Das Risiko der Sekundärnutzung trainierter Modelle als zentrales Problem von Datenschutz und KI-Regulierung im Medizinbereich

Rainer Mühlhoff, Prof. Dr. <rainer.muehlhoff@uni-osnabrueck.de>

Ethik und Kritische Theorien der künstlichen Intelligenz

Universität Osnabrück

Manuskript Version 2023-10-16c, im Erscheinen als:

Mühlhoff, Rainer. Das Risiko der Sekundärnutzung trainierter Modelle als zentrales Problem von Datenschutz und KI-Regulierung im Medizinbereich. In: *KI und Robotik in der Medizin – interdisziplinäre Fragen*, Hrsg. Hannah Ruschemeier & Björn Steinrötter, Nomos Verlag, 2023.

1. Einleitung

In diesem pointierten Beitrag möchte ich aus philosophischer und ethischer Perspektive ausführen,[1] was aus meiner Sicht das zur Zeit vielleicht gravierendste Datenschutzrisiko im Kontext von künstlicher Intelligenz (KI) darstellt: die zu wenig regulierte Gefahr missbräuchlicher Sekundärnutzung trainierter KI-Modelle. Während dieses Problem sektorübergreifend besteht, werde ich es am Beispiel von maschinellem Lernen (ML) in der medizinischen Forschung ausführen, weil in diesen Anwendungsfällen der Kontrast einer potentiellen Gefährdung des Gemeinwohls durch sekundäre zweckentfremdende Nutzung der resultierenden Modelle gegenüber der eigentlich im Sinne des Gemeinwohls verfahrenden Forschung besonders deutlich wird.

In der Debatte um Datenethik und Datenschutz im Zusammenhang mit maschinellem Lernen und Big Data liegt der Fokus überwiegend auf der Datenverarbeitung im Input-Stadium: Welche Daten werden als Trainingsdaten erhoben, liegt hierfür eine Rechtsgrundlage (z.B. Einwilligung von Patient:innen und Proband:innen) vor und werden die Daten ausreichend anonymisiert. Dies *sind* enorm wichtige Punkte zum Schutz der Privatsphäre und der Grundrechte der Datensubjekte, die in den Trainingsdaten repräsentiert sind. Dennoch verfehlt die Verengung dieser Debatte auf das Erfassungsstadium der Datenverarbeitungskette ein gravierenderes, weil potenziell sehr viel mehr Menschen betreffendes Datenschutzproblem, welches nicht die Trainingsdaten, sondern die trainierten Modelle und ihre spätere Verwendung betrifft. Denn unter bestimmten, jedoch realistischen Bedingungen können trainierte Modelle ohne

[1] Ich danke Hannah Ruschemeier für unsere intensive, interdisziplinäre Zusammenarbeit zu diesen Themen und ihre detaillierten Kommentare zu diesem Beitrag.

nennenswerte Datenschutzhürden den Kontext ihres ursprünglichen Einsatzzwecks verlassen und für missbräuchliche, diskriminierende, politisch fragwürdige Zwecke zweitverwendet werden (Sekundärdatennutzung).

So könnte zum Beispiel ein Modell zur Einschätzung eines Krankheitsrisikos anhand von Verhaltensdaten, das zu erstrebenswerten Zwecken im Rahmen der medizinischen Forschung entwickelt wurde, von der Versicherungsbranche zur Preisdiskriminierung weiterverwendet (oder an diese weiterverkauft) werden. Medizinische Forschung, die solche Modelle herstellt, umfasst in vielen Fällen die Verarbeitung besonders sensibler Daten. Typischerweise werden diese Daten zum Training von ML-Modellen anonymisiert und – wenn dies mit geeigneten technischen Mitteln geschieht – sind auch die resultierenden Modelle anonyme Daten. In diesen Fällen greifen die Schutzmechanismen des Datenschutzes in Bezug auf diese Modelle nicht mehr. Wer in den Besitz eines solchen Modells gelangt, ist jedoch in der Lage, über beliebige Individuen anhand „wenig" sensibler Daten (z.B. die Verhaltensdaten) die sensiblen medizinischen Information über Individuen abzuschätzen, woraus eine erhebliche Akkumulation von Wissen und Macht resultiert (siehe Beispiele in Abschnitt 2).

Wenn der Zweck des Datenschutzes darin besteht, Machtasymmetrien zwischen Datenverarbeitern und Einzelpersonen/Gesellschaften auszugleichen,[2] ist das Missbrauchsrisiko in Bezug auf trainierte Modelle der größte blinde Fleck in den derzeitigen Regulierungsprojekten. Denn wie ich in Abschnitt 4 ausführen werde, sind ML-Modelle breit skalierbar, sie ermöglichen eine unbemerkte Eskalation der Sensibilität von Daten und neben strukturellen Effekten durch massenhafte Abschätzung sensibler oder persönlicher Informationen für große Menschenmengen (Diskriminierung, soziale Stratifizierung) tritt noch das individuelle Risiko *falscher* Schätzungen durch das Modell, also das Risiko einer individuell „fehlerhaften" Behandlung von Individuen, hinzu. Der bloße *Besitz* eines trainierten Modells bedeutet deshalb eine Akkumulation eines neuen Typs der Informationsmacht, die noch vor der *Anwendung* des Modells auf konkrete Fälle das Ziel von Regulierung und Kontrolle sein sollte. Die derzeitige Fokussierung des Datenschutzes auf das Input-Stadium spielt im öffentlichen Diskurs eine doppelt problematische Rolle: Betroffene lenkt das Versprechen von Anonymisierung und die informierte Einwilligung von der Gefahr der missbräuchlichen *Sekundärnutzung* trainierter Modelle ab. Für Praktiker:innen und Forscher:innen zum Beispiel im Medizinapparat geht die auf das Input-Stadium fokussierte Praxis des Datenschutzes oft mit enormen bürokratischen Auflagen einher, die leicht zur Verunglimpfung des Datenschutzes in der öffentlichen Diskussion als Innovationshemmnis angeführt werden (Müller-Jung, 2023; Nida-Rümelin und Hilgendorf, 2021).

Ich werde im Folgenden das Problem der Sekundärnutzung von ML-Modellen zuerst anhand von zwei Beispielen (Abschnitt 2) und danach allgemein (Abschnitt 3)

[2] Vgl. Lewinski (2009); Rehak (2022); Rost (2018).

formulieren. Ich werde sodann argumentieren, dass das zu regulierende Problem den Ausgleich einer spezifischen Spielart informationeller Machtasymmetrie bedeutet („Vorhersagemacht"), auf die Grenzen des Datenschutzes angesichts dieses Risikos hinweisen und das Prinzip der Risikoprävention ins Spiel bringen (Abschnitt 4). Danach werde ich zwei zusammenhängende Lösungsansätze andeuten (Abschnitt 5): der erste betrifft die ethische Debatte um prädiktive Privatheit, der zweite, darauf aufbauende, betrifft die regulatorische Idee einer Zweckbindung für Modelle.

2. Missbräuchliche Sekundärnutzung: Zwei Beispiele

Um auf das Problem der missbräuchlichen Sekundärnutzung trainierter Modelle zu führen, seien hier zwei fiktive, jedoch realistische Beispielszenarien gegeben.

Szenario A

Eine psychotherapeutische Forschungsgruppe eines Universitätsklinikums verfolgt die Forschungsfrage, ob sich die psychiatrische Diagnosestellung verbessern lässt, indem Audiodaten (Audio-Mitschnitte von Therapiesitzungen) der Patient:innen mittels KI auf Marker für bestimmte psychische Leiden hin ausgewertet werden. Es geht hierbei nicht darum, den Inhalt gesprochener Sprache, sondern die phonetischen und tonalen Aspekte der Sprache auszuwerten („wie" gesprochen wird), um auf psychische Krankheiten zu schließen.[3] Der Zweck dieses Projektes ist, dass Behandelnde durch ein solches KI-Tool in ihrer Diagnosefindung unterstützt werden können. Es handelt sich hierbei um ein Forschungsprojekt mit offener Erfolgsaussicht. Einige Patient:innen willigen ein, dass für dieses Forschungsprojekt Audiomitschnitte ihrer Therapiesitzungen sowie ihre Krankenakten verwendet werden dürfen. Mit diesen Trainingsdaten trainiert das Universitätsklinikum ein ML-Modell, welches anhand der Audiodaten die in den Krankenakten verzeichneten psychiatrischen Diagnosen vorherzusagen lernen soll. Bei diesem Training des Modells werden Anonymisierungsverfahren nach dem aktuellen Stand der Technik eingesetzt[4], und wir wollen davon ausgehen, dass das in dem spezifischen Fall so gut gelingt, dass das trainierte Modell keinerlei Rückschlüsse mehr auf die in den Trainingsdaten enthaltenen individuellen Fälle erlaubt.[5] Die Trainingsdaten werden

[3] Zur Plausibilisierung dieses gänzlich fiktiven Szenarios siehe: Tian u. a. (2023); Ma u. a. (2016); https://www.psychologytoday.com/intl/blog/different-kind-therapy/202211/ai-can-use-your-voice-detect-depression; https://www.npr.org/2022/10/10/1127181418/ai-app-voice-diagnose-disease.

[4] Siehe dazu insbesondere differential privacy in machine learning (vgl. Abadi u. a., 2016; Dwork, 2006), das ggf. mit federated learning kombiniert werden kann, was besonders im Medizinkontext von Interesse ist (Kaissis u. a., 2020).

[5] Das bedeutet also, dass wir hier annehmen, dass zum Beispiel membership inference attacks (Shokri u. a., 2017) und model inversion attacks (Fredrikson u. a., 2015) ausgeschlossen sind.

