Transparenz

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

Transparenz ist eine zentrale Anforderung für vertrauenswürdige künstliche Intelligenz (vertrauenswürdige KI): Sie bezeichnet die Pflichten zur Schließung der Informationslücken zwischen dem Anbieter eines KI-Systems, dessen Betreiber und den Personen, die von dessen Vorhersagen betroffen sind. Bei maschinelles Lernen (ML)-Verfahren wird Transparenz häufig synonym zu Erklärbarkeit verwendet: Erklärungen, die zusammen mit den Vorhersagen einer gelernten Hypothese geliefert werden, ermöglichen es einem menschlichen Nutzer, diese Vorhersagen vorherzusehen. Der KI-Verordnung macht Transparenz zu einer verbindlichen Entwurfsanforderung. Nach dieser Verordnung muss der Anbieter ein Hochrisiko-KI-System so gestalten, dass dessen Betreiber die gelieferten Vorhersagen interpretieren kann. Personen, die mit einem KI-System interagieren, müssen über diesen Umstand informiert werden. Betroffene Personen können eine Erklärung von Entscheidungen erhalten, die auf den von einem Hochrisiko-KI-System gelieferten Vorhersagen beruhen. Die Verordnung verlangt zudem eine Dokumentation des Algorithmusentwurfs, der Trainings-Datensätze und der Zweckbestimmung eines KI-Systems sowie eine maschinenlesbare Kennzeichnung synthetischer Inhalte als künstlich erzeugt.

Definition

Betrachtet sei eine Bank, die ein KI-System zur Bewertung von Kreditanträgen einsetzt. Der Kreditsachbearbeiter sieht lediglich den Score-Wert, und der Antragsteller weiß möglicherweise nicht einmal, dass ein KI-System beteiligt war. Transparenz bezeichnet die Pflichten zur Schließung solcher Informationslücken; sie ist eine grundlegende Anforderung für vertrauenswürdige künstliche Intelligenz (vertrauenswürdige KI) (High-Level Expert Group on Artificial Intelligence, 2019).

Im Kontext von maschinelles Lernen (ML)-Verfahren wird Transparenz häufig synonym zu Erklärbarkeit verwendet (Gallese, 2023): Erklärungen, die zusammen mit den Vorhersagen einer gelernten Hypothese geliefert werden, ermöglichen es einem menschlichen Nutzer, diese Vorhersagen vorherzusehen (Jung and Nardelli, 2020). Einige ML-Verfahren bieten diese Form der Transparenz von Haus aus. Verfahren der Klassifikation quantifizieren die Konfidenz in eine einzelne Vorhersage über die Abstand des Merkmalsvektors von der

Entscheidungsgrenze. Ein flacher Entscheidungsbaum erfordert keine separat konstruierte Erklärung: Die Darstellung des Baums als Flussdiagramm expliziter, für den Menschen lesbarer Entscheidungsregeln ist bereits eine Erklärung jeder von ihm gelieferten Vorhersage (siehe interpretierbares maschinelles Lernen) (Rudin, 2019).

Im weiteren Rahmen von KI-Systeme geht Transparenz über die Erklärbarkeit hinaus: Sie umfasst die Unterrichtung von Personen darüber, dass ein KI-System im Einsatz ist, sowie die Bereitstellung von Informationen über die Grenzen, die Gesamtzuverlässigkeit und die Zweckbestimmung des Systems als Ganzes. Die nachstehend erörterten verbindlichen Transparenzpflichten sind diejenigen des KI-Verordnung; andere Rechtsordnungen sehen verwandte, aber eigenständige Pflichten vor, z. B. der Colorado Automated Decision-Making Technology Act (Colorado General Assembly, 2026) sowie das Gerichtsverfahren State v. Loomis zum Einsatz eines proprietären Rückfallrisiko-Scores bei der Strafzumessung (Supreme Court of Wisconsin, 2016). Die Verordnung verteilt ihre Pflichten auf drei Rollen: den Anbieter, der ein KI-System entwickelt, den Betreiber, der es einsetzt, und die Personen, die mit dem KI-System interagieren oder von dessen Ausgabe betroffen sind (siehe Betreiber). Der in der Verordnung verwendete Begriff *Ausgabe* bezeichnet die von einer gelernten Hypothese gelieferten Vorhersagen; im Folgenden wird der ML-Begriff *Vorhersage* verwendet. Wie Interpretierbarkeit und Erklärbarkeit, die jeweils auf einen bestimmten Nutzer oder eine bestimmte Nutzergruppe bezogen sind, ist auch die Transparenz relativ zu ihrem Adressaten: Was offengelegt werden muss, hängt von der jeweiligen Rolle ab.

