

FÜHREN UNTERSCHIEDE IN DEN SPRACHFASSUNGEN DER KI-VO ZU UNTERSCHIEDLICHEN TECHNISCHEN UND JURISTISCHEN INTERPRETATIONEN? EINE UNTERSUCHUNG ANHAND AUSGEWÄHLTER TATBESTANDSMERKMALE.

Axel Adrian / Bao Minh Doan Dang / Stephanie Evert / Philipp Heinrich
/ Leonardo Zilio

Axel Adrian, Honorarprofessor und Notar, FAU Erlangen-Nürnberg, Lehrstuhl für Strafrecht, Strafprozessrecht und Rechtsphilosophie

Königstr. 21, 90402 Nürnberg, DE

axel.adrian@fau.de; <https://www.notare-adrian-weiss.de/home>

Bao Minh Doan Dang, Wissenschaftlicher Mitarbeiter, FAU Erlangen-Nürnberg, Lehrstuhl für Korpus- und Computerlinguistik
Bismarckstr. 6, 91054 Erlangen, DE

baominh.d.dang@fau.de; <https://www.linguistik.phil.fau.de/>

Stephanie Evert, Lehrstuhlinhaberin, FAU Erlangen-Nürnberg, Lehrstuhl für Korpus- und Computerlinguistik
Bismarckstr. 6, 91054 Erlangen, DE

stephanie.evert@fau.de; <https://www.stephanie-evert.de/>

Philipp Heinrich, Wissenschaftlicher Mitarbeiter, FAU Erlangen-Nürnberg, Lehrstuhl für Korpus- und Computerlinguistik
Bismarckstr. 6, 91054 Erlangen, DE

philipp.heinrich@fau.de; <https://philipp-heinrich.eu/>

Leonardo Zilio, Assistenzprofessor, UCLouvain, Institut Langage et Communication (ILC)

Place Cardinal Mercier 14/L3.06.02, 1348 Louvain-la-Neuve, BE

leonardo.zilio@uclouvain.be

Schlagworte: *KI-Verordnung, Sprachfassungen, Methodenlehre, Auslegung, Korpuslinguistik, Genauigkeit, Robustheit*

Abstract: *Die KI-Verordnung der EU soll den Einsatz von künstlicher Intelligenz in den Mitgliedsstaaten regulieren. Sie ist, wie alle Verordnungen der EU, in 24 Amtssprachen verbindlich; es ist dementsprechend schwierig, bei einem Text von etwa 80.000 Wörtern (in der deutschen Fassung) stringente Konsistenz über alle Sprachen hinweg zu gewährleisten. Insbesondere unter Berücksichtigung des technischen Kontextes der Tatbestandsmerkmale „Genauigkeit“ und „Robustheit“ beschäftigt sich die vorliegende Untersuchung auf Basis einer Korpusanalyse mit den Unterschieden in den amtlichen Sprachfassungen der Normen und der Frage, ob diese Unterschiede Einfluss auf technische Anforderungen sowie rechtliche Bewertungen haben.*

1. Einleitung

Die europäische Verordnung über künstliche Intelligenz (KI-VO) wurde in insgesamt 24 normativ gleichberechtigten Amtssprachen verfasst.¹ In zwei Aufsätzen haben wir uns bereits mit der Auslegung der

¹ Die Verordnung (EU) 2024/1689 des Europäischen Parlaments und des Rates vom 13. Juni 2024 zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz und zur Änderung der Verordnungen (EG) Nr. 300/2008,

deutschen Fassung der KI-VO – bzw. damals noch mit dem Entwurf dieser Verordnung – beschäftigt. Exemplarisch wurden die Tatbestandsmerkmale „Zweckbestimmung“, „Genauigkeit“, „Risikoabschätzung“ und „Robustheit“ insbesondere im technischen Kontext sowohl symbolischer² als auch subsymbolischer³ KI-Verfahren untersucht. Es wurde dargelegt, dass – frei nach Karl Engisch – bei der Auslegung der KI-VO, der Blick nicht nur zwischen Obersatz (Norm der KI-VO) und Lebenssachverhalt (zu beurteilendes konkretes KI-Verfahren), „hin- und herwandern“ muss, sondern, dass der Blick zwischen Norm (Tatbestandsmerkmal z.B. „Genauigkeit“), Sachverhalt (KI-Verfahren), und Technik (jeweils erst zu bestimmendes technisches Evaluationsverfahren z.B. in Bezug auf „Genauigkeit“ subsymbolischer KI-Verfahren⁴) „umherwandern“ muss. Bei der Auslegung sind also auch und insbesondere technische Begriffe zu berücksichtigen.

