KI-basierte Erkennung beleidigender Äußerungen – Offensive Speech Detection in Social Media

Executive Summary

Offensive Speech, also beleidigende Äußerungen, im Extremfall die Hassrede (Hate Speech), gehören zu den größten Herausforderungen digitaler Kommunikation, insbesondere auf Online-Plattformen, auf denen Inhalt geteilt wird.[2] Die Hassrede grenzt sich von Offensive Speech dahingehend ab, als dass sie auf ganz bestimmte Personen oder Gruppen abzielt, um diese aufgrund spezieller Merkmale wie etwa Herkunft, Religion, Geschlecht, sexueller Orientierung u.a. zu diskriminieren und zu erniedrigen. Ausdruck finden Offensive Speech und Hassrede in Texten, Bildern, Symbolen oder Gesten. Ihr massenhaftes Auftreten sowie das Spannungsfeld zwischen der Verletzung von Menschenrechten und einer Beschränkung des Rechts auf freie Meinungsäußerung machen das Thema ethisch und gesellschaftlich sehr herausfordernd. Aufgrund der großen Menge an problematischen Inhalten und der in Europa bestehenden gesetzlichen Verpflichtung, gegen Hass und Hetze vorzugehen, setzen Online-Plattformen vermehrt KI-Systeme zur automatisierten Erkennung und Entfernung von Offensive Speech und Hassrede ein.[1,5,10,20]

Aus technischer Sicht ist eine klare und faire Bewertung von Sprache durch KI-Systeme herausfordernd, weil die Beurteilung beleidigender Äußerungen im jeweiligen Kontext stattfinden muss und oft selbst für Menschen herausfordernd ist. KI-Systeme haben zudem Schwierigkeiten, Ironie und Sarkasmus zu erkennen. Somit besteht das Risiko eines ungenügenden Schutzes vor problematischen und illegalen Äußerungen. Umgekehrt kann es auch zu einem ungerechtfertigten Filtern und Streichen von Inhalten kommen, was aus Sicht von Nutzer:innen das Recht auf freie Meinungsäußerung beschneiden kann.[10]

Im ALAIT Risikoradar wird der Einsatz von KI-Systemen zur Erkennung und Entfernung beleidigender Äußerungen und Hassrede mit einem Gesamtrisiko von hoch bis mittel (dunkelorange und hellorange) eingestuft. Was den Autonomiegrad des KI-Systems betrifft, weisen die eingesetzten KI-Systeme in der Regel einen mittleren bis hohen Autonomiegrad (Stufe 3 und 4), da nur in Problemfällen, etwa bei Beschwerden von Nutzer:innen oder bei unklaren Befunden, eine Prüfung durch Menschen erfolgt. Das Risiko im Anwendungskontext wird als „mittel" (Stufe 3) eingeordnet, da zwar meist keine unmittelbare Gefahr für Leib und Leben besteht, jedoch fehlerhaft oder ungenau arbeitende KI-Systeme erhebliche Auswirkungen auf die Grundrechte der Menschen wie Schutz vor Diskriminierung, persönliche Sicherheit, körperliche Unversehrtheit oder Meinungsfreiheit haben können.

Bürger:innen können in Österreich gegen digitale Hassrede, die sie beleidigt, verletzt, oder aus anderen Gründen illegal ist, vorgehen und Anzeige erstatten. Der europäische Digital Services Act (DSA) verpflichtet zudem Plattformbetreiber dazu, aktiv gegen Hassrede vorzugehen

ALAIT Risikoradar für KI-basierte Erkennung beleidigender Äußerungen

KI-Anwendungen im Bereich Offensive Speech Detection

Das ALAIT Risikoradar ist ein wissenschaftlich entwickeltes Risikoanalysetool für Künstliche Intelligenz (KI), das KI-Anwendungen kontextbezogen und unter Berücksichtigung ihres technischen Autonomiegrades einstuft und so die Risikosphäre für Anwender:innen auf einen Blick sichtbar macht. Dabei gilt: Je höher das Einsatz-Risiko aus dem Anwendungskontext und je größer der Autonomiegrad des KI-Systems in Bezug auf Entscheidungen, desto riskanter ist der Einsatz einzustufen. Eine erweiterte Klammer weist auf eine Bandbreite in der Risikoeinstufung hin. Ein geringer Autonomiegrad eines KI-Systems bedeutet nicht, dass man sich zurücklehnen kann. Es erfordert eine starke Rolle der Menschen, die es anwenden. (Details zum Stufenmodell s. S. 8f).

**ALAIT ist ein vom Bundesministerium für Innovation, Mobilität und Infrastruktur (BMIMI) initiiertes Forschungsprojekt zur Entwicklung einer neuartigen Form von Technologiebewertung für KI-Technologien mit wissenschaftlichen Methoden.*

und marginalisierte Gruppen zu schützen. Etwa ist sicherzustellen, dass Content-Moderator:innen die Funktionsweise und den Output der Systeme überwachen. Mitarbeitende im Bereich der Content-Moderation sind kontinuierlich zu schulen – sowohl hinsichtlich der Funktionsweise und Grenzen von KI-Systemen als auch in Bezug auf gesellschaftliche Sensibilitäten. Im Januar 2025 wurde der DSA durch den „Code of Conduct on Countering Illegal Hate Speech Online" erweitert, der über freiwillige Selbstverpflichtungen hinaus konkrete Handlungsvorgaben für Plattform-Betreiber:innen schafft.

