www.jusletter-it.eu

David Rosenthal

Datenschutz und KI: Worauf in der Praxis zu achten ist

Der Einsatz von «künstlicher Intelligenz» sorgt seit einigen Jahren vermehrt für Gesprächsstoff. Neu ist das Thema zwar nicht, aber in der Praxis werden immer häufiger Fragen laut, wie damit aus datenschutzrechtlicher Sicht umzugehen ist. Dieser Beitrag beantwortet einige davon und zeigt sowohl über- als auch unterschätzte datenschutzrechtliche Risiken beim Einsatz von KI auf.

Beitragsart: Beiträge
Region: Schweiz
Rechtsgebiete: Datenschutz

Zitiervorschlag: David Rosenthal, Datenschutz und KI: Worauf in der Praxis zu achten ist, in: Jusletter IT 22. April 2022

Inhaltsübersicht

[1] Was «künstliche Intelligenz» (KI) überhaupt ist, wird an dieser Stelle nicht vertieft erörtert. Der Begriff ist gegenwärtig zweifellos ein Modebegriff, mit welchem sich IT-Systeme und -Leistungen sehr gut vermarkten lassen – von Software zur Abwehr von Malware und Ransomware-Attacken über «intelligente» Steuersoftware in Fahrzeugen über Tools zur Steuerung von Marketingaktivitäten oder Analyse von Dokumenten. Wir alle nutzen KI-basierte Services in unterschiedlichster Form – von Übersetzungslösungen wie «DeepL» bis hin zur Spracherkennung von virtuellen Assistenten, ohne uns dessen bewusst zu sein. In anderen Fällen werden die Möglichkeiten der maschinellen «Intelligenz» hoffnungslos überzeichnet. Klar ist jedoch, dass KI auf Konzepten basiert, die in den letzten Jahren aufgrund der gestiegenen Leistungsfähigkeit von Computersystemen praxisrelevanter geworden sind. Die Möglichkeit, über Cloud-Angebote günstig an grosse Mengen von Rechen- und Speicherleistung zu gelangen, hat diesen Entwicklungen Schub verliehen. Immer häufiger werden auch KI-Modelle öffentlich angeboten, d.h. es entsteht ein Markt für Expertensysteme, die auf einer bestimmten Datenbasis dafür trainiert worden sind, bestimmte Fragen zu beantworten. Das können nebst Alltagsanwendungen auch hochspezialisierte Systeme sein, um zum Beispiel Ärzten zu helfen, Leukämie anhand eines Blutbilds besser zuverlässiger diagnostizieren zu können.[1] Andere vielzitierte Beispiele sind Systeme zur Sprach-, Bild- und Gesichtserkennung.

1. Worüber wir wirklich sprechen

[2] Im vorliegenden Beitrag sind mit KI ausschliesslich Verfahren zur *Mustererkennung* gemeint, d.h. Computersysteme, welche in Bildern, Videos, Texten, Transaktionen oder anderen Inhalten bestimmte Muster erkennen können, die auf bestimmte Vorgänge oder Zustände hindeuten. Sie sind dahingehend «dumm», dass sie die Muster weder verstehen noch irgendwelche Kausalitäten begründen können; was sie finden, sind Korrelationen. Das können sie, weil sie mit Trainingsinhalten darauf «konditioniert» wurden; so ähnlich wie ein Drogenspürhund darauf trainiert wurde, Gerüche von bestimmten Stoffen aufzuspüren und anzuzeigen. Womit er es aber zu tun hat,

[1] Vgl. etwa https://sma.org/ai-in-medical-diagnosis/ (zuletzt abgerufen am 13. März 2022).

weiss auch er nicht. In der Fachsprache ist häufig von «machine learning» oder «deep learning» die Rede.

[3] Setzt also beispielsweise ein Online-Shop eine KI zum Aufspüren von Betrügern auf, so muss sie vorgängig mit hinreichend vielen Transaktionsdaten gefüttert werden, bei welchen es zu Betrug kam und solchen, bei welchen das nicht der Fall war, um die betreffenden Muster, also Anzeichen, zu erkennen. Die Anzeichen verstehen auch sie nicht. Sie schlagen nur an, je mehr solcher Anzeichen sie sehen. Ab welcher Schwelle der Shop-Betreiber dann von einem Betrug ausgeht, ist seine Sache. Davon zu unterscheiden sind Systeme, bei welchen der Shop-Betreiber seinen Computer programmiert, auf bestimmte konkrete, von ihm vorgegebene Zeichen zu achten und Alarm zu schlagen, wie z.B. häufige Tippfehler, die Weigerung persönliche Angaben zu machen oder die Bestellung hochpreisiger Waren ohne bisherige Verkaufshistorie.[2] Das ist für vorliegende Zwecke keine KI.

2. Her mit dem Datenfutter!

[4] Wer eine KI aufbauen – sprich: trainieren – will, braucht viele Daten, aus denen sein Computer ein Sinn für Muster entwickeln und den er dann später für seinen Mustervergleich einsetzen kann. Je mehr und bessere Daten für das Training einer KI benutzt werden können, desto deutlicher kennt sie die Muster, nach denen gesucht wird und die wiederum bestimmte Schlüsse (z.B. eine bestimmte Krankheit) zulassen. Anders gesagt: Je höher die Quantität und Qualität der Daten, die für das Training benutzt werden, desto höher ist die Wahrscheinlichkeit, dass die KI eine richtige Beurteilung abgibt. Auch wenn in vielen Fällen darauf gebaut wird, dass mit steigender Quantität auch die Qualität des gesamten Datensatzes steigt, ist Quantität zwar eine Voraussetzung, für sich aber kein Garant für Qualität und damit eine korrekte Mustererkennung: Wer seine KI nur mit Daten von Männern trainiert und nicht von Frauen, wird auch bei hohem Datenvolumen jene Muster, die sich nur bei Frauen finden lassen, nicht erkennen. Die KI hat in diesem Fall einen «bias», also eine unerwünschte Verzerrung, Neigung oder Tendenz in eine bestimmte Richtung. Weil es sich um eine Maschine handelt, ist oft auch von einem «machine bias» die Rede. Sie merkt selbstverständlich nicht, dass ihr «Musterwissen» unausgeglichen ist.

[5] Wie aber gelangen Anbieter von KIs an ihre Daten? Damit sind wir beim ersten datenschutzrechtlichen Problem in der Praxis. Die eine Quelle sind öffentliche Daten, zum Beispiel Texte, Bilder oder andere Inhalte aus dem Internet, aus staatlichen Datensammlungen (Stichwort «Open Government Data»[3]) oder aus Forschungsprojekten[4]. Auch Maschinen generieren immer mehr Daten, die ausgewertet werden können (Stichwort «Internet of Things»); die Europäische Kommission hat mit dem Entwurf eines «Data Act» kürzlich sogar eine Regulierung für den Zugang zu solchen (Sach-)Daten vorgestellt. Die andere Quelle sind immer mehr Unternehmen, die versuchen, das nötige «Datenfutter» aus den Daten zu gewinnen, die ihnen ihre Kunden anvertraut oder die sie im Rahmen ihrer Geschäftstätigkeit angesammelt haben.

---

2 Vgl. https://www.haendlerbund.de/de/downloads/studie-betrugserkennung-im-online-shop.pdf (zuletzt abgerufen am 13. März 2022).

3 https://www.bfs.admin.ch/bfs/en/home/services/ogd.html (zuletzt abgerufen am 13. März 2022).

4 Wer heute Fördergelder erhalten will, muss sich immer häufiger verpflichten, seine Datenbasis zu publizieren, so auch in der Schweiz.

