach Hilfe sucht, findet eher das, was am sichtbarsten ist, nicht unbedingt das, was am besten geprüft ist. Präventive Verantwortung umfasst daher auch die institutionelle Unterstützung geprüfter Angebote bei der Bekanntmachung – etwa durch Einbindung in offizielle bundeweite Informationsportale, Verweise durch Koordinierungs- und Beratungsstellen oder polizeiliche, soziale und medizinische Einrichtungen.

Eng verknüpft damit ist die Frage der Verstetigung und Finanzierung. Digitale Präventionsangebote verursachen fortlaufende Kosten: Serverbetrieb, Cybersicherheit, technische Wartung, Barrierefreiheit, Überset-

zungen und inhaltliche Aktualisierung. Wissenschaftliche Projektförderungen ermöglichen die Entwicklung und erste Evaluation, bieten aber meist keinen verlässlichen Rahmen für den langfristigen Betrieb. Im Gewaltkontext ist ein kommerzielles Geschäftsmodell, das Hilfe „verkauft“ mit Gleichbehandlungsgrundsätzen und Zugangsansprüchen nicht vereinbar. Langfristig bedarf es daher tragfähiger Finanzierungsmodelle, die digitale KI-gestützte Prävention als Teil öffentlicher und politischer Verantwortung verstehen und entsprechend institutionell verankern.

Hinzu kommt der Bedarf an systematischer Evaluation und Weiterentwicklung. Angesichts der hohen Vulnerabilität der Zielgruppe ist zu berücksichtigen, dass jede Chatinteraktion eine Form von Intervention darstellt, deren Wirkungen auf Sicherheitsempfinden, Handlungsentscheidungen und weitere Hilfesuche wissenschaftlich noch wesentlich besser verstanden werden muss. Für den spezifischen Kontext häuslicher Gewalt liegen bislang jedoch nur wenige evaluierte Anwendungen vor. Notwendig sind differenzierte Evaluationsdesigns, die prüfen, inwieweit sich Orientierung, Risikoerkenntnis, Sicherheitsempfinden und tatsächliche Hilfesuche verändern – und ob unerwünschte Effekte auftreten. Partizipative Forschung, die Betroffene und Fachkräfte systematisch einbezieht, ist hierfür zentral.

Mit zunehmender Verbreitung wird auch die Frage nach sinnvollen Kombinationen von KI-gestützter Orientierung und menschlicher Unterstützung bedeutsamer. Ein möglicher Entwicklungspfad liegt in „human-in-the-loop“-Ansätzen, bei denen Chatbots als erstes, niedrigschwelliges Kontaktangebot fungieren und – wo sinnvoll und sicher – strukturierte Übergänge zu Beratung, Online-Sprechstunden oder anderen professionellen Hilfen unterstützen. Solche Modelle erfordern klare Schnittstellen, definierte Übergabepunkte und Schulungen für Fachkräfte, die digitale Informationen im eigenen Handeln einordnen und aufnehmen.

Schließlich ist zu betonen, dass technologische Lösungen in der Prävention häuslicher Gewalt prinzipielle Grenzen haben. Digitale Tools können nicht alle Betroffenen erreichen – etwa, wenn kein sicherer Zugang zu Geräten existiert oder wenn digitale Kontrolle durch Täter*innen zu hoch ist. Sie können zudem emotionale Unterstützung nur in begrenztem Maße leisten; tragfähige Beziehungen und komplexe Entscheidungsprozesse bleiben zutiefst menschliche Aufgaben. Prävention ist sozial eingebettet und rechtlich, institutionell und kulturell gerahmt. KI-gestützte

Systeme können in diesem Gefüge ein ergänzender Baustein sein, aber sie können weder strukturelle Defizite des Hilfesystems kompensieren noch gesellschaftliche Machtverhältnisse eigenständig verändern.