Obwohl die in Fußnote 4 erwähnten Differential-Privacy-Verfahren mathematisch betrachtet eine rigorose Form der Anonymisierung (innerhalb eines gegebenen „privacy budget") gewähren (vgl. Dwork, 2011; Nissim u. a., 2018), sind technische Anonymisierungsverfahren grundsätzlich mit Vorsicht zu betrachten, weil die Geschichte der Informatik zahlreiche Beispiele für das rückwirkende Durchbrechen von Anonymisierungsverfahren kennt [vgl. exemplarisch

überdies direkt nach der Verwendung wieder gelöscht, so dass am Ende nur ein trainiertes Modell übrig bleibt.

Angenommen, das Forschungsprojekt wäre erfolgreich und die Forschungsgruppe hätte dann ein trainiertes ML-Modell, welches anhand von Audiodaten beliebiger Menschen Prognosen über deren potentielle psychiatrische Krankheiten ableiten könnte. Das trainierte Modell selbst kann als ein Datensatz aufgefasst werden, der aus den internen Parametern des Modells besteht (z.B. die Gewichte der Synapsen im Fall eines simulierten neuronalen Netzes etc.); ich verwende fortan hierfür die Bezeichnung *Modelldaten.* Wenn beim Training geeignete Anonymisierungstechniken verwendet wurden, haben die Modelldaten keinen Personenbezug mehr; sie würden als anonyme Daten gelten. Weil die Verarbeitung anonymer Daten nicht in den Geltungsbereich der DSGVO fällt, wäre die Forschungsgruppe nun datenschutzrechtlich in der Lage, eine Kopie des trainierten Modells frei zu zirkulieren.

Angenommen, ein Unternehmen, welches KI-Software für die Unterstützung von Job-Auswahlverfahren entwickelt, interessiert sich für dieses trainierte Modell.[6] Das Unternehmen verwendet bereits KI-Verfahren zur Auswertung der Stimme der Bewerber:innen in digital durchgeführten Bewerbungsverfahren. Das trainierte Modell der Forschungsgruppe würde es dem Unternehmen ermöglichen, die automatisierte Auswertung der Audiodaten von Bewerber:innen noch um eine Bewertung des Risikos psychischer Erkrankungen der Bewerber:innen zu ergänzen.

Diese Zweitverwendung des psychodiagnostischen Modells birgt ein erhebliches Risiko der Diskriminierung von Bewerber:innen mit psychiatrischen Krankheiten. Es gibt zur Zeit keine datenschutzrechtlichen Hürden, die verhindern, dass die Modelldaten, also das trainierte Modell, dieser Sekundärnutzung zugeführt werden kann (siehe genauer Abschnitt 3). Da der Primärzweck des Modells ein gesellschaftlich wünschenswerter ist, insofern er die medizinische Behandlung verbessern soll, steht diesem Vorhaben dennoch das erhebliche Risiko missbräuchlicher Sekundärnutzung der entstehenden Modelle

Sweeney1997WeavingTechnology; Ohm (2010); Narayanan und Shmatikov (2008); Gymrek u. a. (2013)]. Daher sollte die Verwendung technischer Anonymisierungsverfahren nur in speziellen Situationen zur *normativen* Beurteilung eines Zusammenhangs belastet werden. In dem *hier* vorliegenden Kontext ist die Situation jedoch genau umgekehrt – und dies birgt Potenzial für Missverständnisse: *Wenn* die Anonymisierung mittels technischer Verfahren wirklich gelingt (wie hier angenommen), dann stellt dies den hinsichtlich des Risikos der Sekundärnutzung trainierter Modelle rechtlich weniger gut abgedeckten und somit gesellschaftlich riskanteren Fall dar, gerade *weil* für die Verbreitung solcher Modelle dann nicht die DSGVO greift. Um genau diese Konstellation fehlender regulatorischer Auflagen zu adressieren, fokussiere ich in diesem Beitrag auf diesen Fall. Sollten die Modelldaten in einem konkreten Fall keine anonymen Daten sein, handelt es sich um einen Mischfall: Möglicherweise greift dann die DSGVO in Bezug auf die Verarbeitung der Modelldaten; die hier vorgeschlagenen Regulierungsansätze würden ebenso greifen. Zu betonen ist, dass es das Ideal jeder technischen Vorgehensweise ist, anonyme Modelldaten herzustellen; fehlende Anonymisierung wäre also kaum im Interesse der Betreiber und würde auf fragwürdige Arbeitsweise hindeuten.

[6] Zur Plausibilisierung dieses rein fiktiven Vorschlags: https://www.audeering.com/technology/health-ai/, https://www.peakprofiling.com/medical-voice-analytics/, https://arstechnica.com/gadgets/2018/10/amazon-patents-alexa-tech-to-tell-if-youre-sick-depressed-and-sell-you-meds/.

entgegen. Bisher bestehen keine Hindernisse, entsprechende Modelle zu anderen Zwecken zu kommerzialisieren oder aus anderen Gründen anderen Akteuren und Zwecken zugänglich zu machen. Es wäre die Aufgabe effektiver Regulierung, die Bindung des entstehenden Modells an den Primärzweck sicherzustellen, damit Innovationen durch KI in der medizinischen Diagnostik vorbehaltlos gefördert werden können.

Szenario B

Während wir es im ersten Szenario mit einer öffentlich-rechtlichen Forschungseinrichtung (Universitätsklinikum) zu tun hatten, betrachten wir nun ein Szenario, welches einen privatwirtschaftlichen Primärakteur betrifft. Das folgende Szenario wurde ausführlicher in der Studie Mühlhoff und Willem (2023) besprochen und wird hier nur rekapituliert.

Eine medizinische Forschungseinrichtung (egal ob öffentlich-rechtlich oder privatwirtschaftlich) entwickelt ein neues Therapieverfahren für Typ 2 Diabetes (fortan: T2D) und sucht für eine klinische Studie Proband:innen mit dieser Krankheit. Um die Zielgruppe effizient zu erreichen, setzt es dazu Werbung auf Social Media ein, z.B. auf Facebook. Eine Facebook-Werbeanzeige bewirbt den neuen Therapieansatz und enthält einen Button „Sign Up", mit dem man sich als T2D-Patient:in direkt für die Studie anmelden kann.[7]

Der Auftraggeber versieht die Facebook-Anzeige, wie im Bereich der Online-Werbung üblich, zunächst mit bestimmten targeting-Kriterien, von denen er sich verspricht, möglichst genau die Zielgruppe zu erreichen. Die Plattform (hier: Facebook) verwendet einen targeting-Algorithmus, der, z.B. auf Grundlage eines ML-Modells und unter Berücksichtigung der targeting-Kriterien des Auftraggebers, entscheidet, welche Nutzer:innen diese Anzeige zu sehen bekommen. Der Algorithmus berechnet dazu für jede Nutzer:in der Plattform die Wahrscheinlichkeit, dass die betreffende Werbung für sie „relevant ist"; aus wirtschaftlichen Gründen wird die Anzeige dann genau den Nutzer:innen gezeigt, bei denen diese Wahrscheinlichkeit am größten ist.

Um das targeting-Verfahren zu optimieren, wird im Bereich der Online-Werbung die Reaktion der einzelnen Nutzer:innen auf gesehene Anzeigen mittels Tracking-Technologien ausgemessen. Insbesondere registrieren Plattformen, ob die Nutzer:in eine Anzeige zum Beispiel anklickt, abspeichert oder weiterleitet. Diese Daten sind Signale dafür, dass die Anzeige für die Nutzer:in interessant und relevant ist. Diese Daten können als Trainingsdaten für die fortlaufende Verfeinerung des targeting-Modells eingesetzt werden. So beobachteten Forscher:innen im Gesundheitsbereich, dass sich die Genauigkeit des Targetings einer bestimmen Anzeige über die ersten Tage der Laufzeit gravierend verbessert – mutmaßlich durch die Lernkurve des targeting-Algorithmus, der anhand der

[7] Siehe hier für den Screenshot einer solchen im Jahr 2018 gelaufenen Facebook-Anzeige der Firma Trialfacts: https://trialfacts.com/case-study/effective-clinical-trial-recruitment-plan-narrowing-field-from-500-to-24/

Feedback-Daten (tracking des enagements mit der Anzeige) treffsicherer die Nutzer:innen-Profile detektieren kann, für die die Anzeige „relevant" ist (cf. Borodovsky u. a., 2018; Mühlhoff und Willem, 2023).

Nach einigen Tagen Laufzeit besitzt Facebook also ein stark verbessertes ML-Modell für das Targeting der T2D-Anzeige. Dieses Modell stellt anhand der Facebook-Profildaten beliebiger Nutzer:innen eine Prognose darüber aus, ob die Nutzer:in die T2D-Anzeige anklicken wird oder nicht; dieses Modell dient primär dem Zweck, die Werbung besonders effizient an die Zielgruppe zu vermitteln. Das Modell ist aber zugleich ein Modell, welches direkt oder indirekt medizinische Informationen über beliebige Facebook-Nutzer:innen abschätzen kann – nämlich, ob sie an T2D leiden. Es ist davon auszugehen, dass nur die wenigsten von T2D betroffenen Nutzer:innen diese Information freiwillig auf der Plattform angeben würden. Das als Beiprodukt des Werbe-Targetings trainierte Modell erlaubt es der Plattform jedoch, eine Schätzung dieser Information über beliebige Nutzer:innen anzufertigen, auch über jene, die die Anzeige gar nicht gesehen haben, oder jene, die sich erst zukünftig auf Facebook anmelden werden.