Konkret muss der Anbieter gegenüber dem Betreiber ein Hochrisiko-KI-System so konzipieren und entwickeln, dass dessen Betrieb hinreichend transparent ist, damit die Betreiber dessen Vorhersagen interpretieren und angemessen verwenden können (European Parliament and Council of the European Union, 2024, Art. 13). Ein zweites Beispiel, die medizinische Diagnose, veranschaulicht die beiden Rollen: Der Anbieter ist z. B. ein Unternehmen, das ein Werkzeug zur Röntgenbildanalyse entwickelt, und eine Klinik, die dieses Werkzeug einsetzt, fungiert als Betreiber. Der Anbieter kann die Pflicht nach Art. 13 erfüllen, indem er das Werkzeug so gestaltet, dass es dem behandelnden Arzt das Konfidenzniveau für die von einer gelernten Hypothese gelieferten Vorhersagen offenlegt. Diese Pflicht betrifft die Bedeutung der Vorhersagen im Kontext ihrer Verwendung; das AI Risk Management Framework des US National Institute of Standards and Technology reserviert den Begriff *Interpretierbarkeit* für diese kontextbezogene Bedeutung (National Institute of Standards and Technology, 2023). Dieser Sprachgebrauch unterscheidet sich von der Interpretierbarkeit als dem Verständnis des Rechenprozesses des Verfahrens; Art. 13 setzt ein solches Verständnis nicht voraus.

Gegenüber Personen, die direkt mit einem KI-System interagieren, etwa einem Künstliche Intelligenz (KI)-gestützten Chatbot, muss der Anbieter das KI-System so gestalten, dass diese Personen über diesen Umstand informiert werden (European Parliament and Council of the European Union, 2024, Art. 50(1)). Anbieter von KI-Systeme, die synthetische Audio-, Bild-, Video- oder Textin-

halte erzeugen, müssen solche Inhalte in einem maschinenlesbaren Format als künstlich erzeugt kennzeichnen (European Parliament and Council of the European Union, 2024, Art. 50(2)). Diese Kennzeichnungspflicht gilt unabhängig davon, ob es sich bei dem KI-System um ein Hochrisiko-KI-System handelt; sie entfällt z. B., wenn das KI-System lediglich eine unterstützende Funktion für die Standardbearbeitung wahrnimmt (European Parliament and Council of the European Union, 2024, Art. 50(2)). Bei einem Deepfake muss der Betreiber darüber hinaus sichtbar offenlegen, dass der Inhalt künstlich erzeugt oder manipuliert wurde (European Parliament and Council of the European Union, 2024, Art. 50(4)).

Eine Person kann von einem KI-System betroffen sein, ohne direkt mit ihm zu interagieren. Ein Patient, dessen Diagnose durch ein KI-gestütztes System unterstützt wird, hat ausschließlich mit dem behandelnden Arzt zu tun. Gleichwohl muss der Betreiber den Patienten darüber informieren, dass in Bezug auf ihn ein Hochrisiko-KI-System eingesetzt wird (European Parliament and Council of the European Union, 2024, Art. 26(11)). Der KI-Verordnung räumt der betroffenen Person zudem ein Recht auf Erklärung ein: Der Betreiber muss auf Anfrage eine klare und aussagekräftige Erklärung zur Rolle des KI-Systems bei einer auf dessen Vorhersagen beruhenden Entscheidung bereitstellen (European Parliament and Council of the European Union, 2024, Art. 86). Bei der Kreditwürdigkeitsprüfung kann beispielsweise ein Kreditantragsteller, der mit einer nachteiligen automatisierten Entscheidung konfrontiert ist, nach diesem Recht vom Betreiber eine Erklärung der maßgeblichen Faktoren wie Einkommenshöhe oder Kredithistorie erhalten und diese nutzen, um die Entscheidung anzufechten (siehe Abb. 1).

Transparenz umfasst auch eine Dokumentation, in der Zweck und Entwurfsentscheidungen dargelegt werden, die dem KI-System zugrunde liegen. Für ein Hochrisiko-KI-System muss der Anbieter eine Dokumentation bereitstellen, die den Entwurf des Algorithmus und die für das Training verwendeten Datensätze abdeckt (European Parliament and Council of the European Union, 2024, Art. 11). Der Anbieter eines KI-Modell mit allgemeinem Verwendungszweck (GPAI-Modell)s muss zusätzlich eine hinreichend detaillierte Zusammenfassung der für das Training verwendeten Inhalte veröffentlichen (European Parliament and Council of the European Union, 2024, Art. 53). Datenblätter (Datasheets) für Datensätze (Gebru et al., 2021) und Modellkarten (Model Cards) (Mitchell et al., 2019) helfen Anwendern in der Praxis, die vorgesehenen Anwendungsfälle und Grenzen eines KI-Systems zu verstehen.