Im vorliegenden Beitrag untersuchen wir erneut exemplarisch die Tatbestandsmerkmale „Genauigkeit“ und „Robustheit“, diesmal jedoch im Hinblick auf etwaige Unterschiede in verschiedenen amtlichen Sprachfassungen der KI-VO, welche über den zu berücksichtigenden technischen Kontext zu unterschiedlichen juristischen Anforderungen führen könnten. Artikel 15 der KI-VO lautet bspw. folgendermaßen in den fünf meistgesprochenen Amtssprachen der europäischen Union:

- Deutsch (DE): Genauigkeit, Robustheit und Cybersicherheit
- Englisch (EN): Accuracy, robustness and cybersecurity
- Französisch (FR): Exactitude, robustesse et cybersécurité
- Italienisch (IT): Accuratezza, robustezza e cibersecurity
- Spanisch (ES): Precisión, solidez y ciberseguridad

Das deutsche Wort „Genauigkeit“ ist mathematisch unterspezifiziert und kann bspw. im Sinne der technischen Maßzahlen *Precision* oder *Accuracy* ausgelegt werden, während die englischen, französischen und italienischen Varianten („accuracy“, „exactitude“ bzw. „accuratezza“) das Maß als *Accuracy* festlegen (siehe Kapitel 2 für technische Definitionen). Das spanische „precisión“ ist hingegen das fünfte Rad am Wagen: *Precision* (spanisch „precisión“) ist ein Qualitätsmaß, das sich per definitionem grundlegend von *Accuracy* unterscheidet.⁵

(EU) Nr. 167/2013, (EU) Nr. 168/2013, (EU) 2018/858, (EU) 2018/1139 und (EU) 2019/2144 sowie der Richtlinien 2014/90/EU, (EU) 2016/797 und (EU) 2020/1828 (Verordnung über künstliche Intelligenz) ist aufrufbar unter dem folgenden Link: <http://data.europa.eu/eli/reg/2024/1689/oj>

² ADRIAN/KEUCHEN/RAPP/STEEN, Auslegung des KI-VO-E zur Evaluation von symbolischen Deduktionsverfahren der künstlichen Intelligenz, in: Schweighofer/Eder/Costatini/Schmautzer/Pfister (Hrsg.), Sprachmodelle: Juristische Papageien oder mehr?, Tagungsband des 27. Internationalen Rechtsinformatik Symposiums IRIS 2024, Bern 2024, S. 85 ff.

³ ADRIAN/EVERT/HEINRICH/KEUCHEN, Auslegung des KI-VO-E zur Evaluation von Verfahren der künstlichen Intelligenz am Beispiel der automatischen Anonymisierung von Gerichtsentscheidungen, in: Schweighofer/Eder/Costatini/Schmautzer/Pfister (Hrsg.), Sprachmodelle: Juristische Papageien oder mehr?, Tagungsband des 27. Internationalen Rechtsinformatik Symposiums IRIS 2024, Bern 2024, S. 205 ff.

⁴ So argumentieren wir z.B. in ADRIAN/EVERT/HEINRICH/KEUCHEN, a.a.O., dass für die dort konkret untersuchte subsymbolische KI-Anwendung zur automatischen Anonymisierung von Gerichtsurteilen insbesondere das technische Qualitätsmaß des Recall wichtig ist und erläutern den Unterschied zwischen Recall und Precision, als Qualitätsmaße zur Messung der Performance von subsymbolischen KI-Verfahren.

⁵ Siehe bspw. ADRIAN/EVERT/HEINRICH/KEUCHEN, a.a.O. Die beiden Qualitätsmaße werden im folgenden Abschnitt näher definiert.

Muss aufgrund des Wortlautes der spanischen Fassung ein spanisches KI-System notwendigerweise mittels *Precision* evaluiert werden, weil diese in der spanischen Version der KI-VO genannt wird? Erfasst der spanische Wortlaut dann keine KI-Verfahren, die nicht mittels *Precision* evaluiert werden können? Wäre von einem technischen und/oder rechtlichen Unterschied auszugehen, wenn es im Deutschen – statt „Genauigkeit“ – „Akkuratesse“, „Exaktheit“ oder auch „Präzision“ lauten würde? Es stellt sich die Frage, ob nicht in allen genannten Sprachfassungen (auch in der deutschen) besser das englische Wort „accuracy“ als *Terminus technicus* – oder eine jeweilige *wörtliche* Übersetzung – verwendet werden sollte, um eine einheitliche Auslegung der Norm zu gewährleisten.

2. Technische Bewertung von Genauigkeit und Robustheit

Bereits an früherer Stelle wurde festgehalten, dass die Anforderungen der KI-VO anwendungsspezifisch ausgelegt werden müssen.⁶ Wir geben hier einen Überblick, welche KI-Anwendungen auf welche Weise im Hinblick auf „Genauigkeit“ und „Robustheit“ evaluiert werden können (und sollten).

Zunächst ist festzuhalten, dass KI-Systeme Verfahren verwenden, um eine oder mehrere Zielvariable aus Eingabedaten abzuleiten, bspw. die Kategorie eines Texts oder Textsegments aus den einzelnen Zeichen, die (negative) Beschleunigung eines Autos in Abhängigkeit von Messungen des Straßenverkehrs, eine Diagnose abgeleitet aus einem Röntgenbild, etc. Man spricht in diesem Zusammenhang von abhängigen Variablen, die durch eine (oder mehrere) unabhängige Variable *erklärt* werden sollen. KI-Systeme können regelbasiert (symbolisch) oder statistisch (subsymbolisch) sein, mittels *überwachten Lernens* trainiert werden oder *unüberwacht* aus Daten lernen. In allen Fällen ist eine nachträgliche Evaluation mit Hilfe eines Goldstandards (mit einer Sammlung korrekter bzw. zu erwartender Paare von Eingabedaten und Zielvariable) möglich und generell nötig, um eine Aussage über die Leistungsfähigkeit des Systems, nicht nur aber auch im Hinblick auf „Genauigkeit“ und „Robustheit“, treffen zu können.