[6] Das können Provider sein, die die Daten an sich nur im Auftrag ihrer Kunden bearbeiten, also sogenannte Auftragsbearbeiter. Es können aber auch Unternehmen sein, die Daten als Verantwortliche im datenschutzrechtlichen Sinn bearbeiten und bestrebt sind, aus den ihnen so oder so vorliegenden Daten einen Zweitnutzen zu gewinnen, in dem sie sie für das Training von KIs für irgendwelche Spezialaufgaben oder für die Verbesserung ihrer Dienste benutzen. Ein Beispiel für Letzteres könnte eine Bank sein, welche die Transaktionen ihrer Kunden nutzt, um eine KI zur Prognose bestimmter Verhaltensweisen oder Umstände aufzubauen. Ein Beispiel für Ersteres ein Managed Security Services Provider, der die Logfiles der von ihm überwachten Netzwerke seiner Kunden auswerten will, um mit den Ergebnissen ein neues Tool zur Erkennung von bösartigen Aktivitäten für sich und seine Kunden aufzubauen. Ein anderes Beispiel sind sprachgesteuerte Systeme, welche die im Betrieb gesammelten Sprachmuster manuell transkribieren lassen und mit den so gesammelten Audioaufnahmen und erstellten Texten dann ihre Computer trainieren, um so die automatische Spracherkennung ihrer Systeme zu verbessern.

[7] Datenschutzrechtlich ist grundsätzlich beides zulässig und umsetzbar, auch wenn es sich bei den Daten um solche natürlicher Personen handelt. Handelt es sich beim Datensammler selbst um einen Verantwortlichen im Sinne des Datenschutzgesetzes (DSG), der die ihm anvertrauten Daten für solche Zwecke mitnutzen will, liegt eine sog. Sekundärnutzung von Personendaten vor. Sie ist zusammengefasst dann erlaubt, wenn er die betroffenen Personen in seiner Datenschutzerklärung darüber informiert hat und die Daten nicht in personenbezogener Form weitergibt.[5] Widerspricht eine betroffene Person, wird der Verantwortliche in aller Regel versucht sein, sich mit dem Argument über den Einspruch hinwegzusetzen, dass es nur um eine Bearbeitung zu nicht personenbezogenen Zwecken geht, die in der Regel ein überwiegendes Interesse darstellt. Die dafür gemäss Gesetz nötigen Voraussetzungen[6] dürften meist erfüllt sein.[7] Allerdings ist zu beachten, dass auch auf den ersten Blick anonyme, nicht personenbezogene KI-Modelle inhärente Datenschutzrisiken bergen, die der Verantwortliche in den Griff bekommen muss, weil sie sonst alle Annahmen auf den Kopf stellen; dazu weiter hinten mehr.

[8] Etwas schwieriger wird es, wenn es ein Auftragsbearbeiter ist, der das Bedürfnis der Nutzung der Daten für das Training seiner eigenen KI hat. Braucht er hierfür Personendaten, kann er entweder den Kunden (d.h. den Verantwortlichen) dazu bewegen, ihn mit dem Training seiner eigenen KI zu beauftragen und danach quasi das (hoffentlich nicht mehr personenbezogene) Ergebnis zu übernehmen. Oder aber er lässt sich die Nutzung der Personendaten von seinem Kunden erlauben und bearbeitet diese dann in eigener datenschutzrechtlicher Verantwortung. Die französische Datenschutzbehörde CNIL hat sich kürzlich zu diesem Vorgehen geäussert und für die Zwecke der DSGVO nebst der Zustimmung des Verantwortlichen gewisse Zusatzbedingungen formuliert, wie namentlich das Erfordernis eines «Kompatibilitätstests» im konkreten Einzelfall und Information der betroffenen Personen.[8] Diese Zusatzbedingungen liegen insofern auf der Hand, als dass es datenschutzrechtlich um das Thema einer möglichen Zweckentfremdung von Personendaten geht, wie sie auch bei einer Bekanntgabe an einen Dritten vorkommt

5 Im Detail zur Zweinutzung: David Rosenthal, Die rechtlichen und gefühlten Grenzen der Zweitnutzung von Personendaten, in: sic 4/2021 (https://www.rosenthal.ch/downloads/Rosenthal-Zweitnutzung.pdf, zuletzt abgerufen am 13. März 2022).

6 Art. 13 Abs. 2 Bst. e DSG, Art. 31 Abs. 2 Bst. e revDSG.

7 Ebd.

8 Vgl. dazu https://datenrecht.ch/cnil-leitlinien-fuer-die-weiterverwendung-von-daten-durch-auftragsbearbeiter/ (zuletzt abgerufen am 13. März 2022).

(auch wenn eine solche nicht vorliegt[9]). Folglich müssen auch die Bedingungen für eine solche Zweckentfremdung erfüllt sein.

[9] Unter der DSGVO bedeutet dies, dass

- für das KI-Training ein Rechtsgrund nach Art. 6 bzw. 9 DSGVO bestehen muss und die nach Art. 5 Abs. 1 Bst. b DSGVO verlangte «Vereinbarkeit» gegeben ist. Das Training von KIs wird dort, wo es keine negative Auswirkung auf die betroffenen Personen hat und keine besonders schützenswerten Personendaten betrifft und ferner die spezifischen Datenschutzrisiken von KI-Modellen adressiert sind (dazu hinten), in der Regel mit einem «berechtigten Interesse» nach Art. 6 Abs. 1 Bst. f DSGVO begründet werden können und auch vereinbar im Sinne der zitierten Vorschrift[10] sein. Dies muss entsprechend dokumentiert werden, und zwar von demjenigen, der das KI-Training durchführt und demjenigen, der ihm die Personendaten dafür zu Verfügung stellt, da beide (eigenständige) Verantwortliche sind. Letzterer kann dies im Rahmen seines Vertrags mit Ersterem tun.
- die Datenschutzerklärungen beider Verantwortlicher die Verwendung der Personendaten für den Zweck des KI-Trainings reflektieren müssen. Nebst der Nennung des Zwecks im Rahmen der Zweckangabe sind in der Praxis möglicherweise auch die Ausführungen im Bereich der Empfänger anzupassen. Dort wird im Zusammenhang mit Auftragsbearbeitern häufig festgehalten, dass diese die Daten nur für die Zwecke des Verantwortlichen bearbeiten, was im vorliegenden Fall nicht mehr stimmen würde. Zu beachten ist, dass bei einer Zweckänderung, die als mit dem ursprünglichen Zweck vereinbar im Sinne von Art. 5 Abs. 1 Bst. b DSGVO gilt, eine Orientierung vor dem Beginn der neuen Bearbeitung (hier: KI-Training) genügt.[11] Es kann in diesem Fall also auch auf ältere Daten zurückgegriffen werden, was in diesen Fällen oft essenziell ist.

[10] Unter dem DSG gilt grundsätzlich dasselbe, ausser, dass keine Rechtsgrundlage nach Art. 6 bzw. 9 DSGVO erforderlich ist. Das revidierte DSG führt ebenfalls den Begriff der «Vereinbarkeit» ein[12], um betroffene Personen vor unerwarteten, unangebrachten oder beanstandbaren Bearbeitungen zu schützen.[13] Das DSG verlangt Transparenz über die Zwecke der Bearbeitung von Personendaten zwar schon zum Zeitpunkt ihrer Erhebung,[14] doch kann die Verletzung dieser Vorgabe durch ein überwiegendes privates Interesse gerechtfertigt werden, was bei Einhaltung der Vorgaben des Rechtfertigungsgrunds der nicht personenbezogenen Bearbeitung von Personendaten aber wie im Falle der DSGVO in den hier relevanten Fällen machbar sein wird. Für die Datenschutzerklärung gilt grundsätzlich ebenfalls das für die DSGVO gesagte. Es kann somit auch unter dem DSG auf ältere Daten zurückgegriffen werden.

[11] So oder so wird der Verantwortliche, der einen Auftragsbearbeiter mit KI-Ambitionen hat, genau darauf aufpassen müssen, was ihm dieser in den Vertrag schreibt, um nicht unbedacht

9 Datenschutz-technisch liegt jedoch keine Bekanntgabe an einen Dritten vor, weil zwischen dem Auftragsbearbeiter und dem Dritten Personalunion besteht und er die Daten bereits kennt. Wer Daten bereits kennt, dem können sie nicht mehr bekanntgegeben werden.

