Entscheidungsautonomie und KI - Methodische Hinweise zur Untersuchung von KI-Nutzung in Sicherheitsbehörden

Julian Romeo Hildebrandt
hildebrandt@comm.rwth-aachen.de
RWTH Aachen University,
Human-Computer Interaction Center
Aachen, Germany

Martina Ziefle
ziefle@comm.rwth-aachen.de
RWTH Aachen University,
Human-Computer Interaction Center
Aachen, Germany

André Calero Valdez
calerovaldez@imis.uni-luebeck.de
Universität zu Lübeck, Institut für
Multimediale und Interaktive Systeme
Lübeck, Germany

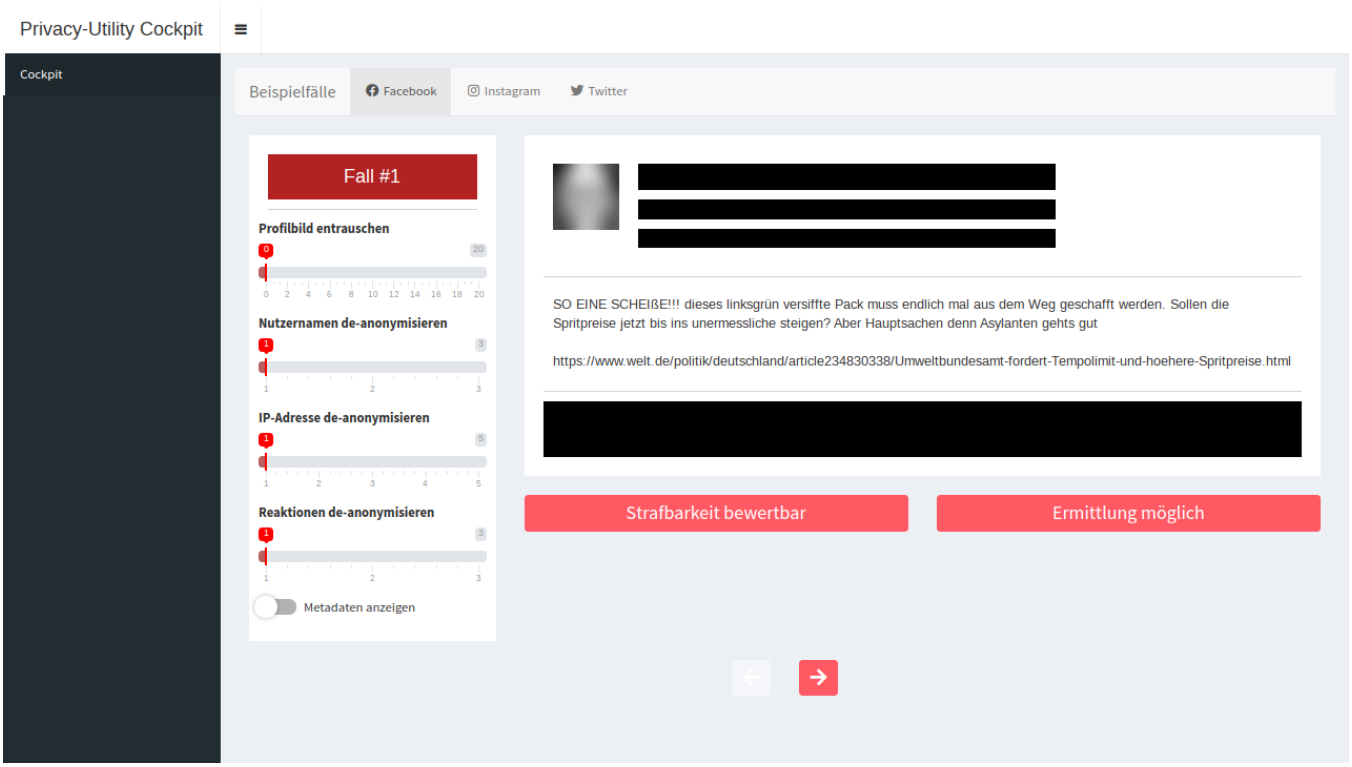

Abbildung 1: Entwurf Privacy-Utility Cockpit für BMBF-Projekt Kistra, 2022. Zustand zeigt maximale Privacy und minimale Utility.

