FORSCHUNG

VERTRAUENSWÜRDIGE KI IM MEDIENKONTEXT

GRUNDLEGENDE ANFORDERUNGEN UND HERAUSFORDERUNGEN

PATRICK AICHROTH , HANNA LUKASHEVICH

Vieles spricht dafür, dass KI besonders dort erfolgreich sein wird, wo sie vertrauenswürdig gestaltet wird. Dieser Artikel liefert einen Überblick über Anforderungen bezüglich vertrauenswürdiger KI und deren Berücksichtigung im Entwicklungsprozess.

There are reasons to assume that AI will be most successful if provided in a trustworthy manner. This article provides an overview over Trustworthy AI requirements and their possible integration into the development process.

Begriffe: KI und maschinelles Lernen

"Difference between machine learning and AI: If it is written in Python, it's probably machine learning. If it is written in PowerPoint, it's probably AI." – Mat Velloso (https://twitter.com/matvelloso/status/1065778379612282885)

„KI" (Künstliche Intelligenz) oder englisch „AI" (Artificial Intelligence) beschreibt allgemein Methoden, Komponenten oder Systeme, die sich „intelligent" verhalten, d. h. einem Computer die Lösung von Aufgaben ermöglichen, die menschliche Intelligenz erfordern. Dabei unterscheidet man zwischen „starker KI", die auf allgemeine Intelligenz und die Lösung allgemein aller möglichen Aufgaben abzielt, und „schwacher KI", die spezielle, ausgewählte Aufgaben löst. Praktisch alle im Medienkontext diskutierten und eingesetzten KI-Verfahren fallen unter „schwache KI".

KI beinhaltet dabei die funktionalen Ebenen **Sensing** (die Übersetzung von sensorischen Informationen aus der physischen Welt in eine für den Computer interpretierbare, konzeptionelle Darstellung), **Reasoning** (Veränderungen der konzeptionellen Darstellung, zum Beispiel Lernen und logische Schlussfolgerungen) und **Acting** (die Übersetzung der konzeptionellen Darstellung in Handlungen in der physischen Welt) [1]. Die Grenzen zwischen diesen Ebenen sind in der Praxis nicht immer trennscharf. Trotzdem kann man feststellen: Viele der im Broadcasting-Umfeld eingesetzten KI-basierten Werkzeuge für automatische Metadaten-Extraktion (AME), wie zum Beispiel Objekt-, Gesichts- und Sprechererkennung, bewegen sich primär auf Sensing-Ebene. Sie erkennen Konzepte in Mediendaten und speichern diese in Form von Metadaten oder Annotationen. Andere Werkzeuge, wie KI-basierte Empfehlungssysteme, beinhalten auch Reasoning und Acting: Sie verwenden Daten über Nutzung, Nutzer und Inhalte, um zum Beispiel Vorhersagen darüber abzuleiten, welche Inhalte einem bestimmten Nutzer zwar noch nicht bekannt sind, aber gefallen könnten, und liefern diese Empfehlungen an die Nutzer.

KI setzt zur Aufgabenerfüllung zwei grundsätzlich verschiedene Methoden ein: einerseits logik- und regelbasierte Ansätze für Machine Reasoning (Optimierung, logisches Schließen etc.), andererseits lernbasierte Ansätze für Machine Learning. Bei Letzterem wird der Computer mittels einer repräsentativen Auswahl von Daten (Trainingsdaten) für eine Aufgabe trainiert, ohne explizit programmiert zu werden. Anschließend kann er (mehr oder weniger gute) Vorhersagen oder Entscheidungen für andere, vergleichbare Daten liefern. „Deep Learning", welches in den vergangenen Jahren große Erfolge bei verschiedenen Aufgaben ermöglicht hat, gehört zu diesen Verfahren. Es verwendet „tiefe" neuronale Netze mit vielen Zwischenschichten, nicht unähnlich der menschlichen Physiologie.

KI im Medienkontext: Vertrauensfragen

"Our technology, our machines, is part of our humanity. We created them to extend ourselves, and that is what is unique about human beings." – Ray Kurzweil

KI-basierte Werkzeuge bieten auch und gerade für Broadcaster ungeheure Chancen: Für **Journalisten und Produzenten** ergeben sich zum Beispiel völlig neue Möglichkeiten, relevante Inhalte zu entdecken, einfacher und schneller zu produzieren und bei Recherchen einen möglichst umfassenden Überblick zu erhalten. **Anbieter** können über Personalisierungs- und Empfehlungsdienste viel besser und zielgerichteter mit Nutzern interagieren, große Mengen von Archivmaterial nutzbar machen und höhere Werbeeinnahmen erzielen. Und **Nutzer** haben Zugriff auf viel mehr und viel relevantere Inhalte.

Um diese Chancen nutzen zu können, müssen aber einige Herausforderungen bewältigt werden, die das Vertrauen in KI untergraben können. Dazu vier Beispiele:

- **Bias** oder Verzerrung kann Probleme bezüglich Fairness und Diskriminierung verursachen. Ein bekanntes Beispiel hierfür ist, dass bei Google Images 2015 Fotos von Afroamerikanern als Gorilla-Bilder klassifiziert wurden [2]. Solche Fehler können unter anderem aufgrund von Disbalancen bei den zugrundeliegenden Trainingsdaten auftreten und werden als „Sample Bias" bezeichnet. Eine große Herausforderung bei der Entwicklung von KI besteht in der Frage, wie Sample Bias erkannt und vermieden werden kann. Bias und Fairness stellen aber auch deshalb Herausforderungen dar, weil sie teilweise

Abbildung 1. "This person does not exist" based on StyleGAN, https://thispersondoesnotexist.com/

politische Implikationen haben und dann auf technischer Ebene alleine kaum zu beantworten sind.