Literatur

- Alberton, A. M., Hertzog, J., & Bila, A. (2025). Insights on intimate partner violence service provision, access, and utilization during the COVID-19 pandemic: A scoping review. *Journal of Family Violence*.
- Albuquerque Netto, L. D., Moura, M. A. V., Queiroz, A. B. A., & Leite, F. M. C. (2017). Isolation of women in situation of violence by intimate partner: a social network condition. *Escola Anna Nery*, 21, e20170007.
- Ascorbe, P., Campos, M. S., Domínguez, C., Heras, J., & Reinares, A. R. T. (2023). prevenIA: A chatbot for information and prevention of suicide and other mental health disorders. In SEPLN (Projects and Demonstrations) (pp. 26-30). <https://ceur-ws.org/Vol-3516/paper06.pdf>
- Babu, D., Joseph, D., Kumar, R. M., Alexander, E., Sasi, R., & Joseph, J. (2025). Emotional AI and the rise of pseudo-intimacy: Are we trading authenticity for algorithmic affection? *Frontiers in Psychology*, 16, 1679324.
- Bauer, T., Devrim, E., Glazunov, M., Jaramillo, W.L., Mohan, B., & Spanakis, G. (2020). #MeTooMaastricht: Building a chatbot to assist survivors of sexual harassment. In P. Cellier & K. Driessens (eds.), *Machine learning and knowledge discovery in databases. ECML PKDD 2019. Communications in Computer and Information Science* (vol. 1167). Springer.
- BKA (2025). *Häusliche Gewalt: Bundeslagebild 2024*. Wiesbaden.
- Council of Europe (2011). *Convention on preventing and combating violence against women and domestic violence (Istanbul Convention)*. Council of Europe Treaty Series, 210. <https://www.coe.int/en/web/conventions/full-list?module=treaty-detail&treaty-num=210>.
- Daley, K., Hungerbuehler, I., Cavanagh, K., Claro, H. G., Swinton, P. A., & Kapps, M. (2020). Preliminary evaluation of the engagement and effectiveness of a mental health chatbot. *Frontiers in Digital Health*, 2, 576361.
- Delpeuch, T., Vassileva, M., Cohade, L., Leconte, L., Delmas, J., Houillot, M., ... & Vogt, C. (2024). Factors influencing the effectiveness of frontline response to domestic violence and potential solutions. IMPROVE project. https://www.improve-horizon.eu/s/IMPROVE_D13_Factors_Influencing_the_Effectiveness_of_Frontline_Response.pdf

- Epstein, D., & Goodman, L. A. (2018). Discounting women: Doubting domestic violence survivors' credibility and dismissing their experiences. *University of Pennsylvania Law Review*, 167, 399-461.
- European Commission (High-Level Expert Group on AI) (2019). Ethics guidelines for trustworthy AI. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
- Farhall, K., Harris, B., & Woodlock, D. (2020). The impact of rurality on women's' space for action in domestic violence. *International Journal of Rural Criminology*, 5, 181-203.
- Fitzsimmons-Craft, E. E., Chan, W. W., Smith, A. C., Firebaugh, M.-L., Fowler, L. A., Topooco, N., DePietro, B., Wilfley, D. E., Taylor, C. B., & Jacobson, N. C. (2022). Effectiveness of a chatbot for eating disorders prevention: A randomized clinical trial. *International Journal of Eating Disorders*, 55(3), 343-353.
- Follingstad, D. R. (2009). The impact of psychological aggression on women's mental health and behaviour. *Trauma, Violence, & Abuse*, 10 (3), 271-289.
- FRA (2020). Getting the future right: Artificial intelligence and fundamental rights. <https://fra.europa.eu/en/publication/2020/artificial-intelligence-and-fundamental-rights>
- FRA, EIGE, Eurostat (2024). EU gender-based violence survey. Key results. https://fra.europa.eu/sites/default/files/fra_uploads/eu-gender-based-violence-survey-key-results.pdf. Accessed 25 March 2025.
- Gabellini, E., Salvatori, A., Greco, M. T., Cattaneo, C., Tambuzzi, S., Costantino, M. A., & Russo, A. G. (2025). Access to health services by women subjected to violence. *BMC Women's Health*, 25, 61.
- Giljohann, S., Vogt, C., Sondern, L., Juszczuk, P., Kersten, J., & Pfeleiderer, B. (2021). Frontline response to high impact domestic violence in Germany. In B. Lobnikar, C. Vogt, & J. Kersten (eds.). *Improving frontline responses to domestic violence in Europe* (pp. 179-200). University of Maribor Press. <https://doi.org/10.18690/978-961-286-543-6>
- Goodey, J. (2017). Violence against women: Placing evidence from a European Union-wide survey in a policy context. *Journal of Interpersonal Violence*, 32, 1760-1791.
- Hajesmaeel-Gohari, S., Khordastan, F., Fatehi, F., Samzadeh, H. & Bahaa-dinbeigy, K. (2022). The most used questionnaires for evaluating satisfaction, usability, acceptance, and quality outcomes of mobile health. *BMC Medical Informatics and Decision Making*, 22, 22.
- Harris, B. A., & Woodlock, D. (2019). Digital coercive control: Insights from two landmark domestic violence studies. *British Journal of Criminology*, 59, 530-50.

