Beispiele für problematische KI-Anwendungen

Karl Hans Bläsius

Hochschule Trier, Fachbereich Informatik
Schneidershof
54293 Trier
E-Mail: blaesius@hochschule-trier.de

Schlüsselwörter

Künstliche Intelligenz, Überwachung, Autonome Waffen, militärische Frühwarnsysteme

ABSTRACT

In diesem Artikel werden drei Beispiele für problematische Anwendungen der Künstlichen Intelligenz (KI) vorgestellt. Bei diesen Beispielen „Bestimmen von persönlichen Eigenschaften“, „autonome Waffensysteme“ und „computerstützte Frühwarn- und Entscheidungssysteme“ geht es um potenziell erhebliche Auswirkungen für die Menschheit. Um einerseits nicht nur einen negativen Blick auf KI-Entwicklungen zu werfen und um andererseits Hinweise auf Erkennungskriterien zu liefern, die zur Beurteilung von problematischen Anwendungen hilfreich sein können, werden auch drei Anwendungen mit eigener Beteiligung vorgestellt. Hierbei werden auch die Aspekte Vagheit und Unsicherheit behandelt, die auch bei den problematischen KI-Anwendungen relevant sind.

1. Einführung

In den letzten Jahren sind einige spektakuläre Erfolge der Künstlichen Intelligenz bekannt geworden. Insbesondere bei der Bilderkennung und dem Sprachverstehen gibt es große Fortschritte. Systeme wie DeepL zum automatischen Übersetzen von Texten liefern inzwischen sehr gute Ergebnisse. Die aktuellen Erfolge basieren vor allem auf „Neuronalen Netzen“ und „Deep Learning“.

Die sprachverstehenden Systeme stützen sich auf riesige Mengen an Mustern, aus denen passende Antworten erzeugt werden, ohne dass der Inhalt wirklich verstanden wird. Um ein gewisses Maß an Verstehen zu erreichen, wären umfangreiches allgemeines Weltwissen und „common sence reasoning“ (Schließen nach gesundem Menschenverstand) erforderlich. Nur in eingeschränkten Anwendungsbereichen ist es derzeit möglich, das erforderliche Wissen in einer geeigneten Form bereitzustellen.

Im Rahmen dieses Beitrags wird nur auf wenige Methoden der KI kurz eingegangen, um eine Grundlage für die Darstellung problematischer KI-Anwendungen zu legen. Dies betrifft logisches Schließen, einschließlich der Aspekte Unsicherheit und Vagheit, sowie Klassifikationsaufgaben.

Logische Sprachen und Kalküle sind wichtige Grundlagen für automatische Schlussfolgerungen bei KI-Anwendungen.

Dazu ein Beispiel:

$$\forall x \forall y \forall z \ (\text{ist\_kind\_von}(x,y) \ \wedge \ \text{ist\_kind\_von}(y,z)) \Rightarrow \text{ist\_enkel\_von}(x,z)$$

Also:
für alle x,y,z: wenn x Kind von y ist und y Kind von z ist,
dann ist x Enkel von z

Eine solche Formel der Prädikatenlogik kann als uneingeschränkt gültig betrachtet werden. In unserem Alltag sind viele Zusammenhänge aber unsicher oder vage. Dies wird an weiteren Beispielen verdeutlicht.

Betrachten wir folgende Regel:

$$\forall x \forall y \ (\text{ist\_auto}(x) \ \wedge \ \text{ist\_besitzer\_von}(y,x)) \Rightarrow \text{ist\_nutzer\_von}(y, x)$$

Also:
für alle x, y: wenn x ein Auto ist und y ist Besitzer von x,
dann ist y Nutzer von x

Eine solche Regel gilt nicht immer, es kann Ausnahmen geben. Der Nutzer eines Autos könnte ein Kind des Besitzers sein. Auch bei Firmen können Besitzer und Nutzer unterschiedlich sein. Eine solche Regel ist also unsicher, sie gilt mit einer gewissen Wahrscheinlichkeit w.

In der Praxis gibt es viele Zusammenhänge, die unsicher sind, also nicht uneingeschränkt gelten. Trotzdem sind auch hier Schlussfolgerungen möglich und für bestimmte Problemlösungen notwendig. In der KI sind verschiedene Methoden zur Behandlung von Unsicherheiten entwickelt worden.

Besonders wichtig sind Methoden des probabilistischen Schließens. Hierbei werden numerische Werte für die Gültigkeit von Formeln verwendet, die dann beim

Schlussfolgern miteinander verrechnet werden. Verschiedene Wahrscheinlichkeitsmodelle unterscheiden sich darin, wie Formeln verknüpft und wie die Wahrscheinlichkeitswerte dann verrechnet werden.

In vielen Fällen kann Unsicherheit auch so behandelt werden, dass zunächst eine normale, „typische" Regelanwendung erfolgt. Typisch ist, dass der Besitzer eines Autos auch ein Nutzer dieses Autos ist. Solange nichts Gegenteiliges bekannt ist und kein Widerspruch entsteht, kann ein entsprechender Schluss gezogen werden. Im Falle eines Konfliktes müssen dann geeignete Maßnahmen zur Auflösung des Konfliktes getroffen werden. Auch für diese Art von Schlussfolgerungen gibt es unterschiedliche Methoden.

Unabhängig vom gewählten Verfahren ist die Behandlung von Unsicherheiten recht komplex und die Schlussfolgerungen sind auch unsicher, das heißt diese können falsch sein. Falsche Annahmen und falsche Schlussfolgerungen führen häufig zu Inkonsistenzen. In diesen Fällen können Korrekturmaßnahmen vorgenommen werden.

Weiteres Beispiel, gegeben sei z.B. folgender Zusammenhang:

Wenn x ein schweres Auto ist, dann benötigt x viel Kraftstoff.

Die Frage ist hier: was bedeutet „schwer", was bedeutet „viel"? Die Prädikatenlogik ist eine zweiwertige Logik, d.h. es gibt nur die Wahrheitswerte „wahr" und „falsch". In diesem Beispiel können die Aussagen „x ist ein schweres Auto" und „x benötigt viel Kraftstoff" nicht einfach mit den Wahrheitswerten wahr und falsch belegt werden. Diese Eigenschaften sind vage, gelten in bestimmtem Maße. Zur Darstellung solcher Aussagen ist eine zweiwertige Logik ungeeignet. Solche Aussagen können z.B. mit Fuzzy-Logic behandelt werden. Hierbei ist der Wahrheitswert ein beliebiger Wert (reelle Zahl) aus dem Intervall [0, 1], wobei 0 für falsch und 1 für wahr steht.