- **Deepfakes** sind mit Hilfe von KI generierte A/V-Inhalte, die authentisch wirken, aber es nicht sind. So kann neben fingierten Bildern (siehe Abbildung 1) und Videos auch Sprache erzeugt werden, die für den Zuhörer nicht mehr von natürlichen Aufnahmen zu unterscheiden ist [3]. Neben vielen sinnvollen Einsatzgebieten ergeben sich daraus erhebliche Probleme bezüglich Betrug [4], politischer Manipulation und Vertrauensverlust bei der Informationskommunikation.
- **Verlust der Privatsphäre:** KI-basierte Systeme verarbeiten nicht selten direkt oder mittelbar sensible oder persönliche Daten. Die Erhebung und Verarbeitung von immer umfangreicheren Daten, oft gepaart mit Sicherheitsdefiziten, bringt große Risiken bezüglich Kontroll- und Datenverlusten mit sich. Entsprechende Vorfälle mit teilweise schwerwiegenden Folgen gab es schon häufig (Beispiel: mehrere Suizide infolge der Veröffentlichung von Nutzerdaten der Seitensprung-Plattform Ashley Madison 2015 [5]). Noch wesentlich schwerer abzuschätzen und zu kontrollieren sind aber die Risiken, die sich durch die Verknüpfung von Daten und Diensten langfristig ergeben werden.
- **Filterblasen:** KI-basierte Personalisierungs- und Empfehlungsdienste können unsere Sicht auf die Welt extrem verengen. Das liegt daran, dass sie oft ausschließlich auf das Erfolgskriterium *Utility* hin entwickelt und evaluiert werden, also auf die Bereitstellung von Informationen, die wir präferieren. Da wir unsere eigene Meinung gern bestätigt sehen (confirmation bias) und konträren Ansichten eher aus dem Weg gehen, befördert KI auf diese Weise eine verengte Wahrnehmung der Realität, mit potenziell dramatischen Folgen für Meinungsbildung und Interessenausgleich in offenen Gesellschaften.

Für diese und andere Herausforderungen gibt es diverse methodische und technische Lösungsansätze, beispielsweise

- Werkzeuge zur Erkennung und Vermeidung von Bias (die im Folgenden noch erwähnt werden),
- Technologien zur Erkennung von A/V-Manipulationen [6] sowie Ansätze zur Erkennung von synthetischem A/V-Material (die sich allerdings noch in den Kinderschuhen befinden),
- Ansätze zur Vermeidung von Filterblasen durch die Beförderung von Neuheit und Diversität bei Personalisierungs- und Empfehlungssystemen [7] oder
- Privacy Enhancing Technologies (PET) zum Schutz der Privatsphäre, die zum Beispiel für Empfehlungsdienste eingesetzt werden können [7], oder Verfahren für dezentrale Analyse und Trainingsprozesse.

Der geschickte Einsatz solcher Werkzeuge erlaubt in vielen Fällen eine Lösung der Probleme, *ohne* dabei die Möglichkeiten von leistungsfähiger Analyse und Personalisierung zu opfern. Besonders Privatsphäre und Datenanalyse werden fälschlicherweise oft als unvereinbare Widersprüche behandelt.

Voraussetzung für die Entwicklung und Nutzung geeigneter Werkzeuge ist aber, dass es (1) allgemeinverbindliche Leitlinien und den Willen für Entwicklung und Einsatz von „vertrauenswürdiger KI" gibt und (2) eine Methodik, wie die Anforderungen in die Entwicklungsprozesse integriert werden können. Beide sollen im Folgenden behandelt werden.

Vertrauenswürdige KI: Definition

"Trust is the glue of life. It's the most essential ingredient in effective communication. It's the foundational principle that holds all relationships." – Stephen Covey

Die Europäische Kommission ist einer der wichtigsten internationalen Akteure bei der Vorgabe von Leitlinien für vertrauenswürdige KI (Trustworthy AI, im Folgenden kurz TAI). Das Dokument "Ethics Guidelines for Trustworthy AI" [8] der zuständigen Expertengruppe schlägt in diesem Zusammenhang v. a. folgende Anforderungen für TAI vor:

- **Human agency and oversight:** KI muss jederzeit unter menschlicher Kontrolle stehen und nutzerzentriert entwickelt werden.
- **Technical robustness and safety:** Dazu gehören Robustheit und Sicherheit für KI einschließlich Zuverlässigkeit, Genauigkeit und Reproduzierbarkeit sowie Notfallplanung. In diese Kategorie gehört auch die Abwehr von Angriffen auf KI-Systeme in Form von subtilen Datenveränderungen.
- **Privacy and data governance:** Dies umfasst die umfangreiche Berücksichtigung von Anforderungen zum Schutz der Privatsphäre und Datenschutz oder Datensouveränität sowie damit verbundene Aspekte von Datensicherheit und Zugriffsschutz zum Beispiel durch den Einsatz von „Privacy Enhancing Technologies".
- **Transparency:** Damit sind Transparenz und zielgruppengerechte Erklärbarkeit bezüglich der Erhebung und Verwendung von Daten, der eingesetzten KI-Methoden und der Prozesse und Geschäftsmodelle gemeint – auch die Erkennung von Deepfakes gehört im weiteren Sinne dazu. Der Erklärbarkeit sind bei lernbasierten Verfahren allerdings gewisse Grenzen gesetzt, was in manchen Fällen für den Einsatz regelbasierter KI-Ansätze sprechen kann.
- **Diversity, non-discrimination and fairness:** Dies zielt auf die Förderung von Informationsvielfalt und diskriminierungsfreiem Zugang, Vermeidung von „Unfair Bias" und Einbindung aller relevanten Stakeholder ab. Die Vermeidung von Filterblasen kann dieser Kategorie zugeordnet werden.
- **Societal and environmental well-being:** Dies ist die Forderung nach Nachhaltigkeit und Umweltfreundlichkeit für heutige und künftige Generationen.
- **Accountability:** Hiermit sind verschiedene Aspekte bezüglich Verantwortlichkeit und Rechenschaft für KI gemeint. Das schließt auch die Überprüfbarkeit von Komponenten und Systemen ein.

Mit der Ausarbeitung von Prinzipien und Anforderungen für TAI hat die Europäische Kommission durchaus eine Vorreiterrolle übernommen, und die Vorgaben und der dahinterstehende politische Wille sind eine wichtige Basis für die Entwicklung von TAI.

Natürlich hängen die Bedeutung und Interpretation der Anforderungen stark vom Anwendungsbereich ab. Tendenziell kann man sagen, dass für Broadcaster vor allem die oben genannten Themen Privatsphäre, Bias und Transparenz relevant sind.

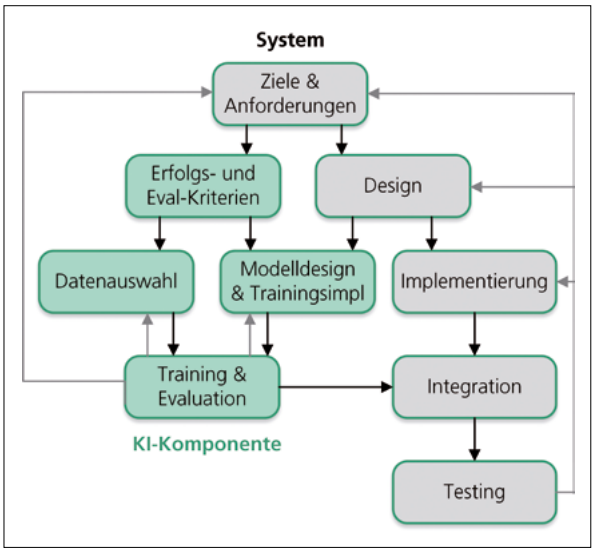

Abbildung 2. Vorschlag: Kombination von KI-Entwicklung und „konventionellem" Software-Engineering

TAI-Anforderungen im Entwicklungsprozess

"The most important thing about goals is having one." – Geoffrey F. Abert

Um TAI zu realisieren, ist – neben geeigneten Ansätzen für spezielle Herausforderungen wie PET oder Deepfake-Erkennung – vor allem ein Prozess erforderlich, der dafür sorgt, dass die genannten Prinzipien und Anforderungen in den Entwicklungsprozess einfließen und ihn steuern. Das ist in der Praxis nicht ganz selbstverständlich, denn KI-Entwicklung folgt im Gegensatz zu konventioneller Softwareentwicklung tendenziell eher einem datenzentrierten als einem anforderungszentrierten Ansatz, und Training und Evaluation laufen verschränkt ab.

TAI erfordert aber Maßnahmen und Evaluation nicht nur auf Ebene von KI-Komponenten, sondern auch auf Ebene der übergeordneten Systeme, in die KI-Komponenten eingebettet sind. Aus diesem Grund bietet es sich an, beide Ansätze in einem gemeinsamen Prozess zu kombinieren (siehe Abbildung 2):

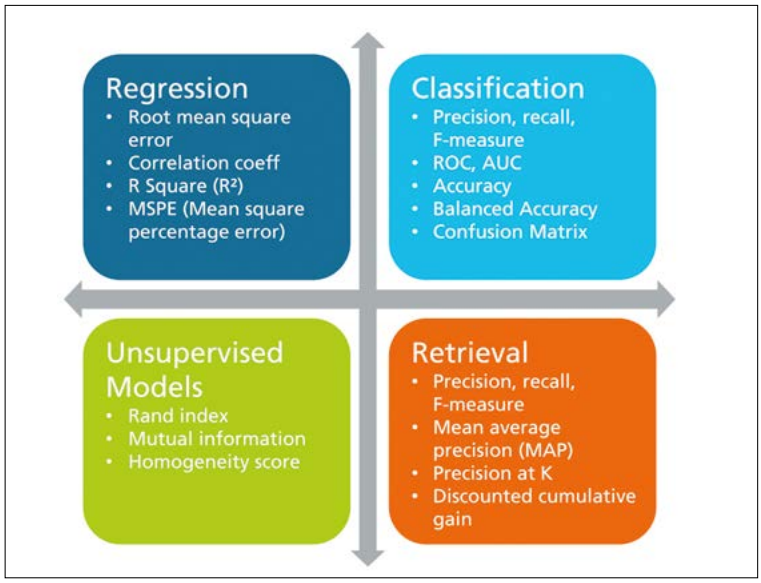

Abbildung 3. Evaluationsmetriken (Auswahl)