Nehmen wir wie in Szenario A an, dass das targeting-Modell mit Anonymisierungstechniken trainiert wurde und die Modelldaten somit anonyme Daten sind. Dann gibt es keine datenschutzrechtlichen Hürden für Facebook, das trainierte targeting-Modell anderweitig zu zirkulieren oder zu verkaufen. Auf hier kommen zahlreiche missbräuchliche und gesellschaftliche schädliche Sekundärnutzungsanliegen in Betracht. Für solche Modelle könnten sich zum Beispiel Unternehmen interessieren, die KI-Tools für Auswahlverfahren entwickeln, um Kandidat:innen, von denen Social Media Daten vorliegen (das ist bei Job-Bewerbungsverfahren nicht ungewöhnlich) hinsichtlich ihres Gesundheitsrisikos zu klassifizieren. Auch hier wäre die Diskriminierung von mutmaßlich kranken Patient:innen die Folge der missbräuchlichen Sekundärnutzung des trainierten Modells.

3. Missbräuchliche Sekundärnutzung: allgemeine Problembeschreibung

Szenarien A und B stellen Beispiele für das Risiko einer Sekundärnutzung trainierter ML-Modelle dar. Mit diesem Risiko ist eine Konstellation adressiert, die allgemein durch folgende grundsätzliche Merkmale charakterisiert ist:

1. Aus Trainingsdaten einer bestimmten Anzahl von Personen P_1 bis P_n, die in einem bestimmten Datenverarbeitungskontext gewonnen werden (z.B. klinische Forschung und Behandlung), wird ein ML-Modell erstellt. Dieses Modell kann, sobald es trainiert ist, mit einer bestimmten Genauigkeit über beliebige Fälle unbekannte Informationen (Szenario A: psychiatrische Diagnose; Szenario B: ob eine Person an T2D erkrankt ist) anhand von verfügbaren Informationen (Szenario A: Audiomitschnitt; Szenario B: Social Media Nutzungsdaten) abschätzen.

2. Die Herstellung des Modells ist datenschutzrechtlich unbedenklich, da die Verarbeitung der Trainingsdaten (Erhebung) auf einer geeigneten Rechtsgrundlage erfolgt (häufig: informierte Einwilligung der Datensubjekte oder „berechtigtes Interesse" der datenverarbeitenden Organisation).

3. Um für unsere Diskussion den Fall mit größtmöglichem Risikos ins Auge zu fassen, sollten wir ferner von der Annahme ausgehen, dass die Modelldaten selbst anonym sind und daher nicht in den Anwendungsbereich der DSGVO fallen. Dass Modelldaten anonym sind, ist eine realistische Annahme, wenn aktuelle Anonymisierungstechniken im Kontext des Maschinellen Lernens verwendet werden (Annahme 3 ist daher ein technischer Punkt, vgl. Fußnoten 4 und 4).

4. Im primären/ursprünglichen Datenverarbeitungskontext verfolgt die Herstellung des ML-Modells einen vertretbaren oder sogar förderungswürdigen Zweck, zum Beispiel die Verbesserung der medizinischen Diagnostik (Szenario A) oder die Erforschung neuer Medikamente (Szenario B).

5. Es sind jedoch Sekundärnutzungen der einmal erstellten Modelle denkbar, die mit erheblichen Risiken für Individuen und die Gesellschaft einhergehen. Solche Sekundärnutzungen können zum Beispiel die Diskriminierung oder soziale Sortierung beliebiger Individuen (Szenarien A und B: aufgrund von Krankheitsprognosen) ermöglichen. Man denke etwa an die Sekundärnutzung zu Scoring- und Klassifikationszwecken beim Zugang zu Ressourcen wie Arbeit, Kredit, Bildung, Immobilien (O'Neil, 2016). Insbesondere denke man an Sekundärnutzungsbereiche wie die KI-gestützte Durchführung von Auswahlverfahren oder die Kreditvergabe, in der häufig eine Einwilligung der Bewerber:innen für die Anwendung von KI-Verfahren auf ihren Fall vorliegt.

Konstellationen dieser Art treten bei der Anwendung von ML-Verfahren *sehr häufig* auf. Charakteristischerweise steht bei der Beurteilung und Thematisierung des Einsatzes von KI der primäre Zweck der Anwendung (Punkt 4) diskursiv im Mittelpunkt. Zum Beispiel werden in der Debatte um den Einsatz von KI in der medizinischen Forschung die wünschenswerten Verbesserungspotenziale von Diagnostik und Therapie betont sowie die Risiken falscher Diagnosen oder von Biases in den Trainingsdatensätzen. Das mutmaßlich deutlich gravierendere Risiko einer missbräuchlichen Zweitnutzung der resultierenden Modelle wird hingegen nicht thematisiert. Auch bei der Information von Patient:innen, die in die Verarbeitung ihrer Daten als Trainingsdaten einwilligen, wird es nicht erwähnt und bei der ethischen oder politischen Bewertung des Vorhabens nur selten einbezogen.

Differenzierung der Verarbeitungsschritte und Datentypen

Die möglicherweise missbräuchliche Sekundärnutzung trainierter Modelle ist nicht nur diskursiv randständig (fehlende öffentliche Diskussion und fehlendes Bewusstsein der

Stakeholder), sondern zugleich stehen ihr keine wirksamen regulatorischen Hürden entgegen. Dies wurde ausführlich in Bezug auf die DSGVO argumentiert (Mühlhoff und Ruschemeier, 2022). Kurz gefasst lässt sich diese Regulierungslücke durch einen Blick auf die unterschiedlichen beteiligten Datenverarbeitungsschritte und Datentypen plausibilisieren, die in Szenarien einer missbräuchlichen Zweitverwendung involviert sind. Die folgende Tabelle zeigt eine schematische Übersicht:

Schritt	**verarbeitete Daten**	**regulatorische Beschränkung**
1: Training des Modells	Daten einer bestimmten Personen- oder Fallgruppe werden als **Trainingsdaten** erfasst. Z.B. Szenario A: Patientendaten, Szenario B: Tracking-Daten der Social Media Nutzer:innen, die die Anzeige gesehen haben. Als Produkt von Schritt 1 entsteht ein trainiertes Modell, repräsentiert durch die **Modelldaten.** Im riskantesten Fall[8] sind die Modelldaten anonym.	DSGVO greift, wenn die Trainingsdaten personenbezogen sind. In geläufiger Vorgehensweise wird als Rechtsgrundlage zur Verarbeitung der Trainingsdaten eine Einwilligung eingeholt, auf „berechtigtes Interesse“ abgestellt und/oder eine Anonymisierung der Daten vorgenommen.
2: Zirkulation des trainierten Modells	Die **Modelldaten** werden in den Bereich einer Sekundärnutzung kopiert oder transferiert. Die Modelldaten können dort ggf. zum Bestandteil eines größeren Modells werden, so dass sie nicht in identischer Form zweitverwendet werden, sondern in ein umfassenderes Vorhersage- oder Scoring-	Hier greift keine Regulierung, sofern die Modelldaten anonym sind.[10]

[8] Wären die Trainingsdaten nicht anonym sondern personenbezogen, würde in Schritt 2 mitunter die DSGVO greifen. Da es technische Methoden gibt, die Trainingsdaten vor dem Training oder während des Trainings des Modells zu anonymisieren, müssen wir, um von einem maximal potenten Angriffsszenario auszugehen und den vollen Umfang des Risikos einer Regulierungslücke zu erkennen, davon ausgehen, dass diese Techniken im Allgemeinen verwendet werden. Das Risiko ist bei anonymen Modelldaten deshalb maximal, weil dann in Schritt 2 keine Verarbeitung personenbezogener Daten vorliegt, die in den Geltungsbereich der DSGVO fällt.

[10] Ob für den Vorgang der Anonymisierung der Daten selbst eine rechtliche Grundlage erforderlich ist, ist umstritten, weil die DSGVO selbst keine Anwendung auf anonymisierte Daten findet. In vielen Fällen willigen Datensubjekte allerdings in die Anonymisierung ihrer Daten ein, da sie vermeintlich das Datenschutzniveau erhöht. Dazu vertiefend: Mühlhoff und Ruschemeier (2022)..

	Modell des Sekundärnutzungsakteurs integriert werden.[9]	
3: Anwendung des Modells	Nutzung des trainierten Modells im Kontext der Zweitverwendung, um über einen neuen Fall eine Prognose zu stellen. Über den neuen Fall liegen **Hilfsdaten** vor (Szenario A: Audiodaten, Szenario B: Social Media Nutzungsdaten). Diese dienen als Eingabedatum des trainierten Modells; das Ausgabedatum ist eine **Prognose, Klassifikation, oder ein Scoring** über den zu beurteilenden Fall. Zu beachten ist, dass das Modell hier auf beliebige Dritte angewendet wird, also auf Personen, die i.A. nicht in den Trainingsdaten von Schritt 1 enthalten sind.	Hier greift die DSGVO, wenn sich die erstellten Prognosen, Klassifikation oder Scorings auf Personen beziehen. In typischen Kontexten (z.B. Bewerbungsverfahren oder Versicherungsbewerbungen) wird als Rechtsgrundlage die Einwilligung des zu beurteilenden Individuums eingeholt. In vielen dieser Kontexte hat das beurteilte Individuum kaum realistischen Spielraum, die Einwilligung zu verweigern.

