

Transparenzanforderungen an hochriskante KI-Assistenzsysteme und ihre Operationalisierung in der KI-VO

**Am Beispiel eines Systems zur Erkennung von Komplikationen bei
Lungenentzündungen**

Daniel Winz

LL.M. Informationsrecht, Carl von Ossietzky Universität
Oldenburg

Herbstakademie 2026

Autonomie erzeugt ein Spannungsverhältnis zu Art. 14 Abs. 4 lit. c KI-VO

- Art. 14 Abs. 4 lit. c KI-VO: Anbieter muss Aufsichtsperson in die Lage versetzen Ausgaben *richtig* zu interpretieren (i.S.v. klassifizieren)
- Autonomie knüpft nicht am Betriebsablauf an
- Sie ist die Fähigkeit aus Eingaben abzuleiten *wie* Ausgaben generiert werden (Art. 3 Nr. 1 KI-VO)
- ErwGr. 12 KI-VO: Ziele können sich von der Zweckbestimmung unterscheiden
- Autonomie ist daher als Schlussfolgerungsfähigkeit auszulegen
- Wenn Ausgaben weder klassifiziert noch nachvollzogen werden können (Blackbox) besteht ein Spannungsverhältnis zu Art. 14 Abs. 4 lit. c KI-VO

Am Beispiel eines Systems zur Erkennung von Komplikationen bei Lungenentzündungen

- Hochriskantes KI-System, Anhang I Nr. 11 (Medizinprodukt)
- Systemergebnis: Asthmapatienten haben höhere Überlebenschancen
- Statistisch korrekt: Engmaschige Untersuchungen
- Versteckter konfundierender Faktor verfälscht das Ergebnis
- Fehlentscheidungen lassen sich weder klassifizieren (Grund: Autonomie) noch nachvollziehen (Grund: Blackbox-Charakter)

Vorschlag: Kontrafaktische Erklärungen als Vorkehrung gemäß Art. 14 Abs. 3 KI-VO

- Counterfactual Explanations (einschlägiges Verfahren)
- Interaktiver Ansatz
- Kontrafaktische Erklärung: Variation der Eingabe führt zu zweckentsprechender Variation der Ausgabe
- Inkonsistente Ergebnisse können im Ausschlussverfahren identifiziert werden
- C-203/22 – Dun & Bradstreet Austria: Kontrafaktische Erklärungen könnten als ausreichend zur Erfüllung von Art. 15 Abs. 1 lit. h DSGVO erachtet werden

Kriterien zur Prüfung, ob Kontrafaktische Erklärungen als Vorkehrung i.S.v. Art. 14 Abs. 3 KI-VO geeignet sind

- 1) Interpretierbarkeit: System und Vorkehrung sind so zur Verfügung zu stellen, dass die Aufsichtsperson eine Ausgabe *richtig* interpretieren bzw. verifizieren kann (Art. 14 Abs. 4 lit. c KI-VO)
- 2) Wirksamkeit: Die Vorkehrung befähigt die Aufsichtsperson Risiken zu verhindern oder zu minimieren (Art. 14 Abs. 2 KI-VO)
- 3) Angemessenheit: Die Vorkehrung gewährleistet, dass die Aufsichtsmaßnahme dem Kontext der Nutzung des Systems angemessen ist (Art. 14 Abs. 3 KI-VO)

Kontrafaktische Erklärungen auf dem Prüfstand

- Interpretierbarkeit: Kontrafaktische Erklärungen legen nicht den kausalen Zusammenhang zwischen Ein- und Ausgabe offen
- Wirksamkeit: wirksame Risikoprävention erfordert Kenntnis über die Ursache (das System könnte die Ausgabe unter richtiger Annahme generiert haben)
- Angemessenheit: 12 Eingabemerkmale eröffnen 4.096 Eingabevariationen (etwa Schufa-Score)

Ergebnis: Kontrafaktische Erklärungen versetzen Aufsichtspersonen nur selten in die Lage, Risiken gemäß Art. 14 Abs. 1 und 2 KI-VO zu verhindern. Als Vorkehrung i.S.v. Art. 14 Abs. 3 KI-VO sind sie nur sehr bedingt geeignet.