- TAI-Anforderungen können zum Beispiel über **Ziele und Anforderungen an das Gesamtsystem** mit formuliert werden und werden so zur Voraussetzung für alle weiteren KI-Entwicklungsschritte. Für die Formulierung bieten sich Standardansätze der konventionellen Softwareentwicklung an. So können Anforderungen an ein datenschutzfreundliches, auf größere Meinungsvielfalt und auf Transparenz abzielendes TAI-Empfehlungssystem beispielsweise als „User Stories" in der Form „Als *<Rolle>* möchte ich *<Ziel/Wunsch>*, um *<Nutzen>*" so formuliert sein:
  - Als Nutzer möchte ich Nachrichten mit möglichst großer Meinungsvielfalt zu einem ausgewählten Thema erhalten, um mir ein umfassendes Bild zu machen.
  - Als Nutzer möchte ich einsehen können, mit welchen Interessengebieten ich vom System aktuell assoziiert werde, um diese zu verstehen und gegebenenfalls anzupassen.
  - Als Anbieter möchte ich Nutzern Empfehlungen ohne Rückschlüsse auf die Nutzeridentität liefern können, um Vertrauen und Nutzerbindung zu vertiefen.
- Auf Basis dieser Anforderungen werden dann **Erfolgskriterien und Evaluationsvorgaben** für die KI-Komponenten abgeleitet. In unserem Beispiel eines TAI-Empfehlungssystems, das auf größere Meinungsvielfalt abzielt, fließen neben dem Erfolgskriterium Nützlichkeit (Utility) mit Metriken wie Precision und Recall (siehe Abbildung 3) bezüglich relevanter Empfehlungen auch noch Kriterien wie Diversität und Neuheit mit geeigneten Metriken ein, um die Modelle in Richtung einer Vermeidung von Filterblasen zu steuern. Dazu kommt das Kriterium Privatheit (Privacy), das zum Beispiel über „Differential Privacy"-Metriken einfließt, die eine quantitative Bewertung des Schutzes der Privatsphäre erlauben.
- Es folgen **Modelldesign und Trainingsimplementierung** (Vorverarbeitung etc.) sowie die **Auswahl geeigneter Trainings- und Testdaten** für die gegebenen Ziele und Anforderungen. Hier liegt ein weiterer entscheidender Punkt bezüglich TAI-Anforderungen: Wenn die Daten nicht in einer dem Kontext angemessenen Variabilität und Menge vorhanden sind, ergeben sich später Probleme zum Beispiel bezüglich Sample Bias. Um solche Disbalancen frühzeitig zu erkennen, bieten sich Werkzeuge wie Google Facets [9] an. Allerdings gibt es weitere Herausforderungen bezüglich Bias, die später noch einmal aufgegriffen werden.
- Schließlich werden Evaluationsvorgaben und Daten verwendet, um die entwickelten KI-Komponenten zu **trainieren** und zu **evaluieren**. Für Evaluation und Fehlersuche eignen sich Tools wie *Google What-if* [10] (siehe Abbildung 4) oder die *INNvestigate toolbox* [11], für das systematische Durcharbeiten relevanter ethischer Fragen Vorlagen wie *Data Ethics Canvas* [12] und *EthicalOS* [13]. Daraus kann sich eine Notwendigkeit für Datenmodifikation und Nachtrainieren ergeben. Erfüllen die KI-Komponenten schließlich die gegebenen Anforderungen, werden sie ins Gesamtsystem integriert.

Generell ist anzumerken: Eine Modularisierung von KI-Systemen kann Evaluation und Fehlersuche erleichtern und die Transparenz verbessern, ist aber nicht immer realisierbar. Und die Evaluation gestaltet sich bei KI-Komponenten leichter, bei denen ein hoher Grad an Übereinstimmung bei menschlichen Experten oder Annotatoren (inter-annotator agreement) gegeben ist, was zum Beispiel bei KI-Komponenten auf Sensing-Ebene oft der Fall ist.

Abbildung 4. Nutzung des What-If Tools zur Analyse von Machine-Learning-Modellen, siehe https://towardsdatascience.com/using-what-if-tool-to-investigate-machine-learning-models-913c7d4118f

Schließlich kann auch der Einsatz von **automatischem Testen** erwogen werden: Die Grundidee von Unit Testing aus dem Bereich der Softwareentwicklung lässt sich grundsätzlich auch auf die Evaluierung von KI-Komponenten übertragen. Voraussetzung dafür sind (1) eine Spezifizierung der Schnittstellen der zu evaluierenden Komponenten, (2) die Definition der relevanten Anforderungen und Test Cases und (3) die Bereitstellung oder automatische Generierung von Testdaten. Sind diese Vorbereitungen einmal getroffen, lassen sich Tests nicht nur während aller Entwicklungsphasen, sondern auch vergleichende Tests verschiedener Komponenten zu einem Anwendungsfall sehr effizient durchführen. Ein solches Framework wurde für die automatische Evaluation von Medienforensik-Tools schon vor einigen Jahren entwickelt (siehe Abbildung 5) [14].

Zusammenfassend lässt sich feststellen, dass ein anforderungsorientierter Prozess, bei dem TAI-Anforderungen in geeignete Erfolgs- und Evaluationskriterien übersetzt werden, um so den Entwicklungsprozess zu steuern, für die Realisierung von TAI entscheidend ist.