Bei vielen KI-Anwendungen geht es um Klassifikation. Hierbei besteht die Aufgabe darin, eine gegebene Situation oder ein Objekt auf sinnvolle Weise einer oder mehreren möglichen Klassen zuzuordnen. Eine gegebene Situation oder ein Objekt kann durch eine Reihe von Merkmalen (Symptomen) beschrieben werden. Aus einer eventuell größeren Menge von gegebenen Klassen (Diagnosen) ist dann eine oder es sind mehrere auszuwählen, zu denen das Objekt, bzw. die Situation passt. Zum Beispiel ist Klassifikation eine wesentliche Aufgabe bei medizinischen Expertensystemen. Aus gegebenen Symptomen (beobachtete oder durch Untersuchung bestimmte Merkmale) ist eine möglichst gute Diagnose (Klasse) zu bestimmen. Auch die Gesichtsverifikation und Gesichtsidentifikation sowie das Bestimmen persönlicher Eigenschaften aus Fotos oder Texten sind Klassifikationsaufgaben.

2. Beispiel-Anwendungen

Zur weiteren Verdeutlichung von Erkennungsaufgaben werden bespielhaft drei Anwendungsbereiche mit eigenem Entwicklungsbezug vorgestellt. Als Programmiersprache wurde jeweils Common-Lisp verwendet.

2.1. Klassifikation juristischer Texte

Eine einfache Vorgehensweise zur Klassifikation wird an einem Beispiel zur Analyse juristischer Texte nach Rechtsgebieten kurz beschrieben. Angenommen es gibt Rechtsgebiete wie z.B. Mietrecht, Familienrecht, Erbrecht, Handelsrecht, Patentrecht, Steuerrecht, usw. Ein gegebener Text, z.B. ein Gerichtsurteil, soll einem oder mehreren dieser Rechtsgebiete zugeordnet werden. Wenn es bereits eine Menge klassifizierter Texte gibt, kann diese als Lerngrundlage verwendet werden. Damit kann bestimmt werden, welche Begriffe und Gesetze bei einem bestimmten Rechtsgebiet sehr viel häufiger vorkommen als bei anderen Rechtsgebieten. So gefundene Kriterien können ein Gewicht erhalten, das sich u.a. aus den positiven (beim betreffenden Rechtsgebiet) und negativen (bei anderen Rechtsgebieten) Vorkommnissen errechnet. Beispielsweise könnten so Begriffe bestimmt werden wie „Ehe", „Kinder", „Unterhalt", die als Kriterien für Familienrecht verwendet werden können. Dies ist inhaltlich sicher sinnvoll. Allerdings könnten bei einem solchen Lernverfahren auch Kriterien gebildet werden, die von der Bedeutung her nicht zu dem Rechtsgebiet passen. Wenn ein Richter mit dem seltenen Namen Mustermann immer nur bei Familienrecht auftaucht, dann wird ein solcher Begriff mitgelernt. Außerdem kann es sein, dass bestimmte Wortfolgen häufig von einer Person (Verfasser eines Urteils) verwendet werden und sonst selten vorkommen. Auch eine solche Wortfolge kann mitgelernt werden.

Solche inhaltlich falschen Kriterien können trotzdem zu guten Erkennungsquoten führen, solange sich die Bedingungen nicht ändern. Solange also Richter Mustermann nur bei Urteilen zu Familienrecht mitwirkt, ist dieses Kriterium in Ordnung und verbessert die Erkennungsraten. Wenn der Richter Mustermann allerdings versetzt wird und danach für ein anderes Rechtsgebiet zuständig ist, kann ein solches Kriterium zu Fehlern führen, solange noch die alte Lerngrundlage verwendet wird. Wie schwerwiegend ein solches falsches Kriterium ist, hängt auch davon ab, wie viele Kriterien für eine Entscheidung zutreffend waren. In manchen Fällen sprechen sehr viele Kriterien für eine Klasse, dann ändert ein falsches Kriterium in der Regel nichts an einer korrekten Zuordnung. In anderen Fällen kann gerade ein solches falsches Kriterium zu einer Fehlklassifikation führen.

Die Entwicklung der Erkennungskriterien für diese Klassifikationsaufgabe führte bei manchen Rechtsgebieten zu einer großen Anzahl automatisch gelernter Kriterien für ein Rechtsgebiet, die unterschiedlich gewichtet waren. Die Folge war, dass bei manchen Beispielen auch viele Kriterien (bis zu 100) zu einem gewissen Grad zutreffend

waren und zu einem Klassifikationsergebnis geführt haben. In solchen Fällen ist es bei einer Fehlklassifikation zeitaufwändig, die ungünstigen Kriterien zu bestimmen, Gewichte anzupassen und so die Erkennung zu verbessern.

2.2. Rechnungsprüfung

Als weiteres Anwendungsbeispiel wird nun die automatische Rechnungsprüfung kurz vorgestellt. Hierbei geht es darum, eingehende Rechnungen (Papier oder elektronisch) automatisch zu analysieren und zu verarbeiten, was in manchen Fällen so weit gehen kann, dass solche Rechnungen automatisch verbucht und bezahlt werden, ohne dass ein Mensch zur Prüfung dazwischen geschaltet ist. Eine solche automatisierte Verarbeitung soll möglich sein, unabhängig von der Art der Rechnungsgestaltung und der dabei verwendeten Begriffe.

Im Falle von Rechnungen in Papierform müssen diese erst gescannt werden. Eine anschließende Texterkennung (OCR) liefert die Grundlage für eine Analyse. Hierbei reicht es nicht nur den Text zu kennen, sondern auch die Positionen der Wörter oder Einzelzeichen werden benötigt. Ausgangspunkt für eine Analyse kann eine Liste mit Einzelzeichen und Positionsangaben sein, wobei die Positionen (umschreibendes Rechteck) sich auf einen Koordinatenursprung (z.B. oben links) beziehen und in 1/10 mm oder Pixel angegeben werden können.