10 Kriterien: Art. 6 Abs. 2 DSGVO.

11 Art. 13 Abs. 3 DSGVO, Art. 14 Abs. 4 DSGVO.

12 Art. 6 Abs. 3 revDSG.

13 David Rosenthal, Das neue Datenschutzgesetz, in: Jusletter, 16. November 2020, Rz. 36.

14 Art. 4 Abs. 4 DSG, Art. 6 Abs. 2 und 3 revDSG.

Sekundärnutzungen seiner Personendaten (oder auch sonstigen Daten) zuzustimmen oder sie gar in Auftrag zu geben, obwohl er sie nicht möchte und vor allem selbst gar nicht die nötigen rechtlichen Voraussetzungen, wie etwa die passende Formulierung seiner eigenen Datenschutzerklärung, geschaffen hat. Sonst kann nebst dem Auftragsbearbeiter auch er als Verantwortlicher für eine DSGVO- bzw. DSG-widrige Nutzung der von ihm an den Auftragsbearbeiter bekanntgegebenen Personendaten zwecks KI-Training zur Verantwortung gezogen werden. Um beim schon erwähnten Beispiel «DeepL» zu bleiben: Wer den Service in der Gratis-Version nutzt, muss damit rechnen, dass die eigenen Texte für weitere Trainings der Übersetzungs-KI verwendet werden und gespeichert bleiben; wer bezahlt, kann das vermeiden.

[12] Ist es möglich, ein Training mit anonymisierten Personendaten durchzuführen, kann dies eine solche Ausgangslage dadurch entschärfen, dass sich der Auftragsbearbeiter mit der Anonymisierung der Personendaten beauftragen lässt, und erst die so anonymisierten Daten übernimmt, die dann nicht mehr unter das Datenschutzrecht fallen. Auch die Anonymisierung von Personendaten ist streng genommen bereits eine Bearbeitung solcher, aber ein Verantwortlicher wird sie in der Praxis in der Regel ohne grosses Risiko vornehmen können. Allerdings sind hier in zweierlei Hinsicht Warnungen angebracht: Erstens ist nicht alles, was als Anonymisierung angepriesen wird, auch tatsächlich eine solche; eine echte Anonymisierung von Daten kann höchst anspruchsvoll sein, sollen die Daten weiterhin brauchbar bleiben. Zweitens stellt ein KI-Training selbst keine Anonymisierung dar. Ein KI-Modell kann seinerseits eine Sammlung von Personendaten sein, d.h. dessen Nutzung zu einer Bekanntgabe von Personendaten führen (dazu hinten mehr). Dies gilt es zu vermeiden, jedenfalls wenn die KI Dritten zugänglich sein soll.

[13] So oder so bleibt die Praxisempfehlung: Wer einen Provider mit der Bearbeitung von Personendaten beauftragt, sollte genau darauf achten, ob der Vertrag eine Klausel für die Zweitnutzung der Daten für KI-Trainingszwecke durch den Provider enthält. Einer solche sollte nur nach eingehender Prüfung zugestimmt werden, soweit sich ein Kunde darauf kommerziell überhaupt einlassen will.

3. Die trickreiche gemeinsame Verantwortlichkeit

[14] Immer häufiger werden Provider nicht nur Daten ihrer Kunden nutzen, um eine KI zu trainieren, sondern eine solche auch für die Zwecke ihrer Services zum Einsatz bringen. Dies wirft sodann die datenschutzrechtlich spannende Frage auf, ob ein Auftragsbearbeiter, der dies tut, wirklich nur Auftragsbearbeiter ist oder ob der Beizug seiner KI nicht dazu führen muss, dass er als gemeinsamer Verantwortlicher gilt.

[15] Die Überlegung ist wie folgt: Ist es die KI, welche mindestens zu einem wesentlichen Teil bestimmt, wie die Personendaten des Kunden bearbeitet werden; kann vertreten werden, dass *sie* es ist, welche die datenschutzrechtlichen Parameter der Datenbearbeitung vorgibt? Ist die KI bzw. deren Training dem Auftragsbearbeiter zuzurechnen, bestimmt letztlich er durch die Ausgestaltung und das Training der KI, wie die von der KI bearbeiteten Personendaten bearbeitet werden bzw. was als Output herauskommt. Wer den Zweck oder die Mittel einer Datenbearbeitung mitbestimmt, wird sowohl unter der DSGVO als auch dem (künftigen) DSG zu ihrem gemeinsamen Verantwortlichen und damit ebenfalls zuständig für die Einhaltung des Datenschutzes. Dies wiederum würde bestehende Auftragsbearbeitungen auf den Kopf stellen; die bisher übliche Rollenverteilung würde nicht mehr gelten und die Verträge müssten neu gemacht werden.

[16] Bisher wurde das Thema in Fachkreisen kaum diskutiert. Es lässt sich allerdings mit guten Gründen eine Gegenposition einnehmen: Wenn eine KI letztlich als Werkzeug verstanden wird, dann bestimmt die Datenbearbeitung nicht derjenige, der es herstellt, sondern derjenige, der über seinen Einsatz entscheidet – und das ist der Kunde, nicht der Provider. Das gilt jedenfalls dort, wo eine KI nur spezialisierte, definierte Berechnungs-, Erkennungs- oder Steuerungsaufgaben übernimmt, d.h. als ein Expertensystem auftritt.

[17] Das ist in der Praxis bisher regelmässig der Fall: Der Input und Output ist in einem bestimmten Umfang festgelegt, und «nur» die Logik dazwischen – quasi die Wissensbasis des Systems – ist für den Kunden nicht vollumfänglich transparent (und für den Auftragsbearbeiter vermutlich auch nicht). Trotzdem ist es der Kunde, der sich in diesem eng definierten Spektrum für den Einsatz einer ganz bestimmten Blackbox entscheidet – und zwar genau jene des Auftragsbearbeiters. Das ist für sich noch kein Grund, die Verantwortlichkeiten für die Datenbearbeitung anders zuzuordnen und auf den Auftragsbearbeiter zu verlagern. Auch der Auftragsbearbeiter, der von seinem Kunden die Instruktion der Bearbeitung von Personendaten nach dem Zufallsprinzip erhält, bleibt letztlich Auftragsbearbeiter, weil es der Kunde ist, der sich für das Zufallsprinzip als das Mittel der Datenbearbeitung entschieden hat. Anders wäre es, wenn der Kunde den Auftragsbearbeiter beauftragen würde, selbst die Parameter der Datenbearbeitung festzulegen. Dann aber würde er mit oder ohne Einsatz von KI zu einem eigenen oder gemeinsamen Verantwortlichen.

[18] Demnach führt der Einsatz der KI eines Auftragsbearbeiters als solcher nicht zur gemeinsamen Verantwortlichkeit des Auftragsbearbeiters.

4. Die Sache mit der Verantwortlichkeit

[19] Das Beispiel wirft freilich eine wichtige Frage auf: Muss jemand, der als Verantwortlicher für eine Datenbearbeitung eine KI einsetzt, genau wissen, wie sie funktioniert? Nein, lautet die kurze, wenn auch kontraintuitive Antwort. Entscheidend ist, welche Auswirkung das Ergebnis des Outputs einer KI für die betroffenen Personen hat. Wie bei jeder Datenbearbeitung, die in einer Analyse durch einen maschinellen Algorithmus oder menschliche Denkleistung besteht, ist es deren Ergebnis, welches nach den Bearbeitungsgrundsätzen geprüft werden muss, also insbesondere, ob es verhältnismässig und richtig ist. Warum? Es ist das *Ergebnis* der Bearbeitung, also ihr Output und nicht die Bearbeitung selbst, welche die Wirkung für die betroffene Person zeitigt.

[20] Ob ein Ergebnis richtig ist, hängt wiederum davon ab, welche Bedeutung dem Ergebnis in einer Anwendung zukommt. Berechnet eine KI einen Score, wie wahrscheinlich ein Kreditnehmer seiner Rückzahlverpflichtung nachkommen wird, so bedeutet der Grundsatz der Richtigkeit zum Beispiel nicht, dass die eine Prognose richtig ausfallen muss – niemand kann die Zukunft vorhersagen. Verlangt wird nur, aber immerhin, dass sie im Hinblick auf den Zweck – die Gewährung eines Kreditkaufs – «richtig», also ein taugliches Kriterium dafür ist, das Ausfallrisiko einzuschätzen. Da der Score keine absolute Zuverlässigkeit für sich in Anspruch nimmt, sondern nur eine Eintrittswahrscheinlichkeit darstellt, muss letztlich entscheidend sein, ob er in einem im Hinblick auf diesen Zweck vernünftigerweise vertretbaren Verfahren zustande kommt. Das wiederum kann entweder empirisch beurteilt werden, indem Erfahrungswerte im Einsatz der Methode bestehen und analysiert werden («Die Methode hatte bisher eine Treffergenauigkeit von XX %»), oder aber – wo solche Erfahrungswerte fehlen – die Methode wird als solche validiert

(«Die Methode erfolgt einem objektiv begründeten, für diesen Zweck vernünftig erscheinenden Modell»).