ZUSAMMENFASSUNG

Künstlicher Intelligenz (KI) werden immense Potentiale zugeschrieben. Dies trifft auch auf den Bereich der zivilen Sicherheit zu und wird u.A. in dem BMBF geförderten Forschungsprojekt *KISTRA: Einsatz von KI zur Früherkennung von Straftaten* erforscht. Ziel des Projektes ist der rechtlich, ethisch und sozial vertretbare Einsatz von KI-Klassifizierern durch das Bundeskriminalamt (BKA) zur Erkennung und Verfolgung von Hasskriminalität in sozialen Netzwerken. Dieser Artikel stellt Forschungsziele aus dem Bereich der Mensch-Technik Interaktion vor und geht hierbei insbesondere auf deren methodische Herangehensweisen und Herausforderungen, sowie erste Ergebnisse des Forschungsprojektes ein. Hierdurch werden für die Forschungsziele *Anforderungsanalyse*, *Konzeption von Nutzerschnittstellen*, und *Usabilityevaluation* Handlungsempfehlungen für zukünftige HCI-Forschung formuliert, wobei insbesondere auf den technischen Kontext der künstlichen Intelligenz und den organisationalen Kontext der Sicherheitsbehörde eingegangen wird.

MuC'22, 04.-07. September 2022, Darmstadt

https://doi.org/10.18420/muc2022-mci-ws10-230

SCHLAGWÖRTER

Künstliche Intelligenz, Entscheidungsautonomie, Behörden, Usability, Sicherheit

1 EINLEITUNG

Die Erkennung und Verfolgung von Hasskriminalität im Internet stellt für Sicherheitsbehörden aufgrund des verstärkten Aufkommens dieser Straftaten eine große Herausforderung dar. Klassifizierer, die auf künstlicher Intelligenz beruhen, könnten hier zu einer Arbeitserleichterung beitragen, indem Sie eine strafrechtliche Vorbewertung sowie eine Klassifikation der Domäne (z. B. Antisemitismus) vornehmen, und somit die strafrechtliche Bewertung in doppelter Weise unterstützen: Durch die Zuordnung in eine spezialisierte Abteilung und durch eine Entscheidungsunterstützung in der Vorbewertung der strafrechtlichen Relevanz.

Die Herausforderungen dieses Vorhabens sind zahlreich. Für eine gelungene Implementierung müssen nicht nur präzise KI-Modelle entwickelt und trainiert werden, es müssen auch rechtliche, ethische und soziale Rahmenbedingungen erforscht und in Einklang gebracht werden. In diesem interdisziplinären Gefüge kommt der HCI-Forschung eine besondere Rolle zu, da Interaktionsfragen im Umgang mit künstlicher Intelligenz wie z. B. die Beibehaltung von menschlicher Entscheidungsautonomie oder der Bedarf an Transparenz und Erklärbarkeit noch nicht hinreichend geklärt sind. In diesem Beitrag möchten wir Zugänge der HCI-Forschung am Beispiel der KI-basierten Klassifikation von Hassrede durch eine Sicherheitsbehörde vorstellen und hieraus methodische Empfehlungen für zukünftige interdisziplinäre Forschung im Systemkontext KI, aber auch Anwenderkontext Sicherheitsbehörde ableiten.

2 POTENTIALE UND HERAUSFORDERUNGEN DES KI-EINSATZES IN SICHERHEITSBEHÖRDEN

Dem Einsatz von KI wird bei der Klassifikation von Hassrede ein großes Potential zugesprochen. Die Vorstellung, dass ein ausgeleitetes Hassposting mit einer KI-Klassifikation versehen wird (z. B. „Hier liegt zu 87% eine Volksverhetzung §130 StGB vor") und beim menschlichen Anwender zu einer sofortigen Arbeitserleichterung führt, wird der Komplexität des Gesamtsystems jedoch nicht gerecht.