Automatisch zu erkennen sind in der Regel alle relevanten Inhalte einer Rechnung wie z.B. Absender, Empfänger, Belegdatum, Rechnungsnummer, alle Beträge, einschließlich Angaben zur Mehrwertsteuer, sowie die Einzelpositionen mit Angaben zu Menge, Einzelpreis und Betrag. Auch Kontoangaben und Zahlungsbedingungen können zu den Erkennungsaufgaben gehören.

Um zu erläutern, welche Erkennungskriterien relevant sind, wird beispielhaft die Erkennung des Nettobetrags (Gesamtbetrag ohne Mehrwertsteuer) betrachtet. Erkennungskriterien für diese Aufgabe sind unter anderem:

- Passt der Wortaufbau, also Ziffern, ein Komma mit zwei Nachkommastellen, eventuell Tausender-Trennpunkte, eventuell ein Minus-Vorzeichen?
- Steht der Betrag auf der letzten Seite, rechts unten?
- Kommen in der Nähe Schlüsselwörter vor, z.B. „Nettobetrag“, „Rechnungssumme“?
- Stehen solche Schlüsselwörter links neben dem Betrag oder darüber?
- Welche Nachbarwörter (links, darüber) gab es vorher bei dem gleichen Absender?
- Steht der Betrag in einer Spalte mit weiteren Betragswerten?
- Sind die Elemente der Spalte rechtsbündig?
- Ist der Betragswert die Summe von Werten darüber?
- Gibt es in der Nähe weitere Wörter, zu denen eine MwSt-Berechnung passt?

Die meisten Erkennungsmerkmale sind unsicher und gelten nur mit einer gewissen Wahrscheinlichkeit. Dies gilt z.B. für Schlüsselwörter, die mehrdeutig sein können. So kann „Gesamtbetrag“ für den Netto-Betrag ohne MwSt. oder für den Brutto-Betrag mit MwSt. verwendet werden.

Das Kriterium der Nachbarschaft gilt eventuell nur in gewissem Maße (ist also vage). Dies kann den Abstand betreffen, aber auch einen Versatz bezüglich „neben“ oder „über“. Auch die Zugehörigkeit zu einer Spalte oder die Rechtsbündigkeit einer Spalte gelten eventuell nur in gewissem Maße. Dies kann z.B. durch Schrägeinzug beim Scannen verursacht sein. Um gute Erkennungsergebnisse zu erzielen, müssen Vagheit und Unsicherheit nicht unbedingt unterschieden werden. Zum Beispiel kann man bei der Rechtsbündigkeit festlegen, dass dieses Merkmal mit einer gewissen Wahrscheinlichkeit zutrifft, in Anhängigkeit davon, wie stark die Pixelabweichung ausfällt.

Die einzelnen Erkennungsmerkmale können unterschiedlich gewichtet werden, da sie unterschiedlich relevant für die Erkennung sind. Zum Beispiel sind bei der Betragserkennung die rechnerischen Zusammenhänge besonders relevant. Bei der Erkennung der Rechnungsnummer haben Schlüsselwörter in der Nähe ein größeres Gewicht. Bei der Erkennung werden für die einzelnen Felder Hypothesen auf Basis vieler unsicherer und gewichteter Merkmale gebildet. Mit bestimmten Formeln wird daraus eine Gesamtbewertung berechnet und schließlich im Vergleich der Alternativen eine Entscheidung getroffen.

Auch wenn hohe Erkennungsraten erzielt werden, sodass in manchen Fällen Rechnungen automatisch gebucht und bezahlt werden, ist die Erkennung grundsätzlich unsicher und es können falsche Ergebnisse vorkommen. Dies kann sogar dann passieren, wenn das System einen Erkennungswert als sehr sicher einstuft. Bei Anwendungen in großen Industrieunternehmen ist es tatsächlich vorgekommen, dass die automatische Erkennung zu einer vollautomatischen Verbuchung und Bezahlung führte, die Erkennung aber falsche Betragswerte lieferte, obwohl die Erkennungsergebnisse von dem System als sicher eingestuft wurden. Diese wenigen Fehlbuchungen wurden von den Anwendern in Kauf genommen, da in der Gesamtbilanz der positive Automatisierungseffekt immer noch überwog.

Besonders schwer zu erkennen sind Abschlagsrechnungen, wie sie im Bauwesen häufig vorkommen, Reisekostenabrechnungen mit komplizierten Provisionsberechnungen, sowie Rechnungen, die andere Rechnungen als Anlage enthalten (z.B. Reisekostenabrechnungen).

2.3. Klassifikation Arzt-, Zahnarztrechnungen

Im Rahmen eines Projekts für private Krankenversicherungen mussten Arztrechnungen und Zahnarztrechnun-

gen unterschieden werden. Ein wesentliches Erkennungskriterium war, dass bei Zahnarztrechnungen auch angegebenen ist, welche Zähne behandelt wurden. Deshalb kommt bei Zahnarztrechnungen der Begriff „Zahn" mehrmals vor. Dieses Kriterium führte aber in einem Fall (bei einer normalen Arztrechnung) zu einer Fehlklassifikation, da die Patientin mit Nachnamen „Zahn" hieß. Der Begriff „Zahn" kam in dieser Rechnung mehrfach vor, bei der Empfängerangabe, bei der Anrede und bei der Patientenangabe. Dieses Problem konnte dann leicht gelöst werden, indem genauer modelliert wurde, in welchem Kontext der Begriff „Zahn" vorkommen soll.

Allerdings zeigt dieses Beispiel, dass es nicht möglich ist, bei der Entwicklung eines solchen Systems alle potenziell vorkommenden Ausnahmesituationen zu kennen und zu berücksichtigen. Es können immer wieder Situationen auftreten, die so nicht bedacht wurden und zu falschen Entscheidungen führen. Dieses grundsätzliche Problem gilt auch für die nachfolgend beschriebenen problematischen KI-Anwendungen und lässt sich auch nicht mit automatischen Lernverfahren lösen, denn auch hierbei ist nicht sichergestellt, dass hinreichend viele Beispiele für solche speziellen Situationen in der Lerngrundlage enthalten sind.