Diese Kette von drei Datenverarbeitungsschritten zeigt regelmäßig eine regulatorische Lücke in Schritt 2. Modelle, die anhand von sensiblen Trainingsdaten legal erstellt werden und somit auch potenziell sensible Informationen über beliebige Dritte abschätzen können, unterliegen als anonyme Modelldaten keinen rechtlichen Beschränkung hinsichtlich ihrer Zirkulation in andere Anwendungsbereiche. Während die Modelle im primären Nutzungskontext häufig einem wünschenswerten „guten" Zweck dienen (z.B. Verbesserung der Gesundheitsversorgung), unterliegen sie einem hohen Risiko, in anderen Anwendungsbereichen missbräuchlich verwendet zu werden, zum Beispiel in Anwendungen, durch die es zu Diskriminierung, social sorting, Manipulation kommt. Die bloße Existenz der Modelle, kombiniert mit dem Zustand fehlender Regulierung und Kontrolle über ihre weitere Zirkulation, stellt daher ein enormes gesellschaftliches Risiko dar.

4. Datenschutz, Datenmacht und Risikoprävention

Das Risiko der Sekundärnutzung trainierter Modelle ist erheblich und sollte einen Hauptfokus der Regulierung von künstlicher Intelligenz bilden. Die Gravität des Problems geht auf das Zusammenspiel zweier Umstände zurück: Erstens können Vorhersagemodelle,

[9] Zum Beispiel in Szenario A könnte das Unternehmen, welches anhand von Audioaufnahmen über die Eignung von Bewerber:innen entscheidet, nicht explizit anhand des erworbenen Modells für psychiatrische Diagnosen eine Krankheitsprognose berechnen, sondern dieses erworbene Modelle als Teilkomponente in ein größeres Modell einbetten, welches insgesamt direkt eine Auswahlempfehlung ausgibt, die dann nur implizit die Prognose über psychische Leiden berücksichtigt.

die schwer zugängliche, sensible oder persönliche Informationen aus leichter zugänglichen Daten wie zum Beispiel Verhaltensdaten, Trackingdaten, Audiodaten oder sozioökonomischen Daten abzuschätzen erlauben, in zahlreichen Kontexten zur Diskriminierung (Preisdiskriminierung z.B. bei Versicherungen), sozialen Sortierung (z.B. Vorabauswahl bei Bewerbungsprozessen, im Bildungssystem, im Sicherheitsapparat) oder Manipulation (z.B. personalisierte politische Wahlwerbung mit demokratieverzerrenden Effekten) verwendet werden [vgl. Barocas und Selbst (2016); Hildebrandt und Gutwirth (2008); O'Neil (2016); Bozdag (2013); Zarsky (2019); Eubanks (2017); Mann-Matzner2019; Mü2020:DZPhil]. Zweitens ist die Zirkulation und Sekundärnutzung trainierter Modelle aktuell nicht reguliert. Dies führt zu einer Vertiefung bestehender informationeller Machtasymmetrien in zweifacher Hinsicht. Zum einen haben Akteure, die über die notwendigen Datengrundlagen und die technischen Möglichkeiten verfügen, nicht nur die Möglichkeit, Vorhersagemodelle für ihre Zwecke zu erstellen, sondern diese auch an Dritte weiterzuverkaufen, was die Nutzungs- und Verwertungsmöglichkeiten multipliziert. Zum anderen vertieft sich die Machtasymmetrie zwischen den Akteuren, welche die Modelle zu Sekundärzwecken nutzen, und den betroffenen Datensubjekten. Denn letztere sind in Bewerbungsverfahren, bei Kreditanträgen und als Nutzer:innen digitaler Medien ohnehin in der Situation, dass ihnen viel weniger Informationen zur Verfügung stehen als den entscheidenden Akteuren.

Zahnloser Datenschutz

Der paradoxale Charakter dieser Konstellation zeigt sich besonders wenn man bedenkt, dass die Anwendung von KI-Methoden, zum Beispiel in der klinischen Forschung mit Patientendaten, bereits jetzt erheblichen Datenschutzhürden unterliegt und von vielen Akteuren in der Praxis als überreguliert empfunden wird.[11] Wir haben es in der Umsetzungspraxis der DSGVO mit einem Datenschutz zu tun, der mit erheblichen prozeduralen Auflagen bei der Behandlung der „Input"-Daten von Verarbeitungsschritt 1 (also der Trainingsdaten) einhergeht, jedoch zahnlos in Bezug auf ein deutlich schwerwiegenderes Datenschutzproblem ist, nämlich das Risiko der (evtl. gar nicht intendierten) Zweitverwendung der resultierenden Modelle (Schritt 2).

Ich bezeichne dieses Risiko als *schwerwiegender,* weil alle mit der Verarbeitung der Trainingsdaten verbundenen Risiken für die Grundrechte der darin enthaltenen Subjekte weiterhin vorliegen und zusätzlich die folgenden drei Risiken einbezogen werden müssten: (1) Breite Skalierbarkeit: Die resultierenden Modelle lassen sich auf beliebig viele Individuen anwenden (und insbesondere auf *deutlich mehr* Individuen, als in den Trainingsdaten abgebildet sind). (2) Eskalation der Sensibilität: Die Anwendung des Modells erlaubt es, über *beliebige Dritte* ähnlich sensible Informationen abzuschätzen, wie über die Trainingsdatensubjekte hineingesteckt wurden. (3) Probabilistische Behandlung:

[11] Vgl. etwa Müller-Jung (2023) und https://www.cr-online.de/blog/2019/05/21/datenschutz-das-zuegellose-recht-teil-ii-der-datenpaternalismus/.

Zu den strukturellen Effekten durch massenhafte Abschätzung sensibler oder persönlicher Informationen für große Menschenmengen (Diskriminierung, soziale Stratifizierung) tritt noch das individuelle Risiko *falscher* Schätzungen durch das Modell, also das Risiko einer individuell „fehlerhaften" Behandlung der Individuen in Schritt 3.[12]

Das Ziel im Umgang mit dem Risiko missbräuchlicher Sekundärnutzung muss eine Regulierung sein, die trennscharf die wünschenswerten, intendierten Datenverarbeitungsvorgänge von den missbräuchlichen unterscheiden kann, um die „guten" zu ermöglichen und die „schlechten" zu verhindern. Hierzu ist es erstens erforderlich, das regulatorische Vorgehen nicht auf das Input-Stadium (Trainingsdaten, die in Schritt 1 hineingesteckt werden) zu fokussieren, sondern auf die in Schritt 2 verarbeiteten anonymen Modelldaten, die aus den Input-Daten (vermittels der Trainingsprozedur) lediglich abgeleitet wurden. Zweitens ist es erforderlich, den Zweck der Herstellung oder Anwendung trainierter Modelle selbst ins Auge zu fassen, ihn ethisch zu bewerten und einer Abweichung von diesem Zweck vorzubeugen. Was wünschenswerte Zwecke sind, kann im Rahmen *dieses* Beitrags nicht geklärt werden, vielmehr ist es mein Anliegen, dass überhaupt eine regulatorische Struktur geschaffen werden muss, die bei den Zwecken trainierter Modelle (und nicht allein der Trainingsdaten) ansetzt – auf diesen Punkt gehe ich weiter unten im Kontext von „Zweckbindung für Modelle" noch einmal ein.

Weil die Notwendigkeit von Regulierung aus der der Aufgabe des Staates resultiert, Machtungleichheiten zu begrenzen, werde ich nun die These ausführen, dass mit der Herstellung von Vorhersage-, Scoring- und Klassifikationsmodellen eine spezifische Form der informationellen Machtasymmetrie einhergeht.

Vorhersagemacht als aktuellste Form der Datenmacht

Um den Umfang des Risikos der missbräuchlichen Zweitnutzung trainierter Modelle theoretisch und philosophisch zu begründen, ist es sinnvoll, den Besitz solcher Modelle als einen *Machtfaktor* aufzufassen (vgl. Lynskey, 2019). Wer über Vorhersagemodelle verfügt, besitzt das, was ich als „Vorhersagemacht" bezeichne, nämlich das Vermögen, Vorhersagen über beliebige Individuen erstellen zu können (Mühlhoff, 2023b; vgl. Mühlhoff und Ruschemeier, 2022). Vorhersagemacht konzentriert sich meist dort, wo Daten aggregiert werden. Die Akkumulation von Vorhersagemacht ist deshalb eng mit der Macht von Plattformunternehmen verknüpft und bildet eine globale Struktur, die es erfordert, die strukturellen Auswirkungen von KI sozialtheoretisch zu beleuchten.

[12] Für das Risiko der missbräuchlichen Zweitverwendung ist es unerheblich, ob abgeschätzte Informationen korrekt oder nicht korrekt sind (vgl. Mühlhoff, 2021). Denn eine unfaire Diskriminierung, Manipulation oder soziale Sortierung ist politisch und ethisch bedenklich, ob sie auf richtigen oder falschen Informationen beruht. Mein Argument an dieser Stelle ist somit, dass neben den strukturellen Risiken bei missbräuchlicher Zweitverwendung trainierter Modelle (z.B. Diskriminierung *vieler* Individuen, soziale Sortierung ganzer Kohorten, Manipulation des Wahlverhaltens *Vieler*, etc.) noch das Risiko der individuellen Fehlbehandlung aufgrund von falsch abgeschätzten Informationen hinzutritt.