2.1. Genauigkeit

Die Quantifizierung der Genauigkeit eines Systems ist abhängig von der Art der durchgeführten Vorhersage. Sind die Ausprägungen der Zielvariablen Klassen (bspw. die Zugehörigkeit zu einer Textdomäne oder ein Test auf eine bestimmte Krankheit), so nennt man das KI-System einen *Klassifikator*. Sind die Ausprägungen Zahlen (bspw. Abstände oder Geschwindigkeiten), so handelt es sich um einen *Regressor*. Wir behandeln im vorliegenden Beitrag auf Grund des begrenzten Umfangs lediglich den Fall der Klassifikation.⁷

		Vorhersage	
		Positiv	Negativ
Korrekter Wert	Positiv	True Positive (TP)	False Negative (FN)
	Negativ	False Positive (FP)	True Negative (TN)

Tabelle 1: Kreuztabelle

⁶ ADRIAN/EVERT/HEINRICH/KEUCHEN, a.a.O.

⁷ Bei der Regression kann Genauigkeit insbesondere durch den mittleren (gewichteten) Vorhersagefehler bestimmt werden, d.h. der Summe der (quadrierten oder absoluten) Differenzen zwischen jeweiligem Vorhersagewert und tatsächlichem Wert.

Die Bewertung einer Klassifikation findet in den meisten Fällen unter Rückführung auf *binäre Klassifikation* statt⁸, in welcher es nur zwei Klassen (o.B.d.A. „positiv“ und „negativ“) gibt. Alle Genauigkeitsmaßzahlen basieren auf einer *Kreuztabelle* (siehe Tabelle 1), in welcher die Vorhersagen des KI-Systems bzgl. des wahren Werts der Zielvariable kreuzklassifiziert werden.

Man unterscheidet primär zwischen richtigen Vorhersagen (TP und TN) und falschen (FP und FN); welche Maßzahl aus der Anzahl von TPs, FPs, TNs und FNs abgeleitet wird, hängt von der Zweckbestimmung ab.

Wenn sowohl die Klassen ausbalanciert sind (d.h. ungefähr gleich viele tatsächlich positive bzw. tatsächlich negative Fälle vorhanden sind) sowie Fehler in beiden Klassen ähnlich „schlimm“ bewertet werden sollen, so kann „Genauigkeit“ am einfachsten als *Accuracy* aufgefasst werden, d.h. der Anteil der richtig klassifizierten Fälle an allen Fällen:

$$Accuracy = (TP + TN) / (TP + FP + TN + FN)$$

Bei unausgewogener Klassenverteilung kann diese Maßzahl jedoch irreführend sein, da hohe *Accuracy* auch rein durch die Dominanz der Mehrheitsklasse entstehen kann.⁹

Wenn die Kosten für falsch-positive Ergebnisse hoch sind, bietet sich die Bewertung mittels *Precision* (dt. auch „Präzision“) an. Diese entspricht dem Anteil der korrekt als positiv klassifizierten Fälle an allen als positiv klassifizierten Fällen.

$$Precision = TP / (TP + FP)$$

Als Beispiele für Anwendungen, die typischerweise mit *Precision* bewertet werden, sind unmoderierte Klassifikationen von unerwünschten, betrügerischen oder beleidigenden Inhalten zu erwähnen, wie bspw. bei der Spam-, Fake-News- oder Hate-Speech-Erkennung.

Sind hingegen die Kosten für falsch-negative Ergebnisse hoch, so wird der *Recall* (dt. auch „Sensitivität“ oder „Trefferquote“ genannt) betrachtet, d.h. der Anteil der korrekt als positiv klassifizierten Fälle an allen tatsächlich positiven Fällen:

$$Recall = TP / (TP + FN)$$

Diese Maßzahl bietet sich bspw. für die automatische Anonymisierung von Gerichtsentscheidungen an; dabei stellen FNs den ungünstigsten Fall dar, nämlich dass eine zu anonymisierende Textstelle vom System übersehen wurde.

Daneben gibt es die *Specificity* (dt. auch „Spezifität“), der Anteil der korrekt als negativ klassifizierten Fälle an allen tatsächlich negativen Fällen:

$$Specificity = TN / (TN + FP)$$

⁸ Die Bewertung der Klassifikation in mehrere Klassen wird meist als mehrfache binäre Klassifikation behandelt (One-vs-Rest bzw. One-vs-One); anschließend kann der (gewichtete oder auch Micro- bzw. Macro-) Durchschnitt als skalare Maßzahl verwendet werden.