Allerdings: Ausgerechnet eine der am hitzigsten diskutierten Anforderungen – die Herstellung von Fairness und Vermeidung von Bias – ist mit zusätzlichen Herausforderungen verbunden.

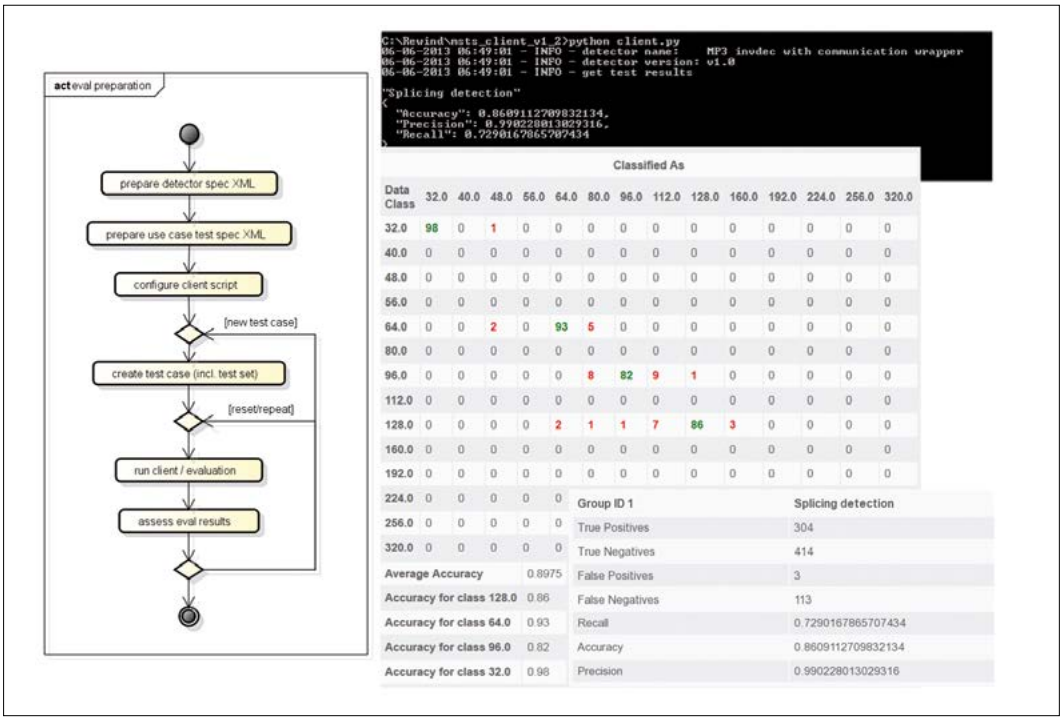

Abbildung 5. Automatisches Testen von Klassifikatoren beim REWIND-Projekt [14]

Bias und Fairness

"A good intention, with a bad approach, often leads to a poor result." – Thomas A. Edison

Bias und Fairness sind Begriffe, die im Kontext von KI-Entwicklung häufig und kontrovers diskutiert werden und alles andere als einfach zu adressieren sind.

Mit Bias ist in diesem Kontext zunächst und vorrangig der bereits erwähnte Sample Bias gemeint. **Sample Bias** kann bei der Entwicklung von KI im Medienkontext praktisch überall auftreten: bei Gesichtserkennung, die für einige Bevölkerungsgruppen viel besser funktioniert als für andere, oder bei Empfehlungssystemen, die bestimmte Arten von Inhalten systematisch bevorzugen, weil sie überrepräsentiert sind.

Die Erkennung von Sample Bias wird unter anderem dadurch erschwert, dass bei der Entwicklung von lernbasierten KI-Modellen die verfügbaren Daten oft in Trainings- und Testdaten aufgeteilt werden: Disbalancen sind in den Testdaten dann ebenso vorhanden wie in den Trainingsdaten und werden beim Testen nicht erkannt. Öffentlich verfügbare Testdaten können hier zwar Abhilfe schaffen, sind aber nicht immer verfügbar. Ein weiteres Problem besteht darin, dass Sample Bias (wie jede Verzerrung) nur anwendungsspezifisch bewertet und vermieden werden kann: Trainingsdaten, die zum Beispiel für die Entwicklung von Musikanalyse in einer bestimmten kulturellen Region ausgelegt und angemessen sind, werden für andere Regionen natürlich „biased" sein. Und dennoch: Sample Bias kann durch Abgleich mit der realen Welt und die bereits genannten Werkzeuge noch relativ gut vermieden werden.

Eine wesentlich schwierigere Form von Bias, teilweise **Prejudice Bias** genannt, meint Verzerrungen, die durch kulturelle Einflüsse oder Stereotype entstehen, und kann sich zum Beispiel auf Geschlechterrollen, Kulturen oder Ethnien beziehen. Im Medienkontext wurde dies zum Beispiel im Zusammenhang mit der Dominanz männlicher CEOs oder mit geschlechtsspezifischen Berufsbildern als Ergebnis entsprechender Google-Suchanfragen diskutiert.

Auf den ersten Blick kann dieser Bias vermieden werden, indem „geschützte Merkmale" wie Geschlecht etc. einfach aus den Daten entfernt werden. Damit ist es aber nicht getan. Um einen solchen Bias zu entfernen, müssen zusätzlich alle Merkmale, die mit den „geschützten Merkmalen" korrelieren (zum Beispiel Körpergröße), entfernt werden. Diese sind aber je nach Anwendungsfall schwer oder überhaupt nicht zu identifizieren (weil die korrelierenden Merkmale in den Daten „verborgen" sind).