- Hegarty, K., Sheehan, M., & Schonfeld, C. (2017). A multidimensional definition of partner abuse: Development and preliminary validation of the Composite Abuse Scale. In M. Natarajan (ed.), *Domestic Violence* (pp. 15-31). Routledge.
- Horn, S., Vogt, C., Wüller, C. & Görgen, T. (2024). Intimate partner homicide: Risk constellations in separation conflicts and points of intervention for the police. *Policing: A Journal of Policy and Practice*, 18, paae029, <https://doi.org/10.1093/police/paae029>
- Kang, K. A., Kim, S. J., Oh B. D., & Kim, Y. H. (2024). Development of a chatbot for school violence prevention among elementary school students in South Korea. *Child Health Nursing Research*, 30(1), 45-53.
- Lee, S., Yoon, J., Cho, Y., & Chun, J. (2024). A systematic review of chatbot-assisted interventions for substance use. *Frontiers in Psychiatry*, 15, 1456689.
- Lian, A. T., Costilla Reyes, A., & Hu, X. (2023). CAPTAIN: An AI-based chatbot for cyberbullying prevention and intervention. In: Degen, H., Ntoa, S. (eds.), *Artificial Intelligence in HCI. HCII 2023. Lecture Notes in Computer Science* (Vol. 14051, pp. 98–107). Springer.
- LKA Niedersachsen (2022). Bericht zu Gewalterfahrungen in Paarbeziehungen: Sonderbericht zur Befragung zu Sicherheit und Kriminalität in Niedersachsen 2021. https://www.lka.polizei-nds.de/download/75823/Sondermodul_Gewalterfahrungen_in_Paarbeziehungen_2021.pdf
- Maeng, W., & Lee, J. (2022, April). Designing and evaluating a chatbot for survivors of image-based sexual abuse. In *Proceedings of the 2022 CHI conference on human factors in computing systems* (no. 344, pp. 1-21). <https://courses.cs.umbc.edu/graduate/691/spring23/pdf/3491102.3517630.pdf>
- Maia, E., Vieira, P., & Praça, I. (2023). Empowering preventive care with GECA chatbot. *Healthcare*, 11(18), 2532.
- Mantler, T., Jackson, K. T., Shillington, K., Walsh, E. J., Tobah, S., Jackson, B., & Davidson, C. A. (2021). Factors influencing rural women's disclosure of intimate partner violence: A qualitative study. *SN Social Sciences*, 1, 1-19.
- Mela, M., Houtsonen, J., Izaguirre Choperena, A., May, A., Juusela, A., Pfleiderer, B., ..., & Koivukoski, U. (2023). Factors leading to low reporting of domestic violence and restricting access to service. IMPROVE Project. <https://doi.org/10.13140/RG.2.2.21029.47849>
- Mendoza-Pinto, R. (2023). Artificial intelligence in the fight against bullying: Integration of ChatGPT in an emotional support chatbot. In *CEUR Workshop Proceedings* (Vol. 3691, p. 30). <https://ceur-ws.org/Vol-3691/paper30.pdf>