[21] Wo ein auf dem Ergebnis eines maschinellen Algorithmus aufbauender Entscheid – hier: die Kreditgewährung – ohnehin im freien Entscheid des Verantwortlichen liegt, wie dies bei der Kreditvergabe heute grundsätzlich der Fall ist, dürfen die Anforderungen an das Verfahren konsequenterweise nicht zu hoch sein. Darf der Händler auf sein Bauchgefühl hören, wem er Kredit gewährt oder sich seinem Computer eine einfache, über die Jahre entwickelte Regel vorgeben, so muss es fairerweise auch erlaubt sein, den dafür benutzten Kreditscore von einer KI berechnen zu lassen. Der Händler weiss zwar nicht, wie sie dies genau tut, aber er weiss, dass sie dies auf Basis von Zahlungserfahrungen tut, indem sie nach Mustern sucht, welche eine Korrelation zu Zahlungsausfällen aufweisen. So ist es heute ohne Weiteres zulässig, dass sich ein Händler auf den Kreditscore einer Kreditauskunftei verlässt, auch wenn sie weder ihm noch den betroffenen Personen offenlegt, wie sie ihn genau berechnet. Sie muss das nicht einmal im Rahmen des Auskunftsrechts tun. Entscheidend ist nur, ob die von der betroffenen Person als Input verwendeten Daten richtig waren.

5. Höhere Anforderungen an eine KI?

[22] Für den Fall, einer KI mehr oder anderes zu verlangen als von anderen Datenbearbeitungsverfahren, kann auch unter dem Aspekt der Richtigkeit und der Verhältnismässigkeit nicht legitimiert werden. Dies zu tun würde bedeuten, an den Einsatz einer KI höhere Anforderungen zu stellen als an andere Datenbearbeitungen mit demselben Zweck und Ergebnis. Dafür gibt es – mit Ausnahme der Regelung zu automatisierten Einzelentscheiden (dazu sogleich) – im heutigen Recht keine rechtliche Grundlage.

[23] In diesem Zusammenhang wird häufig und richtigerweise das Risiko der Diskriminierung angeführt. Auch hierbei ist jedoch aus rein rechtlicher Sicht zur Kenntnis zu nehmen, dass es ein Diskriminierungsverbot, anders als im Staats- und Verwaltungsrecht für Private von spezifischen Regelungen in Gleichstellungsgesetzen, im Straf-, Miet- und im Arbeitsrecht abgesehen in der Schweiz bisher nicht wirklich gibt. Der Bundesrat hat die Einführung einer Diskriminierungsnorm im Privatrecht in Ergänzung zum Persönlichkeitsschutz bisher abgelehnt – und so auch das Parlament.[15] Im privaten Bereich ist es somit schwierig, die Datenschutzwidrigkeit einer KI mit der Gefahr einer Diskriminierung zu begründen. Eine solche kann sich immerhin implizit auf eine datenschutzrechtliche Beurteilung auswirken, sei es über das Kriterium der Richtigkeit des Outputs einer KI, sei es über die tatsächlichen Nachteile, welche eine Datenbearbeitung bzw. ihr Output für eine betroffene Person mit sich bringen kann.

[24] Dem allem unbesehen ist es selbstverständlich möglich, dass eine KI in ihrer Mustererkennung Verzerrungen (*bias*) oder blinden Flecken ausgesetzt ist, weil die für ihr Training verwendete Daten in irgendeiner Weise einseitig oder unvollständig waren. Diesen Risiken ist jedoch auch jedes manuell programmierte Verfahren der Datenbearbeitung ausgesetzt, und wo der

15 Recht auf Schutz vor Diskriminierung, Bericht des Bundesrates in Erfüllung des Postulats Naef 12.3543 vom 14. Juni 2012 vom 25. Mai 2016 (https://www.ejpd.admin.ch/dam/bj/de/data/aktuell/news/2016/2016-05-25/ber-br-d.pdf); vgl. dazu auch https://www.humanrights.ch/de/ipf/menschenrechte/diskriminierung/bericht-bundesrat-diskriminierungsschutz und https://www.parlament.ch/de/ratsbetrieb/suche-curia-vista/geschaeft?AffairId=20163626 (Links zuletzt abgerufen am 13. März 2022).

Mensch selbst die Entscheide trifft, kommen noch sehr viel mehr Störfaktoren hinzu, wie etwa Verrauschungen (*noise*), welche die Entscheide zusätzlich beeinträchtigen. In vielen Fällen sind menschliche Entscheide daher in hohem Masse auch vom Zufall geprägt und damit von wesentlich schlechterer Qualität als maschinelle.[16] Auch das oft gehörte Argument, dass Fehler in maschinellen Entscheiden systematisch erfolgen und sich daher im Gegensatz zu den Zufallsfehlern menschlicher Entscheide addieren oder sogar potenzieren, verkennt, dass sich die Auswirkung zufälliger Fehler ebenso addieren können. Was sie unterscheidet ist, dass diese zufälligen Fehler kaum als solche erkennbar sind und daher möglicherweise nicht im selben Ausmass wahrgenommen werden.[17]

[25] Wird eine KI für eine Prognose eingesetzt, muss daher in einem ersten Schritt ermittelt werden, welche Anforderungen an die Richtigkeit dieser Prognose gestellt werden. Das hat zunächst nichts mit der Methode der Prognose zu tun, sondern nur damit, wie die Prognose verwendet wird und welche Bedeutung ihr Ergebnis für die betroffene Person hat. Dient die Prognose der Ermittlung eines geeigneten Bewerbers für eine Stelle, so steht einem Arbeitgeber ein grösserer Spielraum zu als wenn er prognostizieren muss, wie wahrscheinlich ein Arbeitnehmer seine Leistungen nicht mehr erbringen wird und daher entlassen werden soll. Der Einsatz von Scoring-Techniken zur Unterstützung der Bewerberselektion wurde in einer kürzlichen Umfrage unter Datenschutzexperten beispielsweise als nicht besonders heikel beurteilt.[18]

[26] Erst in einem zweiten Schritt sollte geprüft werden, ob die in Frage kommende Prognosemethode, die auch den Einsatz einer KI beinhalten kann, diesen Anforderungen entspricht. Ergeben sich im ersten Schritt höhere Anforderungen an die Prognosequalität, wird sich das auch auf die Anforderungen an das Vorhersageverfahren auswirken. Im Falle einer KI ist dann zu berücksichtigen, wie sie trainiert wurde, ob sie formalisierten Tests unterzogen wurde, wie erklärbar ihr Output ist und welche weiteren Qualitätsmerkmale sie aufweist.

[27] Realität ist allerdings, dass der Mensch an die Qualität einer Prognose einer Maschine höhere qualitative Anforderungen stellt als an die Qualität der Prognose eines Menschen. Zudem akzeptiert er es als Ausfluss der Menschenwürde wesentlich weniger leicht, von einer Maschine beurteilt zu werden als von einem Menschen, selbst wenn der Entscheid des Menschen qualitativ schlechter ist als derjenige der Maschine, d.h. der Richtigkeit vom Menschen weniger Rechnung getragen wird. Ob das sinnvoll und fair ist, spielt letztlich keine Rolle, denn auch datenschutzrechtliche Beurteilungen sind in der Praxis häufig subjektiver Natur, wie der Autor dieses Beitrags an anderer Stelle bereits gezeigt hat.[19]

[28] Wer eine KI für Entscheide betreffend Menschen einsetzen will, muss sich daher zum heutigen Zeitpunkt auf entsprechende Diskussionen vorbereiten und sich bewusst sein, dass er strenger beurteilt werden wird als beim Einsatz anderer Verfahren. Das wird sich zwar im Laufe der Zeit mit zunehmender Gewöhnung an und Erfahrung im Umgang mit KI ändern, aber noch wird KI-Systemen im Datenschutz – und von Datenschützern – grundsätzlich misstraut. KI-Systeme werden auch häufig in ihrer Funktionsweise und Eigenheit aufgrund bisher mangelnder Erfah-

16 Mit vielen weiteren Hinweisen: Daniel Kahneman, Olivier Sibony, Cass Sunstein, Noise: A Flaw in Human Judgment, New York, 2021.

