KI-Regulierung durch Zweckbindung für Modelle

*Rainer Mühlhoff und Hannah Ruschemeier**

In diesem Artikel stellen wir das Konzept der Zweckbindung für KI-Modelle als neuen Vorschlag zur effektiven Regulierung von KI vor. Die unregulierte (sekundäre) Nutzung bestimmter Modelle birgt immense individuelle und gesellschaftliche Risiken: Diskriminierung von Einzelpersonen oder Gruppen, Verletzung von Grundrechten oder Gefährdung demokratischer Prozesse durch Fehlinformationen. Wir argumentieren, dass der Besitz von trainierten KI-Modellen, die in vielen Fällen aus anonymen Daten bestehen und daher nicht dem Geltungsbereich der DSGVO unterliegen, der Kern einer zunehmenden informationellen Machtasymmetrie zwischen großen Datenunternehmen und der Gesellschaft bilden. Durch die Kombination von ethischen und rechtlichen Argumenten in unserem interdisziplinären Ansatz identifizieren wir das trainierte Modell und nicht die Trainingsdaten als Regulierungsgegenstand für eine Aktualisierung des Zweckbindungsprinzips auf KI-Modelle. Dieser geänderte Fokus ergänzt die bestehenden Datenschutzgesetze, insbesondere die DSGVO, und die KI-Verordnung. Die bestehenden Regelungen sind nicht ausreichend, um den Missbrauch von trainierten Modellen zu verhindern, da sie sich auf die verfahrenstechnischen Aspekte der Verarbeitung von personenbezogenen Daten bzw. Trainingsdaten konzentrieren. Ausgehend von den Wertungen des Risikopräventionsrechts und dem Grundsatz der Verhältnismäßigkeit argumentieren wir, dass bereits die potenzielle Verwendung trainierter Modelle durch mächtige Akteure in einer sozialschädlichen Weise präventive regulatorische Eingriffe rechtfertigt. Zweckbindung für KI-Modelle zielt darauf, eine demokratische Kontrolle darüber zu ermöglichen, wo und wie prädiktive und generative KI-Modelle genutzt und wiederverwendet werden dürfen.

Schlüsselwörter: KI-VO; DSGVO; Zweckbindung; Regulierungsmodelle; Datenmacht; Sekundärnutzung

I. Einführung

Die Technologie der künstlichen Intelligenz (KI) spielt heute in zahlreichen Anwendungsfeldern eine wichtige Rolle. Die meisten gesellschaftlich relevanten Anwendungsfälle von KI beruhen auf Techniken des maschinellen Lernens (ML). Dabei handelt es sich um Algorithmen, die auf der Grundlage großer Datenmengen konfiguriert („trainiert“) werden, um „Muster“ zu erkennen; anschließend können sie verwendet werden, um ebendiese Muster in den Eingabedaten zu detektieren, was die Grundlage für ihre Ausgabe bildet. Dieses Prinzip liegt den verschiedenen Anwendungsbereichen von ML-Technologie gleichermaßen zugrunde: Im Bereich der prädiktiven Analytik wird die Mustererkennung zum Beispiel mit Risikobewertungen oder Vorhersagen über unbekannte Attribute wie gesundheitliche Risiken, sexuelle Orientierung, religiöse und ethnische Zugehörigkeit, politische Ansichten, Bildungshintergrund und finanzieller Status in Verbindung gebracht.[1] Beim Einsatz von ML-Modelle zur

*Rainer Mühlhoff, Professor für Ethik und Künstliche Intelligenz an der Universität Osnabrück. Hannah Ruschemeier, Junior Professorin (tenure W3) für öffentliches Recht, Datenschutzrecht und Recht der Digitalisierung an der Fernuniversität in Hagen. Kontakt: rainer.muehlhoff@uni-osnabrueck.de und

Verarbeitung natürlicher Sprache oder zur Bilderkennung ist die Mustererkennung mit Textproduktion verknüpft, z.B. wenn eine Transkription für ein erkanntes Wort in Audiodaten oder eine Kennzeichnung für ein identifiziertes Objekt in einem Bild erstellt wird. Bei generativen KI-Modellen wird Mustererkennung mit der Fortzeichnung dieser Muster kombiniert, z.B. erweitert ChatGPT den Eingabe-Promptum die wahrscheinlichste Antwort als Ausgabe.

All diesen vielfältigen Anwendungen des maschinellen Lernens haben zwei strukturelle Aspekte gemeinsam: Erstens sind sie auf große Mengen an Trainingsdaten angewiesen, die oftmals von tausenden bis Millionen von Personen und aus verschiedenen Quellen, darunter soziale Medien, Foren, Webseiten und -blogs, sowie alle weiteren digital zugänglichen Text- und Bildrepositorien, gewonnen werden. Zweitens können die Anbieter die trainierten Modelle häufig ohne wesentliche rechtliche Hindernisse für zahlreiche sekundäre Zwecke wiederverwenden, einschließlich riskanter und missbräuchlicher Zwecke, die von personalisierten Preisen für Angebote und Dienstleistungen bis hin zur Verbreitung von Falschinformationen reichen können. In vielen Fällen besteht ein trainiertes Modell aus einer Reihe von aggregierten, anonymen und daher nicht personenbezogenen Daten, die damit nicht unter datenschutzrechtliche Vorschriften fallen und frei zirkuliert, verkauft und veröffentlicht werden können.

Die zentrale These dieses Beitrags lautet, dass die unzureichende Regulierung trainierter Modelle eine ernsthafte Bedrohung für Individuen und die Gesellschaft darstellt und daher dringend regulatorische Maßnahmen erfordert. Trainierte ML-Modelle sind mächtige Werkzeuge, da sie für automatisierte Entscheidungen, Verhaltensbewertungen oder diskriminierende Geschäfts- und Verwaltungspraktiken eingesetzt oder wiederverwendet werden können. Zum Beispiel können Modellen, die psychologische Charaktereigenschaften vorhersagen können, für personalisierte politische Werbung zweitverwendet werden; Modelle, die in der Lage sind, die Prävalenz von Erkrankungen auf der Grundlage von Social-Media-Daten abzuschätzen, könnten in der Versicherungsbranche wiederverwendet werden; oder medizinische Modelle, die Drogenmissbrauch oder psychologische Veranlagungen wie Depressionen vorhersagen können, könnten für KI-gestützte Einstellungsverfahren missbraucht werden.[2] Auch das Missbrauchsrisiko generativer KI-Modelle ist erheblich: Es reicht von der Erstellung gefälschter persönlicher Äußerungen, die in Persönlichkeitsrechte eingreifen, über die Verbreitung synthetischer Fotos oder Videos, den vielfachen Verstoß gegen Urheberrechte, bis hin zur Erzeugung von Hassrede und Fake News in Form von Bildern oder Texten.[3]

In diesem Beitrag stellen wir das Prinzip der *Zweckbindung für KI-Modelle* als Konzept für einen neuen regulatorischen Ansatz dar, der den Risiken einer missbräuchlichen Zweitverwendung trainierter Modelle begegnet. Zweckbindung für Modelle sieht vor, dass die Erstellung und Nutzung von KI-Modellen auf bestimmte Zwecke beschränkt sein muss. Insbesondere müssen die Zwecke des Modells mit den Zwecken kompatibel sein, für die die Trainingsdaten erhoben wurden. Die Zwecke des Modells müssen dabei *ex ante* angegeben und während des gesamten Lebenszyklus eines KI-Modells

---

hannah.ruschemeier@fernuni-hagen.de. Alle Internetlinks wurden zuletzt am 04. Oktober 2024 aufgerufen. Dieser Beitrag ist eine aktualisierte, Deutsche Version des papers: „Regulating AI with Purpose Limitation for Models“, erschienen in: Journal of AI Law and Regulation (AIReg), 2024, S. 24-39.

[1] Siehe zB *Lammerant/de Hert*, in: *van der Sloot/Broeders/Schrijvers (Hrsg.),* Exploring the boundaries of big data, 2016; *Mühlhoff*, Predictive Privacy: Collective Data Protection in Times of AI and Big Data, Big Data Soc. 2023, 1; *Hildebrandt/Gutwirth*, Profiling the European Citizen: Cross-Disciplinary Perspectives 2008; *Wachter/Mittelstadt*, A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI, Columbia Bus. Law Rev. 2019, 1.

[2] Beispielhaft zu missbräuchlicher Sekundärnutzung: *Mühlhoff*, in: *Ruschemeier/Steinrötter (Hrsg.),* Der Einsatz von KI & Robotik in der Medizin, 2024. Zu Cambridge Analytica: *Hern*, Cambridge Analytica: how did it turn clicks into votes? https://www.theguardian.com/news/2018/may/06/cambridge-analytica-how-turn-clicks-into-votes-christopher-wylie.

[3] *Hacker/Engel/Mauer*, Regulating ChatGPT and other Large Generative AI Models, 2023.

dokumentiert und eingehalten werden. Analog zur Zweckbindung bei der Verarbeitung von personenbezogenen Daten, die aus dem Datenschutz bekannt ist, strebt Zweckbindung für Modelle einen Zustand an, in dem trainierte Modelle nicht ohne Einbeziehung von Kontroll- und Aufsichtsmechanismen für Zwecke (zweit-)verwendet werden dürfen, die mit den primären Zwecke inkompatibel sind.

Die detaillierte Ausarbeitung einer grundsätzlichen Positivliste zulässiger Zwecke und geeigneter ethischer Grundsätze zur Validierung konkreter Zwecksetzungen sowie eventueller, ex post anvisierter Zweckerweiterungen ist Gegenstand einer gesonderten Arbeit. In diesem Artikel geht es uns zunächst um zwei Punkte: (1) den theoretischen Rahmen eines Regulierungsansatzes vorzustellen, der das Konzept der Zweckbindung nicht für Trainingsdaten, sondern für trainierte Modelle neu formuliert. (2) darauf hinzuweisen, dass das im europäischen Datenschutzrecht[hier Fußnote 4] verbürgte Zweckbindungsprinzip aktuell für trainierte ML-Modelle in der Regel nicht greift.[4] Im Gegensatz zur Zweckbindung bei der Verarbeitung personenbezogener Daten können im Fall der Zweckbindung für Modelle nicht die Personen, deren Daten in den Trainingsdaten enthalten sind, diejenigen sein, die durch die Zweckbindung für Modelle ermächtigt werden. Vielmehr muss die Stelle, die befugt ist, Entscheidungen über gültige Zwecke zu treffen, ein Akteur sein, der kollektive Auswirkungen und kollektive Interessen berücksichtigen kann, wie z.B. ein Aufsichtsgremium unter demokratischer Kontrolle.

Es gibt derzeit keine spezifischen gesetzlichen inhaltlichen Vorgaben, für welche Zwecke KI-Modelle überhaupt erstellt und verwendet werden dürfen – mit Ausnahme der verbotenen Praktiken des Art. 5 KI-VO. Die Entscheidung, ob, wie und von wem KI-Modelle entwickelt und genutzt werden, liegt ganz überwiegend in den Händen einiger weniger global agierender Wirtschaftsakteure – der Big-Tech-Unternehmen. Zugleich betreffen die Auswirkungen der Art und Weise, wie bestimmte trainierte ML-Modelle verwendet und wiederverwendet werden, in der Regel eine große Anzahl von Menschen oder die Gesamtgesellschaft (z.B. durch strukturelle Diskriminierungen). Dieses Fehlen einer wirksamen Regulierung der Zweckebene von KI führt zu einer Machtverschiebung zugunsten privater KI-Unternehmen. Bei ihnen akkumuliert sich die Fähigkeit, prädiktive und generative KI-Modelle zu trainieren und zu nutzen als eine neue Manifestation von *informationeller Machtasymmetrie* zwischen datenverarbeitenden Unternehmen, betroffenen Personen und Gesellschaften. Zweckbindung für Modelle ist ein Regelungsvorschlag, der direkt auf die transformativen Auswirkungen von KI auf Machtverhältnisse abzielt.[5] Hier ist aus unserer Sicht staatliche *Regulierung* erforderlich, denn die informationellen Machtasymmetrien durch den Besitz von KI-Modellen können nur durch einen Mechanismus ausgeglichen werden, der von den wirtschaftlichen Interessen der Akteure unabhängig ist.[6]

Die Tatsache, dass trainierte KI-Modelle legal für sekundäre Zwecke wiederverwendet werden können, ohne dass dies öffentlich nachvollziehbar wäre, trägt besonders zur unkontrollierten Multiplikation und Vervielfältigung informationeller Machtasymmetrien bei. Wir argumentieren daher für einen regulatorischen Ansatz, der sich nicht nur auf die *beabsichtigte* Erstverwendung eines KI-Modells bezieht, sondern zugleich auf absehbare und zum Zeitpunkt der Modellerstellung unabsehbare Möglichkeiten der Zweitverwendung. Dazu ist eine Berücksichtigung der rechtlichen, wirtschaftlichen und gesellschaftlichen Stellung der Akteure, der potenziell denkbaren Verbreitungswege trainierter Modelle, der potenziell betroffenen Rechtsgüter und der weiteren kollektiven Auswirkungen dieser Modelle über ihren ursprünglichen Anwendungskontext hinaus. Der

[4] Dazu zB *Koning*, The purpose and limitations of purpose limitation, 2020.

[5] *Kalluri*, Don't ask if artificial intelligence is good or fair, ask how it shifts power, Nature 2020, 169.

[6] Die Situation verbessert sich nicht, wenn staatliche Akteure Vorhersagemodelle verwenden, da sie häufig mit privaten Akteuren zusammenarbeiten und die Beziehung zwischen Staat und Bürgern die Machtasymmetrie gleichermaßen erhöht.

Regulierungsansatz einer Zweckbindung für Modelle zielt daher darauf ab, eine fortlaufende demokratische Aufsicht und Kontrolle auf der Ebene der Herstellung, Verwendung und Wiederverwendung trainierter Modelle zu implementieren.