Zudem kann das Entfernen der Merkmale wiederum einen sogenannten **Exclusion Bias** verursachen: Dieser Bias entsteht dadurch, dass Merkmale aus Daten entfernt werden, die für die Analyse wichtig sind. So kann das Entfernen von Merkmalen zum Beispiel beim Trainieren eines Gesichtserkenners dazu führen, dass für bestimmte Personengruppen die Erkennungsraten schlechter werden, was wiederum ebenfalls als Bias betrachtet werden kann. Zwischen Prejudice Bias und Exclusion Bias besteht also ein potenzieller Konflikt [15].

Zudem stellen sich weitere Fragen: Gibt es eine universelle, allgemein akzeptierte Definition von akzeptablen versus inakzeptablen Stereotypen, die als Basis für Entscheidungen dienen kann? Ist es eine gute Idee, Ergebnisse

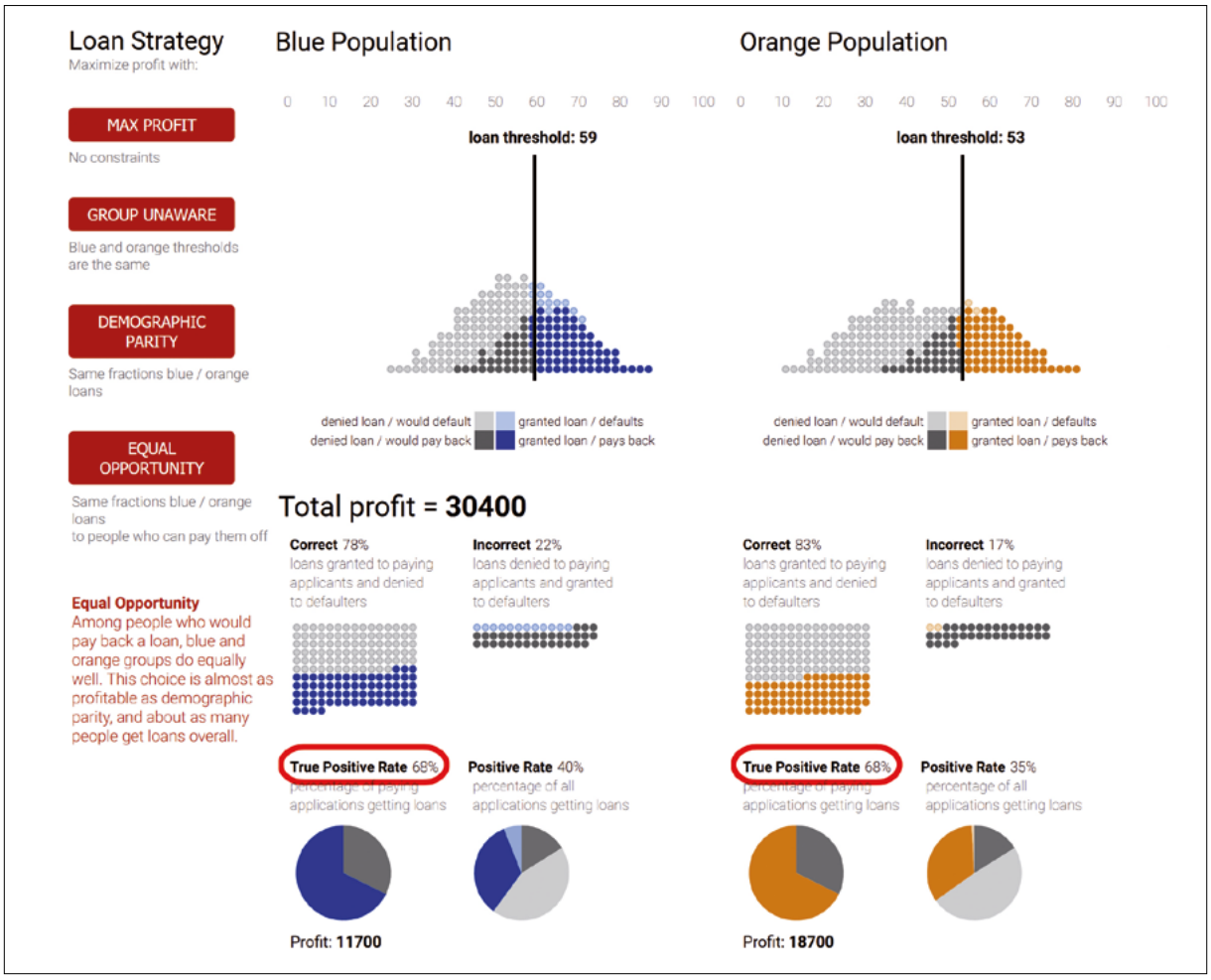

Abbildung 6. Simulation verschiedener Fairness-Strategien, siehe [18]

bezüglich demographischer Parität zu nivellieren, wenn dadurch Realitäten und damit verbundene Probleme „unsichtbar" gemacht werden?

