

SIAK-Journal – Zeitschrift für Polizeiwissenschaft und polizeiliche Praxis



Herrmann, Jessica/Breitinger, Frank

KI in der Rechtsprüfung: Warum XAI allein nicht ausreicht

SIAK-Journal – Zeitschrift für Polizeiwissenschaft und polizeiliche Praxis (1/2026), 85-101.

doi: 10.7396/2026_1_G

Um auf diesen Artikel als Quelle zu verweisen, verwenden Sie bitte folgende Angaben:

Herrmann, Jessica/Breitinger, Frank (2026). KI in der Rechtsprüfung: Warum XAI allein nicht ausreicht, SIAK-Journal – Zeitschrift für Polizeiwissenschaft und polizeiliche Praxis (1), 85-101, Online: https://dx.doi.org/10.7396/2026_1_G.

© Bundesministerium für Inneres – Sicherheitsakademie / Verlag Österreich, 2026

Hinweis: Die gedruckte Ausgabe des Artikels ist in der Print-Version des SIAK-Journals im Verlag Österreich (<https://www.verlagoesterreich.at/>) erschienen.

Online publiziert: 4/2026

KI in der Rechtsprüfung: Warum XAI allein nicht ausreicht

Die juristische Subsumtion gilt seit jeher als Prüfstein für die Leistungsfähigkeit moderner KI-Systeme. Trotz jahrzehntelanger Forschung gelingt es bislang kaum, juristische Prüfschritte in ihrer dogmatischen Tiefe maschinell abzubilden. Frühe Expertensysteme scheiterten an starren Formalisierungen, aktuelle Sprachmodelle eröffnen zwar neue Möglichkeiten, doch ihre Argumentationslogik bleibt probabilistisch und rechtlich nur eingeschränkt nachvollziehbar. Hier zeigt sich die Spannung zwischen technischer Leistungsfähigkeit und normativer Strukturtreue. KI-Modelle können plausibel klingende Texte erzeugen, ohne eine systematische Prüfstruktur einzuhalten. Vor diesem Hintergrund stellt sich die Frage: Was fehlt noch zum KI-Richter und womit müssen wir uns befassen, um eine rechtsstaatliche Einführung zu ermöglichen? Die meisten Verfahren der „Explainable AI“¹ („erklärbare KI“, XAI) liefern bislang lediglich technische Transparenz, nicht aber eine rechtsstaatlich tragfähige Begründung im Sinne einer „Explainability by Design“². Ein juristisch normiertes Lastenheft kann hier Maßstäbe setzen, indem es Anforderungen an Prüfschritte, Alternativdarstellungen, Protokollierung und Zielgruppenadäquanz definiert. Dieser Artikel stellt den aktuellen Forschungsstand entlang von fünf methodischen Hauptlinien dar: regelbasierte Systeme, sprachmodellbasierte Ansätze, fall- und vektorbasierte Verfahren³, wissensrepräsentationsbasierte Modelle sowie hybride Architekturen und diskutiert ihre jeweiligen Stärken und Grenzen im Hinblick auf Transparenz, Flexibilität und dogmatische Steuerbarkeit. Im Zentrum steht die These, dass nur Systeme, die von vornherein so konstruiert sind, dass jeder Subsumtionsschritt explizit nachvollziehbar bleibt, perspektivisch als eigenständige Entscheidungsinstanz diskutiert werden können.

1. EINFÜHRUNG

1.1 Historischer Hintergrund

Die Frage, ob und wie juristische Entscheidungsprozesse durch technische Systeme abgebildet werden können, begleitet die Rechtsinformatik seit ihren Anfängen. Bereits in den 1960er Jahren gab es frühe Versuche, juristische Entscheidungen formal zu erfassen. Statistische Analysen

von Gerichtsentscheidungen (vgl. Lawlor 1963) sollten Muster sichtbar machen und gelten heute als Vorläufer der „Legal Judgment Prediction“⁴. In den 1970er- und 1980er Jahren entstanden erste Expertensysteme wie TAXMAN, LEGOL oder die logikbasierten Modelle von Sergot et al. und Bench-Capon et al., die Rechtsnormen in Entscheidungsbäume überführten (vgl. McCarty 1977; Stamper 1980; Sergot et



JESSICA HERRMANN,
*Polizeibeamtin (BW), forscht zu
Legal AI und rechtsstaatlicher
KI-Implementierung.*



FRANK BREITINGER,
*Lehrstuhlinhaber für Cybersecurity,
Universität Augsburg.*

al. 1986; Bench-Capon et al. 1987). Diese machten Subsumtionslogik erstmals maschinenlesbar und hochgradig erklärbar, erwiesen sich jedoch als starr (vgl. Susskind 1990). Das Problem lag nicht in der Nachvollziehbarkeit, sondern in der fehlenden dogmatischen Tiefe. Über einfache Wenn-Dann-Regeln hinaus ließen sich keine belastbaren Entscheidungen darstellen. In den 1990er Jahren stagnierte die Entwicklung juristischer Expertensysteme, unter anderem wegen fehlender digitaler Falldaten, begrenzter Rechenkapazitäten und mangelnder Interoperabilität juristischer Systeme (vgl. Bench-Capon et al. 2012).

Mit dem Aufkommen großer Sprachmodelle (LLM⁵) erhielt die Debatte neue Dynamik. Modelle wie GPT-2⁶ (2019) und GPT-3 (2020) demonstrierten, dass juristische Texte sprachlich kohärent verarbeitet und erzeugt werden konnten. Parallel entstanden größere juristische Datensätze wie CAIL2018 und Benchmarks⁷ wie LexGLUE, die systematische Evaluierungen ermöglichten (vgl. Zhong et al. 2018; Chalkidis et al. 2022). Das grundlegende Defizit blieb jedoch bestehen: Die Argumentationslogik dieser Modelle folgte statistischen Wahrscheinlichkeiten und nicht einer nachvollziehbaren juristischen Prüfarchitektur. Während einzelne Subsumtionsschritte, also die Anwendung einer Definition auf einen Sachverhalt, von Sprachmodellen durchaus bewältigt werden konnten, scheiterten sie an der konsistenten Abbildung ganzer Prüfschemata. Gerade bei reinem Prompting⁸ wurde deutlich, dass zwar plausible Ergebnisse erzeugt wurden, die dogmatische Struktur jedoch nicht systematisch eingehalten wurde (vgl. Guha et al. 2023). Damit zeigt sich ein wiederkehrendes Muster: Juristische KI-Systeme pendeln zwischen starrer Logik und sprachlicher Flexibilität, ohne die dogmatische Prüfarchitektur in ihrer

Tiefe präzise zu erfassen. Diese Defizite sind nicht nur technischer Natur, sondern berühren unmittelbar die verfassungsrechtlichen Mindestanforderungen, die an richterliche Entscheidungen gestellt werden.

1.2 Verfassungsrechtliche Dimension

Richterliche Entscheidungen unterliegen in allen europäischen Rechtsordnungen rechtsstaatlichen Mindestanforderungen. Art. 6 der Europäischen Menschenrechtskonvention (EMRK) garantiert das Recht auf ein faires Verfahren (vgl. Council of Europe 1950) und das Prinzip des natürlichen Richters, das durch KI-Systeme nicht unterlaufen werden darf (vgl. Council of Europe, Parliamentary Assembly 2020). Die herrschende Meinung geht bislang davon aus, dass richterliche Entscheidungen zwingend von einem Menschen getroffen werden müssen (vgl. Susskind 2019). Gleichzeitig verweist die hohe Zahl von in zweiter Instanz abgeänderten oder aufgehobenen Urteilen auf die Fehlbarkeit (vgl. Posner 2008) bzw. unterschiedliche Interpretationsspielräume (vgl. Engel 2022) menschlicher Richter. Ob automatisierte Entscheidungssysteme zulässig wären, wird je nach nationaler Rechtsordnung unterschiedlich bewertet, nicht zuletzt, weil sich die jeweiligen Verfassungen und justiziellen Traditionen erheblich unterscheiden (vgl. De Visser 2014; Hildebrandt 2020). Sollte es jedoch gelingen, richterliche Qualifikationsmerkmale wie Unabhängigkeit, Unparteilichkeit und fachliche Kompetenz auch technisch einzulösen, stellt sich die Frage, ob konsistentere Ergebnisse durch KI-Systeme von Verfahrensbeteiligten nicht sogar bevorzugt würden und ob es in diesem Fall unethisch wäre, auf ihren Einsatz zu verzichten (vgl. Floridi/Cowls 2019; Lagioia/Sartor 2020; Cole 2024). Hinzu kommt, dass viele Justizsysteme bereits heute an strukturelle Belastungsgrenzen stoßen. Die Überlas-

tung von Gerichten gefährdet nicht nur die Effizienz, sondern auch den verfassungsrechtlich garantierten Zugang zum Recht. Eine maßvoll eingesetzte Automatisierung könnte hier nicht allein als technisches Hilfsmittel, sondern als Bestandteil rechtsstaatlicher Sicherung verstanden werden (vgl. Susskind 2017; ders. 2019).

