Position

Bundesverband Digitale Wirtschaft e.V. · Schumannstraße 2 · 10117 Berlin

29.08.2025

Philipp Hagen, Director Legal Affairs & Data Privacy, **Hagen@bvdw.org**
Janek Kuberzig, Public Affairs Manager Data & (Future) Tech, **Kuberzig@bvdw.org**

Konsultation zum datenschutzkonformen Umgang mit personenbezogenen Daten in KI-Modellen

Der Bundesverband Digitale Wirtschaft (BVDW) e. V. ist die Interessenvertretung für in Deutschland ansässige Unternehmen, die digitale Geschäftsmodelle betreiben oder deren Wertschöpfung auf dem Einsatz digitaler Technologien beruht. Die Grundlage dafür ist die intelligente Verbindung von Daten und Kreativität bei gleichzeitig maßgeblicher Orientierung an ethischen Prinzipien. Mit seinen über 600 Mitgliedsunternehmen – von großen und kleinen Digitalunternehmen über Agenturen bis hin zu Publishern – vertritt der Verband die Belange der digitalen Wirtschaft gegenüber Politik und Gesellschaft. Sein Netzwerk von Expertinnen und Experten liefert mit Zahlen, Daten und Fakten Orientierung zu einem zentralen Zukunftsfeld.

Der BVDW begrüßt ausdrücklich die Konsultation der Bundesbeauftragten für Datenschutz und Informationsfreiheit (BfDI) zum datenschutzkonformen Umgang mit personenbezogenen Daten in KI-Modellen. KI-Modelle, insbesondere Large Language Models (LLMs), stellt eine Schlüsseltechnologie für die digitale Transformation dar und erfordert eine sorgfältige, differenzierte und zugleich innovationsfreundliche Regulierung.

Aus Sicht der Digitalen Wirtschaft ist entscheidend, dass die Diskussion nicht durch abstrakte Worst-Case-Szenarien geprägt wird, sondern an den tatsächlichen technischen Gegebenheiten und realistischen Risiken ansetzt. LLMs sind keine Datenbanken, in denen personenbezogene Daten in abrufbarer Form gespeichert sind. Vielmehr handelt es sich um hochdimensionale mathematische Systeme, die statistische Muster aus Trainingsdaten abstrahieren und daraus probabilistische Vorhersagen generieren. Die Annahme, dass LLMs personenbezogene Daten „enthalten“ oder kontinuierlich „verarbeiten“, greift daher zu kurz und führt zu rechtlich wie praktisch problematischen Konsequenzen. Entscheidend ist, ob unter

Deutsches Lobbyregister R000257 **EU-Transparenzregister** 479540331468-69 **USt-IdNr.** DE196415580 **St-Nr.** 27/620/62930

Berücksichtigung von Aufwand, Kontext und verfügbaren Mitteln eine Identifizierung tatsächlich möglich ist. In den meisten praktischen Szenarien ist eine Re-Identifizierung von Privatpersonen faktisch ausgeschlossen.

Um mögliche Risiken der Memorisierung und unbeabsichtigten Wiedergabe zu minimieren, verfolgen die Unternehmen der Digitalen Wirtschaft einen ganzheitlichen Ansatz, der sich über den gesamten Lebenszyklus der LLMs erstreckt. Schon bei der Datenauswahl werden sensible Quellen ausgeschlossen, Inhalte bereinigt und Duplikate entfernt, damit einzelne Datenpunkte nicht überproportional ins Gewicht fallen. Während des Trainings werden Regularisierungstechniken und Fine-Tuning eingesetzt, um Überanpassungen zu verhindern. In der Bereitstellung sorgen Ausgabefilter und Monitoring dafür, dass sensible Inhalte nicht an Endnutzer*innen gelangen. Ergänzt wird dies durch Red-Teaming, gezielte Tests und neue probabilistische Methoden, die Risiken realitätsnah erfassen. So entsteht ein Schutzsystem, das sich nicht auf ein einzelnes Verfahren stützt, sondern auf die Kombination abgestimmter Maßnahmen, die auf LLMs, Daten und Einsatzzweck zugeschnitten sind.

Empirische Erfahrungen zeigen, dass die tatsächliche Wahrscheinlichkeit einer personenbezogenen Rekonstruktion verschwindend gering ist. Die wenigen dokumentierten Fälle betreffen fast ausschließlich stark überrepräsentierte Daten wie häufig vorkommende Phrasen oder Informationen über Personen des öffentlichen Lebens, deren Bekanntheitsgrad ohnehin hoch ist. Für private Informationen, die in Trainingsdaten nur am Rande vorkommen, gibt es bislang keine realweltlichen Nachweise für ein relevantes Extraktionsrisiko.