Die Herstellung von KI-Modellen ist aus ethischer, politischer und rechtlicher Sicht kein neutraler Schritt. Vielmehr impliziert der Besitz eines trainierten Modells eine starke Form von Daten- und Informationsmacht. Zugleich schützt jedoch der bestehende rechtliche Rahmen der DSGVO nicht effektiv vor dem Risiko eines Transfers trainierter Modelle in sekundäre Anwendungskontexte. Um dies zu erkennen, unterscheiden wir analytisch zwischen (1) der *Herstellung* eines Modells („Training"), (2) der *Weiterverarbeitung* trainierter Modelle (speichern, kopieren, transferieren) und (3) der *Anwendung* eines Modells auf eine bestimmte Person oder einen bestimmten Fall (siehe Abschnitt II.1). Da ein Modell im Allgemeinen zwar aus personenbezogenen Daten trainiert wird, seine internen Parameter jedoch in vielen relevanten Fällen *keine* personenbezogenen Daten darstellen, fallen im Allgemeinen nur Schritte 1 und 3, nicht jedoch Schritt 2 – die Verarbeitung solcher Modelle etwa zum Transfer in andere Anwendungskontexte – in den Geltungsbereich der DSGVO. An der für die Multiplikation von informationeller Macht entscheidenden Stelle besteht hier also eine regulatorische Lücke; um diese zu schließen, ist es unser Vorschlag, trainierten Modellen selbst regulatorische Aufmerksamkeit zu widmen. Denn allein die Existenz eines trainierten Modells, das frei (und legal) zirkuliert und umfunktioniert werden kann, birgt erhebliche gesellschaftliche Risiken. Nach geltendem Recht können weder die betroffenen Personen, deren Daten als Trainingsdaten verwendet werden, noch die Gesellschaft als Ganzes wirksam kontrollieren, in welche Anwendungskontexte die KI-Modelle, die aus ihren Daten erstellt werden, transferiert werden. Wie wir darlegen werden, wäre es ohnehin nicht zweckdienlich, die einzelnen betroffenen Personen mit Kontrollinstrumenten auszustatten, da nur aggregierte Datensammlungen, nicht einzelne Datenpunkte, das Training von ML-Modellen ermöglichen. In dieser inhärent kollektiven Struktur, die maschinelles Lernen ermöglicht, bedarf es kollektiver Kontroll- und Regulierungsmechanismen.

In Abschnitt II legen wir dar, warum und auf welche Weise die unkontrollierte Herstellung und Verarbeitung von trainierten Modellen ein Risiko für die Gesellschaft, Gemeinschaftsinteressen und Grundrechte von Individuen darstellt. Wir werden auf das Konzept der Risikoprävention im Recht zurückgreifen, um unsere Argumente für eine bessere Regulierung normativ zu untermauern. In Abschnitt III führen wir das Konzept der Zweckbindung für Modelle ein. Anschließend an die ethisch motivierte Argumentation erörtern wir, warum die derzeitigen Datenschutzvorgaben nicht ausreichen, um trainierte Modelle hinreichend zu kontrollieren. In Abschnitt IV skizzieren wir die ersten Schritte hin zu einem Regelungsvorschlag zur Zweckbindung von Modellen, der die kollektive Struktur von datenbasierten ML-Modellen berücksichtigt.

II. Was ist das Problem?

1. Datenverarbeitungskette

Der Regulierungsvorschlag einer Zweckbindung für Modelle steht in engem Zusammenhang mit dem Datenverarbeitungszyklus bei der Herstellung und Verwendung von ML-Systemen. Um genau zu bestimmen, wo und wie Zweckbindungsvorgaben greifen sollen, unterscheiden wir drei Schritte in der typischen Datenverarbeitungskette solcher Systeme: (1) Sammlung von Trainingsdaten und Training des Modells, (2) Speicherung, Transfer, Weiterverarbeitung des trainierten Modells und (3) Modellanwendung.

(1) Sammlung von Trainingsdaten, Training: Im ersten Schritt der Erstellung eines ML-Systems werden viele Datenpunkte als Trainingsdaten gesammelt.[7] Diese Datenpunkte können personenbezogene oder nicht-personenbezogene Daten enthalten. In einigen Fällen werden extrem große Datensätze für diesen Schritt verwendet, die durch Scraping oder Web-Crawling extrahiert werden – ChatGPT zum Beispiel wurde mit riesigen Datenmengen trainiert, die aus dem Internet gescrapt wurden.[8] Die schiere Quantität der Trainingsdaten macht es bei den meisten ML-Projekten schwierig, wenn nicht gar unmöglich, zwischen verschiedenen Datenkategorien zu unterscheiden.

Nach der Datenerhebung und Aufbereitung wird ein ML-Modell trainiert. Bei dem Trainingsverfahren handelt es sich um einen Algorithmus, der darauf abzielt, Korrelationen, Muster oder andere Regularitäten in den Trainingsdaten zu detektieren, was zu einem konfigurierten („trainierten“) Modell führt.[9] Ein solches Modell kann ein trainiertes Künstliches Neuronales Netz (KNN) oder eine andere Implementierung maschinell lernender Algorithmen sein.[10] Bei Sammlung der Trainingsdaten und dem Training des Modells handelt es sich um unterschiedliche Datenverarbeitungsvorgänge nach der DSGVO. Für die Zweckbindung von Modellen ist aber primär der zweite, folgende Schritt der Modellverarbeitung relevant:

(2) Modellverarbeitung: Im zweiten Schritt der typischen Datenverarbeitungskette wird das trainierte Modell weiterverarbeitet. Handelt es sich bei dem ML-Verfahren um ein künstliches neuronales Netz, dann besteht das trainierte Modell beispielsweise aus einer Zahlenmatrix der internen Parameter, die durch die Gewichte und andere Parameter (z.B. Aktivierungsschwellen) der „Neuronen“ und ihrer Verbindungen bestimmt wird.[11] Wir bezeichnen diese Daten fortan als *Modelldaten*. Die Modelldaten stellen den trainierten Zustand des Modells dar. Für den vorliegenden Kontext ist entscheidend, dass das trainierte Modell im Rahmen der Weiterverarbeitung gespeichert, transferiert oder mit größeren ML-Modellen kombiniert werden kann, ohne dass die Trainingsdaten dafür benötigt werden. In vielen relevanten Fällen kann im Rahmen des Trainings eines ML-System eine Anonymisierung gewährleistet werden. Das heißt, dass die Modelldaten keine personenbezogenen Daten mehr enthalten, auch wenn die Trainingsdaten aus personenbezogenen Daten bestehen. In diesen Fällen unterliegt Schritt 2 nicht dem Anwendungsbereich der DSGVO.

(3) Modellanwendung: Im dritten Schritt der typischen Datenverarbeitungskette wird das trainierte ML-Modell zum Beispiel für Prognosen zu bestimmten Sachverhalten oder Personen angewendet. Dadurch wird das Modell als „Black Box“ verwendet, welches als Reaktion auf bestimmte, in diesem Schritt neu hergebrachte Eingabedaten eine bestimmte Ausgabe berechnet. Die Daten, die in diesem Verarbeitungsschritt berechnet werden (Modellausgabe) stellen daher Informationen über neue Umstände oder dritte Personen dar. Damit sind Sachverhalte, Informationen und Personen gemeint, die nicht zu den Trainingsdaten gehören, die in Schritt 1 zur Herstellung des Modells verwendet wurden.[12] Bei generativen KI-Systemen bspw. wird dem Modell eine Eingabeaufforderung („prompt“) übermittelt, die als Ausgabe z.B. einen Text oder ein Bild erzeugt. Bei

[7] Vgl. *Murphy*, Machine learning, 2012; *O'Neil*, Weapons of math destruction, 2016.

[8] *Brown/Mann/Ryder/Subbiah/Kaplan/Dhariwal/Neelakantan/Shyam/Sastry/Askell/Agarwal/Herbert-Voss/Krueger/Henighan/Child/Ramesh/Ziegler/Wu/Winter/Hesse/Chen/Sigler/Litwin/Gray/Chess/Clark/Berner/McCandlish/Radford/Sutskever/Amodei*, Language models are few-shot learners, 2020.

[9] Die Verwendung von Begriffen wie "Training" und "Lernen" wurde als Anthropomorphisierung von KI-Systemen kritisiert. Die Formulierung, dass ein Modell "konfiguriert" statt "trainiert" wird, umgeht diesen Fallstrick, hat aber den Nachteil, dass diese Terminologie weniger populär ist. Siehe *Rehak*, in: *Verdegem (Hrsg.),* AI for Everyone? Critical Perspectives, 2021.

[10] Zu den verschiedenen Aufgaben und unterschiedlichen Arten von Algorithmen siehe *Goodfellow/Bengio/Courville*, Deep learning, 2016.

[11] *Goodfellow/Bengio/Courville*, [Fn. 10].

[12] *Mühlhoff*, Predictive privacy: towards an applied ethics of data analytics, Ethics Inf. Technol. 2021, 675; *Mühlhoff/Ruschemeier*, Predictive analytics and the collective dimensions of data protection, Law Innov. Technol. 2024, 1.

Klassifikationsmodellen oder prädiktiven Modellen werden verfügbare sachverhalts- oder personenbezogene Daten als Input für das Modell verwendet. Dies kann bspw. der Lebenslauf einer Bewerber:in sein, der in ein Modell eingegeben wird, welches die Vorauswahl im Rahmen eines Bewerbungsverfahrens trifft, oder die Social-Media-Daten eines Instagram-Nutzers, die in ein Modell eingegeben werden, das den aktuellen emotionalen Zustand des Nutzers vorhersagen soll, um personalisierte Inhalte (Werbung) zu unterbreiten. Die Anwendung des Modells muss nicht unmittelbar dem Modelltraining (Schritt 1) folgen. Vielmehr kann das Modell auch wesentlich später oder von anderen Akteuren verwendet werden. Damit ist der Zugriff auf die Trainingsdaten keine Voraussetzung für die Anwendung des Modells.

2. Das Risiko der Sekundärnutzung

Der Vorschlag einer Zweckbindung für Modelle adressiert das Risiko der Wiederverwendung oder Umfunktionierung trainierter Modelle, die für den Einzelnen oder die Gesellschaft schädlich sind. Dieses Risiko entsteht bereits mit Vollendung des ersten Schritts – des Trainings des Modells.[13] Man stelle sich vor, eine Social-Media-Plattform entwickelt ein Modell, das Alkoholkonsum anhand der Verhaltensdaten ihrer Nutzer:innen (z.B. „gelikte" Inhalte und besuchte Websites) vorhersagen kann. Der ursprüngliche Zweck des Modells ist es, relevanten Nutzer:innen verstärkt Informationen zu Suchtpräventionsprogrammen in ihrem Newsfeed anzuzeigen.[14] Die Art möglicher sekundärer Nutzungen, mit der wir uns in diesem Beitrag befassen, bezieht sich auf Fälle, in denen ein solches Modell für andere Zwecke als zu denen es trainiert wurde, wiederverwendet wird, z.B. für algorithmische Systemen im Human Resource Management, etwa zur Auswahl von Job-Bewerber:innen.

Das Risiko einer unkontrollierten und oft schädlichen Zweitverwendung eines trainierten Modells gerät leicht aus dem Blick und entzieht sich öffentlicher Kontrolle, insbesondere wenn die ursprüngliche Zweckausrichtung des Modells wünschenswert, im öffentlichen Interesse oder sonst förderungswürdig ist, z.B. bei medizinischer Forschung. In dem Beispielszenario einer Vorhersage von Alkoholkonsum für das Targeting mit Informationen zu Suchtprävention, mag zumindest der ursprüngliche und öffentlich kommunizierte Primärzweck für die Erstellung des ML-Modells nicht streitig sein. Die Aufmerksamkeit von Öffentlichkeit, Wirtschaft und Politik richtet sich allerdings häufig allein auf das Innovationspotenzial leistungsfähiger Modelle zur Steigerung von Effizienz, Effektivität, Verbesserung von Forschung etc. Die Möglichkeiten einer missbräuchlichen Zweit- und Weiterverwendung von Modellen mit wünschenswerten Primärzwecken ist hingegen ein blinder Fleck in der Debatte. Diese Weiterverwendung gerät auch deshalb leicht aus dem Fokus, da sie auch Jahre später und durch andere Akteure erfolgen kann (z.B. durch verschiedene Unternehmen nach einer Fusion oder Übernahme des ursprünglichen Unternehmens). Während im nächsten Unterabschnitt (II.3) näher erläutert wird, was wir unter einer schädlichen oder missbräuchlichen Verwendung von KI-Modellen verstehen, skizzieren wir im Folgenden, warum ein reales Risiko besteht, dass trainierte Modelle auf potenziell unvorhergesehene sekundäre Anwendungsfälle übertragen werden.

[13] *Mühlhoff*, [Fn. 2].

[14] Dass die Vorhersage von Drogenmissbrauch und vielen anderen psychosozialen, gesundheitsbezogenen Problemen auf der Grundlage von Daten aus sozialen Medien möglich ist, ist hinlänglich bekannt: *Kosinski u. a.*, Private traits and attributes are predictable from digital records of human behavior, Proc. Natl. Acad. Sci. 2013, 5802; *Merchant/Asch/Crutchley/Ungar/Guntuku/Eichstaedt/Hill/Padrez/Smith/Schwartz*, Evaluating the predictability of medical conditions from social media posts, PLOS ONE, 2019, e0215476. Die Risiken der Wiederverwendung von Modellen, die ursprünglich für gezielte Werbezwecke trainiert wurden, werden insb. erörtert in *Mühlhoff/Willem*, Social Media Advertising for Clinical Studies: Ethical and Data Protection Implications of Online Targeting, Big Data Soc. 2023.