Die europäische und nationale Rechtswissenschaft diskutiert diese Fragen kontrovers. Gerade juristische Berufe verstehen sich traditionell als menschliche Interpretationsinstanz, was eine anhaltende Skepsis gegenüber automatisierten Entscheidungen erklärt (vgl. Wischmeyer 2020; Susskind 2017). Empirische Studien zeigen jedoch, dass die Zustimmung steigt, wenn KI klar als Assistenzsystem positioniert wird, das den Menschen unterstützt, aber nicht ersetzt (vgl. Deeks 2019; Jakesch et al. 2023). Diese Ergebnisse unterstreichen, dass die Akzeptanzfrage eng mit dem Grad der Automatisierung verknüpft ist. Während die Zulässigkeit vollautomatisierter Entscheidungen überwiegend verneint wird, stellt sich zugleich die Herausforderung, wie KI-gestützte Systeme als Assistenztechnik so gestaltet werden können, dass sie rechtsstaatlichen Anforderungen genügen. Gerade hier kommt es auf Transparenz und Erklärbarkeit an. Eine Begründung im Rechtssinne muss nachvollziehbar darlegen, wie Daten in eine normativ vorgegebene Prüfstruktur eingeordnet werden. Dies verweist auf die Notwendigkeit eines verbindlichen Rahmens, der festlegt, welche Anforderungen eine juristisch tragfähige Erklärung erfüllen muss.

2. AKTUELLER FORSCHUNGSSTAND

Vor dem Hintergrund rechtsstaatlicher Mindestanforderungen lassen sich die aktuellen Entwicklungen als unterschiedliche Versuche verstehen, die bekannten

Grenzen früherer Systeme mit neuen Technologien zu überwinden. Klassische Ansätze scheiterten häufig an der begrenzten Formalisierbarkeit rechtlicher Entscheidungsprozesse, neuere Modelle dagegen an der fehlenden dogmatischen Strukturtreue.

Der aktuelle Forschungsstand zur automatisierten juristischen Subsumtion lässt sich in fünf methodische Hauptlinien gliedern. Sie verfolgen unterschiedliche Strategien, um Entscheidungsprozesse formal, nachvollziehbar und zugleich sprachlich flexibel abzubilden. Allen gemeinsam ist, dass sie die juristische Prüfarchitektur nicht vollständig erfassen. Der schrittweise Aufbau einer juristischen Prüfung wird nur unvollständig oder verzerrt nachgebildet, wobei die Ansätze jeweils an anderer Stelle spezifische Defizite aufweisen.

- ▶ Regelbasierte Expertensysteme: Sie bilden juristische Prüfschemata vollständig ab, meist in Form von Entscheidungsbäumen oder deterministischen Wenn-Dann-Regeln. Damit wird der Entscheidungsprozess Schritt für Schritt maschinenlesbar.

Beispiele: TAXMAN (vgl. McCarty 1977) oder LEGOL (vgl. Stamper 1980).

- ▶ Rein sprachmodellbasierte Systeme (LLM-only): Sie steuern Entscheidungen anhand probabilistischer Sprachmuster, ohne feste logische Strukturen. Die Subsumtion erfolgt implizit über Wahrscheinlichkeitsverteilungen der Trainingsdaten. Beispiel: GPT-3 (vgl. Brown et al. 2020).

- ▶ Fall- und vektorbasierte Ansätze: Sie orientieren sich an Fallanalogien statt an expliziten Regeln. Ausgangspunkt ist Hafners Konzept der fallbasierten Wissensorganisation (vgl. Hafner 1987), das in modernen vektorraumgestützten Verfahren fortgeführt wird. Entscheidungen ergeben sich hier aus Ähnlichkeitsmessungen im semantischen Raum, nicht aus expliziten Prüfschemata.

Beispiele: LegalDuet (vgl. Xu et al. 2025), IOT-Match (vgl. Yu et al. 2022).

- ▶ Wissensrepräsentationsbasierte Systeme: Sie erfassen juristische Begriffe und Normzusammenhänge formal in Ontologien⁹ oder semantischen Graphen. Die Subsumtion erfolgt über logische Schlussfolgerungen auf dieser Wissensbasis. Beispiele: LegalRuleML (vgl. Palmirani et al. 2011), JUREX-4E (vgl. Liu et al. 2025), Description Logic Programs (vgl. Grosz et al. 2003).
- ▶ Hybride Architekturen: Sie verbinden feste Argumentationsstrukturen mit der Generationsfähigkeit großer Sprachmodelle. Ziel ist eine Balance zwischen Nachvollziehbarkeit formaler Prüfschemata und sprachlicher Flexibilität.

Beispiele: SyLeR (vgl. Zhang et al. 2025a) oder Beyond Guilt (vgl. Zhang et al. 2025b).

Die fünf Ansätze verdeutlichen unterschiedliche Wege, juristische Prüfstrukturen technisch zu operationalisieren. Sie unterscheiden sich vor allem im Verhältnis von Transparenz, Flexibilität und dogmatischer Steuerbarkeit. Keiner der Ansätze löst den Zielkonflikt vollständig.

Auch hybride und kombinierte Modelle, die Transparenz und Flexibilität in Balance bringen sollen, befinden sich noch im Entwicklungsstadium und sind methodisch nicht ausgereift. Die fünf Ansätze werden im Folgenden im Detail betrachtet.

2.1 Expertensysteme und regelbasierte Subsumtionslogik

2.1.1 Frühe Systeme und ihre Grenzen

Während TAXMAN und LEGOL in den 1970er Jahren noch prototypisch blieben, gelten sie heute als Ausgangspunkt der regelbasierten Systeme. Diese orientierten sich unmittelbar an der klassischen juristischen Prüfstruktur. Ihr Ziel war es, den juristischen Entscheidungsprozess

vollständig in maschinenlesbare Entscheidungsbaume und formalisierte Normlogik zu übertragen. Rechtsnormen wurden dazu in explizite Wenn-Dann-Regeln kodiert, die schrittweise prüften, ob Tatbestandsvoraussetzungen erfüllt waren, strikt deterministisch und regelbasiert (vgl. McCarty 1977; Stamper 1980). Gerade diese Ansätze offenbarten die Grenzen klassischer Regelmodelle bei komplexeren Normkonstellationen. Systeme müssen nicht nur formallogisch korrekt, sondern auch dogmatisch fundiert sein, das heißt die rechtlichen Begriffe, Wertungen und Systematiken korrekt abbilden (vgl. Susskind 1990). Ein Problem vieler früherer Expertensysteme.

2.1.2 Aktuelle Bewertungen und methodische Kritik

Trotz dieser Kritik sind regelbasierte Systeme nicht vollständig obsolet. Ihre Stärke liegt weiterhin in hoher Transparenz und Prüfbarkeit. Jede Einzelfallentscheidung lässt sich schrittweise nachvollziehen und auf normativer Ebene verifizieren (vgl. McCarty 1977; Stamper 1980). Mit zunehmender Komplexität der Rechtsanwendung treten jedoch deutliche Grenzen auf. Unbestimmte Rechtsbegriffe, konkurrierende Auslegungen oder abweichende Argumentationen in der Rechtsprechung lassen sich nur schwer vollständig formal abbilden (vgl. Bench-Capon et al. 2012). Um solche Varianten zu berücksichtigen, müssten sämtliche denkbaren Einzelfälle und Argumentationsmuster bereits im Vorfeld modelliert und kodifiziert werden (vgl. Hafner 1987). Hinzu kommt eine aufwändige Vorverarbeitung und Normzuordnung der Fallbeschreibungen, um sie in die formalisierte Prüfstruktur zu überführen. So kritisieren Gless und Wohlers im Konzept des Subsumtionsautomaten 2.0 explizit den Versuch, richterliche Entscheidungsprozesse vollständig

algorithmisch zu modellieren (vgl. Gless/Wohlers 2019). Die Autoren betonen, dass Subsumtion stets Wertung, Kontextbezug und Begründung erfordert. Eigenschaften, die sich nicht durch schematische Logik erfassen lassen. Diese fehlende Flexibilität und der hohe Modellierungsaufwand haben dazu geführt, dass rein regelbasierte Expertensysteme heute in der rechtspraktischen Anwendung praktisch keine Rolle mehr spielen. In der aktuellen Forschung gelten sie vor allem als methodischer Referenzpunkt, aus dem heraus sprachmodellbasierte und insbesondere hybride Verfahren entwickelt werden.