- Ministerio del Interior (2025). Informe. ENACT: Dissemination of the IMPROVE project's tool. Unveröffentlichter Bericht.
- MyProtectify (2025). Impact report. <https://myprotectify.org/s/myProtectify-Impact-Report-Sep2025.pdf>
- Postmus, J. L., Hoge, G. L., Breckenridge, J., Sharp-Jeffs, N., & Chung, D. (2020). Economic abuse as an invisible form of domestic violence: A multicountry review. *Trauma, Violence, & Abuse, 21*(2), 261-283.
- Pritchard, A. J., Reckdenwald, A., & Nordham, C. (2017). Nonfatal strangulation as part of domestic violence: A review of research. *Trauma, Violence, & Abuse, 18*(4), 407-424.
- Rumpf, T., Horn, S., Vogt, C., Göbel, K., Görgen, T., Zibulsi, K. M., Uttenweiler, V., & Bondü, R. (2024). Leaking among intimate partner homicide perpetrators. A systematic review. *Trauma, Violence, & Abuse, 25*(4), 3005-3019. <https://doi.org/10.1177/15248380241237213>
- Sanz Urquijo, B., Izaguirre Choperena, A., & Lopez Belloso, M. (2024). Empowering change: Unveiling the synergy of feminist perspectives and AI tools in addressing domestic violence. *Communication Papers: Media Literacy & Gender Studies, 13*(27), 50-75.
- Schafer, M., Lachman, J.M., Gardner, F., ..., & Clements, L. (2023). Integrating intimate partner violence prevention content into a digital parenting chatbot intervention during COVID-19. *BMC Public Health 23*, 1708.
- Seligman, M. E., Rockstroh, B., Petermann, F., & Petermann, F. (1979). *Erlernte Hilflosigkeit*. München: Urban und Schwarzenberg.
- Storer, H. L., & Nyerges, E. X. (2023). The rapid uptake of digital technologies at domestic violence and sexual assault organizations during the COVID-19 pandemic. *Violence against Women, 29*, 1085-1096.
- Tananau Blumenschein, E., Hopf, S., Leonhardmair, N., Vogt, C., Kersten, J., Köpsel, N., ... & Vassileva, M. (2023). Victims' mental maps of institutional response to domestic violence and needs regarding AI chatbot. <https://doi.org/10.13140/RG.2.2.36148.42882>.
- UNESCO (2021). Recommendation on the ethics of artificial intelligence. <https://unesdoc.unesco.org/ark:/48223/pf0000381137>
- UNFPA (2023). Guidance on the safe and ethical use of technology to address gender-based violence and harmful practices (Implementation summary). <https://www.unfpa.org/resources/guidance-safe-and-ethical-use-technology-address-gbv-and-harmful-practices>

- Vogt, C. (2020). Interagency cooperation. Building capacity to manage domestic abuse. *European Law Enforcement Research Bulletin*, 19, 153-163. <https://doi.org/10.5281/zenodo.5512011>
- Vogt, C. & Giljohann, S. (2025). Der AinoAid™ Chatbot in Deutschland: Ein alternativer Weg ins Hilfesystem bei häuslicher Gewalt. *Kriminalistik*, 79(8-9), 450-457.
- Vogt, C. & Giljohann, S. & Hopf, S. (im Druck). Digitaler Begleiter für Betroffene und Fachkräfte bei häuslicher Gewalt: AinoAid™ in Österreich. *SIAK Journal*.
- Vogt, C., Giljohann, S., Köpsel, N., González Cabezas, S., Houtsonen, J., Izaguirre Choperena, A., Juusela, A., Tananau Blumenschein, E., Vassileva, M. (2026). From needs assessment to usability testing: Evaluating the AinoAid™ chatbot for domestic violence support. *BMC Women's Health*, 26, 31. <https://doi.org/10.1186/s12905-025-04202-3>
- Vogt, C. & Kersten, J. (2022). Human factors shaping the cooperation of police with other sectors: The example of domestic violence. *Brazilian Journal of Police Sciences*, 13(10), 29-59. <https://doi.org.br/10.31412/rbcp.v13i10.1015>
- Walker, L. E. (2006). Battered woman syndrome: Empirical findings. *Annals of the New York Academy of Sciences*, 1087(1), 142-157.
- WHO (2020). Responding to intimate partner violence and sexual violence against women: WHO clinical and policy guidelines. <https://www.who.int/publications/i/item/9789241548595>
- WHO (2021). Ethics and governance of artificial intelligence for health. <https://www.who.int/publications/i/9789240029200>
- Wood, L., Hairston, D., Schrag, R. V., Clark, E., Parra-Cardona, R., & Temple, J. R. (2022). Creating a digital trauma informed space: Chat and text advocacy for survivors of violence. *Journal of Interpersonal Violence*, 37, NP18960-NP18987.

Zur weiteren Vertiefung

- Leitgöb-Guzy, N., & Bieber, I. (2026). Ergebnisse der Dunkelfeldstudie "Lebenssituation Sicherheit und Belastung im Alltag (LeSu-BiA)" I. Bericht des BMBFSFJ, BMI & BKA (Hrsg.). https://www.bka.de/SharedDocs/Downloads/DE/Publikationen/Publikationsreihen/Forschungsergebnisse/260210_LeSuBiA_Ergebnisse_I.html?nn=261272

- Lobnikar, B., Vogt, C., & Kersten, J. (Hrsg.) (2021). Improving frontline responses to domestic violence in Europe. University of Maribor Press. <https://doi.org/10.18690/978-961-286-543-6>
- Mielismäki, H., & Husso, M. (2025). Ethical implications of AI-driven chatbots in domestic violence support. *Social Inclusion*, 13.