Einleitung

Auf Online-Plattformen, auf denen Nutzer:innen Content wie Texte, Bilder oder Videos teilen, stellen Offensive Speech (beleidigende Äußerungen) und Hate Speech (Hassrede) besondere Herausforderungen dar. Im Unterschied zu Offensive Speech beziehen sich die Beleidigungen bei Hassrede auf eine bestimmte Gruppenzugehörigkeit. Studien[1] zeigen, dass von Hassrede im Internet besonders Frauen, Muslime, Juden oder Menschen of Color betroffen sind. Offensive Speech und Hassrede können zu psychischer Belastung, sozialer Ausgrenzung und im Extremfall sogar zu physischer Gewalt führen.[2] Sie schränken die freie Meinungsäußerung durch Angst vor Angriffen sowie durch unfaire Filterung oder Zensur in sozialen Medien ein. Dadurch wird gesellschaftliche Polarisierung verstärkt und diskriminierende Einstellungen werden normalisiert, sodass diese zunehmend als akzeptabel erscheinen. In einigen gesellschaftlichen Gruppen kann dies Radikalisierungsprozesse fördern, weil feindselige Narrative als legitim erlebt und so weiter verstärkt werden. Hassrede kann sowohl in Form von Sprache, als auch in Form von Bildern, Videos, Karikaturen, Emojis und Zeichen dargestellt werden.

Durch *generative KI-Technologien* steigen die Möglichkeiten zur Erzeugung und Verbreitung von offensive Speech und Hassrede in sozialen Medien, etwa in Form von sogenannten Deepfakes: Realistisch erscheinende Medieninhalte wie Bilder, Videos oder Audiodateien, die als echt wahrgenommen werden, aber tatsächlich mit Hilfe von KI erstellt oder bearbeitet wurden und reale oder fiktive Personen in Situationen oder mit Aussagen zeigen, die erfunden sind.[1,3]

Da sowohl die freie Meinungsäußerung als auch die Sicherheit der Nutzer:innen in Einklang gebracht werden müssen, ist die effektive Erkennung und Verhinderung beleidigender Äußerungen und Hassrede unerlässlich. Automatisierte Inhaltsmoderation, welche eine große Menge an nutzergenerierten Inhalten auf schädliche und illegale Inhalte durchforstet und diese auch gegebenenfalls entfernt, ist in der Praxis unverzichtbar. Die zum Einsatz kommenden KI-Systeme sollen drei Gruppen zugutekommen:

1. Den Betreiber:innen von Online-Plattformen, die damit die Menge der Inhaltsmoderation stark erhöhen können.
2. Den Content-Moderator:innen, die durch Automatisierung von psychisch belastenden Aufgaben geschützt werden können.
3. Den Nutzer:innen sozialer Medien, die diese im Idealfall ohne anstößige, illegale oder gewalttätige Inhalte erleben können.

Diesen drei Gruppen stehen jedoch ebenso drei Gruppen gegenüber, die durch die beschriebene Automatisierung benachteiligt sind:

1. Die Data-Labeler, die oft im globalen Süden unter prekären Umständen psychisch belastende Daten labeln müssen, die als Trainingsdaten für diese KI Systeme dienen.

2. Benachteiligte Gruppen, deren Inhalte durch Verzerrungen in der KI-gestützten Inhaltsmoderation öfters fälschlicherweise als Hassrede eingestuft, und daher häufiger zensiert werden.

3. Die Demokratie insgesamt mit ihren demokratisch legitimierten Institutionen. Denn welche Inhalte als „gut" oder „schlecht" qualifiziert werden sowie was gesagt und was nicht gesagt werden darf, entscheiden die von Plattform- und Foren-Betreiber:innen eingesetzten KI-Systeme, welche üblicherweise kaum transparent gemacht werden und sich damit einer demokratischen Kontrolle entziehen.

Technologiebeschreibung

Die KI-basierte Erkennung von beleidigenden Äußerungen stützt sich auf Fortschritte im maschinellen Lernen (ML), insbesondere im Bereich der Verarbeitung natürlicher Sprache (Natural Language Processing – NLP), die sich auf Textinhalte bezieht. Im Kern umfasst diese Technologie das Training von Algorithmen anhand großer Datensätze mit gekennzeichneten Inhalten (*überwachtes Lernen*), sodass sie lernen, neue Inhalte automatisch als „zulässig" oder „verstoßend" gegen eine bestimmte Richtlinie (z.B. beleidigende Äußerungen vs. normale Sprache) zu klassifizieren.[6] Solche Systeme können offensichtlich hasserfüllte Wörter zwar markieren, haben jedoch Schwierigkeiten mit der korrekten Interpretation von Text etwa den Kontext betreffend, weiters mit Rechtschreibfehlern, Sarkasmus und Ironie sowie harmlos klingenden Codewörtern für spezifisch feindselige Bedeutungen. Generell funktionieren sie aufgrund der Vielzahl englischer Trainingsdateien in englischer Sprache besser als in anderen Sprachen.
Es gibt verschiedene Techniken, die dabei helfen sollen, diese Herausforderungen abzuschwächen. Unter anderem werden die folgenden beiden Techniken eingesetzt:

1. Named Entity Recognition (NER) und Entity Linking (EL) werden verwendet, um Identitätsgruppen in Texten zu kennzeichnen, genau zu bestimmen, welche Wörter sich auf die Zielgruppe beziehen, und sie mit einer externen Wissensdatenbank zu verknüpfen.[26,27] Diese Techniken sollen dabei helfen, zu identifizieren, worum es in dem Beitrag geht und wer/was angegriffen wird.
2. Sentiment- und Emotionsanalysen werden verwendet, um echte Hassäußerungen von anderen Formen der Verwendung beleidigender Sprache zu unterscheiden.[28,29]

Fortschritte in der Leistungsfähigkeit werden insbesondere mit dem Einsatz von Transformer-Modellen (z.B. BERT, RoBERTa, XLM, GPT) erzielt, die kontextuelle Beziehungen zwischen Wörtern, z.B. in Sätzen, erkennen können und sich deshalb für Sprachanalyse eignen.[7,8]

Doch diese Fortschritte beziehen sich hauptsächlich auf die englische Sprache und betreffen weniger andere Sprachen wie etwa Deutsch oder Italienisch. Zudem weisen diese Modelle auch Verzerrungen auf, die oft besonders marginalisierte Gruppen noch weiter benachteiligen.

