Humanistische Union

Fairness, Erklärbarkeit und Transparenz bei KI-Anwendungen im Sicherheitsbereich – ein unmögliches Unterfangen?

In der ethischen, rechtswissenschaftlichen und politischen Diskussion über Anforderungen an KI-Anwendungen spielen Fairness, Erklärbarkeit und Transparenz als Konzepte und Postulate eine zentrale Rolle. Die Erfüllung dieser Postulate wird zumeist als Voraussetzung für Vertrauen in KI-Anwendungen angesehen. Dieser Beitrag diskutiert die Zusammenhänge, Unterschiede und Abhängigkeiten zwischen diesen Konzepten aus einer interdisziplinären rechts- und politikwissenschaftlichen sowie ethischen Perspektive. Im Beitrag wird argumentiert, dass KI-Fairness nicht ausschließlich als statistisches Unterfangen verhandelt werden sollte, anhand dessen Verzerrungen (z.B. gender bias, racial bias) in Modellen evaluiert, verringert oder aufgelöst werden können. Vielmehr steht und fällt die Fairness von KI mit der Güte der Transparenz- und Kommunikationsstrategien, welche die gewählten Fairness-Verfahren verständlich und rechenschaftsfähig machen.[i]

1. Einleitung: ethische und rechtliche Anforderungen an Anwendungen Künstlicher Intelligenz

Anknüpfungspunkte dieses Beitrags über Fairness, Erklärbarkeit und Transparenz als Anforderungen an Anwendungen Künstlicher Intelligenz (KI) bilden die derzeitigen Bestrebungen zur Schaffung eines Rechtsrahmens für die KI-Entwicklung und -Nutzung, insbesondere durch den 2021 veröffentlichten Vorschlag für eine KI-Verordnung der Europäischen Union (Europäische Kommission 2021; im Folgenden: KI-VO-E), deren Aushandlung im Dezember 2023 abgeschlossen werden konnte. Die hier untersuchten normativen Anforderungskategorien an KI-Anwendungen werden damit von einem ethischen (*Soft-law*-)Rahmen in einen rechtsverbindlichen *Hard-law*-Rahmen überführt. Empirisch basiert der Beitrag auf Forschungen zu KI-Anwendungen für die Nutzung durch Sicherheitsbehörden, insbesondere durch die Polizei. Vor diesem Hintergrund geht dieser Beitrag zwei zentralen Fragen nach:

(1) Welche Zusammenhänge und Abhängigkeiten bestehen zwischen Fairness, Transparenz und Erklärbarkeit als normative Kategorien zur Bewertung von KI-Anwendungen?

(2) Wie stellen sich diese Abhängigkeiten im Hinblick auf KI-basierte Technologien im Sicherheitsbereich dar?

2. Fairness, Transparenz und Erklärbarkeit von KI-Anwendungen als verknüpfte

normative Kategorien im Übergang zwischen Ethik und Recht

Fairness und *Transparenz* sind keine neuen Kategorien – in verschiedenen wissenschaftlichen Disziplinen sind sie seit langem ebenso verankert wie im alltäglichen Sprachgebrauch. Dagegen ist *Erklärbarkeit* ein seltener gebrauchter Begriff, der speziell im Zusammenhang mit der Entwicklung von KI-Anwendungen an Bedeutung gewonnen hat. Diese drei normativen Konzepte sollen hier auf ihre Überschneidungen, Zusammenhänge, Abhängigkeiten und auch Unterschiede hin überprüft werden.

2.1 Fairness

Fairness ist im Vergleich zu *Transparenz* und *Erklärbarkeit* die allgemeinste hier betrachtete Kategorie. Auch im deutschen Alltagssprachgebrauch ist der Begriff als Anglizismus fest verankert. Im wissenschaftlichen Gebrauch unterscheiden sich das Verständnis von und der Umgang mit *Fairness* jedoch recht stark, je nach auslegender Fachdisziplin.

In rechtlicher Hinsicht sind insbesondere das EU-Recht und das Antidiskriminierungsrecht im Zusammenhang mit *Fairness* relevant. Das EU-Antidiskriminierungsrecht ist sowohl primärrechtlich als auch sekundärrechtlich ausgeprägt, wobei der weitestgehende Anwendungsbereich primärrechtlich in Art. 21 EU-Grundrechtecharta (GRCh) zu finden ist. Diese Vorschrift besagt, dass jedwede Diskriminierung aufgrund „des Geschlechts, der Rasse, der Hautfarbe, der ethnischen oder sozialen Herkunft, der genetischen Merkmale, der Sprache, der Religion oder der Weltanschauung, der politischen oder sonstigen Anschauung, der Zugehörigkeit zu einer nationalen Minderheit, des Vermögens, der Geburt, einer Behinderung, des Alters oder der sexuellen Ausrichtung […] verboten“ ist. Diese Liste ist sektorneutral und nicht abschließend, gilt allerdings nur für öffentliche Einrichtungen der EU und deren Mitgliedsstaaten, jedoch nicht ohne Weiteres für den Privatsektor (Wachter/Mittelstadt/Russell 2021: 13). Sekundärrechtlich existieren verschiedene Richtlinien, die ebenfalls Nichtdiskriminierung zum Inhalt haben.[ii] Darüber hinaus findet sich der Begriff *Fairness* als tragendes Prinzip in der englischsprachigen Fassung der EU-Datenschutzgrundverordnung – die deutsche Übersetzung mit „Treu und Glauben“ verengt das Gemeinte (Art. 5 Abs. 1 Nr. 1 DSGVO/GDPR). In anderen Bereichen versucht das EU-Recht, Fairnessstandards durch Richtlinien zu etablieren. Hierbei handelt es sich in der Regel um Mindeststandards, welche in den Mitgliedsstaaten in nationales Recht umgesetzt werden müssen. Viele Mitgliedsstaaten bieten, auch im Bereich der Nichtdiskriminierung, einen höheren Schutz als in den Richtlinien festgelegt. Dies führt auch innerhalb der EU zu einem fragmentierten Schutzstandard (Wachter/Mittelstadt/Russell 2021: 15). Neben dieser Herleitung von *Fairness* anhand des Antidiskriminierungsrechts spielen jedoch noch weitere Kriterien wie Nachvollziehbarkeit, Akzeptierbarkeit, hinreichende Berücksichtigung konkreter Einzelfälle und dergleichen eine gewichtige Rolle. Diese Aspekte leiten sich in rechtlicher Hinsicht eher aus Verfahrensrechten ab und umfassen wesentlich mehr als eine Gleichsetzung von *Fairness* und Nichtdiskriminierung. Hier zeigt sich sodann bereits die enge Verknüpfung von *Fairness* mit weiteren Prinzipien wie *Transparenz*. Der Versuch, *Fairness* und Nichtdiskriminierung in KI-Systemen zu automatisieren, erscheint bereits angesichts dessen als ein schwieriges Unterfangen.

War das Konzept von *Fairness* im ursprünglichen KI-VO-E (Europäische Kommission 2021) überraschenderweise gar nicht enthalten, findet es sich im aktuellen Kompromissvorschlag des Europäischen

Parlaments in Art. 4a Abs. 1 lit. e wie folgt wieder:

> „‘diversity, non-discrimination and *fairness*’ means that AI systems shall be developed and used in a way that includes diverse actors and promotes equal access, gender equality and cultural diversity, while avoiding discriminatory impacts and unfair biases that are prohibited by Union or national law” (European Parliament 2023).

Diese Neuerung ist begrüßenswert, weil damit nicht nur definitorische Ergänzungen hinzukommen, sondern Art. 4a generelle Prinzipien, die für alle KI-Systeme gelten sollen, festlegt. Unklar bleibt jedoch, was das konkret bedeutet und wie diese Anforderungen umgesetzt werden sollen.