Auf rein technischer Ebene ist die KI-basierte Klassifikation von Hassrede alles andere als trivial. Trotz immenser Fortschritte in der inhaltlichen Klassifikation von Text [10] ist Hass in seiner Ausgestaltung stark domänenabhängig [2] und nicht nur inhaltlich, sondern auch zeitlich variabel: z. B. verwenden antisemitische Gruppen andere sprachliche Codes und Chiffren als früher, aber auch andere Codes als Einzeltäter, die Personen des öffentlichen Lebens bedrohen. Auch die Neubildung von sich radikalisierenden Gruppen (z. B. sog. „Querdenker") geht mit der Bildung eigener Codes und Chiffren einher. Zudem können Hassbotschaften über verschiedene soziale Netzwerke medial unterschiedlich ausgestaltet sein (Text, Bild, Audio, Video), was immense Implikationen für das Training von KI-Modellen mit sich bringt. Auch Mischformen der genannten Ausgestaltungen wie z. B. ASCII-Art können Hass darstellen, obwohl sie eigener Klassifikatoren bedürfen und weder von einem reinen Bild-Klassifikator noch einem reinen Text-Klassifikator erkannt werden können.

Eine weitere Herausforderung stellt der Kommunikationskontext dar, da Beiträge erst im Kontext überhaupt als Hass erkennbar sein können (z. B. „Euch allen weiterhin ein frohes Osterfest!" als Kommentar unter einem Zeitungsbericht über einen Anschlag). Je nach sozialem Netzwerk kann sich Hass somit erst aus einem Gefüge zahlreicher anderer Beiträge ergeben, die gesammelt nicht durch den Klassifikator abgedeckt werden können oder schlimmstenfalls gar nicht Teil der ausgeleiteten Datenmenge sind. Hassrede führt zudem zu Gegenmechanismen wie Counter-Speech [5] und weiterer Hassrede, die für optimale Ermittlungsvoraussetzungen Teil der Ausleitung sein sollten. Unter Berücksichtigung der aufgezählten Aspekte kommt auch dem Training der KI-Modelle eine besondere Bedeutung zu, da zunächst Trainingsdaten in großer Anzahl vorliegen und händisch annotiert werden müssen. Zudem müssen die Daten von hoher Qualität sein und den Phänomenbereich vollständig ausfüllen, um einen *sample bias* zu vermeiden, bei dem z. B. ausschließlich islamfeindlicher Hass in den Trainingsdaten vorhanden ist und zu einem Modell führt, welches antisemitischen Hass als Resultat nur unzureichend detektieren kann. Unterschiedliche soziale und politische Milieus äußern unterschiedliche leichtßu klassifizierende Hassrede, was zu weiteren Verzerrungen und Ungleichbehandlung führen kann.

Hier knüpfen sich Fragen aus rechtlich-ethisch-sozialer Perspektive an, die aus HCI-Perspektive angegangen werden können. Wie muss das System gestaltet sein, um eine „Scheinprüfung" der Beiträge auszuschließen, bei der menschliche Nutzer sich zu sehr auf das KI-Resultat verlassen und dem eigentlichen Beitrag zu wenig Beachtung schenken? Welche Performancemetriken entscheiden über einen sinnvollen Einsatz und welche Maßnahmen sind zu ergreifen, um Technologieakzeptanz durch die Nutzer sicherzustellen? Bei wem liegt die Entscheidungsverantwortung beim Einsatz semiautomatisierender Klassifikation und wie können Maßnahmen in ihrer Systemwirksamkeit evaluiert werden?

3 FORSCHUNGSZIELE DER HUMAN-COMPUTER INTERACTION

In diesem Kapitel beschreiben wir mögliche Forschungsziele und methodische Ansätze vor dem Hintergrund der skizzierten Forschungsproblematik entlang des in DIN 9241-210 [3] normierten menschzentrierten Gestaltungsprozesses.

