vertrauenswürdige künstliche Intelligenz (vertrauenswürdige KI)

Autor und Herausgeber

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland (ORCID 0000-0001-7538-0990)
Editors: Konstantina Olioumtsevits (ORCID 0000-0003-2697-8659), Ekkehard Schnoor (ORCID 0000-0001-8252-1499), Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Zusammenfassung

Vertrauenswürdige künstliche Intelligenz (vertrauenswürdige KI) bezeichnet ein KI-System, das Anforderungen erfüllt, die über das Liefern präziser Vorhersagen hinausgehen. Aus ingenieurwissenschaftlicher Sicht besteht das oberste Ziel des maschinelles Lernen (ML)s darin, das Label jedes beliebigen Datenpunkts präzise vorherzusagen, d. h. eine Hypothese mit kleinem Risiko zu lernen. Diese Sichtweise ist zu eng, da eine Methode des MLs in ein soziotechnisches System eingebettet ist, das menschliche Nutzende einschließt. Vertrauenswürdige KI formuliert daher zusätzliche Anforderungen an die Methode: menschliche Aufsicht, Robustheit, Datenschutz, Transparenz, Fairness (maschinelles Lernen), gesellschaftliches Wohlergehen und Rechenschaftspflicht. Leitlinien und Verordnungen auf der ganzen Welt legen diese Anforderungen fest, von den OECD-KI-Grundsätzen bis hin zu nationalen Vorschriften in China und Australien sowie Bundes- und einzelstaatlichen Vorschriften in den USA, und die Ethik-Leitlinien für eine vertrauenswürdige KI der Europäischen Kommission sind ein Beispiel unter ihnen. Ihre sieben zentralen Anforderungen sind für das KI-System als Ganzes formuliert, einschließlich der Menschen und der Organisation rund um die darin enthaltene Methode des MLs. Die Anforderungen lassen sich in Entwurfsentscheidungen für die Komponenten der empirische Risikominimierung (ERM) übersetzen: den Trainingsdatensatz, das Modell, die Verlustfunktion und den Algorithmus. Beispielsweise erfordert Robustheit ein Modell, das keine Hypothesenabbildung enthält, die zu empfindlich auf eine kleine Störung des Merkmalsvektors reagiert. Jede solche Anforderung ist eine messbare Eigenschaft der gelernten Hypothese oder der Methode, die sie gelernt hat: eine Schranke für die Änderung einer Vorhersage unter einer Störung des Merkmalsvektors, die Änderung der Vorhersage, wenn nur ein sensitives Merkmal umgekehrt wird, das Kontrafaktum, das eine Ablehnung erklärt, oder der Anteil von Grenzfällen, die an eine Person weitergeleitet werden.

Definition

Betrachtet sei eine Bank, die ein KI-System einsetzt, um über Kreditanträge zu entscheiden. Das System wendet eine Hypothese $\widehat{h}$, die aus den Datensätzen

abgeschlossener Kredite gelernt wurde, auf den Merkmalsvektor $\mathbf{x}$ einer antragstellenden Person an, der das Einkommen, das Geschlecht und weitere Attribute zusammenfasst. Der Kredit wird bewilligt, wenn die vorhergesagte Kreditwürdigkeit $\widehat{h}(\mathbf{x})$ einen Schwellenwert τ überschreitet.

Aus ingenieurwissenschaftlicher Sicht besteht das oberste Ziel einer Methode des maschinelles Lernen (ML)s darin, das Label jedes beliebigen Datenpunkts präzise vorherzusagen. Dieses Ziel wird als kleines Risiko von $\widehat{h}$ formalisiert, also als der erwartete Verlust bei einer antragstellenden Person, die aus derselben Wahrscheinlichkeitsverteilung gezogen wird wie die abgeschlossenen Kredite. Diese Sichtweise ist zu eng. Die Methode des MLs ist in ein soziotechnisches System eingebettet, das menschliche Nutzende einschließt: die Antragstellenden, über deren Kredite sie entscheidet, die Kreditsachbearbeitenden, die auf der Grundlage ihrer Vorhersagen handeln, und die Aufsichtsbehörde, die die Bank prüft. Diese Nutzenden verlassen sich nur dann auf die Entscheidungen, wenn das System Anforderungen erfüllt, die über die Genauigkeit seiner Vorhersagen hinausgehen. Eine kleine Änderung des Einkommens darf die Entscheidung nicht umkehren, das Geschlecht allein darf sie ebenfalls nicht umkehren, eine abgelehnte antragstellende Person muss erfahren, warum, und eine kreditsachbearbeitende Person muss das System überstimmen können. Vertrauenswürdige KI formuliert solche Anforderungen an eine Methode des MLs und bezeichnet ein KI-System, das sie erfüllt.