In der Praxis kommen überwiegend sprachbasierte Systeme zum Einsatz. Da beleidigende Äußerungen jedoch über den rein textuellen Inhalt hinausgehen, bedarf es multimodaler Systeme, die auch Bilder oder Videos analysieren können und für die Computer Vision-Technologien benötigt werden.[9] Die Entwicklung solcher multimodalen Systeme birgt allerdings große Herausforderungen in Bezug auf Fairness, Transparenz und Kontextverständnis.

Chancen

Der Einsatz von KI zur Erkennung von beleidigenden Äußerungen auf Online-Plattformen hat das Potenzial, sowohl für Nutzer:innen als auch für Content-Moderator:innen und Anbieter:innen sozialer Medien von Vorteil zu sein:[6,8,10]

Chancen für Nutzer:innen:

- **Gesteigertes User-Erlebnis in Sozialen Medien:** Die Integration von KI in die Moderation von Inhalten – auch unabhängig von der Erkennung beleidigender Äußerungen – kann dazu beitragen, eine personalisierte Erfahrung zu gewährleisten und die Menge an Spam, Falschinformationen sowie verstörenden bis hin zu illegalen Inhalten zu reduzieren. In Verbindung mit der im Digital Services Act (DSA) verankerten Möglichkeit, die Kriterien, nach denen Content moderiert wird, selbst auszusuchen, kann das Gesamterlebnis verbessert und das Risiko, dass Nutzer:innen mit hasserfüllten und schädlichen Inhalten konfrontiert werden, (idealerweise) reduziert werden.[3,12]

Chancen für Content-Moderator:innen:

- **Geringere psychische Belastung und Arbeitsmenge:** KI-Tools können sensible Bilder unkenntlich machen und beleidigende Sprache in Echtzeit zensieren. Dadurch sinkt die Menge an Material, das an Content-Moderator:innen gelangt bzw. diese werden mitunter vor den schlimmsten Inhalten geschützt. Das reduziert die psychische Belastung, die mit dieser Arbeit verbunden ist.[3,13]

Chancen für Betreiber:innen von Online-Plattformen:

- **Geschwindigkeit und Umfang:** Laut eigenen Angaben werden beispielsweise auf der Plattform X (ehemals Twitter) täglich über 500 Millionen Beiträge generiert.[11] Eine effiziente Moderation dieser Inhalte ist nur durch die Integration von KI-Systemen zu bewerkstelligen, die automatisiert einen Großteil der offensichtlich illegalen und beleidigenden Inhalte innerhalb von Minuten nach ihrer Veröffentlichung erkennen und entfernen können.[10]
- **Konsistenz und Genauigkeit:** KI-Systeme können, sofern sie richtig trainiert und eingestellt wurden, eine hohe Konsistenz auf alle analysierten Inhalte anwenden. Die Forschung zeigt, dass Transformer-Modelle wie RoBERTa und spezialisierte Varianten wie HateBERT (ein BERT-Modell, das auf hasserfüllte bzw. beleidigende Sprache vorab trainiert wurde) bei der Erkennung von Hassreden gute Leistungen erzielen können.[12]

Herausforderungen und Risiken

Trotz der oben genannten Chancen gibt es eine Reihe von Risiken und Herausforderungen, die die Anbieter:innen von Online-Plattformen lösen müssen, um eine vertrauenswürdige und zuverlässige Integration von KI zur Erkennung von beleidigenden Äußerungen zu gewährleisten:

1. **Viele falsch-positive und falsch-negative Ergebnisse sowie** ***Verzerrungen (KI-Bias)***: Obwohl KI-Systeme darauf ausgelegt sind, beleidigende Äußerungen zu erkennen und zu löschen, machen sie häufig Fehler. Insbesondere haben KI-Systeme Schwierigkeiten bei der Erkennung von Ironie und Sarkasmus.[8,15] Bei multimodalen Systemen liegt das Problem darin, dass Bilder keinen klaren semantischen Kontext wie Texte haben, sondern sich auf komplexe visuelle Hinweise, Symbolik und kulturelles Verständnis stützen, die die Systeme nicht erkennen können. Fehler eines Klassifizierungssystems lassen sich als „falsch Positive" (wenn ein harmloser Beitrag fälschlicherweise als Hassrede identifiziert wird) oder als „falsch Negative" (wenn tatsächliche Hassrede vom KI-System nicht erkannt wird) kategorisieren. Solche Fehler treten oft bei marginalisierten oder vulnerablen Gruppen verstärkt auf, und führen zur weiteren Benachteiligung dieser Gruppen. Studien haben gezeigt, dass automatisierte Hassrede-Detektoren Inhalte, die in afroamerikanischem Englisch (AAE) oder Minderheitendialekten verfasst sind, oft fälschlicherweise als „toxisch" kennzeichnen.[8,17] Dies kann zur Verschärfung von Vorurteilen bis hin zur Zensur solcher Gruppen führen. Umgekehrt kann bei manchen Gruppen eine Anhäufung von falsch negativen Ergebnissen dazu führen, dass sie weiterhin viel Hass und Beleidigung ausgesetzt sind, was zu psychischen Belastungen und zum Rückzug aus Online-Foren führen kann.

