Künstliche Intelligenz im Strafverfahren am Beispiel von Kriminalprognosen

Von Simon Eisbach, Prof. Dr. Michael Heghmanns, Prof. Dr. Guido Hertel, Münster*

Der Einsatz künstlicher Intelligenz im Strafverfahren wird für eine ganze Reihe von Einsatzfeldern diskutiert. Dies beginnt bei spezifischen Ermittlungsaspekten, etwa der Überwachung von Maßnahmen nach den §§ 100b, 100c StPO zur Gewährleistung des Kernbereichsschutzes,[1] und geht über Entscheidungshilfen für den Richter im Rahmen der Strafzumessung[2] bis hin zu Fragen der Strafvollstreckung und des Strafvollzuges, insbesondere im Rahmen von Kriminalprognosen zur Beurteilung erforderlicher Vollzugsmaßnahmen und der (Rest-)Strafaussetzung zur Bewährung. Angesichts der aktuell rapiden Fortschritte bei der Entwicklung hochkomplexer digitaler Systeme werden die notwendigen technischen Voraussetzungen, insbesondere Rechnerkapazitäten und geeignete Software, über kurz oder lang verfügbar sein, weshalb die Entwicklung nicht bei theoretischen Überlegungen stehenbleiben wird. Damit ist es an der Zeit, Chancen, Risiken und notwendige normative Rahmenbedingungen des Einsatzes künstlicher Intelligenz im Strafverfahren zu reflektieren, bevor der Reiz des Machbaren die Justiz- und Kriminalpolitik unvorbereitet trifft und vermeintliche Sachzwänge oder Automatismen auslöst.

I. Was verstehen wir unter „Künstlicher Intelligenz"?

Der Begriff der Künstlichen Intelligenz (KI) wird recht uneinheitlich gebraucht und soll hier in einem eher engen Sinn verstanden werden. Gemeint ist nicht eine Digitalisierung im Allgemeinen, die mit Datenbanksystemen und elektronischer Aktenführung entweder bereits Einzug in die Justiz gehalten hat[3] oder in einem absehbaren Zeitraum geplant ist.[4] Was solche, in der Regel relativ simpel strukturierte elektronischen Systeme zum Verfahrensfortgang beitragen, geschieht auf der Basis eines feststehenden Computerprogramms und kann daher im Prinzip durch den Richter stets nachvollzogen und kontrolliert werden. Das Besondere an KI ist die Möglichkeit maschinellen Lernens durch den Einsatz von Algorithmen, die selbstständig Regelmäßigkeiten aus Daten erschließen, und so das System auf der Basis von Erfahrungen und auch Fehlern kontinuierlich anpassen und optimieren. Statt einfacher Wenn-Dann-Regeln nutzen KI-Systeme während ihrer Entwicklung und auch danach bei der Verarbeitung von Daten mathematische Algorithmen, mit denen sie komplexe Muster erkennen, die dem Auge des menschlichen Betrachters entgehen, der zudem meist auf den Einzelfall fokussiert ist. Besonders mächtige Algorithmen sind künstliche neuronale Netze, deren Funktionsweise ähnlich wie Lernprozesse und die Verarbeitung von Feedback im menschlichen Gehirn ablaufen. Auf die Information (bspw. nach eigenständigem Zugriff auf externe Datenbanken), eine vorangegangene Einschätzung sei richtig oder falsch, verändern solche KI-Systeme einzelne oder mehrere ihrer zahlreichen Regelverbindungen innerhalb des Systems („Synapsen" bzw. „neurons"), wodurch künftige Anfragen zu veränderten Einschätzungen führen. Allerdings benötigen solche selbstlernenden KI-Systeme eine Vielzahl von Trainingsdaten, um komplexere Muster zu erkennen und idealiter zu immer sachgerechteren Ergebnissen zu gelangen.

Auf Grund der Vielschichtigkeit und Komplexität entwickelter KI-Systeme lässt sich für den menschlichen Anwender mit der Zeit immer weniger nachvollziehen, an welchen Stellen einzelne Parameter verändert worden sind und warum nach einer ausgedehnten Trainingsphase mit neuen, dem System bis dahin unbekannten Daten ein ausgeworfenes Ergebnis so und nicht anders zustande gekommen ist. Der Entscheidungsprozess erscheint auf diese Weise schnell als „Black Box", die bei komplexen KI-Systemen auch die ursprünglichen Programmierer nicht mehr komplett erklären können. Um dieses Problem der Intransparenz von KI zu adressieren, entwickeln Forschungsarbeiten zu „Explainable AI" in den letzten Jahren erste vielversprechende Methoden[5] (s.u.). Der Transfer auf den hier fokussierten Anwendungsbereich steht aber noch aus. Wegen dieser Besonderheiten sind mit dem Einsatz im Rahmen des Strafverfahrens besondere Herausforderungen verbunden, soweit es darum geht, Einschätzungen, die von einer solchen KI stammen, dem weiteren Verfahren zu Grunde zu legen oder bei Entscheidungen zu verwenden.

II. Einsatz bei der Erstellung von Kriminalprognosen

Angesichts der sehr vielfältigen Einsatzmöglichkeiten von KI-Systemen für Aufgaben innerhalb des Strafverfahrens kann an dieser Stelle keine umfassende Einschätzung von Zulässigkeit und Machbarkeit geleistet werden. Deshalb wid-

* Der Verf. *Heghmanns* ist Inhaber des Lehrstuhls für Strafrecht, Strafprozessrecht, Medienstrafrecht und Strafvollzugsrecht an der Westfälischen Wilhelms-Universität Münster, der Verf. *Hertel* Inhaber des Lehrstuhls für Organisations- und Wirtschaftspsychologie an der Westfälischen Wilhelms-Universität Münster und der Verf. *Eisbach* ist dort Wissenschaftlicher Mitarbeiter. Die Reihung der *Verfasser* in der Überschrift folgt dem Alphabet.

[1] *Gleizer*, in: Beck/Kusche/Valerius (Hrsg.), Digitalisierung, Automatisierung, KI und Recht, 2020, S. 535.

[2] *Kohn*, Künstliche Intelligenz in der Strafzumessung, 2021; ferner *Nink*, Justiz und Algorithmen, 2021, S. 402 ff.; *Giannoulis*, Studien zur Strafzumessung, 2014, S. 368 ff.

[3] So hat die elektronische Akte in der Ziviljustiz bereits das Versuchsstadium verlassen und ist weitgehend etabliert (§§ 130a ff. ZPO).

[4] In der Strafjustiz sind die Vorbereitung zur Nutzung der Ermächtigung des § 32 StPO zur elektronischen Aktenführung noch nicht abgeschlossen, vgl. *Mitterer*, in: Anders/Graalmann-Scheerer/Schady (Hrsg.), Innovative Entwicklungen in den deutschen Staatsanwaltschaften, 2021, S. 353.