Der Begriff „Macht" ist vielseitig und in Sozialtheorie und Sozialphilosophie umfochten. Es ist im vorliegenden Rahmen weder möglich noch erforderlich, hier ins Detail zu gehen (vgl. D'Ignazio und Klein, 2020; Mühlhoff, 2020, 2023a), statt dessen genügt es, ein begriffliches Spektrum abzustecken. Nach einer klassischen Definition von Max Weber bedeutet Macht „jede Chance, innerhalb einer sozialen Beziehung den eigenen Willen auch gegen Widerstreben durchzusetzen, gleichviel worauf diese Chance beruht" (Weber, 1990: 28). Nach einer anderen klassischen, auf Michel Michel Foucault zurückgehenden und dem Weber'schen Verständnis oft entgegengestellten Definition bedeutet Macht „handelndes Einwirken auf Handeln" (Foucault, 2007: 96).[13] Ein Hauptunterschied besteht darin, dass die erste Definition Macht bei bestimmten Akteuren lokalisiert, während sie anderen fehlt (man kann Macht also besitzen). Die zweite Definition hingegen ist die Vorlage für einen strukturellen Machtbegriff, nach dem Macht im (mitunter nicht intendierten und geplanten) Zusammenspiel vieler Akteure liegt und sich auf die subtile Rahmung individueller Handlungshorizonte in multifaktoriellen sozio-technologischen Kontexten (Handlungsnetzwerken) bezieht.[14]

Beide Elemente sind im Fall der Vorhersagemacht gegeben. Erstens ermöglicht es Vorhersagemacht, persönliche Informationen auch gegen den Willen und Widerstand der Betroffenen über sie abzuschätzen. Solche Vorhersagen können zum Beispiel der Kandidat:in in einem Bewerbungsverfahren die Möglichkeit nehmen, ihren Disposition zu depressiven Verstimmungen zu verbergen (Szenario A). Zweitens können Vorhersagemodelle dazu benutzt werden, den Handlungshorizont großer Menschenmengen unterschwellig zu strukturieren. Facebook beispielsweise könnte das Modell zur Abschätzung von Typ 2 Diabetes (Szenario B) auch beim Targeting weiterer Werbeanzeigen verwenden, etwa für Stellenanzeigen oder Lebensversicherungen, die Menschen mit einem (vorhergesagten und den Betroffenen vielleicht selbst unbekannten) Diabetes-Risiko dann nicht angezeigt werden. In diesem Fall wird also anhand abgeschätzter Informationen der Informations- und Handlungshorizont zahlreicher Nutzer:innen manipuliert.

Aufgrund der enormen Verbreitung und wirtschaftlichen Bedeutung von prädiktiven KI-Anwendungen (prädiktiver Analytik) im Zusammenhang mit Konsumenten- und Nutzungsdaten stellt Vorhersagemacht aktuell eine der wichtigsten Spielarten informationeller Machtasymmetrie zwischen datenverarbeitenden Akteuren und Individuen dar. Durch die Konzeptualisierung des Problems als ein Machtphänomen kommt eine entscheidende Eigenschaft des damit verbundenen Regulierungsanliegens zum Vorschein: Was reguliert werden muss, ist ein *Potenzial,* ein *Vermögen* zu bestimmten (missbräuchlichen) Handlungen oder Handlungseffekten [vgl. Mü-Ru2022: 43]. Eine

[13] „In Wirklichkeit sind Machtbeziehungen definiert durch eine Form von Handeln, die nicht direkt und unmittelbar auf andere, sondern auf deren Handeln einwirkt. Eine handelnde Einwirkung auf Handeln, auf mögliches oder tatsächliches, zukünftiges oder gegenwärtiges Handeln." (Foucault, 2007: 96)

[14] Zur Netzwerk-Semantik in Bezug auf Macht bei Foucault vgl. Foucault (1983): 95–97.

Regulierung die, wie die DSGVO, erst in Schritt 3 greift, wenn das zweckentfremdete Modell auf eine *bestimmte* Bewerber:in eines Job-Auswahlverfahrens angewendet wird, greift zu spät. Denn bereits das Potenzial des Missbrauchs, das durch den Besitz des Vorhersagemodells gegeben ist, und nicht erst der vollzogene Missbrauch, sollte unter Kontrolle gebracht werden. Das ist in der vorliegenden Konstellation deshalb so entscheidend, weil von dem bloßen Besitz eines Vorhersagemodells in den falschen Händen ein *breit gestreutes* Diskriminierungspotenzial ausgeht, also ein Potenzial der sekundären Verwendung, das eine Vielzahl nicht näher bestimmbarer Individuen treffen könnte und mit den Zwecken der ursprünglichen Herstellung des Modells nicht mehr viel gemein hat. Diese latente allgemeine Gefahr, die mehr ist als die Summe der einzelnen tatsächlich vollzogenen Diskriminierungshandlungen, konstituiert die Machtposition der Akteure, die das fragliche Modell in einem bestimmten Kontext zu einem bestimmten Zweck zur Anwendung bringen könnten. Macht ist hier also stets als ein *Vermögen* gegeben, unabhängig von seiner Aktualisierung in manifesten Taten oder Handlungen. Regulierung muss direkt bei der Kontrolle und Beschränkung dieser Macht-als-Vermögen einsetzen, denn unreguliert und unkontrolliert wirkt sich Vorhersagemacht – selbst wenn sie latent bleibt oder es bei der „Androhung" bleibt – erheblich auf das soziale und gesellschaftliche Feld aus.

Risikoprävention jenseits des Datenschutzes

In der Geschichte des Datenschutzes wurde gelegentlich der Begriff der „Datenmacht" herangezogen, um das Regulierungsanliegen des Datenschutzes zu formulieren (Lewinski, 2014: 56 ff.; Lewinski, 2009): Der Datenschutz dient dem Ausgleich informationeller Machtasymmetrie zwischen datenverarbeitenden Organisationen und Individuen bzw. Gesellschaft. Vorhersagemacht, können wir nun anfügen, ist eine aktuell besonders relevante Spielart von Datenmacht. Deshalb liegt es nahe, in Bezug auf die Regulierung von Vorhersagemacht an den Datenschutz zu denken und darauf zu sinnen, den Datenschutz zu erweitern, so dass er auch in Schritt 2 (also bei der Zirkulation trainierter Modelle mit anonymen Modelldaten) greift.

Allerdings kommen wesentliche architektonische Grundstrukturen deutscher und europäischer Datenschutzregulierung (nicht nur der DSGVO, sondern auch ihre historischen Vorläufer wie das BDSG und die europäischen Datenschutzrichtlinie) angesichts dieses Problems an prinzipielle Grenzen: die *eigenen* Daten als Anknüpfungspunkt der informationellen Selbstbestimmung (mit der Rechtsgrundlage der Einwilligung als markantestem Auswuchs); die Orientierung am Personenbezug der Daten und die Unterscheidung zwischen personenbezogenen, anonymen und sensiblen[15] Daten; die individuellen Betroffenenrechte (siehe ausführlich Mühlhoff und Ruschemeier, 2022; Purtova, 2018; Wachter, 2019). All dies sind Grundpfeiler des Datenschutzes als

[15] Gemeint sind die besonderen Kategorien personenbezogener Daten nach Art. 9 DSGVO.

Regulierungsansatz, sie lassen sich nicht einfach „korrigieren" und sind zugleich mit der wirksamen Regulierung des Risikos der Zweitverwendung trainierter Modelle inkompatibel.[16]

Da es somit unklar ist, ob und wie weit der Bogen des Datenschutzes noch gespannt werden kann, um immer weitere, seinem ursprünglichen Konzept unbekannte Gefahren wirksam einzubeziehen, ist es geboten, disziplin- und paradigmenübergreifend nach (neuen) Regulierungsansätzen zu suchen. Unserem[17] Vorschlag nach sollte hierbei insbesondere Ansätze der Risiko- und Gefahrenprävention eine größere Rolle spielen, wie sie aus dem Umweltrecht bekannt sind.[18] Diese Ansätze machen das Prinzip der Risikoprävention stark, um den Staat in die Pflicht zu nehmen, die Eintrittswahrscheinlichkeit von Gefahren zu reduzieren. So ist insbesondere im EU-Recht das „precautionary principle" bekannt, welches ursprünglich im *Vertrag über die Arbeitsweise der Europäischen Union* (AEUV) in Bezug auf Umweltschutz veranschlagt wird (siehe 191(2) AEUV). Über die letzten Jahrzehnte ist das Vorsorgeprinzip jedoch verstärkt als ein allgemeines Prinzip des EU-Rechts interpretiert und auch in anderen Bereichen angewandt worden (Girela, 2006).

Gerade weil die Gefahr durch missbräuchliche Zweitverwendung trainierter Modelle breit gestreut ist (sie betrifft prinzipiell große Kohorten von Individuen) und bereits gegeben ist, *bevor* sie sich in der Diskriminierung einzelner Individuen oder in der Hervorbringung sozialer Sortierungsmuster oder gesellschaftlicher Ungleichheiten manifestiert, ist Risikoprävention ein vielversprechender Ansatz im vorliegenden Kontext.

Drei Rückfragen mit Antworten

1. Warum hängt das spezifische Risiko, das hier Gegenstand von Regulierung werden soll, mit ML-Modellen zusammen? Warum sind Modelldaten der Anknüpfungspunkt für die Analyse und Regulierung dieses Risikos? Man könnte ja auch mittels weniger zuverlässigen Methoden, z.B. simpler Heuristiken, Stereotypen und manueller Verfahren, Vorhersagen erstellen und Menschen unterschiedlich behandeln; hier wären dann keine Modelldaten im Spiel. Man denke an Bankberater:innen, die über Kreditvergabe entscheiden, oder Mitarbeiter:innen von Personalabteilungen bei der Bewerberauswahl.