Es ist von zentraler Bedeutung für unsere Argumentation, dass das trainierte Modell, welches aus dem ersten Verarbeitungsschritt hervorgeht, eigenständige Daten („Modelldaten", siehe Schritt 2 oben) beinhaltet, die sich von den Trainingsdaten unterscheiden.[15] Um zu beurteilen, welche rechtlichen Vorgaben für die Verarbeitung von Modelldaten bestehen, ist es relevant, ob die Modelldaten personenbezogene Daten enthalten. Dies lässt sich nicht pauschal, sondern rein einzelfallbezogen beurteilen. Handelt es sich bei den Trainingsdaten um personenbezogene Daten, können die Modelldaten je nach Trainingsverfahren entweder personenbezogene oder anonyme Daten sein. Wenn beispielsweise aktuelle Anonymisierungstechniken wie „Differential Privacy" oder „Federated Learning" verwendet werden, ist es theoretisch möglich, dass in Schritt 1 ein Modell erstellt wird, welches keine Rückverweise auf die Trainingsdaten enthält und damit anonymisiert ist.[16] In diesem Fall würden die rechtlichen Beschränkungen für die Verarbeitung personenbezogener Daten der DSGVO nicht für die Modelldaten gelten. Wichtig ist nun, dass das Potenzial für schädliche Weiterverwendungen eines trainierten Modells (siehe II.3) nicht abnimmt, sobald die Modelldaten anonymisiert sind. Vielmehr besteht ein dringender Regulierungsbedarf in Konstellationen in denen die Trainingsdaten personenbezogene Daten enthalten, die Modelldaten aber anonymisiert sind. Denn dann fällt die Weiterverarbeitung des trainierten Modells (Schritt 2) aus dem Anwendungsbereich der DSGVO. [Passt diese Fußnote noch hier hin?][17]

Im dritten Verarbeitungsschritt, der Anwendung des Modells, kommt – je nach Anwendungskontext – wiederum eine Verarbeitung personenbezogener Daten in Betracht. Dies liegt daran, dass in der typischen Beispielsituation die Eingabedaten in der Anwendungsphase solche sind, die mit einer bestimmten Person oder einem bestimmten Sachverhaltverknüpft sind und die Ausgabe schließlich Daten über diese Person oder diesen Sachverhalt darstellt (z.B. eine Vorhersage, eine Klassifizierung oder ein generierter Text oder ein Bild, das sich auf die Person oder den Umstand bezieht). Die Person oder der Sachverhalt, auf den das Modell angewandt wird, muss dabei gerade *nicht* Teil der Trainingsdaten sein, die in Schritt 1 zur Erstellung des Modells verwendet wurden.[18] Zur Veranschaulichung bietet sich das oben eingeführte Beispiel an: Ein Modell zur Vorhersage von Alkoholmissbrauch aus Social-Media-Daten, das auf den (anonymisierten) Daten von Personen 1–1.000 trainiert wurde, kann auf die Verhaltensdaten von Nutzerin Nr. 1.001 angewandt werden, um die Wahrscheinlichkeit von Alkoholmissbrauch dieser speziellen Person vorherzusagen. Zu beachten ist, dass die Datenverarbeitungskette somit zwei verschiedene Arten von Betroffenen enthält: die Datensubjekte des Trainingsdatensatzes (Schritt 1) und die Datensubjekte des Anwendungsschritts (Schritt 3).[19] In Schritt 2, der sich auf die Speicherung, Verwendung und mögliche Wiederverwendung eines trainierten Modells bezieht, gibt es in der Regel keine betroffenen Personen, da die Modelldaten stark aggregiert und in vielen Fällen sogar anonym sind.

Unserer Ansicht nach reichen deshalb die bestehenden Vorgaben des Datenschutzrechts in Bezug auf einzelne betroffene Personen in den Verarbeitungsschritten 1 und 3 nicht aus, um eine schädliche Sekundärnutzung von Modelldaten in Schritt 2 zu verhindern. Modelldaten, die trainierte Modelle darstellen, sind derzeit keinen spezifischen Regelungen unterworfen und unterfallen bei

---

[15] *Khan/Hanna*, The Subjects and Stages of AI Dataset Development: A Framework for Dataset Accountability, Ohio State Technol. Law J. 2023, 171.

[16] *Abadi/Chu/Goodfellow/McMahan/Mironov/Talwar/Zhang*, Deep Learning with Differential Privacy, Proc. 2016 ACM SIGSAC Conf. Comput. Commun. Secur. - CCS16 2016, 308; *Dwork*, in: *Bugliesi/Preneel/Sassone/Wegener (Hrsg.),* Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 2006.

[17] VO (EU) 2016/679 des Europäischen Parlaments und des Rates vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten, zum freien Datenverkehr und zur Aufhebung der RL 95/46/EG (Datenschutz-Grundverordnung), 2016, ABl. L 119/1.

[18] *Mühlhoff*, [Fn. 12], 675.

[19] *Khan/Hanna*, [Fn. 15], 171.

Anonymisierung nicht der DSGVO. Zudem reguliert die DSGVO einzelne Datenverarbeitungen und keine Modelle. Die betroffenen Datensubjekte sind in den Verarbeitungsschritten 1 und 3 nicht deckungsgleich, weshalb Betroffenenrechte nicht effektiv ausgeübt werden können. Modelldaten stellen ein erhebliches Risiko für die Gesellschaft dar, wenn sie uneingeschränkt verarbeitet und verbreitet werden können, z.B. durch Verkauf und Weitergabe. Dieses Risiko begründet sich daraus, dass das trainierte Modell für jeden Zweck und auf jede Person oder jeden Sachverhalt – in der Vergangenheit, gegenwärtig, und in der Zukunft – einzeln oder parallel (durch Massenverarbeitung) in einer Weise angewendet werden kann, die sich einer rechtlichen, demokratischen Kontrolle entzieht. Es ist deshalb aus unserer Sicht erforderlich, in Schritt 2 der Datenverarbeitungskette (Abschnitt III) eine Zweckbindung für Modelle einzuführen, um von vornherein riskante Anwendungen zu verhindern, die sich nachteilig auf Einzelne und die Gesellschaft auswirken.

3. Gesellschaftliche Risiken und Gefahren im Zusammenhang mit KI-Modellen

Das Spektrum individueller und gesellschaftlicher Risiken im Zusammenhang mit der Anwendung von Big Data und maschinellem Lernen ist breit gefächert. Vielfach dokumentiert sind strukturelle Diskriminierungen, die soziale Ungerechtigkeit verstärken,[20] versteckte biases bei finanzieller Risikobemessung, teilautomatisierten Auswahlverfahren oder Entscheidungen über die Vergabe öffentlicher Leistungen,[21] neue Formen der Verletzung der Privatsphäre,[22] kapitalistische und kolonialistische Ausbeutung menschlicher und natürlicher Ressourcen,[23] Bedrohung der Demokratie durch Desinformation,[24] Verletzung von Urheber- und Persönlichkeitsrechten[25] und vieles mehr.

Aufgrund der Ausrichtung dieses Beitrags können wir auf diese verschiedenen Debatten im Detail nicht eingehen. Wir fokussieren uns vorliegend auf zwei (von mehreren) Kategorien des potenziellen Missbrauchs im Zusammenhang mit trainierten ML-Modellen. Die erste Kategorie ist die prädiktive Analytik und betrifft die Verwendung von ML-Modellen zur Vorhersage unbekannter Informationen über Personen oder Sachverhalte. Das in Abschnitt II.2 erwähnte Beispiel der Vorhersage des Alkoholkonsums anhand von Social-Media-Daten fällt in diese Kategorie. Im Allgemeinen hat sich gezeigt, dass verschiedene Suchterkrankungen und weitere Krankheiten, darunter Depressionen, Psychosen, Diabetes und Bluthochdruck, anhand von Social-Media-Daten vorhergesagt werden können.[26] Faktisch sind Versicherungs- und Finanzunternehmen an diesen Vorhersagemodellen interessiert, da sie eine individuelle Risikobewertung mittels Kreditwürdigkeitsprüfung ermöglichen.[27] In diesen Branchen werden Vorhersagemodelle zweifelsohne genutzt, jedoch ist die Herkunft der Daten oft nicht bekannt. In diesen Bereichen sowie in Bereichen wie der Job-Auswahl könnten Vorhersagemodelle zu einer indirekten Diskriminierung

[20] *O'Neil*, [Fn. 7]; *Verdegem*, AI for Everyone?, 2021; *Mühlhoff*, Automatisierte Ungleichheit: Ethik der Künstlichen Intelligenz in der biopolitischen Wende des Digitalen Kapitalismus, Dtsch. Z. Für Philos. 2020, 867.

[21] *Barocas/Selbst*, Big data's disparate impact, Calif Rev 2016, 671; *Eubanks*, Automating inequality, 2017; *Wachter*, The Theory of Artificial Immutability: Protecting Algorithmic Groups under Anti-Discrimination Law, preprint, Tulane Law Rev. 2023, 149.

[22] *Mühlhoff*, [Fn. 12], 675; *Mühlhoff*, [Fn. 2], 1; *Mühlhoff/Ruschemeier*, [Fn. 12], 1; *Zuboff*, The age of surveillance capitalism, 2019.

[23] *Crawford*, Atlas of AI, 2021; *Coleman*, Digital Colonialism: The 21st Century Scramble for Africa through the Extraction and Control of User Data and the Limitations of Data Protection Laws, Mich. J. Race Law 2018, 417; *Sadowski*, Too smart, 2020.

[24] *Tufekci*, Algorithmic harms beyond Facebook and Google: Emergent challenges of computational agency, Colo Tech LJ 2015, 203.

[25] *Sag*, Copyright Safety for Generative AI, 2023.

[26] *Merchant/Asch/Crutchley/Ungar/Guntuku/Eichstaedt/Hill/Padrez/Smith/Schwartz*, [Fn. 14], e0215476; *Mühlhoff/Willem*, [Fn. 14].

[27] *O'Neil*, [Fn. 7], Kap. 8.

aufgrund sensibler Informationen wie ethnischer Herkunft oder Schwangerschaft führen.[28] Ein weiterer großer Anwendungsbereich von prädiktiver Analytik ist personalisierte Werbung (targeted advertising). Hier kann die Nutzung von Echtzeitinformationen über die Vulnerabilitäten und Emotionen der Nutzer:innen zu manipulativen Praktiken führen. Diese Praktiken – die manchmal als „Hypernudges"[29] oder „Dark Patterns" diskutiert werden – kombinieren Vorhersagetechnologien und psychologische Manipulation. Dies wiederum birgt Risiken für die Autonomie der Nutzer:innen, wie das Beispiel von Facebook zeigt, das emotional „instabile" Teenager mit spezifischen Werbeanzeigen ansprach.[30]

Die unterschiedliche Behandlung von Personen aufgrund von vorhergesagten Merkmalen und Verhaltensweisen untergräbt demokratische Prinzipien, da sie zu Stereotypisierung, zur unfairen Behandlung von Abweichungen und zu epistemischer Ungerechtigkeit führt.[31] Der präventive Schutz individueller Rechte, kollektiver Interessen ist daher notwendig, damit die Risiken von Vorhersagemodellen kontrollierbar werden. Diese Risiken intensivieren sich, wenn trainierte Modelle über ihren ursprünglichen Zweck hinaus weiterverwendet werden. In demokratischen politischen Systemen setzen kollektive Entscheidungsprozesse (wie Wahlen) die Autonomie des Individuums voraus.[32] Diese Autonomie wird durch algorithmisch generierte Zuschreibungen, auf die einzelne Personen keinen Einfluss haben, beeinträchtigt. Darüber hinaus birgt die Anwendung von Vorhersagemodellen in verschiedenen Kontexten Risiken für das Recht auf Privatsphäre und den Schutz vor Diskriminierung. Erkenntnistheoretisch gesehen ist der Übergang zu einer prädiktionsbasierten Wissensordnung, die auf Korrelationen statt auf Kausalitäten beruht, zudem aufgrund fehlender Qualitätssicherungsmechanismen problematisch.[33]

Eine zweite Kategorie des potenziellen Missbrauchs trainierter KI-Modelle betrifft generative KI – insbesondere das Risiko, dass generative Modelle für die Erstellung falscher Informationen und Nachrichtenberichte[34] oder gefälschter Bilder verwendet werden.[35] Es hat sich gezeigt, dass Menschen „weitgehend nicht in der Lage sind, zwischen KI- und von Menschen erzeugten Texten zu unterscheiden".[36] Die mit dieser Technologie verbundenen Gefahren vervielfachen sich aufgrund der leichten Skalierbarkeit von KI-Systemen, die leicht dazu führt, dass im öffentlichen Diskurs extreme Positionen verstärkt werden.[37] Dies kann letztlich zu Wahlmanipulationen und der Verbreitung von Fehlinformationen führen, die reale Prozesse beeinflussen und die Demokratie gefährden.[38]

---

[28] *O'Neil*, [Fn. 7], S. 108, 148.

[29] *Yeung*, 'Hypernudge': Big Data as a mode of regulation by design, Inf. Commun. Soc. 2017, 118.

[30] *Susser/Roessler/Nissenbaum*, Online Manipulation: Hidden Influences in a Digital World, Georget. Law Technol. Rev. 2019, 1; *Zarsky*, Privacy and Manipulation in the Digital Age, Theor. Inq. Law 2019, 157.

[31] *McQuillan*, Predicted benefits, proven harms: How AI's algorithmic violence emerged from our own social matrix, Sociol. Rev. Mag. 2023; *Joque*, Revolutionary mathematics: Artificial intelligence, statistics and the logic of capitalism, 2022; *Mühlhoff*, [Fn. 11], 675.

[32] BVerfG Beschl. des Ersten Senats vom 15. Dezember 1983 - 1 BvR 209/83, BVerfGE 65, 1 – Volkszählungsurteil.

[33] *Joque*, [Fn. 31]; *Mühlhoff*, Die Macht der Daten, 2023.

[34] *Buchanan/Lohn/Musser/Sedova*, Truth, Lies, and Automation: How Language Models Could Change Disinformation, 2021.

[35] *Fallis*, The Epistemic Threat of Deepfakes, Philos. Technol. 2021, 623.

[36] *Kreps u. a.*, All the news that's fit to fabricate: AI-generated text as a tool of media misinformation, J. Exp. Polit. Sci. 2022, 104.

[37] *Buchanan/Lohn/Musser/Sedova*, Truth, Lies, and Automation: How Language Models Could Change Disinformation, 2021.