2.2 Rein sprachmodellbasierte Systeme

2.2.1 Grundprinzip und Defizite

Rein sprachmodellbasierte Systeme (LLM-only) wie GPT-3 nutzen große, allgemein vortrainierte Sprachmodelle, die juristische Entscheidungen auf Grundlage statistischer Sprachmuster und semantischer Ähnlichkeitsbeziehungen abbilden (vgl. Chalkidis et al. 2022; Zhong et al. 2018). Einzelne Subsumtionsschritte, wie etwa die Anwendung einer Definition auf einen Sachverhalt, können von Sprachmodellen mit einer gewissen Wahrscheinlichkeit korrekt abgebildet werden. Was jedoch fehlt, ist die konsistente Umsetzung ganzer juristischer Prüfschemata: Die Modelle folgen statistischen Wahrscheinlichkeiten, nicht einer nachvollziehbaren Prüfarchitektur. Rechtliche Argumentation erscheint damit nicht als dogmatisch strukturierte Prüfkette, sondern als statistisch plausibel erzeugter Text. Solche Systeme liefern zwar kohärent wirkende Ergebnisse, ihre Erklärungen bleiben jedoch oft rein technisch und genügen nicht den Anforderungen einer tragfähigen juristischen Begründung (vgl. Bibal et al. 2021; Valvoda/Cotterell 2024). Sprachmodelle nutzen zudem häufig oberflächliche Muster statt

inhaltlicher Argumentationsstrukturen und begünstigen damit systematische Verzerrungen („Bias“¹⁰) (vgl. Rudin 2019). Modelle erreichen dadurch zwar hohe Genauigkeitswerte, reagieren jedoch teils auf irrelevante Merkmale wie den Namen einer Vorinstanz (vgl. Santosh et al. 2024). Solche Erklärungen sind juristisch wertlos, wenn sie keinen Bezug zu normativen Prüfungspunkten haben. Diese Modelle werden daher häufig als „Black Box“¹¹ bezeichnet, da ihre innere Entscheidungslogik nicht nachvollziehbar ist.

2.2.2 Prompting-Ansätze

Neuere Arbeiten nutzen gezielte Promptgestaltung, um deduktive Argumentationsstrukturen sprachlich zu simulieren. Legal Syllogism Prompting (vgl. Jiang/Yang 2023) arbeitet mit einer dreigliedrigen Struktur (Rule → Fact → Conclusion), die juristische Prüfschritte sprachlich vorgibt. Ähnlich funktioniert „Chain-of-Logic Prompting“¹² (vgl. Servantez et al. 2024), das komplexere Argumentationsketten über mehrstufige logische Prompts simuliert. Auch Beyond Guilt (vgl. Zhang et al. 2025b) folgt diesem Ansatz, indem es die klassischen Subsumtionsstufen des Strafrechts (Tatbestand → Rechtswidrigkeit → Schuld) sequenziell abprüfen lässt. Die Prüfstruktur wird dabei nur sprachlich vorgegeben, eine explizite Entscheidungslogik wird nicht modelliert. Damit nähern sich diese Verfahren zwar hybriden Ansätzen an, verbleiben technisch jedoch im Bereich rein sprachmodellbasierter Systeme, die auf Wahrscheinlichkeiten statt expliziter Entscheidungslogik beruhen.

Die Qualität der vom Modell hervorgerufenen Begründungen (Rationales) schwankt, da die ausgewählten Textstellen häufig problematisch sind. Statt zentrale Argumente zu erfassen, markieren die Systeme oft allgemeine, unvollständige oder rechtlich irrelevante Textstellen. Um

die Auswahlqualität zu verbessern, werden Kriterien wie „Sparsity“ (möglichst wenige, aber wichtige Passagen) und „Comprehensiveness“ (Vollständigkeit der Argumentation) eingeführt (vgl. Chalkidis et al. 2021). Dennoch zeigt sich, dass Modelle, die in klassischen Klassifikationsmetriken wie dem F1-Score bei der Vorhersage von Entscheidungsausgängen über 90 % Genauigkeit erreichen, zentrale juristische Prüfschritte übersehen können (vgl. Steging et al. 2021). Technische Genauigkeit garantiert nicht juristische Nachvollziehbarkeit.

2.3 Fall- und vektorbasierte Ansätze

2.3.1 Grundprinzip

Fall- und vektorbasierte Systeme basieren nicht auf expliziten Regeln oder Ontologien, sondern auf Analogien und semantischen Ähnlichkeiten. Statt Normen vollständig zu formalisieren, vergleichen sie Fälle oder Sachverhalte miteinander, um daraus Argumentationsmuster abzuleiten. Ihre Stärke liegt in erhöhter Flexibilität, ihre Schwäche in der fehlenden Transparenz und der nur eingeschränkten dogmatischen Steuerbarkeit.

Ein fall- oder vektorbasierter Ansatz könnte beispielsweise einen Fall von Wohnungseinbruch mit einem früheren Diebstahlsfall vergleichen, weil in beiden Fällen die Wegnahme fremden Eigentums im Vordergrund steht. Die Ähnlichkeit ergibt sich dabei nicht aus einer expliziten Prüfung einzelner Tatbestandsmerkmale, sondern aus semantischer Nähe im Vektorraum. Das System „lernt“ aus solchen Vergleichen, wie in ähnlichen Konstellationen entschieden wurde, kann die zugrunde liegende dogmatische Struktur jedoch nicht transparent abbilden.

2.3.2 Fallbasierte Ansätze

Ein früher Vertreter dieses Ansatzes war Hafner, der bereits 1981 eine fallbasierte

Organisation juristischer Wissensbasen vorschlug (vgl. Hafner 1987). Anstatt sämtliche Normen in Wenn-Dann-Regeln zu formalisieren, nutzte er Analogien zwischen ähnlich gelagerten Fällen, um flexiblere Argumentationsmuster zu erfassen. Dieses Konzept erhöhte zwar die Ausdrucksfähigkeit, ging jedoch zulasten der Transparenz und gilt heute als konzeptioneller Vorläufer moderner vektorraumgestützter Verfahren.

2.3.3 Vektorbasierte Ansätze

Moderne Verfahren setzen auf latente Repräsentationen in hochdimensionalen Vektorräumen. Ein Beispiel ist LegalDuet (vgl. Xu et al. 2025), das ähnliche Norm-Fall-Paare im semantischen Raum näher zusammenliegen lässt als unähnliche. Dieses Verfahren wird als kontrastives Repräsentationslernen¹³ bezeichnet. Fälle werden dabei über Fallanalogien („Law Case Reasoning“) und Normvergleiche („Legal Ground Reasoning“) positioniert und semantisch gegenübergestellt. Die Subsumtionsentscheidung ergibt sich nicht aus expliziten Prüfstrukturen, sondern aus Ähnlichkeitsmessungen zwischen Vektorrepräsentationen von Sachverhalt und Norm. Um solche Abstände zuverlässig erlernen zu können, sind umfangreiche und sorgfältig annotierte Fall- und Normdatensätze erforderlich. Der argumentative Prozess bleibt jedoch latent und für Außenstehende schwer nachvollziehbar, da sich die Ähnlichkeitsbewertung auf den gesamten Fall bezieht und nicht auf einzelne Prüfschritte (vgl. Bibal et al. 2021; Katz et al. 2024).

2.3.4 Erklärungsorientierte Varianten

Eine stärker erklärungsorientierte Variante des vektorbasierten Ansatzes ist IOT-Match (vgl. Yu et al. 2022). Anstatt nur globale Ähnlichkeiten im Vektorraum zu berechnen, identifiziert dieses Verfahren

satzweise relevante Passagen („Rationales“) zwischen zwei Fällen. Die so ermittelten Zuordnungen („Alignments“) dienen nicht nur der Fallzuordnung, sondern auch der Generierung einer textlichen Begründung.

Damit verbindet der Ansatz die Leistungsfähigkeit vektorbasierten Ähnlichkeitslernens mit einer expliziten, lokal nachvollziehbaren Erklärungsebene. Die normative Prüfung bleibt jedoch auch hier unvollständig. Zwar können lokale Passagen plausibel zugeordnet werden, doch ersetzt dies keine systematische Abbildung der juristischen Prüfstruktur. Einen ähnlichen Ansatz verfolgen aufgabenspezifisch trainierte Sequenz-zu-Sequenz-Modelle¹⁴ (Seq2Seq) (vgl. Ye et al. 2018). Sie transformieren eine Eingabe, etwa eine Sachverhaltsschilderung, schrittweise in eine strukturierte Ausgabe wie eine Anklagebegründung. Technisch nutzen sie Encoder-Decoder-Architekturen¹⁵ mit Aufmerksamkeitsmechanismus („Attention“¹⁶). Der Encoder zerlegt den Eingabetext in einzelne Repräsentationen, gewichtet diese nach Relevanz, und der Decoder erzeugt daraus Schritt für Schritt eine strukturierte Ausgabe. Besonders wichtige Passagen, etwa Tatbestandsmerkmale, werden dadurch stärker berücksichtigt (vgl. Chalkidis et al. 2021).