2. **Einschränkung der Meinungsfreiheit:** Artikel 19 des Internationalen Pakts über bürgerliche und politische Rechte (ICCPR) verankert den Schutz der Meinungsfreiheit, während Artikel 20(2) jede Aufstachelung zu nationalem, rassistischem oder religiösem Hass verbietet, die zu Diskriminierung, Feindseligkeit oder Gewalt führt.[1] Die Filterung und Streichung von Inhalten steht daher direkt im Konflikt mit dem Recht auf Meinungsfreiheit. KI-Moderationsinstrumente können von Regierungen oder Anbieter:innen von sozialen Medien missbraucht werden, um damit grundlegende individuelle Freiheiten erheblich einzuschränken. In sozialen Medien werden mit Hilfe von KI-Systemen nachweislich – aufgrund falsch positiver Klassifikation –auch harmlose Inhalte aus sozialen Medien und damit die Meinungsfreiheit ihrer Nutzer:innen eingeschränkt.[10]

3. **Mangelndes Kontextverständnis:** Eine der größten Schwierigkeiten bei der automatisierten Erkennung von beleidigenden Äußerungen und Hassrede ist das mangelnde Kontextverständnis von KI.[14] Ausdrucksweisen wie Sarkasmus, Metaphern, Ironie oder verschlüsselte Sprache sind für KI-Systeme generell schwierig erkennbar.[15] Verschlüsselte Sprache ist die Nutzung von Codewörtern für ethnische Beleidigungen, die in einer bestimmten Community eindeutig verstanden werden, für Außenstehende aber nicht zu erkennen sind. Hasserfüllte Kommentare können so einer Filterung entgehen.[16] Die Performance eines KI-Modells bezieht sich immer auf Testdaten, mit einem fixen Anteil an Sarkasmus, Ironie, oder verschlüsselter Sprache. In der Praxis aber wissen viele Akteur:innen, wie sie mit Sarkasmus und verschlüsselten Aussagen die KI-gestützte Moderation umgehen können. Somit können KI-Systeme zur Erkennung beleidigender Äußerungen nur bedingt gewährleisten, dass Hassrede erkannt wird.

4. **Mangelnde *Transparenz:*** Die Entscheidungen von KI-Systemen sind in der Regel – sowohl aufgrund der Architektur, der enormen Datenmengen und der Komplexität der Systeme, als auch aufgrund der Geheimhaltung proprietärer Daten und Software – undurchsichtig und nicht nachvollziehbar. Dadurch wird der Überprüfungsprozess arbeits- und zeitaufwändig oder schlicht unmöglich (*Black-Box-Problem*).[18] Das steht im Widerspruch zum Recht von Nutzer:innen gemäß Digital Service Act (DSA) Artikel 17. Auch die Datenschutz-Grundverordnung (DSGVO) erfordert eine Erklärung für das Löschen ihres Beitrags.[6,19]

5. **Verlagerung der psychischen Belastung:** Ein bedeutendes Problem stellt die Verschiebung psychisch belastender Arbeiten in den Globalen Süden dar. Studien zeigen, dass sogenannte Datenarbeiter:innen in Ländern mit geringeren Regulierungen und Arbeitsrechten häufig Aufgaben wie Datenannotation, Inhaltsmoderation oder Labeln übernehmen – oft unter prekären Bedingungen, in zumeist informellen Arbeitsverhältnissen. Dabei müssen sie oft verstörende Bilder und Texte labeln, um Trainingsdaten für die KI-Systeme zur automatischen Erkennung von Hassrede herzustellen. Sie werden schlecht bezahlt und haben keinen angemessenen Schutz vor psychischer Belastung.[32,33,34]

6. **Allgemeine Unsicherheit in Bezug auf beleidigende Äußerungen:** Die meisten Social-Media-Anbieter:innen legen ihre Nutzungsbedingungen sowie die Definitionen für beleidigende Äußerungen selbst fest. Meta hat etwa eine dafür zuständige Abteilung eingerichtet, die sich mit dem Thema beschäftigt.[4,10] Dadurch verlagert sich die Entscheidung, was als akzeptable Sprache/Ansprache gilt, in private Unternehmen, die keinen demokratischen Prozessen unterliegen. Proprietäre, intransparente KI-Modelle setzen damit den Gold-Standard, was erlaubt ist, und was nicht. Beleidigende Äußerungen werden jedoch subjektiv unterschiedlich wahrgenommen und eine genaue, vollständige Auflistung möglicher Hassreden ist daher nicht möglich, da sie sich kombinieren und erfinden lassen. Es sollte daher in der Verantwortung der Gesetzgeber liegen, mit demokratischen Mitteln diesen Wertekonflikt zwischen Meinungsfreiheit auf der einen Seite und dem Schutz der Nutzer:innen vor Hassrede und Gewalt auf der anderen Seite angemessen zu adressieren. Eine klare Definition von Hassrede kann zwar als Grundlage für das Training von KI-Systemen herangezogen werden, benötigt aber jedenfalls Kontextualisierung bei deren Anwendung auf spezifische Bedingungen. Hier könnten z. B. Open-Source-Initiativen eine demokratische Alternative bilden, indem gelabelte Daten von betroffenen Communities für das Training von Open-Source-KI-Modellen zur Erkennung von Hassrede bereitgestellt werden – und sich Nutzer:innen ihr eigenes KI-Modell aussuchen können.

Empfehlung zum Praxiseinsatz

Um die oben genannten Risiken im Praxiseinsatz von KI gestützter Offensive Speech Detection zu minimieren, ist eine Kombination aus regulatorischen Maßnahmen, technischen Vorkehrungen und bewährten Moderationsverfahren erforderlich.

1. **Menschliche Aufsicht einrichten:** Einer der wirksamsten Grundsätze zur Risikominderung ist sicherzustellen, dass KI keine endgültigen Entscheidungen über Inhalte trifft, welche die Grundrechte der Nutzer:innen, beispielsweise das Recht auf freie Meinungsäußerung, beeinträchtigt. KI-Systeme sollten daher zumindest nach dem „Human-on-the-Loop"-Prinzip entwickelt und eingesetzt werden, bei dem KI und menschliches Fachwissen kombiniert werden, z. B. im Rahmen von stichprobenartigem Testen von KI-Entscheidungen durch Moderator:innen. Im Rahmen dieser Strategie sollen Moderator:innen auch geschult werden, Grenzen von KI-Tools und rechtliche Standards für Hassrede zu kennen und nicht blind den Warnmeldungen der KI zu vertrauen. Darüber hinaus kann die Einbeziehung von Moderator:innen oder Berater:innen aus marginalisierten und von Hassreden häufig betroffenen Gruppen das Verständnis für den Kontext erleichtern.[1] In Kombination mit *erklärbarer KI (XAI)* schafft diese Vorgangsweise mehr Transparenz und Effizienz für Moderator:innen.[7,10,23,26]

2. **Kulturelle und sprachliche Kontexte beachten:** Aktuell nimmt Content-Moderation nur wenig auf kulturelle und sprachliche Kontexte Rücksicht, obwohl das Voraussetzung ist, um problematische Inhalte erkennen zu können. KI-Systeme sollten in stärkerem Ausmaß für Kulturräume außerhalb des angloamerikanischen Sprachraums entwickelt und angewendet werden.