[38] *Zhou/Zhang/Luo/Parker/De Choudhury*, Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions, 2023; *Zarouali/Dobber/Pauw/de Vreese*, Using a Personality-Profiling Algorithm to Investigate Political Microtargeting: Assessing the Persuasion Effects of Personality-Tailored Ads on Social Media, Commun. Res. 2022, 1066.

4. Unzureichende Regulierung trainierter Modelle de lege lata: DSGVO und Antidiskriminierungsrecht

Die Zweckbindung bei der Verarbeitung personenbezogener Daten ist ein wichtiger Grundsatz der DSGVO, kodifiziert in den Datenschutzprinzipien des Art. 5 DSGVO.[39] Der Zweckbindungsgrundsatz schreibt vor, dass die für die Datenverarbeitung Verantwortlichen den Zweck der Datenerhebung spätestens zum Zeitpunkt der Erhebung der Daten festlegen müssen, und verbietet die Verarbeitung der Daten in einer Weise, die mit dem ursprünglich angegebenen Zweck gemäß Art. 5 Abs. 1 Buchst. b DSGVO unvereinbar ist. Diese Zwecke wiederum müssen festgelegt, eindeutig und legitim sein, um das Ziel der Datenverarbeitung zu definieren. Daher ist der Grundsatz der Zweckbindung eng verknüpft mit den Grundsätzen der Speicherbegrenzung und der Datenminimierung.[40] Das Prinzip der Zweckbindung verengt die Datenverarbeitung nicht absolut auf den ursprünglichen Zweck. Vielmehr muss die sekundäre Datenverwendung mit dem ursprünglichen Zweck *vereinbar* sein (vgl. Art. 5 Abs. Buchst. b DSGVO). Die Vereinbarkeit wird konkretisiert durch zwei Spezifikationen: Erstens werden gemäß Art. 5 Abs. 1 Buchst. b DSGVO, die genannten privilegierten Verarbeitungszwecke (für Archivzwecke im öffentlichen Interesse, wissenschaftliche oder historische Forschungszwecke oder zu statistischen Zwecken) als mit dem ursprünglichen Zweck im Sinne einer gesetzlichen Fiktion als vereinbar angesehen, gemäß Art. 89 Abs. 1 DSGVO. Zweitens fordert Art. 6 Abs. 4 DSGVO eine Vereinbarkeitsprüfung und stellt eine Reihe von Kriterien auf, um festzustellen, ob die Verarbeitung für einen anderen Zweck als den, für den die personenbezogenen Daten erhoben wurden, mit dem ursprünglichen Zweck vereinbar ist.

Die Zwecke der Datenverarbeitung müssen zu Beginn der Datenverarbeitung ausdrücklich festgelegt sein. Infolgedessen ist der Datenverarbeiter verpflichtet, in einem ersten Schritt die Zwecke der Datenverarbeitung zu definieren. Nur bei definierten Verarbeitungszwecken können weitere Voraussetzung für andere DSGVO-Anforderungen, wie die Datenschutzprinzipien, z.B. Datenminimierung überhaupt überprüft werden. Der zweite Schritt erfordert eine Prüfung, ob es sich bei der Weiterverarbeitung um einen privilegierten Zweck gemäß Art. 5 Abs. 1 Buchst. b DSGVO, Art. 89 Abs. 1 DSGVO handelt. Sollte dies nicht der Fall sein, muss in einem dritten Schritt das Vorliegen der Voraussetzungen des Art. 6 Abs. 4 DSGVO geprüft werden. Diese postulieren, dass die Vereinbarkeitsprüfung nicht anwendbar ist, wenn die Datenverarbeitung a) auf einer Einwilligung oder b) auf einer Rechtsvorschrift der Union oder eines Mitgliedstaats, die in einer demokratischen Gesellschaft eine notwendige und verhältnismäßige Maßnahme zur Wahrung der in Art. 23 Abs. 1 DSGVO genannten Ziele darstellt, beruht. In allen anderen Fällen muss der Datenverarbeiter die Vereinbarkeit der Zwecke nach den in Art. 6 Abs. 4 DSGVO genannten Kriterien prüfen.

Das Ziel der Zweckbindung ist es, den Betroffenen informierte Entscheidungen zu ermöglichen, welche Akteure ihre Daten verarbeiten und zu welchen Zwecken.[41] Der Grundsatz der Zweckbindung berücksichtigt dabei, dass Daten, sobald sie einmal erhoben und gespeichert sind, für jeden beliebigen Zweck weiterverwendet werden können. Dies birgt das Potenzial, die Kontrolle der betroffenen Personen zu unterlaufen und das Recht auf Datenschutz zu verletzen (Art. 8 GrCh). Darüber hinaus müssen die verfolgten Zwecke rechtmäßig sein, also nicht nur mit dem Datenschutzrecht, sondern der gesamten Rechtsordnung konform sein.[42] Dadurch soll es nicht den betroffenen Personen aufgebürdet werden, selbst die Legitimität dieser Zwecke überprüfen zu müssen. Diese Verantwortung liegt

[39] Zur Zweckbindung als Mittel zur datenschutzkonformen Nutzung von KI-Systemen: *Schuh/Weiss*, Die Zweckbestimmung und Zweckbindung als Weichenstellung für die DSGVO-konforme Nutzung von Daten für KI-Systeme, ZfDR 2024, 225.

[40] *Finck/Biega*, Reviving Purpose Limitation and Data Minimisation in Data-Driven Systems, Technol. Regul. 2021, 44.

[41] Artikel-29-Datenschutzgruppe, WP203/569/13, 2013. Dazu auch ausführlich: *Schuh/Weiss* [Fn.39], 225 (229).

[42] *Herbst*, in: *Kühling/Buchner* DSGVO, 4. Aufl. 2024, Art. 5 Rn. 37.

vielmehr bei den Verantwortlichen der Datenverarbeitung, vgl. auch Art. 5 Abs. 2 DSGVO. Im Sinne einer grundrechtskonformen Auslegung ist das Hauptziel der Zweckbindung, die betroffene Person zu schützen und die Kontrollierbarkeit der weiteren Datenverarbeitung und deren die Einhaltung des Datenschutzrechts zu ermöglichen.

Der Zweckbindungsgrundsatz ist „altes Datenschutzrecht".[43] Mit Wurzeln in Art. 8 GrCh ist die Zweckbindung ein Kernprinzip des Datenschutzrechts, und zwar schon vor Inkrafttreten der DSGVO. Verschiedene Stimmen haben Bedenken geäußert, dass in Bezug auf KI und Big-Data-Technologien die Zweckbestimmung, und insbesondere die Zweckbindung, in der Praxis zu Schwierigkeiten führt.[44] Wie bei allen Grundsätzen des Datenschutzes bestehen auch bei der Zweckbindung erhebliche Vollzugsdefizite („enorme Diskrepanz zwischen Recht und Wirklichkeit").[45] Auch die zahlreichen Bußgeldverfahren und die jüngste Rechtsprechung des EuGH vermögen das Durchsetzungsdefizit nicht wesentlich aufzufangen, da diese Entscheidungen nicht zu einer wesentlichen Änderung der datenintensiven Geschäftsmodelle geführt haben, die unseres Erachtens nach nicht mit der DSGVO vereinbar sind.[46] Datenmächtige Akteure verarbeiten Daten für hunderte von vagen und nicht spezifizierten Zwecke.[47] Eine gängige Praxis dieser Unternehmen und von Datenhändler:innen ist es, Daten zu sammeln und anschließend erst ihre Verwendung festzulegen.[48]

Wir argumentieren, dass die beschriebenen Herausforderungen nicht allein durch eine effektivere Durchsetzung des bestehenden Datenschutzgesetzes gelöst werden können; gleichwohl ist es notwendig, die bestehenden Vollzugsdefizite abzubauen.[49] Aus dreierlei Gründen geht die derzeitige Gesetzeslage nicht ausreichend auf die Zweckbindung für Modelle ein:

(1) Erstens, Modelle, die aus anonymisierten Daten bestehen, fallen nicht in den Anwendungsbereich der DSGVO, welche die Verarbeitung personenbezogener Daten adressiert.[50] Die Annahme, dass die Anonymisierung selbst eine erlaubnispflichtige Datenverarbeitung sein kann, löst dieses Problem nicht, da jede Zweckbindung oder -einschränkung nach der Anonymisierung verloren geht.

(2) Zweitens stimmen die Annahmen der DSGVO über Datenverarbeitungspraktiken oft nicht mehr mit der Realität überein. Grundsätzlich ist die DSGVO neben bisher in der Rechtspraxis wenig entscheidungsrelevanten systemischen Vorgaben wie der Datenschutzfolgenabschätzung in Art. 35 DSGVO in erster Linie ein individualrechtsbezogenes Regelwerk, was sich auch durch die Wichtigkeit der Betroffenenrechte in der Praxis widerspiegelt. Damit überantwortet die DSGVO allerdings auch der einzelnen betroffenen Person große Teile ihrer Durchsetzung – dies steht in deutlicher Spannung zu dem vorliegenden Umfeld aus Massendatenverarbeitungen, kollektiven Auswirkungen und struktureller informationeller Machtasymmetrie. In diesen Konstellationen ist die Durchsetzung von Betroffenenrechten erheblich erschwert und durch die routinierte Verknüpfung der Daten des

---

[43] *Finck/Biega*, [Fn. 40], 44.

[44] *Hildebrandt*, Slaves to Big Data. Or Are We?, Ipd Rev. Internet Derecho Política 2013, 7; *Zarsky*, Incompatible: the GDPR in the age of big data, Seton Hall Rev 2016, 995.

[45] *Koops*, The trouble with European data protection law, Int. Data Priv. Law 2014, 250.

[46] *Mühlhoff/Ruschemeier*, [Fn. 12], 1; *Ruschemeier*, Generative AI and Data Protection, in: *Calo/Ebers/Poncibo/Zou (Hrsg.),* Generative AI and the Law 2024 i.E.

[47] *Hahn*, Purpose Limitation in the Time of Data Power: Is There a Way Forward?, Eur. Data Prot. Law Rev. 2021, 31.

[48] *Ruschemeier*, Data Brokers and European Digital Legislation, Eur. Data Prot. Law Rev. 2023, 27.

[49] *Hahn*, [Fn. 47], 31.

[50] Wenig überzeugend im Kontext von Large Language Models: *Hamburgische Beauftragte für Datenschutz und Datensicherheit*, Diskussionspapier: Large Language Models und personenbezogene Daten; abrufbar unter: https://datenschutz-hamburg.de/fileadmin/user_upload/HmbBfDI/Datenschutz/Informationen/240715_Diskussionspapier_HmbBfDI_KI_Modelle.pdf. Zu einer grundrechtlichen Zweckbindung für private Akteure, die dieses Problem aber auch nicht löst: *Schuh/Weiss*, [Fn. 39], 225 (247 f.).

Einzelnen mit den Daten vieler anderer Betroffener schlicht nicht zielführend, um mehr Datenschutz zu erreichen. Denn „die eigenen Daten" gibt es im Kontext großer KI-Systeme nicht mehr, Daten stehen stets in Relation zu Daten von anderen. Darin liegt nicht nur ein Vollzugsdefizit, sondern ein *strukturelles* Problem der DSGVO: selbst eine optimale Durchsetzung aller Betroffenenrechte kann bspw. die Benachteiligung anderer Datensubjekte durch diskriminierende Prädiktionen nicht verhindern. Fakt ist, dass viele KI-Modelle im Einsatz *sind*, die entweder nicht mit der DSGVO vereinbar sind oder deren Vereinbarkeit ungeklärt ist. In der bisherigen Aufsichtsstruktur in der Union und den Mitgliedstaaten wird sich dies nicht auf absehbare Zeit ändern. Wir regen deshalb an, einen demokratischen Diskurs darüber zu führen, welche KI-Modelle und Anwendungszwecke wünschenswert sind, anstatt die DSGVO mit gesamtgesellschaftlichen und rechtspolitischen Ausrichtungen zu überfrachten, die dort so nicht angelegt sind.[51] Wir argumentieren deshalb, systemische Lösungen zu entwickeln, die sich in bestimmten Situationen informationeller Machtasymmetrien und massenhafter Datenverarbeitung nicht auf die Durchsetzung von Individualrechten beschränken. Dazu sollten Behörden – ggf. auch nicht nur Datenschutzbehörden – mit effektiveren Instrumenten ausgestattet werden, als vorgelagerte Verpflichtungen wie die Datenschutzfolgenabschätzung es sind, die maßgeblich durch die Verantwortlichen selbst vorgenommen werden.

In seinem Urteil in der Rechtssache Meta[52] hat der EuGH anerkannt, dass die Unterscheidung zwischen personenbezogenen und nicht-personenbezogenen Daten *de facto* obsolet ist.[53] Am Beispiel von Large Language Models wie GPT-3 und GPT-4 zeigt sich, dass es, wenn große Mengen von Daten aus dem Internet abgerufen werden, nicht mehr möglich ist, die Daten *ex post* nach normativen Kategorien zu differenzieren.[54] Der regulatorische Bezugspunkt des einzelnen Datenverarbeitungsvorgangs ist in diesen Fällen kein tauglicher Anknüpfungspunkt mehr. Denn wenn es zunehmend schwieriger wird, das Regelungsobjekt der Verarbeitung von personenbezogenen Daten zu identifizieren, verlieren individualschützend verstandene Mechanismen wie die Zweckbindung ihre Wirkung. Diese Probleme werden insbesondere durch die in der Praxis am weitesten verbreitete Rechtsgrundlage für die Verarbeitung verschärft: die Einwilligung. Die DSGVO sieht die Einwilligung nicht als Einschränkung, sondern als Ausübung des Grundrechts auf Datenschutz an. Viele Autor:innen, uns eingeschlossen, haben bereits argumentiert, dass die Einwilligung ein ungeeignetes Rechtsinstrument für die Legitimation der Datenverarbeitung in digitalen Kontexten ist.[55] Der EuGH selbst ging allerdings nicht davon aus, dass eine freiwillige Einwilligung allein aufgrund der marktbeherrschenden Stellung einer Social Media Plattform wie Meta ausgeschlossen sei.[56] In einem wettbewerbsrechtlichen Kontext ist dies in der konkreten Situation verständlich. Das Problem der Einwilligung im digitalen Kontext ist aber nicht nur auf die Marktstellung eines Akteurs zurückzuführen, sondern auch auf die schiere Informationsflut, die Menschen nicht verarbeiten

[51] *Engeler/Rolfes*: Datenschutzrechtliche Korrekturansprüche bei Erzeugung von Falschinformationen durch LLMs, Z. Für Datenschutz 2024, 423 (428).