Auch die Frage der Betroffenenrechte lässt sich im Einklang mit der DSGVO pragmatisch lösen. Da LLMs keine editierbaren personenbezogenen Datensätze enthalten, ist eine nachträgliche Löschung oder Berichtigung innerhalb des LLMs nicht möglich. Effektiver Schutz wird stattdessen dadurch erreicht, dass konkrete Outputs, die personenbezogene Inhalte betreffen, unterdrückt oder blockiert werden können. Dies entspricht der Logik der EuGH-Rechtsprechung zu Suchmaschinen: Nicht das LLM selbst, sondern seine nutzerbezogene Ausgabe ist der Ort, an dem Rechte besser gewahrt werden.

Position

Aus Sicht der Digitalen Wirtschaft ist es wichtig, die Diskussion nicht auf theoretische Risiken zu verengen, sondern praktikable Schutzstandards zu etablieren. Dazu gehören gemeinsame Definitionen von Schlüsselbegriffen wie Anonymisierung oder Memorisierung, transparente Dokumentation der Datenverarbeitungsschritte, kontinuierliche Risikobewertungen und die Entwicklung interoperabler Standards auf europäischer Ebene. Ein solcher Rahmen fördert Innovation bei gleichzeitigem Setzen von klaren Leitplanken für verantwortungsvolle Datenverarbeitung und den Schutz der Rechte der Bürger*innen.

Ki-Modelle eröffnen vielfältige Chancen für Innovation, Wirtschaftswachstum und gesellschaftlichen Fortschritt. Gleichzeitig bieten sie die Möglichkeit, Datenschutzprinzipien von Beginn an wirksam zu integrieren und so Vertrauen in neue Technologien zu stärken.

Mit den folgenden Antworten möchten wir dazu beitragen, eine praxisgerechte und zukunftsorientierte Ausgestaltung der DSGVO im Kontext von LLMs zu unterstützen.

Antworten

1. **Nach Erwägungsgrund 26 Satz 3 DSGVO sollten bei der Prüfung, ob eine natürliche Person identifizierbar ist, alle Mittel berücksichtigt werden, die von dem Verantwortlichen oder einer anderen Person nach allgemeinem Ermessen wahrscheinlich genutzt werden, um die natürliche Person direkt oder indirekt zu identifizieren. Unter Berücksichtigung der in der EDSA Stellungnahme 28/2024 Rn. 35ff. gelisteten Vorgehen, unter welchen Umständen könnte ein LLM als anonym erachtet werden?**

   Die Frage, ob ein LLM als anonym gelten kann, muss im Lichte von Erwägungsgrund 26 DSGVO und unter Berücksichtigung der Stellungnahme 28/2024 des Europäischen Datenschutzausschusses (EDSA) fortlaufend überprüft werden. Eine pauschale Betrachtung greift dabei zu kurz. Vielmehr ist zu prüfen, ob unter den gegebenen Umständen eine identifizierbare natürliche Person mit vertretbarem Aufwand durch den Verantwortlichen oder Dritte ermittelt werden kann. Der EDSA selbst betont, dass der Kontext des Einsatzes eines Modells, insbesondere der

Unterschied zwischen öffentlich zugänglichen und rein intern betriebenen Modellen, die Bewertung der Anonymität maßgeblich beeinflusst.

In technischer Hinsicht handelt es sich bei LLMs nicht um Datenbanken, in denen personenbezogene Informationen als abrufbare Einträge gespeichert sind. Vielmehr erfassen diese Modelle statistische Wahrscheinlichkeitsmuster, indem sie Trainingsdaten in abstrakte Vektoren übersetzen. Diese Vektoren repräsentieren nicht einzelne Personen, sondern semantische Relationen und Sprachkontexte. Ausgaben eines LLMs sind damit das Ergebnis probabilistischer Vorhersagen und stellen neue, synthetisch erzeugte Inhalte dar. Sie sind keine exakte Reproduktion konkreter Trainingsdaten, sondern vielmehr mathematische Approximationen sprachlicher Regelmäßigkeiten. Dieses Verständnis wird sowohl vom Hamburgischen Beauftragten für Datenschutz als auch von Aufsichtsbehörden in anderen europäischen Ländern, etwa in Dänemark und Norwegen, geteilt.