2.3.5 Grenzen erklärungsorientierter Varianten

In ihrem spezifischen Bereich erreichen Seq2Seq-Modelle häufig gute Ergebnisse in NLP¹⁷-Benchmarks, etwa BLEU¹⁸- und ROUGE¹⁹-Werte, bei der Generierung von Anklagebegründungen (vgl. Ye et al. 2018) oder Micro-F1-Scores²⁰ von rund 70 % bei der Extraktion von Begründungspassagen (vgl. Chalkidis et al. 2021). Diese Kennzahlen wirken auf den ersten Blick hoch, erfassen aber nur die Übereinstimmung mit Referenztexten und sagen wenig über

die dogmatische Qualität der Begründung aus. Die Modelle bieten damit eine eingeschränkte Form der Nachvollziehbarkeit, verlieren jedoch bei Anwendung auf andere Rechtsgebiete oder Sprachen. Ihre Erklärbarkeit bleibt eng an den jeweiligen Anwendungsfall gebunden. Die erzeugten Begründungen liefern lediglich textuelle Approximationen einzelner Argumentationsschritte.

2.4 Wissensrepräsentationsbasierte Systeme

2.4.1 Grundprinzip und Beispiele

Wissensrepräsentationsbasierte Systeme verfolgen einen grundsätzlich anderen Ansatz als Expertensysteme, LLM-basierte Verfahren oder fall- bzw. vektorbasierte Ansätze. Während Expertensysteme die Prüflogik als starre Abfolge von Entscheidungsregeln modellieren und vektorbasierte Verfahren auf latente Ähnlichkeitsbeziehungen setzen, strukturieren wissensrepräsentationsbasierte Systeme das juristische Fachwissen explizit in Ontologien, Begriffen und Relationen und leiten daraus logische Schlussfolgerungen ab. Im Mittelpunkt steht nicht der Prüfablauf oder die Fallanalogie, sondern die Modellierung juristischer Begriffe, Normen und ihrer Beziehungen in Ontologien, semantischen Graphen oder logikbasierten Wissensbasen. Die Subsumtion erfolgt anschließend durch logische Schlussfolgerungen („Reasoning“), die aus diesen Wissensstrukturen abgeleitet werden. So lassen sich Normen, Begriffsdefinitionen, Tatbestandsmerkmale und ihre Abhängigkeiten explizit erfassen und flexibel kombinieren. Damit werden juristische Prüfarchitekturen formal nachgebildet, jedoch nicht in dynamischen Entscheidungspfaden, sondern in Wissensstrukturen, die erst durch logische Schlussfolgerungen aktiviert werden. Beispiele sind Systeme

wie ChatLaw (vgl. Cui et al. 2023), „Ontology-Driven Reasoning about Property Crimes“ (vgl. Navarrete et al. 2025) oder JUREX-4E (vgl. Liu et al. 2025), die strafrechtliche Tatbestände und Subsumtionsstrukturen formal in Wissensbasen überführen.

2.4.2 Stärken und Grenzen

Die Stärke wissensrepräsentationsbasierter Systeme liegt in ihrer hohen Transparenz und logischen Nachvollziehbarkeit. Entscheidungsstrukturen werden formalisiert und sind damit für Außenstehende nachvollziehbar. In der Praxis liefern die Systeme jedoch häufig nur lokale Erklärungen für den Einzelfall, etwa indem sie hervorheben, welche einzelnen Merkmale (z.B. bestimmte Tatbestandsmerkmale) das Ergebnis besonders beeinflusst haben. Solche Hinweise geben keinen Überblick über alternative Entscheidungswege oder mögliche Begründungslinien (vgl. Deng et al. 2024). Hinzu kommt, dass ihre vollständige Modellierung erheblichen Vorabaufwand bei der Normstrukturierung und Wissensmodellierung erfordert und nur eingeschränkt auf andere Rechtsordnungen übertragbar ist (vgl. Bench-Capon et al. 2012). Auch in sprachlich variablen oder komplexen, argumentativ geprägten Fällen stoßen die Systeme schnell an Grenzen (vgl. Palmirani et al. 2011; Morris 2021). Reinen Ontologie-basierten Ansätzen fehlt zudem oft die Fähigkeit, unstrukturierte Fallbeschreibungen automatisch in die Entscheidungslogik einzubinden. Das bedeutet, auch wenn diese Systeme logisch konsistente Ergebnisse liefern, ist die zugrunde liegende dogmatische Wertung nicht zwangsläufig auch korrekt erfasst (vgl. Bibal et al. 2021). Rein technische Evaluationsmetriken reichen daher auch hier nicht aus.

2.5 Hybride Systeme mit Entscheidungslogik

2.5.1 Generische hybride Ansätze

Hybride Ansätze betten Sprachmodelle in explizite Entscheidungsrahmen ein, um den Subsumtionsprozess schrittweise rekonstruierbar und nachvollziehbar zu machen. Ziel ist es, die Ausdruckstärke großer Sprachmodelle mit der formalen Struktur juristischer Prüfung zu verbinden. Ein Beispiel ist SyLeR (vgl. Zhang et al. 2025a), das den Subsumtionsprozess entlang einer festen Argumentationsstruktur aus „Legal Rule“, „Fact“, „Reasoning“ und „Conclusion“ organisiert. Damit knüpft es an didaktisch etablierte Schemata wie IRAC²¹ („Issue – Rule – Application – Conclusion“) an, die eine klare Zerlegung des juristischen Prüfungswegs vorsehen (vgl. Bench-Capon 2020). Anders als klassische Entscheidungsbäume bildet SyLeR jedoch keine vollständige Prüfkette ab. Stattdessen wählt es relevante Ober- und Untersätze kontextgesteuert aus und präsentiert diese dem Sprachmodell. Die eigentliche Subsumtion erfolgt durch das LLM, das über „Reinforcement Learning“²² systematisch auf argumentative Kohärenz und Plausibilität optimiert wird. Einen verwandten, jedoch noch stärker rechtsgebietsbezogenen Ansatz verfolgt Beyond Guilt (vgl. Zhang et al. 2025b), das speziell für das Strafrecht die klassischen Subsumtionsstufen Tatbestand, Rechtswidrigkeit und Schuld als eigenständige Entscheidungsstufen abbildet. Während IRAC oder SyLeR eher generische Strukturen vorsehen, orientiert sich Beyond Guilt enger an der Architektur des Strafgesetzbuchs. Auch hier übernimmt das Sprachmodell die Bewertung der Einzelfrage, während die Prüfstruktur durch die gesetzliche Systematik vorgegeben ist. Damit nähern sich hybride Systeme der dogmatischen Prüfarchitektur

weiter an. Sie machen die Struktur sichtbar, überlassen die eigentliche Subsumtion jedoch weiterhin einem Sprachmodell, das nicht in jedem Schritt konsequent an die juristische Logik gebunden ist.

2.5.2 Strukturierte Hybridmodelle

Einen anderen Weg verfolgt „Explainable Legal Judgment Prediction via Concept Tree and Concept Forest Reasoning with Collegiate Bench Mechanism“²³ (vgl. Deng et al. 2024). Hier werden juristische Argumentationen in baumartige Teilstrukturen zerlegt, sodass unterschiedliche Aspekte eines Falls durch spezialisierte Modelle analysiert werden. Die Ergebnisse werden anschließend durch einen Collegiate Bench Mechanism zusammengeführt, um argumentative Konsistenz und abweichende Bewertungen zu berücksichtigen. Das Verfahren greift damit eine in der juristischen Fachkultur zentrale Praxis auf. Auch Kollegialgerichte fällen ihre Entscheidungen nicht einstimmig, sondern spiegeln häufig ein Spannungsfeld aus herrschenden und abweichenden Meinungen wider. Während der Collegiate Bench Mechanism diese Pluralität technisch nachbildet, bleibt die normative Bewertung, welcher Argumentation zu folgen ist, weiterhin eine offene Herausforderung. Auch hier gilt: Die dogmatische Prüfarchitektur wird nur teilweise abgebildet. Zwar werden Argumentationslinien strukturiert zusammengeführt, doch die verbindliche Festlegung, welche Linie dogmatisch tragfähig ist, bleibt ungelöst.

2.5.3 Stärken und Grenzen hybrider Systeme

Allen hybriden Verfahren ist gemeinsam, dass sie Elemente juristischer Prüfschemata explizit nachbilden, die dogmatische Tiefe jedoch nur teilweise erfassen. Sie bieten einen Mittelweg zwischen Flexibilität und Nachvollziehbarkeit und erhöhen

die Transparenz gegenüber rein sprachmodellbasierten Ansätzen, bleiben aber mit zentralen Defiziten behaftet. Empirische Untersuchungen zeigen, dass Modelle trotz formaler Rahmung nicht konsequent an juristische Prüfstrukturen gebunden sind.

Ein System kann etwa korrekt eine Wegnahme als Tatbestandsmerkmal des Diebstahls identifizieren, dabei jedoch die Prüfung der Zueignungsabsicht auslassen. Das Ergebnis wirkt plausibel, ist juristisch aber unvollständig (vgl. Steging et al. 2021). Hinzu kommt, dass Evaluationsmetriken wie F1-Score oder BLEU zwar sprachliche Übereinstimmung messen, aber wenig über die juristische Qualität der Begründung aussagen (vgl. Ashley 2017). Erst die Kombination technischer Metriken mit Expertenbewertungen erlaubt eine belastbare Einschätzung der tatsächlichen Tragfähigkeit solcher Systeme (vgl. Bibal et al. 2021; Valvoda/Cotterell 2024).