[52] EuGH Urt. v. 4.7.2023 – C-252/21 = ZD 2023, 664 – Meta v Bundeskartellamt.

[53] Vgl. zur Kritik an der Kategorie selbst *Purtova*, The law of everything. Broad concept of personal data and future of EU data protection law, Law Innov. Technol. 2018, 40.

[54] *Ruschemeier*, Squaring the Circle, Verfassungsblog 2023, abrufbar unter https://verfassungsblog.de/squaring-the-circle/.

[55] *Ben-shahar/Schneider*, The Failure of Mandated Disclosure, Univ. Pa. Law Rev. 2011, 647; *Borgesius/Kruikemeier/Boerman/Helberger*, Tracking Walls, Take-It-Or-Leave-It Choices, the GDPR, and the ePrivacy Regulation, Eur. Data Prot. Law Rev. 2017, 353; *De/Imine*, Consent for targeted advertising: the case of Facebook, AI Soc. 2020, 1055; *Nguyen/Backes/Stock*, Freely Given Consent? Studying Consent Notice of Third-Party Tracking and Its Violations of GDPR in Android Apps, 2022; *Holland*, Privacy Paradox 2.0, Widener Law J. 2010, 893; *Ruschemeier*, Privacy als Paradoxon, in: *Friedewald, Roßnagel* (*Hrsg*.) 2022 - Künstliche Intelligenz 2022, S. 211 ff.

[56] EuGH Urt. v. 4.7.2023 – C-252/21 = ZD 2023, 664 (674 f.) – Meta v Bundeskartellamt.

können, dass sie faktisch keine informierte Entscheidung treffen.[57] Im Hinblick auf Informationsasymmetrien, haben die gleichen mächtigen Akteure Geschäftsmodelle geschaffen, die für die einzelne Person nicht zu überblicken und zu verstehen sind. Dazu gehören Konstellationen, in denen die Einwilligung an 300 verschiedene Datenverarbeiter gleichzeitig erteilt wird. Hinzu kommen potenzielle Ableitungen neuer personenbezogener Daten, welche die betroffene Person zum Zeitpunkt der Einwilligung schon gar nicht vorhersehen kann.

(3) Drittens ist auch bei Modellen, die personenbezogene Daten verarbeiten, die sekundäre Datennutzung durch Weitergabe und Weiterverkauf nicht wirksam geregelt.[58] In vielen Fällen der sekundären Datenverwendung ist der Grundsatz der Zweckbindung nicht mehr nachvollziehbar. Selbst wenn die Datenschutzrichtlinien öffentlich verfügbar sind und Zwecke wie „Personalisierung von Inhalten" oder „Verbesserung von Dienstleistungen" formulieren, gibt es faktisch keine Kontrolle darüber, ob dies der tatsächlichen Datenverarbeitungspraxis entspricht.[59] Dies liegt unter anderem an der erforderlichen Differenzierung zwischen Fällen einer Zweckänderung durch denselben Datenverarbeiter und der Weiterverwendung durch Dritte. Im ersten Fall müssen die Anforderungen für eine Zweckänderung gemäß Art. 5 Abs. 1 Buchst. b DSGVO, Art. 6 Abs. 4 DSGVO bestimmt werden. Im Gegensatz dazu wird der zweite Fall, bspw. der Verkauf von Datensätzen, als neue Datenverarbeitung durch einen dritten Verantwortlichen eingeordnet. Diese neue Verarbeitung kann auf eine neue Rechtsgrundlage im Einklang mit Art. 6 DSGVO gestützt werden, ohne dass sie Einschränkungen durch die ursprüngliche Zweckbindung unterliegt. Diese weitere Verarbeitung muss dann nur mit dem/den neuen Zweck(en) vereinbar sein.[60] Nach dem derzeitigen Verständnis besteht Vereinbarkeit grundsätzlich bei Anonymisierung oder Pseudonymisierung, da die Risiken für die Betroffenen hier als gering eingeschätzt werden. In jedem Fall können auch unvereinbare Zwecke durch die Einwilligung der betroffenen Person überwunden werden, was aber wegen der Ungeeignetheit der Einwilligung keinen ausreichenden Schutz bietet. Auf den ersten Blick ist es unklar, ob die Funktion von Art. 6 Abs. 4 DSGVO auf eine Vereinbarkeitsprüfung beschränkt ist oder ob Art. 6 Abs. 4 DSGVO auch einen Erlaubnistatbestand zur Weiterverarbeitung von personenbezogenen Daten für einen anderen Zweck darstellt. Nach dem Wortlaut und der Systematik kann sich Art. 6 Abs. 4 DSGVO aber nur auf die Auslegung des Erfordernisses der Vereinbarkeit nach Art. 5 Abs. 1 Buchst. b DSGVO beziehen, da es sich um eine Frage der Vereinbarkeit handelt und keine Ausnahme von der allgemeinen Regel des Art. 6 Abs. 1 DSGVO darstellen kann. Art. 6 Abs. 1 DSGVO bezieht sich nämlich nur auf die Buchstaben a bis f und nicht auf Abs. 4 der Bestimmung. Daher spezifizieren die Anforderungen aus Art. 6 Abs. 4 DSGVO den Art. 5 Abs. 1 Buchst. b DSGVO und schaffen keinen eigenen Erlaubnistatbestand.

Selbst wenn nationale oder EU-Antidiskriminierungsvorschriften[61] für die Schritte 2 und 3 der Datenverarbeitungskette gelten würden, würden sie unter den gleichen Durchsetzungsdefiziten wie

---

[57] *Ruschemeier*, [Fn. 54].

[58] *Ruschemeier*, Data Brokers and European Digital Legislation, Eur. Data Prot. Law Rev. 2023, 27.

[59] *Finck/Biega*, [Fn. 40], 44.

[60] Artikel-29-Datenschutzgruppe, WP203/569/13 (2013). Zu den Schwierigkeiten der Zweckbestimmung bei den unterschiedlichen Rechtsgrundlagen des Art. 6 Abs. 1: *Schuh/Weiss*, [Fn. 39], 225 (230 ff.).

[61] ZB das deutsche Allgemeine Gleichbehandlungsgesetz (AGG) vom 14. August 2006 (BGBl. I, S. 1897), zuletzt geändert durch Art. 4 des Gesetzes vom 19. Dezember 2022 (BGBl. I, S. 2510); RL 2000/43/EG des Rates vom 29. Juni 2000 zur Anwendung des Gleichbehandlungsgrundsatzes ohne Unterschied der Rasse oder der ethnischen Herkunft [2000] ABl. L 180/22; RL 2000/78/EG des Rates vom 27. November 2000 zur Festlegung eines allgemeinen Rahmens für die Verwirklichung der Gleichbehandlung in Beschäftigung und Beruf [2000] L 303/16; RL 2006/54/EG des Europäischen Parlaments und des Rates vom 5. Juli 2006 zur Verwirklichung des Grundsatzes der Chancengleichheit und Gleichbehandlung von Männern und Frauen in Arbeits- und Beschäftigungsfragen (Neufassung) [2006] ABl. L 204, 23; RL 2004/113/EG des Rates vom 13. Dezember 2004 zur Verwirklichung des Grundsatzes der Gleichbehandlung von Männern und Frauen beim Zugang zu und bei

das Datenschutzrecht leiden.[62] Aufgrund der kollektiven Dimensionen datenintensiver Technologien sind diese Fälle eine Form der „opferlosen Diskriminierung“: Das Gesetz geht davon aus dass einzelne Betroffene ihre Rechte geltend machen werden. Jedoch sind sie nicht mehr identifizierbar, und selbst wenn sie es wären, sind die Hürden für die Durchsetzung aufgrund der beschriebenen Machtasymmetrien zu hoch.

5. Gesetzlicher Rahmen für anonyme Daten

Der derzeitige Rechtsrahmen für anonymisierte Daten adressiert den Aspekt der informellen Machtasymmetrien durch Vorhersagemodelle nicht. Vielmehr folgt die Gesetzgebung bisher einer Dichotomie zwischen dem Schutz der Einzelnen durch das Datenschutzrecht und dem Schutz von nicht-individuellen Gütern, wie dem freien Verkehr von Daten zur Unterstützung des Binnenmarkts. Insbesondere die Verordnung über einen Rahmen für den freien Verkehr nicht-personenbezogener Daten in der Europäischen Union (2018/1807) verfolgt das Ziel der Entwicklung der Datenwirtschaft und der Stärkung der Wettbewerbsfähigkeit der Industrie in der EU. Dazu müsste sich auch mit Fragen der datenmächtigen Akteure auseinandergesetzt werden, die wir hier diskutieren. Die Verordnung adressiert nicht die Auswirkungen auf (nicht-gewerbliche) Endnutzer:innen, da sie sich auf die Anforderungen an die Datenlokalisierung bezieht, die Verfügbarkeit von Daten für die zuständigen Behörden, und die Übertragung von Daten für berufliche Nutzer (siehe Art. 1 VO (EU) 2018/1807). Da jede Regulierung von nicht-personenbezogenen Daten der Zweiteilung zwischen personenbezogenen und nicht-personenbezogenen Daten folgt, wird sie bei großen Datensätzen oder großen Sprachmodellen faktisch nicht durchsetzbar. Auch wenn die Verordnung ausdrücklich erwähnt, dass sie nur für den nicht-personenbezogenen Teil eines gemischten Datensatzes gelten und die Anwendung der DSGVO nicht berühren soll (Art. 2 Abs. 2 VO (EU) 2018/1807), ist die Unterscheidung in der Praxis schwierig, da sie eben eine Identifikation der Datentypen voraussetzt. Die Anbieter eines Produkts wie ChatGPT werden aufgrund der riesigen Datenbanken nicht in der Lage sein, die Differenzierung für alle verarbeiteten Daten vorzunehmen. Die gleichen Überlegungen gelten für die Identifizierung besonderer Kategorien von personenbezogenen Daten, vgl. Art. 9 Abs. 1 DSGVO.

6. Regulierung von Trainingsdaten

Die Regulierung von KI-Trainingsdaten jenseits der DSGVO adressiert, wenn überhaupt, nur einen Teil des hier beschriebenen Problems. Die Regulierung sollte bei den KI-Modellen selbst und ihrem potentiellen Verwendungskontext ansetzen. Die alleinige Regulierung der Daten, nicht aber des Kontexts ihrer Verwendung, hat sich als nicht besonders effektiv herausgestellt. Daher sind wir der Ansicht, dass die Risiken im Zusammenhang mit Trainingsdaten (z.B. Qualitätsrisiken, Diskriminierungsrisiken) sich aus der Anwendung der Modelle auf unbekannte und unbegrenzte Zwecke ergeben. Trainingsdaten sind inzwischen explizit in der KI-VO adressiert. Die KI-VO sieht in Art. 10 KI-VO Anforderungen für Trainings-, Validierungs- und Testdaten für Hochrisiko-KI-Systeme vor (deren Risikoeinstufung sich nach Art. 6 Abs. 1, 2 KI-VO i.V.m. Anhang III KI-VO richtet). Die Vorschriften zur Daten-Governance des Art. 10 KI-VO fordern, dass diese Daten „relevant, repräsentativ und so weit wie möglich fehlerfrei und vollständig sein sollen“ (Abs. 3). Zudem müssen die Datensätze, „soweit

der Versorgung mit Gütern und Dienstleistungen [2004]. ABL. L 373/37. Weitere Informationen zur Problematik der künstlichen Intelligenz und des Antidiskriminierungsrechts: *Gerards/Borgesius*, Protected Grounds and the System of Non-Discrimination Law in the Context of Algorithmic Decision-Making and Artificial Intelligence Articles and Essays, Colo. Technol. Law J. 2022, 1; *Wachter u. a.*, Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI, Comput. Law Secur. Rev. 2021, 105567.

[62] *Hacker*, A legal framework for AI training data - from first principles to the Artificial Intelligence Act, Law Innov. Technol. 2021, 257.

dies für die Zweckbestimmung erforderlich ist, die entsprechenden Merkmale oder Elemente berücksichtigen, die für die besonderen geografischen, kontextuellen, verhaltensbezogenen oder funktionalen Rahmenbedingungen, unter denen das Hochrisiko-KI-System bestimmungsgemäß verwendet werden soll, typisch sind“ (Abs. 4).

Damit ist jedoch nicht das Risiko adressiert, dass ein trainiertes KI-Modell, selbst wenn es aus relevanten, repräsentativen und „unvoreingenommenen“ Trainingsdaten erstellt wurde, verwendet oder wiederverwendet werden könnte für gesellschaftlich riskante und schädliche Zwecke. Beispielsweise sind bei der Entwicklung eines Modells zur Erkennung bösartiger Hautläsionen anhand von Fotografien menschlicher Haut die Qualität und Repräsentativität der Trainingsdaten entscheidend für ein System, das für alle Hauttypen gleichermaßen zuverlässig arbeitet. [63] Ein solches System könnte zwar in seiner primären Anwendung ein wertvolles Hilfsmittel in der medizinischen Versorgung darstellen. Aber das Risiko der sekundären Verwendung dieses Systems für diskriminierende Zwecke, zum Beispiel in der Risikobewertung von Versicherungen, bezieht sich nicht auf die Qualität der Trainingsdaten, sondern auf den nachfolgenden Anwendungskontext, sodass Vorgaben für Trainings- und Validierungsdaten dieses Problem nicht adressieren.