Konkret verortet Abb. 1 diese Pflichten für das Beispiel der Kreditwürdigkeitsprüfung in einem Streudiagramm von Datenpunkte des Trainingsdatensatzs zusammen mit einer gelernten Hypothese. Die durchgezogene Kurve stellt die gelernte Hypothese $\widehat{h}$ dar, die das KI-System anwendet: Sie bildet das Einkommen x eines Antragstellers auf einen vorhergesagten Kredit-Score $\widehat{h}(x)$ ab, und ein Kredit wird bewilligt, wenn der vorhergesagte Score einen Bewilligungsschwellenwert (gepunktete Linie) überschreitet. Die Pflicht nach Art. 13 betrifft die Verwendung dieser Abbildung: Hier muss der Betreiber in der Lage sein, die Vorhersage $\widehat{h}(x')$ für einen Antragsteller mit

dem Einkommen x' (offenes Quadrat) sowie deren Abstand vom Schwellenwert anzugeben.

Die ausgefüllten Kreise stellen das Trainingsdatensatz $\mathcal{D}^{(\text{train})}$ abgeschlossener Kredite dar, mit dem Einkommen als Merkmal x und dem Kredit-Score als Label y, aus dem $\widehat{h}$ gelernt wurde, z. B. mittels empirische Risikominimierung (ERM) durch Minimierung des durchschnittlichen Verlustes der Hypothese über $\mathcal{D}^{(\text{train})}$. Die Dokumentation nach Art. 11 deckt beide Komponenten dieses Trainings ab: den Datensatz $\mathcal{D}^{(\text{train})}$ und den Entwurf des Algorithmus, einschließlich des verwendeten Modells und des verwendeten Verlustes. Einkommenshöhen außerhalb des durch $\mathcal{D}^{(\text{train})}$ abgedeckten Bereichs (schattierter Bereich) stellen eine Einschränkung dar, die in der Dokumentation angegeben werden muss, da keine aus $\mathcal{D}^{(\text{train})}$ berechnete Schätzung des Risikos diesen Bereich abdeckt.

Das Recht auf Erklärung nach Art. 86 wird für eine einzelne Vorhersage ausgeübt: Der Antragsteller mit dem Einkommen x', dessen vorhergesagter Score unter den Schwellenwert fällt, kann nach dem Grund fragen. Eine mögliche Antwort ist ein Kontrafaktum: die kleinste Änderung des Merkmals, welche die Entscheidung umkehrt, also das Einkommen x'' (offene Raute, Pfeil), bei dem $\widehat{h}$ den Schwellenwert erreicht.

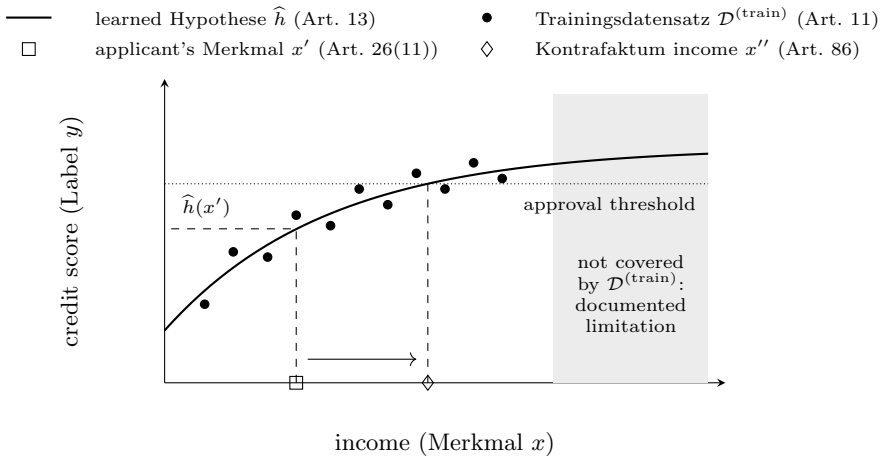


Abbildung 1: Transparenzpflichten des KI-Verordnung für eine ML-Anwendung zur Kreditwürdigkeitsprüfung. Die durchgezogene Kurve stellt die gelernte Hypothese $\widehat{h}$ dar, deren Vorhersagen der Betreiber interpretieren können muss (Art. 13): Die Vorhersage $\widehat{h}(x')$ für einen Kreditantragsteller mit dem Einkommen x' (offenes Quadrat) fällt unter den Bewilligungsschwellenwert (gepunktete Linie). Ausgefüllte Kreise sind die Datenpunkte des Trainingsdatensatzs $\mathcal{D}^{(\text{train})}$, aus dem $\widehat{h}$ mittels ERM gelernt wurde; der Anbieter muss diesen Datensatz und den Entwurf des Algorithmus dokumentieren (Art. 11), und der schattierte Bereich von Einkommen, der nicht durch $\mathcal{D}^{(\text{train})}$ abgedeckt ist, stellt eine Einschränkung dar, die in der Dokumentation angegeben werden muss. Der Antragsteller wird darüber informiert, dass das KI-System eingesetzt wird (Art. 26(11)), und kann eine Erklärung der Entscheidung erhalten (Art. 86), z. B. das Einkommen x'' (offene Raute, Pfeil) als Kontrafaktum, bei dem $\widehat{h}$ den Schwellenwert erreicht

Abb. 2 fasst diese Transparenzpflichten als Informationsflüsse zwischen dem Anbieter eines KI-Systems, dessen Betreiber und den seinen Vorhersagen ausgesetzten Personen zusammen.

