heise Developer

Künstliche Intelligenz trifft Datenschutz

Know-how 15.03.2019 08:55 Uhr
Martin Rost

Wer Daten mit Machine Learning verarbeitet, muss auf das Einhalten der Datenschutzanforderungen achten. Künstliche Intelligenz bringt neue Risikotypen für den Datenschutz mit sich.

Organisationen, die mit einer Komponente wie einem künstlichen neuronalen Netz personenbezogene Daten verarbeiten, erzeugen hohe Risiken für Betroffene. Bei hohem Risiko verlangt die Datenschutz-Grundverordnung (Art. 35 DSGVO) das Durchführen einer Datenschutz-Folgenabschätzung (DSFA).

Künstliche Intelligenz und DSGVO

Das Verständnis von künstlicher Intelligenz (KI) und Machine Learning (ML) reicht von regelbasierten Entscheidungsmodellen auf Grundlage gut erforschter Regressionsanalysen bis zu den subsymbolischen Strukturen der künstlichen neuronalen Netze (KNN). Bei den KI-Modellen lassen sich Lernstile, -aufgaben und -verfahren unterscheiden, die von linearer Regression über Bayessche Inferenz, Clustering und K-Means bis zur klassischen Backpropagation reichen. Zur Klassifikation von KI-Systemen ist eine **Abhandlung der Fraunhofer Gesellschaft [1]** hilfreich.

Wenn im Folgenden von KI die Rede ist, ist deren aktuell stärkste Ausprägung gemeint, nämlich KNN mit Deep Learning. Die Datenschutz-Herausforderungen dabei sind immens, wenn man **dem Mathematiker und KI-Spezialisten Gunter Laßmann [2]** folgt: "Deep-Learning-Systeme sind also nicht prüfbar, nicht evaluierbar, ändern Ihre Eigenschafen, liefern keinerlei Begründung, sind leicht zu manipulieren, willkürlich aufgebaut und immer ungenau." [1]. Einer solchen Technik sollen Menschen anvertraut werden?

Bei den erstaunlichen Leistungen der KI in den letzten beiden Jahrzehnten bei Spielen wie Schach, Jeopardy, Go oder Poker sind die riskanten Eigenschaften weitgehend irrelevant. Im militärischen, industriell-produktiven und Alltagskontext von Menschen spielen sie dagegen eine große Rolle [2] und [3]. Wenn sich für KNN Datenschutzmaßnahmen entwickeln lassen, dann auch für leichter beherrschbare Automationsverfahren.

Verständnis von Datenschutz

Jede Verarbeitung personenbezogener Daten durch eine Organisation ist ein Grundrechtseingriff für davon betroffene Personen und erzeugt ein bestimmtes Set an Risiken, die es ohne diese Verarbeitung nicht gäbe.

Die Risiken, die von Hackern oder illoyalen Mitarbeitern ausgehen können, sind davon nur eine Untermenge. Insofern gilt als generalisiertes Angreifermodell im Datenschutz: "Jede Organisation, gerade und auch die rechtlich zur Verarbeitung befugte, ist ein Angreifer!"

Organisationen nehmen unvermeidlich in den Transaktionen gegenüber ihren Bürgern, Kunden, Patienten, Lieferanten und Mitarbeitern eine Fremdbestimmung des Handelns, Denkens und Fühlens dieser Personen vor. Bei dieser Kontrolle organisierter Fremdbestimmung geht es nicht um die Bekämpfung organisierter Boshaftigkeit, auch nicht um Schuld durch Organisation, sondern um den sachgerecht organisierten Umgang mit dem, was die Organisation von Rechts wegen sachlich etwas angeht und was nicht.

Zudem muss eine Organisation vermeiden, Personen zu Objekten von (KI-)Automaten zu machen, weil es dann sogar gänzlich an der Legitimation für eine Datenverarbeitung fehlt. Eine Konstellation, die aus Subjekten Objekte macht, ist nicht grundrechtskonform und einwilligungsfähig. Artikel 22 DSGVO legt fest: "Die betroffene Person hat das Recht, nicht einer ausschließlich auf einer automatisierten Verarbeitung – einschließlich Profiling – beruhenden Entscheidung unterworfen zu werden." Datenschutz formuliert die Risiken beziehungsweise den Schutzbedarf von betroffenen Menschen, während die IT-Sicherheit den Schutzbedarf der eigenen Leute und Organisation adressiert.