3. **Verbesserte Datensatzqualität/kontinuierlich lernende Datensätze sicherstellen:** Die Sprache auf Online-Plattformen entwickelt sich rasant weiter, wobei ständig neue Slangausdrücke und nuancierte Ver-

wendungsweisen entstehen. Kontinuierliche Lernframeworks, bei denen Modelle schrittweise neu gekennzeichnete Daten (einschließlich Kontext) integrieren, stellen sicher, dass die Erkennungsalgorithmen auf dem neuesten Stand bleiben. Um *Verzerrungen (KI-Bias)* zu vermeiden, sollten Unternehmen außerdem sicherstellen, dass ihre Trainingsdatensätze repräsentativ und sorgfältig annotiert sind und mehr Beispiele für nicht hasserfüllte Inhalte aus Minderheitensprachen und -dialekten enthalten. So lernt das Modell z. B., wann es sich um eine aggressiv verwendete rassistische Beleidigung handelt. Bei der Erkennung von Kontext mit NLP-Modellen wie HateBERT wurden erhebliche Fortschritte erzielt, sodass diese Modelle eine gute Grundlage für die Weiterentwicklung von KI-Systemen bilden. Dadurch wurde die Gesamtgenauigkeit verbessert und systematische Vorurteile gegenüber geschützten Gruppen verringert.[8,21]

4. **Open-Source-Modelle nutzen:** Die Verwendung von ausgewogenen und transparenten Open-Source-Modellen verbessert nicht nur die Gesamtgenauigkeit und verringert systematische Vorurteile gegenüber geschützten Gruppen sondern hilft auch, das Vertrauen der Nutzer:innen zu stärken.[8,2]

5. **Regelmäßige Audits veröffentlichen und Ergebnisse sinnvoll nutzen:** Gemäß den Artikeln 34 und 37 des Digital Service Acts (DSA) sind Plattformen dazu verpflichtet, systemische Risiken wie die Verbreitung illegaler Inhalte, negative Auswirkungen auf Grundrechte sowie potenzielle Schäden für den zivilgesellschaftlichen Diskurs, Wahlprozesse, die öffentliche Sicherheit, geschlechtsspezifische Gewalt, die öffentliche Gesundheit und Minderjährige zu ermitteln, zu analysieren und zu bewerten.[30] Ebenso sind Plattformen dazu verpflichtet, die zum Training benutzten Datensätze offenzulegen sowie die KI-Modelle, damit die Performance unabhängig getestet werden kann. Diese Audits sind zur Erhöhung der Rechenschaftspflicht zu veröffentlichen und müssen zu konkreten Maßnahmen und Anpassung der Modelle führen.[20]

6. **Multimodale Analyse zur Identifizierung von Deepfakes entwickeln:** Im Kontext von Hassrede erfordert die Erkennung von *Deepfakes* die Entwicklung neuer Technologien, die neben der Textanalyse auch die Bildanalyse einschließt. Organisationen wie die „Coalition for Content Provenance and Authenticity" (C2PA) und Praktiken wie „Open Source Intelligence" (OSINT) setzen dazu bereits erste Schritte. Sie bekämpfen die Verbreitung irreführender Informationen im Internet, indem sie technische Standards zur Zertifizierung der Quelle und Geschichte von Medieninhalten schaffen.[24,25]

7. **Beschwerdemöglichkeiten für Nutzer:innen einrichten:** Plattform-Anbieter:innen müssen Nutzer:innen gemäß Digital Services Act (DSA) die Möglichkeit geben, eine Überprüfung zu beantragen, wenn diese mit der Entscheidung des Unternehmens, kontroverse Inhalte zu löschen oder zu belassen, nicht einverstanden sind und alle Rechtsmittel ausgeschöpft haben.[31] In Deutschland gibt es dazu auch einen Rechtsrahmen. Das Netzwerkdurchsetzungsgesetz (NetzDG) verpflichtet seit 2017 soziale Netzwerke, angemessen auf Beschwerden von Nutzer:innen zu reagieren.[10]

Wichtige Begriffe

Black-Box-Systeme: Bezeichnet alle Systeme, deren interne Entscheidungsprozesse für Menschen nicht transparent und nachvollziehbar sind. Das heißt, dass nur Eingaben und Ausgaben beobachtet werden können, ohne zu verstehen, wie die Verarbeitung dazwischen genau abläuft.

Datenschutz in der KI: Die Gesamtheit der Praktiken und Bedenken im Zusammenhang mit der ethischen Erfassung, Speicherung und Nutzung personenbezogener Daten durch Systeme der künstlichen Intelligenz.

Deepfakes: Sind realistisch wirkende Medieninhalte (Fotos, Audiodateien oder Videos), die durch Techniken der KI verändert, erstellt bzw. verfälscht wurden.

Erklärbare KI (XAI): Bezieht sich auf Methoden und Techniken, die es ermöglichen, die Entscheidungen und Ergebnisse von KI-Systemen besser zu verstehen und nachzuvollziehen.

Generative KI: Künstliche Intelligenz, die neue Inhalte wie Texte, Bilder, Musik oder Code erzeugen kann, basierend auf erlernten Mustern aus Trainingsdaten.

Gesetz der Europäischen Union über künstliche Intelligenz (EU AI-Act): Gesetz der Europäischen Union über künstliche Intelligenz (EU AI-Act): Eine europäische Verordnung über künstliche Intelligenz (KI) – die erste umfassende Verordnung über KI von einer großen Regulierungsbehörde. Sie konzentriert sich insbesondere auf KI-Systeme mit hohem Risiko.