Im Vergleich zu rein sprachmodellbasierten Verfahren zeigt sich jedoch, dass hybride Systeme trotz aller Grenzen einen Schritt weiter gehen. Sie bilden die normative Struktur juristischer Argumentation zumindest teilweise ab. Gleichwohl bleiben sie abhängig von sorgfältiger Modellierung und domänenspezifischer Anpassung und laufen Gefahr, bei unvollständiger Abdeckung der Prüfschritte eine trügerische Scheinkohärenz zu erzeugen (vgl. Chung et al. 2024).

3. FORSCHUNGSBEDARF

3.1 Stärken und Grenzen bestehender Systeme

Die bisherigen Ansätze verdeutlichen ein Spannungsfeld: Entweder sind Systeme transparent und dogmatisch steuerbar, dann aber starr, oder sie sind flexibel und leistungsfähig, bilden jedoch die juristische Prüfarchitektur nicht konsistent ab.

- ▶ Regelbasierte Expertensysteme gewährleisten maximale Nachvollziehbarkeit, sind jedoch starr und nur begrenzt auf komplexe oder offene Rechtsfragen übertragbar.
- ▶ Rein sprachmodellbasierte Systeme (LLM-only) bieten hohe Flexibilität und Ausdrucksstärke, entziehen sich aber der dogmatischen Steuerbarkeit und sind nur eingeschränkt überprüfbar.
- ▶ Fall- und vektorbasierte Ansätze erhöhen die Ausdrucksfähigkeit durch Analogien und Ähnlichkeitsmessungen, gehen jedoch zulasten der Transparenz, da die Begründungslogik latent bleibt.
- ▶ Wissensrepräsentationsbasierte Systeme ermöglichen dogmatische Steuerbarkeit und semantische Präzision, erfordern jedoch erheblichen Modellierungsaufwand und bleiben sprachlich wenig flexibel.
- ▶ Hybride Architekturen verbinden formale Prüfstrukturen mit den Generationsfähigkeiten von LLMs, befinden sich jedoch noch im Anfangsstadium und sind methodisch nicht ausgereift.

Der Forschungsbedarf ergibt sich damit weniger aus einem Mangel an Ansätzen, sondern aus der Notwendigkeit, ihre jeweiligen Stärken systematisch zu kombinieren. Entscheidend ist, formale Strukturtreue, sprachliche Flexibilität und dogmatische Nachvollziehbarkeit in einer kohärenten Architektur der Prüfschemata zusammenzuführen.

3.2 Technische Defizite und Explainability by Design als rechtsstaatliches Gebot

In sensiblen Bereichen sollten sogenannte „Black-Box-Modelle“, deren innere Entscheidungslogik für Außenstehende nicht nachvollziehbar ist, durch intrinsisch interpretierbare Systeme ersetzt werden (vgl. Rudin 2019).

Gemeint sind Verfahren, deren Entscheidungslogik von vornherein transparent ist, etwa Entscheidungsbäume oder regelba-

sierte Systeme, nicht bloß nachträglich erzeugte Erklärungen. Generative Sprachmodelle wie GPT-3 lassen sich zwar nicht vollständig öffnen, können aber in feste juristische Prüfschemata eingebettet werden. Dabei übernimmt das LLM nur die Subsumtion einzelner Schritte, während Definitionen, Ontologien, Rechtsprechung und bisherige Prüfschritte vorgegeben sind.

Die zentrale Herausforderung liegt jedoch nicht in der Reproduktion korrekter Ergebnisse, sondern in der Abbildung des juristischen Prüfungswegs. Rechtliche Normen sind über Jahrzehnte zu konsistenten, logisch geschlossenen Strukturen entwickelt worden. Aufgabe technischer Systeme ist es daher, diese Prüfschritte transparent nachzuvollziehen, nicht nur plausible Resultate zu liefern. Gängige Benchmarks erfassen dies kaum und viele XAI-Methoden können die rechtliche Anforderung an Transparenz und Rechenschaftspflicht nicht erfüllen (vgl. Waller et al. 2024). Sie messen Ergebnisgenauigkeit, nicht aber die Nachvollziehbarkeit des juristischen Prüfprozesses (vgl. Bibal et al. 2021; Eriksson et al. 2025).

Deshalb muss Erklärbarkeit von Beginn an in die Architektur integriert werden. Erst die Kombination technischer und juristischer Evaluationsmaßstäbe erlaubt eine belastbare Einschätzung (vgl. Valvoda/Cotterell 2024). „Explainability by Design“ darf nicht nachträglich simuliert werden, sondern muss ein konstitutives Element technischer Systeme sein, analog zu etablierten Prinzipien wie „Security by Design“, „Safety by Design“ oder „Ethics by Design“ (vgl. Council of Europe, Parliamentary Assembly 2020; Amariles/Troussel 2025; Gold et al. 2025). Jenseits der Technik bleibt jedoch offen, welche Maßstäbe die Rechtswissenschaft selbst an die Qualität einer Begründung anlegt. Ohne solche Standards kann Explainability by Design nicht eingelöst werden.

3.3 Strukturelle Defizite und offene Diskussionslinien

3.3.1 Maßstäbe der rechtlichen

Begründung

Während in der Softwareentwicklung technische Systeme auf klar definierte Anforderungen hin entwickelt werden, fehlt es in der KI-gestützten Rechtsanwendung noch an vergleichbaren Maßstäben. Insbesondere ist unklar, was als ausreichende Begründung gelten soll (vgl. Coeckelbergh 2019; Atkinson/Bench-Capon 2021). Ohne ein solches juristisches Lastenheft laufen technische Lösungen Gefahr, sich an den rechtsstaatlichen Bedürfnissen vorbeizuentwickeln (vgl. Hildebrandt 2020; Bibal et al. 2021). Doch selbst wenn Systeme künftig in der Lage wären, normative Prüfschritte vollständig und transparent abzubilden, bleibt zu klären, wann eine Argumentation als hinreichend gilt. Reicht es, wenn sie der herrschenden Meinung entspricht, oder müssten auch Mindermeinungen berücksichtigt werden? Soll ein System nur eine Entscheidung präsentieren oder mehrere Varianten mit Wahrscheinlichkeiten? Wer legt fest, welcher Maßstab als Vergleich dient, einzelne Richter, eine bestimmte Instanz oder die Mehrheitspraxis? Solche Fragen lassen sich nicht technisch lösen, sondern erfordern klare Vorgaben der Rechtswissenschaft (vgl. Gless/Wohlers 2019; Wischmeyer 2020; Deeks 2019).

Allgemeine XAI-Kataloge wie der von Fresz et al. umfassen Kriterien wie „Correctness“, „Completeness“, „Context“ und „Counterability“ (vgl. Fresz et al. 2024).

Diese Ansätze lassen sich auf juristische Prüfschemata übertragen und eröffnen damit einen möglichen Weg zu einem rechtsstaatlich tragfähigen Lastenheft. „Explainability by Design“ muss dabei über technische Transparenz hinaus auch rechtsstaatliche Begründungspflichten, Zielgruppenadäquanz und Anfechtbarkeit

berücksichtigen (vgl. Maxwell/Dumas 2023; Herrewijnen et al. 2024). Ein juristisch normiertes Lastenheft, das solche Kriterien systematisch übersetzt, könnte einen Ausgangspunkt für Standardisierung und empirische Evaluation bieten. Darin könnten etwa Mindestanforderungen an Prüfschemata (explizite, maschinenlesbare Abbildung der Prüfschritte), Pflichten zur Alternativdarstellung (sichtbare Mindermeinungen bei Auslegungsvarianten) oder Anforderungen an einen prüfbareren Audit-Trail²⁴ (maschinell nachvollziehbare Protokolle) verankert sein. Forschungsbedarf besteht vor allem in der Ausgestaltung eines solchen Anforderungskatalogs, einschließlich der Frage, in welchen Einsatzszenarien Berufsgruppen oder Bürger Zustimmung signalisieren oder Ablehnung äußern (vgl. Casey/Lemley 2020) und wie Systeme zugleich den Zugang zum Recht erleichtern können.

Während Fresz et al. primär technische Eigenschaften erklärbarer KI-Verfahren aus bestehenden Rechtsregimen ableiten, überführt das hier vorgeschlagene Lastenheft diese Überlegungen in eine juristisch-normative Perspektive. Es adressiert nicht nur die Eigenschaften einzelner Erklärverfahren, sondern die strukturellen Anforderungen an ein Systemdesign, das rechtliche Nachvollziehbarkeit und Begründungspflichten intrinsisch absichert. Damit verschiebt sich der Fokus von der bloßen „Explainability of Methods“ hin zu einer „Explainability by Design“, die technische, organisatorische und rechtsstaatliche Komponenten gleichermaßen integriert. Abbildung 1 (siehe Seite 96) skizziert beispielhaft zentrale Bausteine eines solchen Lastenhefts, die juristische Anforderungen mit technischen Umsetzungshinweisen verbinden. Die Darstellung ist kein vollständiger Standard, sondern ein Vorschlag, der als Grundlage für künftige Standardisierung und empirische Evaluation dienen kann.