[5] Vgl. u.a. *Goodfellow/Bengio/Courville*, Deep Learning, 2018; *Humm/Buxmann/Schmidt*, Künstliche Intelligenz in der Forschung, 2022, S. 13 ff.; *Greiner/Reinhart/Mayer*, Künstliche Intelligenz – eine Einführung, 2021, S. 17 ff.; *Kohn* (Fn. 2), S. 25 ff.; *Steinbach*, Regulierung algorithmenbasierter Entscheidungen, 2021, S. 55 ff.; *F. Puppe*, in: Beck/Kusche/Valerius (Fn. 1), S. 121.

men sich die folgenden Überlegungen exemplarisch denkbaren Einsatzszenarien im Rahmen der Erstellung einer Kriminalprognose. Zum einen handelt es sich dort um eine vergleichsweise noch einfach strukturierte Entscheidung (Risiko/kein Risiko). Zum anderen existieren schon praktische Erfahrungen im Ausland, wenn auch noch nicht mit einer KI im oben genannten Sinne, da die bislang genutzten Systeme zwar trainiert wurden, aber ihre Parameter im Einsatz nicht eigenständig fortlaufend weiter optimieren.[6] Derartige Prognoseentscheidungen fallen im Rahmen der richterlichen Entscheidung über eine Straf- oder Maßregelaussetzung im Urteil (§§ 56, 67b StGB) oder in der Strafvollstreckung an (§§ 67c, 67d Abs. 2 und 3 StGB). Soweit es den Maßregelvollzug anbelangt, dürfte ein KI-Einsatz freilich wenig realistisch sein, da häufig psychiatrische Einschätzungen maßgebend sind, die zu stark einzelfallbezogen und zu wenig empirisch erfassbar sind, um in eine elektronische Entscheidungsroutine eingepasst zu werden. Zahlenmäßig deutlich häufiger und – jedenfalls im Bereich kleinerer und mittlerer Kriminalität – auch stärker von statistisch messbaren und elektronisch verarbeitbaren Kriterien (wie Vorstrafen, Vollzugserfahrungen, Kriminalitätsart, Alter) abhängig ist hingegen die Prognose über (Rest-)Strafaussetzungen. Idealtypisch entscheidet der Richter, soweit er nicht ausnahmsweise sachverständig beraten ist (§ 246a Abs. 2, § 454 Abs. 2 StPO), hier anhand der Erhebung und Abwägung von tatsächlichen Umständen des Einzelfalls vor dem Hintergrund seiner mehr oder weniger großen Berufserfahrung und seiner vermeintlichen Kenntnisse über Kriminalitätsursachen und -verläufe, letztlich aber dann doch intuitiv.

Diese Entscheidungsstruktur – Erhebung und Bewertung von Daten auf der Basis statistisch-mathematischer Zusammenhänge – ließe sich durchaus mittels einer Datenverarbeitung simulieren, die im Ergebnis eine Risikoeinschätzung ausgeben könnte, welche sodann der Richter bei seiner wertenden Entscheidung berücksichtigen könnte, ob dieses Risiko zu einer „Erwartung" i.S.v. § 56 Abs. 1 StGB führt bzw. einer „Verantwortbarkeit" i.S.v. § 57 Abs. 1 Nr. 2 StGB entspricht. Ein darüberhinausgehender Einsatz, d.h. die eigentliche Strafaussetzungsentscheidung, kann nach dem gegenwärtigen Normprogramm wegen der erforderlichen wertenden Elemente der Strafaussetzungsentscheidung jedoch nicht durch eine KI getroffen werden. Zudem ist es evident, dass rechtsprechende Gewalt i.S.v. Art. 92 GG auszuüben wäre, welche „den Richtern anvertraut" ist und nicht etwa einem von Technikern entworfenen Computerprogramm.[7] Somit besteht ein Konsens, dass die letztliche Entscheidung stets eine richterliche sein muss,[8] wobei an dieser Stelle zwei Diskussionsebenen zu unterscheiden sind. Zum einen handelt es sich um die sachorientierte Frage nach der Intensität und Effizienz humaner Kontrolle eines potenziell nicht stets fehlerfrei arbeitenden Systems,[9] zum anderen um den normativen Aspekt des Richtervorbehalts oder allgemeiner: des Vorbehalts eines menschlichen Entscheiders. Schon einfachgesetzlich verbietet im allgemeinen Datenschutzrecht Art. 22 Abs. 1 DSGVO[10] (jedenfalls auch und mit einigen Ausnahmen) eine vollautomatisierte Entscheidung.[11] Für den hier im Mittelpunkt stehenden Sektor der Gefahrenabwehr und Strafverfolgung ist ferner die RL (EU) 2016/680 einschlägig, die eine entsprechende Regelung in Art. 11 Abs. 1 enthält, welche explizit die Möglichkeit eines Eingreifens des (menschlichen) Verantwortlichen verlangt.[12] Umgesetzt wurde diese Bestimmung in § 54 Abs. 1 BDSG[13] mit einigen Modifikationen: So reduziert sich das Verbot vollautomatisierter Entscheidung im Ergebnis auf ein Verbot mit Erlaubnisvorbehalt qua Rechtsnorm, während sich die Mindestgarantie des Rechts auf Eingreifen des Verantwortlichen dort erstaunlicherweise nicht wiederfindet, aber im Wege richtlinienkonformer Auslegung gleichwohl mit hineinzulesen sei.[14] Der damit beschriebene Mindeststandard einer Eingriffsmöglichkeit impliziert indessen eine (richterliche) Überwachung des Entscheidungsprozesses, was im Ergebnis – zumal angesichts

[6] Siehe dazu näher unter III.

[7] Eingehend *Nink* (Fn. 2), S. 261 ff., 287 f.

[8] *Nink* (Fn. 2), S. 354 f.; *Kaspar/Höffler/Harrendorf*, NK 2020, 35 (51); *Kohn* (Fn. 2), S. 184 ff.; *Staffler/Jany*, ZIS 2020, 164 (170).

[9] *Eisele/Böhm*, in: Beck/Kusche/Valerius (Fn. 1), S. 519 (532).

[10] Art. 22 Abs. 1 DSGVO lautet: „Die betroffene Person hat das Recht, nicht einer ausschließlich auf einer automatisierten Verarbeitung […] beruhenden Entscheidung unterworfen zu werden, die ihr gegenüber rechtliche Wirkung entfaltet oder sie in ähnlicher Weise erheblich beeinträchtigt."

[11] *Steinbach* (Fn. 5), S. 120.

[12] Art. 11 Abs. 1 RL (EU) 2016/80 lautet: „Die Mitgliedstaaten sehen vor, dass eine ausschließlich auf einer automatischen Verarbeitung beruhende Entscheidung – einschließlich Profiling –, die eine nachteilige Rechtsfolge für die betroffene Person hat oder sie erheblich beeinträchtigt, verboten ist, es sei denn, sie ist nach dem Unionsrecht oder dem Recht der Mitgliedstaaten, dem der Verantwortliche unterliegt und das geeignete Garantien für die Rechte und Freiheiten der betroffenen Person bietet, zumindest aber das Recht auf persönliches Eingreifen seitens des Verantwortlichen, erlaubt."

[13] Die Vorschrift lautet: „Eine ausschließlich auf einer automatischen Verarbeitung beruhende Entscheidung, die mit einer nachteiligen Rechtsfolge für die betroffene Person verbunden ist oder sie erheblich beeinträchtigt, ist nur zulässig, wenn sie in einer Rechtsvorschrift vorgesehen ist."

