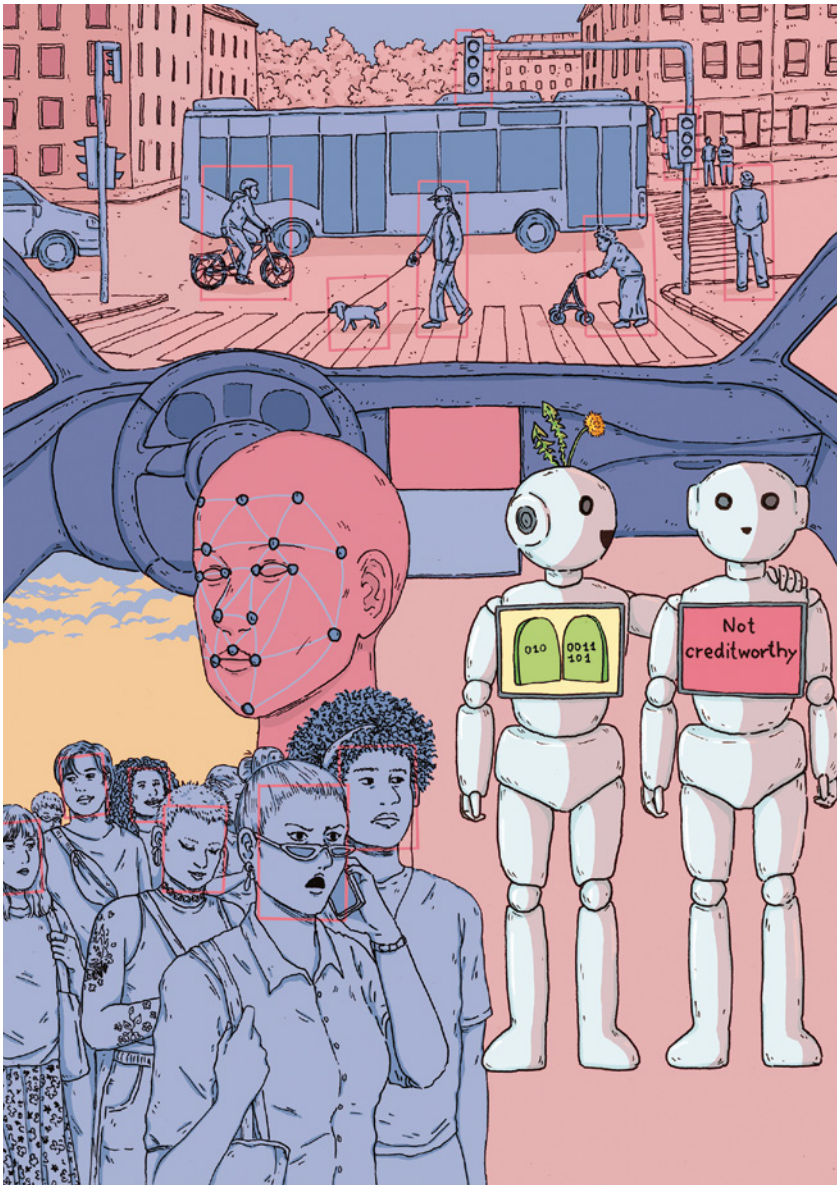
Gerechte Künstliche Intelligenz

Von Alexander Steen


010
0011
101
Not
creditworthy

Gerechte Künstliche Intelligenz

Von Alexander Steen

Methoden der Künstlichen Intelligenz (KI) finden breite Anwendung in Automatisierungsprozessen; und das sowohl in privatwirtschaftlichen, öffentlichen und individuellen Bereichen. Können wir sicherstellen, dass KI-Systeme gerechte Entscheidungen treffen?