Quelle: Herrmann/Breitinger (eig. Darstellung)

Baustein	Kernanforderung	Hinweis zur Umsetzung
Mindestanforderungen an Prüfschemata	Juristische Standardprüfstrukturen müssen explizit, maschinenlesbar und schrittweise abgebildet sein.	Prüfschritte als formale Knoten/Kanten, LLM nur für Einzelsubsumtionen.
Kriterien für juristische Erklärbarkeit	Begründung muss Normbasis, erfüllte Tatbestandsmerkmale, Streitfragen und ggf. abweichende Meinungen sichtbar machen.	Kurzschema: Welche Norm? Welche Merkmale? Welche Streitfragen? Warum diese Linie?
Pflicht zur Alternativdarstellung	Bei Mehrdeutigkeit sind Alternativbegründungen auszuweisen (mit Gründen, Wahrscheinlichkeiten optional).	Nebenlösung/Abweichung als eigener Pfad mit kurzer Begründung.
Prüfprotokoll/Audit Trail	Jeder Prüfschritt ist maschinenlesbar protokolliert (Eingaben, Quellen, Heuristiken, Entscheidung).	Exportierbares JSON/XML-Protokoll, externe Nachprüfbarkeit.
Ontologien und Benchmarks	Begriffe, Definitionen, Normrelationen müssen in Ontologien definiert sein. Benchmarks koppeln technische Scores mit juristischen Bewertungen.	Referenz-Ontologie (Versionierung, Zitationsfähigkeit); „Tech+Jura“-Doppelmetrik.
Governance und Rollenverständnis	Es muss geklärt sein, wer Standards setzt, pflegt und ändert, Einsatzmodus (Assistenz vs. Entscheidung) muss klar definiert sein.	Fachgremium, Änderungsprotokoll, UI-Kennzeichnung.

Abb. 1: Beispielhafte Bausteine eines juristisch normierten Lastenhefts für „Explainability by Design“

3.3.2 Maschinenlesbare Gesetzgebung als Spiegel rechtsstaatlicher Qualität

Wenn KI-Systeme juristische Subsumtion nur schwer abbilden können, weil Gesetze unklar, widersprüchlich oder übermäßig komplex formuliert sind, verweist das weniger auf eine Schwäche der Technik als auf ein Defizit des Rechts. Schon frühe Arbeiten wie das „British Nationality Act“-Projekt zeigten, dass sich rechtliche Strukturen grundsätzlich formal modellieren lassen (vgl. Sergot et al. 1986). Neuere Ansätze wie „Rules as Code“²⁵ (vgl. Morris 2021) oder standardisierte Repräsentationsformate wie LegalRuleML (vgl. Palmirani et al. 2011) und das Web-Ontology-Language-Modell (OWL) (vgl. Antoniou/van Harmelen 2009) verdeutlichen, dass maschinenlesbare Gesetzgebung technisch längst umsetzbar ist. Statt ausschließlich zu fragen, wie KI bestehendes Recht nachvollziehen kann, wäre die Gegenfrage naheliegend: Warum gestalten wir Gesetze nicht so, dass sie auch maschinell interpretierbar sind? Eine rechtsstaatlich normierte, ma-

schinenlesbare Gesetzgebung könnte nicht nur KI-gestützte Prüfverfahren erleichtern, sondern zugleich den Zugang zum Recht verbessern. Angesichts aktueller Regelwerke wie dem EU-AI-Act, die selbst für Juristinnen und Juristen nur schwer durchschaubar sind (vgl. Hanif et al. 2024), stellt sich die Frage, ob nicht die wachsende Komplexität des Rechts selbst zum Hemmnis rechtsstaatlicher Verständlichkeit geworden ist.

3.3.3 Technologische Waffengleichheit im Rechtsstaat

Die technologische Entwicklung schreitet schneller voran als ihre rechtliche Einbettung (vgl. Susskind 2017; ders. 2019). KI-gestützte Systeme werden sich nicht in theoretischen Pilotprojekten, sondern in der Praxis bewähren, etwa auf Seiten der Verteidigung oder des privaten Sektors (vgl. Schwarcz et al. 2025). Wenn staatliche Akteure zögern, ihre Verfahren technisch aufzurüsten, könnte sich die Überlegenheit solcher Systeme genau dort zeigen, wo sie den Staat faktisch ins Hintertreffen

geraten lassen, nicht durch Rechtsbruch, sondern durch Effizienz, Präzision und den gezielten Einsatz von Informationsvorsprung.

Die entscheidende Frage lautet daher nicht mehr, ob KI-Systeme in der Rechtsanwendung eingesetzt werden, sondern wer sie zuerst beherrscht. Verteidiger und Großkanzleien werden diese Technologien ohnehin nutzen, um Schwachstellen in Argumentationen zu identifizieren oder Verfahrensstrategien zu optimieren (vgl. Gold et al. 2025). Daraus kann ein Ungleichgewicht entstehen, das ohne technische Unterstützung von Staatsanwaltschaften und Gerichten kaum auszugleichen ist. Ein einfaches Beispiel verdeutlicht dies: Mit einem Prompt an ein Sprachmodell lassen sich in Sekunden Angriffspunkte gegen eine Anklageschrift simulieren. Wenn nur eine Partei über solche Werkzeuge verfügt, droht eine Verletzung des Prinzips der Waffengleichheit (vgl. Council of Europe 1950; 2020).

4. CONCLUSIO

Während sich die rechtswissenschaftliche Debatte noch überwiegend auf die Klärung der Maßstäbe konzentriert, eröffnet ein Ansatz, der Erklärbarkeit intrinsisch im Design berücksichtigt, die Möglichkeit, Standards nicht nur theoretisch zu diskutieren, sondern praktisch zu erproben. Damit verschiebt sich die Diskussion vom Ob und Wo des KI-Einsatzes zur Frage,

wie Systeme gestaltet werden können, um rechtsstaatlichen Ansprüchen zu genügen. Die meisten XAI-Verfahren liefern lediglich technische Transparenz, nicht aber eine rechtsstaatlich tragfähige Begründung. Eine Erklärung im Rechtssinne muss zeigen, wie Daten in eine normativ vorgegebene Prüfstruktur eingeordnet werden. Genau hier liegt die Sollbruchstelle zwischen technischer XAI und juristischer Praxis. Die Fähigkeit, Fakten zu erkennen, ersetzt nicht die Fähigkeit, sie systematisch in ein Prüfschema einzufügen. Diese Lücke markiert den eigentlichen Entwicklungsbedarf. Statt Erklärbarkeit als nachgelagerte Zusatzschicht zu behandeln, muss die rechtliche Prüfstruktur selbst Teil des Modells sein. Erst wenn jeder Prüfschritt, von der Tatsachenfeststellung bis zur rechtlichen Würdigung, in klarer, maschinenlesbarer Form vorliegt, können Systeme Begründungen liefern, die sowohl technisch als auch juristisch tragbar sind. Doch Technik allein kann diese Maßstäbe nicht definieren. Erforderlich sind rechtliche und gesellschaftliche Vorgaben, die bestimmen, welche Argumentationstiefe und Begründungstiefe rechtsstaatlich hinreichend ist. „Explainability by Design“ ist daher nicht nur ein technisches Konzept, sondern ein rechtspolitisches Gestaltungsprinzip. Es zwingt dazu, die normative Logik des Rechts selbst explizit zu modellieren und wirkt so zugleich als Stresstest für bestehende Normstrukturen.

¹ *Explainable Artificial Intelligence (XAI) bezeichnet eine Forschungsrichtung, die darauf abzielt, Entscheidungen von KI-Systemen nachvollziehbar und überprüfbar zu machen.*

² *Explainability by Design stellt einen Ansatz dar, nach dem Erklärbarkeit nicht nachträglich hinzugefügt, sondern von*

Beginn an als Strukturprinzip in das Systemdesign integriert wird.

³ *Vektorrepräsentationen bilden Wörter oder Sätze als Punkte in einem mehrdimensionalen Raum ab, um Ähnlichkeiten zwischen Begriffen zu berechnen.*

⁴ *Legal Judgment Prediction (LJP) bezeichnet die Aufgabe, gerichtliche Ent-*

scheidungen oder Urteile auf Basis der Falldarstellung vorherzusagen.

⁵ *Unter einem Large Language Model (LLM) versteht man ein sehr großes Sprachmodell, das mithilfe neuronaler Netze trainiert wird, um menschenähnlichen Text zu verstehen und zu erzeugen.*

⁶ *GPT-Modelle (Generative Pretrained*

Transformer) sind eine Familie großer Sprachmodelle, die auf dem sogenannten Transformer-Architekturprinzip beruhen und in mehreren Versionen (z.B. GPT-2, GPT-3, GPT-4) veröffentlicht wurden. Sie werden auf umfangreichen Textkorpora vortrainiert, um anschließend menschenähnliche Texte zu erzeugen und Aufgaben wie Übersetzung, Zusammenfassung oder Argumentation zu bewältigen.