3.1 Anforderungsanalyse

Die Erhebung und Spezifikation von Anforderungen stellt den Beginn einer jeden technischen Entwicklung dar und kann bekanntlich beliebig viel Zeit in Anspruch nehmen. Im Hinblick auf künstliche Intelligenz ist zu berücksichtigen, dass es sich bei KI um eine Technologie handelt und nicht um ein System. Diese Unterscheidung ist wichtig, da potentielle Nutzer viel Erfahrung im Umgang mit KI haben können, ohne sich dessen bewusst zu sein (z. B. Bildbearbeitung und Empfehlungssysteme). Auf der anderen Seite haben Nutzer mentale Modelle von künstlicher Intelligenz, die teilweise stark medial geprägt sind. So verweisen Nutzende mit Termini wie „Transparenz" auf andere dahinter liegende Konzepte als KI-Entwickler es tun würden, sodass hiermit nicht auf die Möglichkeit des Entwicklers verwiesen wird, interne Strukturen und Mechanismen des KI-Modells zu betrachten (vgl. [6]), sondern eher der Bedarf des Nutzers formuliert wird an einer Möglichkeit, das Ergebnis der Klassifizierung nachvollziehen zu können.

Im Kontext von Sicherheitsbehörden muss darauf geachtet werden, dass klassische Methoden zur Erhebung von Anforderungen

wie Interviews, Workshops, Benutzerbefragungen, uvm. aus mehreren Gründen nicht streng nach Lehrbuch durchgeführt werden können. Zum einen kann ein intensiver Einblick in die Arbeitsweisen von Sicherheitsbehörden wie z. B. bei contextual inquiry als zu sicherheitskritisch empfunden werden oder an hohe bürokratische Auflagen gebunden sein. Eine ähnliche Problematik besteht bei der Audioaufzeichnung von Interviews, da Gäste in Sicherheitsbehörden oft schriftlich zusichern müssen, dass im Gebäude keine Aufnahmen vorgenommen werden. Zusätzlich kann erschwerend dazu kommen, dass es noch gar keine „echten" Nutzer gibt, die für die spätere Systemnutzung vorgesehen sind, da sich entsprechende Abteilungen noch im Aufbau oder in der Planung befinden. Die Empfehlung lautet grundsätzlich, zusätzlich zu den o. g. Methoden in jedem Kontakt mit der Sicherheitsbehörde auf Anforderungen zu achten und in der Anfangsphase des Projektes entsprechend in sämtlichen Workshops um Teilnahme zu bitten. Für die Möglichkeiten der Aufzeichnung sollte direkt zu Projektbeginn genügend Zeit eingeplant werden, eine von allen Seiten abgestimmte Möglichkeit der Aufzeichnung zu erarbeiten. Weitere gut funktionierende Methoden zur Erhebung von Anforderungen sind der schriftliche Fragenkatalog und die Einholung von Stellenausschreibungen und Organigrammen. Diese öffentlich zugänglichen Dokumente benötigen jedoch einer Reflexion mit der Sicherheitsbehörde, um die darin kodifizierten Anforderungen an die Systemgestaltung zu konkretisieren (z.B. erwartbares technisches Vorwissen oder Entscheidungsprozesse). Sollte die Spezifikation des Nutzungskontextes die Anfertigung von Personas umfassen, sollte das Ziel dieser Darstellungsform aufgrund des vorhandenen Kreativanteils und der Ähnlichkeit zu Fahndungstexten gründlich erklärt werden. In eventueller Ermangelung echter Nutzer sollten Personas zunächst mit weiteren Interessensvertretern validiert und zur Projektlaufzeit re-evaluiert werden.