3. Beispiele für problematische KI-Anwendungen

3.1. Bestimmen persönlicher Eigenschaften

Menschen hinterlassen im Internet eine riesige Menge an Daten in unterschiedlichen Formen. Dazu gehören verfasste Texte, Fotos, Ton- und Videoaufnahmen. Diese Daten werden automatisch analysiert, um hieraus weitere Informationen abzuleiten. Dies können auch persönliche Eigenschaften der betroffenen Personen sein.

Bezüglich der Bestimmung persönlicher Eigenschaften aus Texten oder Fotos ist besonders Michel Kosinski bekannt geworden. In Wang, Kosinski 2017 wird ein Verfahren beschrieben, um aus Fotos von Personen Eigenschaften wie die sexuelle Orientierung zu bestimmen. Bei Männern wird eine Trefferquote von 81%, bei Frauen von 71% angegeben. Auch andere Eigenschaften will Kosinski allein aus Fotos bestimmen können, wie z.B. Intelligenz oder kriminelle Ambitionen. Aus Internet-Spuren, wie z.B. Facebook-Likes, bestimmt Kosinski u.a. politische Einstellungen. Angeblich hat Cambridge Analytica diese Verfahren genutzt, um Wahlen zu beeinflussen. Viele weitere Forscher-Gruppen haben Anwendungen realisiert, um aus Fotos, Texten oder Sprachaufnahmen Charaktereigenschaften von Menschen abzuleiten. Das Literaturverzeichnis enthält Hinweise zu einigen dieser Arbeiten. Automatisch bestimmt werden können z.B. auch Abhängigkeiten von Drogen oder Alkohol, sowie die Neigung zur Depression oder auch andere psychische Erkrankungen. Veränderungen der Sprache lassen auch Rückschlüsse über die Wirksamkeit von Medikamenten zu. Solche Eigenschaften können z.B. aus Sprachsignalen bestimmt werden, wie sie von Systemen wie Alexa erfasst werden. Auch der aktuelle emotionale Zustand einer Person ist ein Ziel solcher Untersuchungen.

Für die Bestimmung solcher Merkmale kann es vielfältige Anwendungen geben. Psychische Erkrankungen lassen sich einfacher und vielleicht sogar zuverlässiger diagnostizieren, als dies durch Psychotherapeuten auf der Basis von Gesprächen oder Fragebogen möglich ist. Wenn Betroffene solchen Maßnahmen zustimmen, können diese Anwendungen hilfreich sein. Allerdings kann die automatische Erkennung eines Gefühlszustands ohne Kenntnis der Betroffenen auch von digitalen Medien für spezielle Werbemaßnahmen oder zur Manipulation verwendet werden. Auch in Zusammenhang mit Stellenbewerbungen werden bereits Stimmanalysesysteme eingesetzt, um Persönlichkeitsmerkmale eines Bewerbers zu erkennen.

Die Bestimmung persönlicher Eigenschaften aus Fotos, Texten oder Tonaufnahmen ist eine Klassifikationsaufgabe, wobei es in der Regel viele Kriterien gibt, die zu einer Entscheidung beitragen. Diese Kriterien sind rein statistischer Natur und kaum inhaltlich begründet, stellen also keine kausalen Zusammenhänge dar.

Die Algorithmen zur Bestimmung persönlicher Eigenschaften aus Fotos werden kritisiert, weil kausale Zusammenhang suggeriert werden, die es nicht gibt. Dies spielt aber keine Rolle, solange die neuen zu untersuchenden Fälle zur Lerngrundlage passen. Statistische Verfahren funktionieren, auch ohne dass kausale Zusammenhänge vorliegen oder bestimmt werden. Es gibt viele Anwendungen, bei denen sehr gute Ergebnisse auf der Basis einer statistischen Auswertung von Symptomen erzielt werden, ohne dass kausale Zusammenhänge vorliegen.

Es kann auch passieren, dass sich bei einer späteren Anwendung herausstellt, dass die Erkennung nicht funktioniert, da eine ungünstige Lerngrundlage verwendet wurde, aus der irrelevante Merkmale mit hohem Gewicht bestimmt wurden. Dies kann aber mit einer neuen Lerngrundlage korrigiert werden und ändert nichts an der prinzipiellen Anwendbarkeit solcher Verfahren.

Die Bestimmung persönlicher Eigenschaften ist unsicher und vage. Ähnlich wie bei den Begriffen „schwer" und „viel" ist eine „Neigung zur Depression" nicht nur wahr oder falsch, sondern eine solche Eigenschaft gilt in bestimmtem Maße. Ähnliches gilt für andere persönliche Eigenschaften, wie politische Einstellung, Neigung zur Kriminalität, usw.

Die Anwendung von Verfahren für die Bestimmung persönlicher Eigenschaften ist aus folgenden Gründen sehr problematisch:

Keine Unterscheidung Unsicherheit - Vagheit: Unsicherheit und Vagheit werden häufig nicht unterschieden. Stattdessen wird einfach von Erkennungsraten gesprochen. Wenn bei der automatischen Rechnungsprüfung

z.B. das vage Kriterium der Rechtsbündigkeit einfach als Unsicherheit behandelt wird, wenn also statt dem Kriterium „die Rechtsbündigkeit gilt in gewissem Maße“ das Kriterium „die Rechtsbündigkeit gilt mit gewisser Wahrscheinlichkeit“ verwendet wird, wobei abhängig vom Grad des Zutreffens eine Wahrscheinlichkeit für die Gültigkeit verwendet wird, dann ist dies kein Problem. Die Erkennung funktioniert noch genauso gut. Bei persönlichen Eigenschaften wie Neigung zur Depression ist die Situation anders. Es ist ein Unterschied, ob für eine Person als Ergebnis geliefert wird, dass sie mit hoher Wahrscheinlichkeit schwach depressiv ist oder dass sie mit geringer Wahrscheinlichkeit stark depressiv ist. Es wird in der Regel auch nicht möglich sein, für die Lerngrundlage vage Werte genau zu bestimmen, also für Personen festzulegen, in welchem Maße eine Eigenschaft wie Neigung zur Depression vorliegt. Stattdessen werden Personen einfach nur in zwei Kategorien (gilt - gilt nicht) eingeteilt, ohne zu bestimmen, in welchem Maße eine Eigenschaft gilt.