[14] *Helfrich*, in: Sydow (Hrsg.), Nomos Kommentar, Bundesdatenschutzgesetz, 2019, § 54 Rn. 5; *Paschke*, in: Gola/Heckmann (Hrsg.), Bundesdatenschutzgesetz, Kommentar, 13. Aufl. 2019, § 54 Rn. 9; abweichend *Kamlah*, in: Plath (Hrsg.), DSGVO/BDSG, Kommentar, 3. Aufl. 2018, § 54 Rn. 7; *Frenzel*, in: Paal/Pauly, Datenschutz-Grundverordnung, Bundesdatenschutzgesetz: DS-GVO BDSG, Kommentar, 3. Aufl. 2021, § 54 Rn. 6, die die entsprechenden Schutzmaßnahmen eher im (allgemeinen) Verfassungs- und Verwaltungsrecht angesiedelt sehen; *Herbst*, in: Auernhammer, Datenschutz-Grundverordnung, Bundesdatenschutzgesetz und Nebengesetze, Kommentar, 7. Aufl. 2020; § 54 Rn. 9, verlangt hingegen eine Implementation des in der RL genannten Eingriffsrechts in der jeweiligen Rechtsnorm.

der weiterhin notwendigen wertenden Entscheidungsaspekte – auf eine lediglich entscheidungsvorbereitende KI-Unterstützung hinausläuft: Der Richter prüft die von der KI gelieferte Kriminalprognose, um sie notfalls zu korrigieren, und legt sie sodann ggf. seiner Entscheidung über „Erwartung" bzw. „Verantwortbarkeit" zu Grunde.

III. Chancen und Risiken

Das soeben beschriebene Entscheidungsmodell wäre also zunächst theoretisch konstruierbar, ohne von vornherein vor unüberwindlichen rechtlichen Hürden zu stehen. Damit ist indessen nur ein erster Schritt getan, denn zwei ineinander verwobene Fragestellungen können durchaus noch dazu führen, selbst eine solche Entscheidungsunterstützung durch eine KI auszuschließen. Die erste dieser Fragen betrifft die Durchführbarkeit und Effizienz einer richterlichen Kontrolle der KI-Einschätzung. Lassen sich hier unbehebbare Schwachstellen ausmachen, führt dies zu der erneuten Frage der normativen Anforderungen an die Qualität einer richterlichen Entscheidung: Handelt es sich (noch) um eine solche, wenn bestimmte Grundvoraussetzungen der Entscheidung aufgrund der Komplexität der KI nicht hinterfragt werden können? Zudem gelten Bedenken, die auch bereits für den Einsatz nichtlernender Informationssysteme gelten, nämlich dass Arbeitsbelastung, Technikgläubigkeit oder andere Gründe dazu führen, allzu kritiklos technische Einschätzungen zu übernehmen. Damit verwandt, aber ebenso zu bedenken wäre der umgekehrte Effekt, nämlich ein undifferenziertes Misstrauen gegen KI-Einschätzungen, das dazu führt, technische Unterstützung aus Prinzip zu ignorieren und durch (womöglich schlechtere) intuitive Einschätzungen zu ersetzen. Damit verkehrten sich die Vorteile einer KI-Unterstützung ins Gegenteil, was am Ende schlechtere Entscheidungen zur Folge hätte.

Vor diesem Hintergrund lohnt sich ein Blick auf die bisherigen, im Ausland gewonnenen Erfahrungen mit softwaregestützten Kriminalprognoseinstrumenten. In den USA ist mittlerweile wohl in allen Bundesstaaten (aber nicht auf Bundesebene) der Einsatz algorithmenbasiert erstellter Risikoprofile von Straftätern verbreitet.[15] Das derzeit bekannteste dieser Systeme,[16] COMPAS[17], stützt sich auf 136 Angaben, die zum Teil per Interview erhoben werden.[18] Die Erfahrungen mit derartigen Systemen sind jedoch zwiespältig. Ein wesentlicher Kritikpunkt ist die mögliche Diskriminierung bestimmter Ethnien, da allein auf Grund der Hautfarbe die Gefahr bestand, in andere Risikogruppen eingestuft zu werden.[19] Zudem dürfen angebliche Erfolge in der Risikovorhersage nicht überschätzt werden, da sich die verwendeten Systeme in ihrer Wirksamkeit kaum unterscheiden und durchweg nur mit mäßiger Genauigkeit arbeiten.[20]

Mit deutlich geringerer Klassifizierungsleistung wird in der Schweiz die Software FaST[21] eingesetzt, die allerdings keine Prognose abgibt, sondern die Probanden in lediglich drei Risikogruppen einordnet und nur auf etwaigen näheren Abklärungsbedarf hinweist.[22] Verwendet werden hier aus vier Merkmalsbereichen insgesamt 17 Kriterien betreffend Straftat und Person, die teilweise vom Anwender noch zu gewichten sind, jedoch weder sonderlich tief gehen noch Raum für Einzelfallbesonderheiten lassen.[23]

Auf Basis dieser Erfahrungen ist die Einsatzfähigkeit der zurzeit verfügbaren Programme zur effektiven Unterstützung der richterlichen Entscheidung noch mit Skepsis zu betrachten. Wie aber ist die potenzielle Tauglichkeit von KI-Systemen in diesem Kontext zu beurteilen?

---

[15] *Steinbach* (Fn. 5), S. 85.; *Eisele/Böhm* (Fn. 9), S. 527 f.

[16] Eine Übersicht bietet Electronic Privacy Information Center (epic.org), Liberty at Risk: Pre-trial Risk Assessment Tools in the U.S., 2020, S. 2 ff.

[17] COMPAS = Correctional Offender Management Profiling for Alternative Sanctions, entwickelt von Northpointe Inc. und inzwischen als COMPAS-R Core Teil der Northpointe Suite Pretrial (https://www.equivant.com/northpointe-suite-pretrial-2/ [19.6.2022]) des Unternehmens Equivant (Volaris Group).

[18] Der Fragebogen ist abgedruckt in Electronic Privacy Information Center (Fn. 16), S. 26 ff.

[19] *Nink* (Fn. 2), S. 381 ff.; *Eisele/Böhm* (Fn. 9), S. 527 f., vgl. auch *Becerril u.a.*, Validation and Assessment of Pennsylvania's Risk Assessment Instrument, 2019, abrufbar unter https://pennstateoffice365.sharepoint.com/:b:/s/PCSFileshare/EZR_-I6PipJHp0trR3hX9bwB0g3SJClRuwPjSnsak02iqw?e=6mMYQB (19.6.2022).

[20] *James*, Risk and Needs Assessment in the Federal Prison System, 10.7.2018, S. 4, abrufbar unter https://sgp.fas.org/crs/misc/R44087.pdf (19.6.2022).

[21] FaST = Fall-Screening-Tool; dazu *Schwarzenegger/Manzoni/Baur*, Modellversuch Risikoorientierter Sanktionenvollzug (ROS) – Ergebnisevaluation Schlussbericht, 2013, abrufbar unter https://www.researchgate.net/profile/Christian-Schwarzenegger/publication/299564758_Modellversuch_Risikoorientierter_Sanktionenvollzug_ROS_Ergebnisevaluation_Schlussbericht/links/5899f3184585158bf6f8a583/Modellversuch-Risikoorientierter-Sanktionenvollzug-ROS-Ergebnisevaluation-Schlussbericht.pdf (19.6.2022); Bundesamt für Justiz (Hrsg.), Schlussbericht Modellversuch Risikoorientierter Sanktionenvollzug, 2014, abrufbar unter https://www.bj.admin.ch/dam/bj/de/data/sicherheit/smv/modellversuche/evaluationsberichte/ros-schlussber-d.pdf (19.6.2022); *Kilias/Brüngger*, Modellversuch: Risikoorientierter Sanktionsvollzug – Bemerkungen und Analysen zum Projekt des Amtes für Justizvollzug des Kantons Zürich, 2016, abrufbar unter https://www.krc.ch/krcwp/wp-content/uploads/2016/03/Schlussbericht-ROS-KRC.pdf (19.6.2022).