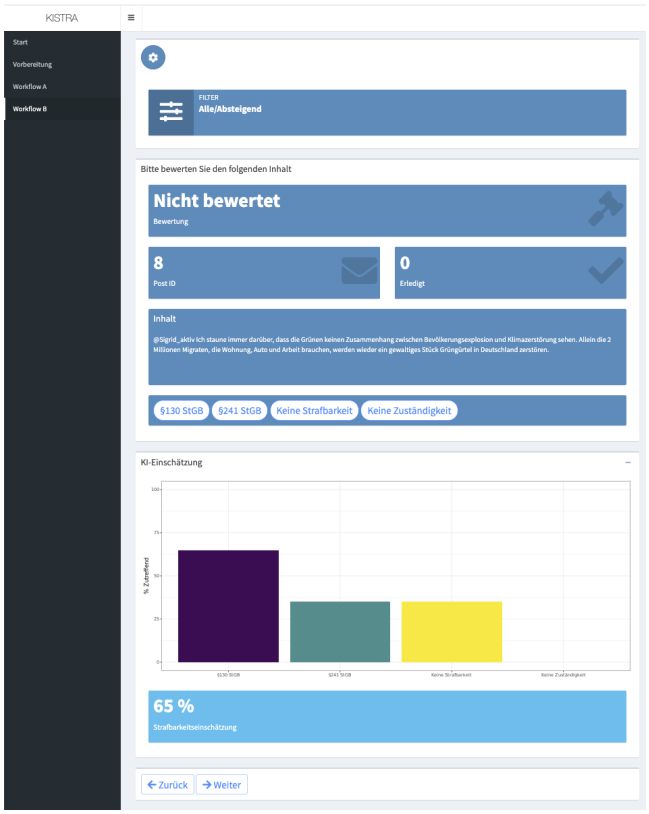

Abbildung 2: High-fidelity Prototyp der Nutzerschnittstelle eines KI-basierten Klassifikators zur Erkennung von Hassrede.

3.2 Erzeugung von Gestaltungslösungen: Konzeption von Nutzerschnittstellen

Bei der Konzeption von low- oder high-fidelity Prototypen sollte die Wizard of Oz-Methode gewählt werden, um das KI Modell als Teil des Interaktiven Systems methodisch isolieren zu können. Sorgfältig abgewogen werden sollte hierbei die Frage, ob man eine gut funktionierende, eine schlecht funktionierende oder eine absichtlich defekte Empfehlung imitiert [1], da die grundsätzliche Empfehlung so früh wie möglich zu testen einer realistischen Einschätzung der zukünftigen Performance des KI-Modells widerspricht. Das wissenschaftliche Experiment, dass den HCI-Forscher hier brennend interessiert, löst beim zukünftigen Nutzer ggfs. Vorbehalte gegen die spätere Nutzung, die auch bei hinreichender Erklärung verbleiben könnten.

Zur Klärung einiger der o. g. Forschungsfragen haben wir im Projektkontext neben einem high-fidelity Prototypen (s. Abb. 2) ein *Privacy-Utility Cockpit* entwickelt. Diese Form der Nutzerschnittstelle impliziert einen Privacy-Utility Trade-Off [8] und ermöglicht eine graduelle Verrauschung ausgeleiteter Beiträge über Slider entlang der einzelnen Datentypen zwischen den Polen Privacy und Utility. Maximale Privacy zeigt hierbei nur den Inhalt des Beitrags und blendet sämtliche weiteren Datenpartikel wie beispielsweise (s. Abb. 1) das Profilbild, den Nutzernamen, die IP-Adresse und die Reaktionen innerhalb des sozialen Netzwerks (am Beispiel Facebook: Anzahl der Likes, Herzen, Lach-, Wut- und Trauersmileys, Kommentare und Weiterleitungen), sowie die Metadaten aus. Dieser Modus kann den Ermittlungsprozess erschweren und widerspricht einigen Nutzungsanforderungen, gewährleistet dafür aber einen hohen Datenschutz und eine niedrige Informationsmenge für den Nutzer. Maximale Utility auf der anderen Seite zeigt alle diese Daten. Hierbei ist die Ermittlung maximal unterstützt, es gibt jedoch eine eventuelle Datenschutzproblematik und es ist fraglich, ob die genannten Informationen überhaupt zur Verfügung stehen.