Autonome Systeme, insbesondere basierend auf KI-Methoden (Künstliche Intelligenz; englisch AI, artificial intelligence), bestimmen bereits heute viele Bereiche des täglichen Lebens. Als autonomes System kann jeder automatisierte Prozess oder jede automatisierte Komponente komplexerer Strukturen angesehen werden: Während eines Suchvorgangs mit einer Suchmaschine werden – autonom – potenzielle Treffer selektiert und priorisiert. In Onlineshops werden aus der Menge aller angebotenen Produkte gewisse Produkte zuerst angezeigt. In sozialen Medien werden einer konkreten Benutzerin einige Beiträge angezeigt, andere aber nicht oder nur nach expliziter Suche. KI-Systeme sollen auf Internetplattformen zudem automatisiert Falschinformationen (Fake News) und Hassnachrichten (Hate Speech) erkennen und entfernen. Im Rahmen der Prüfung der Kreditwürdigkeit eines (potenziellen) Bankkunden können KI-Systeme eine Bewertung abgeben und damit die letztendliche Entscheidung der Bank signifikant beeinflussen. Verdächtige Zahlungsströme im Finanzbereich sollen ebenfalls automatisiert überwacht werden.

Das frühe Feld der KI wurde maßgeblich in den 1950er und 1960er Jahren geprägt. Erste KI-Systeme waren etwa sogenannte Expertensysteme, die explizit kodiertes Wissen als Regelsystem darstellen und automatisiert anwenden (zum Beispiel Klassifizierungs- oder Planungssysteme). Das Feld der KI versprach viel, konnte aber in folgenden Jahren wenig große Durchbrüche erzielen. Nach dieser Enttäuschung und dem dadurch ausgelösten »KI-Winter« wurden dann seit den 1990ern und 2000ern durch immer leistungsfähigere Computer-Hardware insbesondere lernbasierte Verfahren (Machine Learning) erfolgreich und erzielten erstaunliche Ergebnisse (zum Beispiel in der Echtzeiterkennung von Objekten in Videos oder Fotos oder der Verarbeitung natürlicher Sprache). Diese Verfahren funktionieren nach dem Prinzip des *induktiven Schließens* – also aus vielen einzelnen Beobachtungen möglichst allgemeingültige Verhaltensmuster zu produzieren, um automatisiert Entscheidungen zu treffen. Viele KI-Systeme basieren heutzutage auf genau diesem Prinzip. Können wir ihren Entscheidungen vertrauen, »lernen« sie gegebenenfalls etwas Falsches? Sind sie gerecht?

Gerechtigkeit als zentraler Aspekt sicherer KI

Das Schlagwort »sichere KI« (engl. trusted AI) bezeichnet Bestrebungen, KI-Methoden und autonome Systeme, die diese Methoden nutzen, so zu gestalten, dass unter anderem (ungewollte) Gefahren oder Fehlfunktionen, die durch den Einsatz dieser Systeme entstehen könnten, von vornherein ausgeschlossen werden oder sehr unwahrscheinlich sind.

Eine ganz konkrete und unmittelbare Gefahr geht zum Beispiel von autonomen (letalen) Waffensystemen aus. Ich bin zwar persönlich davon überzeugt, dass der Einsatz von autonomen letalen Waffensystemen prinzipiell verboten sein sollte; ebenso sinnig klingt es aber, diese Systeme so zu konstruieren, dass immerhin keine unbeteiligten Parteien (ungewollt) geschädigt werden oder andere gefährdende Nebeneffekte eintreten. Aber auch andere KI-Systeme können eine Gefahr für den gesamtgesellschaftlichen Frieden darstellen, und zwar dann, wenn diese Systeme nicht entsprechend unserer freiheitlich-demokratischen Grundordnung agieren und allgemein anerkannte gesellschaftliche Grundwerte missachten. So ist ein zentraler Aspekt von nachhaltig sicherer und vertrauenswürdiger KI, dass autonome Entscheidungen in gerechter Art und Weise getroffen werden, also etwa keine Gruppen oder Individuen grundsätzlich benachteiligt werden. Weitere abgeleitete Aspekte umfassen technische Robustheit, Transparenz und Erklärbarkeit. Die letztgenannte Anforderung ist beispielsweise nötig, um in Einzelfällen überhaupt beurteilen und prüfen zu können, ob und warum die konkrete Entscheidung des Systems obige Prinzipien untergräbt, um danach gegebenenfalls Gegenmaßnahmen ergreifen zu können.