KI-Autonomie: Die Fähigkeit eines KI-Systems, eine Reihe von Zielen unter einer Reihe von Unsicherheiten in ihrer Umgebung selbstständig und ohne externe Eingriffe zu erreichen.

KI-Genauigkeit (AI-Accuracy): bezieht sich auf die Fähigkeit eines KI-Systems, korrekte Vorhersagen oder Entscheidungen zu treffen. Sie ist ein wichtiger Maßstab für ihre Leistung und entscheidend für die Bestimmung ihrer Wirksamkeit und Zuverlässigkeit.

Systematische Verzerrungen (KI-Bias): Bias ist eine systematisch unterschiedliche Behandlung bestimmter Objekte, Personen oder Gruppen im Vergleich zu anderen. Behandlung ist jede Art von Handlung, einschließlich Wahrnehmung, Beobachtung, Darstellung, Vorhersage oder Entscheidung.

Transparenz: Bedeutet, dass die Funktionsweise, Entscheidungsprozesse und Einsatzbereiche eines KI-Systems nachvollziehbar, erklärbar und offen zugänglich sind – für Entwickler:innen, Nutzer:innen und andere Stakeholder.

Überwachtes Lernen (supervised learning): Ein Teilgebiet des maschinellen Lernens, bei der die Trainingsdaten vorab gekennzeichnet werden und das System das Muster zwischen dem Inhalt und der Kennzeichnung erlernt. Die Aufgabe eines solchen KI-Systems besteht darin, eine Beziehung zu finden, die jede Eingabe des Trainingssatzes (die Daten) einer Ausgabe (der Kennzeichnung) zuordnet.

Erklärung Stufenmodell des ALAIT Risikoradars

Im ALAIT KI-Risikoradar wird die Beziehung zwischen Anwendungsrisiko und Autonomie eines KI-Systems dargestellt. Die Risikostufen stützen sich auf das EU KI-Gesetz (EU AI-Act), insbesondere auf Artikel 6 und Annex III, die sich mit risikoreichen Anwendungsbereichen von KI befassen. Geringere System-Autonomie und Anwendungsrisiken werden durch kältere Farben (blau) und höhere System-Autonomie und Anwendungsrisiken durch wärmere Farben (rot) dargestellt.

Der Farbwechsel vermittelt das erhöhte Risiko solcher Entscheidungen. Mithilfe dieser Farbskala lässt sich das Gesamtrisiko erkennen: Violett und dunkelrot – sehr hoch, rot und dunkelorange - hoch, hellorange und Gelbtöne – mittel, Blautöne – geringes Gesamtrisiko. Im Idealfall sollten hohe Anwendungsrisiken und System-Autonomie vermieden oder nur nach sehr sorgfältiger Abwägung eingesetzt werden.

Autonomiegrad des KI-Systems

Stufe 1: Keine Autonomie

KI ist ein passives Werkzeug; Menschen treffen alle Entscheidungen und leiten Maßnahmen ein.

Beispiel: Diagnosesysteme, die medizinische Rohdaten anzeigen oder die Daten analysieren (ohne Empfehlungen!)

Empfohlene Anwendungsfälle: Szenarien mit hohen Risiken oder bei denen ethische Entscheidungen von entscheidender Bedeutung sind (z. B. medizinische Diagnostik, Justizsystem).

Stufe 2: Geringer Autonomiegrad (Human-in-the-Loop)

Die KI gibt Empfehlungen oder Optionen, aber der Benutzer:innen bleibt für die Auswahl und Genehmigung von Maßnahmen verantwortlich.

Beispiel: KI schlägt optimale Routen für die Logistik vor oder Empfehlungssysteme im E-Commerce.

Empfohlene Anwendungsfälle: Aufgaben mittlerer Komplexität mit mäßigen Risiken (z. B. Optimierung der Lieferkette).

Stufe 3: Mittlerer Autonomiegrad (Human-on-the-Loop)

Die KI führt bestimmte Aufgaben autonom aus, wobei Menschen in Ausnahmefällen eingreifen.

Beispiel: KI-gestützte Fertigungsprozesse, bei denen das System Maschinen steuert, aber Nutzende bei Anomalien eingreifen.

Empfohlene Anwendungsfälle: Szenarien, in denen eine kontinuierliche menschliche Beteiligung nicht erforderlich ist, kritische Risiken jedoch eine menschliche Überwachung erfordern (z.B. industrielle Automatisierung, Überwachung von Finanztransaktionen).

Stufe 4: Hoher Autonomiegrad (Human in Control)

Das KI-System arbeitet weitgehend autonom, erlaubt es den Benutzern jedoch, es selbst zu übersteuern, um unerwünschte Ergebnisse zu vermeiden.

Beispiel: Autonome Fahrzeuge

Empfohlene Anwendungsfälle: Umgebungen mit geringem bis mittlerem Risiko (z.B. Logistik, einfaches Verkehrsmanagement).

Stufe 5: Vollständige Autonomie mit minimaler Aufsicht

Das KI-System arbeitet unabhängig und erfordert nur minimale oder gar keine menschliche Intervention.
Die Beteiligung des Menschen beschränkt sich auf die langfristige Aufsicht (Audits).

Beispiele: Autonome landwirtschaftliche Maschinen, KI für die Stromnetzverteilung, U-Bahnen, Flughafenbahnen

Empfohlene Anwendungsfälle: Umgebungen mit geringen Sicherheits- oder ethischen Risiken und hoher Zuverlässigkeit des KI-Systems (z.B. sich wiederholende Aufgaben in kontrollierten Umgebungen).

Anwendungsbereich-Risiko

Stufe 1: Minimales Risiko

Das KI-System hat keine Auswirkungen auf den Benutzer:innen oder die Entscheidungsfindung.

Beispiele: Filter, NPCs, Empfehlungsalgorithmen ohne schwerwiegende Folgen (DeepL, andere Übersetzungssysteme)

Kriterien: Keine direkte Auswirkung auf die Hochrisikobereichen des ***EU AI-Acts***.