[22] *Eisele/Böhm* (Fn. 9), S. 525 f.

[23] Wegen der Einzelheiten siehe das Manual des Fall-Screening-Tools, Version 6 aus Januar 2018, abrufbar unter https://www.srf.ch/static/srf-data/data/2018/ros/fast_manual_und_gewichte.pdf (19.6.2022).

Auch wenn die Erwartungen an automatisierte Entscheidungen oft sehr hoch sind,[24] so bilden sie – wie auch menschliche Urteile – immer nur einen Teil der Umwelt ab[25]. Handelt es sich dabei um eine hochkomplexe Umwelt, so werden oft relevante Entwicklungsfaktoren von Personen „übersehen". Das kann zum einen an der Vielfältigkeit und Dynamik der Lebensumstände Betroffener liegen, die es nicht zulässt, im Vorhinein alle Faktoren auszumachen, die einmal Bedeutung für eine Prognoseentscheidung haben. Es mag aber – wie bei FaST – auch auf einer bewussten Reduktion auf wenige Faktoren beruhen, über deren Wirkung man meint, hinlängliches Wissen zu besitzen; streng evidenzbasierte Vorgaben fehlen hier nicht selten. Andere Faktoren mag man bewusst eliminiert haben, um Diskriminierungen zu vermeiden.[26] Welche Faktoren bei dem Training eines Algorithmus in den Vordergrund rücken, ist bis zu einem gewissen Grad auch von dem Weltbild (und Kriminalitätsbild) der Entwickler und Auftraggeber abhängig.[27] Zudem können situativ tatauslösende oder -hemmende Einflüsse nur bei einer expliziten Erfassung und Speisung in ein KI-Modell berücksichtigt werden, was wiederum nur generell kriminalitätsbegünstigende oder -hemmende Umstände identifiziert; ob diese auch im Einzelfall wirken, bleibt schon dank der menschlichen Entscheidungsfreiheit offen. Die Prognosekompetenz einer KI stößt damit zwangsläufig an Grenzen; Perfektion ist auch durch sie nicht zu erreichen.

Gleichwohl ist Perfektion auch nicht zwangsläufig das Ziel, sondern Fortschritt bemisst sich daran, ob eine Prognose besser funktioniert als der Status quo. Hier dürfte der Einsatz von KI in mehreren Aspekten den bisherigen Prognosemethoden überlegen und damit hilfreich sein. Die menschliche Entscheidungsfindung unterliegt bekanntlich einer Reihe von (unbeabsichtigten) Verzerrungen („biases") in der Wahrnehmung, Informationsverarbeitung und Entscheidungsfindung. Im juristischen Kontext relevant sind beispielsweise die Tendenz, Informationen entsprechend anfänglicher Vermutungen zu gewichten und widersprechende Informationen zu ignorieren („Confirmation bias"), oder aber die Schwierigkeit eines Richters, unzulässiges Beweismaterial vollends bei einer Entscheidungsbildung zu ignorieren.[28] Selbst die Höhe von (teilweise irrelevanten) Zahlen während des Prozesses kann die Höhe des letztendlichen Strafmaßes maßgeblich beeinflussen[29] („Anker-Effekt"). Dies sind nur einige Beispiele für subjektive Einflüsse auf menschliche Entscheidungsfindung. Diese Handicaps haben selbstlernende KI-Systeme nicht: Sie treffen jede ihrer Entscheidungen auf der Basis datenbasierter Muster, unbeeinflusst von anfänglichen Vermutungen, irrelevanten Informationen, Reihenfolge der Daten oder auch Tageszeit, Hunger oder sonstigen Befindlichkeiten. Damit kann ein hohes Maß an Fairness realisiert werden, bei dem die stets gleichen evidenzbasierten Regeln zum Einsatz kommen.

Ein weiterer Vorteil der KI ist die Nutzung größtmöglicher Daten als Entscheidungsbasis,[30] welche die singuläre menschliche bzw. richterliche Erfahrungsbasis weit übertrifft. Zusätzlich zu der höheren Rechenkapazität nutzen KI-Systeme auch (selbstständiges) Data Mining, also die Suche nach statistisch signifikanten Zusammenhängen zwischen Faktoren (hier zwischen der Kriminalitätsentstehung einerseits und den persönlichen Umständen wie den sozialen Umweltbedingungen andererseits). KI ist deshalb in der Lage, Kriminalitätsrisiken aus dem Zusammenhang mehrerer, unabhängig nebeneinander wirkenden Ursachen zu erkennen, die sich dem menschlichen Betrachter des Einzelfalls nicht unbedingt erschließen.

Grundvoraussetzung eines erfolgversprechenden Einsatzes der KI ist eine hinlänglich große und valide Datenbasis, die vorwiegend jedoch nur für die häufiger oder gar massenhaft begangenen Delikte darstellbar sein wird.[31] Um tragfähige Prognosen für Delikte mit hohem Dunkelfeldanteil erstellen zu können, bedarf es zudem entsprechender Dunkelfeldforschung, um auch insoweit aussagefähige Parameter für kriminalitätsbegünstigende Lebens- und Persönlichkeitsumstände zu gewinnen. Während daher eine automatisierte Prognose für Kriminalität im Allgemeinen wohl noch Zukunftsmusik sein dürfte, kann sie für gut erforschte oder erforschbare Kriminalitätssektoren wie Ladendiebstahl, BtM-Kriminalität oder Gewaltdelinquenz in absehbarer Zeit durchaus einsatzfähig werden.

Wichtig wird dabei sein, KI-Systeme nur in den Fällen einzusetzen, die von ihrem Kriminalitätsprofil her auch passen. Andernfalls würden sie als Ergebnisse einer als treffsicher bewerteten Methodik Vertrauen erwecken, das sie im konkreten Fall nicht verdienen, weil der Algorithmus auf zu wenig validen Daten basiert. Vergleichbares gilt auch für die Entscheidungsobjektivität bzw. -neutralität[32].

IV. Strukturen einer richterlichen Kontrolle

1. Vergleichbare Entscheidungssituationen

Blickt man auf die notwendige richterliche Kontrolle einer KI-gestützt ermittelten Kriminalitätsprognose, so drängt sich strukturell zunächst ein Vergleich mit der Überprüfung eines Sachverständigengutachtens auf. In beiden Fällen bedient

[24] *Dietvorst/Simmons/Massey*, Journal of Experimental Psychology: General 2015, Bd. 144, Heft 1, 114.
[25] *Steinbach* (Fn. 5), S. 33.
[26] Siehe etwa die Bemühungen in Pennsylvania, ethnisch begründete Risikomodifikationen auszublenden, z.B. in den Richtlinien der Pennsylvania Commission on Sentencing (https://pennstateoffice365.sharepoint.com/:b:/s/PCSFileshare/EaDAtZdmvAJEtLQZTNyTBWcB53phArs6m6I1t8KNGmblqQ?e=IvyVsB [19.6.2022]) und die übrigen offiziellen Dokumente der Kommission (https://pcs.la.psu.edu/guidelines-statutes/risk-assessment/ [19.6.2022]).
[27] *Steinbach* (Fn. 5), S. 34.
[28] *Peer/Gamliel*, Court Review: The Journal of the American Judges Associaton 49 (2013), 114; siehe ferner die Nachweise in Fn. 52.
[29] *Englich/Mussweiler*, Journal of Applied Social Psychology 31 (2001), 1535.
[30] *Steinbach* (Fn. 5), S. 30.
[31] *Eisele/Böhm* (Fn. 9), S. 531.
[32] *Steinbach* (Fn. 5), S. 31 ff.