Schlüsse aus unsicherem Wissen sind unsicher: Die Ergebnisse der Bestimmung persönlicher Eigenschaften aus Fotos oder Texten sind immer unsicher. Unsicheres Wissen kann mit sicheren oder unsicheren Regeln verknüpft werden. Das Ergebnis ist in jedem Fall wieder unsicher. In einer Kette von Schlussfolgerungen können so weitere unsichere Daten erzeugt werden. Alle so erzeugten Daten können falsch sein. Wenn unsichere Ausgangsdaten falsch sind, sind auch alle daraus gezogenen Schlüsse ungültig. Wenn bei den Schlussfolgerungen keine Konflikte (Widersprüche) auftreten, kann die Ungültigkeit nicht automatisch festgestellt werden, egal welches Verfahren zur Behandlung von Unsicherheiten verwendet wird.

Keine Korrektur von falschen Werten: Auch beim Datenaustausch zwischen Unternehmen kann es unsichere Daten geben, wie z.B. bei der automatischen Erkennung und Verarbeitung von Rechnungen oder der Klassifikation von Gerichtsurteilen. Hierbei gibt es aber hinreichend viele Möglichkeiten für Plausibilitätsprüfungen. Des Weiteren werden bei solchen Anwendungen unsichere bzw. falsche Werte manuell überprüft und gegebenenfalls korrigiert. Bei der automatischen Bestimmung persönlicher Eigenschaften aus Texten oder Bildern wird eine Korrektur von falschen Werten kaum möglich sein. Die betroffenen Personen wissen in der Regel auch nicht, welche Informationen auf eine solche Weise bestimmt werden und können somit auch keine Korrektur veranlassen. Das Merkmal wird als gültig angenommen und angewendet. Die Analyse-Ergebnisse werden nicht auf Wahrheit überprüft, sondern werden selbst zur Wahrheit.

Kein Lernfortschritt: Lernende Systeme sind auf Feedback angewiesen. Da falsche Werte in der Regel nicht überprüft und korrigiert werden, gibt es auch keinen Lernfortschritt. Die Erkennungsraten bleiben auf dem Stand, der aus der ursprünglichen Datengrundlage resultierte.

Massenanwendung: Riesige Mengen von Texten, Fotos und Tonaufnahmen haben die großen Internetkonzerne z.B. auf der Basis von Alexa oder Siri gesammelt. Viele dieser Daten, z.B. Fotos sind im Internet auch frei verfügbar und können für entsprechende Analysen verwendet werden. Mit Hilfe von inzwischen sehr erfolgreicher Gesichtserkennung sind auch Fotos mit mehreren Personen (Gruppenaufnahmen) für diese Zwecke nutzbar. Das Missbrauchspotential durch Staaten und auch durch Unternehmen (Verkauf der Daten an andere Unternehmen, Diskriminierung bei Stellenbewerbungen und Bankgeschäften) ist sehr hoch.

Das Bestimmen von persönlichen Eigenschaften aus Texten, Bildern, Tonaufnahmen oder Videosequenzen ist daher äußerst problematisch und abzulehnen. Die Ergebnisse sind immer unsicher, d.h. sie gelten nur mit gewisser Wahrscheinlichkeit, also nur für einen Teil der Personen, denen sie zugeordnet sind. Außerdem wird kein Grad bestimmt, zu dem eine gewisse Eigenschaft zutrifft. Die Verfahren zur Bestimmung persönlicher Eigenschaften können massenhaft auf den Daten in sozialen Netzwerken angewendet werden und die Nutzer in Kategorien einteilen. Dies kann zu Vorurteilen und Diskriminierung führen. Automatisch bestimmte Eigenschaften wie z.B. sexuelle Orientierung, Neigung zur Kriminalität oder politische Präferenzen können zu erheblichen Nachteilen für die Betroffenen führen und in manchen Staaten sogar die Freiheit und das Leben bedrohen.

3.2. Autonome Waffensysteme

Große Fortschritte im Gebiet „Künstliche Intelligenz“ haben auch zu entsprechenden Fortschritten in der Militärtechnik geführt. Solche Entwicklungen können zu autonomen Waffen führen, wobei Entscheidungen über Leben oder Tod von Menschen durch Maschinen getroffen werden könnten. Auf der Basis einer automatischen Bilderkennung mit guter Objektklassifikation könnten feindliche Ziele automatisch identifiziert und attackiert werden. Autonome Waffensysteme werden von mehreren Staaten entwickelt und erprobt. Solche Entwicklungen betreffen verschiedene Waffenkategorien, wie Landfahrzeuge, Luftfahrzeuge und auch U-Boote. Am stärksten in der Diskussion sind hierbei Roboter und Drohnen.

Eine genaue allgemein akzeptiere Definition eines „autonomen Waffensystems“ gibt es bisher nicht. Ob ein Waffensystem autonom oder halbautonom agieren kann, hängt nicht nur von den technischen Eigenschaften des Waffensystems, sondern auch von der Komplexität der Einsatzumgebung und eventuell von (vorherigen) Interaktionen mit menschlichen Akteuren ab.

Besonders in der Diskussion sind tödliche autonome Waffensysteme. In der Literatur wird hierfür meist die Bezeichnung LAWS (lethal autonomous weapon systems) verwendet. LAWS werden von den Vereinten Nationen definiert als Waffensysteme, die menschliche

Ziele ohne menschlichen Eingriff aufspüren, erfassen und eliminieren.

Drei Typen von autonomen Waffensystemen werden unterschieden:

- In-the-Loop-Systeme
- On-the-Loop-Systeme
- Out-of-the-Loop-Systeme

Bei In-the-Loop-Systemen behält der Mensch die Entscheidung über die Zielführung und die Ausführung eines Angriffs. On-the-Loop-Systeme können selbständig agieren und die Operationen ausführen. Ein Mensch kann das Verhalten aber kontrollieren und bei Bedarf eingreifen. Out-of-the-Loop-Systeme agieren selbständig, sobald ein Start vollzogen ist. Menschen haben dann keine Kontroll- und Eingriffsmöglichkeit mehr. On-the-Loop-Systeme und Out-of-the-Loop-Systeme gelten als autonome Systeme, denn sie können eigenständig agieren.