⁹ Für die Evaluation eines in KI-Systemen zur Diagnose einer seltenen Krankheit, die nur 1% der Bevölkerung betrifft, ist *Accuracy* bspw. ungeeignet. Wenn das Modell alle Patienten als „gesund“ klassifiziert, erreicht es eine hohe *Accuracy* von 99%, erkennt jedoch keinen einzigen kranken Patienten.

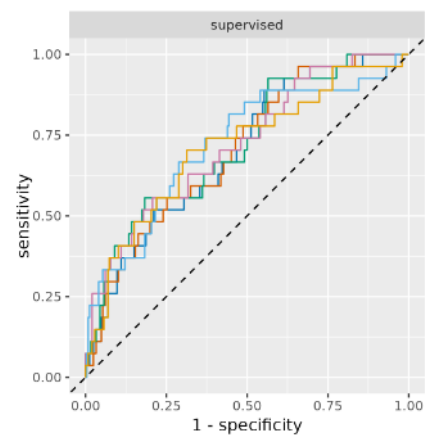
Specificity bietet sich als Bewertungskriterium an, wenn es gilt, negative Fälle korrekt zu klassifizieren und die Falsch-Positiv-Rate gering zu halten, z.B. wenn es insbesondere unerwünscht ist, gesunde Menschen fälschlicherweise als krank zu diagnostizieren.

Es ist zu erwähnen, dass *Precision* und *Recall* (und auch *Specificity* und *Recall*) in gegenseitiger Abhängigkeit stehen: *Recall* kann meist auf Kosten der *Precision* (oder *Specificity*) optimiert werden. Wenn ein Gleichgewicht zwischen *Precision* und *Recall* wichtig ist, d.h. in Fällen, in denen sowohl falsch-positive als auch falsch-negative Ergebnisse vermieden werden sollen, so verwendet man typischerweise das harmonische Mittel aus *Precision* und *Recall*, kurz *F1-Score*:

$$F1 = 2 * Precision * Recall / (Precision + Recall)$$

Diese Maßzahl ist jedoch schwer interpretierbar, wenn sich *Precision* und *Recall* stark unterscheiden.

Der oben erwähnte Trade-off zwischen *Precision* und *Recall* bzw. *Specificity* und *Recall* entspricht gewissermaßen der Trennschärfe des KI-Systems und kann anhand einer *Precision-Recall*- bzw. *Receiver-Operating-Characteristic-Kurve* (ROC-Kurve) veranschaulicht werden, siehe nebenstehende Abbildung¹⁰. Ein (nahezu) perfekter Klassifikator hat eine ROC-Kurve, die am oberen linken Rand verläuft, mit 100% Recall bei 100% Spezifität (und eine Fläche unter dieser Kurve, ROC-AUC, von 1). Das System ist hingegen nicht besser als zufälliges Raten, wenn die Kurve nahe der Winkelhalbierenden verläuft (gestrichelte Linie, ROC-AUC von 0.5).



Alle genannten Maßzahlen (und weitere ungenannte) sind Möglichkeiten, die „Genauigkeit“ eines KI-Systems zu quantifizieren. Welche dieser Maßzahlen genommen werden sollte, ist anwendungsspezifisch zu bestimmen und leitet sich primär aus der Risikobewertung unterschiedlicher Fehlertypen ab – sollte jedoch wohlüberlegten rechtlichen Richtlinien folgen.

2.2. Robustheit

„Robustheit“ bezieht sich auf das Verhalten von KI-Systemen, wenn das System in unterschiedlichen Domänen benutzt wird. Wir veranschaulichen das am Beispiel einer computerlinguistischen Anwendung. Wenn die Anwendung auf einer bestimmten Domäne trainiert wurde, so ist das KI-Verfahren auch nur auf diese bestimmte Textsorte spezialisiert. Die meisten Modelle mit *tieferen neuronalen Netzwerken* benötigen zudem eine große Menge an hochwertigen Trainingsdaten. Da diese häufig nicht zur Verfügung stehen, werden einige trainierte Modelle aufgrund ihrer Komplexität stark an die Trainingsdaten angepasst, was zu einer Verringerung der Leistungsstärke in der realen Anwendung führen kann.

¹⁰ Abbildung entnommen aus HEINRICH/BLOMBACH/DOAN

DANG/ZILIO/HAVENSTEIN/DYKES/EVERT/SCHÄFER, Automatic Identification of COVID-19-Related Conspiracy Narratives in German Telegram Channels and Chats, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Turin, Italien, S. 1932–1943.

Von „Robustheit“ kann erst dann die Rede sein, wenn ein trainiertes Lernverfahren auch auf Daten, die sich von den ursprünglichen Trainingsdaten unterscheiden, konstant gute Vorhersagen liefert. Ein „robustes“ KI-System erzielt also auch bei unbekannten Daten weiterhin verlässliche Ergebnisse. In der KI-Forschung spricht man hier auch von *invarianten Modellen*, bei denen kleinere Transformationen in den Testdaten nicht unmittelbar zu erheblichen Abweichungen in der Vorhersage führen sollten.