Grundsätzlich geht das Gefährdungsrisiko von der *Skalierungsfähigkeit* des Verfahrens in seinem Verwendungskontext aus. Wird ein – wie auch immer geartetes – Vorhersageverfahren in einen Bereich transferiert, in dem es zum Beispiel überhaupt nur für *eine* Zielperson angewendet werden kann (z.B. die private Kopie eines Modells wird von einer Privatperson eingesetzt, die nicht systematisch Zugriff auf die nötigen Hilfsdaten

[16] Zur Unvereinbarkeit des Datenschutzes mit Big Data siehe weiterhin Zarsky (2016); Hildebrandt (2013).

[17] Dieses Argument verdanke ich der intensiven Zusammenarbeit mit Hannah Ruschemeier, es soll daher hier nur in aller kürze angedeutet werden, siehe ausführlicher (Mühlhoff und Ruschemeier, 2023).

[18] Siehe https://www.umweltbundesamt.de/vorsorgeprinzip

vieler anderer Individuen hat), dann ist das Risiko klein. Skaliert die Anwendbarkeit eines Verfahrens sehr stark, zum Beispiel weil es ein rechnergestütztes Verfahren ist, das in die Hände von Unternehmen fällt, die über die Daten großer Menschenmengen verfügen oder ihre Dienste an zahlreiche Abnehmer verkaufen, dann ist das Risiko sehr groß. Digitale Verfahren, insbesondere ML-Modelle, stellen unter dem Aspekt der Skalierbarkeit grundsätzlich ein erheblich höheres Risiko dar als manuelle Verfahren (z.B. die erfahrene Ärzt:in, die viele Fälle gesehen hat und daher gut einschätzen kann, ist ein geringes Risiko). Auch Datensicherheitsaspekte – eine sichere Speicherung des trainierten Modells, die vor unautorisierten Zugriffen schützt – sind Faktoren, die das Risiko bestimmen.

2. Besteht das diagnostizierte Risiko auch bei Modellen, die in ihren Vorhersagen nur wenig Genauigkeit aufweisen? Hängt die Dringlichkeit einer Regulierung nicht davon ab, wie genau ein Modell ist?

Tatsächlich dürfte von *sehr* ungenauen Modellen ein geringeres Risiko ausgehen, wenn es aufgrund ihrer Ungenauigkeit nicht wirtschaftlich ist, sie einzusetzen. Umgekehrt stellen jedoch bereits Modelle mit einer „mittleren" Genauigkeit in hochskalierungsfähigen Anwendungen ein hohes Risiko dar. Denn ab einer bestimmten Genauigkeit ist der Einsatz automatisierter Verfahren wirtschaftlich, auch wenn zahlreiche falsch eingeschätzte Fälle einen Kollateralschaden bilden, der aber aus Sicht der Betreiber nicht ausreichend ins Gewicht fällt.[19] Deshalb ist es bei einer Regulierung entscheidend, dass sie in der Konzeption des Risikos der missbräuchlichen Zweitverwendung trainierter Modelle nicht auf die Genauigkeit des Modells abhebt.

3. Die Kategorie des „Risikos" ist auch in der geplanten KI-Verordnung der Europäischen Union („AI Act") ein zentrales Konzept. KI-Systeme werden dort in verschiedene Risikogruppen eingeteilt und ihr Betrieb entsprechend mit regulatorischen Auflagen versehen.[20] Schließt die KI-Verordnung somit die hier herausgestellte Regulierungslücke?

Das Risiko einer missbräuchlichen Zweitverwendung trainierter Modelle wird durch die KI-Verordnung nach dem aktuellen Stand nicht adressiert und nicht wirksam bekämpft. Denn Gegenstand der Klassifikation nach Risikogruppen der KI-Verordnung ist bisher der Primärzweck des Systems.[21] Der Parlamentsvorschlag hingegen möchte unabhängig vom Einsatzzweck auf die Einsatzmöglichkeiten des KI-Systems abstellen, vgl. Art. 7 Abs. 2 KI-VO-EP.[22] Die Gefahr der missbräuchlichen Zweitverwertung bedeutet jedoch den

[19] Zum Beispiel haben die Entwickler:innen von KI-Systemen zur Bewerberauswahl bei Jobs mit hohen Bewerberzahlen Anreize dazu, die Quote retrospektiv betrachtet fälschlich angenommener Bewerber:innen zu minimieren, während die eine höhere Quote fälschlich abgelehnter Bewerber:innen nicht so sehr ins Gewicht fällt (Mühlhoff, 2021). Diese „Auf Sicherheit spielen"-Strategie kann hohe Diskriminierungsfolgen für Individuen haben, die als unklare Kandidat:innen eingestuft werden.

[20] Zur Problematik von KI als Regulierungsobjekt im Kontext der KI-Verordnung im Zusammenhang mit der Risikoklassifikation: Ruschemeier (2023).

[21] So der Vorschlag von Kommission und Rat, Art. 7 Abs. 2 a) KI-VO-E, COM 2021/206 final; 2021/0106(COD).

[22] Ich beziehe mich auf die folgende Version: Abänderungen des Europäischen Parlaments vom 14. Juni 2023 zu dem Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung

Transfer eines Systembestandteils – der Modelldaten – in einen anderen Anwendungskontext – dieser kann auch zukünftig, nach einer Firmenübernahme, oder aufgrund eines unauthorisierten Datenzugriffs geschehen. Das Risiko des Transfers in einen anderen Anwendungskontext ist in der Bewertung durch die KI-Verordnung nicht berücksichtigt. So bezieht sich die Hochrisikoliste des Annex III des KI-VO-E auf einen bestimmten Anwendungskontext, der Transfer in andere Kontexte ist selbst kein Risikofaktor. Dass dabei ein Modell trainiert werden könnte, das medizinische Informationen in einem anderen Kontext als dem der Erstellung des Modells abzuschätzen erlaubt (siehe Szenario B), ist ein Nebeneffekt, der bei der Risikobewertung zurzeit nicht einbezogen wird. Allerdings kann die sekundäre Anwendung des Modells, z.B. im Bereich von targeted adverstisements für Jobs, selbst eine Hochrisiko Anwendung darstellen.[23] Inwieweit der KI-VO-E effektiv vor Grundrechtsgefährdungen wie Diskriminierung schützen wird, ist Gegenstand laufender Diskussionen (Smuha u. a., 2021).

5. Lösungsansätze und Ideen für Regulierung

Ausgehend von der vorausgegangenen Bestimmung des regulatorischen Problems und der Begründung, warum aktuelle Regulierungsansätze nicht ausreichen, möchte ich im Folgenden zwei Denkansätze in Richtung einer Lösung vorstellen. Beide sind Gegenstand aktueller Forschung und stellen nur erste Ansätze dar. Die potenziellen Machtungleichheiten und die damit verbundenen Risiken, die von der unkontrollierten Zweitnutzung trainierter KI-Modelle ausgehen, sind gravierend und zugleich im Schema bestehender Ansätze in Ethik, Recht und Politik schwer zu fassen. Deshalb bedarf es neuer Denkbewegungen auf mehreren Ebenen zur Behandlung des Problems. Ich werde im Folgenden zuerst unter dem Stichwort „prädiktive Privatheit" eine *ethische* Perspektive skizzieren, die auf eine gesellschaftliche Wertedebatte abhebt, die dringend zur allgemeinen Bewusstseinsbildung und öffentlichen Positionierung gegenüber dieses Problems vonnöten ist. Zweitens werde ich unter dem Titel „Zweckbindung für Modelle" einen regulatorischen Denkansatz anführen.

A) Prädiktive Privatheit als Schutzgut

Es ist grundsätzlich die Frage, welche ethischen Argumente gegen die Zweitverwendung trainierter Modelle in Anschlag gebracht werden können, oder anders formuliert: Was steht ethisch dabei auf dem Spiel? Um die bereits genannten strukturellen Argumente (Diskriminierung, soziale Ungleichheit) durch ein Argument zu ergänzen, das aus individueller Perspektive zugänglich ist, habe ich anderswo unter dem Begriff der „prädiktiven Privatheit" argumentiert, dass die breite Verwendung von

harmonisierter Vorschriften für künstliche Intelligenz (Gesetz über künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))(1)

[23] Vgl. Annex III Nr. 4 KI-VO-EP.

Vorhersagetechnologie in zahlreichen wirtschaftlichen und gesellschaftlichen Bereichen zu einer Konstellation führt, in der *Privatheit* auf neue Weise gefährdet ist (Mühlhoff, 2021, 2023b). Es kommt durch die Anwendung prädiktiver Analytik zu einer neuen Form der Verletzbarkeit von Privatheit, die darin besteht, dass Informationen über Individuen *abgeschätzt* werden. Das heißt, im Sinne eines normativen Vorschlags sollten wir Privatsphäre so verstehen, dass sie auch durch Vorhersage und Abschätzung von Informationen unterwandert werden kann, nicht nur durch den missbräuchlichen Umgang mit Informationen, die das Individuum explizit bereitgestellt hat. Im Sinne einer zunächst *negativen* Definition von prädiktiver Privatheit lässt sich somit festhalten, dass die prädiktive Privatheit einer Person oder Gruppe *verletzt* ist,

> „wenn [persönliche oder gar] sensible Informationen über diese Person oder Gruppe gegen ihren Willen oder ohne ihr Wissen auf der Grundlage von Daten vieler anderer Personen vorhergesagt werden, sofern diese Vorhersagen zu Entscheidungen führen [könnten], die das soziale, wirtschaftliche, psychologische, physische, … Wohlbefinden oder die Autonomie einer Person beeinträchtigen". (vgl. Mühlhoff, 2021: 5)

Während die Übersetzbarkeit des Prinzips prädiktiver Privatheit in Regulierungsansätze nicht unmittelbar evident ist, besteht das Ziel dieser ethischen Debatte primär in einer gesellschaftlichen Wertediskussion. Die Verletzung der Privatsphäre durch Vorhersagen, insbesondere solche, die auf Trainingsdaten *anderer* Menschen basieren, wird bisher kaum öffentlich diskutiert. Für viele Menschen ist dies kein Bestandteil ihres moralischen Bewusstseins bezüglich Datenschutz und Privatsphäre im Internet und auch akademisch wurde dieses Thema kaum ethisch untersucht. Übrigens ist es für eine Verletzungen prädiktiver Privatheit unerheblich, ob die vorhergesagten Informationen korrekt sind; denn auch falsche Vorhersagen über persönliche Attribute können zu nachteiligen Handlungsfolgen führen (vgl. Mann und Matzner, 2019; Noble, 2018; Viljoen, 2021; Wachter und Mittelstadt, 2019).