Stufe 2: Begrenztes Risiko

KI-Systeme, die mit Benutzer:innen interagieren, aber keine Entscheidungen mit hohen Risiken treffen. Das Risiko steigt, wenn es an Transparenz über die Beteiligung von KI mangelt.

Beispiele: Chatbots und KI-generierte Inhalte ohne Offenlegung, einfache Automatisierungsaufgaben.

Kriterien: Bereiche, die nicht in der Liste der „hohen Risiken" des EU AI-Acts enthalten sind.

Stufe 3: Mittleres Risiko

KI-Systeme haben keine besonderen Auswirkungen auf einzelne Personen, aber sie entfalten Wirkung auf kollektiver oder gesellschaftlicher Ebene.

Beispiele: Generative KI wie ChatGPT und andere Systeme, die indirekt die Umgebung beeinflussen können, in der sie eingesetzt werden.

K**riterien:** KI-Systeme, die für die öffentliche Nutzung verfügbar sind und das Potenzial haben, bestehende Gepflogenheiten zu beeinflussen und langfristig zu verändern.

Stufe 4: Hohes Risiko

Jeder Algorithmus, der in den laut EU AI-Act „Hochrisikobereichen" angewendet wird oder direkte Auswirkungen auf einzelne Personen hat.

Beispiele: Medizin, Biometrie, kritische Infrastruktur, Bildung und Berufsausbildung, Beschäftigung, Zugang zu Dienstleistungen im öffentlichen Sektor, Strafverfolgung, Migration.

Kriterien: Zugehörigkeit zum „Hochrisikobereich" des EU AI-Acts, nur wenn die Regeln für Transparenz und Datenqualität eingehalten werden.

Stufe 5: Extremes Risiko

Jeder Algorithmus, der in den laut EU AI-Act „Hochrisikobereichen" angewendet wird.

Beispiele: Medizin, Biometrie, kritische Infrastruktur, Bildung und Berufsausbildung, Beschäftigung, Zugang zu Dienstleistungen im öffentlichen Sektor, Strafverfolgung, Migration.

Kriterien: Zugehörigkeit zum „Hochrisikobereich", wenn die Regeln für Transparenz und Datenqualität NICHT eingehalten werden.

Quellen

1 European Union Agency for Fundamental Rights (Ed.). (2023). Online content moderation: Current challenges in detecting hate speech. Publications Office. https://doi.org/10.2811/332335

2 Nations, U. (2025, June 27). What is hate speech? United Nations; United Nations. https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech

3 Horan, S. (2025, March 12). Deepfake Dangers: How Content Moderation Teams Can Fight AI-Generated Misinformation. Zevo Health. https://www.zevohealth.com/blog/deepfake-dangers-how-content-moderation-teams-can-combat-ai-generated-misinformation/ Accessed 22 July 2025

4 Meta. (2020, July 1). Sharing Our Actions on Stopping Hate. Meta for Business. https://www.facebook.com/business/news/sharing-actions-on-stopping-hate Accessed 22 July 2025

5 Information provided by the IT companies about measures taken to counter hate speech, 2022. (n.d.). Retrieved 6 June 2025, from https://commission.europa.eu/system/files/2022-12/Information%20provided%20by%20the%20IT%20companies%20about%20measures%20taken%20to%20counter%20hate%20speech%20%E2%80%93%202022.pdf

6 Cortiz, D., & Zubiaga, A. (2020). Ethical and technical challenges of AI in tackling hate speech. The International Review of Information Ethics, 29. https://doi.org/10.29173/irie416

7 Mehta, H., & Passi, K. (2022). Social Media Hate Speech Detection Using Explainable Artificial Intelligence (XAI). Algorithms, 15(8), 291. https://doi.org/10.3390/a15080291

8 Yin, W., & Zubiaga, A. (2021). Towards generalisable hate speech detection: A review on obstacles and solutions. PeerJ Computer Science, 7, e598. https://doi.org/10.7717/peerj-cs.598

9 Barragán-Montero, A., Javaid, U., Valdés, G., Nguyen, D., Desbordes, P., Macq, B., Willems, S., Vandewinckele, L., Holmström, M., Löfman, F., Michiels, S., Souris, K., Sterpin, E., & Lee, J. A. (2021). Artificial intelligence and machine learning for medical imaging: A technology review. Physica Medica, 83, 242–256. https://doi.org/10.1016/j.ejmp.2021.04.016

10 Dietrich, F. (2024). AI-based removal of hate speech from digital social networks: Chances and risks for freedom of expression. AI and Ethics. https://doi.org/10.1007/s43681-024-00610-7

11 Shepherd, J. (2025, June 5). 21 Essential Twitter (X) Statistics You Need to Know in 2025. The Social Shepherd. https://thesocialshepherd.com/blog/twitter-statistics

12 Abusaqer, M., Saquer, J., & Shatnawi, H. (2025). Efficient Hate Speech Detection: Evaluating 38 Models from Traditional Methods to Transformers. Proceedings of the 2025 ACM Southeast Conference, 203–214. https://doi.org/10.1145/3696673.3723061

13 Teo, D. M. (2024, November 10). Protecting Content Moderators' Wellbeing. Zevo Health. https://www.zevohealth.com/blog/moderating-harm-maintaining-health-protecting-the-wellbeing-of-content-moderators/ Accessed 14 June 2025

14 Shen, T., Jin, R., Huang, Y., Liu, C., Dong, W., Guo, Z., Wu, X., Liu, Y., & Xiong, D. (2023). Large Language Model Alignment: A Survey (No. arXiv:2309.15025). arXiv. https://doi.org/10.48550/arXiv.2309.15025

15 Potamias, R. A., Siolas, G., & Stafylopatis, A.-G. (2020). A transformer-based approach to irony and sarcasm detection. Neural Computing and Applications, 32(23), 17309–17320. https://doi.org/10.1007/s00521-020-05102-3