Es ist anzumerken, dass der heutige Stand der Technik keine einheitliche Metrik zur Messung von Robustheit definiert, obwohl das Thema für die Entwicklung sowie den Einsatz von KI-Systemen von entscheidender Wichtigkeit ist. In der Regel werden die oben geschilderten Evaluationsmetriken (wie beispielsweise *F1* oder *Accuracy*, siehe Kapitel 2.1.) herangezogen, um die Leistungsfähigkeit des Lernverfahrens auf den Testdaten zu bestimmen oder dessen Robustheit auf unbekannten und nicht domänenspezifischen Daten zu prüfen.

3. Inkonsistenzen in den verschiedenen Sprachfassungen der KI-VO

Es ist festzuhalten, dass Bewertungsmaße für „Genauigkeit“ sowie Tests auf „Robustheit“ anwendungsspezifisch unter Berücksichtigung der Zielvorgaben zu wählen sind: Welche Fehler sind *teuer*, welche können billigend in Kauf genommen werden? Inwieweit unterscheiden sich die potenziellen Anwendungsdomänen von der Trainingsdomäne? Bei der Berücksichtigung von „Robustheit“ ist außerdem anzumerken, dass praktisch immer eine Verschlechterung der *In-Domain-Leistungsfähigkeit* in Kauf genommen werden muss, um eine domänenübergreifende Verbesserung gewährleisten zu können. Das ist insbesondere dann der Fall, wenn „Robustheit“ mittels *data augmentation*, d.h. durch künstliche Erzeugung von abgewandelten bzw. verrauschten Trainingsdaten aus dem vorhandenen Goldstandard erreicht wird.

3.1. Korpus unterschiedlicher Sprachfassungen der KI-VO

Für eine systematische Analyse der unterschiedlichen Sprachfassungen wurde ein Korpus mit den deutschen, englischen, französischen, italienischen und spanischen Fassungen der KI-VO erstellt. Wir stellen sowohl die rohen Textdateien als auch eine alignierte Variante mit Englisch als Pivot-Sprache zur Verfügung.¹¹ Das Korpus hat insgesamt einen Umfang von 566.360 Wörtern. Die französischen sowie spanischen Fassungen sind mit jeweils über 100.000 Wörtern die längsten; wobei die italienischen und portugiesischen Versionen 95.549 bzw. 96.928 Wörter haben. Die deutsche Fassung ist mit lediglich 78.513 Wörtern die kürzeste unter den ausgewählten Amtssprachen, gefolgt von Englisch mit 86.927 Wörtern.

Mit geeigneter Software¹² können in der alignierten Variante Sprachpaare parallel gelesen werden. Absatz 50 der Erwägungsgründe lautet in der deutschen und englischen Version der Verordnung bspw. (unsere Hervorhebungen):

DE: In Bezug auf KI-Systeme, die Sicherheitsbauteile von Produkten oder selbst Produkte sind, die in den Anwendungsbereich bestimmter, im Anhang dieser Verordnung aufgeführter Harmonisierungsrechtsvorschriften der Union fallen, *ist es angezeigt, sie im Rahmen dieser Verordnung als hochriskant einzustufen*, wenn das betreffende Produkt gemäß den einschlägigen Harmonisierungsrechtsvorschriften der Union dem

¹¹ Verfügbar unter <https://github.com/fau-klue/ai-act-corpus>. Das Korpus enthält neben den fünf erwähnten Sprachen auch portugiesisch (PT). Die halbautomatische Alignierung wurde mit Hilfe von *Trados Studio 2024 aligner tool* (<https://www.trados.com/product/studio/>) mit anschließender manueller Überprüfung durchgeführt.

¹² Wir verwenden hierzu SketchEngine (<https://www.sketchengine.eu/>).

Konformitätsbewertungsverfahren durch eine als Dritte auftretende Konformitätsbewertungsstelle unterzogen wird.

EN: As regards AI systems that are safety components of products, or which are themselves products, falling within the scope of certain Union harmonisation legislation listed in an annex to this Regulation, *it is appropriate to classify them as high-risk* under this Regulation if the product concerned undergoes the conformity assessment procedure with a third-party conformity assessment body pursuant to that relevant Union harmonisation legislation.

Absätze 50 bis 62 beschreiben Fallbeispiele, wann Systeme als Hochrisiko-KI eingestuft werden sollten bzw. wann eine solche Einstufung angebracht ist. Es fällt hier bereits auf, dass die Bedeutung des Begriffs „Hochrisiko-KI-System“ im Gesetz nicht technisch definiert, sondern ostensiv (durch Aufzeigen einiger Beispiele) eingeführt wird.

3.2. Genauigkeit und Robustheit in der KI-VO

Eine Auflistung der Häufigkeiten der Begriffe „Genauigkeit“ und „Robustheit“ sowie ihrer Übersetzungen (Tabelle 2) lässt einige Diskrepanzen erkennen, auf die wir in den folgenden Abschnitten näher eingehen. Anzumerken ist, dass der Begriff „precisión“ im Spanischen insgesamt 21 mal erwähnt wird, davon aber zweimal in anderen Kontexten. In der deutschen Version ist bei drei von 17 Erwähnungen von „Robustheit“ von „technischer Robustheit“ die Rede.