⁷ Benchmarks sind standardisierte Vergleichstests zur Bewertung der Leistungsfähigkeit von KI-Modellen anhand festgelegter Aufgaben oder Datensätze.

⁸ Prompting meint die gezielte Eingabe von Anweisungen („Prompts“), um Sprachmodelle zu bestimmten Antworten zu veranlassen.

⁹ Ontologien sind formale, maschinenlesbare Darstellungen von Begriffen und deren Beziehungen innerhalb eines Wissensbereichs, etwa des Rechts.

¹⁰ Bias bezeichnet eine systematische Verzerrung in Daten oder Modellen, die zu unfairen oder fehlerhaften Ergebnissen führen kann.

¹¹ Ein Black-Box-Modell ist ein System, dessen innere Entscheidungslogik für Nutzerinnen und Nutzer nicht nachvollziehbar ist.

¹² Chain-of-Logic Prompting ist eine spezielle Form davon, bei der logische Zwischenschritte explizit dargestellt werden, um die Nachvollziehbarkeit zu erhöhen.

¹³ Beim kontrastiven Lernen wird ein Modell darauf trainiert, ähnliche Paare von Eingaben näher zusammen und unähnliche weiter auseinander zu platzieren.

¹⁴ Sequence-to-Sequence-Modelle (Seq2Seq) beruhen auf einer Architektur neuronaler Netze, die Eingabesequenzen (z.B. Texte) schrittweise in Ausgabesequenzen umwandelt; häufig verwendet für Übersetzung oder Textgenerierung.

¹⁵ Eine Encoder-Decoder-Architektur ist die Grundstruktur vieler moderner

Sprachmodelle, bei der ein Encoder Eingabetexte in Repräsentationen umwandelt und ein Decoder daraus neue Texte generiert.

¹⁶ Ein Attention-Mechanismus stellt ein Verfahren in neuronalen Netzen dar, das besonders relevante Textteile stärker gewichtet, um die Qualität der erzeugten Ausgabe zu verbessern.

¹⁷ NLP (Natural Language Processing) ist das Teilgebiet der Künstlichen Intelligenz, das sich mit der automatischen Verarbeitung und Analyse natürlicher Sprache durch Computer beschäftigt.

¹⁸ Der BLEU-Wert (Bilingual Evaluation Understudy) ist eine automatische Metrik zur Bewertung der Qualität maschinell erzeugter Texte, insbesondere Übersetzungen, im Vergleich zu menschlichen Referenztexten.

¹⁹ Der ROUGE-Wert (Recall-Oriented Understudy for Gisting Evaluation) ist eine Metrik zur Bewertung von automatisch generierten Zusammenfassungen durch Vergleich mit Referenztexten.

²⁰ Der Micro-F1-Score kombiniert Präzision und Trefferquote (Recall) und gewichtet dabei alle Klassen gleich stark, unabhängig von deren Größe.

²¹ Das IRAC-Schema ist ein klassisches didaktisches Schema der juristischen Falllösung: „Issue – Rule – Application – Conclusion“ (Frage – Regel – Anwendung – Schlussfolgerung).

²² Reinforcement Learning bezeichnet das Lernverfahren, bei dem ein Modell durch Belohnung oder Bestrafung trainiert wird, gewünschte Entscheidungen zu verstärken.

²³ Der Collegiate Bench Mechanism simuliert in KI-Systemen den Austausch und die Abstimmung mehrerer Modelle analog zu Richtergerichten in Kollegialgerichten.

²⁴ Ein Audit Trail ist eine nachvollziehbare, maschinenlesbare Protokollspur,

die dokumentiert, wie ein System zu einer bestimmten Entscheidung gelangt ist.

²⁵ Rules as Code beschreibt das Konzept, juristische Normen so zu formulieren, dass sie unmittelbar von Maschinen verarbeitet und geprüft werden können.

Quellenangaben

Amariles, David Restrepo/Troussel, Aurore (2025). *The Regulatory Approach of the European Union's Artificial Intelligence Act*, in: Verein für Recht und Digitalisierung e.V./Universität Trier/Institut für Recht und Digitalisierung Trier (Hg.), *Artificial Intelligence and Fundamental Rights: The AI Act of the European Union and its implications for global technology regulation*, Trier, 111–128.

Antoniou, Grigoris/van Harmelen, Frank (2009). *Web Ontology Language: OWL*, in: Staab, Steffen/Studer, Rudi (Eds.), *Handbook on Ontologies*, Berlin/Heidelberg, 91–110.

Ashley, Kevin D. (2017). *Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age*, Cambridge.

Atkinson, Katie/Bench-Capon, Trevor (2021). *Argumentation Schemes in AI and Law, Argument & Computation*, 12 (3), 417–434.

Bench-Capon, Trevor et al. (2012). *A History of AI and Law in 50 Papers: 25 Years of the International Conference on AI and Law, Artificial Intelligence and Law*, 20 (3), 215–319.

Bench-Capon, Trevor (2020). *Explaining Legal Decisions Using IRAC, Computational Models of Natural Argument 2020*, Perugia, *CEUR Workshop Proceedings* 2669, 74–83.

Bench-Capon, Trevor J. M. et al. (1987). *Logic Programming for Large Scale Applications in Law: A Formalisation of Supplementary Benefit Legislation, Proceedings of the 1st International Con-*

- ference on Artificial Intelligence and Law (ICAIL), Boston, 190–198.
- Bibal, Adrien et al. (2021). Legal requirements on explainability in machine learning, *Artificial Intelligence and Law*, 29 (2), 149–169.
- Brown, Tom et al. (2020). Language Models Are Few-Shot Learners, in: Larochelle, Hugo et al. (Eds.), *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 1877–1901.
- Casey, Bryan/Lemley, Mark A. (2020). You Might Be a Robot, *Cornell Law Review*, 105 (2), 287–318.
- Chalkidis, Ilias et al. (2021). Paragraph-level Rationale Extraction through Regularization: A case study on European Court of Human Rights Cases, in: Toutanova, Kristina et al. (Eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2021)*, 226–241.
- Chalkidis, Ilias et al. (2022). LexGLUE: A Benchmark Dataset for Legal Language Understanding in English, in: Muresan, Smaranda et al. (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, 4310–4330.
- Chung, Neo Christopher et al. (2024). False Sense of Security in Explainable Artificial Intelligence, *IJCAI 2024 Workshop „AI Governance: Alignment, Morality and Law“*, Online: <https://openreview.net/forum?id=FKdewXF9t7> (27.10.2025).
- Coeckelbergh, Mark (2019). Artificial Intelligence: Some Ethical Issues and Regulatory Challenges, *Technology and Regulation*, 31–34.
- Cole, Michael (2024). From Code to Conduct: Ethical Considerations for AI in Legal Practice, *Westlaw Today*, Online: <https://www.reuters.com/legal/legalindustry/code-conduct-ethical-considerations-ai-legal-practice-2024-08-13/> (27.10.2025).
- Council of Europe (1950). *Convention for the Protection of Human Rights and Fundamental Freedoms (European Convention on Human Rights)*, Art. 6, Council of Europe Treaty Series No. 005, Online: https://www.echr.coe.int/documents/convention_eng.pdf (27.10.2025).
- Council of Europe, Parliamentary Assembly (2020). *Justice by Algorithm: Artificial Intelligence, Information Technology and Fundamental Rights*, Report Doc. 15282, Online: <https://pace.coe.int/en/files/28723/html> (27.10.2025).
- Cui, Jiayi et al. (2023). ChatLaw: Open-Source Legal Large Language Model with Integrated External Knowledge Bases, *arXiv Preprint*, Online: <https://arxiv.org/abs/2306.16092> (27.10.2025).
- Deeks, Ashley (2019). The Judicial Demand for Explainable Artificial Intelligence, *Columbia Law Review*, 119 (7), 1829–1850.
- Deng, Wei et al. (2024). Explainable Legal Judgment Prediction via Concept Tree and Concept Forest Reasoning with Collegiate Bench Mechanism, *Artificial Intelligence and Human-Computer Interaction*, Amsterdam, 194–201.
- Engel, Christoph (2022). Judicial Decision-Making: A Survey of the Experimental Evidence, *Max Planck Institute for Research on Collective Goods*, Online: <https://hdl.handle.net/21.11116/0000-000A-E4EF-8> (27.10.2025).
- Eriksson, Maria et al. (2025). AI Benchmarks: Interdisciplinary Issues and Policy Considerations, European Commission, Joint Research Centre, Luxemburg.
- Floridi, Luciano/Cowls, Josh (2019). A Unified Framework of Five Principles for AI in Society, *Harvard Data Science Review* (1), Online: <https://doi.org/10.1162/99608f92.8cd550d1> (27.10.2025).
- Fresz, Benjamin et al. (2024). How Should AI Decisions Be Explained? Requirements for Explanations from the Perspective of European Law, *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, Montreal, 438–450.
- Gless, Sabine/Wohlers, Wolfgang (2019). Subsumtionsautomat 2.0 – Künstliche Intelligenz statt menschlicher Rechtsanwendung?, in: Böse, Martin et al. (Hg.), *Festschrift zum 70. Geburtstag von Urs Kindhäuser*, Baden-Baden, 147–165.
- Gold, Valentin et al. (2025). Milestone 1: Abschlussbericht – Forschungsprojekt „Künstliche Intelligenz und richterliche Entscheidungsfindung“, Niedersächsisches Justizministerium in Kooperation mit der Universität Göttingen, Online: https://reusz.eu/upload/maki_m1abschlussbericht.pdf (27.10.2025).
- Grosz, Benjamin N. et al. (2003). Description Logic Programs: Combining Logic Programs with Description Logic, *Proceedings of the 12th International Conference on World Wide Web (WWW '03)*, Budapest, 48–57.
- Guha, Neel et al. (2023). LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models, in: Oh, Alice et al. (Eds.), *Advances in Neural Information Processing Systems 36*, New York, 44123–44279.
- Hafner, Carole D. (1987). *Conceptual Organization of Case Law Knowledge Bases*, *Proceedings of the 1st International Conference on Artificial Intelligence and Law (ICAIL '87)*, Boston, 35–42.
- Hanif, Hilmy et al. (2024). Navigating the EU AI Act Maze using a Decision-Tree Approach, *ACM Journal on Responsible Computing*, 1 (3), 1–16, Online: <https://doi.org/10.1145/3677174> (27.10.2025).
- Herrewijnen, Elize et al. (2024). Requirements and Attitudes towards Explainable AI in Law Enforcement, *Proceedings of the 2024 ACM Designing Interactive*