Um diesen Anforderungen an KI-Systeme gerecht zu werden und einen (teilweise vorbeugenden) Rahmen für KI-Systeme in der Europäischen Union zu setzen, sieht der aktuelle Vorschlag der Europäischen Kommission für ein »Gesetz über Künstliche Intelligenz« (AI Act) vor, bestimmte Anwendungen von KI-Systemen vollständig zu untersagen und für KI-Systeme in Hochrisiko-Kontexten strenge Vorgaben zu

machen. In der Begründung des Gesetzesvorschlags heißt es dabei insbesondere, dass gewisse KI-Systeme »zu diskriminierenden Ergebnissen und zur Ausgrenzung bestimmter Gruppen führen [können]« und daher »die Menschenwürde und das Recht auf Nichtdiskriminierung sowie die Werte der Gleichheit und Gerechtigkeit verletzen [können]« (aus AI Act, Grund (17)). Verbotene Praktiken umfassen nach Artikel 5 des Entwurfs die unterschwellige Beeinflussung von Personen oder das Ausnutzen einer Schwäche oder Schutzbedürftigkeit von Personen(-Gruppen), sodass diesen ein physischer oder psychischer Schaden zugefügt werden kann. Kategorisch verboten sollen ebenso Klassifizierungs- und Bewertungssysteme sein, die die Vertrauenswürdigkeit von Personen bewerten, sodass durch ein bestimmtes soziales Verhalten Benachteiligungen entstehen. Ob der AI Act tatsächlich eine abschließende Lösung der Probleme bereitstellt, ist fraglich[1] – er stellt aber wohl einen wichtigen Schritt in die richtige Richtung dar.

Vertrauen in KI-Methoden

Selbstverständlich dürfen und sollen autonome KI-Systeme eine Person nicht aufgrund ihres Geschlechts, ihrer Abstammung, Hautfarbe, Religion oder weiterer Merkmale nachteilig behandeln. Wie dies sicherzustellen ist, ist allerdings bei induktiven KI-Systemen (z. B. basierend auf Machine Learning) alles andere als offensichtlich. Lernbasierte KI-Systeme könnten genau mithilfe der oben beschriebenen Merkmale Entscheidungsmuster lernen, sofern während des Lernprozes-

ses Zugriff auf diese Daten besteht. Aber auch wenn Daten über Geschlecht, Herkunft und andere Merkmale nicht explizit vorliegen, können personengruppenbezogene Benachteiligungen über Umwege erlernt werden, zum Beispiel wenn die Lerndaten entsprechende Verzerrungen (engl. *bias*) bereits indirekt enthalten. Mehrabi et al.[2] fassen verschiedene Arten von Beeinflussungen zusammen, die sowohl Datenerhebung, Datenqualität als auch deren algorithmische Verarbeitung betreffen.

Dass vom unreflektierten Einsatz von KI-Systemen reelle und ernstzunehmende Gefahren ausgehen, ist wenig strittig. Ein beeindruckendes Beispiel ist die sogenannte COMPAS-Software (Correctional Offender Management Profiling for Alternative Sanctions), die von US-amerikanischen Gerichten unter anderem zur Unterstützung von Entscheidungen zu vorzeitigen Haftentlassungen genutzt wird. Einer Studie nach waren dabei die von COMPAS geschätzten Rückfallquoten für afroamerikanische Häftlinge wesentlich höher als bei anderen Gruppen.[3] Hier hat also ein KI-System, das man als ungerecht bezeichnen kann, unmittelbaren Einfluss auf den Lebensverlauf von Individuen. Weitere Verzerrungseffekte können etwa bei Empfehlungssystemen oder bei Gesichtserkennungs-Software auftreten.

Sind symbolische KI-Methoden die Lösung?

Symbolische KI-Methoden sind historisch eher mit dem Prinzip der Deduktion verbunden. Bei der Deduktion werden aus allgemeinen und bekannten Gesetzmäßigkeiten (oder Regeln)

und aus konkreten Situationen spezielle Schlussfolgerungen oder Handlungsanweisungen für die Situation abgeleitet. Im Gegensatz zur Induktion wird also vom Allgemeinen auf Spezielles geschlossen (und nicht umgekehrt). Dies hat natürlich den Vorteil, dass aus allgemein anerkannten Prinzipien nur sich daraus ergebende Handlungsanweisungen abgeleitet werden können, womit ungewollte Schlussfolgerungen zumindest stark begrenzt werden. Allerdings benötigen deduktive Systeme aktuell noch mehr menschliche Vorarbeit, da die genutzten Regeln üblicherweise explizit von Menschen kodiert und eingepflegt werden. Auch von den Regeln nicht abgedeckte Situationen können üblicherweise schlichtweg nicht bearbeitet beziehungsweise eingeschätzt werden.