Solche Anforderungen werden in Leitlinien und Verordnungen festgehalten, die seit 2019 weltweit verabschiedet wurden. Beispiele sind die OECD-KI-Grundsätze (OECD, 2019), Australiens KI-Ethik-Grundsätze (Department of Industry, Science, Energy and Resources, 2024), das AI Risk Management Framework des US-amerikanischen National Institute of Standards and Technology (National Institute of Standards and Technology, 2023), Chinas vorläufige Maßnahmen für Dienste generativer Künstliche Intelligenz (KI) (Cyberspace Administration of China, 2023) und die Managementsystemnorm ISO/IEC 42001 (International Organization for Standardization and International Electrotechnical Commission, 2023). Diese Dokumente benennen weitgehend dieselben Anforderungen und unterscheiden sich in ihrer Rechtskraft. Die Ethik-Leitlinien für eine vertrauenswürdige KI der Hochrangigen Expertengruppe der Europäischen Kommission (High-Level Expert Group on Artificial Intelligence, 2019) dienen hier als Beispiel, da sie die Anforderungen explizit aufzählen. Ihre sieben zentralen Anforderungen (Key Requirements, KRs) sind Vorrang menschlichen Handelns und menschliche Aufsicht (KR1), technische Robustheit und Sicherheit (KR2), Schutz der Privatsphäre und Daten-Governance (KR3), Transparenz (KR4), Vielfalt, Nichtdiskriminierung und Fairness (maschinelles Lernen) (KR5), gesellschaftliches und ökologisches Wohlergehen (KR6) sowie Rechenschaftspflicht (KR7). Die zugehörige Bewertungsliste übersetzt jede Anforderung in Fragen, die der Anbieter und der Betreiber des KI-Systems über ihr System beantworten (High-Level Expert Group on Artificial Intelligence, 2020).

Diese Anforderungen sind für ein KI-System als Ganzes formuliert, wozu die Datenpipeline, die Benutzerschnittstelle und die betreibende Organisation

gehören, und nicht nur für die darin enthaltene Methode des MLs. KR6 und KR7 betreffen den Zweck und die Steuerung des Systems und haben kein Gegenstück in der Methode des MLs. KR1 bis KR5 lassen sich hingegen in Entwurfsentscheidungen für die Komponenten der empirische Risikominimierung (ERM) übersetzen: den Trainingsdatensatz $\mathcal{D}^{(\text{train})}$, das Modell $\mathcal{H}$, die Verlustfunktion und einen Algorithmus, der den durchschnittlichen Verlust über $\mathcal{D}^{(\text{train})}$ minimiert (Pfau and Jung, 2026). KR2 erfordert z. B. ein Modell $\mathcal{H}$, das keine Hypothesenabbildung enthält, die zu empfindlich auf eine kleine Störung des Merkmalsvektors reagiert. Das Ausschließen solcher Abbildungen aus $\mathcal{H}$ ist Modellpruning, einer der drei Wege, die die Regularisierung unterscheidet. Jede dieser Anforderungen ist eine messbare Eigenschaft der gelernten Hypothese $\widehat{h}$ oder der Methode, die sie gelernt hat.