Das Privacy-Utility Cockpit kann als Erhebungswerkzeug verwendet werden, um die Frage nach entscheidungsrelevanten Datenpartikeln zu beantworten. Es dient dazu Parameter für mögliche Anonymisierungsverfahren (z. B. differential privacy [4] oder k-Anonymität [7]) bewertbarer zu machen. Die Notwendigkeit einer ethischen und rechtlichen Betrachtung dieser Parameter entfällt hierdurch keineswegs. Es kann aber zumindest ermöglicht werden, dass keine Daten über die für die Ermittlungsarbeiten notwendigen hinaus betrachtet werden (müssen). Zusätzlich kann ein Privacy-Utility Cockpit es als Sensibilisierungs- und Schulungswerkzeug eingesetzt werden, um im Zuge der Einarbeitung einen sensiblen

Umgang mit personenbezogenen Daten unbeteiligter Dritter zu vermitteln. Zudem kann es als alternative Nutzerschnittstelle bereitgestellt werden, um Nutzern eine freiwillige und bewusste Reduktion visueller Komplexität zu ermöglichen.

3.3 Usability-Evaluation

Vor einer Usability-Evaluation müssen die in der Anforderungsanalyse erarbeiteten Indikatoren operationalisiert werden. Im KI-Kontext ist hierbei zu beachten, dass angestrebte Systemeigenschaften wie Transparenz oder Entscheidungsautonomie aus dem Forschungsstand heraus nicht abschließend definiert sind und an an den jeweiligen Kontext adaptiert werden müssen. Hierbei bietet es sich an, aus dem Forschungsstand abgeleitete Definitionen als Grundlage zu verwenden und in Nutzertests und anderen empirischen Verfahren induktiv zu erweitern.

Eine weitere Informationsquelle in Sicherheitsbehörden stellen interne Publikationen wie interne Rundschreiben oder Leitlinien dar, die auch zur Oprationalisierung von Evaluationskriterien hilfreich sind. Ein weiterer pragmatischer Zugang zur Entwicklung einer Evaluationsskala stellt die Entwicklung von reflektiven Selbstauskunftsskalen dar, die auf den Äußerungen der Probanden in qualitativen Vorstudien oder ersten Nutzertests basieren, sowie die grundsätzliche Empfehlung, in ersten Schritten auf qualitative Methoden wie z. B. *think-aloud* zurückzugreifen.

Unsere ersten Erfahrungen mit Evaluationsstudien zeigen, dass es KI-Klassifikatoren eine sehr spezifischen Einarbeitung verlangen, um *overtrust* oder *undertrust* [9] zu vermeiden, der aus Zuschreibungen zur grundsätzlichen KI-Funktionsweise resultiert: Nutzer unterstellen in unserem Falle der KI z. B. tiefergehende Kompetenzen und vermuten Kontextinformationen, die nur der KI vorlagen, nicht aber dem Nutzer. Hier wird unter anderem der Bedarf deutlich, die Interpretation und das Verständnis von *confusion matrices* bzw. *false positives* und *false negatives* nutzergerecht zu vermitteln und deren Implikation für den Ermittlungsalltag zu verstehen.

Gleichzeitig konnten wir beobachten, dass die Präsentation einer KI-Klassifikation zu keiner direkten Zeitersparnis führt, da die Menge der dargestellten Information steigt. Eine Problematik in der Entscheidungsautonomie konnten wir zunächst nicht beobachteten, da Probanden von sich aus zuerst eine eigene Einschätzung vornahmen bevor Sie das KI-Ergebnis betrachteten. Die KI diente meistens als kurze Rückversicherung für die eigene Entscheidung. Im Falle eines von der Nutzereinschätzung abweichenden Resultats konnten wir beobachten, dass ein aus Nutzersicht „false positive", bei dem die KI eine strafrechtliche Relevanz angab, die der Nutzer nicht nachvollziehen konnte, zu wesentlich mehr Zeitverlust und Verwirrung führte als ein aus Nutzersicht „false negative", bei dem der Nutzer eine strafrechtliche Relevanz erkannte, die KI jedoch nicht. Im erstgenannten Fall neigten Nutzer dazu, den Grund für die vermeintlich falsche KI-Einschätzung zu suchen, was jedoch bei längeren und sprachlich inkohärenten Beiträgen (was möglicherweise das false positive verursacht hatte) zu erheblichem Aufwand führte. Im letzteren Fall wurde im Gegensatz dazu nicht nach Gründen gesucht, warum die KI einen vermeintlich strafbaren Beitrag nicht als solchen erkannt hat. Implizit deutet das darauf hin, dass Nutzende eine sehr hohe Spezifität erwarten, jedoch nicht notwendigerweise eeine hohe Sensitivität.