Fairness kann sodann aus den Blickwinkeln anderer Disziplinen betrachtet werden. In den Fachgebieten der Informatik und Data Science bezieht sich *Fairness* gemeinhin auf statistische Methoden und Metriken, deren Ziel die Verringerung oder Beseitigung diskriminierender Verzerrungen in KI-Systemen ist. Der Begriff der *Fairness* wird hier als Beschreibung des Ziels (ein faires KI-Modell), sowie auch als statistische Bezifferung des Maßes an erreichter *Fairness* (die *Fairness* des Modells beträgt XY) genutzt. Zur Errechnung des Maßes an *Fairness* kommen verschiedene Verfahren zum Zuge, die normativ auf unterschiedlichen Prämissen aufbauen. Bei der statistischen Interpretation von *Fairness* beispielsweise als *demographic parity*, wird die erreichte Abbildung der Verteilung sozialer Gruppen in einer Gesamtbevölkerung im Modell beschrieben. Wenn *Fairness* durch die Metrik der *equalized odds* errechnet wird, steht dagegen die Chancengleichheit im Vordergrund, was durch die Veränderung der Treffer- und Abweisungsquoten eines Systems erreicht wird. Eine weitere Unterscheidung betrifft Verfahren individueller und gruppenbezogener *Fairness* (Mehrabi et al. 2021: 11ff.). Individuelle *Fairness* versucht zu gewährleisten, dass die Ergebnisse für vergleichbare Individuen auch gleich sind und Gruppenfairness, dass Ergebnisse eines KI-Systems so angeglichen werden, dass diese für unterschiedliche vordefinierte Gruppen gleich oder zumindest ähnlich sind (Brandner/Hirsbrunner 2023: 27). Ferner stellen sich jedoch auch Fragen, die über solche technischen Definitionen hinausgehen und andere gesellschaftliche Vorstellungen von *Fairness* und Gerechtigkeit beinhalten. Auch die Wahrnehmung der Betroffenen bezüglich „fairer“ Ergebnisse einer KI-Anwendung ist hier relevant (vgl. Aden/Kleemann 2023: 57).

In der kritischen Literatur verschiedener Fachdisziplinen wurde darauf hingewiesen, dass die Engführung von *Fairness* auf eine statistische Kategorie problematisch sein kann (Binns et al. 2018). Oftmals wird *Fairness* beispielsweise verkürzt als Schließung von Leistungslücken in den unterschiedlichen Ergebnissen zwischen demografischen Gruppen verhandelt, wobei die Genauigkeit des Systems bestmöglich erhalten bleiben soll (Mittelstadt/Wachter/Russell 2023: 2). Diese Vereinfachung führt dazu, dass viele Fairnessmaßnahmen eine Leistungsverschlechterung zur Folge haben. Dies wird auch als *levelling down* beschrieben (Mittelstadt/Wachter/Russell 2023). *Fairness* soll demzufolge dadurch erreicht werden, dass leistungsstärkere Gruppen auf das Niveau der am schlechtesten gestellten Gruppen herabgesetzt werden. *Levelling down* ist ein Symptom dafür, dass *Fairness* ausschließlich auf Gleichheit oder Ungleichheit zwischen Gruppen in Bezug auf Performanz und Ergebnisse reduziert wird; ebenfalls relevante Merkmale von Fragen distributiver Gerechtigkeit (Verteilungsgerechtigkeit) – wie Wohlfahrt oder andere schwierig zu quantifizierende Kriterien – werden dagegen ignoriert (Mittelstadt/Wachter/Russell 2023: 2). Das größte Problem des Versuchs, *Fairness* durch algorithmische oder mathematische Gleichungen zu gewährleisten, besteht demnach bereits darin, dass dieses Ziel unmöglich zu erreichen erscheint. Anders ausgedrückt besteht eine deutliche Diskrepanz zwischen statistischen Maßstäben für *Fairness* und den

kontextabhängigen, oft intuitiven und mehrdeutigen rechtlichen Diskriminierungsmaßstäben und Beweisanforderungen, etwa im Rahmen eines Gerichtsprozesses (Wachter/Mittelstadt/Russell 2021). Das Ziel, *Fairness* zu quantifizieren und ausschließlich statistisch zu optimieren, erscheint somit nicht zielführend, um *Fairness* im rechtlichen oder auch ethischen Verständnis herzustellen. Vereinheitlichte *Fairness*-Metriken und -Benchmarks stellen zwar wichtige Bewertungsverfahren und Mindeststandards für automatisierte Systeme dar, können jedoch nicht garantieren, dass KI-Anwendungen in konkreten Fällen von Betroffenen und lokalen Entscheidungsträger*innen (zum Beispiel Richter*innen) auch als fair wahrgenommen und bewertet werden.

2.2 Transparenz

Transparenz (englisch: *transparency*) ist ein Begriff, der in der Terminologie vieler wissenschaftlicher Disziplinen, in der Politik ebenso wie im alltäglichen Sprachgebrauch, fest verankert ist. Im Kontext digitaler Technologieentwicklung bezieht sich *Transparenz* oft auf den Zugang zu Informationen, die absichtlich offengelegt wurden. Aus ethischer Perspektive handelt es sich bei *Transparenz* nicht um einen Wert an sich, sondern um eine zentrale Vorbedingung, welche die Realisierung oder Beeinträchtigung verschiedener anderer ethischer Werte begründet (Turilli/Floridi 2009). Auch aus rechtlicher Perspektive hat *Transparenz*, beispielsweise ausgehend vom Primärrecht der Europäischen Union, eine hohe Gewichtigkeit (Fährmann/Aden/Arzt 2022: 311). Die EU-GRCh verleiht zentralen Elementen der *Transparenz* im Kontext des Datenschutzes (Art. 8 Abs. 2 S. 2) den Status eines Grundrechts: Jede Person hat „das Recht, Auskunft über die sie betreffenden erhobenen Daten zu erhalten und die Berichtigung der Daten zu erwirken“. Auch die EU-Datenschutzgrundverordnung (DSGVO) erwähnt in Art. 5 Abs. 1 lit. a *Transparenz* als Datenschutzgrundsatz.

Bezüglich aktueller Regulierungsbestrebungen für KI findet sich im vorgeschlagenen Art. 4a Abs. 1 lit. d des Kompromissvorschlags des Europäischen Parlaments *Transparenz* definiert als:

> „‘transparency’ means that AI systems shall be developed and used in a way that allows appropriate traceability and explainability, while making humans aware that they communicate or interact with an AI system as well as duly informing users of the capabilities and limitations of that AI system and affected persons about their rights.” (European Parliament 2023: 143)

Rechtlich normiert werden Transparenzpflichten für KI-Systeme unter anderem in Art. 13 KI-VO-E, wonach Nachvollziehbarkeit und *Transparenz* als Anforderungen definiert werden. Hochrisiko-KI-Systeme müssen demnach so konzipiert und entwickelt werden, dass der Betrieb hinreichend transparent ist und eine angemessene Interpretierbarkeit gegeben ist, die eine sinnvolle Verwendung der Ergebnisse durch die Nutzenden ermöglicht. *Transparenz* muss also auf geeignete Art und in angemessenem Maß gewährleistet werden (Aden/Kleemann 2023: 58). Artikel 7 der KI-Konvention des Europarats geht in eine ähnliche Richtung und besagt:

> „Each Party shall take appropriate measures to ensure that adequate oversight mechanisms as well as transparency requirements tailored to the specific contexts and risks are in place in respect of design, development, use and decommissioning of artificial intelligence systems." (Europarat 2023: 6)

Auffällig ist, dass alle Regelungen unbestimmte Rechtsbegriffe wie „angemessen" oder „geeignete Art" verwenden. Diese bedürfen jedoch der weiteren Konkretisierung, die die Regulierungsentwürfe nicht bereithalten. Weiterhin fokussiert sich beispielsweise Art. 13 KI-VO-E insbesondere auf die Interpretierbarkeit des Outputs von KI-Systemen durch deren Nutzer*innen (Ebers et al. 2021: 533 f.). Die Begriffe Interpretierbarkeit und *Erklärbarkeit* sowie deren Zusammenspiel bedürfen ebenfalls weiterer Präzisierung. Der KI-VO-E überantwortet die Konkretisierungsanforderungen spezifischer Transparenzpflichten in hohem Maß an die Anbieter*innen solcher Systeme, wobei diese lediglich an deren Selbsteinschätzung festgemacht werden (Ebers et al. 2021: 533 f.). Die technischen Maßnahmen zur Interpretation der Systemleistung und der *Transparenz* den Anbieter*innen von KI-Systemen zu überlassen erscheint mit Blick auf die zu schützenden Rechtsgüter von Personen, die von KI-basierten Entscheidungen betroffen sind, problematisch. Ist die Funktionsweise eines Modells nicht transparent, kann die Fehlerdiagnose einerseits und die Durchsetzung von Rechenschaftspflichten andererseits nicht gewährleistet werden (Busuioc/Curtin/Almada 2022: 21).