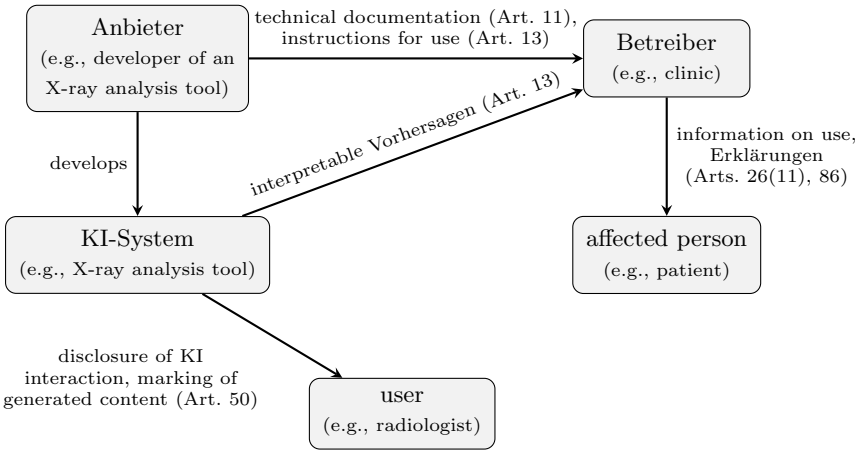


Abbildung 2: Transparenzpflichten des KI-Verordnung als Informationsflüsse, wobei die Rollen dem Beispiel der medizinischen Diagnose entnommen sind. Der Anbieter eines Hochrisiko-KI-Systems liefert die technische Dokumentation, die den Entwurf des Algorithmus und die Trainings-Datensätze abdeckt, zusammen mit einer Betriebsanleitung. Der Betreiber muss in der Lage sein, die Vorhersagen des KI-Systems zu interpretieren. Personen, die direkt mit dem KI-System interagieren, also dessen Nutzer, müssen über diesen Umstand informiert werden, und erzeugte Inhalte müssen eine maschinenlesbare Kennzeichnung tragen, aus der hervorgeht, dass sie künstlich sind (Art. 50). Eine betroffene Person, die nie mit dem KI-System interagiert, etwa ein Patient, dessen Diagnose das System unterstützt, wird vom Betreiber über dessen Einsatz informiert und hat ihm gegenüber Anspruch auf eine Erklärung einer nachteiligen Entscheidung (Art. 26(11), 86)

Siehe auch: vertrauenswürdige KI, Erklärbarkeit, Interpretierbarkeit, interpretierbares maschinelles Lernen, Erklärung, Kontrafaktum, Recht auf Erklärung, KI-Verordnung, Hochrisiko-KI-System, KI-System, GPAI-Modell, Anbieter, Betreiber, Wasserzeichen, Deepfake, Inhaltsprovenienz.

Querverweise

Literatur

1. Mitchell et al. (2019). Model Cards for Model Reporting. In: Proc. Conf. Fairness, Accountability, Transparency, pp. 220–229.

2. Colorado General Assembly (2026). Senate Bill 26-189 — Automated Decision-Making Technology. Colorado General Assembly, 2026 Regular Session, signed May 14, 2026. URL: https://leg.colorado.gov/bills/sb26-189.

3. Gebru et al. (2021). Datasheets for datasets. Commun. ACM, 64(12), pp. 86–92.

4. 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.

5. 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.

6. Jung and Nardelli (2020). An Information-Theoretic Approach to Personalized Explainable Machine Learning. IEEE Signal Process. Lett., 27, pp. 825–829.

7. Supreme Court of Wisconsin (2016). State v. Loomis, 2016 WI 68, 881 N.W.2d 749. Supreme Court of Wisconsin, opinion filed July 13, 2016. URL: https://www.wicourts.gov/sc/opinion/DisplayDocument.pdf?content=pdf&seqNo=171690.

8. 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.

9. Gallese (2023). The AI Act proposal: A new right to technical interpretability?. SSRN Electron. J.

10. Rudin (2019). Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead. Nature Mach. Intell., 1(5), pp. 206–215.