Abb. 1 verortet KR1, KR2, KR4 und KR5 im Kredit-Scoring-Beispiel. Die horizontale Achse ist eine eindimensionale Skizze des Merkmalsvektors $\mathbf{x}$: das Einkommen, das Geschlecht und die weiteren Attribute einer antragstellenden Person, zusammengefasst in einer Koordinate. Die vertikale Achse ist die Kreditwürdigkeit, das Label y. Die ausgefüllten Kreise sind der Trainingsdatensatz $\mathcal{D}^{(\text{train})}$ aus $m = 5$ abgeschlossenen Krediten, die durchgezogene Kurve eine Hypothese $\widehat{h}$, die zur Veranschaulichung anstelle einer aus $\mathcal{D}^{(\text{train})}$ gelernten eingezeichnet ist, und die gepunktete Linie der Bewilligungsschwellenwert $\tau = 3.8$. Die antragstellende Person mit dem Merkmalsvektor $\mathbf{x}'$ (offenes Quadrat) erhält die vorhergesagte Kreditwürdigkeit $\widehat{h}(\mathbf{x}') = 2.94$ und wird abgelehnt.

KR2 verlangt, dass eine kleine Störung des Merkmalsvektors die Vorhersage nur wenig verändert (Pfau and Jung, 2026, Abschn. V). Die Störung von $\mathbf{x}'$ um $\delta = 0.5$ (horizontaler Pfeil) verschiebt die vorhergesagte Kreditwürdigkeit von 2.94 auf 3.19 (vertikaler Pfeil). Die Steigung von $\widehat{h}$ überschreitet niemals 1.225, sodass die Hypothese mit dieser Konstante Lipschitz-Stetigkeit ist: Keine Störung um δ verändert die Vorhersage um mehr als $1.225\,\delta$ (siehe Robustheit). KR2 deckt auch Störungen von $\mathcal{D}^{(\text{train})}$ ab, z. B. durch Datenvergiftung, wogegen die Stabilität der Methode des MLs schützt.

KR5 verlangt, dass die Entscheidung nicht von einem sensitives Merkmal s wie dem Geschlecht abhängt. Das offene Dreieck stellt dieselbe antragstellende Person dar, bei der lediglich das Geschlecht umgekehrt wurde, den Merkmalsvektor $\mathbf{x}^\star$. Die vorhergesagte Kreditwürdigkeit steigt auf $\widehat{h}(\mathbf{x}^\star) = 3.93$, also über τ, sodass das System den Kredit bewilligt (als Grenzfall wird er zudem an die kreditsachbearbeitende Person weitergeleitet): Die Entscheidung hängt vom Geschlecht ab, was das kontrafaktische Kriterium der Fairness (maschinelles Lernen) verletzt. Die Differenz $\widehat{h}(\mathbf{x}^\star) - \widehat{h}(\mathbf{x}') = 0.99$ misst den Verstoß. Sie wird durch eine Entwurfsentscheidung für eine Komponente der ERM geschlossen: das Modell (Ausschluss des sensitives Merkmals und der damit korrelierten Merkmale), die Verlustfunktion (ein Strafterm für die Differenz) oder $\mathcal{D}^{(\text{train})}$ (Neuabtastung der Gruppen) (Pfau and Jung, 2026, Abschn. VIII).

KR4 verlangt Transparenz: Der abgelehnten antragstellenden Person wird mitgeteilt, dass ein KI-System eingesetzt wurde, und sie kann eine Erklärung erhalten. Das Kontrafaktum $\mathbf{x}''$ (offene Raute), bei dem $\widehat{h}$ den Wert τ erreicht, ist

eine solche Erklärung. Es ist die kleinste Änderung des Einkommens bei festem Geschlecht, die die Entscheidung umkehrt, und es nennt der antragstellenden Person das Einkommen, bei dem der Kredit bewilligt würde. Die Umkehrung des Geschlechts $\mathbf{x}^\star$ kehrt die Entscheidung ebenfalls um, ist aber keine Änderung, die die antragstellende Person vornehmen kann; genau das macht sie zu einem Verstoß gegen KR5 statt zu einer Erklärung.