Die Transparenzregelungen im KI-VO-E zielen insbesondere auf die Nutzenden (Art. 13 KI-VO-E) und weniger auf die von KI-basierten Entscheidungen Betroffenen ab. Durch das Zusammenspiel beziehungsweise das Adressieren von KI-Entwickler*innen und -Nutzenden erinnert der KI-VO-E eher an ein haftungsrechtliches Regelungsregime. Dies ist mit Blick auf eines der maßgeblichen Ziele des Entwurfs, welches – neben der Gewährleistung der Produktsicherheit und der Einhaltung des EU-Rechts im Allgemeinen – insbesondere im Schutz der Grundrechte liegt, verwunderlich. Problematisch ist, dass eine solche Ausgestaltung grundsätzlich zu Misstrauen gegenüber der Anwendung von KI-Technologien führen könnte. Zunächst erscheint es, als wäre dies im Kompromissvorschlag des Europäischen Parlaments erkannt und darauf reagiert worden. Dieser bezieht sich in den Erwägungsgründen und auch in den konkreten Normen nun an etlichen Stellen mehr als zuvor auf *fundamental rights* (Grundrechte) und sieht sogar in Art. 29a KI-VO-Kompromissvorschlag ein *Fundamental rights impact assessment for high-risk AI systems* (European Parliament 2023: 42 ff.) vor. Allerdings wirkt die Art des *copy-paste*-Einfügens der Worte *fundamental rights* im gesamten Verordnungsentwurf wenig durchdacht. An keiner Stelle wird klargemacht, was dies *in concreto* bedeutet, worin der Unterschied zur Version der Kommission besteht und wie diese Anforderungen realisiert werden sollen.

2.3 Erklärbarkeit

Erklärbarkeit ist im allgemeinen Sprachgebrauch deutlich weniger etabliert als *Fairness* und *Transparenz*, hat sich aber im Kontext der KI-Debatte als zentrales Prinzip zur Herstellung von Vertrauenswürdigkeit etabliert. Aus ethischer Perspektive beschreibt *Erklärbarkeit* die Fähigkeit, anderen zu erklären, warum man etwas getan oder eine Entscheidung getroffen hat. Entsprechend verwoben ist diese Qualität auch mit der Eigenschaft, Verantwortung und Verantwortlichkeit für etwas anzuzeigen (Coeckelbergh 2020).

Im KI-VO-E fehlte der Begriff zunächst und wurde erst mit den Änderungsvorschlägen des Europäischen Parlaments am Ende der 1. Lesung explizit eingeführt (European Parliament 2023: 143). Bezüglich der *Erklärbarkeit* von KI-basierten Entscheidungen wurde bisher vor allem zwischen *Transparenz* und sogenannter post-hoc-Interpretierbarkeit unterschieden (im Detail: Ebers 2021: 110f.). Anknüpfungspunkte hierzu finden sich etwa in Art. 13 Abs. 1 KI-VO-E, wonach Hochrisiko-KI-Systeme so konzipiert sein müssen, dass es den Nutzer*innen möglich ist, die Ergebnisse angemessen zu interpretieren und zu verwenden. Im Kompromissvorschlag des Europäischen Parlaments wurde im Zusammenhang mit dem neu vorgeschlagenen Art. 4a Abs. 1 lit. d *Transparenz* definiert und in diesem Zusammenhang ein starker inhaltlicher Bezug zu *Erklärbarkeit* hergestellt (siehe oben, Abschnitt 2.2).

Um dies weiter zu konkretisieren, erscheint es ratsam, insbesondere auf die Disziplin der *Explainable AI* (XAI) zurückzugreifen. *XAI* beschreibt solche KI-Modelle, die auch von außen verständlich und erklärbar sind und trotz dieser *Erklärbarkeit* eine hohe Lernleistung aufweisen (Linardatos 2022: 65). Diese Techniken, die ursprünglich von Ingenieur*innen entwickelt wurden, um das Verhalten von Modellen zu untersuchen und Fehler zu beheben, werden jetzt als Lösung für *Black-Box*-Problematiken und diesmal in institutionellen und Governance-Kontexten weiterentwickelt. Hierbei wird, vereinfacht gesagt, ein *Black-Box* -Modell mit einem zweiten *Post-hoc*-Erklärungsmodell verknüpft und im Wesentlichen ein einfacherer Algorithmus zur Erklärung der *Black-Box* verwendet, um das *Black-Box*-Modell in etwas Verständlicheres umzuwandeln (Busuioc/Curtin/Almada 2022: 11). Das *Post-hoc*-Erklärungsmodell ist dabei oftmals ein völlig anderes Modell als die zugrundeliegende *Black-Box*. Die dadurch erzeugte Erklärung ist damit nicht ohne Weiteres für die Nutzenden oder Betroffenen selbst nachvollziehbar. Sie ist vielmehr eine von den Anbieter*innen der Systeme entwickelte Erklärung, wodurch die Anbieter*innen eine Art Vermittlerrolle zwischen den Nutzenden von KI-Systemen und den von KI-basierten Entscheidungen Betroffenen einnehmen. Für die Anbieter*innen besteht der Vorteil offenkundig darin, dass sie es vermeiden die *Black Box* öffnen zu müssen. Außerdem können dadurch die zugrundeliegenden Modelle unter Wahrung des Schutzes des geistigen Eigentums geheim bleiben und dennoch durch das Erklärungsmodell ein bestenfalls verständliches Ergebnis liefern (Busuioc/Curtin/Almada 2022: 11). Die starke Rolle der Anbieter*innen in diesem Prozess ist jedoch durchaus auch kritisch zu betrachten. Entscheiden diese allein, wie die Erklärung und die Offenlegung entscheidungsrelevanter Parameter aussehen, besteht hier sogar die Gefahr von Intransparenz und womöglich unfairen Entscheidungen. Die Qualität, die Genauigkeit und der Wahrheitsgehalt der bereitgestellten Informationen sollten nicht von der Vertrauenswürdigkeit eines Vermittlers abhängen, der ein eigenes großes Interesse am Inhalt der präsentierten Informationen hat.

2.4 Transparenz als Voraussetzung für Fairness und Erklärbarkeit

Transparenz dahingehend zu interpretieren, dass lediglich alles offengelegt wird, führt im Zusammenhang mit KI keineswegs automatisch zu mehr Verständnis oder *Erklärbarkeit*. Modelle des *Machine Learning* (ML) oder neuronale Netze sind extrem komplex, haben Hunderte von Schichten (*Layers*), Tausende Merkmale und Milliarden verschiedene Gewichtungen, die zu einem Vorhersageergebnis führen können (Busuioc/Curtin/Almada 2022: 9). „Seeing inside a system does not necessarily mean understanding its behavior and origins" (Ananny/Crawford 2018: 980). *Transparenz* muss also neu verstanden, interpretiert und an die Rahmenbedingungen von KI angepasst werden.

Der Diskurs über *Transparenz* hat sich von der ursprünglichen Vorstellung von einer reinen

Zurverfügungstellung aller – auch für viele Stakeholder irrelevanter – Informationen hin zu einer Debatte über *Erklärbarkeit* durch *Transparenz* verschoben (Busuioc/Curtin/Almada 2022: 8). *Transparenz* soll also nicht mehr nur dadurch verwirklicht werden, dass Informationen bereitgestellt werden, was schnell zu einem *information overflow* führen kann. Vielmehr soll sie das Verstehen von KI-Systemen erleichtern. Dieses Postulat findet sich in vielen der neueren politischen und regulatorischen Dokumente zu KI wieder.[iii]

Diese Entwicklung hin zu *Transparenz* durch Erklärung wird in technischer Hinsicht oftmals versucht in Form von *XAI* zu realisieren. *XAI* dient jedoch lediglich als Oberbegriff für diverse Techniken mit denen versucht wird, der *Black-Box*-Problematik von KI-Modellen Herr zu werden. Die Idee von *Transparenz* durch Erklärung ist zu einem gewissen Grad sinnvoll. Werden die Regulierungs- und Governance-Ansätze jedoch einzig auf diesen Ansatz konzentriert, tun sich neue Probleme auf. Dann könnte sogar *Transparenz* verloren gehen, weil eine Fokussierung auf *Explainability*-Methoden Beeinflussungsmöglichkeiten hinsichtlich der Umsetzung von Transparenzpflichten öffnet (Busuioc/Curtin/Almada 2022: 11). Auch hier gilt es zu klären, wer diese Pflichten erfüllen muss (häufig die Anbieter*innen von KI-Systemen) und welche Interessen dabei verfolgt werden. Wie beschrieben, wird den Anbieter*innen diese Pflicht übertragen, aber auch die genaue Ausgestaltung zu großen Teilen überlassen. Das führt dazu, dass die Anbieter*innen mit ihren technischen Mitteln kontrollieren, was offengelegt wird. Die Stakeholder, die diesen Prozess auf diese Weise technisch in der Hand haben, sind zugleich diejenigen mit den stärksten Geheimhaltungsinteressen (Busuioc/Curtin/Almada 2022: 11).

