AUSLEGUNG DES KI-VO-E ZUR EVALUATION VON VERFAHREN DER KÜNSTLICHEN INTELLIGENZ AM BEISPIEL DER AUTOMATISCHEN ANONYMISIERUNG VON GERICHTSENTSCHEIDUNGEN

Axel Adrian / Stephanie Evert / Philipp Heinrich / Michael Keuchen

Prof. Dr. Axel Adrian, Notar, Honorarprofessor für Rechtstheorie und Rechtsgestaltung, axel.adrian@fau.de

Prof. Dr. Stephanie Evert, Inhaberin des Lehrstuhls für Korpus- und Computerlinguistik, stephanie.evert@fau.de

Philipp Heinrich, M.Sc. und Doktorand, philipp.heinrich@fau.de

Michael Keuchen, Rechtsassessor und Doktorand, michael.keuchen@fau.de

Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Bismarckstr. 6, 91054 Erlangen, DE

Schlagworte: ***KI-Verordnung, Evaluation, automatische Anonymisierung von Gerichtsurteilen, Datenschutz, Legal-Tech***

Abstract: *Die Qualitätssicherung eines KI-Systems ist für den Grundrechtsschutz elementar. Der Entwurf zur KI-VO der EU versucht das sicherzustellen, indem verschiedene Anforderungen an KI-Systeme, wie die Erfüllung der Zweckbestimmung, Genauigkeit, Risikoabschätzung und die Robustheit abgesteckt werden. Aufgrund der Vielfalt an KI-Systemen, insbesondere mit einem hohen Risiko beim Einsatz in der Justiz, müssen die Anforderungen anwendungsspezifisch ausgelegt werden. Am Beispiel der automatischen Anonymisierung von Gerichtsentscheidungen zeigen wir, dass durch die Erstellung eines Goldstandards und einer strengen kontinuierlichen Evaluation den Anforderungen nachgekommen werden kann.*

1. Einleitung

1.1. Output und Qualität von KI-Systemen

„Errare humanum est“ – Irren ist menschlich. Und tatsächlich rechnen wir immer damit, dass andere Menschen Fehler machen: § 1 I StVO sieht bspw. vor, rücksichtsvoll zu fahren und mit Fehlern anderer Verkehrsteilnehmer zu rechnen; bei schwierigen Aufgaben wird das „Vier-Augen-Prinzip“ angewandt; etc. Werden Aufgaben an KI-Systeme übertragen, ist die Erwartungshaltung jedoch eine andere: KI-Systeme sollten nahezu unfehlbar sein und obwohl ChatGPT für viele Menschen einen „Quantensprung“ der Technik darstellte, so wurde schnell klar, dass auch „Large Language Models“ Schwachstellen haben. Zwar mag ChatGPT auf alle Fragen eine selbstbewusst formulierte Antwort haben, doch ist sie auch richtig?[1] Berichte von falschen Antworten und oft sogar erfundenen Angaben (sog. Halluzinationen) finden sich schnell.[2] Die Herausforderung besteht also darin, die Qualität des Outputs eines KI-Systems korrekt einschätzen zu können. Dafür bedarf es Angaben über das KI-System, um Ergebnisse einordnen zu können. Denn Systeme wie ChatGPT

[1] Auch ChatGPT ist keine erklärbare bzw. transparente KI, obwohl es seine Antworten meist ausführlich erläutert. Diese Erläuterungen sind lediglich eine zusätzliche, ebenfalls aus Trainingsdaten gelernte Funktion des KI-Systems. Sie stehen in keinem Zusammenhang mit den internen Prozessen, die zu der eigentlichen Antwort von ChatGPT geführt haben. Inwieweit diese Erläuterungen überhaupt richtig und nützlich sind, müsste erst durch eine separate Evaluation überprüft werden.

[2] Hartung, Smartlaw, ChatGPT und das RDG, RDi 2023, 209 (212).

verweigern nicht die Antwort, wenn sie nicht auf eine bestimmte Frage trainiert sind, sondern liefern dennoch eine (meist auf den ersten Blick überzeugend klingende) „Antwort".[3] Werden nun Rechtsfolgen an den Output eines solchen KI-Systems geknüpft, weil der Staat dadurch die Rechtsfindung vollzieht, können in einem erheblichen Maße Grundrechtsverletzungen eintreten. Die Qualitätssicherung eines KI-Systems ist für den Grundrechtsschutz elementar.

1.2. Entwurf zur KI-Verordnung als Ordnungsrahmen zur Qualitätssicherung?

Diese Gefahren anerkennend, befindet sich u.a. auf der europäischen Ebene der Entwurf der „Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für Künstliche Intelligenz (Gesetz über Künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union" vom 21.4.2021, samt überarbeitetem Ratsvorschlag vom 6.12.2022, („KI-VO-E")[4] in Entwicklung.[5] Der Beitrag wird zeigen, inwieweit die hier angerissenen Problemstellungen zur Qualitätssicherung und Qualitätsbestimmung für die Anwendung von KI-Verfahren im Bereich der Rechtswissenschaft adäquat vom KI-VO-E Anerkennung finden.[6] Dafür stellen sich Fragen, ob rechtliche Vorgaben aus dem KI-VO-E zu einer Evaluation zur Qualitätssicherung verpflichten, Vorgaben zur Genauigkeit von KI-Systemen machen oder sogar zur Offenlegung von Goldstandards oder Trainingsdaten verpflichten. Neben der rechtlichen Betrachtung einer Auswahl relevanter Vorgaben werden technische Verfahren zur Evaluation von KI-Systemen herangezogen und untersucht, ob der KI-VO-E den Stand der Technik adäquat abbildet. Dabei beschränkt sich der Beitrag auf von Justizbehörden eingesetzte KI-Systeme, welche natürlichsprachlichen Text verarbeiten. Zum einen scheint eine Qualitätssicherung beim staatlichen Einsatz zur Rechtsfindung wegen der erheblichen Grundrechtsrelevanz besonders wichtig. Zum anderen stellen sich bei der Verarbeitung von Bildern, biometrischen Daten, strukturierten Daten, Finanzdaten usw. ganz andere Anforderungen und Schwierigkeiten bei der Qualitätssicherung.

