Themenblock 10 Robust/Hybrid/Robust AI 10_06Diskussion_Explainer

Diskussion zur Interpretierbarkeit von KI

Erarbeitet von
Marc Feger M.Sc.

Die Inhalte dieses Videos stellen keinesfalls eine Rechtsberatung in irgendeiner Form dar oder rechtliche Leitlinien. Ziels dieses Videos ist es, ein Problembewusstsein im Umgang mit KI zu schaffen und für rechtliche Fragen in diesem Kontext zu sensibilisieren. Vor dem Einsatz von KI-Systemen im Rahmen deines Projekts oder deiner Arbeit wende dich an die jeweiligen Fachstellen deiner Universität oder deines Unternehmens, um die rechtlichen Rahmenbedingungen für den Einsatz von KI zu besprechen.

Lernziele

- Du lernst die grundlegenden gesetzlichen Regelungen für KI kennen
- Du erkennst die Bedeutung ethischer und verantwortungsvoller Herausforderungen im Umgang mit KI
- Du kannst verstehen, wie Expertisen in rechtlichen Auseinandersetzungen um KI-Systeme angewendet werden

Projekt ai4all „KI für Alle"
HeiCAD –Heine Center for Artificial Intelligence and Data Science

Inhalt

Einstieg

Marc: Willkommen zu unserem heutigen Video, in dem wir anhand des praktischen Beispiels der Kreditvergabe die konkrete Anwendbarkeit von „Explainable AI“ diskutieren werden. Mein Name ist Marc und ein wichtiger Hinweis vorweg: Die Inhalte dieses Gesprächs stellen keine Rechtsberatung dar. Dieser Hinweis wird auch während Gesprächs als Erinnerung eingeblendet. Heute freue ich mich aber, Andreas Müller begrüßen zu dürfen. Andreas ist Jurist und Doktorand, der sich intensiv mit der Transparenz von künstlicher Intelligenz auseinandersetzt.

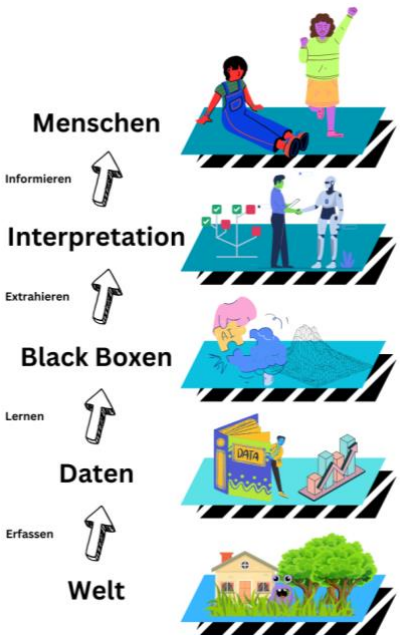
Interview

Quelle [1, 2, 3, 4]

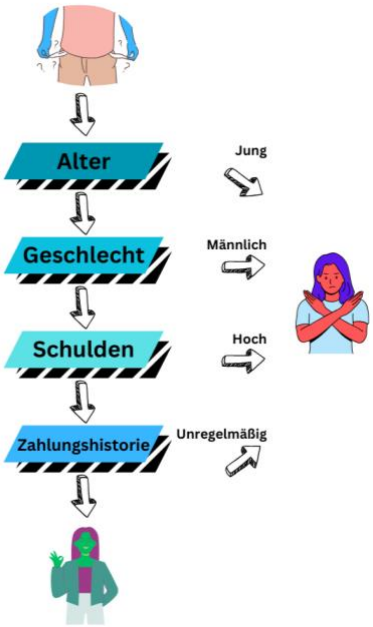
Marc: Andreas, schön, dass du da bist. Beginnen wir mit einer grundlegenden Frage. Andreas, könntest du uns erklären, was es mit der Kreditwürdigkeit auf sich hat und warum manche Menschen laut KI als nicht kreditwürdig eingestuft werden.