Amtssprache	Begriff: Genauigkeit	Begriff: Robustheit
Deutsch	Genauigkeit (n=20)	Robustheit (n=17)
Englisch	accuracy (n=20)	robustness (n=18)
Französisch	exactitude (n=19)	robustesse (n=16)
Italienisch	accuratezza (n=19)	robustezza (n=18)
Spanisch	precisión (n=19)	solidez (n=17)

Tabelle 2: Auflistung der Übersetzungen von „Genauigkeit“ sowie „Robustheit“ in den ausgewählten Amtssprachen mit ihrer Vorkommenshäufigkeit in der jeweiligen Sprachversion in runden Klammern.

Im unteren Beispiel wird *accuracy* in der englischen Fassung der KI-VO für „sachliche Richtigkeit“ anstelle von Genauigkeit verwendet:

DE: Bei einer solchen Nutzung, sofern sie zulässig ist, müssen auch die in Artikel 4 Absatz 1 der Richtlinie (EU) 2016/680 festgelegten Grundsätze geachtet werden, einschließlich Rechtmäßigkeit, Fairness und Transparenz, Zweckbindung, *sachliche Richtigkeit* und Speicherbegrenzung.

EN: Such use, when authorised, also needs to respect the principles laid down in Article 4 (1) of Directive (EU) 2016/680 including lawfulness, fairness and transparency, purpose limitation, *accuracy* and storage limitation.

Hervorzuheben ist Artikel 15 der KI-VO, welcher darauf abzielt, Parameter für die Zuverlässigkeit von Hochrisiko-KI-Systemen zu definieren, und die Begriffe „Genauigkeit“ und „Robustheit“ bereits im Titel enthält. In den folgenden Absätzen der KI-VO finden sich allerdings keine eindeutigen Definitionen dieser Begriffe — und auch nirgendwo sonst in der Verordnung, insbesondere nicht in Artikel 3 (Begriffsbestimmungen).

Es ist außerdem anzumerken, dass die KI-VO sowohl „Robustheit“ allein als auch „technische Robustheit“ erwähnt. Zwar erhält nur letzterer Begriff in den einleitenden Absätzen eine bedeutungsklärende Beschreibung, doch ist davon auszugehen, dass die KI-VO durchgehend *technische Robustheit* meint, auch wenn das Attribut „technisch“ nicht explizit genannt wird (diese Interpretation wird durch eine manuelle Überprüfung aller Vorkommnisse im Korpus gestützt). Erwägungsgrund 27 (Grundsätze für eine vertrauenswürdige und ethisch vertretbare KI) definiert technische Robustheit folgendermaßen (unsere Hervorhebungen):

DE: „*Technische Robustheit und Sicherheit*“ bedeutet, dass KI-Systeme *so entwickelt und verwendet werden, dass sie im Fall von Schwierigkeiten robust sind und widerstandsfähig gegen Versuche, die Verwendung oder Leistung des KI-Systems so zu verändern, dass dadurch die unrechtmäßige Verwendung durch Dritte ermöglicht wird, und dass ferner unbeabsichtigte Schäden minimiert werden.*

EN: “*Technical robustness and safety*” means that AI systems are *developed and used in a way that allows robustness in the case of problems and resilience against attempts to alter the use or performance of the AI system so as to allow unlawful use by third parties, and minimise unintended harm.*

Diese Definition *technischer Robustheit* unterscheidet sich grundsätzlich von der Definition von Robustheit wie sie bspw. in der Computerlinguistik und Informatik üblich ist; hier werden Lernverfahren robust genannt, wenn sie auch auf Daten, die sich von den ursprünglichen Trainingsdaten unterscheiden, zuverlässige Vorhersagen liefern (siehe Kapitel 2.2.).

In einigen anderen Abschnitten (wie Artikel 15 der KI-Verordnung, in dem die Begriffe „Genauigkeit“, „Robustheit“ und „Cybersicherheit“ erläutert werden) wird in der französischen Version „solidité“ anstelle von „robustesse“ für *Robustheit* gewählt:¹³

DE: dass die KI-Reallabore die Entwicklung von Instrumenten und Infrastruktur für das Testen, das Benchmarking, die Bewertung und die Erklärung der Dimensionen von KI-Systemen erleichtern, die für das regulatorische Lernen Bedeutung sind, wie etwa Genauigkeit, *Robustheit* und Cybersicherheit, sowie Maßnahmen zur Risikominderung im Hinblick auf die Grundrechte und die Gesellschaft als Ganzes fördern.

EN: that AI regulatory sandboxes facilitate the development of tools and infrastructure for testing, benchmarking, assessing and explaining dimensions of AI systems relevant for regulatory learning, such as accuracy, *robustness* and cybersecurity, as well as measures to mitigate risks to fundamental rights and society at large.