Vielfach gefordert wird, dass ein Mensch die letzte Entscheidung haben muss, oft ausgedrückt durch „human in the loop" (Mensch in der Entscheidungsschleife). Hierbei stellt sich aber die Frage, inwieweit der Mensch die vorliegenden Informationen in der verfügbaren Zeit bewerten und damit eine geeignete Grundlage für seine Entscheidung haben kann. Es kann schwierig sein, die Vorschläge der Maschine rational in Zweifel zu ziehen. Kersten Lahl schreibt dazu (Lahl 2021, Seite 50): „Solange es nicht eine „starke KI" gibt, ist pro forma die Kontrolle des Menschen über den Einsatz von Waffen gegeben. Allerdings beruhigt dieser Befund nicht restlos, da er zugleich einer Illusion unterliegt. Denn die formale Kompetenz ist das eine, die tatsächliche Eingriffschance das andere. Je komplexer ein kollektiver Verbund teilautonomer Waffensysteme ist, desto unmöglicher wird es dem kontrollierenden Menschen, die „Black Box" zu durchschauen und Fehler oder Manipulation zu erkennen – also die von Algorithmen gelieferten Ergebnisse nachzuvollziehen, zu bewerten und notfalls auch zu korrigieren. In hochintensiven Lagen unter extremem Zeitdruck reduziert sich seine Rolle dann de facto auf eine Scheinkontrolle."

Die Befürworter von autonomen Waffen argumentieren, dass aufgrund höherer Zielgenauigkeit zivile Opfer leichter vermieden werden können. Insbesondere wird das Leben von Soldaten der eigenen Streitkräfte nicht gefährdet, wenn die autonomen Systeme eigenständig ohne Beteiligung von Menschen agieren können.

Kritiker befürchten, dass bei vollautonomen Waffensystemen unvorhersehbare Interaktionen zwischen den automatischen Systemen vorkommen und eine Kettenreaktion von autonom geführten Angriffen und Gegenangriffen auslösen können. In sehr kurzen Zeitabschnitten kann hierbei eine Eskalationsspirale entstehen, die von Menschen in der Kürze der Zeit nicht beherrscht werden kann. Dieses Risiko wird auch als „flash war" bezeichnet, in Analogie zu einen „flash crash", wobei es im Hochfrequenzhandel zwischen verschiedenen Algorithmen zu unvorhergesehenen Interaktionsprozessen kommen kann, die innerhalb von Sekunden zu Kursabstürzen und finanziellen Verlusten führen können.

Vollautomatische Waffensysteme gefährden Völkerrechtsvereinbarungen, des Weiteren kann es im Falle von Tötungen durch autonome Waffen schwer sein festzulegen, wer hierfür verantwortlich gemacht werden kann, es können Verantwortungslücken entstehen.

Im Bericht für den Bundestag zur Technikfolgen-Abschätzung für autonome Waffen schreiben die Autoren zum Zusammenhang zwischen autonomen Waffen und Nuklearwaffen: „Auf der globalen Ebene spielt das strategische Gleichgewicht zwischen den Nuklearwaffenstaaten nach wie vor eine herausragende Rolle. Es basiert wesentlich auf der gesicherten Fähigkeit eines Zweitschlags und der daraus resultierenden Abschreckung eines möglichen Erstschlags. Es wäre vorstellbar, dass sehr potente AWS (autonomous weapon systems) zukünftig als konventionelle Erstschlagwaffen zur Zerstörung gegnerischer Nuklearwaffenarsenale eingesetzt werden könnten, die mögliche Ziele (Raketensilos oder mit Nuklearwaffen bestückte U-Boote) selbstständig aufklären, in deren Nähe unentdeckt verweilen und auf Befehl koordiniert diese Ziele angreifen und zerstören. AWS könnten auch als Trägerplattformen für Nuklearwaffen verwendet werden, beispielsweise in Form von autonomen Unterwasserfahrzeugen. Diese könnten schneller, überraschender und koordinierter als bisherige Trägersysteme zuschlagen und vorhandene Verteidigungsmaßnahmen aushebeln. Eine solche Nutzung von AWS würde die strategische Stabilität massiv infrage stellen."

In Russell 2020 (Seite 122) wird argumentiert, dass autonome Waffensysteme als Massenvernichtungswaffen einzustufen sind, da sie „skalierbar" also in beliebig großer Anzahl produzierbar sind. Das heißt, auch kleine Staaten oder Terrororganisationen könnten solche Waffen in großer Zahl produzieren und einsetzen. Für Stuart Russell ist dies ein wesentliches Argument dafür, dass autonome Waffen verboten werden müssen.

Das „Future of Life Institute" wurde 2014 gegründet und hat das Ziel, existenzielle Risiken für die gesamte Menschheit zu verringern. Auf der Weltkonferenz für „Künstliche Intelligenz", der „International Joint Conference on Artificial Intelligence" wurde am 28.7.2015 ein offener Brief zu autonomen Waffen in einer Pressekonferenz vorgestellt, der von vielen führenden KI-Forschern unterschrieben wurde. Zur Unterstützung einer Kampagne zum Verbot autonomer Waffen hat das Future of Life Institute im November 2017 einen knapp achtminütigen Film „Slaughterbots" veröffentlicht (https://www.youtube.com/watch?v=9CO6M2HsoIA), in dem die Risiken autonomer Waffen verdeutlicht werden. Auch wenn der Film als Science Fiction angesehen werden kann, sind die Techniken für eine Realisierung

heute bereits vorhanden. Solche Waffen könnten in recht kurzer Zeit konstruiert und eingesetzt werden. Der Film zeigt eine Minidrohne, die ihren Weg mit automatischer Bilderkennung sucht und ein Ziel mit automatischer Gesichtserkennung identifiziert. Nach Erreichen des Ziels kann eine tödliche Sprengladung gezündet werden. Am Ende des Films spricht der renommierte KI-Froscher Start Russell. Er warnt eindringlich vor der Entwicklung von autonomen Waffen und schreibt in Russell 2020 (Seite 122), dass nach seiner Kenntnis das Schweizer Verteidigungsministerium bereits einen solchen „Slaughterbot" gebaut, getestet und für einsatztauglich befunden habe.

Seit 2014 gibt es auf UN-Ebene Gespräche, um Verhandlungen zu einem Verbot autonomer Waffen zu starten. Bisher gibt es aber kaum Fortschritte. Die großen Militärmächte blockieren Verbotsverhandlungen.