Andreas: Banken müssen im Rahmen ihrer geschäftlichen Tätigkeit, genauso wie jeder andere Betrieb, der untechnisch gesprochen in Vorleistung geht, zuvor eine Risikoabwägung für jedes einzelne Geschäft vornehmen. Bei der Frage nach der Kreditwürdigkeit geht es darum, dass ermittelt wird, ob ein potenzieller Kunde, der bei der Bank ein Darlehen aufnehmen will, das heißt also sofort eine große Menge Geld erhalten und dafür in Zukunft eine gewisse Menge Tilgung und Zinns Zahlung leisten will, die Gewehr dafür bietet, dieser Verpflichtung auch tatsächlich nachzukommen. Denn nur wenn davon auszugehen ist, dass die Bank ihr Geld auch zurückerhält, ist es wirtschaftliche Risiko vertretbar und ein Deal sinnvoll. Die Beurteilung der Kreditwürdigkeit hat sich seit Anbeginn des Kreditwesens immer weiter objektiviert. Während es historisch häufig vor allem eine Frage war, ob die Kreditnehmer dem Sachbearbeiter bei der Bank bekannt waren, richtet sich die Ermittlung der Kreditwürdigkeit im heutigen Geschäftsalltag nach den vorliegenden Daten. Diese Daten, unabhängig davon, ob sie bereits bei der Bank oder einem Dritten, wie etwa der Schufa, als spezialisierter Dienstleister vorliegen oder erst im Rahmen der Kreditanfrage als Selbstauskunft vom Kreditnehmer verlangt werden, können nicht nur manuell, sondern auch durch KI-Systeme analysiert werden. Die Fortschritte im Hinblick auf Objektivierung und Standardisierung unter anderem Getriebenen durch rechtliche oder anderweitige Vorgaben, führen zu immer geringeren Ermessensspielräumen und damit einem hohen Potenzial für Automatisierung. So kann es geschehen, dass eine KI entweder unmittelbar oder mittelbar den Ausschlag bei einer negativen Beurteilung der individuellen Kreditwürdigkeit gibt. KI wird zukünftig als effektive Entscheidungshilfe zunehmend eine Rolle spielen, nicht nur bei der Kreditvergabe, die lediglich als ein mögliches Beispiel aus einer ganzen Fülle von Einsatzszenarien dienen soll. Man denke nur an behördliche Verfahren, die Verarbeitung von Kunden im Einzelhandel, Steuerungen von Stromnetzen und industriellen Anlagen oder auch gar die Hilfe im gerichtlichen Verfahren. Unabhängig vom Zweck kann festgestellt werden, dass es aus verschiedenen wichtigen Gründen

wünschenswert ist, das KI transparent ist. Für diese zentrale Feststellung ist die folgende Grafik anschaulich. Es ergibt sich eine Art Datenverarbeitungshierarchie, ausgehend von der physischen Welt, bis hin zum menschlichen Verständnis. Die echte Welt bildet die Grundlage für die Welt der Daten, welche ihrerseits wieder Auswirkungen in der echten Welt hat. Um mögliche Black-Box-Effekte zu vermeiden und effektiven wie angemessen Einsatz von KI zu ermöglichen, bedarf es des Einsatzes von Ansätzen zur Interpretierbarkeit, welche die menschliche Einschätzung von KI, beziehungsweise das Verständnis darüber, unterstützen. Der Mensch ist dabei Dreh- und Angelpunkt aller Überlegungen. KI ist eine menschliche Entwicklung, KI lernt auf Basis von Daten, die Menschen liefern und KI wirkt sich auf menschliche Rechte aus. Gerade deswegen ist es wichtig, dass Entwickler von KI -Systemen sich früh mit den Auswirkungen ihrer Entscheidungen für die verschiedenen Stakeholder befassen.



Marc: Sehr interessant, könntest du an einem Beispiel veranschaulichen, wie Alter, Geschlecht, Schulden und Zahlungshistorie, die Kreditwürdigkeit beeinflussen und warum das problematisch sein kann?



Andreas: Und an der Fülle an Daten, die zur Ermittlung der individuellen Kreditwürdigkeit ran gezogen werden, sind die Zahlungshistorie und bestehende Schulden, die wohl sachnächsten Kriterien. „Fool me once, shame on me, fool me twice, shame on me" hat auch im Bankwesen Bedeutung. Wer schon einmal seine eingegangene Zahlungsverpflichtung nicht ernst genommen hat, könnte dies in Zukunft wieder tun. Dieses steigert das Risiko, ebenso wie ein besonders hoher Schuldenstand, vor allem wenn dieser ohne Gegenwerte besteht. Alter und Geschlecht hingegen erscheinen nicht unmittelbar relevant, können dies aber durchaus sein.
Junge Kreditnehmer könnten zur Selbstüberschätzung neigen, während das Gleiche bei männlichen Kreditnehmern der Fall sein könnte. Auch dies würde zu der negativen Risikobeurteilung führen oder jedenfalls beitragen. Probleme können hier vor allem aus der datengetriebenen Natur von KI-Systemen, die zur Abgabe derartiger Kreditwürdigkeitsbeurteilung eingesetzt werden, ergeben. Wenn beispielsweise zum Training ein Datensatz eingesetzt wird, in dem die versteckten Biases der Bankangestellten kodiert sind, können die Entscheidungen des KI-Systems unfair ausfallen. Ein solcher Bias kann etwa darin liegen, dass die Bankangestellten zuvor regelmäßig und ohne guten Grund die Kreditwürdigkeit von Menschen mit ausländisch klingenden Nachnamen negativ beurteilt haben. Denn dann dieses Merkmal auch für KI zum Anknüpfungspunkt entweder direkt oder auf Umwegen über Zwischeneigenschaften.