2. Vorgaben aus dem Entwurf zur KI-Verordnung

2.1. Hochrisiko-KI

Eingangs ist festzustellen, dass nach Art. 6 II, Anhang III Nr. 8 lit. a KI-VO-E KI-Systeme, die bestimmungsgemäß Justizbehörden bei der Ermittlung von Sachverhalten und Auslegung von Rechtsvorschriften oder bei der Anwendung des Rechts auf konkrete Sachverhalte unterstützen sollen, als Hochrisiko-KI einzuordnen sind.[7] Die Folgen dieser Einordnung sind insbesondere ein Risikomanagement-System (Art. 9 KI-VO-E), Anforderungen an Daten (Art. 10 KI-VO-E), eine technische Dokumentation (Art. 11 KI-VO-E), Transparenzanforderungen (Art. 13 KI-VO-E), eine menschliche Aufsicht (Art. 14 KI-VO-E) und Pflichten für Hersteller und Benutzer (Art. 16 ff. KI-VO-E). Grundlegend für die weitere Einordnung der mannigfaltigen Anforderungen an Hochrisiko-KI-Systeme sind die nachfolgenden Begriffsbestimmungen.

---

3 Hartung, Smartlaw, ChatGPT und das RDG, RDi 2023, 209 (212).

4 https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX%3A52021PC0206 – abgerufen am 4.3.2023.

5 Am 9. Dezember 2023 wurde die politische Einigung im Trilogue zwischen Parlament, Rat und Kommission erzielt. Nachfolgend wurde der Text in Interinstitutionellen Technischen Meetings (ITMs) weiter ausverhandelt. Eine Endversion wurde am 22. Jänner 2023 inoffiziell zugänglich gemacht. Die Ergebnisse konnten nicht mehr berücksichtigt werden.

6 Z.B. nach Gemin, Die Regulierung Künstlicher Intelligenz, ZD 2021, 354 ff. zeigt sich nach dessen Auffassung aufgrund einer Gesamtschau, „dass deutlicher Nachbesserungsbedarf besteht, um die Handhabbarkeit einer solchen Verordnung zu gewährleisten und überschießende Effekte zu verhindern".

7 Rostalski/Weiss, Der KI-Verordnungsentwurf der Europäischen Kommission, ZfDR 2021, 329 (345).

2.2. Begriffsbestimmung zur Zweckbestimmung und Leistung des KI-Systems

So sieht der KI-VO-E u.a. Begriffsbestimmungen für die Zweckbestimmung des KI-Systems gem. Art. 3 Nr. 12 KI-VO-E, die Leistung des KI-Systems gem. Art. 3 Nr. 18 KI-VO-E und von Trainings-, Validierungs- und Testdaten gem. Art. 3 Nr. 29 bis Nr. 31 KI-VO-E vor, damit eine unabhängige Bewertung des trainierten und validierten KI-Systems erfolgen kann, um die erwartete Leistung dieses Systems vor dessen Inverkehrbringen oder Inbetriebnahme zu bestätigen. An dieser Stelle vollzieht der KI-VO-E einen grundlegenden Dreischritt. Als erstes wird über die Zweckbestimmung des Anbieters die bestimmungsgemäße Verwendung des KI-Systems, einschließlich der besonderen Nutzungsumstände und -bedingungen entsprechend der Anbieterangaben in der Gebrauchsanweisung (Art. 3 Nr. 15 KI-VO-E), im Werbe- oder Verkaufsmaterial und in diesbezüglichen Erklärungen sowie in der technischen Dokumentation (Art. 11, Anhang IV KI-VO-E) ‚festgelegt. Ob nun nach einer Entwicklung eines KI-Systems die Soll-Vorgaben erfüllt werden, beschreibt die Leistung eines KI-Systems gem. Art. 3 Nr. 18 KI-VO-E als die Fähigkeit eines KI-Systems, seine Zweckbestimmung zu erfüllen. Der Abgleich vom Soll- zum Ist-Zustand vollzieht sich über die Testdaten, die eine unabhängige Bewertung des trainierten und validierten KI-Systems erlauben, um die erwartete Leistung dieses Systems vor dessen Inverkehrbringen oder Inbetriebnahme zu bestätigen.

2.3. Risikomanagementsystem

Ein weiterer großer Baustein des KI-VO-E ist das Risikomanagementsystem, dessen Anforderungen an die Zweckbestimmung nach Art. 8 II KI-VO-E bei Hochrisiko-KI-Systemen essenziell sind. Das Risikomanagementsystem nach Art. 9 II KI-VO-E verpflichtet nicht nur u.a. zur Risikoabschätzung, sondern auch zum kontinuierlichen und iterativen Testen des KI-Systems während des gesamten Lebenszyklus eines KI-Systems, um gem. Art. 9 V KI-VO-E zu prüfen, ob die Zweckbestimmung des KI-Systems und die rechtlichen Vorgaben erfüllt werden. Das Testen von Hochrisiko-KI-Systemen erfolgt nach Art. 9 VII KI-VO-E zu jedem geeigneten Zeitpunkt während des gesamten Entwicklungsprozesses und in jedem Fall vor dem Inverkehrbringen oder der Inbetriebnahme anhand vorab festgelegter Parameter und probabilistischer Schwellenwerte, die für die Zweckbestimmung des Hochrisiko-KI-Systems geeignet sind. Weiters verpflichtet Art. 9 VI KI-VO-E, dass die Testverfahren geeignet sein müssen, die Zweckbestimmung des KI-Systems unabhängig zu bewerten (vgl. Art. 3 Nr. 31 KI-VO-E), allerdings nicht über das hierfür erforderliche Maß hinauszugehen brauchen. Erst nach diesem Testverfahren lässt sich feststellen, ob ein angemessenes Maß an Genauigkeit zur Erreichung der Zweckbestimmung des Hochrisiko-KI-Systems nach Art. 15 Nr. 1 KI-VO-E erfüllt ist.

An dieser Stelle zeigt sich die starke Subjektivierung der Zweckbestimmung durch den Anbieter, der über minimalistische Maßgaben in der Zweckbestimmung eine hohe Leistungsfähigkeit des KI-Systems „erkaufen“ kann. Aus dem Blickwinkel eines freien Wettbewerbs und des stetigen technischen Fortschritts erscheint die Herangehensweise nachvollziehbar, da andere konkurrierende Anbieter ambitioniertere Zwecke und bessere Leistungen anpreisen können. Wettbewerber können so immer wieder in Zugzwang geraten, die Zweckbestimmungen des KI-Systems doch zu erweitern und müssen dann dennoch auch die Qualitätsanforderungen des Systems technisch erreichen, um keine Marktanteile zu verlieren. Damit kann der Wettbewerb einer kontinuierlichen Optimierung der auf dem Markt befindlichen Systeme dienen. Eine gewisse Objektivierung und Qualitätssicherung erfolgt über die Anforderung des Art. 9 III 1 KI-VO-E, wonach die Risikomanagementmaßnahmen dem allgemein anerkannten Stand der Technik Rechnung tragen müssen, wie er auch in einschlägigen harmonisierten Normen oder gemeinsamen Spezifikationen zum Ausdruck kommt.