Die genannten Vorarbeiten stellen heraus, dass ein besonderes ethisches Problem prädiktiver Privatheit in der „prediction gap" besteht: Zu diesem Problem kommt es, wenn algorithmische Vorhersagen in Handlungsentscheidungen umgesetzt werden. Das Ausgabedatum eines Vorhersagemodells ist im Allgemeinen keine eindeutige Zuordnung oder Entscheidung, sondern stets eine Wahrscheinlichkeitsaussage (z.B. „70% Wahrscheinlichkeit, an Depression zu leiden", „40% an Angststörungen" in Szenario A). Eine Handlungsroutine, die auf solchen Vorhersagen basiert, muss sich für eine der möglichen Merkmale entscheiden (z. B. den mit dem höchsten Wahrscheinlichkeitsgewicht) und behandelt dann die Person so, als ob sie diese Eigenschaft *sicher* besitzt. Dieser Prozess beinhaltet die Umwandlung einer statistischen Inferenz, die immer auf populationsbezogenem Wahrscheinlichkeitswissen beruht (bezieht sich auf die Gesamtheit aller Individuen in den Trainingsdaten), in eine Vorhersage für

einen Einzelfall (Punkt-Prädiktion). Dieser Sprung vom Gruppenbezug zum Individualbezug geht über die klassische statistische Argumentationsweise hinaus und bedeutet, dass eine Wette über das Individuum eingegangen wird.[24] Dieses „Festnageln" des Individuums auf die Vorhersage mit dem höchsten Wahrscheinlichkeitswert stellt ein für Vorhersagewissen spezifisches ethisches Problem dar, da in einem Kontext von Möglichkeiten und Unsicherheiten die Behandlungsweise des Individuums vereindeutigt und seine Zukunft vorausgefasst wird (siehe ausführlich Mühlhoff, 2021).

Die negative Definition prädiktiver Privatheit taugt gewiss dazu, eine öffentliche Debatte zu befruchten und auf eine spezifische Gefahr hinzuweisen. Doch ihr haftet dasselbe Problem an wie der bisherigen Datenschutzregelung: Sie greift erst in Verarbeitungsschritt 3, wenn das Vorhersagemodell auf ein konkretes Individuum angewendet wurde, und nicht, wie für das regulatorische Projekt nötig, bereits in Schritt 2. Um einen präventiven Regulierungsansatz ethisch zu untermauern, muss das ethische Problem von der individualistischen Perspektive der möglichen prädiktiven Verletzung der eigenen Privatsphäre deshalb auf das kollektive ethische Problem der *potenziellen* Verletzung prädiktiver Privatheit beliebiger (und beliebig vieler) Individuen ausgeweitet werden. Eine *positive* Definition prädiktiver Privatheit ermöglicht es dann, ein *Schutzgut* zu artikulieren, welches der Gefahrenlage durch Vorhersagemacht entgegengestellt werden kann. In Mühlhoff und Ruschemeier (2022) wurde deshalb definiert:

> „Prädiktive Privatheit als gesellschaftliches Schutzgut bezeichnet den Schutz des Gemeinwesens vor negativen Auswirkungen von Vorhersagemacht großer datenverarbeitender Organisationen. Prädiktive Privatheit formuliert somit den – zunächst ethisch begründeten – Anspruch, Individuen und die Gesellschaft im Ganzen gegen die unkontrollierte Akkumulation von Vorhersagemacht als Ausformung informationeller Machtasymmetrie zu schützen." Mühlhoff (2023b)

Diese positive Definition von prädiktiver Privatheit als ein Schutzgut im kollektiven Interesse bietet das ethische Fundament für die Regulierung missbräuchlicher Anwendungen einer Technologie, die vielen Individuen und damit der Gesellschaft im Allgemeinen strukturell schaden können. Eine Bedrohung prädiktiver Privatheit in diesem kollektiven und nicht individuellen Sinn bezieht sich auf eine politisch, wirtschaftlich und technologisch fundierte Machtasymmetrie, die soziale Ungleichheit, automatisierte Ausnutzung individueller Schwachstellen und die datenbasierte sozioökonomische Selektion auf struktureller Ebene durch den Einsatz von Vorhersagemodellen fördert.

B) Zweckbindung für Modelle

Wie in Kapitel 3 ausgeführt, sollte eine Regulierung, die auf das Risiko der missbräuchlichen Zweitnutzung trainierter KI-Modelle abstellt, insbesondere *trennscharf*

[24] Hier steht ein Übergang von einem frequentistischen zu einem subjektiven Wahrscheinlichkeitsbegriff im Hintergrund, wie ihn die Bayes'sche Statistik stark macht, vgl. (Joque, 2022).

die gesellschaftlich und politisch für wünschenswert erachteten Anwendungszwecke von den für „schlecht" oder „missbräuchlich" erachteten trennen. Die Idee ist, dass der Schutz der Gesellschaft vor der Zweckentfremdung trainierter Modelle die Durchführung nützlicher Maßnahmen – z.B. in der medizinischen Forschung – sogar fördern könnte, da man aktuell eigentlich wegen des schlecht regulierten Zweitnutzungsrisikos vor der Durchführung einiger Maßnahmen (wie z.B. den Szenarien A und B) zurückschrecken sollte. Aktuell ist wegen mangelnden öffentlichen Bewusstseins der politische und moralische Druck in diese Richtung noch nicht besonders hoch und die Regulierung der Sekundärnutzung von Modellen und Modelldaten wird bisher nur vereinzelt diskutiert (Mühlhoff und Ruschemeier, 2023; Ruschemeier, 2023). In der Debatte um KI-Regulierung stehen primär das Konzept „künstliche Intelligenz" und der primäre Einsatzzweck im Zentrum.

Heuristisch betrachtet sind mindestens zwei verschiedene Regulierungsansätze, die bei den Modelldaten anknüpfen, denkbar: Erstens könnte man eine Positiv- und Negativliste von erlaubten und verbotenen Zwecken für den Einsatz von KI-Modellen erstellen. Diese Vorgehensweise, die ich im folgenden als „statisch" bezeichne, ist zum Beispiel in dem Entwurf einer Regulierung des European Health Data Space (EHDS – Com2022/197-final) erkennbar, die in Art. 34/35 erlaubte und verbotene Sekundärnutzungsweisen von Gesundheitsdaten definiert.[25] Ein zweiter, damit ggf. kombinierbarer Ansatz besteht in unserem (vgl. Mühlhoff und Ruschemeier, 2023) Vorschlag einer *Zweckbindung* für Modelle. Damit ist gemeint, dass „[d]ie Erstellung und Verwendung von KI-Modellen auf bestimmte Zwecke beschränkt sein muss, die im Voraus festgelegt und während des gesamten Lebenszyklus eines KI-Modells durchgesetzt werden" (Mühlhoff und Ruschemeier, 2023: 2). Dies könnte zum Beispiel durch Einführung einer Aufsichtsinstanz umgesetzt werden, gegenüber der die Organisationen, die ML-Modelle trainieren, *vorher* den Anwendungszweck dieses Modells definieren und genehmigen lassen müssen. Des Weiteren müsste das Modell bei dieser Kontrollinstanz registriert werden und eine Zweitverwendung zu anderem Zwecke wäre unter Androhung von Sanktionen (z.B. analog den Bußgeldern nach DSGVO) untersagt.

Anders als der statische Regulierungsvorschlag direkten Festsetzung einer Positiv-/Negativliste stellt Zweckbindung für Modelle zunächst nur eine Kontrollprozedur bereit, die eine *Beschränkung* bzw. *Bindung* an den Primärzweck vorschreibt, um eine schleichende Ausweitung des Zwecks zu verhindern. In Bezug auf die Entscheidung, welche Primärzwecke überhaupt zulässig sein sollten, wäre dieser Vorschlag noch weiter auszubauen. Um der Komplexität der anstehenden Entscheidungssituationen gerecht zu werden, wäre ein partizipatives und dynamisches Verfahren wünschenswert, indem die Kontrollinstanz etwa verschiedene gesellschaftliche Gruppen und Stakeholder anhören muss, um ggf. Einzelfallentscheidungen zu treffen. Natürlich ist der Vorschlag zugleich

[25] Ich verdanke den Hinweise Hannah Ruschemeier, vgl. (Mühlhoff und Ruschemeier, 2023: 15).

mit einer Positiv-/Negativliste kombinierbar, und *sollte* im Sinne einer schnellen Bearbeitung von Standardfällen mit einer solche kombiniert werden. Selbst dann besteht jedoch der Unterschied zur bloß statischen Liste der erlaubten vs. verbotenen Zwecke darin, dass Zweckbindung für Modelle vor allem die Festsetzung auf den *einen* vorab benannten Zweck festschreibt. So wird etwa eine schleichende Ausweitung des Zweckes innerhalb der Positivliste unterbunden, ohne die Kontrollinstanz dabei einzubeziehen.