Zur Bestimmung der Höhe der Risiken betroffener Personen haben sich die Datenschutz-Aufsichtsbehörden europaweit auf einen neun Kriterien umfassenden Katalog geeinigt (vgl. **Art. 29, Gruppe 2017 [3]**):

1. Bewerten oder Einstufen,
2. automatische Entscheidungsfindung,
3. systematische Überwachung,
4. vertrauliche oder höchst persönliche Daten,
5. Datenverarbeitung im großen Umfang,
6. Abgleichen oder Zusammenführen von Datensätzen,
7. Daten zu schutzbedürftigen Betroffenen,
8. innovative Nutzung oder Anwendung neuer technologischer oder organisatorischer Lösungen,
9. Betroffene werden an der Ausübung eines Rechts oder der Nutzung einer Dienstleistung beziehungsweise Durchführung eines Vertrags gehindert.

Die Regel zur Entscheidung lautet: Wenn aus diesem Katalog mindestens zwei Kriterien zutreffen, besteht ein hohes Risiko für betroffene Personen und eine DSFA ist durchzuführen. Bei Verfahren mit KI-Komponenten treffen häufig gleich alle neun Kriterien zu. Daraus folgt wiederum die Regel: Wenn KI bei einer Verarbeitung personenbezogener Daten zum Einsatz kommt, ist eine DSFA obligatorisch.

Der Zweck einer DSFA besteht darin, die Risiken zu bestimmen und mit der Gestaltung des Verfahrens sowie flankierender Schutzmaßnahmen auf das geringst mögliche Maß zu mildern – nicht gemeint sind die Risiken von Haftungsschäden durch Unfälle oder Vergleichbares; das wäre ein vollkommen anderes Thema. Artikel 5 DSGVO listet die materiellen Anforderungen des Datenschutzrechts auf. Das Risiko bei KI-basierter Verarbeitungstätigkeit besteht somit darin, dass Organisationen gegenüber Personen die Anforderungen aus Artikel 5 nicht wirksam erfüllen. Was unter einer Verarbeitungstätigkeit und einem personenbezogenem Datum zu verstehen ist, definiert Artikel 4 DSGVO.

Mittlerweile wurden diese normativen Anforderungen nach dem methodischen Vorbild der IT-Sicherheit des Grundschutzes nach BSI durch Schutz-beziehungsweise Gewährleistungsziele im Kontext **eines Standard-Datenschutzmodells [7]** in funktionale Anforderungen transformiert.

Das heißt für die Praxis von KI-Entwicklern, dass sie für eine DSFA ihrer KI erstens einen Anwendungsfall der

Verarbeitung formulieren müssen und zweitens die dafür zu treffenden Schutzmaßnahmen nicht aus dem komplexen Datenschutzrecht extrahieren, sondern aus den funktional ganz gut verstandenen sechs Schutzzielen herleiten können.

Anschließend müssen die Verfügbarkeit, Integrität, Vertraulichkeit, Transparenz, Nichtverkettbarkeit (inkl. Datenminimierung) und Intervenierbarkeit einer Verarbeitung gewährleistet und sichergestellt werden, mit Resilienz als einer zusätzlichen Anforderung bezüglich aller Schutzziele [4]. Einzelne Datenschutz-Aufsichtsbehörden haben begonnen, zu allen Zielen **konkrete Referenz-Schutzmaßnahmen [8]** auszuweisen.

Spezifische Risiken durch KI

Das Auflisten aller Gewährleistungsziele [5] und der Anforderungen konventioneller IT-Sicherheit auf dem nicht kognitiven Layer der IT würden im Rahmen des Artikels zu weit gehen. Für eine Übersicht sollen die Antworten auf folgende drei Fragen helfen:

1. Was ist zu prüfen?
2. Wie lässt sich die Zweckbindung sichern?
3. Wie kann eine KI gestoppt werden?

Prüfbarkeit einer KI herstellen

Eine wesentliche Anforderung an Verarbeitungen mit KI-Komponenten ist die Transparenz, genauer nach deren Prüffähigkeit. So verlangt Artikel 13 DSGVO: Es sind "(...) aussagekräftige Informationen über die involvierte Logik sowie die Tragweite und die angestrebten Auswirkungen einer derartigen Verarbeitung für die betroffene Person" zu geben. Das heißt: KI-Systeme müssen für eine Soll-Ist-Bilanzierung im Hinblick auf bestimmte Eigenschaften zugänglich sein.

Genauer formuliert: KI-Systeme sind auf der Grundlage der mit den Schutzzielen verbundenen Maßnahmen zu spezifizieren, der Betrieb ist unter Ausweis einer Prüfmethodik zu dokumentieren und anhand von Protokollen – durch aktive Selbstauskünfte und bei Interaktion mit anderen Systemen durch Fremdprotokolle – nachvollziehbar zu gestalten. Ein Datenschutz-Managementsystem hat schließlich dafür zu sorgen, dass Datenschutzdefizite integer festgestellt und behebbar sind und tatsächlich vom verantwortlichen Systembetreiber behoben werden.