2.4. Dokumentations- und Bereitstellungspflichten

Neben dem Risikomanagement zählen zu den rechtlichen Anforderungen an ein Hochrisiko-KI-System nach Art. 8 I, Art. 11 und 13 KI-VO-E eine technische Dokumentation sowie Bereitstellungspflichten von Informationen zugunsten der Nutzer. Die verpflichtende technische Dokumentation besteht zum einen nach Art. 11 I 2

KI-VO-E aus dem Nachweis der Einhaltung der Anforderungen des 2. Kapitels (Risikomanagement usw.). Zum anderen enthält die Dokumentation die Mindestangaben zur technischen Dokumentation nach Maßgabe des Anhangs IV. Exemplarisch umfasst das nach Art. 11 I 3, Anhang IV Nr. 1 lit. a KI-VO-E die allgemeine Beschreibung der Zweckbestimmung, nach Nr. 2 lit. g eine detaillierte Beschreibung der verwendeten Validierungs- und Testverfahren mit Angaben zu den verwendeten Validierungs- und Testdaten und deren Hauptmerkmalen, sowie Parameter, die zur Messung der Genauigkeit, Robustheit, Cybersicherheit und Erfüllung anderer einschlägiger Anforderungen verwendet wurden. Des Weiteren müssen detaillierte Informationen über die Überwachung, Funktionsweise und Kontrolle des KI-Systems, insbesondere in Bezug auf seine Fähigkeiten und Leistungsgrenzen, mit dem Genauigkeitsgrad für bestimmte Personen oder Personengruppen, auf die das System angewandt werden soll, und dem insgesamt erwarteten Genauigkeitsgrad in Bezug auf seine Zweckbestimmung gem. Anhang IV Nr. 3 KI-VO-E bereitgestellt werden. Weiters verpflichtet Art. 15 II KI-VO-E die Genauigkeitsgrade und die relevanten Genauigkeitskennzahlen der Hochrisiko-KI-Systeme in der ihnen beigefügten Gebrauchsanweisung anzugeben. Daneben müssen nach Art. 9 IV 1, 2 KI-VO-E die verbleibenden Restrisiken den Nutzern mitgeteilt werden. Erst in der Gesamtschau aus den Evaluationsdaten des jeweiligen Anwendungskontextes lassen sich die erreichten Genauigkeitswerte eines Modells einordnen und bewerten.[8]

2.5. Auslegung von Rechtsbegriffen

Die im KI-VO-E genannten Anforderungen sind notwendigerweise allgemein gehalten. Die Testverfahren werden daher lediglich über ihre Zweckbestimmung definiert; die in Art. 15 Nr. 1 KI-VO-E genannten Maße der „Genauigkeit, Robustheit und Cybersicherheit“ müssen anwendungsspezifisch ausgelegt werden. Dabei ist auch zu beachten, dass KI-Systeme auf Daten trainiert (und evaluiert) werden. Die Datenauswahl hat dabei (neben der Architektur) maßgeblichen Einfluss auf Genauigkeit und Robustheit bei gegebener Zweckbestimmung. Ein verzerrter Datensatz kann schlimmstenfalls zur Diskriminierung bestimmter Anwendergruppen führen (sog. „Bias“ des KI-Systems). Wir zeigen in diesem Beitrag am Beispiel der Anonymisierung von Gerichtsentscheidungen – d.h. natürlichsprachlichen Textdaten, die von den Justizbehörden durch ein Hochrisiko-KI-System[9] verarbeitet werden könnten –, wie eine solche Evaluation fundiert durchgeführt werden kann, um dem Stand der Technik und den obenstehenden Anforderungen des KI-VO-E gerecht zu werden. Dabei konzentrieren wir uns auf vier zentrale Aspekte der Evaluation, die sich aus unseren Ausführungen ergeben: (i) Erfüllung der Zweckbestimmung, (ii) Genauigkeit, (iii) Risikoabschätzung, und (iv) Robustheit.[10]

3. Automatische Anonymisierung von Gerichtsentscheidungen

Im Rahmen eines Forschungsprojekts im Auftrag des Bayerischen Staatsministeriums der Justiz forschen wir seit 2020 zur Möglichkeit einer automatischen Anonymisierung von Gerichtsentscheidungen. Als Datenmaterial standen uns ursprünglich amtsgerichtliche Urteile zur Verfügung (aus zwei unterschiedlichen

---

8 Grabmair, in: Ebers (Hrsg.), Stichwortkommentar Legal Tech, Baden-Baden 2023, Nr. 26 Rn. 5.

9 Nach Erwg. 40 S. 3 KI-VO-E soll sich die Einstufung als hochrisikoreich nicht auf KI-Systeme erstrecken, die für rein begleitende Verwaltungstätigkeiten bestimmt sind, und die die tatsächliche Rechtspflege in Einzelfällen nicht beeinträchtigen, wie die Anonymisierung oder Pseudonymisierung gerichtlicher Urteile, Dokumente oder Daten, die Kommunikation zwischen dem Personal, Verwaltungsaufgaben oder die Zuweisung von Ressourcen. Diese Erwägungen stehen im Widerspruch zum eindeutigen Wortlaut in Anhang III Nr. 8 lit. a KI-VO-E zur Bestimmung eines Hochrisiko-KI-Systems, worin ausdrücklich von „Justizbehörden“ und der „Anwendung des Rechts auf konkrete Sachverhalte“ die Rede ist und demnach nicht nur Gerichte und ihre Rechtsprechungstätigkeit erfasst sind. Die anderen Aufzählungen im Anhang III erfassen ebenso Behörden wie Strafverfolgungs-, Asylbehörden usw., d.h. die Argumentation, dass nur eine „ungefährliche“ Justizverwaltungstätigkeit vorliegt, trägt nicht. Die Gefährdungslage für Grundrechtsverletzungen bei der Anonymisierung von Gerichtsentscheidungen ist gleichermaßen gegeben, wie bei den anderen genannten Verwaltungstätigkeiten. Sicherheitshalber wird von einem Hochrisiko-KI-System ausgegangen.

10 Weitere Aspekte wie Verzerrungen (Bias), menschliche Aufsicht, Cybersicherheit usw. können hier aus Platzgründen nicht ausgeführt werden.

Rechtsgebieten: Miet- und Verkehrsrecht). Mittlerweile stehen uns ebenso Gerichtsentscheidungen von Oberlandesgerichten zur Verfügung (aus mehr als zehn Rechtsgebieten: Bau-, Familien-, Banken-, Handels-, Kapitalanlage-, Kosten-, Immaterialgüter-, Schieds-, Versicherungs-, und allgemeine Zivilsachen sowie straf- und bußgeldrechtliche Beschwerdeverfahren).