Nichtsdestotrotz kann *Transparenz* der Grundstein für *Erklärbarkeit* und letztlich auch für *Fairness* sein. Notwendig ist dafür eine smarte Regulierung, die beispielsweise eine adressatenspezifische Differenzierung vorsieht (näher hierzu unten in Abschnitt 4.).

3. Elemente erklärbarer Fairness von Künstlicher Intelligenz in Sicherheitsanwendungen

Um reale *Fairness* im KI-Kontext zu erreichen, sind eine Reihe von Transparenzmethoden denkbar. Hier werden diese typologisiert und in den Anwendungsbereich der polizeilichen Ermittlungstätigkeit eingeordnet.

3.1 Verpflichtung zur Entwicklung und zum Einsatz fairnesssensibler KI

Wegen des hohen Stellenwerts, welchen die Risiken algorithmischer Diskriminierung über die letzten Jahre in der öffentlichen und medialen Debatte erhalten haben, sind Anbieter*innen von KI-Systemen dazu übergegangen, die Ethik und *Fairness* in ihren Produktbeschreibungen und -bewerbungen zu adressieren. Beispielsweise deklariert Palantir, ein marktführender Anbieter digitaler Sicherheitssoftware, dass bei der Auswahl von Anwendungsfällen und der Entwicklung von KI-Systemen verschiedene ethische Risiken berücksichtigt werden. So bekennt sich die Firma in ihrer KI-Ethik-Deklaration zu einem bewussten und vorsichtigen Umgang mit sensitiven Attributen wie *race* und *gender*:

> „Where such sensitive categories are implicated in programs that use AI either by necessity or choice, their inclusion should be identified and, where appropriate, contextually justified, in addition to complying with any formal requirements under applicable law.“ (Palantir o. D.)

Viele Anbieter*innen adressieren *Fairness* jedoch nur auf sehr viel allgemeinerem Niveau und deklarieren oft unspezifisch, dass ihr KI-System *fair* oder *un-biased* sei. Ein detaillierter ausgearbeitetes und öffentlich zugängliches Statement zum Ansatz des Umgangs mit algorithmischer Diskriminierung sollte aber zu den Mindestanforderungen für KI-Entwickler*innen gehören.

3.2 Begründung der gewählten Fairness-Metriken, De-Biasing-Verfahren und erweiterte Technikfolgenabschätzung

Auch wenn sich KI-Anbieter*innen zu fairer und ethischer KI bekennen, beschreiben sie selten konkret, welche technischen und organisatorischen Maßnahmen getroffen wurden, um den Schutz vor algorithmischer Diskriminierung in ihrer Produktentwicklung und -vermarktung zu berücksichtigen. In einer konsequenten Darstellung der Maßnahmen für algorithmische *Fairness* müssten die genutzte *Fairness*-Metrik, das konkrete Vorgehen für das *De-Biasing*, die erreichten Schwellenwerte und die Qualität dieser Schwellenwerte gegenüber anderen Anbieter*innen und Systemen vermittelt werden. Ebenso müsste kommuniziert werden, in welche organisatorischen Maßnahmen die technischen Verfahren eingebettet werden. Im Zusammenhang mit aktueller KI-Regulierung finden sich hier auch Anknüpfungspunkte an rechtliche Governance- und Data-Management Vorgaben. Diese finden sich sowohl im KI-VO-E und allen auf EU-Ebene in der Trilog-Aushandlung genutzten Dokumenten (vgl. bspw. Art 10, Europäische Kommission 2021 und European Parliament 2023) als auch im Vorschlag für eine KI-Konvention des Europarats (vgl. bspw. Art. 10, 15, Präambel Nr. 17 Option C: Europarat 2023: 4, 7f.). Künftig werden insbesondere im Zusammenhang mit dem Training von KI-Modellen und den dafür notwendigen Daten Verfahren etabliert werden müssen, die beispielsweise die Datenerfassung, Datenaufbereitungsvorgänge, aber auch Untersuchungen bezüglich möglicher Verzerrungen betreffen. Für diese und weitere Verfahren müssen geeignete Daten-Governance und Datenverwaltungsverfahren geschaffen werden.

Damit die dargestellten Maßnahmen von Außenstehenden bewertet werden können, müssten die Transparenzmaßnahmen gegenüber den bestehenden Risiken begründet werden. Entsprechend wäre stets eine allgemeine Technikfolgenabschätzung des Systems notwendig, bevor überhaupt statistische Lösungen ausgewählt und operationalisiert werden. Für diese Art auf den Anwendungsfall bezogener Dokumentation

gibt es bereits Formate wie das von der Firma Microsoft entwickelte R*esponsible AI Impact Assessment Template* (Microsoft 2022).

3.3 Technische Dokumentation des Systems

Um die Passung zwischen Risiken und Adressierungsmaßnahmen im Anwendungsfeld zu evaluieren, müssten Externen technische Informationen zu Input-Daten, eingesetzten Modellen und Informationen zum Lernprozess (Aufbau des Verfahrens, Einstellung der Hyperparameter, etc.) zur Verfügung stehen. Auch hierzu gibt es bereits ausgearbeitete Dokumentationsformate wie *Data Sheets* (im Detail: Gebru et al. 2021) und *Model Cards* (im Detail: Mitchell et al. 2018). In diesem Zusammenhang könnten Anforderungen an KI-Trainingsdatensätze definiert werden, die rechtliche und ethische Standards erfüllen. Die Trainingsdatensätze können entweder im Nachhinein damit abgeglichen werden, oder besser, die Anforderungen bereits beim Anlegen neuer Datensätze umgesetzt werden. Allgemeine Inhalte können dabei unter anderem die gesetzlichen Vorgaben bezüglich des Datenschutzes sein. Zur Konkretisierung können beispielsweise Informationen zu Zeitpunkt und Methode von Anonymisierung oder Pseudonymisierung, zur Herkunft der Daten und zu einer informierten Einwilligung geliefert werden (vgl. Aden/Kleemann 2023: 61f.). Die Etablierung einer solchen Praxis könnte dazu führen, künftige Trainingsdatensätze rechtlich und ethisch sicherer zu gestalten, eine Vergleichbarkeit herzustellen und dies auch nachweisen zu können. Hier könnten ebenfalls abgestufte Versionen (vgl. auch Abschnitt 4) bereitgestellt werden, bei denen ein *Data-Sheet*, welches möglicherweise in ein Gerichtsverfahren eingebracht wird, speziellere Informationen enthält als eines, welches der breiteren Öffentlichkeit zugänglich gemacht werden soll (Aden/Kleemann 2023: 62).

3.4 Zugang zum Source Code und zu Echtdaten

Der Zugang zum *Source Code* des Systems würde es ermöglichen, dieses nachzustellen und die kommunizierten *Fairness*-Angaben technisch zu überprüfen oder zu reproduzieren. Auch wenn der Zugang zum *Source Code* als Mittel der Transparenzherstellung großes Potential aufweist, ist die Umsetzung gerade für Applikationen im Sicherheitsbereich schwierig. So kann erstens die Zugänglichmachung des Programmiercodes zu einer Vulnerabilität gegenüber externen Angriffen führen (*adversarial attacks*) (Qiu et al. 2019: 909). Zweitens ist im Fall einer Entwicklung von Software durch die Privatwirtschaft aus urheber- und nutzungsrechtlichen Gründen fraglich, ob diese eine *Open-Source*-Entwicklung unterstützen und vorantreiben würden. Drittens bestehen KI-Systeme häufig aus vielen verschachtelten Systemen, so dass die Interpretier- und *Erklärbarkeit* offener Codes fraglich scheint.

Mit der starken Verbreitung vortrainierter Transformer-Modelle in den vergangenen Jahren verschieben sich jedoch hier die Bedingungen für *Transparenz*. Oft werden nun vortrainierte *Open-Source*-Modelle genutzt und mittels *Transfer-Learning* auf neue Aufgaben umtrainiert. Die *Transparenz* erschließt sich entsprechend über die Offenheit des zugrundeliegenden Modells, nicht unbedingt über das angepasste Modell. Allerdings gibt es hier durch die Verschiebung des Trends hin zu proprietären *Large Language Models* (LLMs) auch wieder eine Gegenbewegung, deren Konsequenzen für den Sicherheitsbereich noch nicht abzuschätzen sind.