Mediathek



IMPROVE: Die Rolle der Künstlichen Intelligenz in Fällen von häuslicher Gewalt



IMPROVE Webinar — Breaking barriers: How can AI support victims of domestic violence?



Dipl.-Psych. Stefanie Giljohann ist Referentin in der Personalentwicklung an der Technischen Universität Berlin. Bis Februar 2026 war sie zudem wissenschaftliche Mitarbeiterin in der Arbeitsgruppe „Cognition and Gender“ an der Universität Münster in den beiden EU-Projekten IMPROVE und VIPROM, in denen sie sich insbesondere mit der Erstellung der Trainingsplattform für Ersthelfende im Bereich häuslicher Gewalt (<https://training.improve-horizon.eu/>)

befasste. Zuvor war sie bereits in einem weiteren EU-Projekt zur Bekämpfung häuslicher Gewalt (IMPRODOVA) beim LKA Berlin sowie in weiteren (inter)nationalen Forschungsprojekten beschäftigt. Ihre Forschungsschwerpunkte liegen in der Gewaltprävention und Rechtspsychologie. Zusätzlich ist Frau Giljohann in Selbständigkeit als zertifizierte Trainerin, Coachin und Schreibberaterin tätig, nachdem sie zuvor 8 Jahre an der TU Berlin in der wissenschaftlichen Weiterbildung zwei Programmbereiche für Nachwuchswissenschaftler*innen verantwortet hatte.



Dr. Catharina Vogt, ist wissenschaftliche Mitarbeiterin im Fachgebiet Kriminologie und interdisziplinäre Kriminalprävention an der Deutschen Hochschule der Polizei in Münster. Hier leitete sie das EU-geförderte Projekt IMPROVE zur Prä- und Intervention bei häuslicher Gewalt. Im Rahmen von IMPROVE wurde der – mit dem Security Innovation Award der EU ausgezeichnete– Chatbot AinoAid™ für Betroffene häuslicher Gewalt (<https://ainoaid.fi/>) und die kostenfreie IMPROVE Trainingsplattform zur Verbesserung der Ausbildung und Zusammenarbeit professioneller Ersthelfender (<https://training.improve-horizon.eu/de/>) entwickelt. Nach ihrem Psychologiestudium an der TU Dresden promovierte sie zum Thema „Respektvolle Führung“. Ihre Forschung konzentriert sich auf Führung und Konfliktmanagement und wurde durch den DAAD und die Kühne Logistics University gefördert. Von 2014 bis 2018 leitete sie die RespectResearchGroup an der Universität Hamburg. Nebenberuflich arbeitet sie als Beraterin und Trainerin.

» AUS FACHLICHER PERSPEKTIVE ERSCHEINT ES ZUNEHMEND BEDEUT- SAM, DIE ENTWICKLUNG KÜNSTLICHER INTELLIGENZ NICHT NUR ANALYTISCH ZU BEGLEITEN, SONDERN AUCH PRÄVENTIV MITZUGESTALTEN «

Wissenschaftliche Begleitschrift

Der DPT – Deutscher Präventionstag gGmbH ist der weltweit größte Jahreskongress für Kriminalprävention und angrenzende Präventionsbereiche. Seit 1995 bietet der DPT eine internationale Plattform für Information, Wissenstransfer und interdisziplinären Dialog zwischen Präventionspraxis, Präventionsforschung und Präventionspolitik.

Jeder Jahreskongress steht unter einem besonderen Schwerpunktthema. Hierzu lässt der DPT jeweils eine wissenschaftliche Begleitschrift erstellen. In einer interdisziplinären Perspektive werden die Hintergründe dargestellt und im Hinblick auf das Fachgebiet der Gewalt- und Kriminalprävention analysiert.

Der 31. Deutsche Präventionstag widmet sich neben allen anderen Themen im weiten Feld der Gewalt- und Kriminalprävention dem Schwerpunktthema „KI in der Prävention“. Die wissenschaftliche Begleitschrift wird bereits im Vorfeld des Kongresses veröffentlicht. In Koordination durch Prof. Dr. Gina Rosa Wollinger wird darin das Schwerpunktthema aus verschiedenen wissenschaftlichen Perspektiven aufbereitet. Neben dem einleitenden und rahmenden Text von Frau Prof. Wollinger und einem Geleitwort von Dr. Michael Fübi werden sieben ausführliche Expertisen geboten.