sich der Richter einer fremden Kompetenz, über die er selbst nicht verfügt, um Grundlagen für seine Entscheidung zu ermitteln. Datenbasis und Methodik sind ihm jeweils vielleicht in Grundzügen bekannt, aber er könnte ebensowenig die Computerberechnungen nachvollziehen, wie er z.B. eine DNA-Auswertung selbst vornehmen oder eine schizoide Persönlichkeitsstörung diagnostizieren könnte. Gleichwohl wird von ihm verlangt, ein Sachverständigengutachten nachzuvollziehen und sich nicht blind auf das mitgeteilte Resultat zu verlassen.[33] Ebenfalls strukturell vergleichbar ist die im Bußgeldverfahren häufig thematisierte Frage einer ordnungsgemäßen Geschwindigkeitsmessung durch automatisiert arbeitende Messgeräte. Hier bestehen zwar bei standardisierten Messverfahren nur dann eingehendere Überprüfungspflichten, wenn Zweifel an der Richtigkeit des mitgeteilten Ergebnisses bestehen.[34] Allerdings wird man eine sich fortentwickelnde KI jedenfalls vorläufig kaum als standardisierte Methodik begreifen können. Im Vergleichsfall nicht standardisierter Geschwindigkeitsmessungen wäre es notwendig, deren Richtigkeit und die darauf aufbauenden Berechnungen im Einzelnen durch Einholung eines Sachverständigengutachtens nachzuvollziehen.[35] Überträgt man dies wiederum auf die richterliche Überprüfung einer KI-Prognose, so ergäben sich vermutlich auch dabei recht weitgehende Überprüfungserfordernisse. Es erscheint freilich zweifelhaft, ob ein vollständiges Nachvollziehen der Prognoseresultate einerseits zu leisten und andererseits notwendig ist, um eine fehlerfreie Entscheidung zu treffen. Denn im Unterschied zum Sachverständigengutachten und zur Geschwindigkeitsmessung stünde der Richter, wenn er die KI-Prognose wegen beachtlicher Zweifel an ihrer Richtigkeit verwerfen sollte, nicht mit leeren Händen da, denn ihm blieben ja immer noch die traditionellen Prognosemethoden bis hin zur intuitiven Risikoeinschätzung auf der Basis von Berufs- und Lebenserfahrung. Eine interessante Option für die Zukunft ist hier sicher auch die Möglichkeit, KI-basierte Prognosen als Teil der Ausbildung von Richtern mitzudenken, etwa in Form von Fort- und Weiterbildungen.

2. Kontrolle des Programms?

Geht es um die Überprüfung eines KI-Ergebnisses, so richtet sich der Blick in erster Linie auf die Software. Die Realisierbarkeit einer hierauf gestützten Überprüfung stößt jedoch auf erhebliche Schwierigkeiten. Die Entwicklung von KI-Systemen zur Kriminalitätsprognose befindet sich überwiegend in der Hand von gewinnorientiert arbeitenden Privatunternehmen.[36] Zwar sind in den USA auch einzelne Bundesstaaten tätig geworden, jedoch zumeist mit einfacheren statistischen Programmen, was angesichts des notwendigen Programmierungs- und Trainingsaufwandes wenig verwunderlich ist. Für Deutschland darf angesichts seiner föderalen Struktur ebenfalls nicht damit gerechnet werden, dass einzelne oder mehrere Bundesländer leistungsfähige Programme zur Kriminalitätsprognose in Eigenregie erstellen können. Die Erfahrungen mit der Einführung diverser Datenverarbeitungsprogramme innerhalb der Länderjustizen lassen vielmehr erwarten, dass man auch bei der Erstellung von KI-Systemen im Zweifel auf kommerzielle Anbieter zurückgreifen wird. Gewinnorientierte Softwareunternehmen sehen indessen nicht ohne Berechtigung ihre Marktposition in Gefahr, wenn sie ihre Software im Rahmen einer richterlichen Kontrolle bekannt geben müssten. Geschähe dies im Rahmen eines Strafverfahrens, so wäre infolge bestehender Akteneinsichtsrechte nicht verlässlich auszuschließen, dass neuralgische Informationen bis hin zur Konkurrenz dringen. Nach allen bisherigen Erfahrungen weigern sich die Entwickler daher aus gutem Grund, solche Daten zu offenbaren,[37] was zunächst die Forderung provoziert, nur solche Systeme zu akzeptieren, die mit Open Source-Algorithmen operieren.[38]

Aber selbst wenn ein Rückgriff auf Open Source-Software realistisch sein sollte, erscheint es fraglich, ob deren Verwendung eine signifikante Steigerung der Nachvollziehbarkeit bewirken könnte. Bei hinlänglich tiefer Struktur der selbstlernenden KI-Systeme dürfte nämlich die Ausgangsstruktur des Systems eine vergleichsweise geringe Rolle für das später ausgegebene Einzelfallergebnis spielen, das sich als Resultat eines fortlaufend optimierten maschinellen Lernprozesses darstellt. Denn damit entscheidet weniger der Ausgangszustand als Volumen und Struktur der Trainingsdaten sowie die von ihnen bewirkte Veränderung innerhalb des Systems, wie dieses auf eine konkrete Abfrage reagiert. Ob man diese, im Laufe der Zeit erfolgten und stetig weiter geschehenden Veränderungen durch eine Offenlegung der Software, der Neuronen und ihrer jeweiligen Gewichte mit vertretbarem Aufwand nachvollziehen kann,[39] erscheint mindestens fragwürdig. Dies ist jedenfalls nicht routinemäßig zu leisten, womit im Regelfall ein Verzicht auf eine Prüfung durch den Richter einhergehen müsste, was jedoch mit dem Erfordernis einer richterlichen Endentscheidung unvereinbar erscheint.

Wenn eine KI-gestützte Einzelfallprognose weniger von der ursprünglichen Programmierung des Systems als von dem datenbasierten Training abhängt, dann rückt die Qualität dieser Daten in den Fokus. Die Mahnung des „garbage in,

[33] BGHSt 8, 113 (118); 12, 311 (314 f.); 34, 29 (31); BGH NJW 1993, 3081 (3082); BGH NStZ 2020, 294 (297).

[34] Dazu statt vieler BGH NJW 1993, 3081 (3083); OLG Naumburg NJ 2021, 513 f.; OLG Bremen, Beschl. v. 15.4.2020 – 1 SsRs 16/20, Rn. 7 f.; OLG Stuttgart NStZ-RR 2022, 60 f.

[35] OLG Oldenburg, Beschl. v. 19.7.2021 – 2 Ss (OWi) 170/21, Rn. 19; OLG Koblenz, Beschl. v. 15.12.2021 – 3 OWi 32 SsRs 108/21, Rn. 8.

[36] Vgl. für die USA die Übersicht von Electronic Privacy Information Center (Fn. 16), S. 5 ff.