3.3. Frühwarn- und Entscheidungssysteme

Computergestützte Frühwarn- und Entscheidungssysteme basieren auf Sensoren, sehr komplexen Computersystemen und Netzwerken und dienen der Vorhersage und Bewertung von möglichen Angriffen durch Atomraketen. Dabei kann es zu Fehlalarmen kommen, die ganz unterschiedliche Ursachen haben können (z.B. Hardware-, Software-, Bedienungsfehler oder falsche Bewertung von Sensorsignalen). In Friedenszeiten und Phasen politischer Entspannung sind die Risiken sehr gering, dass die Bewertung einer Alarmmeldung zu einem atomaren Angriff führt. In solchen Situationen werden im Zweifelsfall Fehlalarme angenommen.

Die Situation kann sich drastisch ändern, wenn politische Krisensituationen vorliegen, eventuell mit gegenseitigen Drohungen oder wenn in zeitlichem Zusammenhang mit einem Fehlalarm weitere Ereignisse eintreten. Hierfür werden bei einer Bewertung Ursachen gesucht, d.h. es wird versucht, kausale Zusammenhänge zu finden. Wenn solche kausalen Zusammenhänge gefunden werden und logisch plausibel sind, besteht die große Gefahr, dass diese als gültig angenommen werden, d.h. dass die Alarmmeldung als gültig angenommen wird, auch wenn es um zufälliges zeitliches Zusammentreffen von unabhängigen Ereignissen geht.

Fehler können in einem komplexen System nie ausgeschlossen werden und können sowohl durch Menschen als auch durch Computer verursacht werden. Bei komplexen Anwendungen ist es technisch nicht möglich eine fehlerfreie Software zu realisieren. Selbst wenn eine Software mit Techniken der Programmverifikation als korrekt bewiesen wird, sind solche Beweise nur auf Basis einer formalen Spezifikation möglich, die aber selbst wieder Fehler enthalten kann. Ein wichtiges Mittel zur Fehlerreduzierung bei der Softwareentwicklung ist Testen. Aber das Testen eines Frühwarnsystems wird unter realen Bedingungen kaum möglich sein.

Des Weiteren werden die Risiken eines Atomkriegs aus Versehen nicht geringer, wenn es weniger Fehlalarme gibt, denn seltene Fehler sind schwer zu bewerten und werden eher ernst genommen. Wenn es in der Vergangenheit sehr viele Alarmmeldungen gab und diese sich alle als falsch herausstellten, ist die Wahrscheinlichkeit hoch, dass auch die nächste Alarmmeldung als Fehlalarm eingestuft wird. Wenn es also gelingt, Frühwarnsysteme so zu verbessern, dass Fehlalarme nur noch sehr selten auftreten, wird damit die Sicherheit nicht erhöht. Die nur noch selten vorkommenden Alarmmeldungen sind dann ungewöhnlich und schwer interpretierbar, werden also eher als gültig angenommen.

Die Gefahr eines Atomkriegs aus Versehen wird sich in Zukunft deutlich erhöhen. Der Klimawandel wird vermutlich dazu führen, dass verschiedene Regionen unbewohnbar werden und damit vermehrt Klimaflüchtlinge verursachen. Der verfügbare Lebensraum wird kleiner, wichtige Ressourcen, wie zum Beispiel Wasser, knapper. Dadurch wird es in Zukunft häufiger politische Krisen und eventuell sogar kriegerische Konflikte geben. Als Folge werden Raketenangriffsmeldungen deutlich gefährlicher.

Cyberattacken können gefährliche und unkalkulierbare Wechselwirkungen mit Frühwarnsystemen sowie den Nuklearstreitkräften erzeugen und damit das Risiko eines Atomkriegs aus Versehen erheblich erhöhen. Kritisch können hierbei das zeitliche Zusammentreffen eines Cyberangriffs mit einem Fehlalarm in einem Frühwarnsystem oder die Manipulation von Komponenten oder Daten eines Frühwarnsystems sein.

Das Ende des INF-Vertrages wird zu einem neuen Wettrüsten führen. Völlig unkalkulierbar sind hierbei die Auswirkungen der geplanten Bewaffnung des Weltraums, sowie die Entwicklung von Hyperschallwaffen, die offenbar schwer zu lokalisieren sind und die Vorwarnzeiten extrem verkürzen werden.

Anzahl und Vielfalt an Objekten im Luftraum werden weiter steigen (z.B. Drohnen, Satelliten, Hyperschallraketen). Die Bewertung von Sensorsignalen wird damit schwieriger und es werden immer mehr Verfahren der Künstliche Intelligenz erforderlich sein, um für gewisse Teilaufgaben Entscheidungen automatisch zu treffen. Auch die Weiterentwicklung der Waffensysteme mit höherer Treffsicherheit und kürzeren Flugzeiten wird zunehmend Techniken der Künstlichen Intelligenz erforderlich machen. Es gibt bereits Forderungen in Zusammenhang mit Frühwarnsystemen autonome KI-Systeme zu entwickeln, die vollautomatisch eine Alarmmeldung bewerten und gegebenenfalls einen Gegenschlag auslösen, da für menschliche Entscheidungen keine Zeit mehr bleibt.

Die für eine Entscheidung verfügbaren Daten sind jedoch vage, unsicher und unvollständig. Deshalb können auch

KI-Systeme in solchen Situationen nicht zuverlässig entscheiden. Helligkeit und Größe von Sensorsignalen sind vage Werte. Die Klassifikationsergebnisse zur Objekterkennung sind unsicher und es können auch wesentliche Informationen fehlen, z.B. wegen Störaktionen durch elektronische Kampfmittel.

Ein Testen solcher Erkennungssysteme unter realen Bedingungen ist kaum möglich. Auch wird es im Vergleich zu anderen KI-Anwendungen (z.B. autonomes Fahren) deutlich weniger „Lerndaten" geben, um die nötigen Erkennungskriterien zu erzeugen. Dies kann zu unvorhersehbaren Effekten führen, die eventuell von Menschen nicht bewertet und kontrolliert werden können. In der kurzen verfügbaren Zeit wird es in der Regel auch nicht möglich sein, Entscheidungen der Maschine zu überprüfen. Dem Menschen bleibt nur zu glauben, was die Maschine liefert. Aufgrund der unsicheren und unvollständigen Datengrundlage werden weder Menschen noch Maschinen in der Lage sein, Alarmmeldungen zuverlässig zu bewerten.