17 Ebd.

18 David Rosenthal, Die Tücken spontaner Datenschutzbeurteilungen und was sich dagegen tun lässt, in: Jusletter 28. Februar 2022, Anhang, Fallbeispiel Nr. 8.

19 Rosenthal, Tücken (Fn 18).

rungswerte nicht wirklich verstanden, was das Misstrauen verstärkt. Im Bereich der Entscheidfindung kommt es daher vor allem in folgenden drei Situationen zu datenschutzrechtlicher Kritik am Einsatz einer KI:

- Wenn das Datenmodell des Inputs der KI atypisch ist, d.h. für eine Mustererkennung Datenkategorien verwendet werden, die für die zu beurteilende Frage nicht erklärbar sind. Beispiel: Die Farbe des Autos ist ein Faktor für die Kreditwürdigkeit;
- Wenn der Output der KI angesichts der Datenlage nicht erklärbar ist. Beispiel: Jemand, der noch nie eine negative Zahlungserfahrung hatte, wird als kreditunwürdig eingestuft;
- Es ist unklar, ob die für das Training der KI verwendete Datenbasis Schwächen aufweist, die unerwünschte Effekte (*bias*, *blind spots*) ergeben, die wiederum zur Diskriminierung bestimmter Gruppen führen können. Beispiel: Eine Kredit-Scoring-KI wurde nur mit Daten bestimmter Kundengruppen oder mit zu wenig Daten trainiert, was dazu führen kann, dass Mitglieder anderer Gruppen in der Tendenz schlechter oder besser beurteilt werden als sie sollten.

[29] Der dritte Fall wird vorläufig nicht selten gegeben sein, weil zwar immer häufiger damit geworben wird, dass Techniken der künstlichen Intelligenz zum Einsatz kommen, aber nichts über deren Training kommuniziert wird. Die Anbieter vermitteln vielmehr den Eindruck, dass dessen Qualität eine Selbstverständlichkeit ist, die keines weiteren Nachweises bedarf (vgl. nachfolgend zur Frage des Nachweises). Funktioniert die KI in alltäglichen Tests einigermassen gut, so wird ihr Benutzer vermuten, dass sie dies auch in besonderen Fällen tut. Dass eine solche Annahme trügerisch ist, bedarf keiner Erläuterung, denn eine KI kann letztlich nur das prognostizieren, was ihr im Training gezeigt wurde. War dies auf Standardfälle fokussiert, wird sie sich auch in ihren Prognosen darauf beschränken. Die Tücke liegt darin, dass eine KI ihre Prognosen auch abseits vom Standardfall liefern wird. Dass sie dabei nur noch «am Raten» ist, wird ihr Nutzer bestenfalls daran erkennen können, wenn die KI nebst der Prognose auch angibt, wie zuversichtlich sie dabei ist. Solche Angaben sind daher wichtig.

[30] Bei den ersten beiden Fällen handelt es sich wiederum um Kausalitätsprobleme, d.h. der Kausalzusammenhang zwischen Input und Output ist für den Betrachter nicht nachvollziehbar. Möglicherweise gibt es eine Erklärung, die sogar zutreffend sein mag, aber der Betrachter erkennt nicht, wie die Dinge zusammenhängen. In solchen Fällen wird die datenschutzrechtliche Beurteilung in aller Regel zum Ergebnis führen, dass mindestens der Grundsatz der Richtigkeit und vermutlich auch der Verhältnismässigkeit verletzt ist.

[31] Dem kann der Benutzer der KI faktisch nur entgegentreten, indem er das Gegenteil beweist. Das wiederum bedeutet, dass er mindestens eine Korrelation zwischen Input und Output aufzeigt, weil dies gemeinsam als Hinweis auf Kausalität oder mindestens gemeinsamer Faktoren gewertet werden wird. Die Funktionsweise seiner KI braucht er dazu nicht offenzulegen; es würde in diesem Fall genügen, dass er ihr eine genügende Anzahl Fälle vorlegt, um anhand des Outputs zu zeigen, dass sie hinreichend richtig liegt – auch wenn er nicht weiss, warum. Was hinreichend ist, wird in der Praxis wiederum der Ergebnisvergleich üblicherweise praktizierter Alternativerfahren bestimmen. Liefert die KI also mindestens so viele Treffer wie ein anderes, akzeptiertes Verfahren, so dürfte sie auch datenschutzrechtlich kaum mehr in Frage gestellt werden – ohne,

dass die Beurteiler genau werden erklären können, warum. Auch Datenschutz ist im Alltag keine exakte Wissenschaft, sondern in hohem Masse dem Zufall ausgesetzt.[20]

6. Risiken prüfen ohne Beurteilungskriterien

[32] Selbst eine Datenschutz-Folgenabschätzung (DSFA)[21] wird in der Praxis nicht wirklich weiter gehen können. Das Datenschutzrecht schreibt sie in gewissen Fällen vor, aber der Einsatz von KI zählt nicht zu den Fällen, die sie per se erforderlich machen. Auch hier wird eine DSFA nur (aber immerhin) dann nötig werden, wenn zusätzliche Faktoren vorliegen, die ein hohes Risiko für die betroffenen Personen mit sich bringen können – wie zum Beispiel, dass folgenschwere Entscheide gestützt auf die Prognose der KI erfolgen sollen.

[33] Eine DSFA erfordert eine schriftlich dokumentierte Auseinandersetzung mit den Risiken einer Datenbearbeitung für eine betroffene Person. Kommt eine KI zum Einsatz, verlangt eine DSFA daher regelmässig eine Einschätzung darüber, wie wahrscheinlich eine dafür eingesetzte KI einen wie auch immer gearteten fehlerhaften Output generiert und darauf mit Folgen für die betroffene Person abgestellt wird. In Konstellationen, wo der Einsatz einer KI negative Folgen für eine Person haben kann, ist die Pflicht zur Durchführung einer DSFA zugleich oft eine faktische Pflicht zur Beurteilung der Qualität einer KI. Eine solche Beurteilung kann selbst dann erforderlich sein, wenn es nicht zu automatisierten Einzelentscheiden kommt, das Ergebnis der KI aber sonst von erhöhter Relevanz für die betroffene Person ist und die Prognosequalität der KI daher zum Risikofaktor für den menschlichen Entscheid wird. Oder anders gesagt: Wenn der Mensch zwar entscheidet, aber trotzdem auf die KI hört, wird sie ein relevanter Teil der Datenbearbeitung und kann negative Auswirkungen haben – auch ohne, dass ein nach DSG automatisierter Einzelentscheid vorliegt.

[34] Eine solche Beurteilung der Qualität einer KI wird in der Praxis oft eine Herausforderung sein. In der Wissenschaft wird derzeit intensiv daran geforscht, wie KI-Modelle erklärbar und damit letztlich prüfbar gemacht werden können (Stichwort «machine learning explainability»); es gibt bereits diverse Ansätze.[22] Ein anderes Praxis- und Forschungsgebiet beschäftigt sich mit der Frage, wie sich die Qualität des Trainings einer KI sicherstellen und überprüfen lässt. Hierzu wurden ebenfalls bereits diverse Verfahren und Strategien für die Praxis entwickelt.[23] In einer datenschutzrechtlichen Beurteilung sollte folglich vorzugsweise überprüft werden, welche Methoden zur Sicherstellung der Qualität der Trainingsdaten verwendet worden sind und mit welchem Erfolg.