Ein wesentlicher Bestandteil des Forschungsprojekts war (und ist) die Erstellung eines sog. Goldstandards, d.h. eines manuell kuratierten Datensatzes, in welchem alle zu anonymisierenden Textstellen möglichst fehlerfrei nach festen Richtlinien markiert sind. Dazu mussten wir zunächst rechtswissenschaftlich untersuchen, welche Textstellen in Urteilen eines gegebenen Rechtsgebietes nach allen denkbarerweise relevanten rechtlichen Vorschriften, wie z.B. dem allgemeinen Persönlichkeitsrecht, der DSGVO, § 30 AO, dem GeschGehG, dem Unternehmenspersönlichkeitsrecht, §§ 203 f. StGB oder § 35 SGB I, zu anonymisieren sind. Das Ergebnis wurde in Annotationsrichtlinien festgehalten, mit deren Hilfe mehrere Hilfskräfte unabhängig voneinander die vorliegenden Urteile annotieren, d.h. kritische Textstellen (solche, die sensible Informationen enthalten) markieren und mit einer entsprechenden Kategorie (Name einer natürlichen oder juristischen Person, Adressenangabe, Kfz-Kennzeichen, etc.) sowie einer subjektiven Risikoeinschätzung (niedrig, mittel, hoch) versehen. Trotz sorgfältiger Schulung der Annotator:innen stimmen sie niemals in allen Entscheidungen überein. Neben Flüchtigkeitsfehlern gibt es immer auch Auslegungsspielraum bei Grenzfällen. Die unterschiedlichen, potenziell konfligierenden, Annotationen wurden daher in einem abschließenden Adjudikationsprozess zum endgültigen Goldstandard zusammengeführt.[11] Schließlich wurde der Goldstandard pseudonymisiert, indem alle markierten Textstellen durch informationserhaltende realistische Fantasieangaben ersetzt wurden. Da so keinerlei Verbindung zu den tatsächlich betroffenen Personen mehr herzustellen ist, konnte der pseudonymisierte Goldstandard auch außerhalb besonders geschützter Räume im Rechenzentrum der Universität genutzt werden.

Ein derartiger Goldstandard ist in jedem Fall notwendig, um automatische Verfahren zur Erkennung kritischer Textstellen fundiert evaluieren zu können. Zudem benötigen alle modernen KI-Verfahren Trainingsmaterial, aus dem sie lernen können, die Aufgabenstellung automatisch zu lösen.[12] Im Rahmen unserer Forschung haben wir verschiedene Lernverfahren und einige Standardwerkzeuge („off the shelf"-Lösungen) ausführlich miteinander verglichen und konnten zeigen, dass das Finetuning eines auf großen Textmengen vortrainierten neuronalen Sprachmodells (LLM = „large language model") die besten Ergebnisse erzielt. Der von uns entwickelte Demonstrator „LeAK 2022" erkennt knapp 99% aller Hochrisikostellen in den Testdaten.[13]

Wir werden im nächsten Kapitel ausführlich auf die Evaluation automatischer Verfahren gemäß den oben dargestellten Anforderungen der KI-VO-E eingehen. Hingewiesen sei an dieser Stelle bereits darauf, dass ein KI-System nicht auf denjenigen Daten evaluiert werden darf, auf denen es trainiert wurde. Nur so kann eine unabhängige und zuverlässige Bewertung garantiert werden. Im Extremfall eines vollkommen „übertrainierten" Systems würde die KI nämlich alle Textstellen im Trainingsdatensatz korrekt erkennen, jedoch auf neuen Daten extrem schlechte Ergebnisse liefern. Zu diesem Zweck wird der Goldstandard in einen Trainings-, einen Validierungs- und einen Evaluationsdatensatz aufgeteilt. Die Trainingsdaten dienen zum Trainieren der maschinellen Lernverfahren, die Validierungsdaten zur Auswahl und Optimierung der Verfahren. Lediglich die abschließende Evaluation erfolgt auf dem Testdatensatz, um jede Form von Überanpassung zu verhindern.

---

[11] Für Details siehe Adrian/Dykes/Evert/Heinrich/Keuchen, Automatische Anonymisierung von Gerichtsurteilen. In: Schweighofer/Zanol/Eder (Hrsg.), Rechtsinformatik als Methodenwissenschaft des Rechts – Tagungsband des 26. Internationalen Rechtsinformatik Symposions IRIS 2023, Bern 2023, S. 211–220; Adrian/Evert/Keuchen/Heinrich/Dykes, Anonymisierung von Gerichtsurteilen – Eine wesentliche Voraussetzung für E-Justice –. In: Schweighofer/Kummer/Saarenpää/Eder/Hanke (Hrsg.), Cybergovernance – Tagungsband des 24. Internationalen Rechtsinformatik Symposions IRIS 2021, Bern 2021, S. 137–147.

[12] Es gibt auch regelbasierte Verfahren, die ohne Trainingsmaterial auskommen. Ebenso ist es möglich, für andere Zwecke entwickelte Systeme (z.B. sog. Named Entity Recognition zur Identifikation von Personen- und Firmennamen, Adressen, Datumsangaben, etc.) mit entsprechenden Anpassungen einzusetzen. Diese Ansätze sind aber für eine spezielle Aufgabenstellung trainierten Lernverfahren in der Regel deutlich unterlegen, was sich auch in unserer Evaluierung gezeigt hat.

[13] Adrian/Dykes/Evert/Heinrich/Keuchen, a.a.O.

In unserem Fall ist es entscheidend, die Einteilung auf Ebene ganzer Urteile vorzunehmen: würden einige Sätze desselben Urteils den Trainingsdaten zugeschlagen und andere Sätze den Testdaten, so könnte das KI-System einen spezifischen Personennamen lernen und ihn dann leicht in den Testdaten wiedererkennen. Die strenge Abgrenzung von Trainings- und Testdaten ist in der Computerlinguistik und KI-Forschung so selbstverständlich geworden, dass wir in Kapitel 4 nicht gesondert darauf eingehen.

4. Evaluation

Wir wenden uns nun den verschiedenen Aspekten der Evaluation automatischer Verfahren zu, v.a. unter dem Gesichtspunkt, was die Evaluation leisten muss, um aus unserer Sicht dem KI-VO-E gerecht zu werden. Wir konzentrieren uns hierbei exemplarisch auf das Beispiel der Anonymisierung von Gerichtsurteilen und zeigen daran auf, welche Verfahren und Ansätze nach dem derzeitigen Stand der Technik angemessen sind.