Um einen Atomkrieg aus Versehen bei einem Fehlalarm in einer Krisensituation zu verhindern, müssen alle beteiligten Personen in der sehr kurzen Entscheidungszeit nach geltenden Regeln und logisch vernünftig handeln. Es ist äußerst fraglich, ob dies immer gewährleistet werden kann. Unfälle kommen häufig vor, oft hat dabei ein Beteiligter Regeln verletzt oder unvernünftig gehandelt. Ein „Atomkrieg aus Versehen" ist nicht direkt vorhersehbar. Wie bei sonstigen Unfällen in technischen Systemen gibt es keine Vorwarnung. Wie ein „normaler Unfall" kann ein Atomkrieg aus Versehen plötzlich innerhalb weniger Minuten als Folge einer Eskalationsspirale und falscher Einschätzungen geschehen. Danach ist keine Korrektur mehr möglich. Bei normalen Unfällen werden hinterher oft Maßnahmen getroffen, um solche Risiken in Zukunft zu vermeiden. Nach einem atomaren Schlagabtausch wird es eine solche Zukunft kaum noch geben. Beim Atomkriegsrisiko können wir mit Maßnahmen zur Reduzierung dieser Risiken nicht warten, bis es einen ersten „Unfall" in Form eines „Atomkriegs aus Versehen" gegeben hat.

Das Problem automatischer Entscheidungen in Zusammenhang mit Atomwaffen wird in Timm et.al., 2020 behandelt. Zur Darstellung der Gefahren wurden diese Seiten eingerichtet: www.atomkrieg-aus-versehen.de. Der Unterpunkt des Buttons „Unterstützer" zeigt, dass die Warnungen auch auf Zustimmung bei KI-Experten stoßen.

4. Fazit

Die allermeisten KI-Anwendungen sind positiv und für den Menschen nützlich. Dazu gehören zum Beispiel Systeme die auf Basis automatischer Erkennungsverfahren wie der Rechnungserkennung Verwaltungsabläufe verbessern oder auf der Grundlage von Dokumentanalyse und Wissensmanagement wichtige Informationen besser zugänglich machen. Auch im medizinischen Bereich sind große Fortschritte zum Beispiel durch automatische Bilderkennung für die Diagnose erzielt worden.

Es darf aber nicht missachtet werden, dass es auch problematische KI-Anwendungen geben kann. Eine erfolgreiche KI-Forschung kann erhebliche Auswirkungen auf den Menschen und die Zukunft der Menschheit als Ganzes haben. Dies kann auch eine mögliche Superintelligenz betreffen, auf die hier nicht eingegangen wurde. Problematische KI-Anwendungen kann es in Zusammenhang mit der Überwachung und Manipulation von Menschen geben, wobei das automatische Bestimmen von Charaktereigenschaften besonders hohes Mißbrauchspotenzial hat. Bezüglich militärischer Anwendungen der KI werden vor allem mögliche Auswirkungen autonomer Waffensysteme diskutiert. Wenig bekannt sind die Gefahren, die von immer mehr KI in militärischen Frühwarnsystemen zur Erkennung von Angriffen mit Nuklearwaffen ausgehen können. Die Erwartungen, die bei Frühwarnsystemen von Teilen der Politik und des Militärs in die KI gesteckt werden, können dabei nicht erfüllt werden. Aufgrund einer unsicheren und unvollständigen Datengrundlage können weder Menschen noch Maschinen im Falle von Alarmmeldungen sicher entscheiden. In Zusammenhang mit den Nuklearstreitkräften hing das Leben der gesamten Menschheit in der Vergangenheit in einigen Fällen von der Entscheidung eines einzelnen Menschen ab. Weder ein Mensch noch eine Maschine sollte eine solche Fähigkeit haben.

LITERATUR

Benton Adrian, Mitchell Margaret, Hovy Dirk: Multitask Learning for Mental Health Conditions with Limited Social Media Data, Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2017

Grünwald Reinhard, Kehl Christoph: Autonome Waffensysteme – Endbericht zum TA-Projekt, Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag, Arbeitsbericht Nr. 187, Okt. 2020, https://dip21.bundestag.de/dip21/btd/19/236/1923672.pdf

Kosinski Michal, Stillwell David, Graepel Thore: Private traits and attributes are predictable from digital records of human behavior, PNAS, 2013

Lahl Kersten: Autonome Waffensysteme als Stresstest für internationale Sicherheitspolitik, Politikum, Heft 1, 2021, Seite 46-53

Reece Andrew G., Danforth Christopher M.: Instagram photos reveal predictive markers of depression. https://arxiv.org/pdf/1608.03282v2.pdf, 2017

Russell Stuart, Norvig Peter: Künstliche Intelligenz - ein moderner Ansatz, 3. Auflage, Pearson, 2012

Russell Stuart: Human Compatible – Künstliche Intelligenz und wie der Mensch die Kontrolle über superintelligente Maschinen behält. Mitp Verlag, 2020

Schuller Björn: Mensch, Maschine, Emotion: Erkennung aus sprachlicher und manueller Interaktion, VDM-Verlag, 2007
Schuller Björn, Batliner Anton: Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing, Wiley, 2013
Timm Ingo J., Siekmann Jörg, Bläsius Karl Hans: KI in militärischen Frühwarn- und Entscheidungssystemen, 2020, https://www.fwes.info/fwes-ki-20-1.pdf
Wang Yilun, Kosinski Michal: Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. Journal of Personality and Social Psychology, 2017
Wu Xiaolin, Zhang Xi: Automated Inference on Criminality Using Face Images, https://arxiv.org/pdf/1611.04135v1.pdf, 2016
Wu Xiaolin, Zhang Xi: Responses to Critiques on Machine Learning of Criminality Perceptions, https://arxiv.org/pdf/1611.04135v3.pdf, 2017

KONTAKT

Prof. Dr. Karl Hans Bläsius
Hochschule Trier, Fachbereich Informatik
https://www.hochschule-trier.de/informatik/blaesius/
E-Mail: blaesius@hochschule-trier.de