Die Bewertung der Nützlichkeit des Gesamtsystems war trotz zahlreicher Abweichungen zwischen Wizard-of-Oz-KI und der menschlichen Bewertung sehr hoch, sodass wir davon ausgehen, dass die Performance des KI-Klassifizierers nicht den ausschlaggebenden Faktor für die Akzeptanz des Systems darstellt. Probanden lobten z. B. die Usability der Nutzerschnittstelle und empfanden nicht vorhandene Kontextinformationen als größeres Problem als eine KI-Einschätzung, die vermeintliche false positives oder negatives enthielt.

4 FAZIT

Für zukünftige HCI-Forschung mit Sicherheitsbehörden sollte insbesondere berücksichtigt werden, dass ein Einbezug echter Nutzer in frühe Iterationszyklen nicht immer möglich ist. Die Empfehlung lautet daher, auch andere Erhebungsverfahren wie schriftliche Fragenkataloge oder Stellenausschreibungen als Basis zur Spezifikation des Nutzungskontextes zu nutzen und bei einer ersten Validierung auf Interessensvertreter zurückzugreifen. Zudem sollte direkt zu Projektbeginn abgeklärt werden, dass Interviews und Nutzertests für Projektzwecke aufgezeichnet werden dürfen. Im Kontext von KI-Textklassifikation sollte zudem kontextabhängig das Verhältnis zwischen KI-Performance und anderen Nutzungsanforderungen geklärt werden, da auch ein KI-Modell mit schlechter Performance in ein insgesamt akzeptiertes Gesamtsystem integriert sein kann, sofern das System zur effektiven, effizienten und zufriedenstellenden Erreichung der Nutzerziele geeignet ist.

5 FÖRDERHINWEIS

Das diesem Bericht zugrunde liegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 13N15341 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autoren.

LITERATUR

[1] Philipp Brauner, Ralf Philipsen, André Calero Valdez, and Martina Ziefle. 2019. What happens when decision support systems fail?—The importance of usability on performance in erroneous systems. *Behaviour & Information Technology* 38, 12 (2019), 1225–1242.

[2] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 11. 512–515.

[3] ENISO DIN. 2011. Ergonomie der Mensch-System-Interaktion-Teil 210: Prozess zur Gestaltung gebrauchstauglicher interaktiver Systeme.

[4] Cynthia Dwork. 2008. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*. Springer, 1–19.

[5] Lilian Kojan, Hava Melike Osmanbeyoglu, Laura Burbach, Martina Ziefle, and André Calero Valdez. 2020. Defend your enemy. A qualitative study on defending political opponents against hate speech online. In *Multidisciplinary International Symposium on Disinformation in Open Online Media*. Springer, 80–94.

[6] Zachary C Lipton. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[7] Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems* 10, 05 (2002), 557–570.

[8] André Calero Valdez and Martina Ziefle. 2019. The users' perspective on the privacy-utility trade-offs in health recommender systems. *International Journal of Human-Computer Studies* 121 (2019), 108–121.

[9] Christopher D Wickens, William S Helton, Justin G Hollands, and Simon Banbury. 2021. *Engineering psychology and human performance*. Routledge.

[10] Frederike Zufall, Marius Hamacher, Katharina Kloppenborg, and Torsten Zesch. 2020. A Legal Approach to Hate Speech: Operationalizing the EU's Legal Framework against the Expression of Hatred as an NLP Task. *arXiv preprint arXiv:2004.03422* (2020).