4.1. Erfüllung der Zweckbestimmung

Eine Evaluation der Leistung eines KI-Systems ist nicht trivial, da die dafür zu erfüllende „Zweckbestimmung" ein subjektiv geprägter und qualitativer Begriff ist, mit dem die angestrebte Verwendung beschrieben (Art. 3 Nr. 12 KI-VO-E) wird. Eine Evaluation ist hingegen i.S.d. Art. 9 V KI-VO-E nur geeignet, wenn sie mit quantitativ messbaren Kriterien erfolgt. Der Begriff „Zweckbestimmung" muss daher operationalisiert werden. Das hat der KI-VO-E vor Augen, wenn die Leistung des KI-Systems anhand vorab festgelegter Parameter und probabilistischer Schwellenwerte getestet werden muss (Art. 9 VII). Auch bei der Anonymisierung von Gerichtsurteilen steht a priori kein universal anerkanntes Maß zur Verfügung, ab wann Texte „ausreichend anonymisiert" sind. Ein erster Schritt unseres Forschungsprojekts war daher die Aufarbeitung der Rechtsdogmatik: Was muss Anonymisierung aus rechtsdogmatischer Sicht leisten? Auf dieser Basis fand anschließend eine Operationalisierung der Zweckbestimmung in zwei Schritten statt. Zunächst wurde durch die Erstellung von Annotationsrichtlinien *intensional* festgelegt, welche Informationen für eine erfolgreiche Anonymisierung maskiert werden müssen.[14] Im Anschluss wurden diese Richtlinien durch Erstellung eines Goldstandards *extensional* operationalisiert, indem in konkreten Urteilen alle relevanten Textstellen annotiert wurden. Erst durch die intensionale Operationalisierung wird die Zweckbestimmung präzise und vergleichbar definiert – und nur mittels der extensionalen Operationalisierung in Form eines Goldstandards kann die Erfüllung dieser Zweckbestimmung quantitativ gemessen werden.

Wie bereits angemerkt, stimmen Annotator:innen nicht in allen Entscheidungen überein, weshalb eine Mehrfachannotation mit anschließender Adjudikation unvermeidbar ist, um eine hohe Qualität des Goldstandards sicherzustellen. Die Qualitätsanforderungen gehen dabei weit über die gängige Praxis in der Computerlinguistik hinaus, wo oft Goldstandards mit einer Fehlerquote von über 3% herangezogen werden.[15] Das bedeutet aber, dass selbst ein KI-System mit perfekten Evaluationsergebnissen immer noch über 3% Fehler machen könnte (nämlich diejenigen, die bereits im Goldstandard enthalten sind). Für die Evaluation von Hochrisiko-KI ist dieses Maß an Unsicherheit in den allermeisten Fällen nicht akzeptabel (d.h. erfüllt nicht die an probabilistische Schwellenwerte zu stellenden Anforderungen), wie in Abschnitt 4.3 am Beispiel ausgeführt wird.

[14] Neben offensichtlich zu maskierenden Textstellen wie Personennamen, Adressangaben, Telefonnummern, Autokennzeichen usw. gibt es auch zahlreiche andere potenziell identifizierende Merkmale (eine auffällige Außenfarbe eines Hauses, die Nummer eines Stellplatzes, ein besonderer Beruf, etc.), die zu einer Deanonymisierung beitragen können. Dazu Deuber/Keuchen/Christin, Assessing Anonymity Techniques Employed in German Court Decisions: A De-Anonymization Experiment, 32nd USENIX Security Symposium, Anaheim 2023, S. 5199 (5210); Vokinger/Mühlematter, Re-Identifikation von Gerichtsurteilen durch «Linkage» von Daten(banken), Jusletter 2. September 2019, S. 16.

[15] Manning, in: Gelbukh (Hrsg.), Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2011), LNCS 6608, Berlin 2011, S. 172.

Die Qualität eines Goldstandards wird nach dem Stand der Technik anhand des sog. Inter-Annotator Agreements (IAA) beurteilt, d.h. welcher Prozentsatz von Diskrepanzen sich im (meist paarweisen) Vergleich der unabhängig arbeitenden Annotator:innen findet. Dieser Wert ist aber nicht mit der Fehlerquote des Goldstandards gleichzusetzen, da es z.B. sein kann, dass beide Annotator:innen dieselbe Textstelle übersehen haben. Wir haben ein Verfahren entwickelt, mit dem durch eine Extrapolation des IAA abgeschätzt werden kann, wie viele unabhängige Annotator:innen notwendig sind, damit der Goldstandard mit hoher Wahrscheinlichkeit keine übersehenen kritischen Textstellen mehr enthält.[16] Nur so ist der Goldstandard für die Evaluation einer Hochrisiko-KI geeignet und wird Art. 10 III 1 KI-VO-E gerecht, dass Testdatensätze fehlerfrei sein müssen.

4.2. Genauigkeit

Unter der „Genauigkeit" eines KI-Systems verstehen wir hier die quantitative Evaluation gegenüber einem Goldstandard, d.h. die Angabe von Genauigkeitskennzahlen gem. Art. 15 II KI-VO-E. Die dazu herangezogenen Kennzahlen sollen dem Stand der Technik entsprechen (Art. 9 III KI-VO-E). In unserem Beispiel ist die gängige Praxis in der computerlinguistischen Forschung als Stand der Technik anzusetzen. Ein verbreiteter Ansatz besteht darin, die Erkennung kritischer Textstellen als sog. Tagging-Aufgabe zu verstehen, d.h. für jedes Token[17] eines Urteils ist zu entscheiden, ob dieses Token maskiert werden muss („positiv") oder nicht („negativ"). Es handelt sich dabei also um eine binäre Klassifikation, die anhand einer sog. Konfusionsmatrix evaluiert wird. Dazu werden alle Token der Testdaten in die vier folgenden Kategorien eingeteilt.

1. True Positives (TP) sind Token, die das KI-System korrekt als positiv vorhergesagt hat; d.h. das System wird diese Token richtigerweise maskieren.
2. False Positives (FP) sind Token, die das KI-System fälschlicherweise als positiv vorhergesagt hat; d.h. das System wird diese Token unnötig maskieren.
3. True Negatives (TN) sind Token, die das KI-System korrekt als negativ vorhergesagt hat, d.h. das System wird diese Token richtigerweise nicht maskieren.
4. False Negatives (FN) sind Token, die das KI-System fälschlicherweise als negativ vorhergesagt hat, d.h. das System hat diese zu maskierenden Token übersehen. FN sind für die automatische Anonymisierung von entscheidender Bedeutung, da sie unmittelbar zu Anonymisierungsfehlern führen.