KR1 verlangt, dass eine Person die Entscheidungen beaufsichtigen und überstimmen kann. Das schattierte Band markiert eine vorhergesagte Kreditwürdigkeit innerhalb von $\varepsilon = 0.3$ um τ. Anträge in diesem Band werden an eine kreditsachbearbeitende Person weitergeleitet, die entscheidet; einer der fünf Kredite in $\mathcal{D}^{(\text{train})}$ fällt hinein. Ein breiteres Band leitet mehr Anträge an die sachbearbeitende Person weiter und überlässt dem System weniger Grenzfälle (siehe menschliche Aufsicht).

Schließlich hat KR3 keinen Platz in der Abbildung, da sie die Daten und nicht eine Vorhersage betrifft: Sie verlangt, dass die Datensätze abgeschlossener Kredite, bei denen es sich um personenbezogene Daten handelt, während des Trainings geschützt werden und nicht aus $\widehat{h}$ rekonstruiert werden können; differenzielle Privatheit (DP) und föderiertes Lernen (FL) sind hierfür Entwurfsentscheidungen für den Algorithmus und für $\mathcal{D}^{(\text{train})}$ (siehe Daten-Governance).

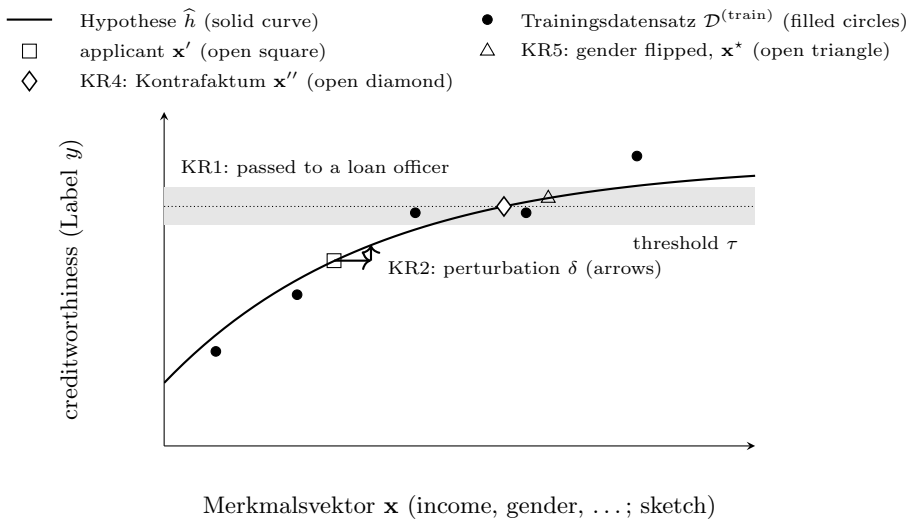

Abbildung 1: Vier Kernanforderungen für vertrauenswürdige KI im Kreditbewertungsbeispiel. Die horizontale Achse ist eine eindimensionale Skizze des Merkmalsvektors. Die durchgezogene Kurve ist eine Hypothese $\widehat{h}$, die zur Veranschaulichung neben den Trainingsdatensatz $\mathcal{D}^{(\text{train})}$ aus fünf abgeschlossenen Krediten (ausgefüllte Kreise) gezeichnet ist; ein Kredit wird bewilligt, wenn die vorhergesagte Kreditwürdigkeit den Schwellenwert τ (gepunktete Linie) überschreitet. Der Antragsteller mit dem Merkmalsvektor $\mathbf{x}'$ (offenes Quadrat) wird abgelehnt. KR1: Vorhersagen im schattierten Bereich werden an einen Kreditsachbearbeiter weitergeleitet. KR2: Eine Störung von $\mathbf{x}'$ um δ (horizontaler Pfeil) verändert die Vorhersage nur geringfügig (vertikaler Pfeil). KR4: Das Kontrafaktum $\mathbf{x}''$ (offene Raute), die kleinste Änderung des Merkmalsvektors, bei der $\widehat{h}$ den Wert τ erreicht, erklärt die Ablehnung. KR5: Derselbe Antragsteller mit umgekehrtem Geschlecht, $\mathbf{x}^\star$ (offenes Dreieck), wird über τ bewertet, sodass die Entscheidung vom Geschlecht abhängt; da der Wert im schattierten Bereich liegt, wird auch dieser Fall an den Kreditsachbearbeiter weitergeleitet