Und schließlich stellen sich auch grundsätzliche Fragen zu **Fairness**: Zielt man auf Chancen- oder auf Ergebnisgleichheit ab, auf Gruppenfairness (das heißt im Mittel werden Gruppen gleichbehandelt) oder auf individuelle Fairness? Diese Ziele stehen im Widerspruch zueinander, es gibt keine „perfekte Fairness". Eine Verdeutlichung dieser Problematik liefert zum Beispiel eine Website von Google, bei der man verschiedene Fairness-Strategien anhand eines einfachen Beispiels durchspielen kann (siehe Abbildung 6) [16]: Die dort dargestellten Beispielstrategien (Profitmaximimierung, gruppenagnostische Entscheidung, demographische Parität, Chancengleichheit) weisen alle mehr oder minder gravierende Nachteile auf. Es gibt kein universelles „richtiges" Verständnis von Fairness, und es ist deshalb erforderlich, anwendungsspezifische Definitionen für Fairness zu finden und zu kommunizieren. In kritischen Bereichen wird es vermutlich auch notwendig sein, eine gesellschaftliche Einigung über das anzuwendende Fairnessprinzip zu erzielen.

Bei der Diskussion um Bias bei KI sollte man nicht vergessen, dass wir Menschen selbst mit vielen Arten von Bias zu kämpfen haben, die mit Blick auf die Evolution sinnvoll und tief in uns verankert sind, aber teilweise überhaupt nicht mehr in unsere Zeit passen [17]. Wie schlecht unsere Intuition schon in relativ einfachen Fällen ist, zeigt zum Beispiel die sogenannte „Base Rate Fallacy" (zum Selbsttest siehe Abbildung 7), und die Problematik des im Zusammenhang mit Filterblasen schon angesprochenen „Confirmation Bias" erleben wir alle ständig. Bias und Generalisierungen sind notwendiger Teil unseres Lebens, sonst wären wir nicht zu Entscheidungen fähig. Sie sind nicht per se schlecht. Aber wir brauchen bessere Werkzeuge, um für einen gegebenen Kontext problematischen von unproblematischem Bias zu unterscheiden.

Hier bietet KI große Chancen: Da sich lernbasierte KI nur über die Daten Bias aneignen kann, kann Bias bei KI letztlich kontrolliert, getestet und vergessen werden – Dinge, die bei Menschen schwierig bis unmöglich sind. KI ist insofern perfekt geeignet, um uns den Spiegel vorzuhalten und zu einem tieferen Verständnis von Zielkonflikten und Problemen bezüglich Bias und Fairness zu gelangen, diese differenzierter zu betrachten und letztlich bessere Lösungen zu finden. Auf dieser Basis könnte dann eine Generation von TAI-Komponenten und -Systemen entwickelt werden, die diese Ziele vermutlich besser realisieren kann, als uns Menschen das ohne Hilfe jemals möglich wäre. Das setzt allerdings voraus, dass wir bereit sind, unsere eigenen Beschränkungen und Verzerrungen zu akzeptieren und uns auf diese Diskussion einzulassen – eine Diskussion, die nicht nur KI-Entwickler, sondern die ganze Gesellschaft betrifft.

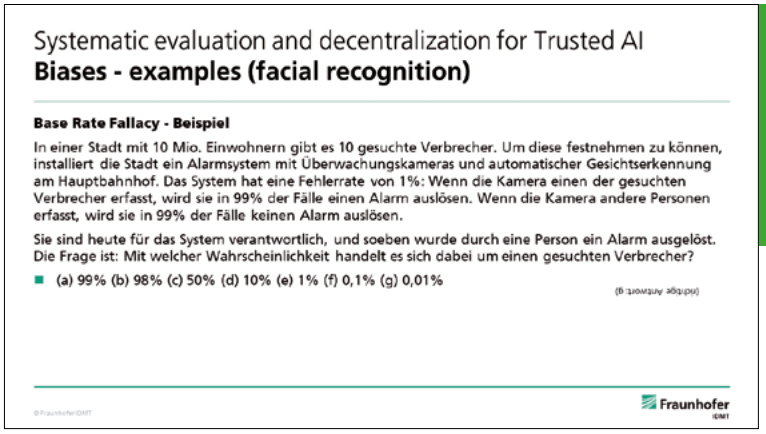

Abbildung 7. Base Rate Fallacy – ein Beispiel für menschlichen Bias

Zusammenfassung

If we do not hang together, we will all hang separately." – Benjamin Franklin

Im Wettlauf um Entwicklung und Einsatz von KI droht das eigentlich aus einer guten Ausgangslage kommende Europa den Anschluss gegenüber den USA und China zu verlieren. Und es gibt nicht wenige in Politik und Gesellschaft, die das auch und vor allem auf Themen wie den Schutz der Privatsphäre und andere TAI-Aspekte zurückführen, die in Europa besonderes Gewicht haben. Diese Annahme ist aber möglicherweise ein fataler Trugschluss.

Denn erstens liegt nahe, dass Europas Probleme primär in ganz anderen Bereichen liegen, wie zum Beispiel einem Mangel an global bedeutenden Unternehmen für KI und Halbleiterherstellung und an vergleichsweise schlechteren Finanzierungsmöglichkeiten. Oder, was möglicherweise damit in Verbindung steht, weil es eine messbar pessimistischere Grundhaltung gegenüber KI und eine erstaunliche Gelassenheit bezüglich der Dringlichkeit von KI in Europa gibt [18], die sich zu einer selbsterfüllenden Prophezeiung entwickeln könnte. Denn wer sich in der Diskussion um KI auf Kritik und Skepsis beschränkt anstatt Lösungen zu entwickeln und zu testen, der verabschiedet sich davon, die Zukunft zu gestalten.