III. Zweckbindung für Modelle

Im vorherigen Abschnitt haben wir erörtert, dass der unregulierte Einsatz von KI-Modellen zu sekundären Zwecken erhebliche gesellschaftliche Risiken mit sich bringen kann, die unter der aktuellen Gesetzeslage nicht hinreichend adressiert werden. In diesem Abschnitt stellen wir das Prinzip der Zweckbindung für KI-Modelle als potenziellen Lösungsvorschlag vor.

1. Der Grundgedanke

Aus ethischer und konzeptueller Perspektive argumentieren wir für eine *präventive* Regulierung, die den unbegrenzten Möglichkeiten der potentiell gesellschaftsschädlichen Nutzung und Wiederverwendung trainierter ML-Modelle Grenzen setzt. Die *Modelldaten* – das trainierte Modell, repräsentiert durch einen Datensatz, der aus Schritt 2 der typischen, in II.1 skizzierten Datenverarbeitungskette hervorgeht – identifizieren wir als regulatorischen Eingriffspunkt. Hier setzt ein für KI-Modelle erweiterter Grundsatz der Zweckbindung an, der für die Verarbeitung (z.B. Speicherung, Weitergabe und Nutzung) dieser Modelldaten gilt. Die Modelldaten sind von den Trainingsdaten eines Modells zu differenzieren. Deshalb unterscheidet sich der Vorschlag einer Zweckbindung für die Modelldaten von dem datenschutzrechtlichen Konzept der Zweckbindung für Trainingsdaten, da die Trainingsdaten nur einmal verwendet werden (in Schritt 1) zur Erstellung des trainierten Modells und dann verworfen werden können.

In dem in II.2 beschriebenen Beispiel wurde ein Vorhersagemodell für Alkoholkonsum für einen Zweck erstellt, der im besten Fall als nützlich, förderungswürdig oder wünschenswert angesehen werden kann. Sobald das Modell jedoch trainiert ist, gibt es Konstellationen, in denen die Modelldaten aus anonymen Daten bestehen, die nicht in den Geltungsbereich der DSGVO fallen. Sie können daher auf unkontrollierte, auch schädliche und missbräuchliche Weise weiterverwendet werden. Weil diese Wiederverwendung bisher gänzlich im Ermessen der Modelleigentümer (in erster Linie großer Technologieunternehmen) liegt, unterliegt die Wiederverwendung von trainierten Modellen aktuell keiner externen Kontrolle. Diese Praxis muss stärker in den Blickpunkt der Öffentlichkeit gerückt und

[63] *Guo u. a.*, Bias in, bias out: Underreporting and underrepresentation of diverse skin types in machine learning research for skin cancer detection - A scoping review, J. Am. Acad. Dermatol. 2022, 157; *Buolamwini/Gebru*, Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, Conf. Fairness Account. Transpar. 2018, 77.

kontrolliert werden, da sie einen wesentlichen Aspekt der informationellen Machtasymmetrie zwischen Daten- und KI-Unternehmen auf der einen Seite und einzelnen betroffenen Personen sowie der Gesellschaft auf der anderen Seite darstellt. Die Regulierung dieser Machtasymmetrie bildet den Kern unseres Vorschlags.

Entscheidend ist aus unserer Sicht, dass diese Asymmetrie nicht hinreichend durch einen Regulierungsansatz eingehegt werden kann, der sich nur auf Verarbeitungsschritt 1 (die Verarbeitung von Trainingsdaten, inkl. Training des Modells) oder nur für den Verarbeitungsschritt 3 (die Anwendung des Modells auf einen Einzelfall, siehe II.1) konzentriert. Dies veranschaulicht der beschriebene Fall eines Modells, welches bspw. das Hepatitis-B-Risiko anhand von Social-Media-Nutzungsdaten vorhersagen kann.[64] Ohne eine Regulierung der Zweckbindung kann das trainierte Vorhersagemodell weitergegeben oder verkauft werden, um Teil eines Systems zu werden, das bei Auswahlverfahren eingesetzt wird und dort Menschen mit bestimmten Krankheitsdispositionen diskriminiert. Regulierungsbedürftig ist aus unserer Sicht bereits die das Vorhandensein eines trainierten Modells, welches zweckentfremdet werden könnte (Verarbeitungsschritt 2), nicht erst seine tatsächlich Anwendung auf Einzelfälle (Schritt 3). Denn Vorhandensein und Verbreitung eines Modells begründen informationelle Machtasymmetrie und die damit einhergehenden Missbrauchsrisiken. Jede Regelung, die erst die Anwendung eines Modells auf Einzelfälle beschränkt, vernachlässigt den präventiven Ansatz, der für die Kontrolle von Machtasymmetrien essentiell ist. Die KI-VO stuft KI-Systeme zur Personalauswahl als Hochrisikosysteme nach Art. 6 Abs. 2 KI-VO i.V.m. Anhang III Nr. 4a KI-VO) ein. Damit wird das Problem der Verknüpfung zwischen einem für andere Zwecke trainierten Modells (Schritt 1) und einer Sekundärnutzung nicht adressiert, sondern nur der dritte Schritt der finalen (zweckentfremdeten Modellanwendung) reguliert. Zudem wird bei der Verwendung von prädiktiven Modellen für die Auswahl von Bewerber:innen die Notwendigkeit eines präventiven Ansatzes besonders deutlich, da es für die von der Auswahl betroffenen Personen keine zumutbare Möglichkeit gibt, in diese Form der Datenverarbeitung nicht einzuwilligen.

Für die beschriebenen Beispielsituationen ist es typisch, dass die (missbräuchliche) Sekundärnutzung eines trainierten Modells auch mit einer Modifikation des betroffenen Modells einhergeht, so dass das System im sekundären Verwendungskontext möglicherweise nicht mehr die gleichen personenbezogenen oder sogar sensiblen Informationen generiert. Um bei dem Beispiel eines Modells zu bleiben das die Prävalenz von Hepatitis B vorhersagen kann: Wenn dieses Modell in einem größeren System zur Beurteilung von Stellenbewerber:innen als Modul zur Risikobemessung verwendet wird, könnten die Betreiber dies auf eine Art und Weise tun, die die „Prävalenz von Hepatitis B“ niemals explizit anzeigt, auswertet oder in einer internen Variable des Computersystems speichern lässt. Die Modelldaten des ursprünglichen Modells könnten vielmehr in ein größeres Modell einfließen, das schlichte Ja/Nein-Entscheidungen darüber fällt, ob eine Bewerber:in zu einem Vorstellungsgespräch eingeladen werden soll. Es ist daher für die betroffenen Personen in Schritt 3 schwer bis nahezu unmöglich, zu beweisen, und für die Entwickler:innen und Betreiber:innen des Systems leicht zu verbergen, dass das ursprüngliche Hepatitis-B-Vorhersagemodell bei der Erstellung des Einstellungsentscheidungsmodells wiederverwendet wurde. Dieses Beispiel verdeutlicht die Bedeutung eines regulatorischen Ansatzes, der die Möglichkeiten der Verarbeitung von Modelldaten in Verarbeitungsschritt 2 begrenzt.

Zweckbindung für Modellen bedeutet, dass die Erstellung und Nutzung von Modellen nur dann zulässig sind, wenn der Zweck, zu dem dies erfolgt, legitim ist und im Voraus benannt wird. Dadurch trennen wir in unserem Vorschlag die Regulierung von KI-Modellen von einer Identifikation einzelner Datenvearbeitungsvorgänge und betroffener Personen. Angesichts der kollektiven und überindividuellen Risiken halten wir es nicht für ausreichend, dass überwiegend betroffene Personen die Einhaltung der Zweckbindung durch den Auftragsverarbeiter durch individuelle Rechte

[64] *Mühlhoff/Willem*, [Fn. 14].

kontrollieren müssen, denn diese müssen identifizierbar sein und ihre Betroffenenrechte auch wahrnehmen und durchsetzen. Die strukturellen Vollzugsdefizite im Datenschutz zeigen, dass dieses Konzept an seine Grenzen stößt (siehe dazu oben 4 (2)) Stattdessen sollte eine demokratisch legitimierte Institution entscheiden, welche Zwecke angesichts der Risiken des jeweiligen Modells erstrebenswert sind. Daher ist es unser Ziel, den Fokus auf Individualrechte und individuelle Datenverarbeitungen, welcher derzeit die Datenverarbeitung in Schritt 1 und 3 rechtlich strukturiert, durch eine davon unabhängige Zweckbindung für KI-Modelle als eigenständige Regulierungsobjekte zu ergänzen. Daher schlagen wir eine *ex ante* Regelung mit kollektiven Interessen als Maßstab für die Bestimmung der zulässigen Zwecke vor. Denn die Hoffnung, dass Machtasymmetrien, die aus Big-Data-Praktiken entstehen, durch eine *ex-post*-Regulierung wirksam einzudämmen sind, hat sich bisher nicht erfüllt.[65] Gegenüber der Schaffung neuer Betroffenenrechte sehen wir daher die Zweckbindung für Modelle als ein Instrument zum Schutz individueller Rechte und Interessen der Gesellschaft an. Im Zusammenhang mit Datenpraktiken, die die Daten von Millionen von Menschen ausnutzen und potenziell jeden betreffen können (das heißt insbesonder, auch Dritte, die in diesen Daten nicht enthalten sind), erscheint es unangebracht, die Verantwortung einzelnen Personen aufzubürden. Stattdessen muss eine demokratische Beteiligung verschiedener Interessengruppen ermöglicht und das politische Kollektiv befähigt werden, über die gewünschten Zwecke von KI-Modellen zu regulieren und zu kontrollieren.

2. Zweckbindung für Modelle als Risikoprävention

Unser Ausgangspunkt ist die normative Struktur des Risikopräventionsrechts, das sich ex ante mit Risiken für individuelle Rechte, kollektive Interessen und Gesellschaften befasst. Im Umweltrecht gibt es zahlreiche Belege für die Notwendigkeit, Risiken, die durch große und mächtige Akteure entstehen, durch Rechtsvorschriften zur Risikoprävention zu kontrollieren und zu begrenzen. Regulierungsfragen in Bezug auf generative und prädiktive KI-Modelle liegen ähnlich: Erstens gibt es viele Hinweise darauf, dass prädiktive Modelle individuelle und gesamtgesellschaftliche Risiken mit sich bringen. Aufgrund der Funktionsweise der Technologie sind ihre Ermöglichungsstruktur und ihre Wirkungen kollektiv – sie beruhen auf den Daten vieler Datensubekte und betreffen in ihrer Anwendung potenziell eine große Anzahl von Individuen, dadurch entfalten sie erhebliche Spillover- und Hebeleffekte.[66] Zweitens sind diese globalen Technologien nur schwer durch individuelle Rechtssysteme und nationale Durchsetzungsmechanismen zu handhaben. Drittens: Die Wirkung von Marktmechanismen ist in digitalen Umgebungen nicht vergleichbar mit analogen Märkten; sie gründet sich auf Netzwerkeffekte, Monopole und die Kontrolle von Infrastrukturen. Diese Charakteristika digitaler Märkte werden nicht nur durch das Verhalten der Akteure beeinflusst, sondern auch durch Größen- oder Verbundvorteile, Netzeffekte, Umstellungskosten, asymmetrische und begrenzte Informationsverfügbarkeit und Verzerrungen im Verbraucherverhalten.[67]

Auch der Einsatz risikoreicher Modelle, die auf der aggressiven Extraktion von Daten zahlloser Datensubjekte beruhen, unterliegt aktuell keinem ex ante bestehenden Rechtfertigungserfordernis. Obwohl der Vorteil personalisierter Werbung gegenüber nicht-personalisierter Werbung gar nicht nachgewiesen ist, sind diese Praktiken allgegenwärtig und erzeugen massenhafte Datenschutzverstöße.[68] Wir plädieren deshalb für die stärkere Übernahme der theoretischen

[65] *Zarsky*, Incompatible: The GDPR in the Age of Big Data, Seton Hall Law Rev. 2017, 995 (1011).

[66] *Ben-Shahar*, Data Pollution, J. Leg. Anal. 2019, 104 (105).

[67] *Laux u. a.*, Taming the few: Platform regulation, independent audits, and the risks of capture created by the DMA and DSA, Comput. Law Secur. Rev. 2021, 105613.

[68] *Marotta/Abhishek/Acquisti*, Online tracking and publishers' revenues: An empirical analysis, 2019.

Grundlagen des Verhältnismäßigkeitsgrundsatzes[69] aus dem Gefahrenabwehrrecht für die Regulierung von KI. Die normative Struktur der Prüfung, ob ein Mittel seinen Zweck erreicht, im Verhältnis zur Erheblichkeit der Beschränkung, umfasst den Schutz von individuellen, kollektiven und politischen Interessen auf der Seite der Zweckerreichung. Zugleich begrenzen Informationsasymmetrien die Interessen der Akteure von vornherein, weshalb von einer geringeren Eingriffstiefe ausgegangen werden sollte. Mit anderen Worten: Je mehr Menschen und wichtige Rechtsgüter betroffen sind, desto mehr ist eine Regulierung gerechtfertigt. Je mehr die Regulierung gerechtfertigt ist, desto höher sollte das Niveau der demokratischen Legitimation auf der Ebene der strukturellen und konkreten Entscheidungen sein. Neben dem normativen theoretischen Rahmen ist für den Einsatz von KI-Modellen die faktische Ausgangsposition entscheidend: Der gesellschaftliche Nutzen ist in den meisten Bereichen noch spekulativ, aber die Schäden sind empirisch bewiesen.