Während die Zugänglichmachung von Daten im Sinne von *Open Data* (Fischer/Hirsbrunner/Teckentrup 2022) für viele KI-Anwendungsbereiche von enormem Wert sein wird, ist sie für viele Applikationen im Sicherheitsbereich nicht praktikabel. Daten von Sicherheitsbehörden, wie etwa Falldaten der Polizei, unterliegen starken Datenschutzbestimmungen und polizeitaktischen Vertraulichkeitsanforderungen. Es ist illusorisch, dass Echtdaten hier im großen Umfang zur Überprüfung von Systemen verfügbar gemacht werden. Allerdings ist der Datenschutz eine Problematik, die nicht erst im Rahmen der Überprüfung von Systemen auftritt, sondern bereits in der Entwicklungsphase. Da Echtdaten aus dem Polizeikontext meist auch den kommerziellen oder nicht-kommerziellen KI-Entwickler*innen nicht zur Verfügung stehen, behelfen sich diese mit offenen Datensätzen mit ähnlichen Eigenschaften. Echtdaten kommen erst in der Testphase der KI-Systeme und auch dort nur punktuell zum Einsatz. Oft werden die sich im Test befindlichen Systeme hierzu auf Server der Sicherheitsbehörden gespielt und aus Datenschutzgründen erst dort mit Echtdaten versorgt. Hinzu kommt, dass die Nutzung polizeilicher Echtdaten im Rahmen von KI-Entwicklung durchaus problematisch sein kann. Je nachdem, in welchem Stadium sich die KI-Entwicklung befindet und mit welchen Ansätzen trainiert wird, muss bedacht werden, dass polizeiliche Echtdaten die Sichtweise der Ermittler*innen aus vorherigen Fallkonstellationen widerspiegeln und notwendigerweise nur einen unvollständigen Teil der Wirklichkeit darstellen. Daraus folgt, dass solche Daten in hohem Maße Verzerrungen enthalten können, die sich bereits aus dem Ermittlungsansatz, den subjektiven Wahrnehmungen der Ermittelnden selbst und generell dem sozialen Kontext ableiten (Barocas/Hardt/Narayanan 2019: 11ff.; González Fuster 2020: 41f.). Dennoch gibt es Bereiche, in denen Echtdaten nützlich erscheinen. Im Zusammenhang mit automatisierter Texterkennung könnte es beispielsweise zielführend sein, die Modelle nicht mit wohlformulierten oder offiziellen Texten zu trainieren, sondern gerade Slang, gemischte Sprachen, *Messenger-Chats*, Umgangssprache und fehlerhafte Sätze zu nutzen. Hier könnte jedoch, anstatt auf reale Texte mit personenbezogenen Daten zurückzugreifen, auch mit synthetisch generierten und den beschriebenen Anforderungen entsprechender Textdaten, trainiert werden. Ähnliche Konstellationen könnten sich im Bereich der Objekterkennung ergeben. Je nach Ausgestaltung und Einsatzzweck des KI-Systems könnte auch mit Objektdaten trainiert werden, die aus einem Ermittlungsverfahren stammen, wenn es bei der Objekterkennung gerade nicht um Personen, sondern um einzelne Objekte wie Schmuck oder Warengüter geht, die keinen Personenbezug aufweisen, aber in bestimmten Konstellationen ähnlich auftauchen (bspw. Schmuckauslage für den Weiterverkauf gestohlener Waren).

4. Adressatenspezifische Ausdifferenzierung von Fairness, Transparenz und Erklärbarkeit im sicherheitsbehördlichen Kontext

Hinsichtlich der Anforderungen (*Fairness*, *Transparenz*, *Erklärbarkeit*) an vertrauenswürdige KI sollte berücksichtigt werden, dass verschiedene Akteur*innen unterschiedliche Bedarfe und Interessen beim Umgang mit einer KI-Anwendung haben können. Hier ist eine Ausdifferenzierung erforderlich, die Möglichkeiten zur Einnahme individueller Perspektiven zulässt.

Im Kontext polizeilicher KI-Nutzung bestehen bezüglich der effektiven Gewährleistung von *Erklärbarkeit*, *Transparenz* und *Fairness* besondere Herausforderungen. Verschiedene Akteur*innen, etwa KI-Fachleute, Ermittler*innen, Gerichtssachverständige, Anwält*innen, Richter*innen, Beschuldigte, sonst Betroffene und die Öffentlichkeit haben unterschiedliche Bedürfnisse und Interessen, die nicht nur berücksichtigt werden müssen, sondern teilweise in einem Spannungsverhältnis stehen, welches womöglich nicht ausreichend aufgelöst werden kann. Um diesen Problemen dennoch bestmöglich zu begegnen, kann eine Risikomatrix

oder bildlich beschrieben, ein „Zwiebelmodell“ entworfen werden, welches eine gruppenspezifische Gewährleistung und Offenlegung verschiedener Informationen ermöglicht (Aden et al. 2023: 302).

Ein „Zwiebelmodell“ würde bedeuten, dass beispielsweise KI-Fachleute, denen der Betrieb der KI-Anwendung obliegt, oder Gerichtssachverständige umfangreichere technische und inhaltliche Informationen erhalten, um die Entscheidungen des KI-Systems fundiert und kritisch bewerten zu können. Ermittler*innen erhalten nach diesem Prinzip die Informationen, die für ihre Arbeit notwendig sind, aber auch eine rechtmäßige Nutzung gewährleisten, ohne dass darüber hinausgehende Expertise zwingend erforderlich ist. Die allgemeine Öffentlichkeit befände sich in der äußeren Schicht und würde bei Bedarf oder Interesse im erforderlichen Umfang allgemeinere Informationen erhalten (Aden et al. 2023: 302). Von KI-basierten Entscheidungen Betroffene müssen indes gesondert betrachtet werden. Eine vorherige Filterung oder Begrenzung von Informationen könnte menschenrechtlichen Prinzipien und damit einem effektiven Menschenrechtsschutz entgegenstehen. Zur Gewährleistung eines solchen Konzepts bedarf es differenzierter Anforderungen, die bereits bei der Entwicklung des KI-Systems mitbedacht und von Beginn an, auch im Rahmen der Hard- und Softwareentwicklung, in ein holistisches Konzept integriert werden. Partizipationsoffenheit und eine interdisziplinäre, integrierte Technikentwicklung sind dafür maßgeblich zu erfüllende Voraussetzungen (Gressel/Orlowski 2019: 71f.).

Eine solche konkrete Ausdifferenzierung könnte anhand dreier Akteursgruppen beispielhaft dargestellt werden. Wenn im polizeilichen Kontext KI-Anwendungen genutzt werden, sind unterschiedliche Stakeholder und Zielgruppen involviert. Zur Verdeutlichung der Abhängigkeiten der Konzepte und der zielgruppenspezifischen Transparenzanforderungen können beispielsweise (1) Ermittler*innen, (2) die digitale Ermittlungsunterstützung, also Expert*innen in den Sicherheitsbehörden, und (3) Aufsichtsbehörden als Gruppen von KI-Nutzer*innen definiert werden. Klassischerweise wird diesen Gruppen eine Vielzahl an zielgruppenunspezifischen Informationen zur Verfügung gestellt, was zur Folge hat, dass die für die einzelnen Beteiligten wirklich relevanten Informationen und Erklärungen nicht mehr selektiert werden können. Ausgehend von den Überlegungen zu *Human-Centered Design* (Baumer 2017) und *Human-Centered Machine-Learning* (Gillies et al. 2016) können aus Sicht der Gruppen von Beteiligten verschiedene Fragen aufkommen, die unterschiedlicher Lösungsansätze bedürfen.