Zweitens sollte man nicht (richtige) Zielsetzung mit (noch unzureichender) Umsetzung verwechseln: Es gibt viele Möglichkeiten, TAI mit der Entwicklung leistungsfähiger KI zu verbinden. Die eigentliche Frage lautet also, ob man Entwicklung und Einsatz dieser Möglichkeiten bisher im erforderlichen Umfang fördert und fordert. Denn tatsächlich liegt in TAI, die viele für eine europäische Schwäche halten, vermutlich eine sehr große Chance:

Vieles spricht dafür, dass KI besonders dort erfolgreich sein wird, wo es gelingt, sie vertrauenswürdig und nachvollziehbar zu gestalten. Europa hat sich bei dieser Frage früh und deutlich positioniert, aber Ziele und Vorgaben alleine genügen nicht. TAI-Anforderungen müssen jetzt auch mit Effizienz und Pragmatismus in der Praxis umgesetzt werden, ohne sich dabei in unrealistischen Vorstellungen und Tech-

Bild: Fraunhofer IDMT

PATRICK AICHROTH

ist seit 2003 wissenschaftlicher Mitarbeiter und seit 2006 Leiter der Gruppe „Mediendistribution und Sicherheit" am Fraunhofer-Institut für Digitale Medientechnologie IDMT in Ilmenau.

www.idmt.fraunhofer.de

Bild: Fraunhofer IDMT

HANNA LUKASHEVICH

ist seit 2006 wissenschaftliche Mitarbeiterin und seit 2014 Leiterin der Gruppe „Semantische Musiktechnologien" am Fraunhofer-Institut für Digitale Medientechnologie IDMT in Ilmenau.

www.idmt.fraunhofer.de

nikfeindlichkeit zu verlieren und ohne dabei die Leistungsfähigkeit von KI zu opfern.

Vertrauenswürdigkeit und Leistungsfähigkeit sind keine Widersprüche – ihre Kombination ist möglich und bietet eine große Chance. Die Ansätze und F&E-Potenziale dafür existieren, auch und gerade in Europa, wo „Talent" immer noch in großem Umfang vorhanden ist. Es werden aber vielleicht etwas mehr Selbstbewusstsein, Optimismus und Mut für unkonventionelle Lösungen notwendig sein, um etwas aus dieser Chance zu machen.

Referenzen

[1] Introduction to Artificial Intelligence, https://cs.lmu.edu/~ray/notes/introai/, [abgerufen am 18.11.2019]

[2] Google engineer apologizes after Photos app tags two black people as gorillas, https://www.theverge.com/2015/7/1/8880363/google-apologizes-photos-app-tags-two-black-people-gorillas, [abgerufen am 18.11.2019]

[3] Google's voice-generating AI is now indistinguishable from humans, https://qz.com/1165775/googles-voice-generating-ai-is-now-indistinguishable-from-humans/, [abgerufen am 18.11.2019]

[4] Mit künstlicher Intelligenz 220.000 Euro erbeutet, https://www.golem.de/news/social-engineering-mit-kuenstlicher-intelligenz-220-000-euro-erbeutet-1909-143638.html, [abgerufen am 18.11.2019]

[5] Hackers Finally Post Stolen Ashley Madison Data, https://www.wired.com/2015/08/happened-hackers-posted-stolen-ashley-madison-data/, [abgerufen am 18.11.2019]

[6] P. Aichroth, H. Lukashevich. Audioforensik, partielles Audio-Matching und Audio Phylogenie-Analyse: Technologien für Medienverifikation und Medienmanagement. FKT 9/2019

[7] P. Aichroth. Hybride, datenschutzfreundliche Empfehlungssysteme – Mehr als nützlich. FKT 11/2019.

[8] European Commission. Ethics guidelines for trustworthy AI, https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai, [abgerufen am 18.11.2019]

[9] Google. Facets - know your data, https://pair-code.github.io/facets/, [abgerufen am 18.11.2019]

[10] What If...you could inspect a machine learning model, with minimal coding required?, https://pair-code.github.io/what-if-tool/, [abgerufen am 18.11.2019]

[11] A toolbox to iNNvestigate neural networks' prediction, https://github.com/albermax/innvestigate [abgerufen am 18.11.2019]

[12] Open Data Institute. What is the Data Ethics Canvas?, https://theodi.org/article/data-ethics-canvas/, [abgerufen am 18.11.2019]

[13] Ethical OS Toolkit, https://ethicalos.org/, [abgerufen am 18.11.2019]

[14] REWIND Project, https://sites.google.com/site/rewindpolimi/home, [abgerufen am 18.11.2019]

[15] https://towardsdatascience.com/5-types-of-bias-how-to-eliminate-them-in-your-machine-learning-project-75959af-9d3a0, [abgerufen am 18.11.2019]

[16] Attacking discrimination with smarter machine learning, https://research.google.com/bigpicture/attacking-discrimination-in-ml/, [abgerufen am 18.11.2019]

[17] T. Hochma. The Ultimate List of Cognitive Biases: Why Humans Make Irrational Decisions, https://humanhow.com/en/list-of-cognitive-biases-with-examples/, [abgerufen am 18.11.2019]

[18] Center for Data Innovation. Who Is Winning the AI Race: China, the EU or the United States?, https://www.datainnovation.org/2019/08/who-is-winning-the-ai-race-china-the-eu-or-the-united-states/, [abgerufen am 18.11.2019]