Marc: Das führt zu einer wichtigen Frage. Was bedeutet es, fair bewertet zu werden und wie können wir feststellen, ob unsere Kreditwürdigkeit fair bewertet wurde?

Andreas: Zunächst sollten wir zwischen zwei Ansätzen unterscheiden. Während fair ein recht dehnbarer Begriff und zudem in besonderem Maße subjektiv geprägt ist, gibt das Recht ebenfalls gewisse Maßstäbe für angemessene Entscheidungen vor. Diese lassen sich in der gesamten Rechtsordnung für viele spezifische Bereiche und auch ganz allgemein auffinden. Konkretisiert werden sie durch Rechtsnorm. Eine rechtmäßige Entscheidung ist daran zu sehen, dass keine rechtlichen Vorgaben verletzt werden, während eine faire Entscheidung anderweitig zu bewerten ist. Um dies umfassend bewerten zu können, ist ein Blick in das grundlegende System unerlässlich, weshalb Transparenz zum Gebot wird.

Marc: Können Betroffene rechtliche Schritte einleiten, wenn sie eine ungerechte Behandlung feststellen. Welche Lösungsansätze gibt es, um die Fairness der KI-Systeme zu verbessern?

Andreas: Richtig, auch in der rechtlichen Sphäre kann sich Transparenz auszahlen. Um auf unser Kreditwürdigkeitsbeispiel zurückzukehren, sollte ein KI-System aufgrund der zugrundeliegenden Trainingsdaten etwa objektiv übermäßigerweise das Alter zum Anknüpfungspunkt für die Kreditwürdigkeit machen. Beispielsweise, da es alle Kreditanfragen von über 65-Jährigen pauschal ablehnt, ohne sonstige Umstände einzubeziehen, kann es sich für die Betroffenen ein Anspruch nach § 19 des Allgemeinen Gleichbehandlungsgesetztes ergeben. Davon umfasst ist insbesondere Schadenssatz, der in der weiteren Folge auch Regressforderungen gegenüber dem Entwickler des fehlerbehafteten KI-Systems als Konsequenz haben kann. Um derartige Geschehnisse zu vermeiden, können Transparenzansätze bereits im Entwicklungs- und dem Erprobungsstadium systematische Probleme aufdecken, bevor es zu Rechtsverletzungen kommt. Aber auch in der Situation eines Gerichtsprozesses kann mit Transparenzmethoden ein Entlastungsbeweis geführt werden, falls die Diskriminierungsvorwürfe fälschlicherweise gegen ein KI-System erhoben werden. Die angemessene Implementation von Transparenzmethoden im Hinblick auf KI-Systeme schützt daher in allgemeiner und in individueller Hinsicht vor nachteiligen rechtlichen und tatsächlichen Folgen. Wichtige Lösungsansätze sind dabei die regelmäßige Überprüfung und Kalibrierung der KI, z. B. De-Biasing, wobei Explainability zum Finden und Begründen von Mängeln notwendig ist, sowie die intentionell erklärbar ausgestaltete Entscheidungsfindung durch und mittels KI.

Marc: Blicken wir in die aktuelle dynamische Lage. Welche juristischen Weichenstellungen sind notwendig, um die Entwicklung und Anwendung von KI in unterschiedlichen Bereichen verantwortungsvoll zu gestalten?

Andreas: KI wird schon heute von bestehenden rechtlichen Regelungen umfasst. Dennoch sind verschiedene, spezifisch auf KI zugeschnittenen Regelwerke absehbar. Diese basieren teilweise auf international abgestimmten Soft-Law, das verdichtet auch in den umsetzungsfähigen Rechtsnormen zu finden ist. Und sie sind sehr facettenreich, so dass nur auf die insbesondere für Transparenz wichtigen Grundlagen eingegangen werden soll. Zentrale Punkte, die von der KI der Zukunft in vielen zentralen Bereichen gefordert werden, sind unter anderem die Bereitstellungen von Handbüchern und damit Dokumentationen,