Bibliographie

Abadi M, Chu A, Goodfellow I, u. a. (2016) Deep Learning with Differential Privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security - CCS'16*: 308–318.

Barocas S und Selbst AD (2016) Big data's disparate impact. *Calif. L. Rev.* 104: 671.

Borodovsky JT, Marsch LA und Budney AJ (2018) Studying Cannabis Use Behaviors With Facebook and Web Surveys: Methods and Insights. *JMIR Public Health and Surveillance* 4(2): e48.

Bozdag E (2013) Bias in algorithmic filtering and personalization. *Ethics and Information Technology* 15(3): 209–227.

D'Ignazio C und Klein LF (2020) *Data feminism*. Strong ideas series. Cambridge, Massachusetts: The MIT Press.

Dwork C (2006) Differential Privacy. In: Bugliesi M, Preneel B, Sassone V, u. a. (Hrsg.) *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10–14, 2006, Proceedings, Part II*. Lecture Notes in Computer Science 4052. Berlin and Heidelberg: Springer, S. 1–12.

Dwork C (2011) A firm foundation for private data analysis. *Communications of the ACM* 54(1): 86–95.

Eubanks V (2017) *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. First Edition. New York, NY: St. Martin's Press.

Foucault M (1983) *Der Wille zum Wissen: Sexualität und Wahrheit 1* (Übers. U Raulff und W Seitter). Frankfurt am Main: Suhrkamp.

Foucault M (2007) Subjekt und Macht. In: Defert D und Ewald F (Hrsg.) *Ästhetik der Existenz: Schriften zur Lebenskunst*. Frankfurt am Main: Suhrkamp, S. 81–104.

Fredrikson M, Jha S und Ristenpart T (2015) Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, Denver Colorado USA, 12 Oktober 2015, S. 1322–1333. ACM. Available at: https://dl.acm.org/doi/10.1145/2810103.2813677 (zugegriffen 22 Januar 2022).

Girela MAR (2006) Risk and Reason in the European Union Law. *European Food and Feed Law Review* 1: 270.

Gymrek M, McGuire AL, Golan D, u. a. (2013) Identifying Personal Genomes by Surname Inference. *Science* 339(6117): 321–324.

Hildebrandt M (2013) Slaves to Big Data. Or Are We? *IDP. REVISTA DE INTERNET, DERECHO Y POLÍTICA* 17: 7–44.

Hildebrandt M und Gutwirth S (Hrsg.) (2008) *Profiling the European Citizen: Cross-Disciplinary Perspectives*. New York: Springer.

Joque J (2022) *Revolutionary Mathematics: Artificial Intelligence, Statistics and the Logic of Capitalism*. London New York: Verso.

Kaissis GA, Makowski MR, Rückert D, u. a. (2020) Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence* 2(6): 305–311.

Lewinski K von (2009) Geschichte des Datenschutzrechts von 1600 bis 1977. In: *Freiheit – Sicherheit – Öffentlichkeit: 48. Assistententagung Öffentliches Recht, Heidelberg 2008*. Baden-Baden: Nomos, S. 196–220.

Lewinski K von (2014) *Die Matrix des Datenschutzes Besichtigung und Ordnung eines Begriffsfeldes*. Tübingen: Mohr Siebeck. Available at: http://public.eblib.com/choice/PublicFullRecord.aspx?p=6624481 (zugegriffen 15 Januar 2022).

Lynskey O (2019) Grappling with „Data Power“: Normative Nudges from Data Protection and Privacy. *Theoretical Inquiries in Law* 20(1). De Gruyter: 189–220.

Ma X, Yang H, Chen Q, u. a. (2016) DepAudioNet: An Efficient Deep Model for Audio based Depression Classification. In: *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, New York, NY, USA, 16 Oktober 2016, S. 35–42. AVEC ’16. Association for Computing Machinery. Available at: https://doi.org/10.1145/2988257.2988267 (zugegriffen 14 Oktober 2023).

Mann M und Matzner T (2019) Challenging algorithmic profiling: The limits of data protection and anti-discrimination in responding to emergent discrimination. *Big Data & Society* 6(2): 2053951719895805.

Mühlhoff R (2020) Automatisierte Ungleichheit: Ethik der Künstlichen Intelligenz in der biopolitischen Wende des Digitalen Kapitalismus. *Deutsche Zeitschrift für Philosophie* 68(6): 867–890.

Mühlhoff R (2021) Predictive privacy: towards an applied ethics of data analytics. *Ethics and Information Technology* 23: 675–690.

Mühlhoff R (2023a) *Die Macht der Daten: Warum künstliche Intelligenz eine Frage der Ethik ist*. V&R unipress, Universitätsverlag Osnabrück. Available at: https://www.vr-elibrary.de/doi/book/10.14220/9783737015523.

Mühlhoff R (2023b) Predictive Privacy: Collective Data Protection in Times of AI and Big Data. *Big Data & Society*: 1–14.

Mühlhoff R und Ruschemeier H (2022) Predictive Analytics und DSGVO: Ethische und rechtliche Implikationen. In: Gräfe H-C und Telemedicus e.V. (Hrsg.) *Telemedicus – Recht der Informationsgesellschaft, Tagungsband zur Sommerkonferenz 2022*. Frankfurt am Main: Deutscher Fachverlag, S. 38–67.

Mühlhoff R und Ruschemeier H (2023) Purpose Limitation for Models. 4599869, SSRN Scholarly Paper. Rochester, NY. Available at: https://papers.ssrn.com/abstract=4599869 (zugegriffen 13 Oktober 2023).

Mühlhoff R und Willem T (2023) Social Media Advertising for Clinical Studies: Ethical and Data Protection Implications of Online Targeting. *Big Data & Society*. i. E. 2023. DOI: 10.1177/20539517231156127.

Müller-Jung J (2023) „Bürokratie kann tödlich wirken“ (Interview mit Michael Hallek, Axel Ockenfels, Wiebke Rösler). *Frankfurter Allgemeine Zeitung*, 9 August. 183. Aufl.

Narayanan A und Shmatikov V (2008) Robust De-anonymization of Large Sparse Datasets. In: *2008 IEEE Symposium on Security and Privacy (sp 2008)*, Oakland, CA, USA, Mai 2008, S. 111–125. IEEE. Available at: http://ieeexplore.ieee.org/document/4531148/ (zugegriffen 21 Januar 2022).

Nida-Rümelin J und Hilgendorf E (2021) Grundrechte: Unser Datenschutz verhindert eine wirksame Corona-Warn-App. Available at: https://www.welt.de/debatte/kommentare/plus224695267/Grundrechte-Unser-Datenschutz-verhindert-eine-wirksame-Corona-Warn-App.html (zugegriffen 14 Oktober 2023).

Nissim K, Steinke T, Wood A, u. a. (2018) Differential Privacy: A Primer for a Non-Technical Audience. *SSRN Electronic Journal*. i. E. 2018. DOI: 10.2139/ssrn.3338027.

Noble SU (2018) *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.

O’Neil C (2016) *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. First edition. New York: Crown.

Ohm P (2010) Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. *UCLA Law Review* 57: 1701–1777.

Purtova N (2018) The law of everything. Broad concept of personal data and future of EU data protection law. *Law, Innovation and Technology* 10(1). Routledge: 40–81.

Rehak R (2022) When ethics demands the already present: How ethics undermines effective data protection in the case of the Corona-Warn-App in Germany. In: Krämer D, Haltaufderheide J, und Vollmann J (Hrsg.) *Technologien der Krise: Die Covid-19-Pandemie als Katalysator neuer Formen der Vernetzung*. 1. Aufl. Digitale Gesellschaft. Bielefeld, Germany: transcript Verlag, S. 89–108. Available at: https://www.transcript-open.de/isbn/5924 (zugegriffen 16 Oktober 2023).

Rost M (2018) Risiken im Datenschutz. *Vorgänge – Zeitschrift für Bürgerrechte und Gesellschaftspolitik* 57(1/2): 79–92.

Ruschemeier H (2023) AI as a challenge for legal regulation – the scope of application of the artificial intelligence act proposal. *ERA Forum* 23: 361–376.

Shokri R, Stronati M, Song C, u. a. (2017) Membership Inference Attacks against Machine Learning Models. *arXiv:1610.05820 [cs, stat]*. i. E. 31. März 2017.

Smuha NA, Ahmed-Rengers E, Harkens A, u. a. (2021) How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act. 3899991, SSRN Scholarly Paper. Rochester, NY. Available at: https://papers.ssrn.com/abstract=3899991 (zugegriffen 14 Juli 2023).

Tian H, Zhu Z und Jing X (2023) Deep learning for Depression Recognition from Speech. *Mobile Networks and Applications*. i. E. 26. Januar 2023. DOI: 10.1007/s11036-022-02086-3.

Viljoen S (2021) A Relational Theory of Data Governance. *Yale Law Journal* 131(2): 573–654.

Wachter S (2019) Data protection in the age of big data. *Nature Electronics* 2(1): 6–7.

Wachter S und Mittelstadt B (2019) A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI. *Columbia Business Law Review* 2019(1): 1–130.

Weber M (1990) *Wirtschaft und Gesellschaft: Grundriss der verstehenden Soziologie* (Hrsg. J Winckelmann). 5., rev. Aufl., Nachdr., Studienausg. Tübingen: Mohr.

Zarsky TZ (2016) Incompatible: the GDPR in the age of big data. *Seton Hall L. Rev.* 47: 995.

Zarsky TZ (2019) Privacy and Manipulation in the Digital Age. *Theoretical Inquiries in Law* 20(1): 157–188.