[37] Electronic Privacy Information Center (epic.org), AI in the Criminal Justice System, abrufbar unter https://epic.org/issues/ai/ai-in-the-criminal-justice-system/ (19.6.2022); *Jiang*, in: Beck/Kusche/Valerius (Fn. 1), S. 557 (560 f., 568); *Nink* (Fn. 2), S. 385; *Steinbach* (Fn. 5), S. 87 f.

[38] *Jiang* (Fn. 37), S. 569 f.

[39] So die Forderung von *Jiang* (Fn. 37), S. 585.

garbage out“ legt nahe, auf die Wahrung bestimmter Standards zu dringen, was theoretisch auch kontrollierbar erscheint. So müssen die Trainingsdaten quantitativ der kriminologisch verifizierten Realität entsprechen und dürften keine diskriminierenden Tendenzen aufweisen. Praktisch allerdings stellen sich hier wiederum erhebliche Hürden, da es keineswegs selbstverständlich ist, diese Daten überhaupt zu bekommen. Die Erfahrungen in den USA gehen vielmehr dahin, dass sowohl aus Wettbewerbs- als auch aus Kostengründen keine Kooperation und Transparenz zu erwarten ist.[40] Aber selbst bei verfügbaren Trainingsdaten ist es nicht trivial, potenziell kritische Datenmuster zu erkennen. Beispielsweise lassen sich diskriminierende Muster (die etwa die häufigere Festnahme von Personen mit Migrationshintergrund im Vergleich zu Personen ohne Migrationshintergrund zeigen) nur als solche identifizieren, wenn man die Überrepräsentation erkennt, was wiederum ein vergleichsweise präzises Wissen darum voraussetzt, wie es mit der Kriminalitätsneigung von diesbezüglichen Personengruppen in Wahrheit bestellt ist. Ob die jeweiligen Fakten überhaupt hinlänglich erforscht und bekannt sind, erscheint jedenfalls nicht für alle relevanten Parameter gewährleistet. Der simple Ausweg, diskriminierungsanfällige Fakten wie die Herkunft aus der Analyse vorsorglich komplett auszublenden,[41] birgt die Gefahr, damit auf möglicherweise relevante Hinweise für die Vorhersage zu verzichten. Erforderlich wäre vielmehr, Trainingsdaten bestimmter Kategorien nur in dem Maße zu verwenden, wie sie der – hinlänglich erforschten – Realität entsprechen. Wo die Realität dagegen unerforscht ist, die KI aber Korrelationen feststellt, müssten diese erkennbar sein, damit ihre Auswirkungen auf die Prognose der richterlichen Kontrolle zugänglich bleiben.

Nicht zu vergessen ist die Relevanz der Qualität der Daten des konkreten Falles. Soweit die Falldaten gezielt eingegeben und nicht automatisiert erhoben werden,[42] bleibt der Standard und die Möglichkeit zur Eingabefehlerkontrolle immerhin noch derjenige einer nichtunterstützten Entscheidung, bei welcher der Richter unter den Fakten wählt, was er für relevant und was er für irrelevant hält. Zusätzliche Fehlerquellen oder Kontrolllücken ließen sich an dieser Stelle daher gut vermeiden.

3. Ansätze zur Erklärung eines KI-Ergebnisses

Der Blick in die Daten- und Softwarebasis dürfte also aus mehreren Gründen für eine richterliche Kontrolle kaum nutzbar sein. Es steht vielmehr zu befürchten, dass je entwickelter eine KI ist und auf je mehr „Erfahrungen“ sie zurückgreifen kann, desto unzugänglicher sich ihre Entscheidungsfindung darstellen wird. Der Versuch, sie gewissermaßen zu sezieren und die Entscheidungsprozesse sichtbar und verständlich zu machen,[43] dürfte deshalb kein sehr gangbarer Weg sein. Das muss jedoch nicht das Ende aller Bemühungen sein, Ergebnisse einer KI wenigstens im Kern zu verstehen, zumal die Informatik gerade auf diesem Gebiet zur Zeit sehr intensiv forscht und bereits etliche Ansätze entwickelt hat.[44]

Um die *Transparenz* einer KI zu steigern, wäre es denkbar, sie so zu programmieren, dass sie die wesentlichen Parameter eines Falles, die zu einem bestimmten Ergebnis geführt haben, offenlegt. Eine vor kurzem entwickelte Methode namens „SHAP“ erlaubt einen solchen Einblick selbst in bisher recht intransparente neuronale Netze. Dabei kann für jede Entscheidung angezeigt werden, welche Faktoren am ausschlaggebendsten waren. Auch ob ein Faktor positiv oder negativ für das Ergebnis gewertet wurde, lässt sich dadurch nachvollziehen. Dies wäre ein großer Schritt hin zur Nachvollziehbarkeit komplexer KI-Systeme und es ließen sich damit diskriminierend wirkende Umstände potentiell aufdecken.

Um das *Vertrauen* in eine KI und seine *Akzeptanz* zu steigern, könnte dem richterlichen Nutzer ermöglicht werden, mit den konkreten Falldaten zu „spielen“, also einzelne oder mehrere Parameter zu verändern oder wegzulassen, um so zu lernen, inwiefern sich die KI-Einschätzung ändert. Auf diese Weise ließe sich beispielsweise feststellen, ob ein vom Richter intuitiv als wesentlich angesehener Aspekt eines Falles überhaupt eine Rolle für die automatisierte Prognose gespielt hat. Wäre ein solcher Umstand von der KI offenbar als irrelevant gewertet worden, so wäre der Richter jedenfalls veranlasst, sich zu fragen, ob seine intuitiv anderslautende Einschätzung möglicherweise verfehlt war oder aber er zu der Auffassung gelangen sollte, dass die KI an dieser Stelle offenbar an ihre Grenzen gestoßen und ihr diesmal nicht zu folgen ist. Ebenso ließen sich diskriminierende Tendenzen möglicherweise aufdecken, etwa durch das Austauschen eines ausländisch klingenden Namens gegen einen „unverdächtigen“.[45] Diese Möglichkeiten bieten nicht zuletzt interessantes Potenzial für die Ausbildung zukünftiger Richter.

---

[40] *Jiang* (Fn. 37), S. 563.

[41] Sympathien für eine solch radikale Lösung scheinen bei *Eisele/Böhm* (Fn. 9), S. 530, durch.

[42] Denkbar wäre natürlich auch ein Scan der Akten bzw. künftig der elektronischen Akten. Allerdings erscheint der dadurch erzielbare Vorteil in Gestalt einer Arbeitserleichterung eher gering, weshalb sich diese zusätzliche Fehlerquelle ohne größeren Aufwand vermeiden ließe.

[43] Auf die damit verbundenen datenschutzrechtlichen Probleme im Lichte der Art. 5 Abs. 1 lit. a DSGVO (für die betroffene Person nachvollziehbare Verarbeitung) und der dies konkretisierenden Art. 13–15 DSGVO soll an dieser Stelle nicht eingegangen werden; siehe dazu *P. Vogel*, in: Beck/Kusche/Valerius (Fn. 1), S. 645 (648).

[44] Einen anschaulichen Überblick liefert die Studie von *Kraus/Ganschow/Eisenträger/Wischmann*, Erklärbare KI – Anforderungen, Anwendungsfälle und Lösungen, 2021, S. 24 ff., abrufbar unter https://www.digitale-technologien.de/DT/Redaktion/DE/Downloads/Publikation/KI-Inno/2021/Studie_Erklaerbare_KI.pdf;jsessionid=08916AD64A4F7FB9442B9F1AE03E5EA8?__blob=publicationFile&v=17 (19.6.2022).