Anhand der Konfusionsmatrix können verschiedene quantitative Maßzahlen für die Qualität des KI-Systems berechnet werden, was dann im Rahmen der juristischen Auslegung des Rechtsbegriffs der „Genauigkeit" nach dem Wortlaut des KI-VO-E zur Beurteilung des KI-Systems als maßgeblich angeordnet werden könnte. Zunächst ist aber festzustellen, dass die Evaluation anhand einer Konfusionsmatrix auf Tokenebene für KI-Systeme zur automatischen Anonymisierung ungeeignet ist. Übersieht das KI-System auch nur ein einziges Token einer längeren Textstelle, so kann dies bereits für eine Deanonymisierung genügen. Wären z.B. bei dem Personennamen „Prof. Dr.-Ing. Max von und zu Mustermann" alle Token bis auf den Nachnamen vom KI-System erkannt worden, so würden 6 von 7 Token als TP gewertet und nur eines als FN. Der übersehene Nachname stellt aber einen kritischen Anonymisierungsfehler dar, so dass die komplette Textstelle als FN zu werten ist.

Somit muss im Rahmen der Auslegung des Rechtsbegriffs „Genauigkeit" im Wortlaut des KI-VO-E – in Abhängigkeit vom Zweck des konkreten KI-Systems, Urteile automatisch zu anonymisieren – eine Evaluation auf der Ebene von Textstellen gefordert werden, um den Anforderungen des Gesetzgebers gerecht zu werden.

[16] Heinrich/Dykes/Evert, Annotator agreement in the anonymization of court decisions. Vortrag bei der Corpus Linguistics 2021 Conference, online.

[17] Der Begriff „Token" bezeichnet die Grundeinheiten, in die Texte bei der computerlinguistischen Verarbeitung zerlegt werden. Traditionelle handelt es sich dabei um einzelne Wörter, Zahlen, Satzzeichen, etc. Aktuell vortrainierte LLM zerlegen aber auch längere Wörter in mehrere Einzeltoken.

Dabei handelt es sich nicht mehr um eine binäre Klassifikation, sondern um ein Erkennungsproblem. Dadurch ist zwar keine Evaluation anhand einer Konfusionsmatrix mehr möglich, die zentralen Kategorien TP, FP und FN lassen sich aber übertragen. Eine vom KI-System identifizierte Textstelle gilt nur dann als TP, wenn sie exakt mit der entsprechenden Textstelle im Goldstandard übereinstimmt. Wird eine Textstelle gar nicht oder nur teilweise vom System erkannt, so gilt sie als FN (und in letzterem Fall zugleich als FP). Auf dieser Basis können nun gängige Evaluationsmaße berechnet werden. Die zwei wichtigsten Maße für unsere Anwendung sind Precision und Recall.

Der Recall $R = TP / (TP + FN)$ quantifiziert die „Trefferquote" des KI-Systems, d.h. den prozentualen Anteil der zu anonymisierenden Textstellen im Goldstandard, der vom System gefunden wurde. Ein hoher Recall bedeutet, dass das System den Großteil der zu anonymisierenden Textstellen gefunden hat, ein niedriger Recall bedeutet, dass viele kritische Textstellen vom System übersehen wurden.

Die Precision $P = TP / (TP + FP)$ quantifiziert die „Genauigkeit" des KI-Systems im engeren Sinn, d.h. welcher Anteil der vom System gefundenen Textstellen tatsächlich zu anonymisieren ist. Eine hohe Precision bedeutet, dass ein Großteil dieser Textstellen zurecht vom System vorgeschlagen wurde, eine niedrige Precision bedeutet, dass das System viele Textstellen unnötigerweise maskiert (und damit der Lesbarkeit des Urteils schadet). Recall und Precision hängen voneinander ab[18]: Eine hohe Precision wird in der Regel durch einen niedrigeren Recall erkauft und umgekehrt.[19] In der Computerlinguistik wird daher oft der harmonische Mittelwert $F_1 = 2 \times P \times R / (P + R)$ berechnet, um verschiedene Systeme anhand einer einzigen kombinierten Maßzahl vergleichen zu können. In unserem Fall ist aber in erster Linie ein hoher Recall wichtig: jede übersehene Textstelle bedeutet einen Anonymisierungsfehler. Dafür können niedrigere Precision-Werte in Kauf genommen werden, solange diese nicht deutlich unter 95% fallen und damit die Lesbarkeit der anonymisierten Urteile erheblich mindern würden. Diese Entscheidung, „im Zweifel für Recall und gegen Precision" ist eine anwendungsspezifische Entscheidung, die sich aus einer juristischen Abwägung zwischen der Veröffentlichungspflicht aufgrund des Rechtsstaatsprinzips, des Demokratieprinzips sowie des Justizgewährungsanspruchs einerseits und dem Persönlichkeitsschutz Betroffener andererseits ergibt.[20]

Gem. Art. 9 VI KI-VO-E müssen Testverfahren geeignet sein, die Leistung eines KI-Systems gemäß seiner Zweckbestimmung zu überprüfen, aber nicht über das hierfür erforderliche Maß hinausgehen. In unserem Fall bedeutet dies, dass – u.a. aufgrund einer teleologischen Auslegung des in Rede stehenden Wortlautes des KI-VO-E – zwar, wie gezeigt, eine strenge Evaluation auf Ebene kompletter Textstellen erforderlich ist, die Berechnung von Precision und Recall aber ein wenig abgeschwächt werden kann. Wird nämlich eine Textstelle vom KI-System mit zusätzlichen Token vorhergesagt (z.B. „Zeuge Max Mustermann" gegenüber nur „Max Mustermann" im Goldstandard), so gilt diese in der strengen Evaluation als falsch. Für die Berechnung des Recall wäre es aber sinnvoll, sie als TP zu werten, da ja alle kritische Information vom System maskiert wurde. Für die Berechnung der Precision gilt sie hingegen weiterhin als FP, da unnötige Token maskiert werden. Ähnliches gilt, wenn eine Stelle vom KI-System in zwei Teilen erkannt wird (z.B. statt der vollständigen Adresse „Bismarckstr. 6, 91054 Erlangen" im Goldstandard zwei Teiladressen „Bismarckstr. 6" und „91054 Erlangen"). Da hier weder relevante Information übersehen noch unnötige Token maskiert wurden, ist dieser Fall sowohl für Recall als auch für Precision als TP zu werten. Das eingangs erwähnte Beispiel eines Personennamens, der unvollständig erkannt wurde, kann schließlich bei der Berechnung der Precision als TP gezählt werden, da keine unnötigen Token maskiert wurden, nicht aber für den Recall.

---

[18] Buckland/Gey, The Relationship between recall and precision, Journal of the American Society for Information Science 1994 45(1), 12–19.