Können also symbolische KI-Methoden die Lösung für vertrauenswürdige, aber eben auch gleichzeitig leistungsfähige und flexible KI-basierte Systeme sein? Die Antwort liegt wahrscheinlich, wie so oft, irgendwo in der Mitte: Sowohl induktive als auch deduktive Methoden haben Vor- und Nachteile, könnten sich aber gut ergänzen. Der für mich aktuell überzeugendste Ansatz ist der eines *ethical governors* (zu Deutsch etwa »ethischer Regulator«):[4] Hierbei sollen (nicht unbedingt vertrauenswürdige) lernbasierte KI-Systeme weiterhin das autonome System steuern; dessen Anweisungen werden aber durch den ethical governor, einer Art ethischen Kontrollinstanz, geprüft und gegebenenfalls untersagt. Die deduktive Beurteilung der Kontrollinstanz basiert dann auf einer explizit kodierten Menge von regulatorischen Prinzipien (zum Beispiel ethische Prinzipien oder Rechtsnormen), deren Angemessenheit von Expertinnen und Experten ex ante untersucht und diskutiert werden kann. So sollen ungewollte

Entscheidungen von autonomen Systemen bereits untersagt werden, bevor sie zur Ausführung kommen und einen Schaden anrichten können.

Der Weg zur Konzeption und Umsetzung von nachhaltig vertrauenswürdigen KI-Systemen ist noch lang; aber ich bin mir sicher, dass eine enge Kooperation der heutzutage noch weitgehend getrennten »KI-Lager« (induktiv versus deduktiv) essenziell für die Erreichung dieses Ziels ist.

Literaturhinweise

1 Siehe z.B. die Zusammenfassung der kritischen Auseinandersetzungen mit dem AI Act unter artificialintelligenceact.eu/analyses/

2 Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. ACM Comput. Surv. 54, 6, Article 115 (July 2021), 35 pages. doi.org/10.1145/3457607

3 www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

4 R.C. Arkin, P. Ulam, and A. R. Wagner. Moral Decision Making in Autonomous Systems: Enforcement, Moral Emotions, Dignity, Trust, and Deception. Proceedings of the IEEE, 100(3):571–589, 2012. A. Dennis, M. Fisher, M. Slavkovik, and M. P. Webster. Formal Verification of Ethical Choices in Autonomous Systems. Robotics and Autonomous Systems, 77:1–14, 2016

Mojib Latif (Hg.)

Gerechtigkeit im 21. Jahrhundert

Zwischen Klimawandel und Künstlicher Intelligenz

HERDER

FREIBURG · BASEL · WIEN

Herausgeber: Prof. Dr. Mojib Latif, für die Akademie der Wissenschaften in Hamburg
Redaktion: Wolfgang Denzler, Akademie der Wissenschaften in Hamburg
Illustration: Luise Mirdita, https://www.luisemirdita.com
Finanziert aus Mitteln der Freien und Hansestadt Hamburg.

Akademie der Wissenschaften in Hamburg
Edmund-Siemers-Allee 1
20146 Hamburg
Deutschland
organisation@awhamburg.de
https://www.awhamburg.de/essays

Umschlaggestaltung: Verlag Herder
Umschlagmotiv: © Andriy Onufriyenko, © fhm,
© Guido Dingemans, De Eindredactie, © NikonShutterman,
© Olga Rolenko, © Paul Souders, © photo by Mike Lanzetta,
© Portra Images, © the_burtons, © Westend61/GettyImages,
© photosaint/AdobeStock

E-Book-Konvertierung: Carsten Klein, Torgau

ISBN Print 978-3-451-39584-0
ISBN E-Book (PDF) 978-3-451-83163-8
ISBN E-Book (EPUB) 978-3-451-83162-1