[45] So sollen Algorithmen in der Versicherungsbranche bei einer bestimmten nationalen Herkunft zu Angeboten mit höheren Versicherungsprämien neigen, vgl. die Website der Eidgenössischen Kommission gegen Rassismus, abrufbar unter

Ein Softwaredesign, das derartige Gegenproben erlaubt, ist zweifelsfrei konstruierbar und in Ansätzen auch schon vorhanden.[46] Es basiert auf dem Grundsatz von SHAP,[47] bei dem der Einfluss einzelner auszuwählender Faktoren auf ein konkretes Ergebnis des KI-Systems bestimmt wird. Hierzu berechnet die KI zuerst eine Prognose unter Berücksichtigung aller Faktoren. Anschließend werden weitere Prognosen erstellt, die jedoch jeweils einzelne Faktoren ausblenden. Auf diese Weise kann ein fraglicher Unterschied zwischen den Ergebnissen berechnet und damit der konkrete Einfluss eines Faktors identifiziert werden. Nachteil ist ein hoher Rechneraufwand, sofern nicht nur einige wenige Faktoren überprüft werden sollen.[48] Einen anderen Ansatz bieten Counterfactual Explanations. Hierbei handelt es sich um ein Konzept, das für ein konkretes Klassifikationsergebnis (z.B. eine schlechte Kriminalprognose) eine möglichst kleine Änderung in den Eingabewerten zu identifizieren sucht, die zu einer anderen Einstufung führen würde. Auf diese Weise zeigt bereits die Software die neuralgischen Punkte auf.[49] Dies kann hilfreich sein, erlaubt aber nicht, die zu prüfenden Faktoren selbst zu definieren. Die Sensitivitätsanalyse als dritte denkbare Methodik verändert systematisch einzelne Eingabeparameterwerte, um durch diese Variationen zu ermitteln, welche Eingabeparameter den größten Einfluss auf ein Klassifikationsergebnis haben.[50] Ihr Ziel ist allerdings die Identifikation der relevanten Faktoren und nicht die Überprüfung der Wirksamkeit einzelner, isoliert betrachtet möglicherweise weniger bedeutsamer Parameter.

4. Leistungsfähigkeit richterlicher Kontrolle

Angesichts dieser bereits vorhandenen Erklärungsmodelle und des weiterhin zu erwartenden Fortschritts der KI-Technik dürfte es keine unrealistische Zukunftsmusik sein, von einer künftigen Verfügbarkeit relativ simpler Möglichkeiten zur Überprüfung des Einflusses einzelner Fallparameter auch für Nutzer ohne eingehende Informatikkenntnisse auszugehen. Doch mit der Verfügbarkeit von KI-Systemen zur Prognoseerstellung und entsprechenden Kontrollmöglichkeiten ist es nicht getan. Beim Nebeneinander von menschlich-richterlicher Entscheidung und KI-Entscheidungsvorschlag bleibt vielmehr zu bedenken, dass beide Entscheidungssysteme nicht unabhängig voneinander agieren. Es ist einerseits ungeklärt, ob eine der richterlichen Rolle entsprechende, neutrale und unbeeinflusste Kontrolle überhaupt zu leisten ist, insbesondere wenn von einer grundsätzlichen Zuverlässigkeit und Treffsicherheit der KI-Prognose auszugehen und dies dem Richter bewusst ist.[51] Andererseits handelt es sich bei der Richterschaft um eine Personengruppe, die möglicherweise nicht dem Normaltypus des IT-Nutzers entspricht. Richter sind es gewohnt, unabhängig Entscheidungen zu fällen, und auf Grund ihrer beruflichen Sozialisation sind sie eventuell auch misstrauischer als andere Berufsgruppen gegenüber Einschätzungen, wie ein von ihnen zu entscheidender Sachverhalt zu verstehen sei, die ihnen vorgesetzt, aber nicht näher erläutert werden. Inwieweit diese gegenläufigen Tendenzen tatsächlich wirken, bedarf noch einer Klärung. Je nach Auftreten bzw. Überwiegen einer dieser beiden denkbaren Haltungen droht entweder eine unkritische IT-Hörigkeit, welche eine dem Richtervorbehalt unterliegende Entscheidung der Sache nach durch die Technik treffen ließe, oder aber eine der Entscheidungsqualität abträgliche Missachtung hochwertiger Informationen. Beides gilt es zu vermeiden, um die Vorzüge einer KI-Unterstützung zu nutzen, ohne ihrem Einfluss von vornherein zu unterliegen.

Dass Richter nicht unbeeinflusst entscheiden, sondern durch Aktenkenntnis, Parteivortrag und Vorbefassung geprägt sein können, ist u.a. durch die einschlägigen Publikationen von *Schünemann*[52] bekannt, der bspw. das Frageverhalten von Richtern in der Hauptverhandlung in Abhängigkeit von ihrer Aktenkenntnis untersucht und festgestellt hat, dass Vorinformationen zu einer geringen Nachfragebereitschaft geführt haben.[53] Es kann deshalb nicht ausgeschlossen werden, dass auch eine Vorinformation in Gestalt einer KI-Kriminalprognose zu einem veränderten Richterverhalten führt.[54] Das wäre nicht per se nachteilig, solange es darauf hinausliefe, einen KI-Vorschlag in angemessenem, kritisch reflektiertem Maße zu beachten. Weitere Überlegungen hierzu blieben ohne empirische Überprüfung allerdings spekulativ.

Der richterliche KI-Anwender benötigt zudem Vertrauen in die Fähigkeit der KI, im Regelfall sachgerechte Entscheidungen vorzuschlagen; andernfalls wird er sie nicht effektiv einsetzen. Die wahrgenommene Vertrauenswürdigkeit einer KI umfasst zum einen Aspekte der *Leistungsfähigkeit* in Form von zutreffenden Empfehlungen, die es dem Anwender – durchaus rational erfassbar – erlauben, sich auf das System zu verlassen.[55] Als weitere Komponente spielen prozessorien-

https://www.rechtsratgeber-rassismus.admin.ch/lebensbereiche/d254.html# (19.6.2022).

[46] Zahlreiche Modelle, die eine KI zu erklären suchen, sind allerdings auf KI-Systeme zur Bild- (oder Gesichts-)Erkennung ausgerichtet (wie LRP [Layer-Wise Relevance Propagation]) oder für Nutzer nicht anwendungssicher (wie DeepLIFT [Deep Learning Important FeaTures] oder Activation Maximization) und scheiden daher für die hier in Rede stehenden Zwecke aus.

[47] SHAP = SHapley Additive exPlanations.

[48] *Kraus/Ganschow/Eisenträger/Wischmann* (Fn. 44), S. 27 f. m.w.N.

[49] *Kraus/Ganschow/Eisenträger/Wischmann* (Fn. 44), S. 26 f. m.w.N.

[50] *Kraus/Ganschow/Eisenträger/Wischmann* (Fn. 44), S. 31 m.w.N.

[51] Ähnlich für den Einsatz im Rahmen medizinischer Diagnose und Behandlung *Lohmann/Schömig*, in: Beck/Kusche/Valerius (Fn. 1), S. 345 (355).

[52] *Schünemann*, StV 2000, 159 = *ders.*, Strafprozessrecht und Strafprozessreform, 2020, S. 239; *ders*., ebenda, S. 215.

[53] *Schünemann*, StV 2000, 159 (161 f.).