[19] Eine hohe Precision wird dadurch erreicht, dass das KI-System nur Textstellen zur Anonymisierung vorschlägt, bei denen es sich sehr sicher ist. Dadurch erhöht sich natürlich die Wahrscheinlichkeit, dass kritische Stellen übersehen werden. Im gegensätzlichen Extrem könnte ein System leicht einen Recall von 100% erreichen, indem es den kompletten Text maskiert; dann würde aber die Precision extrem niedrig ausfallen.

[20] Vgl. BVerwG, NJW 1997, 2694 (2695); BVerfG, NJW 2015, 3708 (3710); BGH, NJW 2017, 1819.

Abschließend ist noch festzuhalten, dass die reine Angabe von (quantitativ ermittelten) Maßzahlen nicht ausreichend ist, um die Erfüllung der Zweckbestimmung einer Hochrisiko-KI abschließend zu bewerten. In unserem Beispiel ist a priori nicht klar, welcher Recall-Wert erzielt werden muss, um das System in der Praxis einsetzen zu können. Als Basis für ein sorgfältiges Risikomanagement ist daher eine Risikoabschätzung erforderlich, die nur anhand einer detaillierten Aufschlüsselung der Evaluationsergebnisse erfolgen kann.

4.3. Risikoabschätzung

KI-Systeme sollen typischerweise dann Verwendung finden, wenn eine manuelle Bearbeitung auf Grund der Größe der Datenmengen nicht in Frage kommt. Bei der Beurteilung von Evaluationsergebnissen und der Festlegung probabilistischer Schwellenwerte muss daher die beabsichtigte Skalierung des KI-Einsatzes stets mitgedacht werden, um eine solide Risikoabschätzung zu gewährleisten. Wie in Kapitel 3 beschrieben, erzielt unser KI-System („LeAK 2022“) bei der automatischen Anonymisierung von Gerichtsurteilen knapp 99% Recall auf Hochrisikostellen. Obwohl dies für computerlinguistische Verhältnisse – und auch im Vergleich zur menschlichen Annotation – als nahezu perfekt angesehen werden kann, würden bei der Veröffentlichung von bspw. einer Million Urteilen[21] schätzungsweise mehr als 135.000 Anonymisierungsfehler erfolgen.[22] Daher sind für den massenhaften Einsatz solcher KI-Systeme höchste Anforderungen an den Recall zu stellen.

Voraussetzung für eine derartige Risikoabschätzung ist die genaue Aufschlüsselung der Evaluationsergebnisse nach verschiedenen Kriterien. Im obigen Beispiel hat erst die Aufschlüsselung übers Risikoniveau ermöglicht, die Anzahl übersehener Hochrisikostellen hochzurechnen. Auch eine Aufschlüsselung über verschiedene Informationskategorien ist in unserem Beispiel sinnvoll. So sind indirekte Identifikatoren (z.B. die Außenfarbe eines Hauses oder eine Berufsangabe) i.d.R. weniger kritisch als direkte Identifikatoren (wie Eigennamen und Adressen). Man mag auch öffentliche Firmennamen etwas weniger kritisch einschätzen als Personennamen.

Wichtig ist zudem eine Bestimmung des Fehlerrisikos auf Urteilsebene, d.h. welcher Prozentsatz der Urteile mindestens einen kritischen Anonymisierungsfehler enthält. Es ist offensichtlich weniger problematisch, wenn ein KI-System in einem einzelnen Urteil den gleichen Personennamen fünfmal nicht erkennt, als wenn es fünf verschiedene Personennamen in unterschiedlichen Urteilen übersieht.[23] Im Fall von „LeAK 2022“ verteilen sich insgesamt 26 Hochrisiko-FN auf 18 verschiedene Urteile; z.B. wurde ein Firmenname im gleichen Urteil dreimal nicht erkannt. Eine manuelle Überprüfung der FN zeigt aber, dass in den meisten Fällen der kritische Teil der Textstelle erkannt wurde und nur partielle Information auf einem niedrigeren Risikoniveau sichtbar blieb. So wurden etwa Berufsbezeichnungen wie „Dipl.-Psychologin“ oder „Rechtsanwalt“ nicht dem folgenden Personennamen zugeschlagen; bei Adressen nur Straßenname und Hausnummer maskiert, nicht aber die zugehörige PLZ und Stadt; oder am häufigsten bei Kfz-Kennzeichen der Verwaltungsbezirk nicht mit erkannt (was aber nur eine sehr grobe räumliche Eingrenzung erlaubt). Echte Hochrisiko-Fehler

[21] Dieser Wert ist eher niedrig angesetzt, da schon bei der aktuellen sehr lückenhaften Veröffentlichungspraxis in der openJur-Datenbank mehr als 600.000 Urteile zur Verfügung stehen (https://openjur.de, abgerufen am 03.10.2023). Keuchen/Deuber, Öffentlich zugängliche Rechtsprechung für Legal Tech, RDi 2022, S. 229 (233) gehen von ca. 1,6 Millionen veröffentlichungsfähigen Gerichtsentscheidungen pro Jahr aus.

[22] In dem in Kapitel 3 beschriebenen Goldstandard finden sich im Mittel etwas über 13,5 Hochrisikostellen je Urteil. Hochgerechnet auf eine Million Urteile ergibt dies insgesamt 13,5 Millionen kritische Textstellen, von denen bei einem Recall von 99% ungefähr 1% – also 135.000 – übersehen werden.

[23] Im ersten Fall ist bereits durch die erste übersehene Textstelle die Anonymität der Person aufgehoben. Weitere übersehene Stellen „verschlimmern“ den Anonymisierungsfehler nicht. Im zweiten Fall wird für alle fünf Urteile eine Deanonymisierung ermöglicht. Aus diesem Grund sind auch die unvermeidlichen Flüchtigkeitsfehler bei einer manuellen Anonymisierung höchst problematisch einzustufen.

finden sich lediglich in 5 von 141 Urteilen, was einer Fehlerrate auf Urteilsebene von 3,5% entspricht. Bei einer Anwendung dieses KI-Systems auf eine Million Urteile wäre also damit zu rechnen, dass ca. 35.000 Urteile mindestens einen kritischen Anonymisierungsfehler enthalten.

4.4. Robustheit

Ein gründliches Risikomanagement muss auch die Robustheit des KI-Systems in Betracht ziehen. So verlangt Art. 15 KI-VO-E u.a. ein angemessenes Maß an Robustheit bei Hochrisiko-KI-Systemen. In der Computerlinguistik gilt ein automatisches Verfahren insbesondere dann als „robust", wenn es gleichbleibend gute Ergebnisse erzielt, auch wenn es auf Daten angewandt wird, die sich maßgeblich von seinen Trainingsdaten unterscheiden (z.B. aus einer anderen Textsorte oder Fachdomäne stammen). Wie Erwg. 50 S. 2 KI-VO-E erläutert, gewährleistet die technische Robustheit eine Widerstandsfähigkeit gegenüber Risiken im Zusammenhang mit den Grenzen des Systems, was im Einklang mit dem computerlinguistischen Verständnis des Begriffs steht.