4.1 Ermittler*innen

Ermittler*innen sind Polizeibeamt*innen, die in ihrer Funktion als Ermittlungspersonen KI-Anwendungen nutzen bzw. von spezialisierten Dienststellen oder externen Dienstleistern mit KI-basierten Ermittlungsansätzen „versorgt“ werden. Als Akteur*innen, die oft schnell, aber dennoch fundiert entscheiden müssen, benötigen Ermittler*innen einerseits klare und eindeutige Informationen. Andererseits müssen die Verantwortlichkeiten vorab klar definiert sein. Bei ersterem ist immer darauf zu achten, dass die KI-basierte Entscheidung zwar leicht zugänglich, aber deutlich erkennbar als Unterstützungshilfe ausgegeben werden muss. Dabei müssen die Leistungsgrenzen des Systems klar aufgezeigt werden. Für die Ermittlungsarbeit ebenfalls ausschlaggebend sind klare Informationen bezüglich der Verantwortung. Um dies zu gewährleisten, bietet sich die Entwicklung einer Verantwortungskaskade an. Oftmals bestehen die praktischen Probleme der Ermittlungsarbeit in unklaren Verantwortlichkeiten. Dies kann auch im Rahmen von KI dazu führen, dass nicht klar ist, wer welche Systeme wie einsetzen darf und wer letztlich verantwortlich ist. Dazu zählen Fragen im Vorfeld, etwa wer entscheidet, welches Tool überhaupt angeschafft wird, welches geeignet ist und wie das Training an diesen Systemen vonstatten geht. Für die

Ermittler*innen wird bei der Frage, ob sie ein Tool einsetzen, letztlich entscheidend sein, dass sie sicher gehen können, dieses rechtskonform und bestimmungsgemäß zu verwenden. Dies kann durch eine klare Verantwortungskaskade speziell für die Nutzung KI-basierter Systeme innerhalb der Behörde realisiert werden. Bezüglich der technischen Umsetzung sollten Anforderungen wie eine intuitive *Interface-G* estaltung und gute Handhabbarkeit implementiert werden, ebenso bedarf es Klarheit darüber, dass es sich lediglich um ein Unterstützungs- und kein Entscheidungstool handelt. Hierbei sind auch Maßnahmen gegen die „Technikgläubigkeit“ von Anwender*innen avancierter Technologien erforderlich, etwa durch die Entwicklung von Irritationen und die Darlegung möglicher Gegenargumente im Hinblick auf die Interpretation KI-basierter Ermittlungsansätze. Die Gefahren eines *information overflow* müssen insbesondere bei dieser Gruppe von Akteur*innen mitbedacht werden.

4.2 Digitale Ermittlungsunterstützung

Die digitale Ermittlungsunterstützung ist die zweite, hier exemplarisch betrachtete Gruppe von Akteur*innen. Dabei handelt es sich um Teams aus IT-Fachleuten innerhalb der Sicherheitsbehörden, deren Aufgabe darin besteht, Ermittler*innen bei der Auswertung großer Datenmengen und digitaler Asservate zu unterstützen. Mit Blick auf die wachsenden digitalen Kommunikationsströme handelt es sich hierbei um ein aufstrebendes Tätigkeitsfeld innerhalb der Polizeibehörden, welches bisweilen auch an Fachbereiche der digitalen Forensik oder der Bekämpfung von Cyberkriminalität angegliedert ist. Die hier tätigen Expert*innen verfügen typischerweise über eine fundierte Ausbildung in Bereichen wie der Informatik, der Datenwissenschaft, der Computerlinguistik oder der digitalen Forensik. Entsprechend sind sie nicht nur in der Lage, KI-basierte Anwendungen zur Datenanalyse zu nutzen, sondern sie ebenso technisch zu bewerten und zu überprüfen.

Der Aufgabenbereich dieser Gruppe von Akteur*innen kann eine Vielzahl von IT-bezogenen Leistungen miteinschließen, die mit bestimmten Transparenzanforderungen an Technik einhergehen. Zum einen kann eine Aufgabe der IT-Fachleute darin bestehen, auf Basis frei verfügbarer Software (etwa Python-Scripts und -bibliotheken) kleinere Computerprogramme selbst zu entwickeln, die für die Auswertung von Daten eingesetzt werden. Zum anderen beinhaltet der Aufgabenbereich technische Bewertungen und Risikoeinschätzungen. Solche Einschätzungen können unter anderem im Beschaffungsprozess für neue Software eine Rolle spielen, aber auch im Falle des Auftretens von Problemen mit bereits verwendeter Software.

Um diese Aufgaben bewältigen zu können, sind die Fachleute auf verschiedene Transparenzeigenschaften und -maßnahmen angewiesen. Viele dieser Eigenschaften lassen sich unter dem Feld der *XAI* subsummieren. Dies kann etwa *Post-hoc*-Erklärungsmodelle beinhalten, die zur Erklärung der Funktionsweise des komplexeren, aber opaken Primärmodells verwendet werden. Ebenso relevant können jedoch technische Komponenten sein, die Input-Daten, die Performanz des Systems oder andere Eigenschaften vermitteln (Liao/Gruen/Miller 2020). Im Kontext der *Fairness*-Diskussion sind hier beispielsweise Techniken im Einsatz, die über einfache Erklärungen hinausgehen und komplexere „Was wäre wenn“-Experimente ermöglichen.

4.3 Aufsichtsbehörden und parlamentarische Gremien

Aufsichtsbehörden sind öffentliche Behörden und Stellen wie Datenschutz- oder Polizeibeauftragte, die Kontrollaufgaben wahrnehmen. Sie haben Informationsbedarfe, die sich stärker auf das Funktionieren und die Verlässlichkeit einer KI-Anwendung im Allgemeinen beziehen. Auch parlamentarische Gremien mit Kontrollaufgaben, etwa Fach- und Untersuchungsausschüsse, haben vergleichbare Informationsbedarfe. Im KI-VO-E ist außerdem in den Art. 56ff. die Einrichtung eines Europäischen Ausschusses für KI (*Artificial Intelligence Board*) vorgesehen. Die genaue Ausgestaltung überlässt der Verordnungsentwurf allerdings den Mitgliedstaaten, wo zu entscheiden sein wird, ob die Kontrolle von KI-Anwendungen beispielsweise in die bestehende Datenschutzaufsicht integriert oder neu zu etablierenden Aufsichtsbehörden übertragen wird (Aden/Kleemann 2023: 57). Der ursprüngliche KI-VO-E sah ferner überhaupt keine Möglichkeit für individuelle Beschwerden von Betroffenen vor. Der Kompromissvorschlag des Europäischen Parlaments erkennt jedoch die Notwendigkeit eines solchen Mechanismus an, woraufhin ein eigens dafür ergänztes Kapitel 3a (*Chapter 3a Remedies*) mit den Art. 68a-68e vorgeschlagen wird (European Parliament 2023: 106ff.). Dabei ist bereits jetzt absehbar, dass diese Beschwerdemechanismen als Bindeglied zwischen den KI-Entwickler*innen, den Nutzenden (hier: Sicherheitsbehörden) und den von KI-basierten Entscheidungen Betroffenen dienen werden. Dies führt dazu, dass hier eine gewisse Übersetzungsleistung erforderlich sein wird. Die verschiedenen Interessengruppen, die mit diesen Kontrollbehörden interagieren, sollten bestenfalls in den Behörden selbst vertreten sein. Das bedeutet, dass technisch versierte KI-Expert*innen dort mit Betroffenenvertreter*innen, juristischen Expert*innen sowie Vertreter*innen der Zivilgesellschaft zusammenarbeiten müssen. Aufsichtsbehörden müssen KI-basierte Systeme dahingehend prüfen können, ob diese den gesetzlichen Anforderungen entsprechen und hohe Schutzstandards für die Grundrechte Betroffener gewährleisten. Dafür bedarf es wesentlich weitergehender Informationen, also Transparenzmaßnahmen, um *Erklärbarkeit*, eine effektive Kontrolle und letztlich faire Entscheidungen zu gewährleisten.

Durch Entwickler*innen und Anbieter*innen von KI-Software im Sicherheitsbereich muss ein standardisiertes Format geschaffen werden, welches es den Expert*innen in den Aufsichtsbehörden ermöglicht, notwendige Informationen bezüglich eines KI-Systems zu erhalten. Dazu zählen zum Beispiel Fragen nach der Herkunft des Datensatzes, der zu KI-Trainingszwecken verwendet wurde, wobei bereits hier unterschiedliche Interessen berücksichtigt werden müssen. Es könnte beispielsweise ausreichen, aggregierte Informationen (Statistiken, Visualisierungen) über die Trainingsdatensätze und die Datensubjekte zu erhalten. Werden diese standardisiert, könnte eine Vergleichbarkeit hergestellt werden. Dies könnte mittels *Data Sheets* erfolgen. Dadurch könnte Akzeptanz auf Seiten der Industrie hergestellt werden, die in vielen Fällen aus manchmal guten Gründen eine komplette Offenlegung von Trainingsdaten vermeiden möchte, etwa im Hinblick auf die Cybersicherheit. Denn auf der Basis zu detaillierter offengelegter Informationen könnten Angriffsvektoren für Cyberattacken geschaffen werden. Werden mittels Wissens über die Trainingsdaten Schwachstellen des ML-Modells herausgefunden und ausgenutzt, handelt es sich hier nicht um eine direkte Manipulation des ML-Modells, sondern es werden mittels *exploratory attacks* Eingaben genutzt, die schlussendlich zu Fehlklassifizierungen des Modells führen. Weitere Gefahren bestehen darin, dass durch Wissen über die Trainingsdaten die Eigenschaften (Integrität, Vertraulichkeit, Verfügbarkeit) des Systems angegriffen werden können. Bei einem Angriff auf die Eigenschaften des Systems können beispielsweise durch *reverse engineering* Informationen über das Modell oder über die Trainingsdaten herausgefunden werden. (Sultanow et al. 2022: 148f.). Werden jedoch mittels Statistiken über Trainingsdaten die für die *Erklärbarkeit* relevanten Informationen aufbereitet und so bereitgestellt, dass die Trainingsdaten selbst robust geschützt sind, kann diese Gefahr verringert werden und gleichzeitig der Kritik von möglicher Überregulierung begegnet werden. Eine solche Art der Aufbereitung von Statistiken über Trainingsdaten könnte auch weitergehend den Weg der Daten nachvollziehbar machen und lediglich die

bereits bewerteten Daten auswerten.