Durch KI beziehungsweise ML ist eine neue Klasse an Transparenz in der Spezifikationsphase bezüglich der Daten entstanden, nämlich die Qualität der Aufbereitung der Daten für eine KI, das sogenannte Kuratieren, zu sichern. In diesem Sinne müssen Entwickler die folgenden Eigenschaften dokumentieren, um bei einer DSFA für ein KI-Verfahren das Schutzziel Transparenz zu erfüllen [6]:

- die Herkunft der Daten,
- die Form der Veredlung (Definieren, Sammeln, Selektieren, Umwandeln, Verifizieren) und Anreicherung der Rohdaten zu Modell- oder Trainingsdaten,
- der Lernstil (Supervised Learning, Unsupervised Learning, Reinforcement Learning),
- die verwendeten Lernmodelle (von Regressionsmodell bis KNN mit ML),
- der potenzielle Einsatz einer speziellen KI-Komponente,
- menschliche Beteiligung an den Entscheidungsfindungen innerhalb einer Verarbeitung,
- die Institutionen, die die Komponenten des KI-Systems hergestellt und über die Auswahl, Konfiguration, Implementation und Betrieb der verwendeten KI-Technik, das Kuratieren der Daten, das Training und der Auswahl der Modelle entschieden haben,
- ein Gutachten zur Vollständigkeit der Repräsentativität der von der KI beherrschten Wissensdomäne (die sich historisch ändert),
- die Implementierung des KI-Algorithmus, insbesondere der regelbasierten Instruktionen und Entscheidungen,

- der Einbau von Prüfankern, Prüfagenten, Selbstdokumentationsmechanismen.

Zweckbindung einer KI sicherstellen

Die Zwecksetzung für die Nutzung einer KI geschieht durch den Verantwortlichen und muss legitim sein. Die nachfolgende Zweckdefinition für die Verarbeitung muss rechtskonform erfolgen, die Zwecktrennung von anderen, inhaltlich benachbarten Verarbeitungstätigkeiten muss scharf und entschieden sein, damit sich die Zweckbindung der Datenverarbeitung über alle Weisungshierarchien der Organisation und alle Ebenen der technischen Infrastruktur hinweg überprüfen beziehungsweise nachweisen lässt.

Die wesentliche generische Maßnahme zum Beherrschen durch Zweckbindung ist die funktionale Kapselung, Isolation beziehungsweise Trennung von Komponenten, um kleinteilige Prüfungen für Teilfunktionen durchführen und Bedingungen für Aktoren formulieren zu können. Die generelle Strategie dabei ist die, das unvermeidliche Maß an Nichtkalkulierbarkeit beziehungsweise die erwartete Unsicherheit möglichst sicher zu isolieren.

Anders formuliert geht es darum, Inseln zu bilden, deren Vertrauensniveaus beispielsweise auf der Grundlage eines statistischen Fehlerverteilungsmodells kalkulierbar sind [7]. Ein schwerer Fehler in einer Komponente darf sich bei einem komplexen Automaten nicht auf das gesamte System ausbreiten können (Konzept Brandmauer oder Schiffsschott). Auf keinen Fall darf bei einem KNN passieren, dass durch geringfügige Änderungen der Trainingsdaten das "katastrophische Vergessen" von zuvor stabil Abgebildetem einsetzt. Für Verantwortliche und Betroffene muss zudem jederzeit klar sein, in welchem Zustand sich alle Komponenten eines größeren IT-Gesamtsystems befinden, das in der Praxis zumeist aus verschiedenen Typen von KI-Modellen besteht.

Um die Kalkulierbarkeit zu verbessern, lassen sich zwei gegensätzliche Strategien verfolgen: Trivialisierung und Komplexitätssteigerung. Für Ersteres sollte die Modellierung weg von KI/ML hin zu Entscheidungsbäumen gehen, die beispielsweise auf linearer Regression oder Cluster-Bildungen basieren. Überspitzt lautet die Strategie "Weg von der bloßen Korrelation durch Musteradaptionen und hin zur theoriegestützten, regelbeherrschbaren Kausalität". KI-Entwickler müssen insofern nachweisen können, dass ihre Entscheidungskomponenten nicht weniger riskant als mit KNN/ML umsetzbar sind, selbst wenn die Entstehungskosten dafür um vieles höher sind.