[54] *Staffler/Jany*, ZIS 2020, 164 (175).

[55] *Solberg et al.*, Group & Organization Management 2022, Bd. 47, Heft 2, 187, abrufbar unter https://journals.sagepub.com/doi/abs/10.1177/10596011221081238 (19.6.2022); *Thielsch/Meeßen/Hertel*, PeerJ 6:e5483 (2018), abrufbar unter

tierte Aspekte wie die Nutzbarkeit (Usability) einer KI und die bereits diskutierte *Transparenz* der Arbeitsprozesse eine wichtige Rolle, um die Zuverlässigkeit des Systems einzuschätzen. Eine dritte Komponente ist das Vertrauen in die generelle *Zielsetzung* einer KI und die zugrundeliegenden *Werte* und Ausrichtung der Programmierung. Zusätzlich zu der solcherart eingeschätzten Vertrauenswürdigkeit eines KI-Systems wird das richterliche Vertrauen zudem von der technikbezogenen Vertrauensneigung bestimmt, die sowohl generell (in der Gesellschaft im Ganzen wie in ihrer Untergruppe der Richterschaft) als auch individuell eine wandelbare Größe und abhängig von Erfahrungen mit der KI bzw. mit der Digitalisierung im Allgemeinen (Technikvertrauen) ist. Mit dem Maß des Vertrauens steigt das Einsatzpotenzial der KI, weil Anwender sich ihrer gerne bedienen und damit ihre Vorzüge nutzen werden.[56] Vertrauen ist dabei nichts uneingeschränkt Positives, weil es als blindes Vertrauen dazu führen kann, sich der KI zu bedienen, wo dies nicht mehr angezeigt ist, was wiederum zu potenziell nicht mehr sachgerechten Entscheidungen führen kann. Dem entgegen steht das Phänomen der „algorithmischen Aversion", demnach Personen einem System einen Fehler sehr viel weniger nachsehen, als wenn dieser von einem Menschen gemacht wurde.[57]

Es gilt also zu explorieren, wie überhaupt Vertrauen in die Fähigkeiten einer KI in der speziellen richterlichen Zielgruppe entsteht, und welche weiteren, teils sachfremden Faktoren die richterliche Akzeptanz der KI-Entscheidung beeinflussen. Denkbar wären etwa hohe Arbeitslasten, die keine Zeit für fundierte eigene Entscheidungen lassen, oder persönliches Misstrauen in automatisierte Entscheidungsprozesse aus Technikangst oder der Befürchtung heraus, überflüssig zu werden. Sodann wird zu ermitteln sein, ob und wie sich Vertrauen gewissermaßen begrenzen lässt, ohne es zugleich zu zerstören, und wie es sich also auf ein „gesundes" Maß reduzieren lässt. Beide Fragestellungen sind solche, welche auch mit Hilfe empirischer Forschung zu beantworten wären. Es wird darauf ankommen, welches Vertrauen der richterliche Anwender der KI gegenüber aufbringen wird, und zwar unter variablen Ausgangsbedingungen, was die Kenntnisse über die konkret eingesetzte KI anbelangt. Denn da es nicht den Idealzustand einer vollständigen Kenntnis über alle bekannten Trainings- und Erfahrungsdaten sowie den Einfluss möglicherweise unbekannter Fakten des Einzelfalls geben wird, ist klärungsbedürftig, was die Mindestbedingungen eines Vertrauens in die KI sind. Außerdem wäre es wichtig zu erfahren, wann ein vorhandenes Vertrauen schwindet bzw. wie lange ein Richter der KI noch vertraut, obschon ihm augenscheinlich gegenläufige – relevante, irrelevante oder in ihrer Relevanz kaum zu beurteilende – Fakten des Einzelfalls auffallen. Diese Fragen sind optimalerweise Gegenstand interdisziplinärer, vor allem auch psychologischer Forschung.

Zum Anwendervertrauen in KI-Systeme generell sind erste Untersuchungen vorhanden,[58] sie betreffen aber nicht die spezifische Gruppe richterlicher Anwender. Diese weist unter anderem durch ihre Unabhängigkeit eine Besonderheit auf, die zahlreiche Entscheider in anderen Organisationen (Wirtschaft, Verwaltung) nicht besitzen. Andererseits folgen Richter einem strikten Entscheidungsprogramm, welches die gesetzlichen Normen vorgeben, das für sie indisponibel ist. Eine dritte Besonderheit liegt darin, keine übergeordneten Interessen (wie einen wirtschaftlichen Erfolg oder politische Ziele) zu verfolgen, sondern lediglich die Richtigkeit ihrer Entscheidung nach Maßgabe des Entscheidungsprogramms als Endziel anzustreben. Diese Spezifika lassen es prima facie nicht zu, Erkenntnisse zu anderen Berufsgruppen ungeprüft zu übernehmen. Vielmehr erscheint es notwendig, empirische Daten zur Vertrauensgenese und zur Vertrauensgrenze in der Richterschaft zu erheben.

V. Zusammenfassung

Der Einsatz einer KI kann im Strafverfahren langfristig zu einer schnelleren und langfristig ökonomischeren Entscheidung von höherer Qualität jedenfalls dort führen, wo – wie bei der Erstellung von Kriminalprognosen – eine Vielzahl von Faktoren des Einzelfalls empirisch nachvollziehbar das Ergebnis bestimmen. Vorläufig sind die verfügbaren KI-Anwendungen allerdings noch nicht hinlänglich entwickelt, um den dabei anstehenden inhaltlichen und normativen Anforderungen zu genügen. Grundsätzlich kann eine richterliche Prognoseentscheidung auf eine KI-Prognose gestützt werden, sofern der Richter in der Lage ist, die Entscheidung nachzuvollziehen, damit seine Entscheidung eine auch materiell richterliche Entscheidung bleibt. Dazu muss ihm die Software die Möglichkeit bieten, die Relevanz bestimmter Fallparameter für eine konkrete Prognoseentscheidung zumindest prinzipiell nachzuvollziehen. Zugleich müssen die KI-Anwendung und die sonstigen Rahmenbedingungen seiner Tätigkeit so konzipiert werden, dass der Richter ein gesundes Maß an Vertrauen in eine (im Regelfall zutreffende Einschätzungen auswerfende) KI-Anwendung entwickeln kann, ohne die Fähigkeit und die Bereitschaft zu kritischer Nachfrage zu verlieren. Wie eine solche Arbeitsumgebung zu gestalten wäre, bedarf noch eingehender interdisziplinärer empirischer Forschung. Bis diese vorliegt, ist der Justizverwaltung und dem Gesetzgeber Zurückhaltung anzuraten.

https://peerj.com/articles/5483.pdf (19.6.2022).

[56] *Meeßen/Thielsch/Hertel*, Zeitschrift für Arbeits- und Organisationspsychologie 2020, Bd. 64, Heft 1, 6, abrufbar unter https://doi.org/10.1026/0932-4089/a000306 (19.6.2022).

[57] *Dietvorst/Simmons/Massey*, Journal of Experimental Psychology: General 2015, Bd. 144, Heft 1, 114.

[58] *Höddinghaus/Sondern/Hertel*, Computers in Human Behavior 116 (2021), 106635 m.w.N.; *Meeßen/Thielsch/Hertel*, Zeitschrift für Arbeits- und Organisationspsychologie 2020, Bd. 64, Heft 1, 6, abrufbar unter https://doi.org/10.1026/0932-4089/a000306 (19.6.2022).