Um nachzuweisen, dass ein System wie vorgesehen funktioniert, könnten Aufsichtsbehörden auch geheime Datensätze von unbeteiligten Dritten zur Verfügung gestellt werden. Diese Methoden würden maßgeblich zur *Erklärbarkeit* der Systeme beitragen, was wiederum die Konsequenzen für die Betroffenen nachvollziehbarer machen würde. Der zweckmäßige Einsatz des Systems könnte so nachgewiesen werden, wobei an dieser Stelle die zwingend notwendige Übersetzungsleistung für Menschen ohne spezielle KI-Ausbildung ansetzen muss.

5. Schlussfolgerungen und Ausblick

Dieser Beitrag hat gezeigt, dass *Fairness*, *Erklärbarkeit* und *Transparenz* bei der Entwicklung und Nutzung von KI-Anwendungen, auch im Sicherheitsbereich, eng miteinander zusammenhängen. Eine speziell auf Gruppen von Akteur*innen wie Ermittler*innen oder Aufsichtsbehörden zugeschnittenes Transparenzkonzept ist Voraussetzung dafür, dass diese Gruppen für ihre Aufgaben die passenden und erforderlichen Informationen bekommen, ohne mit einer Flut unverständlicher Informationen konfrontiert zu werden. Diese *Transparenz* ist notwendige Bedingung für eine erklärbare KI. Allerdings reicht *Transparenz* für die *Erklärbarkeit* von KI-basierten Ergebnissen noch nicht aus, sondern bedarf einer inhaltlichen Anreicherung, die technisch zu implementieren ist. Damit könnte die Herstellung von *Erklärbarkeit* der Ergebnisse von KI-Anwendungen im Sicherheitsbereich zugleich einen Beitrag zur Rechtmäßigkeit staatlichen Handelns leisten, wenn es gelingt, die verwendete Soft- und Hardware so auszugestalten, dass sie rechtmäßiges Handeln nicht nur unterstützt, sondern auch durch technische Vorkehrungen „erzwingt“. Auf diese Weise könnten *Transparenz* und *Erklärbarkeit* auch einen Beitrag zur *Fairness* staatlicher Entscheidungen leisten, die mit KI-Tools unterstützt wurden. Dabei geht es um *Fairness* im Sinne von Verhinderung von Diskriminierung, aber auch von akzeptablen und nachvollziehbaren Entscheidungen, in denen Besonderheiten des Einzelfalls Berücksichtigung finden, ähnlich wie bei der menschlichen Ausübung von Verwaltungsermessen.

Wie dargelegt, können rechtliche Vorschriften allein kaum ein hohes Maß an *Fairness*, *Transparenz* und Akzeptanz neuer Technologien erzwingen. Im EU-Datenschutzrecht wurde in diesem Zusammenhang beispielsweise der Ansatz von *privacy by design and by default* (vgl. Art. 25 DSGVO) gewählt, um die Lücke zwischen den rechtlichen Anforderungen und deren praktischer Umsetzung zu schließen. Die Einhaltung von Gesetzen und Rechtsgrundsätzen sollte nicht von den Nutzenden eines technischen Geräts abhängen, sondern die Technologie ausschließlich so gestaltet sein, dass sie idealerweise nur auf legale Weise genutzt werden kann (vgl. Fährmann/Aden/Arzt 2022: 318ff.). Dieser Ansatz kann auch auf andere neue Technologien jenseits der Datenverarbeitung angewandt werden. Ein genereller *legality-by-design*-Ansatz (vgl. Hildebrandt 2020) kann auch für die Entwicklung und Nutzung von KI-Systemen relevant sein. Hierbei gilt es jedoch in einem ersten Schritt zu identifizieren, an welcher Stelle der Entwicklung dies umgesetzt und integriert werden kann, um in einem nächsten Schritt die konkrete Ausgestaltung zu erarbeiten. Während bei der Entwicklung von Modellen hier große Schwierigkeiten der Implementierung eines solchen Ansatzes gesehen werden können, ist insbesondere der im Kompromissvorschlag des Europäischen Parlaments zum KI-VO-E neu eingeführte Adressat*innenkreis der *Deployer* von KI Systemen auch für diese Überlegungen interessant. Hier besteht demnach weiterer Forschungsbedarf.

Letztlich kann davon ausgegangen werden, dass *Transparenz*, *Erklärbarkeit* und *Fairness* bei der KI-Nutzung auch im Sicherheitsbereich keine gänzlich unerreichbaren Ziele sind. Allerdings bleibt dabei die Schnittstelle zur Technik und die Implementation dieser Ziele im technischen Design die größte Herausforderung – sowohl wegen des rasanten Tempos der KI-Entwicklung als auch wegen der Notwendigkeit, Entwickler*innen und Anwender*innen von der hohen Priorität dieser Ziele zu überzeugen. Im Zusammenhang mit KI-Systemen treten zwar Spannungen zwischen rechtlicher Normierbarkeit von *Fairness*, *Transparenz* und *Erklärbarkeit* und den Konkretisierungsbedarfen durch die Hard- und Softwareentwicklung und letztlich deren technischer Umsetzung auf, allerdings kann hier ebenfalls eine integrierte und interdisziplinäre Technikentwicklung sowie Partizipationsoffenheit zu grundrechtsorientierten Lösungen beitragen.

Steven Kleemann *ist Jurist, wissenschaftlicher Mitarbeiter am Forschungsinstitut für Öffentliche und Private Sicherheit (FÖPS Berlin) der Hochschule für Wirtschaft und Recht Berlin (HWR) und Doktorand an der Universität Potsdam.*

Dr. Simon david hirsbrunner *leitet am Institut für Ethik in den Wissenschaften (IZEW) der Universität Tübingen Projekte zu den Themen KI-Ethik, Datenethik sowie algorithmische Polizeiarbeit. Er ist Sozial- und Medienwissenschaftler und forschte bisher im Bereich Human-Centered Computing an der Freien Universität Berlin, sowie bei der Wikimedia, am Potsdam Institut für Klimafolgenforschung, der Universität Potsdam und der Universität Siegen.*

Prof. Dr. Hartmut aden *ist Jurist und Politikwissenschaftler. Er ist Professor für Öffentliches Recht, Europarecht, Politik- und Verwaltungswissenschaft an der Hochschule für Wirtschaft und Recht Berlin (HWR), zugleich Vizepräsident für Forschung (seit 2020) und Mitglied des Forschungsinstituts für Öffentliche und Private Sicherheit (FÖPS Berlin). Zudem ist er Mitglied der Redaktion der* ***vor****gänge. Website: www.hwr-berlin.de/prof/hartmut-aden.*

Literatur[iv]

Aden, Hartmut/Kleemann, Steven 2023: Die Verantwortlichkeit für die Nutzung Künstlicher Intelligenz im Sicherheitsbereich – Regelungsansätze und Problemfelder des KI-Verordnungsentwurfs der EU-Kommission, in: Pfeffer, Kristin (Hrsg.): *Smart Big Data Policing – Chancen, Risiken und regulative Herausforderungen*, Göttingen, S. 51-64.

Aden, Hartmut et al. 2023: Faire globale Daten-Governance im Sicherheitsbereich? Risiken bei der internationalen Zusammenarbeit von Sicherheitsbehörden und eine mögliche Rolle der Europäischen Union. In: Friedewald, Michael et al. (Hrsg.): *Daten-Fairness in einer globalisierten Welt, Baden-Baden*, S. 285-315; doi.org/10.5771/9783748938743.

Ananny, Mike/Crawford, Kate 2018: Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. In: *New Media & Society* , Jg. 20, H.3, S. 973–989.

Barocas, Solon/Hardt, Moritz/Narayanan, Arvind 2019: *Fairness* and machine learning. Limitations and Opportunities, fairmlbook, https://fairmlbook.org/.