Für den gegenteiligen Ansatz der Komplexitätssteigerung ließe sich eine zweite KI, die durchaus ebenfalls auf KNN/ML basieren kann, auf das Einhalten des Zwecks der Produktions-KI ansetzen. Die zweite KI warnt oder greift besser noch unmittelbar regulierend ein. Diese Strategie ließe sich bezeichnen als "Feuer mit Gegenfeuer unter Kontrolle halten" Es zeichnet sich ab: Die Vielschichtigkeit einer grundrechtskonformen Regulation komplexer Verarbeitungstätigkeiten ist dermaßen groß, dass ein tatsächlich wirksamer Datenschutz auf die Entwicklung von Prüf-KI angewiesen sein wird.

Für eine Prüf-KI ist zu fordern, dass sie unabhängig von der Produktions-KI agiert. Diese Forderung nach Unabhängigkeit durch Trennung und Isolation besteht streng genommen für den Hersteller der Hardware, des Betriebssystems und der Middleware bis hin zu den kognitiven Ebenen und deren Kuratoren, Customizers und Trainern.

Es ist somit geboten, dass gerade innerhalb einer Domäne unterschiedliche Ökosysteme für KI – neben dem amerikanischen und dem chinesischen mindestens noch ein europäisches – ausgebildet werden, um zumindest über integre Prüfverfahren zu verfügen, sollte die KI der Produktionsebene auf Systemen bekannter Monopolhersteller laufen. Die Einhaltung des definierten Zwecks und das Durchsetzen der Zweckbindung für eine KI zu sichern und nachzuweisen, dürfte die Hauptschwierigkeit einer DSFA bilden.

Intervenierbarkeit bei einer KI

Je smarter ein Automat assistiert, desto dringlicher stellt sich eine normative Frage: Soll die Maschine letztlich den Piloten oder der Pilot die Maschine führen (instruktiv die sechsstufige Automationsskala bezüglich automatisierten Fahrens)? Die aus der Antwort ableitbare Regel hängt abstrakt vom Prüfkriterium ab und lautet konkret: Systeme sind so einzurichten, dass bei wechselnden Anforderungen die Steuerung beziehungsweise Assistenz wechseln kann.

Aus Datenschutzsicht gilt dabei dogmatisch zu fordern, dass sich ein KI-System mit Personenbezug ausschalten lässt, ohne das Vorgehen als Notfall zu gestalten oder dem Nutzer besondere Haftungskosten aufzubürden. Ein Ausschaltknopf operationalisiert perfekt die Einwilligung für den unmittelbar Betroffenen. Dabei kann die Skala der Intervention beim nachträglichen Ausfüllen eines Beschwerdeformulars beginnen und über den Ausschaltknopf an jedem KI-System bis zur Totmannschaltung reichen, mit der eine KI nur dann läuft, wenn ein Mensch diese aktiv fortlaufend überwacht.

Intervenierbarkeit muss zudem für Organisationen und die Gesellschaft insgesamt sichergestellt sein. Dabei reicht die Skala von Maßnahmen zur obligatorischen Prüfung und Freigabe riskanter KI durch Kontrollbehörden beispielsweise analog zur Freigabe von Medikamenten oder aus dem Umweltschutz- und Kartellbereich, bis zu einer KI-"Feuerwehr" oder der Polizei, die passende rechtliche Befugnisse erhalten müssen.

Auch für solche Interventionen müssen die KI-Systeme den Akteuren entgegenkommen. Wieder lohnt ein Blick in die Autofahrautomation, denn dort werden derzeit überzeugend skalierbare Maßnahmen entwickelt und getestet, um KI-Systeme allseits verträglich herunterzufahren. Derart starke Grundrechtseingriffe sollten staatlichen Institutionen vorbehalten sein und keinesfalls privaten Interessenten wie Herstellern oder Versicherungen zugestanden werden.

Bei der Gestaltung von KI sollten Nutzer sowohl an der Festlegung der Risikomodelle als auch beim Kuratieren der Daten beteiligt werden, um unter anderem das Risiko von Diskriminierungen zu verringern. Das heißt konkret, dass bei Architekturentscheidungen im Kontext der KI tatsächlich alle gesellschaftlich relevanten Interessenverbände zu beteiligen sind. Und bei persönlichen Asisstenzsystemen sollte das Paradigma "nutzerkontrolliertes Kuratieren" gelten, wonach das Training der Systeme mit den Nutzerdaten unter ausschließlicher Kontrolle der Betroffenen erfolgt.