welche die Nutzung des Systems auch durch unerfahrene Anwender ermöglicht, in dem etwa die Funktionsweise beschrieben wird und eine Einschätzung der Grenzen des Systems angelegt ist, sowie die Durchführung von Risikoabschätzung beim Einsatz. Maßnahmen zur Risikominimierung erlangen ebenfalls Relevanz. Das Urheberrecht wird insbesondere durch die Marktpräsenz von generativer KI herausgefordert, welche einerseits auf Seite der Trainingsdaten urheberrechtlich geschütztes Material einbezieht, andererseits aber auch urheberrechtlich relevante Nachahmung als Output bereitstellen kann. In dieser Frage ist die Rechtslage bisher noch nicht vollständig klar, weshalb Reform oder jedenfalls Klarstellung zu erwarten sind. Bei Überlegungen zur Expandability sollten diese Aspekte daher zwingend berücksichtigt werden. Bei der Nachvollziehung der Entwicklung rechtlicher KI-Einigungsversuche ist zu betonen, dass sich das Verständnis von Transparenz gemeinsam mit der Technologie KI fortentwickelt. Es handelt sich bisher nicht um einen feststehenden definierten Begriff, sondern um ein überformendes Prinzip, welches durch konkrete Ansätze ausgefüllt werden kann und muss.

Marc: Stichwort Handbuch. Wo könnte ich mich denn bereits heute schon informieren, wenn ich mich für diese regulatorischen Vorgaben interessiere?

Andreas: Es gibt schon jetzt eine Vielzahl an Quellen, die sich zur Selbstinformation eignen. Die meisten Rechtsakte im Bereich der KI-Regulierung werden durch Begleitwerke ergänzt, die rechtliche Vorgaben für die Praxis handhabbar machen. Diese Vorgehensweise ist im Bereich des Datenschutzrechts bereits erprobt. Aufgrund der besonderen Datenbezogenheit von KI-Systemen bieten entsprechende Veröffentlichungen, etwa vonseiten der Datenschutzbeauftragten, wichtige Informationen. Im Hinblick auf Transparenz und Explanability, werden zukünftig die KI-Behörde der Europäischen Union sowie die Mitgliedstaatlichen KI-Behörden Empfehlungen abgeben. Auch sind delegierte Rechtsakte der Kommissionen absehbar. Relevant für die USA sind die Erarbeitungen des NIST, wie etwa dessen Risikomanagementrahmenwerk für KI. Ein abschließender Punkt soll die Empfehlung des Konzepts „Explanability by Design“ sein. Transparenz und Explanability haben, wie deutlich geworden ist, sowohl auf Makro als auch Mikroebene, gewichtige Vorteile. Bereits im Entwicklungsstatus sollten daher diese Belange entsprechend beachtet werden, da die Einbeziehung dann leichter fällt und die kontinuierliche Überwachung und Anpassung von KI-Systemen, wie sie nötig ist, erleichtert wird, so wie im Fall eines Rechtsstreits, massive Vorteile zu gewärtigen sind. Die individuelle, rechtskundige Beratung im Hinblick auf die jeweils zu beachtenden regulatorischen Anforderungen und empfehlenswerten Praktiken ist allerdings nicht zu ersetzen, auch und insbesondere nicht durch generative KI. Weshalb nur dazu geraten werden kann, einen solchen Anspruch zu nehmen, sollten sich Pläne zur wirtschaftlichen Verwertung von KI ergeben.

Marc: Andreas, vielen Dank für diese tiefgreifenden Einblicke. Es zeigt sich, dass das Thema KI und Explainability vielseitig ist und sowohl technische als auch rechtliche Herausforderungen bürgt. Für unsere Zuschauer, wir hoffen, diese Diskussion hat euch wertvolle Einblicke gegeben. Wenn ihr euch weiter informieren möchtet, schaut in die Beschreibung dieses Videos für Quellen, zum Datenschutz, dem EU-AI-Office und anderen relevanten Informationen. Denkt daran, „Explainability by Design“ ist der Schlüssel zu transparenten und fairen KI-Systemen. Danke, dass ihr heute dabei wart und bis zum nächsten Mal.

Quellen

Quelle [1] European Commission. (2024). Europäisches AI-Büro. https://digital-strategy.ec.europa.eu/de/policies/ai-office

Quelle [2] European Commission. (2024). Europäischer Ansatz für künstliche Intelligenz. https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence

Quelle [3] National Institute of Standards and Technology. (2024). AI Risk Management Framework. https://www.nist.gov/itl/ai-risk-management-framework

Quelle [4] Future of Life Institute. (2024). Das EU-Gesetz zur künstlichen Intelligenz: Aktuelle Entwicklungen und Analysen des EU AI-Gesetzes. https://artificialintelligenceact.eu/de/

Disclaimer

Transkript zu dem Video „Themenblock 10 Robust/Hybrid/Robust AI 10_06Diskussion_Explainer“, Marc Feger.