16 Taylor, J., Peignon, M., & Chen, Y.-S. (2017). Surfacing contextual hate speech words within social media (No. arXiv:1711.10093). arXiv. https://doi.org/10.48550/arXiv.1711.10093

17 Coldewey, D. (2019, August 14). Racial bias observed in hate speech detection algorithm from Google. TechCrunch. https://techcrunch.com/2019/08/14/racial-bias-observed-in-hate-speech-detection-algorithm-from-google/

18 Castiglioni, I., Rundo, L., Codari, M., Di Leo, G., Salvatore, C., Interlenghi, M., Gallivanone, F., Cozzi, A., D'Amico, N. C., & Sardanelli, F. (2021). AI applications to medical images: From machine learning to deep learning. Physica Medica, 83, 9–24. https://doi.org/10.1016/j.ejmp.2021.02.006

19 Art. 1 DSGVO – Gegenstand und Ziele. (2016). Datenschutz-Grundverordnung (DSGVO). https://dsgvo-gesetz.de/art-1-dsgvo/

20 European Union. (2025, May 27). How the Digital Services Act enhances transparency online | Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/policies/dsa-brings-transparency

21 Rawat, A., Kumar, S., & Samant, S. S. (2024). Hate speech detection in social media: Techniques, recent trends, and future challenges. WIREs Computational Statistics, 16(2), e1648. https://doi.org/10.1002/wics.1648

22 Anjum, & Katarya, R. (2024). Hate speech, toxicity detection in online social media: A recent survey of state of the art and opportunities. International Journal of Information Security, 23(1), 577–608. https://doi.org/10.1007/s10207-023-00755-2

23 Calabrese, A., Neves, L., Shah, N., Bos, M., Ross, B., Lapata, M., & Barbieri, F. (2024). Explainability and Hate Speech: Structured Explanations Make Social Media Moderators Faster. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp.398–408). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-short.38

24 Coalition for Content Provenance and Authenticity. (2025). Overview—C2PA. https://c2pa.org/

25 Sharma, A., Maier, F., & Breeden, J. (2025, April 17). Open Source Intelligence: Die besten OSINT Tools. Computerwoche. https://www.computerwoche.de/article/2795282/wie-viel-wissen-hacker-ueber-sie.html Accessed 19 June 2025

26 Carvallo, A., Quiroga, T., Aspillaga, C., & Mendoza, M. (2024). Unveiling Social Media Comments with a Novel Named Entity Recognition System for Identity Groups. arXiv preprint arXiv:2405.13011.

27 Lin, J. (2022). Leveraging world knowledge in implicit hate speech detection. arXiv preprint arXiv:2212.14100.

28 Mnassri, K., Rajapaksha, P., Farahbakhsh, R., & Crespi, N. (2023, May). Hate speech and offensive language detection using an emotion-aware shared encoder. In ICC 2023-IEEE International Conference on Communications (pp.2852–2857). IEEE.

29 Martins, R., Gomes, M., Almeida, J. J., Novais, P., & Henriques, P. (2018, October). Hate speech classification in social media using emotional analysis. In 2018 7th Brazilian Conference on Intelligent Systems (BRACIS) (pp.61–66). IEEE.

30 Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and Amending Directive 2000/31/EC (Digital Services Act) (Text with EEA Relevance), 277 OJ L (2022). http://data.europa.eu/eli/reg/2022/2065/oj/eng

31 Meta Oversight Board Bylaws: (2024). https://www.oversightboard.com/wp-content/uploads/2024/03/Oversight-Board-Bylaws.pdf Accessed 22 July 2025

32 Data Workers' Inquiry. (n.d.). Data Workers' Inquiry. Retrieved 18 September 2025, from https://data-workers.org/

33 Qhala. (2025, June 10). Data Workers in AI: A New Frontier of Labour Exploitation in the Global South. Medium. https://qhalahq.medium.com/data-workers-in-ai-a-new-frontier-of-labour-exploitation-in-the-global-south-362e22eae01b

34 Tasin, J. J. and F. (2024, July 22). How the global South may pay the cost of AI development. OMFIF. https://www.omfif.org/2024/07/how-the-global-south-may-pay-the-cost-of-ai-development/

Projekt ALAIT

Das Austrian Lab for AI Trust (ALAIT) ist ein vom österreichischen Bundesministerium für Innovation, Mobilität und Infrastruktur (BMIMI) initiiertes Forschungs- und Entwicklungs-Projekt zur Schaffung von Vertrauen durch Wissen im Bereich Künstliche Intelligenz (KI). Das Projekt ALAIT zielt darauf ab, Interessierte und wichtige gesellschaftliche Gruppen zu befähigen, KI-Technologien verantwortungsvoll zu nutzen und ethische sowie qualitativ hochwertige Standards für den Einsatz von AI zu etablieren.

Das Projekt wird von **winnovation** geleitet (Gertraud Leimüller und Lena Müller-Kress) und im Konsortium mit **leiwand.ai** (Rania Wazir und Silvia Wasserbacher-Schwarzer), **TU Wien** (Sabine Köszegi und Ilya Faynleyb) und **Austria Presse Agentur – APA** (Verena Krawarik und Sophia Marecek) umgesetzt.

Die ALAIT-Dossiers sind auf der Projekthomepage abrufbar: https://science.apa.at/project/alait/

Die Inhalte des Dossiers entsprechen dem aktuellen Stand der Technik und wurden sorgfältig nach wissenschaftlichen Kriterien erstellt. Sie dienen jedoch nicht als rechtsverbindliche Auskunft oder Beratung.

Impressum

Medieninhaberin und Herausgeberin:
winnovation consulting gmbh
Linke Wienzeile 42/1, Top 5
1060 Vienna

Danksagung:

Wir danken folgenden Expert:innen für ihre Unterstützung zu Vorversionen dieses Dossiers:
Florian Schmidt (APA), Ingrid Brodnig, Mira Reisinger (leiwand.ai)

Veröffentlicht 2026