- Systems Conference (DIS '24)*, Copenhagen, 995–1009.
- Hildebrandt, Mireille (2020). *Law for Computer Scientists and Other Folk*, Oxford.
- Jakesch, Maurice et al. (2023). *Co-Writing with Opinionated Language Models Affects Users' Views*, *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*, Hamburg.
- Jiang, Cong/Yang, Xiaolei (2023). *Legal Syllogism Prompting: Teaching Large Language Models for Legal Judgment Prediction*, *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law (ICAIL '23)*, Braga, 417–421.
- Katz, Daniel Martin et al. (2024). *GPT-4 passes the bar exam*, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 382 (2270), Art. 20230254, Online: <https://royalsocietypublishing.org/doi/full/10.1098/rsta.2023.0254> (26.10.2025).
- Lagioia, Francesca/Sartor, Giovanni (2020). *AI Systems Under Criminal Law: A Legal Analysis and a Regulatory Perspective*, *Philosophy & Technology* (33), 433–465.
- Lawlor, Reed C. (1963). *What Computers Can Do: Analysis and Prediction of Judicial Decisions*, *American Bar Association Journal*, 49 (4), 337–344, Online: <https://www.jstor.org/stable/25722338> (27.10.2025).
- Liu, Huanghai et al. (2025). *JUREX-4E: Juridical Expert-Annotated Four-Element Knowledge Base for Legal Reasoning*, *arXiv Preprint*, Online: <https://arxiv.org/abs/2502.17166> (28.10.2025).
- Maxwell, Winston/Dumas, Bruno (2023). *Meaningful XAI Based on User-Centric Design Methodology: Combining Legal and Human-Computer Interaction (HCI) Approaches to Achieve Meaningful Algorithmic Explainability*, *CERRE – Centre on Regulation in Europe*, Brussels, Online: <https://doi.org/10.2139/ssrn.4520754> (27.10.2025).
- McCarty, L. Thorne (1977). *Reflections on Taxman: An Experiment in Artificial Intelligence and Legal Reasoning*, *Harvard Law Review*, 90 (5), 837–893.
- Morris, Jason (2021). *Constraint Answer Set Programming as a Tool to Improve Legislative Drafting: A Rules as Code Experiment*, *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law (ICAIL '21)*, São Paulo, 262–263.
- Navarrete, Francisco et al. (2025). *Ontology-Driven Automated Reasoning About Property Crimes*, *Business & Information Systems Engineering* (67), 687–710.
- Palmirani, Monica et al. (2011). *LegalRuleML: XML-Based Rules and Norms*, in: Olken, Frank et al. (Eds.), *Rule-Based Modeling and Computing on the Semantic Web*, Berlin/Heidelberg, 298–312.
- Posner, Richard A. (2008). *How Judges Think*, Cambridge/London.
- Rudin, Cynthia (2019). *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*, *Nature Machine Intelligence*, 1 (5), 206–215.
- Santosh, T. Y. S. S. et al. (2024). *Towards Explainability and Fairness in Swiss Judgement Prediction: Benchmarking on a Multilingual Dataset*, in: Calzolari, Nicoletta et al. (Eds.), *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Turin, 16500–16510.
- Schwarcz, Daniel et al. (2025). *AI-Powered Lawyering: AI Reasoning Models, Retrieval-Augmented Generation, and the Future of Legal Practice*, *Minnesota Legal Studies Research Paper No. 25–16/University of Michigan Public Law Research Paper No. 24–058*, Minneapolis/Ann Arbor, Online: <https://doi.org/10.2139/ssrn.5162111> (28.10.2025).
- Sergot, Marek J. et al. (1986). *The British Nationality Act as a Logic Program*, *Communications of the ACM*, 29 (5), 370–386.
- Servantez, Sergio et al. (2024). *Chain of Logic: Rule-Based Reasoning with Large Language Models*, *arXiv Preprint*, Online: <https://arxiv.org/abs/2402.10400> (28.10.2025).

- Stamper, Ronald (1980). *LEGOL: Modelling Legal Rules by Computer*, in: Walter, Charles/Parker, Donald B. (Eds.), *Computer Science and Law*, London, 45–71.
- Steging, Cor et al. (2021). *Discovering the Rationale of Decisions: Experiments on Aligning Learning and Reasoning*, *arXiv Preprint*, Online: <https://arxiv.org/abs/2105.06758> (28.10.2025).
- Susskind, Richard E. (1990). *Artificial Intelligence, Expert Systems and Law*, *The Denning Law Journal*, 5 (1), 105–116.
- Susskind, Richard (2017). *Tomorrow's Lawyers: An Introduction to Your Future*, Oxford.
- Susskind, Richard (2019). *Online Courts and the Future of Justice*, Oxford.
- Valvoda, Josef/Cotterell, Ryan (2024). *Towards Explainability in Legal Outcome Prediction Models*, *arXiv Preprint*, Online: <https://arxiv.org/abs/2403.16852> (28.10.2025).
- De Visser, Maartje (2014). *Constitutional Review in Europe: A Comparative Analysis*, London.
- Waller, Madeleine et al. (2024). *Can XAI methods satisfy legal obligations of transparency, reason-giving and legal justification?*, *ELSA – European Lighthouse on Secure and Safe AI*, Brüssel, Online: <https://elsa-ai.eu/can-xai-methods-satisfy-legal-obligations-of-transparency-reason-giving-and-legal-justification/> (28.10.2025).
- Wischmeyer, Thomas (2020). *Artificial Intelligence and Transparency: Opening the Black Box*, in: Wischmeyer, Thomas/Rademacher, Timo (Eds.), *Regulating Artificial Intelligence*, Cham, 75–101.
- Xu, Buqiang et al. (2025). *LegalDuet: Learning Fine-Grained Representations for Legal Judgment Prediction via a Dual-View Contrastive Learning*, in: Yoshikawa, Masatoshi et al. (Eds.), *Advanced Data Mining and Applications (ADMA 2025)*, Singapur, 337–352.
- Ye, Hai et al. (2018). *Interpretable Charge Predictions for Criminal Cases: Learning to Generate Court Views from Fact Descriptions*, in: Walker, Marilyn A. et al. (Eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018, Volume 1: Long Papers)*, New Orleans, 1854–1864.
- Yu, Weijie et al. (2022). *Explainable Legal Case Matching via Inverse Optimal Transport-based Rationale Extraction*, *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22)*, Madrid, 657–668.
- Zhang, Kepu et al. (2025a). *An Explicit Syllogistic Legal Reasoning Framework for Large Language Models*, *arXiv Preprint*, Online: <https://arxiv.org/abs/2504.04042> (28.10.2025).
- Zhang, Kepu et al. (2025b). *Beyond Guilt: Legal Judgment Prediction with Trichotomous Reasoning*, *arXiv Preprint*, Online: <https://arxiv.org/abs/2412.14588> (28.10.2025).
- Zhong, Haoxi et al. (2018). *Overview of CAIL2018: Legal Judgment Prediction Competition*, *arXiv Preprint*, Online: <https://arxiv.org/abs/1810.05851> (28.10.2025).