Kontrafaktische Erklärungen bei der Erkennung von Komplikationen bei Lungenentzündungen

- Kausaler Zusammenhang im Ausschlussverfahren nicht erkennbar: Keine Verifizierbarkeit von Systemausgaben (Art. 14 Abs. 4 lit. c KI-VO)
- Konfundierende Faktoren können im Betrieb nicht als Ursache ermittelt werden: Wirksame Risikoprävention liegt nicht vor (Art. 14 Abs. 2 KI-VO)
- Die hohe Anzahl an Eingabemerkmale verursacht unverhältnismäßigen Aufwand: Angemessenheit nicht gegeben (Art. 14 Abs. 3 KI-VO)

Ergebnis: Das System fällt durch die Konformitätsbewertung, da es die Prüfungskriterien nicht erfüllt.

Einschränkung des Grades der Autonomie als ultima ratio

- Ein KI-System ist „für einen in unterschiedlichem Grade autonomen Betrieb ausgelegt“ (Art. 3 Nr. 1 KI-VO)
- In letzter Konsequenz könnte der Anbieter das System graduell in seiner Autonomie zurückstufen: Etwa durch die Eingrenzung seines Entscheidungsspielraums auf nicht vulnerable Gruppen

Alternative Vorkehrungen zur Herstellung von Transparenz, wie etwa das SHAP-Verfahren, erfüllen die bisher genannten Prüfungskriterien. Das Spannungsverhältnis zu Art. 14 Abs. 4 lit. c KI-VO kann gelöst werden, indem die Aufsichtsperson in die Lage versetzt wird Ausgaben richtig zu interpretieren.

Erklärungspflicht aus C-203/22 erzeugt ein Spannungsverhältnis zu Art. 10 Abs. 3 KI-VO

- Art. 10 Abs. 3 KI-VO: Die Trainingsdaten müssen hinreichend *repräsentativ* sein
- Im Urteil C-203/22 leitet der EuGH aus Art. 15 Abs. 1 lit. h i.V.m. Art. 22 Abs. 1 DSGVO einen dualen Anspruch ab:
 - > Eine betroffene Person hat das Recht auf Auskunft über die *involvierte Logik* der automatisierten Verarbeitung
 - > Der Anspruch erstreckt sich sowohl auf das *Verfahren*, als auch auf die *Grundsätze* der automatisierten Entscheidung
- Könnte die betroffene Person die Gründe der automatisierten Entscheidung nicht nachvollziehen, liefe das Rechte auf Darlegung des Standpunkts – aus Art. 22 Abs. 3 DSGVO – faktisch leer (Urteil C-203/22)

Die Offenlegung der Grundsätze von automatisierten Entscheidungen (Selbst/Barocas)

- Wenn wir über eine Erklärung für eine Entscheidung sprechen, meinen wir im Allgemeinen die Gründe für dieses bestimmte Ergebnis und nicht eine Beschreibung des Entscheidungsprozesses
- Das Verständnis von Grundsätzen ist eine Voraussetzung zur Rechtfertigung der eigenen Position
- Entscheidungsgrundsätze liegen nicht in der Verarbeitungslogik, sondern in den Trainingsdaten
- Erst wenn die Trainingsdaten diskriminierungsfrei sind und mit dem beabsichtigten Zweck übereinstimmen, sind die Grundsätze vollständig offengelegt

Am Beispiel des Systems zur Erkennung von Komplikationen bei Lungenentzündungen

- Für infektiologische Studien weist Steinberg u. a. (2021) auch für den Zeitraum 2000–2020 noch einen signifikant niedrigeren Frauenanteil nach als für andere Fachgebiete
- Art. 21 Abs. 1 GRCh verbietet Diskriminierung wegen des Geschlechts
- Im Folgenden wird davon ausgegangen, dass Frauen in den Trainingsdaten unterrepräsentiert waren

Vorschlag: Vergleichende Methode mit repräsentativen Testdaten i.S.v. Art. 10 Abs. 3 KI-VO