Im vorliegenden Beispiel der Anonymisierung von Gerichtsurteilen geht es hierbei insbesondere um die Übertragbarkeit des KI-Systems auf andere Instanzen und Rechtsgebiete. Dies ist Teil unserer aktuellen Forschung: Wie gut funktioniert ein System, das nur auf amtsgerichtlichen Urteilen des Wohnraummietrechts und des Verkehrsrechts trainiert wurde, bei der Anonymisierung von Urteilen von Oberlandesgerichten ganz anderer Rechtsgebiete. Das Problem ist hierbei nicht nur, dass dem KI-System möglicherweise die Expertise fehlt, um diese Urteile zu bearbeiten, sondern auch, dass das System in diesem Fall ohne Fehlermeldung schlechte Ergebnisse liefert. Die in der ursprünglichen Evaluation ermittelten Werte von Recall und Precision haben für eine sachfremde Anwendung des KI-Systems keine Aussagekraft mehr. In diesem Fall ist daher stets eine erneute Evaluation auf geeigneten Testdaten erforderlich.

Es ist anzumerken, dass sich die Frage nach der Robustheit eines KI-Systems zur automatischen Anonymisierung nicht nur für die Übertragung auf andere Rechtsgebiete oder andere Instanzen stellt, sondern auch für einen längerfristigen Einsatz des Systems relevant ist. Bereits kleinere stilistische Änderungen können ein KI-System durcheinanderbringen. Im vorliegenden Beispiel betrifft dies die Übertragbarkeit auf Urteile anderer Amtsgerichte – im gleichen oder in anderen Bundesländern, selbst wenn die trainierten Rechtsgebiete Wohnraummiet- und Verkehrsrecht betroffen sind. Auch kann sich die gängige Formulierung und Formatierung der Urteile mit der Zeit verändern. Ein stetiges Risikomanagement durch regelmäßig aktualisierte Evaluation ist daher unbedingt notwendig (vgl. Art. 9 II KI-VO-E). Alternativ müsste das KI-System selbst erkennen, ob es noch „robust" anwendbar ist. Prinzipiell erscheint dies möglich, da es sich bei modernen KI-Systemen um statistische Verfahren handelt, die probabilistische Konfidenzwerte für ihre Vorhersage ausgeben können. So könnten Endanwender im Falle von Urteilen, in denen Textstellen bei der automatischen Annotation auffallend geringe Konfidenzwerte enthalten, entsprechend gewarnt werden, um eine menschliche Aufsicht gemäß Art 14 IV KI-VO-E zu ermöglichen. Erfahrungswerte aus unserer Forschung legen aber nahe, dass aktuelle KI-Systeme auch bei falschen Entscheidungen meist eine sehr hohe Konfidenz ermitteln.[24] Für FN, die vom System ganz übersehen wurden, sind zudem keine Konfidenzwerte verfügbar. In jedem Fall müsste auch diese zusätzliche Fähigkeit des KI-Systems durch eine entsprechende gründliche Evaluation nachgewiesen werden.

[24] Man vergleiche hierzu das in Kapitel 1 genannte Phänomen, dass ChatGPT auch frei halluzinierte „Antworten" selbstbewusst formuliert und keine Anzeichen von Unsicherheit zeigt.

5. Fazit

Zusammenfassend lässt sich festhalten, dass der KI-VO-E grundlegende rechtliche Konzepte zur Bemessung der Qualität, wie insbesondere der Genauigkeit von KI-Systemen, zu erfassen versucht. Dennoch kann der KI-VO-E keinen allgemeinen Qualitätsmaßstab für verschiedenste KI-Systeme normieren. Zwar hat der KI-VO-E an vielen Stellen die Qualitätssicherung über ein Risikomanagement des KI-Systems vor Augen, damit die Zweckbestimmung erfüllt und Risiken reduziert werden. Dafür müssen die Zwecke und zugehörigen Genauigkeitswerte für die mannigfaltigen Anwendungsfälle durch die Nutzenden aber erst weiter spezifiziert werden. Nur in Bezug auf konkrete Datensätze und Anwendungsszenarien lässt sich eine Aussage über die Qualität eines konkreten KI-Systems treffen.[25] In diesem Beitrag haben wir gezeigt, dass eine ausführliche Evaluation eines KI-Systems zur automatischen Anonymisierung von Gerichtsurteilen, bei denen der Normwortlaut sachgerecht und technisch fundiert juristisch ausgelegt wurde[26], die aufgeworfenen Anforderungen des KI-VO-E erfüllen kann. Wir gehen davon aus, dass sich dieser Ansatz auch auf andere Hochrisiko-KI-Systeme zur Verarbeitung natürlichsprachlicher Texte übertragen lässt.

[25] Ebers, Standardisierung Künstlicher Intelligenz und KI-Verordnungsvorschlag, RDi 2021, 588 (593).

[26] So müssen z.B. die „richtigen" technischen Evaluationsverfahren für das jeweilige KI-System ermittelt werden, um den konkreten Fall unter das Tatbestandsmerkmal „Genauigkeit" subsumieren zu können. Des Weiteren bestimmt sich dann auch die Auslegung des Rechtsbegriffs der „Risikoabschätzung" wiederum nach den vorgreiflichen Auslegungsergebnissen zum Rechtsbegriff der „Genauigkeit". Es handelt sich also um einen komplexeren Auslegungsvorgang als sonst üblich, der einer Anwendung der künftigen KI-VO vorgeschaltet werden muss. Man könnte sagen: das berühmte Wort von Karl Engisch vom „Hin- und Herwandern des Blickes", zwischen Obersatz (Norm) und Lebenssachverhalt, muss nunmehr ergänzt werden, um deutlich zu machen, dass hier der Blick zwischen Sachverhalt (konkretes KI-System), Norm (Tatbestandsmerkmal „Genauigkeit") und Technik (jeweils erst zu bestimmendes Evaluationsverfahren) „umherwandern" muss. Siehe Karl Engisch: Logische Studien zur Gesetzesanwendung, 3. Aufl., Heidelberg 1963, S. 15: „Sieht man aber näher zu, so handelt es sich nur um eine ständige Wechselwirkung, ein Hin- und Herwandern des Blickes zwischen Obersatz und Lebenssachverhalt …".