FR: que les bacs à sable réglementaire de l'IA facilitent le développement d'outils et d'infrastructures pour la mise à l'essai, l'étalonnage des performances, l'évaluation et l'explication des aspects des systèmes d'IA pertinents pour l'apprentissage réglementaire, tels que la précision, la *solidité* et la cybersécurité, ainsi que les mesures d'atténuation des risques d'atteinte aux droits fondamentaux et à la société au sens large.

Die Terminologiedatenbank IATE (Interactive Terminology for Europe), die seit „Sommer 2004 den EU-Institutionen und -Einrichtungen zur Sammlung, Verbreitung und Verwaltung EU-spezifischer Termino-

¹³ In den deutschen, französischen und spanischen Versionen wird Robustheit ebenfalls adjektivisch als „robust“ (bspw. Erwägungsgrund 8), „robustes“ (bspw. Erwägungsgrund 27) und „sólidos“ (auch Erwägungsgrund 27) verwendet.

logie“ dient¹⁴, listet tatsächlich diese beiden französischen Übersetzungen („robustesse“ und „solidité“) für „robustness“ (und analog italienisch „robustezza“ und „solidità“) auf. Eintrag 3544336¹⁵ aus dem Bereich Statistik, Wirtschaft und Informatik definiert Robustheit als „Unempfindlichkeit gegenüber Anwendungsvoraussetzungen“, was der oben aufgeführten computerlinguistischen Definition entspricht, jedoch nur schwer mit den Ausführungen in der KI-VO in Einklang zu bringen ist.

Beim Begriff „Genauigkeit“ ist das Problem noch komplexer, da sich die Definitionen in den verschiedenen Fachbereichen vervielfachen (und das KI-Gesetz selbst keine Definition liefert), während die Sprachen auch nicht konsistent festlegen, was „Genauigkeit“ bedeutet. Der Text des KI-Gesetzes verwendet „Genauigkeit“ im Sinne einer Metrik, die zur Messung der Leistungsfähigkeit eines KI-Systems angewandt werden kann, wie in Artikel 15, Absatz 3 (unsere Hervorhebungen):

DE: Die *Maße an Genauigkeit* und die relevanten *Genauigkeitsmetriken* von Hochrisiko-KI-Systemen werden in den ihnen beigelegten Betriebsanleitungen angegeben.

EN: The *levels of accuracy* and the relevant *accuracy metrics* of high-risk AI systems shall be declared in the accompanying instructions of use.

ES: En las instrucciones de uso que acompañen a los sistemas de IA de alto riesgo se indicarán los *niveles de precisión* de dichos sistemas, así como los parámetros pertinentes para medirla.

Wie eingangs erwähnt, ist das deutsche Wort „Genauigkeit“ mathematisch unterspezifiziert; ein Blick auf die englische Sprachfassung (sowie die französische und italienische) suggeriert die Verwendung von *Accuracy*, während das spanische „precisión“ die Verwendung von *Precision* nahelegt.

Diese Interpretation wird auch durch einen Blick in die IATE-Datenbank unterstützt. Im Eintrag für *Accuracy* aus dem Bereich der technischen Vorschriften wird ausdrücklich erwähnt, dass der Begriff „nicht verwechselt werden darf mit: [...] Präzision [und] Richtigkeit [...]“.¹⁶ Im spanischen Äquivalent für „accuracy“ finden wir den Begriff „exactitud“, während „precisión“ als Äquivalent für „precision“ angegeben wird. Es ist allerdings festzuhalten, dass es sich bei dieser Definition um *Messgenauigkeit* (v.a. in den Naturwissenschaften) handelt. Es ist fraglich, ob diese Definition auf computerlinguistische bzw. KI-Sachverhalte übertragen werden kann.

3.3. Rechtliche Diskussion

Wenn man berücksichtigt, dass alle Sprachversionen der KI-VO im EU-Gebiet gleichermaßen gültig sind, führt terminologische Diskrepanz zu einem erheblichen Problem hinsichtlich der angemessenen Bewertung von KI-Systemen nach Maßgabe der KI-VO.

Die Frage, wie mit etwaigen Unterschieden in verschiedenen Sprachfassungen einheitlich geltender Normen rechtsmethodisch zu verfahren ist, ist nicht neu.¹⁷ Es bestehen mindestens elf verschiedene theoreti-

¹⁴ <https://iate.europa.eu/home> – abgerufen am 31.10.2024

¹⁵ <https://iate.europa.eu/entry/result/3544336/> – abgerufen am 31.10.2024

¹⁶ <https://iate.europa.eu/entry/result/1406252>, unsere Übersetzung aus dem Englischen – abgerufen am 31.10.2024

¹⁷ Aktuell zu verschiedenen Sprachfassungen der Definition von „KI-System“ in der KI-VO siehe z.B. BRAEGEL-MANN, KI-VO und Compliance – aktuell Brennpunkte, in *Künstliche Intelligenz und Recht* Heft 2/2024, S. 39 ff.

sche Lösungsansätze¹⁸, die hier nicht alle im Einzelnen wiedergegeben werden können. Ansätze, die z.B. davon ausgehen, dass der Gesetzgeber selbst einen Vorrang einer Sprachfassung definieren könnte, oder dass man auf die Sprachfassung eines Urentwurfes, zu dem z.B. die Verhandlungen im Normsetzungsverfahren erfolgt sind, zurückgreift, also z.B. auf die Arbeitssprache o.ä., scheiden schon deswegen aus, da die KI-VO in allen 24 Amtssprachen gleichermaßen rechtsverbindlich ist. Methodisch auf die engste Sprachfassung abzustellen oder umgekehrt auf die weiteste wird vorgeschlagen. Die vermittelnde Lösung will schließlich auf die Sprachfassung abzielen, die der Mehrheit der Sprachfassungen entspricht. Auch diese Vorschläge können aus dem oben genannten Grund nicht überzeugen.