In diesem Zusammenhang verstehen wir Zweckbindung für Modelle als ein Instrument, um präventiv den Risiken durch spezifische KI-Modelle zu regulieren. Zweckbindung für gesellschaftlich riskante Modelle bestimmter mächtiger Akteure oder risikoreichen Anwendungskontexten würde helfen, Machtasymmetrien auszugleichen und den jeweiligen Kontext der KI-Nutzung zu demokratisieren. Durch die Anwendung des Verhältnismäßigkeitsgrundsatzes und der normativen Grundlagen der Risikoprävention können Parameter identifiziert werden, die eine Auswahl zulässiger Zwecke ermöglichen. Im Gegensatz zur KI-VO schlagen wir vor, nicht auf den primären Verwendungszweck abzustellen, sondern auf die Position der Akteure, die Möglichkeiten und Risiken der weiteren Verbreitung, die potentiell durch weitere Verbreitung betroffenen Rechtsgüter und vor allem auf die kollektiven Auswirkungen solcher Sekundärnutzungsweisen.[70] Je wahrscheinlicher es ist, dass durch eine ungeplante Sekundärnutzung eine große Zahl von Individuen oder ganze Gesellschaften betroffen sind, desto eher ist eine Bindung des Zwecks an den primären Nutzungskontext gerechtfertigt. Dieses Risiko kann auch dadurch entstehen, dass die beteiligten Akteure besonders mächtig sind, Zugang zu umfangreichen Datenressourcen haben und eine ähnliche Macht ausüben wie staatliche Stellen, jedoch aufgrund ihres Status als private Akteure nicht grundrechtsverpflichtet sind. Die Zwecke der Datenverarbeitung sollten nicht von diesen datenmächtigen Akteuren selbst festgelegt und kontrolliert werden, sondern durch demokratisch legitimierte Prozesse und Regeln.

3. Kritik an der Zweckbindung nach der DSGVO im Kontext von Big Data

Vielfach wurde kritisiert, dass Big Data und das Prinzip der Zweckbindung unvereinbar seien, insbesondere im Kontext der DSGVO.[71] Wenn man als primäres Ziel ansieht, dass die DSGVO Big-Data-Praktiken ermöglichen und den freien Datenverkehr fördern soll (vgl. Art. 1 Abs. 2 DSGVO), überzeugt die Feststellung der Unvereinbarkeit. Dies würde aber der grundrechtsschützenden Ausrichtung der

[69] Siehe dazu: *Engle*, The History of the General Principle of Proportionality: An Overview, Dartm. Law J. 2012, 1; *Duarte/Sampaio*, Proportionality in law, 2018. Aus verfassungsrechtlicher Perspektive: *Klatt/Meister*, The Constitutional Structure of Proportionality, 2012. Zu den philosophischen Grundlagen: *Andreescu/Puran*, The Philosophical Basis of the Principle of Proportionality, Chall. Knowl. Soc. 2022, 188.

[70] Zur vorgeschlagenen KI-VO siehe auch III.4. Eine umfassende Analyse, warum die vorgeschlagene KI-VO die Risiken einer unberücksichtigten Sekundärnutzung trainierter Modelle nicht ausreichend berücksichtigt, wird in einer separaten Veröffentlichung besprochen, pre-print: *Mühlhoff/Ruschemeier*, Updating Purpose Limitation for AI: A normative approach from law and philosophy, 2024; https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4711621.

[71] *Hildebrandt*, [Fn. 44], 7; *Moerel/Prins*, Privacy for the Homo Digitalis: Proposal for a New Regulatory Framework for Data Protection in the Light of Big Data and the Internet of Things, 2016; *Mayer-Schönberger/Padova*, Regime Change? Enabling Big Data through Europe's New Data Protection Regulation, Sci. Technol. Law Rev. 2016, 315; *Zarsky*, [Fn. 65], 995. Zu Auflösung dieses Konflikts: *Schuh/Weiss* [Fn.39], 225 (228 ff.).

DSGVO nicht gerecht. Grundsätzlich sind KI-Systeme, die personenbezogene Daten verarbeiten, die Zweckbindung aber nicht einhalten, rechtswidrig. Wie bereits erläutert, bestehen aber Regulierungslücken aufgrund der möglichen Anonymisierung, des zwischen Training und Anwendung von KI-Modellen stattfindenden Wechsels des betroffenen Datensubjekts, und der Regulierungsstruktur der DSGVO, die über Vollzugsdefizite hinausgehen. Um diese Sackgasse zu vermeiden, geht unser Regulierungsvorschlag das Problem nicht von der Big-Data-Seite an, d.h. der Erhebung und Verarbeitung von Trainingsdaten, wie in Verarbeitungsschritt 1 (siehe Abschnitt II.1). Vielmehr adressieren wir die Verarbeitung einer völlig neuen Art von Daten – der Modelldaten –, die den Gegenstand von Verarbeitungsschritt 2 bilden (siehe Abschnitt II.1). Denn das trainierte Modell ist oft der Kern einer daran geknüpften informationellen Machtasymmetrie. In Anbetracht dieser Klarstellung ist die Kritik am Grundprinzip der Zweckbindung für unseren Ansatz nicht. Siehe zum Beispiel Hildebrandt, die ausführt:

> „Wenn Big Data von Interesse ist, weil es Muster erzeugt, die nicht vorhersehbar waren, und somit eine Nutzung ermöglicht, die nicht voraussagbar ist, dann ist eine Zweckbindung anmaßend und geht von der falschen Prämisse aus. Wir wissen nicht im Voraus, welche Nutzung ermöglicht wird, und um das herauszufinden, müssen wir zuerst die Daten auswerten [...]. Der Wert von Big Data kann nur freigesetzt werden, wenn wir die Neuartigkeit des gewonnenen Wissens anerkennen und die Zweckbindung entsprechend dem innovativen Potenzial seiner Ergebnisse überdenken.“[72]

Diese Argumentationslinie unterstützt im Ergebnis unseren Vorschlag, den regulatorischen Eingriffspunkt von Big Data (Trainingsdaten) auf die Modelldaten und deren Verwendungskontext zu verlagern. Zweckoffene Datenextraktion oder Data Mining kann als ein explorativer Prozess betrachtet werden, der nicht unmittelbar an einen Zweck gebunden ist. Wenn das Data Mining das Ziel der explorativen Erstellung von Modellen beinhaltet, würde unser Verfahren zur Zweckbindung so etwas wie „Grundlagenforschung“ als gültigen Zweck einschließen. Dadurch ergeben sich *keine* Einschränkungen für den Primärzweck der Forschung, die das Potenzial von Big Data nutzen kann. Gleichzeitig stellt unser Vorschlag jedoch sicher, dass anwendbare Modelle, die aus der Grundlagenforschung hervorgehen, nicht sofort in der Praxis eingesetzt werden können. Wenn entsprechende Modelle in anderen Anwendungsbereichen verwendet werden sollen, wäre nach unserem Dafürhalten eine neue Überprüfung dieses sekundären Zwecks erforderlich.

Eine andere Facette der Kritik an Zweckbindung im Kontext von Big Data verwendet Helen Nissebaums philosophischen Rahmen einer „Privatsphäre als kontextuelle Integrität“ als Ausweg aus der vermeintlichen Sackgasse der Zweckbindung für Big Data.[73] Hahn führt dazu aus, dass kontextuelle Integrität potenziell durch eine zweckfreie Datenverarbeitung verletzt werden könnte.[74] Da das Prinzip der kontextuelle Integrität es ermögliche, die im Erhebungskontext von Big Data gültigen „informationellen Normen“ gegeneinander abzuwägen, behauptet Hahn, dass dieser Ansatz überall dort eine nuanciertere Bewertung der Zulässigkeit von Big-Data-Praktiken erlaube, wo der Grundsatz der Zweckbindung gegenüber großen Datenunternehmen versage (vgl. S. 41-42):

> „Daher wird das Argument vorgebracht, dass der kontextbezogene Integritätsrahmen als Ausgangspunkt für eine strengere Durchsetzung des Grundsatzes der Zweckbindung in Bezug auf Unternehmen mit Datenmacht im Besonderen fungieren kann. Es wird vorgeschlagen, dass

[72] *Hildebrandt*, [Fn. 44], 7.

[73] *Nissenbaum*, Privacy in Context, 2009; *Hildebrandt*, [Fn. 44], 7 (37 ff.).

[74] *Hahn*, [Fn. 47], 31.

der Rahmen verwendet wird, um die Konsequenzen der Nichteinhaltung der Zweckbindung zu evaluieren, um zu zeigen, dass diese die Erwartungen des Datensubjekts verletzen."[75]

Wenn es auch zutreffend sein mag, dass die kontextuelle Integrität eine differenziertere Analyse des Schutzes der Privatsphäre der Datensubjekte in den Trainingsdaten erlaubt, so gilt doch ein ähnlicher Einwand wie zuvor: Unser Vorschlag zielt nicht darauf ab, Verarbeitungsschritt 1 (siehe II.1) zu regeln, d.h. die Erhebung und Verarbeitung von Big Data als Trainingsdaten. Vielmehr schlagen wir eine Zweckbindung vor, die die *Modelldaten* aus Schritt 2 (siehe II.1) betrifft. Diese regelt dann, wie die Daten im Hinblick auf mögliche Anwendungen in Schritt 3 (siehe II.1) verarbeitet werden. Da bei den Modelldaten davon ausgegangen werden kann, dass es sich um anonyme und hochaggregierte Daten handelt,[76] gibt es kein Datensubjekt auf das das Hahn'sche Argument zutreffen könnte. Stattdessen geht es uns um die Notwendigkeit eines präventiven Schutzes *aller* Personen vor den potenziellen Schäden, die aus der Anwendung der Modelldaten resultieren können.

4. Herausforderungen bei der Regulierung von Zwecken: KI-VO

Die Regulierung von Zwecken ist ein anspruchsvolles Ziel. Denn Zwecke können auf verschiedenen Abstraktionsebenen und aus unterschiedlichen Perspektiven formuliert werden.[77] Deshalb ist es wichtig zu bestimmen, wie und von wem die Zwecke von bestehenden und zukünftigen Modellen definiert werden sollen. In Betracht kommen objektive Dritte, Interessengruppen oder die Nutzer:innen des Modells. Der Regulierungsansatz der KI-VO bestimmt das Risiko eines KI-Modells einerseits nach einer produktsicherheitsrechtlichen Klassifizierung, andererseits durch den Anwendungskontext, der durch den intendierten Verwendungszweck der Anbieter nach Art. 6 Abs. 1, 2, 7 Abs. 2 Buchst. a KI-VO i.V.m. Anhang III KI-VO.[78] Wir sind stehen dem aus mehreren Gründen kritisch gegenüber:

Die KI-VO ist von ihrem Regelungsansatz her stark produktsicherheitsrechtlich orientiert und entgegen ihrer Ziele, vgl. Art. 1 Abs. 1 KI-VO, nicht als primäres Instrument zur Sicherung von Grundrechten oder zur Bewältigung gesellschaftlicher Risiken konstruiert.[79] Dies zeigt sich auch darin, dass die KI-VO die Vertrauenswürdigkeit eines Systems als entscheidend für die Akzeptanz der damit verbundenen Risiken einstuft. Die gesellschaftlichen Risiken, die dieser Artikel adressiert, können aber nicht am Maßstab der Vertrauenswürdigkeit von Modellen *innerhalb* ihrer primären Anwendungskontexte bemessen werden. Das Problem von Sekundärnutzungen und dadurch multiplizierten informationellen Machtasymmetrien wird durch die KI-VO nicht explizit aufgegriffen. Ihr Risikoklassifizierungsschema berücksichtigt insbesondere die Stellung der privaten Akteure und die daraus resultierenden Machtverhältnisse nicht; mit Ausnahme einiger weniger sektoraler Regelungen für kleine und mittlere Unternehmen. Die in Anhang III KI-VO aufgeführten Verwendungszwecke decken sich in einigen Fällen mit den problematischen Zwecken, für die Modelle verwendet werden. Viele der grundrechtsrelevanten Anwendungskontexte von Hochrisikosystemen des Anhangs III KI-VO liegen im staatlichen Bereich, z.B. der Einsatz von Behörden bei der Vergabe öffentlicher Leistungen (Nr. 5 a), im Rahmen der Strafverfolgung (Nr. 6), im Bereich Asyl, Migration und Grenzkontrollen (Nr. 7) oder der Rechtspflege (Nr. 8).

Die Klassifizierung von Risiken anhand der Verwendungsabsicht nach Art. 6 Abs. 2 KI-VO i.V.m. Anhang III KI-VO steht in enger Verbindung mit Normen, die die Normungsorganisationen für KI-

[75] *Hahn*, [Fn. 47], 31 (43).

[76] Darstellend zu „verallgemeinertem Wissen", *Loi/Christen*, Two Concepts of Group Privacy, Philos. Technol. 2020, 207.

[77] *Ruschemeier*, Der additive Grundrechtseingriff 2019, S. 145 f., S. 190 ff.

[78] Dazu auch *Ruschemeier*, in: *Martini/Wendehorst*, KI-VO, Art. 7 Rn. 83 ff.

[79] Vgl. *Almada/Petit*, The EU AI Act: a medley of product safety and fundamental rights?, 2023; Art. 6 I KI-VO.

Systeme noch zu definieren haben (vgl. Art. 42 Abs. 1 KI-VO).[80] Dieser Ansatz spiegelt nicht in adäquater Weise die komplexen Machtstrukturen des Einsatzes von KI-Modellen wider, denn ein Vertrauen auf Selbstregulierung anhand von harmonisierten Normen lässt machtpolitische Erwägungen außer Acht.[81] In dieser Regulierungssystematik fehlt es zudem an einer Risikobewertung durch eine unabhängige Stelle. In diesem Kontext, wie bereits kritisiert wurde,[82] sind die zentralen Akteure, auf die sich die KI-VO bezieht, nicht die Anwender:innen, sondern die europäischen Normungsorganisationen des Europäischen Komitees für Normung (CEN) und des Europäisches Komitees für elektronische Normung (CENELEC). Diese Organisationen sind zuständig für die Entwicklung harmonisierter Normen (vgl. Art. 40 ff. KI-VO). Infolgedessen fehlen im Verordnungstext selbst die materiellrechtlichen Anforderungen und der soziotechnische Kontext der Systeme: Wann soll Diskriminierung verboten sein? Wann ist menschliche Aufsicht sinnvoll? Welche ethischen Standards sollten für Systeme gelten?[83] Als bewusste politische Entscheidung lagert die KI-VO damit die ethischen und regulatorischen Kernfragen an private Normsetzungsorganisationen aus. Dies ist problematisch, weil es diesen Organisationen an Vorgaben zur Stakeholder-Beteiligung und demokratischer Legitimation fehlt.[84]

Am Beispiel der Vorgaben für KI mit allgemeinem Verwendungszweck (Art. 51 ff. KI-VO) zeigt sich, dass eine Orientierung an von Anbieter:innen *ex ante* definierten Verwendungszwecken bei Systemen, die für verschiedene Zwecke eingesetzt werden können, an ihre Grenzen stößt. Für Modelle mit allgemeinem Verwendungszweck (Art. 3 Nr. 63) sieht die KI-VO nun eine eigene Kategorie des systemischen Risikos vor (vgl. Art. 55 KI-VO). Anhang XIII benennt dazu Kriterien für Klassifizierung von Modellen mit allgemeinem Verwendungszweck als Modelle mit systemischem Risiko, wobei insbesondere technische Parameter aufgeführt sind. Bezüge zu Missbrauchsrisiken zu unterschiedlichen Zwecken finden sich nicht.