[35] Wer eine KI nicht selbst herstellt, wird diese Prüfungen erfahrungsgemäss selten vornehmen können. Er wird diesbezüglich nur auf entsprechende Zusicherungen des Anbieters der betref-

---

20 Vgl. Rosenthal, Tücken (Fn 18).

21 Art. 35 DSGVO, Art. 22 revDSG.

22 Vgl. Jianlong Zhou, Amir H. Gandomi, Fang Chen, Andreas Holzinger, Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. Electronics 2021, 10, 593. https://doi.org/10.3390/electronics10050593 (zuletzt abgerufen am 13. März 2022); Christoph Molnar, Interpretable Machine Learning, A Guide for Making Black Box Models Explainable, https://christophm.github.io/interpretable-ml-book/ (zuletzt abgerufen am 13. März 2022).

23 Vgl. etwa statt vieler https://www.anolytics.ai/blog/how-to-ensure-quality-training-data-for-ai/ und https://hackernoon.com/how-to-measure-quality-when-training-machine-learning-models-cc9196dd377a (beide zuletzt abgerufen am 13. März 2022).

fenden KI abstellen können, die ihm bestenfalls auch vertraglich gewährt werden. Es ist daher zu hoffen und darauf hinzuwirken, dass sich für KI-Systeme in Zukunft Standards für Qualitätsmessungen und -audits etablieren, mit denen kommerzielle Nutzer selbst ohne entsprechendes Fachwissen abschätzen können, ob eine bestimmte KI gewissen Qualitätsanforderungen entspricht.

[36] Kann die Qualität einer KI bzw. ihres Outputs als solche nicht formal gemessen und beurteilt werden und ist ihrem Verwender auch Beurteilung der Qualität ihrer Herstellung versagt, wird er – wie bereits dargelegt – auf Erfahrungen mit alternativen Methoden abstellen und die Performance der KI im Vergleich dazu bewerten müssen. Dieses «Benchmarking» ist die in der Praxis darum derzeit die beliebteste und einfachste Methode der Überprüfung einer KI: Die KI wird mit Testfällen ausprobiert und das Ergebnis mit einem «Gold Standard» oder dem Ergebnis anderer Beurteilungsverfahren verglichen.

[37] Wer auch dies nicht kann, wird in einer DSFA nicht umhinkommen, die Unsicherheit in der Qualität einer KI entweder als Restrisiko so weit tragbar zu akzeptieren, oder auf andere Weise auf die Risikofaktoren der Eintrittswahrscheinlichkeit oder Schadensschwere zu wirken – etwa durch eine Information der betroffenen Person mit dem Angebot der Überprüfung maschineller Entscheide durch einen Menschen, wie das Datenschutzrecht dies auch vorsieht (dazu hinten).

7. Die KI-Regulierung, die wir schon haben

[38] Es wird gegenwärtig viel darüber diskutiert, ob und wie der Gesetzgeber den Einsatz von KI regulieren sollte. Die Europäische Kommission hat hierzu im April 2021 einen konkreten Vorschlag gemacht,[24] in der Schweiz ist eine generelle Regulierung von KI bisher abgelehnt worden, weil der allgemeine Rechtsrahmen in der Schweiz zum jetzigen Zeitpunkt «grundsätzlich geeignet und ausreichend» sei.[25] Die kritische helvetische Haltung ist nach der hier vertretenen Ansicht völlig richtig: Der EU-Ansatz einer horizontalen Regulierung macht einen unüberlegten, übereilten Eindruck. Die Schweiz braucht kein generelles KI-Gesetz. Wenn überhaupt, sollte zusätzliche Regulierung sektorspezifisch und technologieneutral geprüft werden. In diesem Sinne haben sich eine Reihe von Schweizer Wissenschaftler im Rahmen eines Positionspapiers der Digital Society Initiative geäussert; es fasst die Herausforderungen des Einsatzes von KI in den fünf Bereichen Erkennbarkeit und Nachvollziehbarkeit, Diskriminierung, Manipulation, Haftung und Datenschutz/Datensicherheit zusammen.[26]

[39] Was in der Diskussion häufig übersehen wird ist, dass die Schweiz wie auch der EWR mit dem Datenschutzrecht bereits eine die KI betreffende Regulierung hat, da ihr Einsatz in Bezug auf natürliche Personen in aller Regel auch eine Bearbeitung von deren Daten beinhaltet und darum Grundsätze wie die bereits zitierte Richtigkeit und Verhältnismässigkeit einhalten muss. Mit der Pflicht zur DSFA besteht sogar eine Regelung, die ihren Verwender dazu zwingt, in Anwendungsfällen mit hohen Risiken für eine betroffene Person sich Gedanken über die Qualität

24 https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence (zuletzt abgerufen am 13. März 2022).

25 Vgl. u.a. https://www.admin.ch/gov/de/start/dokumentation/medienmitteilungen.msg-id-77514.html; https://www.sbfi.admin.ch/sbfi/de/home/dienstleistungen/publikationen/publikationsdatenbank/s-n-2020-6/s-n-2020-6f.html (beide zuletzt abgerufen am 13. März 2022).

26 Florent Thouvenin et al., Ein Rechtsrahmen für Künstliche Intelligenz, Balsthal 2021 (https://bit.ly/3vtHLtn, zuletzt abgerufen am 13. März 2022).

seiner KI zu machen, diese zu dokumentieren und Massnahmen gegen entsprechende Mängel oder unerwünschte Auswirkungen zu treffen.

[40] Darüber hinaus sieht das künftige DSG wie schon die DSGVO einen Anspruch von betroffenen Personen vor, bei wichtigen Wertentscheiden *nicht* von einer KI beurteilt zu werden. Es handelt sich um die Regelung, die in der Schweiz und dem EWR im Wesentlichen festhält, dass wo Computer folgenschwerere Wertentscheide betreffend eines Menschen treffen, dieser die Anhörung und Überprüfung des Maschinenentscheids durch einen Menschen verlangen kann.[27] Dieser Regelung zugrunde liegt das bereits erwähnte Misstrauen gegenüber automatisierten Entscheidungen, auch wenn diese in vielen Fällen qualitativ besser sind als jene eines Menschen. Immerhin lässt sich diese Regelung dadurch umgehen, dass eine KI «nur» dazu benutzt wird, einen Entscheid vorzubereiten, dieser aber formal noch von einem Menschen aus Fleisch und Blut getroffen wird, auch wenn dieser von der Empfehlung der Maschine nicht abweicht. Dies ist mit ein Grund dafür, dass die neuen Regelungen über automatisierte Einzelentscheide in der Praxis kaum zum Tragen kommen.

[41] Derselbe Abwehrreflex gegenüber Maschinen liegt überdies dem Begriff des «Profilings» zugrunde. Es handelt sich gemäss Legaldefinition um eine maschinelle Bewertung eines Menschen punkto irgendwelcher Aspekte seiner Person. Wer eine andere Person aus dem Bauch heraus oder einem Schema im Kopf oder auf Papier bewertet, nimmt kein Profiling vor. Tut dies eine Maschine, ist es Profiling. Das Schweizer Parlament hat sich in den Beratungen zum revidierten DSG vor allem um die Legaldefinitionen in diesem Bereich gestritten und nicht realisiert, dass das Gesetz den Begriff selbst praktisch nicht verwendet, jedenfalls im privaten Bereich nicht. Bundesorgane hingegen brauchen für ein Profiling im Grundsatz eine formelle gesetzliche Grundlage – und diese wurde an vielen Orten schlicht vergessen. Auch in diesem Sinne reguliert der Datenschutz den Einsatz von KI jedenfalls bei Bundesorganen, wo er damit genaugenommen weitgehend ausgeschlossen ist, soweit es um Personendaten geht. Eine inhaltliche Diskussion fand hierzu freilich nicht statt.

8. Die Sache mit der Transparenz

[42] Interessant sind die Regelungen des Datenschutzrechts zu automatisierten Einzelentscheiden schliesslich deshalb, weil sie darauf abzielen, das Bedürfnis nach Nachvollziehbarkeit maschineller Entscheide zu befriedigen. Der Gesetzgeber versucht es dadurch zu lösen, dass er beim Einsatz von automatisierten Einzelentscheiden verlangt, dass der Verantwortliche die Existenz einer solcher Entscheidung offenlegt, aber auch eine Information über die «Logik, auf der die Entscheidung beruht».[28] Während im revidierten DSG über die Logik nur auf eine konkrete Anfrage hin im Rahmen des Auskunftsrechts informiert werden muss, sieht die DSGVO eine solche Information auch im Rahmen der Datenschutzerklärung einschliesslich von Angaben zur Tragweite und angestrebten Auswirkung des Einsatzes solcher automatisierten Einzelentscheide vor.[29]

---

27 Art. 21 DSGVO, Art. 21 revDSG; zum Ganzen vgl. David Rosenthal, Das neue Datenschutzgesetz, in: Jusletter 16. November 2020, Rz. 92 ff.