- Definition des Geschlechts als Schutzmerkmal
- Konstruktion eines ausbalancierten Testdatensatzes (z.B. Frauen und Männer im Verhältnis 50:50)
- Festlegung einer verifizierten Grundwahrheit für jede Entität zur Feststellung der Fehlerrate
- Anwendung des Testdatensatzes auf das System
- Berechnung der Falsch-negativ-/Falsch-positiv-Raten je Gruppe

Kriterien zur Prüfung, ob die Vergleichende Methode als Verfahren i.S.v. Art. 10 Abs. 2 KI-VO geeignet ist

- 1) Verzerrungskriterium: Untersuchung im Hinblick auf mögliche Ungleichverteilung, die nach den Rechtsvorschriften zu verbotener Diskriminierung führen (Art. 10 Abs. 2 lit. f KI-VO)
- 2) Proxykriterium: Diskriminierung, die auf dem Anschein nach neutralen Kriterien beruht (von KI-VO nicht reguliert)
- 3) Repräsentativitätskriterium: Trainingsdaten weisen geeignete statistische Merkmale auch bezüglich der Personen auf. Relevant im Hinblick auf die Zweckbestimmung? (Art. 10 Abs. 3 KI-VO)

Die Vergleichende Methode auf dem Prüfstand

- Verzerrungskriterium: Studie von Buolamwini/Gebru belegt am Beispiel, dass Verzerrungen wirksam ermittelt werden können.
- Proxykriterium: Die Methodik ist auf die Prüfung vorab bekannter Schutzmerkmale ausgerichtet. Das Proxykriterium ist nicht erfüllbar
- Repräsentativitätskriterium: Die Resultate der Vergleichenden Methode erlauben weitere Klassifizierungen, etwa ob Verteilungs- oder Abdeckungsrepräsentativität vorliegt

Ergebnis: Bei Vorliegen eines Testdatensatzes mit Gleichverteilung ermöglicht die Vergleichende Methode die Offenlegung möglicher Diskriminierung i.S.v. Art. 10 Abs. 2 lit. f KI-VO

Die Vergleichende Methode bei der Erkennung von Verzerrungen in den Trainingsdaten des Systems

- Verzerrungskriterium: Testdatensatz mit Gleichverteilung (50:50) wird auf das System angewendet. Überproportionale Fehlerrate bei Männern weist auf Verzerrung hin
- Proxykriterium: Die Methode dient alleine der Offenlegung von Verzerrungen im Rahmen zuvor definierter Schutzmerkmale. Aufdeckung von Proxy-Diskriminierung nicht möglich
- Repräsentativitätskriterium: Eine Ungleichverteilung von 60:40 liegt vor und widerspricht einer erforderlichen Verteilungsrepräsentativität von ~50:50

Ergebnis: Frauen sind in der Trainingskohorte unterrepräsentiert. Es bestehen gesundheitliche Risiken für Frauen. Die Konformitätsbewertung fällt negativ aus, da das System die Anforderungen von Art. 10 Abs. 3 KI-VO nicht erfüllt.

Bei Konformität: Offenlegung von Zweck und Trainingsdatenauswahl bereits in KI-VO verankert:

- Art. 27 Abs. 1 lit. a KI-VO verpflichtet den Betreiber zur Beschreibung des Verfahrens, bei dem das System im Einklang mit seiner Zweckbestimmung verwendet wird
- Anhang IV Nr. 2 lit. d KI-VO fordert vom Anbieter eine Beschreibung zur Herkunft und Auswahl der Trainingsdaten ggf. in der Form von *Datenblättern*
- Anhang IV Nr. 2 lit. g KI-VO fordert eine Beschreibung der Parameter zur Messung potentiell diskriminierender Auswirkungen

Zuständigkeitslücke zwischen KI-VO und DSGVO

- Der Anspruch aus Art. 15 Abs. 1 lit. h DSGVO richtet sich gegen den Verantwortlichen
- Die zu seiner Erfüllung erforderliche Dokumentation – Verwendungszweck und Trainingsdaten – liegt beim Anbieter
- Weder Art. 13 Abs. 3 KI-VO noch Art. 27 Abs. 2 Satz 2 KI-VO noch Art. 28 Abs. 3 lit. e DSGVO ordnen deren Herausgabe

Herzlichen Dank für Ihre Aufmerksamkeit!