Ein Framework für DSFA

Ein in der Praxis bewährtes Framework zur systematischen Durchführung einer DSFA gemäß Artikel 35 DSGVO hat **das Privacyforum entwickelt [9]**. Es gliedert den Prozess zur Durchführung einer DSFA in vier Abschnitte:

- Abschnitt A fordert zur Klärung der Voraussetzungen auf wie das Durchlaufen rechtlicher Prüfungen oder den Aufbau eines Projektmanagements, denn gemäß DSGVO sind die Datenschutzbeauftragten nicht für die Durchführung einer DSFA verantwortlich.
- Abschnitt B strukturiert die Durchführung der eigentlichen Risikoabschätzung entlang der sechs Schutzziele des Datenschutzes, die im DSFA-Bericht für die Leitungsebene mündet.
- Artikel 35 verlangt darüber hinaus die Implementierung von Schutzmaßnahmen, die Gegenstand des Abschnitts C ist.
- Abschnitt D umfasst Maßnahmen, mit denen die Verfahrensverantwortlichen die Wirksamkeit der Schutzmaßnahmen und damit auch die Compliance der Verarbeitung insbesondere gegenüber Datenschutz-Aufsichtsbehörden nachweisen können.

Fazit

Operativer Datenschutz ist ein Projekt der Moderne. Mit Datenschutz wurden Instrumente entwickelt, um proaktiv und konstruktiv beherrschbare KI-Systeme zu entwickeln und zu betreiben. In diesem Kontext rein auf ethische Prinzipien zu setzen statt auf prüfbare und vor Gericht einklagbare Datenschutzanforderungen,

nützt lediglich denjenigen, die an einer Verteidigung der bürgerrechtlich verfassten modernen Gesellschaft mit selbstbewussten Bürgern kein Interesse haben. Was bislang vollständig fehlt, aber unerlässlich ist, sind Aktivitäten zum Entwickeln einer Datenschutz-Prüf-KI. (**rme [10]**)

Martin Rost
ist stellvertretender Leiter des Technikreferats des Unabhängigen Landeszentrums für Datenschutz, Schleswig-Holstein, sowie Leiter der Unterarbeitsgruppe Standard-Datenschutzmodell des Arbeitskreis Technik der Konferenz der Datenschutzbeauftragten Deutschlands.

Literatur

1. Günter Laßmann; **Asimovs Robotergesetze – Was leisten sie wirklich? [11]**; E-Book; Heise Medien, 2017
2. Max Tegmark; Leben 3.0 – Mensch sein im Zeitalter Künstlicher Intelligenz; 2. Aufl., Ullstein 2017
3. Thomas Ramge; Mensch und Maschine – Wie künstliche Intelligenz und Roboter unser Leben verändern; 2. Aufl., Reclam2018
4. Susan Gonscherowski, Marit Hansen, Martin Rost; Resilienz – eine neue Anforderung aus der Datenschutz-Grundverordnung, in: DuD – Datenschutz und Datensicherheit, 42. Jahrgang, Heft 7: 442-446; 2018
5. Martin Rost; Künstliche Intelligenz, in: DuD – Datenschutz und Datensicherheit, 42. Jahrgang, Heft 9: 558-565; 2018
6. Nicholas Diakopoulos, Oliver Deussen: Brauchen wir eine Rechenschaftspflicht für algorithmische Entscheidungen? In: Informatik-Spektrum, 40. Jahrgang, Nr. 4: 362-366; 2017
7. Florian Müller; **Richtig entscheiden – Einführung in die probabilistische Programmierung [12]**; iX 2019/02: 108-114

URL dieses Artikels:

`http://www.heise.de/-4337027`

Links in diesem Artikel:

[1] `https://www.bigdata.fraunhofer.de/content/dam/bigdata/de/documents/Publikationen/Fraunhofer_Studie_M`
[2] `https://www.heise.de/tp/buch/telepolis_buch_3912357.html`
[3] `https://ec.europa.eu/newsroom/document.cfm?doc_id=47711`
[4] `https://www.m3-konferenz.de/?source=12`
[5] `https://www.m3-konferenz.de/programm.php?source=12`
[6] `https://www.m3-konferenz.de/lecture.php?id=7707&source=12`
[7] `https://www.datenschutz-mv.de/datenschutz/datenschutzmodell/`
[8] `https://www.datenschutz-mv.de/datenschutz/datenschutzmodell/`
[9] `https://www.forum-privatheit.de/forum-privatheit-de/publikationen-und-downloads/veroeffentlichungen-des-forums/themenpapiere-white-paper/Forum-Privatheit-WP-DSFA-3-Auflage-2017-11-29.pdf`
[10] `mailto:rme@ct.de`
[11] `https://www.heise.de/tp/buch/telepolis_buch_3912357.html`
[12] `https://www.heise.de/select/ix/2019/2/1549106182574415`