Baumer, Eric PS 2017: Toward human-centered algorithm design, in: *Big Data & Society*, Jg. 4, H. 2, S. 1-12.

Binns, Reuben et al. 2018: It's Reducing a Human Being to a Percentage: Perceptions of Justice in Algorithmic Decisions, in: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18, Association for Computing Machinery, New York, NY, USA, Paper 377), S. 1–14. https://doi.org/10.1145/3173574.3173951.

Brandner, Lou Therese/Hirsbrunner, Simon David 2023: Algorithmische Fairness in der Polizeilichen Ermittlungsarbeit: Ethische Analyse von Verfahren des Maschinellen Lernens zur Gesichtserkennung, in: *TATuP - Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis*, Jg. 32, H.1, S. 24-29. https://doi.org/10.14512/tatup.32.1.24.

Busuioc, Madalina/Curtin, Deirdre/Almada, Marco 2023: Reclaiming transparency: contesting the logics of secrecy within the AI Act, in: *European Law Open*, Jg. 2, H. 1, S. 79-105. https://doi.org/10.1017/elo.2022.47

Coeckelbergh, Mark 2020: AI Ethics, Cambridge.

Ebers, Martin 2021: Regulating Explainable AI in the European Union. An Overview of the Current Legal Framework(s), in: Liane Colonna/Stanley Greenstein (Hrsg.): *Nordic Yearbook of Law and Informatics 2020: Law in the Era of Artificial Intelligence*, Stockholm, S. 103-132.

Ebers, Martin et al. 2021: Der Entwurf für eine EU-KI-Verordnung: Richtige Richtung mit Optimierungsbedarf, in: *Recht Digital* (RDi), S. 528-537.

Europarat, Committee on Artificial Intelligence (CAI) 2023: Consolidated Working Draft of the Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law (KI-Konvention), CAI(2023)18.

European Parliament 2023: European Parliament, Draft Compromise Amendments on the Draft Report, Proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9 0146/2021 – 2021/0106(COD))', KMB/DA/AS, 11 May 2023, https://www.europarl.europa.eu/resources/library/media/20230516RES90302/20230516RES90302.pdf.

Europäische Kommission 2019: Generaldirektion Kommunikationsnetze, Inhalte und Technologien, (2019). Ethik-Leitlinien für eine vertrauenswürdige KI, Publications Office.

Europäische Kommission 2021: Entwurf einer Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz (Gesetz über künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union, COM(2021) 206 final, 2021/0106(COD).

Fährmann, Jan/Aden, Hartmut/Arzt, Clemens 2022: Transparenz der polizeilichen Datenverarbeitung: Defizite und technische Lösungsansätze, in: Friedewald Michael/Kreutzer, Michael/Hansen, Marit (Hrsg.): *Selbstbestimmung, Privatheit und Datenschutz - Gestaltungsoptionen für einen europäischen Weg*, Wiesbaden, S. 303-326. https://doi.org/10.1007/978-3-658-33306-5_15.

Fischer Caroline/Hirsbrunner Simon David/Teckentrup Vanessa 2022: Producing Open Data, in: *Research Ideas and Outcomes*, Jg. 8, S. e86384, https://doi.org/10.3897/rio.8.e86384.

Gebru, Timnit et al. 2021: Datasheets for datasets, in: *Communications of the ACM*, Jg. 64, H. 12, S. 86-92.

Gillies, Marco et al. 2016: Human-Centred Machine Learning, in: *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (CHI EA '16), New York, S. 3558–3565. https://doi.org/10.1145/2851581.2856492.

González Fuster, Gloria 2020: Artificial Intelligence and Law Enforcement: Impact on Fundamental Rights. Study requested by the LIBE Committee of the European Parliament PE 656.295.

Gressel, Céline/Orlowski, Alexander 2019: Integrierte Technikentwicklung: Herausforderungen, Umsetzungsweisen und Zukunftsimpulse, in: *TATuP – Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis*, Jg. 28, H. 2, S. 71–72.

Hildebrandt, Mireille 2020: 'Legal by Design' or 'Legal Protection by Design'?, Law for Computer Scientists and Other Folk, in: *Oxford online edition*, Oxford Academic, 23 July 2020.

Kusner, Matt J. et al. 2017: Counterfactual Fairness, in: *Advances in Neural Information Processing Systems, Bd. 30.*, Curran Associates, Inc., 2017. https://proceedings.neurips.cc/paper/2017/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html.

Laux, Johann/Wachter, Sandra/Mittelstadt, Brent 2023: Three Pathways for Standardisation and Ethical Disclosure by Default under the European Union Artificial Intelligence Act (February 20, 2023), http://dx.doi.org/10.2139/ssrn.4365079.

Liao, Q. Vera/Gruen, Daniel/Miller, Sarah 2020: Questioning the AI: Informing Design Practices for Explainable AI User Experiences, in: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (CHI '20), New York, S. 1–15. https://doi.org/10.1145/3313831.3376590.

Linardatos, Dimitrios 2022: Auf dem Weg zu einer europäischen KI-Verordnung – ein (kritischer) Blick auf den aktuellen Kommissionsentwurf, in: *Zeitschrift für das Privatrecht der Europäischen Union (GPR)*, Jg. 19, H. 2, S. 58-70.

Mehrabi, Ninareh et al. 2021: A Survey on Bias and Fairness in Machine Learning, in: *ACM Computing Surveys (CSUR)*, 54 (6), Article 115, S. 1-35.

Microsoft 2022: Responsible AI Impact Assessment Template, https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-RAI-Impact-Assessment-Template.pdf.

Mitchell, Margaret et al. 2019: Model Cards for Model Reporting, in: Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19), New York, S. 220–229. https://doi.org/10.1145/3287560.3287596.

Mittelstadt, Brent/Wachter, Sandra/Russell, Chris 2023: The Unfairness of Fair Machine Learning: Levelling down and strict egalitarianism by default (January 20, 2023), in: *Michigan Technology Law Review* (forthcoming).

Palantir o. D.: Palantir Technologies' Approach to AI Ethics, https://www.palantir.com/pcl/palantir-ai-ethics/.

Qiu, Shilin/Qihe Liu/Shijie Zhou/Chunjiang Wu 2019: Review of Artificial Intelligence Adversarial Attack and Defense Technologies, in: *Applied Sciences*, Jg. 5, H. 9, 909. https://doi.org/10.3390/app9050909.

Sultanow, Eldar/Bauckhage, Christian/ Knopf, Christian/Piatkowski, Nico 2022: Sicherheit von Quantum Machine Learning, in: *Wirtschaftsinformatik & Management*, Jg. 14, H. 2, S. 144-152.

Turilli, Matteo/Floridi, Luciano 2009: The ethics of information transparency, in: E*thics and Information Technology*, Jg. 1, H. 2, S. 105–112, https://doi.org/10.1007/s10676-009-9187-9.

Wachter, Sandra/Mittelstadt, Brent/Russell, Chris 2021: Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI (March 3, 2020), in: *Computer Law & Security Review*, Jg. 41, 105567, https://dx.doi.org/10.2139/ssrn.3547922.

Anmerkungen:

i Dieser Beitrag basiert auf Erkenntnissen aus dem Verbundforschungsprojekt VIKING (Vertrauenswürdige Künstliche Intelligenz für polizeiliche Anwendungen, 2021 bis 2024/25), gefördert vom Bundesministerium für Bildung und Forschung (BMBF, Förder-Kennzeichen 13N16242 (HWR/FÖPS Berlin) bzw. 13N16243 (IZEW Univ. Tübingen).

ii Bspw.: Antirassismusrichtlinie 2000/43/EG; Gleichbehandlungsrichtlinie 2006/54/EG; Vierte Gleichstellungsrichtlinie 2004/113/EG; Gleichbehandlungsrahmenrichtlinie 2000/78/EG.

iii Vgl. dazu bspw. Europäische Kommission (2019), Ethik-Leitlinien für eine vertrauenswürdige KI, S. 22, wo *Transparenz* als „eng mit dem Grundsatz der Erklärbarkeit verbunden“ beschrieben wird oder in der KI-Konvention auf S. 4 anerkannt wird, dass es notwendig ist „to promote transparency, explainability, […] in the design, development, use and decommissioning of artificial intelligence systems“, genauso wie der KI-VO-E in Art. 13 vorsieht, dass „Hochrisiko-KI-Systeme […] so konzipiert und entwickelt [werden], dass ihr Betrieb hinreichend transparent ist, damit die Nutzer die Ergebnisse des Systems angemessen interpretieren und verwenden können“.

[iv] Alle Online-Links sind auf dem Stand Oktober 2023.