Im Ergebnis sind nach der autonomen europarechtlichen juristischen Methodenlehre alle Sprachfassungen „gleich autoritativ“ und die Wortlautauslegung muss alle Sprachfassungen berücksichtigen.¹⁹ Die Mehrzahl an gleichwertigen Sprachfassungen erweitert daher tendenziell die Wortsinnngrenze.²⁰ Wörter stehen im Verhältnis des Spezielleren zum Allgemeineren.²¹ So könnte man das Verhältnis folgendermaßen beschreiben:

Genauigkeit (DE) $\supset$ Accuracy (EN) = Exactitude (FR) = Exatidão (PT) $\neq$ Precisión (ES)

4. Fazit

Im Ergebnis schlagen wir vor, auch zur Lösung etwaiger Sprachdifferenzen der KI-VO den technischen und praktischen Kontext der Tatbestandsmerkmale mit einzubeziehen. Die Grundmenge möglicher Wortbedeutungen der jeweiligen Norm soll daher aus der Summe der in allen Textfassungen der Norm²² verwendeten Begriffen und deren jeweiligen Wortbedeutungen unter Berücksichtigung des jeweiligen technischen Kontextes²³ zu bestimmen sein. Aus dieser maximalen, demokratisch noch legitimierten Grundmenge möglicher Wortbedeutungen ist dann im weiteren Fortgang der Interpretation die „richtige“ Bedeutung aus der Menge möglicher Bedeutungen zu ermitteln, und zwar insbesondere durch Zusammenschau von Norm, konkret zu beurteilendem KI-Verfahren, dafür geeignetem technischen Evaluationsverfahren sowie Anwendungskontext und die daraus abzuleitenden „Kosten“ unterschiedlicher Fehlertypen.

Es bleibt festzuhalten, dass auch technische Begriffe wie „Genauigkeit“ und „Robustheit“ in natürlichen Sprachen unterdeterminiert sind und daher ausgelegt werden müssen. In unserem Beitrag haben wir gezeigt, dass das unter Berücksichtigung der technischen Gegebenheiten und der Kosten verschiedener Fehlertypen erfolgen muss. Zusätzlich vergrößern divergierende Sprachfassungen die Menge der möglichen Auslegungen und daraus ist wieder wie oben die passendste Auslegung zu bestimmen. So muss konkret bei der Auslegung der spanischen Sprachfassung über die irreführende Verwendung des Wortes „precisión“ hinausgegangen und unter Verweis auf den technischen Kontext und die übrigen Sprachfassungen gefolgert werden, dass eigentlich „exactitud“ gemeint ist. Nur so kann die Verordnung auf alle KI-

¹⁸ Ausführlich hierzu bereits ADRIAN, Grundprobleme einer juristischen (gemeinschaftsrechtlichen) Methodenlehre, 1. Aufl., Berlin 2009, S. 363 ff. m.w.N.

¹⁹ RIESENHUBER, a.a.O., § 10 Rdn 14, S. 293 m.w.N. aus der Rechtsprechung des EuGH.

²⁰ RIESENHUBER, a.a.O., § 10 Rdn 18, S. 296.

²¹ ADRIAN, a.a.O., S. 380 ff. m.w.N.

²² ADRIAN, a.a.O., S. 375 ff.

²³ ADRIAN/KEUCHEN/RAPP/STEEN, a.a.O., S. 85 ff.; ADRIAN/EVERT/HEINRICH/KEUCHEN, a.a.O., S. 205 ff.

Verfahren sinnvoll angewendet werden. Rechtsmethodisch liegt darin aus der Perspektive der spanischen Sprachfassung eine (teleologische) Extension.²⁴

Hinsichtlich der Sprachdifferenzen in Bezug auf das Tatbestandsmerkmal „Robustheit“ wird gerade unter Bezugnahme auf den dargestellten technischen Kontext vorgeschlagen, die spanische und die portugiesische Sprachfassung, also „solidez“, zu Gunsten der technisch „gemeinten“ Konzepte von „Robustheit“, „robustness“, bzw. „robustess“ zu vernachlässigen, um nicht zu sagen, zu ignorieren. Rechtsmethodisch liegt darin eine (teleologische) Reduktion des Gesamtwortlautes um die portugiesische und spanische Sprachfassung.²⁵

²⁴ ADRIAN, a.a.O., S. 924 f. m.w.N.

²⁵ ADRIAN, a.a.O., S. 925 f. m.w.N.