28 Art. 21 revDSG, Art. 25 Abs. 2 Bst. f revDSG.

29 Art. 13 Abs. 2 Bst. f DSGVO, Art. 14 Abs. 2 Bst. g DSGVO.

[43] Weil sich die «Logik» im Falle einer KI streng genommen im Hinweis auf den Einsatz von Mustererkennung erschöpft, führt diese Regelung im Datenschutzrecht nicht wirklich weiter. In der Praxis wird ein Verantwortlicher daher gut beraten sein, unter dem Titel der Logik auch Angaben über den Input und Output der KI zu machen, wobei sich dies in aller Regel auf Kategorien von Daten beschränken kann, allenfalls ergänzt durch erläuternde Beispiele. Mehr kann wohl auch unter dem Grundsatz der Transparenz nicht verlangt werden, der parallel zur Informationspflicht gilt und vom Verantwortlichen verlangt, dass die wesentlichen Eckwerte seiner Datenbearbeitung für eine betroffene Person erkennbar sind.[30] Die wissenschaftlichen Forschungen auf dem Bereich der Erklärbarkeit des Outputs von KI wird in der Praxis vor allem der Qualitätssicherung dienen und weniger der Transparenz der betroffenen Personen.

9. Datensicherheit als unterschätzte Herausforderung

[44] Das führt zum letzten datenschutzrechlichen Themenkomplex: Den Massnahmen zur Datensicherheit. Sie werden in der Diskussion datenschutzrechtlichen Risiken von KI oft unterschätzt, weil die Frage der Nachvollziehbarkeit einer KI und das Risiko der Diskriminierung als wichtiger erscheint. Von mangelnder Datensicherheit kann jedoch ein ebenso grosses Risiko ausgehen.

[45] Dieses Risiko zeigt sich in zweierlei Ausprägungen. Das klassische Risiko der Datensicherheit beim Einsatz von KI ist das Risiko einer unbefugten Offenlegung von Personendaten während des Trainings und in der produktiven Anwendung eines KI-Systems. Wird eine KI mit Input in Form von Personendaten benutzt, so müssen diese an die KI übermittelt werden, d.h. an die Trainings-Schnittstelle («Training API») oder an jene für die Generation von Output («Prediction API»). Diese Systeme und ihre Schnittstellen befinden sich in der Regel in einer Cloud oder auf anderen Systemen eines Drittanbieters, möglicherweise sogar verteilt auf zahlreichen Systemen, damit der Input dort bleiben kann, wo er anfällt, und damit Rechnerkapazitäten kombiniert werden können («Federated Learning Algorithms»). Viele Nutzer von KI-Systemen werden selbst nicht über die nötigen Ressourcen verfügen, um solche Systeme auf eigener Infrastruktur zu betreiben, sei es auf eigenen Systemen oder in einer Private Cloud. Auch ist der Einsatz eines «privaten» KI-Systems mitunter gar nicht erwünscht, da die Qualität einer KI mit wachsender Menge an guten Trainingsdaten in aller Regel steigt und es daher sinnvoll sein kann, solche Daten aus verschiedenen Quellen zu beziehen. Kommt dabei noch das Bedürfnis hinzu, diese Trainingsdaten für Re-Trainings über längere Zeiten aufzubewahren, können grosse Datensammlungen mit hoch-sensiblen Daten entstehen. Diese sind entsprechend vor unbefugten Zugriffen zu schützen, ganz abgesehen von den weiteren datenschutzrechtlichen Herausforderungen in solchen Fällen (Verhältnismässigkeit, Wahrung der Betroffenenrechte).

[46] Zum klassischen Risiko der Datensicherheit beim Einsatz von KI gehört ferner die Manipulation der Trainingsdaten. Über sie kann die spätere Funktion der KI beeinflusst werden. Eine solche Manipulation wird zwar den von den Trainingsdaten betroffenen Personen schaden, aber sie tangiert das KI-Modell als solches. Mittelbar kann dies daher jene Personen treffen, deren Daten die KI später analysiert und aufgrund eines manipulierten Trainings zu einem Ergebnis kommt, das im Vergleich zum Sollergebnis bei korrektem Training unrichtig ist. Damit liegt zugleich eine Persönlichkeitsverletzung vor.

30 Art. 5 Abs. 1 Bst. a DSGVO, Art. 4 Abs. 2 und 4 DSG, Art. 6 Abs. 2 revDSG.

10. KI-Modelle als Personendaten

[47] Das KI-spezifische Risiko der Datensicherheit ist herausfordernder. Es besteht im Risiko von Angriffen auf das KI-Modell selbst. Wenn wir vom KI-Modell sprechen, dann ist damit quasi das Hirn der KI gemeint, ihre Wissensbasis oder Denklogik. Es sind die Algorithmen, die aus dem Input (Bilder, Text, andere Daten) einen Output (typischerweise eine Vorhersage) erzeugen. Dieses Modell ist keine klassische Datenbank der Trainingsdaten, sondern die vom «Lernalgorithmus» abgeleiteten Erkenntnisse.

[48] Während früher die Ansicht vorherrschte, dass diese Erkenntnisse von den Trainingsdaten abstrahiert vorliegen und daher anonym sind, d.h. sich die ihnen zugrundeliegenden Personendaten nicht mehr als solche wiederfinden, wissen wir heute, dass dem häufig nicht so ist. Mit den geeigneten Methoden lassen sich die zum Training von KI-Modellen verwendeten Personendaten teilweise wieder extrahieren.

[49] Eine dieser Methoden, die «Membership Inference Attack», besteht darin herauszufinden, ob ein bestimmter Datenpunkt für das Training einer KI verwendet wurde.[31] Dies geschieht dadurch, dass der KI bestimmte Fragen gestellt und ihre Antworten analysiert werden, so etwa wie zuversichtlich die KI ist, dass ihre Prognose richtig ist. Um die datenschutzrechtliche Relevanz des Erfolgs eines solchen Angriffs zu verstehen, stelle man sich folgendes Szenario vor: Ein Krankenhaus hat anhand von zahlreichen Patientendaten eine KI zur Diagnose von Krebs entwickelt. Die Patienten haben dafür ihr Einverständnis gegeben, wie dies auch in Schweizer Krankenhäusern üblich ist. Ihnen wurde dabei versichert, dass sie anonym bleiben. Mit ihren Daten wurde das KI-Modell trainiert. Hierzu wurden von jedem Patienten zahlreiche Eckwerte eingegeben, damit diese zur Mustererkennung verwendet werden können. Das trainierte Modell kommt wiederum an Krankenhäusern weltweit zum Einsatz. Somit können eine beliebige Anzahl von Personen es nutzen. Das KI-Modell kann mit den Werten eines Patienten gefüttert werden (Input) und es berechnet basierend darauf eine Diagnose (Output). Wurde das KI-Modell nicht gegen eine Membership Inference Attack geschützt, kann es möglich sein, mit bestimmten Eckwerten einer Person herauszufinden, ob ihre Daten für das Training des Modells benutzt worden waren und folglich auch, ob sie Krebs hatte. Dies geht vereinfacht gesagt so: Wurde der KI beigebracht, dass jemand mit den Parametern genau dieser Person Krebs hat und ist eine Kombination genau dieser Parameter höchst unwahrscheinlich, gibt die KI aber trotzdem an, dass sie sich sicher ist, so darf davon ausgegangen werden, dass die KI die betreffende Person «wiedererkannt» hat, d.h. sie Teil ihres Trainingsdatensatzes war und die Diagnose nicht nur eine Prognose, sondern eine Tatsache ist.