Über Leitlinien hinaus legt verbindliches Recht dieselben Anforderungen fest. Der KI-Verordnung ist ein Beispiel: Er macht KR1 bis KR4 für ein Hochrisiko-KI-System verpflichtend, eine Klasse, zu der auch das Kredit-Scoring natürlicher Personen gehört (European Parliament and Council of the European Union, 2024, Anhang III, Art. 9, 10 und 13–15). Chinas vorläufige Maßnahmen verpflichten die Anbieter von Diensten generativer KI zu Transparenz und Nichtdiskriminierung (Cyberspace Administration of China, 2023). In den USA erlassen einzelne Bundesstaaten Gesetze. Kalifornien gewährt Verbraucherinnen und Verbrauchern, die einer automatisierten Kreditentscheidung unterliegen, eine Benachrichtigung, eine Widerspruchsmöglichkeit und Zugang zur Logik der Entscheidung (California Privacy Protection Agency, 2025). Arizona schreibt vor, dass ein zugelassener Gesundheitsdienstleister jede Ablehnung eines Versi-

cherungsanspruchs überprüft, die eine KI aufgrund einer medizinischen Beurteilung aussprechen würde (Arizona Legislature, 2025).
Diese Anforderungen können miteinander in Konflikt stehen: Das Schließen der Geschlechterdifferenz erhöht den durchschnittlichen Verlust auf $\mathcal{D}^{(\text{train})}$, und das Rauschen, das DP dem Training hinzufügt, macht Erklärungen ungenauer (Pfau and Jung, 2026). Vertrauenswürdigkeit ist daher eine Menge gemessener Eigenschaften, die für die jeweilige Anwendung gegeneinander abgewogen werden, und keine einzelne Zahl.

Siehe auch: KI-System, KI-Verordnung, Hochrisiko-KI-System, Anbieter, Betreiber, menschliche Aufsicht, Robustheit, Stabilität, Datenvergiftung, Daten-Governance, DP, Transparenz, Erklärbarkeit, Kontrafaktum, Fairness (maschinelles Lernen), sensitives Merkmal, ERM.

Querverweise

Literatur

1. High-Level Expert Group on Artificial Intelligence (2020). The Assessment List for Trustworthy Artificial Intelligence (ALTAI): For self assessment. European Commission. URL: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

2. European Parliament and Council of the European Union (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). Official Journal of the European Union, L series, Jul. 12. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.

3. High-Level Expert Group on Artificial Intelligence (2019). Ethics Guidelines for Trustworthy AI. European Commission. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

4. International Organization for Standardization and International Electrotechnical Commission (2023). ISO/IEC 42001:2023 — Information technology — Artificial intelligence — Management system. International Standard, 1st edition. URL: https://www.iso.org/standard/42001.

5. National Institute of Standards and Technology (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. URL: https://doi.org/10.6028/NIST.AI.100-1.

6. Department of Industry, Science, Energy and Resources (2024). Australia's AI Ethics Principles. URL: https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles.

7. Arizona Legislature (2025). Claims; Prior Authorization; Conduct. URL: https://www.azleg.gov/legtext/57leg/1r/bills/hb2175h.htm.

8. Cyberspace Administration of China (2023). Interim Measures for the Management of Generative Artificial Intelligence Services. URL: https://www.chinalawtranslate.com/en/generative-ai/.

9. California Privacy Protection Agency (2025). CCPA Updates, Insurance, Cybersecurity Audits, Risk Assessments, and Automated Decisionmaking Technology (ADMT) Regulations. URL: https://cppa.ca.gov/regulations/.

10. OECD (2019). Recommendation of the Council on Artificial Intelligence. URL: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

11. Pfau and Jung (2026). Engineering Trustworthy AI: A Developer Guide for Empirical Risk Minimization. IEEE Trans. Artif. Intell., 7(5), pp. 2560–2576.