KI-Modelle mit allgemeinem Verwendungszweck, die ein systemisches Risiko darstellen (z.B. aufgrund ihrer Reichweite, Parameter, Benchmarks etc., Anhang XIII), sollten unserer Auffassung nach ebenfalls einer Zweckbindung unterliegen, wodurch zweckoffene Modelle ausgeschlossen sind. Im Sinne eines alternativen Ansatzes schlagen wir eine Auflistung zulässiger Zwecke, relevanter Akteure und akzeptabler Einsatzkontexte von KI-Modellen mit systemischen Risiken vor, welche in demokratischen Prozessen zu definieren sind. Anstatt das mit einem System verbundene Risiko nach individuellen und beabsichtigten Einsatzzwecken, wie sie von den Anbieter:innen deklariert werden, zu bestimmen, löst unser Ansatz die Konflikte zwischen beabsichtigten Zwecken und KI-Modelle mit allgemeinem Verwendungszweck, indem wir gesellschaftlich nützliche Zwecke identifizieren die mit dem Grundsatz der Verhältnismäßigkeit in Einklang stehen.

[80] Dazu auch *Wachter*, Limitations and Loopholes in the EU AI Act and AI Liability Directives: What This Means for the European Union, the United States, and Beyond, Yale J Tech, 671.

[81] *Kuziemski/Misuraca*, AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings, Telecommun. Policy 2020, 101976.

[82] *Laux/Wachter/Mittelstadt*, Three Pathways for Standardisation and Ethical Disclosure by Default under the European Union Artificial Intelligence Act, 2023; *Veale/Borgesius*, Demystifying the Draft EU Artificial Intelligence Act: Analysing the good, the bad, and the unclear elements of the proposed approach, Comput. Law Rev. Int. 2021, 97.

[83] Vgl. *Ruschemeier/Mühlhoff*, Daten, Werte und der AI Act: Warum wir mehr Ethik für bessere KI-Regulierung brauchen, Verfassungsblog 2023, abrufbar unter https://verfassungsblog.de/daten-werte-und-der-ai-act/.

[84] *Veale/Borgesius*, [Fn. 82], 97; *Smuha u. a.*, How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act, SSRN Electron. J. 2021, 33899991.

IV. Fazit und Ausblick

KI birgt Risiken und Chancen, die Milliarden von Menschen weltweit betreffen. Der allgemeine Ruf nach einer wirksamen Regulierung dieser Technologie entspringt der berechtigten Sorge, dass im status quo private Unternehmen, die von wirtschaftlichen Interessen geleitet werden, die uneingeschränkte Kontrolle über den Einsatz dieser Technologie haben. In diesem Beitrag haben wir das Konzept der Zweckbindung von Modellen als Regulierungsansatz für ein schwerwiegendes Risiko im Zusammenhang mit der Technologie des maschinellen Lernens eingeführt: das Risiko einer nicht kontrollierten und potenziell schädlichen sekundären Verwendung trainierter Modelle. Die rechtliche Umsetzung zur Abschwächung dieses Risikos sollte sich auf staatliche, verwaltungsrechtliche Strukturen einer effektiven Aufsicht konzentrieren, statt Individuen die Durchsetzung von subjektiven Rechten zu überlassen.

Die folgenden Fragen einer konkreten rechtlichen Umsetzung konnten in diesem Beitrag nur aufgeworfen werden. Unser Vorschlag zielt zunächst auf die Etablierung eines demokratischen Diskurses über den legitimen Zweck des Einsatzes von KI. Jede Anwendung mächtiger KI-Modelle sollte verschiedene Interessengruppen einbeziehen, um bestimmte Kategorien und Normen für die Zweckbindung von Modellen festzulegen. Während eine detaillierte Positivliste von erwünschten Zwecken und die zugrunde liegenden ethischen Prinzipien zur Beurteilung nachträglicher Zweckerweiterungen Gegenstand einer separaten Ausarbeitung sind, war es in diesem Beitrag unser Ziel, die konzeptuelle Idee eines Regulierungsansatzes für trainierte Modelle hinsichtlich einer Kontrolle ihrer Verwendungszwecke vorzustellen. Im Folgenden skizzieren wir einige Eckpunkte einer rechtlichen Umsetzung der Zweckbindung von Modellen.

Die alleinige Berufung auf die Rechte der einzelnen Betroffenen ist unzureichend, um die mit KI verbundenen und insbesondere aus der unkontrollierten Wiederverwendung von Modellen resultieren Risiken zu adressieren. Daher sollte die Zweckbindung für KI-Modelle nicht nur durch Maßnahmen zum Schutz des Individuums durchgesetzt werden wie sie in der DSGVO enthalten sind, sondern auf einer systemischen Ebene verankert werden. Dieser Ansatz ermöglicht ein Regulierungssystem, das datenmächtigen Akteuren besondere Verpflichtungen auferlegt, die strukturell Datenschutz und Privatsphäre verletzen, auch außerhalb des Wettbewerbsrechts.[85] Akteure, die grundsätzlich und strukturell Verantwortungsstrukturen bedrohen und untergraben, sollten besonderen Anforderungen unterworfen werden. Die Implementierung dieses systemischen Ansatzes würde zudem die Privilegien für (nicht kommerzielle) statistische Zwecke und die wissenschaftliche Forschung (vgl. Art. 5 Abs. 1 Buchst. b DSGVO, Art. 89 DSGVO) über die Grenzen der DSGVO hinaus ausweiten.

Der Vorschlag eines Mechanismus zur demokratischen Kontrolle der Sekundärnutzungszwecke für gesellschaftlich riskante KI-Modelle, insbesondere wenn dieser Mechanismus effektiver sein soll als der bestehende Datenschutzgrundsatz der Zweckbindung für die Verarbeitung personenbezogener Daten, ist keineswegs radikal. Denn ein ähnlicher Ansatz wird bereits in der vorgeschlagenen Verordnung über den Europäischen Gesundheitsdatenraum (EHDS – European Health Data Space) verfolgt.[86] Der EHDS-Vorschlag definiert eindeutige Zwecke für die Verarbeitung elektronischer

[85] Zur Wechselbeziehung zwischen Datenschutzrecht und Wettbewerbsrecht siehe EuGH Urt. v. 4.7.2023 – C-252/21 = ZD 2023, 664 – Meta v Bundeskartellamt.; *Hacker*, Manipulation by algorithms. Exploring the triangle of unfair commercial practice, data protection, and privacy law, Eur. Law J. 2021, 1; *Lynskey/Costa-Cabral*, Family ties: The intersection between data protection and competition in EU law, Common Mark. Law Rev. 2017; *Ruschemeier*, Competition law as a powerful tool for effective enforcement of the GDPR, Verfassungsblog 2023, abrufbar unter https://verfassungsblog.de/competition-law-as-a-powerful-tool-for-effective-enforcement-of-the-gdpr/.

[86] KOM(2022) 197 endg.; das Ergebnis der Verhandlungen im Trilog wurde vom Rat veröffentlicht: EUCO(2024) 7553.

Gesundheitsdaten zur Sekundärnutzung in Art. 34 EHDS und nennt zusätzlich verbotene sekundäre Datenverwendungszwecke in Art. 35 EHDS. Die "Positivliste" umfasst zum Beispiel Zwecke, die Tätigkeiten im öffentlichen Interesse betreffen. Dazu gehören die Überwachung der öffentlichen Gesundheit und der Schutz vor grenzüberschreitenden Bedrohungen Buchst. a, die Unterstützung öffentlicher Stellen Buchst. b, die Erstellung von Statistiken Buchst. c sowie Bildung oder Unterricht Buchst. d. Einerseits erlaubt Art. 34 EHDS ausdrücklich Entwicklungs- und Innovationstätigkeiten für die Qualität und Sicherheit der Gesundheitsversorgung Buchst. e, inklusive der Erprobung und Evaluierung von Algorithmen, auch in medizinischen Geräten, KI-Systemen und digitalen Gesundheitsanwendungen, die einen Beitrag zur öffentlichen Gesundheit oder zur sozialen Sicherheit beitragen Buchst. e i, ii und personalisierte Gesundheitsfürsorge, die darin besteht, den Gesundheitszustand zu bewerten, die Erhaltung oder Wiederherstellung des Gesundheitszustands natürlicher Personen auf der Grundlage der Gesundheitsdaten anderer natürlicher Personen Buchst. h. Andererseits schließt Art. 35 EHDS Zwecke aus, wie das Treffen von Entscheidungen über natürliche Personen oder Gruppen von natürlichen Personen, die sie von den Vorteilen eines Versicherungsvertrags ausschließen oder ihre Beiträge und Versicherungsprämien ändern würden (Art. 35 Buchst. b EHDS), sowie Werbe- oder Marketingaktivitäten, Art. 35 Buchst. c EHDS.

Die Dokumentation und Registrierung von Modellen, die ein besonderes Risiko für die Gesellschaft darstellen, ist ein notwendiger erster Schritt zur Gewährleistung einer Qualitätskontrolle. Im Anschluss daran sollte auf der Ebene der Europäischen Union ein Aufsichtsgremium eingerichtet werden, das Leitlinien zur Identifizierung von Zweckbestimmungen, rechtliche Umsetzung und Durchsetzung etabliert. Die KI-VO sieht etwas Ähnliches in Form von Reallaboren für KI auf mitgliedstaatlicher Ebene vor (vgl. Art. 57-63 KI-VO). Auf diese Weise können Regulierungsbehörden innovative KI-Anwendungen für einen begrenzten Zeitraum testen. Derartige Ansätze sollten jedoch nicht nur der Förderung von Innovation, sondern auch der des regulatorischen Lernens dienen.[87] Es ist von entscheidender Bedeutung, die Verwendungskontexte von KI laufend zu evaluieren, z.B. durch die Entwicklung von Verfahren zur Zweckbindung, die erfolgreich in den politischen und legislativen Prozess eingebunden werden können. Infolgedessen könnten sich sowohl eine sektorspezifische Regulierung[88] als auch die Schaffung einer neuen Aufsichtsbehörde[89] oder eine Anbindung an bestehende Governance-Strukturen der KI-VO, wie bspw. das KI-Büro, als praktikable Optionen erweisen, ebenso wie die Stärkung kollektiver Rechtsschutzmechanismen.[90] Die Überwachung der Einhaltung des erklärten Zwecks ist eine Dauerpflicht. Eine potenzielle konkrete Umsetzung findet sich im Digital Services Act (DSA) durch vertrauenswürdige Hinweisgeber, Transparenz- und Meldepflichten sowie die Überwachung von systemischen Risiken.

Schließlich ist eine Positivliste für die zulässige Nutzung von trainierten Modellen zu entwickeln. Kombiniert mit dem von uns vorgeschlagenen Regulierungsansatz einer Zweckbindung für Modelle ermöglicht dies einen Ausgleich zwischen Interessen der Anbieter und den individuellen, kollektiven und gesellschaftlichen Risiken. Dieser Ansatz würde die Position der Akteure innerhalb der informationellen Machtasymmetrien im Hinblick auf KI und die globalen Auswirkungen ihrer Anwendungen berücksichtigen. Bei der Interaktion zwischen globalen Big-Tech-Unternehmen und Nutzer:innen kann nicht mehr davon ausgegangen werden, dass sich beide Akteure auf gleicher Augenhöhe gegenüberstehen. Daher sollten die gesellschaftlichen Auswirkungen eine viel größere Rolle bei der Risikoklassifizierung spielen als bisher.

---

[87] *Ruschemeier*, Thinking Outside the Box? Regulatory Sandboxes as a Tool for AI Regulation; https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4787008.

[88] *Ohm*, Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization, UCLA Law Rev. 2009, 1701.

[89] *Tutt*, An FDA for Algorithms, Adm. Law Rev. 2017, 83.

[90] *Ruschemeier*, Kollektiver Rechtsschutz und strategische Prozessführung gegen Digitalkonzerne. Viele Davids gegen Goliath?, MMR 2021, 942.

Abstract: This article proposes the concept of purpose limitation for AI models as an approach to effectively regulate AI. Unregulated (secondary) use of specific models creates immense individual and societal risks, including discrimination against individuals or groups, infringement of fundamental rights, or distortion of democracy through misinformation. We argue that possession of trained models, which in many cases consist of anonymous data (even if the training data contains personal data), is at the core of an increasing asymmetry of informational power between data companies and society. Combining ethical and legal aspects in our interdisciplinary approach, we identify the trained model, rather than the training data, as the object of regulatory intervention. This altered focus adds to existing data protection laws and the Artificial Intelligence Act. These are inefficient in preventing the misuse of trained models due to their focus on the procedural aspects of personal data or training data. Drawing on the concept of risk prevention law and the principle of proportionality, we argue that the potential use of trained models by powerful actors in ways that are damaging to society warrants preventive regulatory interventions. Thus, we seek to balance the asymmetry of power by enabling democratic control over where and how predictive and generative AI capabilities may be used and reused.