31 Vgl. etwa Reza Shokri, Marco Stronati, Congzheng Song, Vitaly Shmatikov, Membership Inference Attacks against Machine Learning Models, IEEE Symposium on Security and Privacy, 2017, arXiv:1610.05820 (https://arxiv.org/abs/1610.05820); Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu, and Wenqi Wei, Demystifying Membership Inference Attacks in Machine Learning as a Service, arXiv:1807.09173v2 [cs.CR] 1 Feb 2019 (https://arxiv.org/pdf/1807.09173.pdf); Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, Florian Tramer, Membership Inference Attacks From First Principles, arXiv:2112.03570v1 [cs.CR] 7 Dec 2021 (https://arxiv.org/pdf/2112.03570.pdf) (alle Links zuletzt abgerufen am 13. März 2022).

11. Datenlecks als konkretes Risiko

[50] Das sind nicht bloss theoretische Überlegungen, und nebst der Membership Inference Attack existieren auch andere Methoden, um an Personendaten (oder Geschäftsgeheimnisse) in einem KI-Modell zu gelangen.[32] Mit diesen ist es je nach Modell möglich, ursprüngliche Personendaten zu rekonstruieren. Die Forscher sprechen dabei von «Datenlecks», welche die betroffenen KI-Modelle aufweisen. Entsprechende Experimente wurden mit Texten wie auch mit Bildern durchgeführt. So gelang Forschern aus Frankreich durch Auswertung der Website *thispersondoesnotexist.com*, auf welcher eine KI Bilder nicht existierender Personen erzeugt, die für ihr Training verwendeten Ursprungsbilder zu rekonstruieren.[33] Anderen Forschern gelang es, die für das Training einer Sprach-KI verwendeten Texte zu extrahieren.[34]

[51] Solche Beispiele widerlegen die Vorstellung, wonach KI-Modelle stets abstrakt und daher anonym sind. Forscher unter anderem der EPFL haben obendrein gezeigt, dass es je nach KI-Modell sogar möglich ist, durch ein gezieltes Austesten einer KI die ihnen antrainierte «Erfahrung» mehr oder weniger vollständig zu replizieren.[35] Wer seine KI der Öffentlichkeit zugänglich macht, muss unter vergleichbaren Voraussetzungen damit rechnen, dass sie mitsamt eines Teils der davon extrahierbaren Daten gestohlen werden kann.

[52] Das alles lässt auch deshalb aufhorchen, weil es zeigt, dass es keineswegs unproblematisch ist, Personendaten (oder auch Geschäftsgeheimnisse) für das Training einer fremden oder für Dritte zugänglichen KI bereitzustellen. Lassen sich Personendaten aus einem fertigen KI-Modell extrahieren, genügen auch die üblichen datenschutzrechtlichen Vorkehrungen nicht mehr. Der Rechtfertigungsgrund der Bearbeitung von Personendaten für nicht personenbezogene Zwecke[36] setzt beispielsweise voraus, dass die Personendaten so rasch wie möglich anonymisiert und die Ergebnisse (hier: das KI-Modell) in jedem Fall so veröffentlicht werden, dass die betroffenen Personen nicht bestimmbar sind. Das Kriterium der Vereinbarkeit der Datenbearbeitung mit dem Zweck, für welche die Daten erhoben wurden, verlangt im Ergebnis meist dasselbe. Sind KI-Modelle im erwähnten Sinne angreifbar, sind diese Voraussetzungen möglicherweise nicht mehr gegeben und KI-Modelle müssen als Personendaten qualifiziert werden. Werden sie in der Folge Dritten zur Verfügung gestellt, kann dies einer Bekanntgabe von Personendaten gleichkommen – jedenfalls wenn angenommen werden muss, dass die Nutzer entsprechende Angriffsmethoden einsetzen.

[53] Immerhin gibt es inzwischen auch Ansätze, wie eine Extraktion von Personendaten aus einer KI oder das Wiedererkennen von betroffenen Personen verhindert werden kann. Einer dieser

---

32 Vgl. etwa Milad Nasr, Reza Shokri, Amir Houmansadr, Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning, 2019 IEEE Symposium on Security and Privacy (https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8835245, zuletzt abgerufen am 13. März 2022).

33 Ryan Webster, Julien Rabin, Loic Simon, Frederic Jurie, This Person (Probably) Exists. Identity Membership Attacks Against GAN Generated Faces, arXiv:2107.06018v1 [cs.CV] 13. Juli 2021 (https://arxiv.org/pdf/2107.06018.pdf, zuletzt abgerufen am 13. März 2022).

34 Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, Colin Raffel, Extracting Training Data from Large Language Models, 2020, arXiv:2012.07805 (https://arxiv.org/abs/2012.07805);

35 Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, Thomas Ristenpart. Stealing Machine Learning Models via Prediction APIs. USENIX (https://arxiv.org/abs/1609.02943) (alle Links zuletzt abgerufen am 13. März 2022).

36 Art. 31 Abs. 2 Bst. e revDSG.

Ansätze ist als «Differential Privacy» bekannt[37], welchem vereinfacht gesagt der Gedanke zugrunde liegt, dass der Input so manipuliert wird, dass es beim späteren Einsatz der KI nie einen vollständigen Treffer mit realen Daten geben kann. Der Nachteil dieses Ansatzes ist die verminderte Genauigkeit des KI-Modells. Er wird auch bei der Publikation von Daten verwendet, um die Identifizierbarkeit einzelner Personen zu verhindern. Methoden zur Quantifizierung des Risiko eines Inference Attacks bzw. zur Auditierung von KI-Modellen sind ebenfalls bereits vorgestellt worden.[38]

12. Fazit

[54] Der Einsatz von KI-Systemen wirft erwartungsgemäss viele datenschutzrechtliche Fragen auf. Diese können jedoch zu einem guten Teil durch eine entsprechende Anwendung des Schweizer Datenschutzrechts vernünftig beantwortet werden. Die praktische Herausforderung besteht vor allem darin, erstens die Qualität eines KI-Modells zu beurteilen und zu dokumentieren und zweitens zu verhindern, dass sich die für das Training von KI-Modellen verwendeten Personendaten aus diesen Modellen extrahieren lassen.

[55] Beides erfordert einiges an Spezialwissen und damit den Beizug von Spezialisten. Darüber hinaus ist der Einsatz von KI-Systemen nicht a priori problematischer als andere Prognosemethoden, auch wenn das Misstrauen gegenüber Entscheiden einer Maschine ungleich grösser ist als gegenüber Entscheiden eines Menschen. Doch auch derjenige, der keine KI-Systeme einsetzen will, muss inzwischen datenschutzrechtlich auf der Hut sein. Denn immer häufiger versuchen Service Provider und andere Geschäftspartner, die Daten ihrer Kunden auch für den Aufbau ihrer KI-Modelle zu verwenden. Hierbei ist Vorsicht geboten, damit dies angesichts der bereits genannten Herausforderungen nicht zu unerwünschten Datenlecks führt. Und kommt es zu einem missbräuchlichen Einsatz von Daten für das Training einer KI, stellt sich die Frage, wie solche Verstösse wirksam geahndet werden können, selbst wenn das KI-Modell nicht mehr als Sammlung von Personendaten gelten sollte. Einen konsequenten Ansatz vermitteln erste Verfahren aus den USA: Die dortige Wettbewerbsbehörde FTC verpflichtete bereits mehrere Unternehmen, die auf unzulässige Weise Daten für den Aufbau einer KI verwendet hatten, kurzerhand zur Löschung nicht nur der Daten, sondern auch der damit trainierten KI.[39]

David Rosenthal, lic. iur., ist Lehrbeauftragter der ETH Zürich und Universität Basel und Partner einer Schweizer Wirtschaftskanzlei. Er ist erreichbar unter drosenthal@vischer.com.

37 https://en.wikipedia.org/wiki/Differential_privacy (zuletzt abgerufen am 13. März 2022).

38 Vgl. Reza Shokri, Auditing Data Privacy For Machine Learning, Präsentation vom 2. Februar 2022 (https://www.usenix.org/conference/enigma2022/presentation/shokri, zuletzt abgerufen am 13. März 2022).

39 Vgl. https://www.ftc.gov/news-events/news/press-releases/2021/01/california-company-settles-ftc-allegations-it-deceived-consumers-about-use-facial-recognition-photo, https://www.ftc.gov/system/files/ftc_gov/pdf/wwkurbostipulatedorder.pdf (beide zuletzt abgerufen am 2. April 2022).