Die Tatsache, dass ein LLM unter bestimmten Bedingungen Informationen über Personen, insbesondere öffentliche Personen, ausgeben kann, ist nicht gleichbedeutend mit dem Vorliegen gespeicherter personenbezogener Daten. Vielmehr lässt sich dies auf die hohe Prävalenz solcher Informationen im öffentlich zugänglichen Trainingsmaterial zurückführen.

Im Hinblick auf die rechtliche Beurteilung ist entscheidend, ob eine identifizierbare Person mit zumutbarem Aufwand ermittelt werden kann. Der EuGH hat hierzu klargestellt, dass die bloße theoretische Möglichkeit der Identifizierung nicht ausreicht. Vielmehr muss eine Identifikation mit "vernünftigerweise einzusetzenden Mitteln" erfolgen können, wobei Kosten, Zeitaufwand und verfügbare Technologien zu berücksichtigen sind. Wenn eine Identifizierung "praktisch unmöglich" erscheint, liegt kein Personenbezug vor. Diese Schwelle wird bei modernen LLMs grundsätzlich nicht erreicht, sofern diese auf unstrukturierten Daten ohne eindeutige Identifikatoren trainiert wurden, keine Feinabstimmung auf Einzelnutzer enthalten, und gleichzeitig technische und organisatorische Maßnahmen zur Risikominimierung eingesetzt werden. Solche Maßnahmen können unter anderem die Deduplizierung von Trainingsdaten, die Reduktion überrepräsentierter Textpassagen, Filtermechanismen gegen die Ausgabe sensibler Inhalte sowie Zugriffsbeschränkungen auf das Modell beinhalten (dazu mehr zur Antwort in Frage 2). Ein weiteres zentrales Argument für die Anerkennung

von Anonymität bei LLMs ist die Unverhältnismäßigkeit strukturierter Rückverfolgung. Bei unstrukturierten Trainingsdaten wäre die Herausarbeitung personenbezogener Datenpunkte nur durch eine umfangreiche Nachverarbeitung personenbezogener Daten und die Verknüpfung mit externen Identifikationsdaten denkbar, was die Auswirkungen auf die Privatsphäre des Einzelnen vervielfachen würde.

Die Frage der Anonymität eines LLMs kann nicht pauschal mit Verweis auf die theoretische Möglichkeit von Datenrekonstruktion beantwortet werden. Maßgeblich ist, ob das Modell unter Berücksichtigung seines Einsatzzwecks, seines Zugriffsrahmens, der Beschaffenheit seiner Trainingsdaten sowie der implementierten Schutzmaßnahmen ein realistisches Risiko der Re-Identifizierung birgt. Wenn dies nicht der Fall ist, ist von Anonymität im Sinne der DSGVO auszugehen. Eine kontext- und risikobasierte Betrachtung entspricht nicht nur dem Regelungsgehalt des Erwägungsgrundes 26, sondern auch dem übergreifenden Prinzip der Verhältnismäßigkeit, das der DSGVO zugrunde liegt.

2. **Welche technischen Maßnahmen setzen Sie bereits ein bzw. planen Sie einzusetzen, um die Memorisierung von Daten zu verhindern (wie z.B. Deduplikation, Verwendung anonymer bzw. anonymisierter Trainingsdaten, Fine-Tuning ohne personenbezogene Daten, Differential Privacy, etc.)? Welche Erfahrungen haben Sie damit gemacht?**

Zur Vermeidung der ungewollten Speicherung personenbezogener Daten setzen die die Anbieter (i.S.d. KI-VO) von LLMs auf eine Vielzahl technischer und organisatorischer Maßnahmen während des gesamten Lebenszyklus. Diese reichen von der Auswahl und Aufbereitung der Trainingsdaten über das Modelltraining bis hin zu Maßnahmen im Rahmen der Bereitstellung und Ausgabekontrolle.

Bereits in der Phase der Datenaufbereitung kommen systematisch deduplizierende und datensparsame Verfahren zum Einsatz. Die Deduplizierung erfolgt häufig mehrstufig, etwa auf URL-, Dokument- oder Zeilenebene. Dies soll verhindern, dass wiederholt auftretende Informationen, darunter potenziell personenbezogene Inhalte, überproportionale Gewichtung im Modell erhalten. Ergänzend erfolgt die Anonymisierung bzw. Entpersonalisierung der Daten, häufig durch Named-Entity-

Recognition(NER)-gestützte Maskierung. Dabei werden etwa Eigennamen oder Adressangaben automatisiert erkannt und durch Platzhalter oder synthetische Äquivalente ersetzt.

Darüber hinaus ist der Einsatz unstrukturierter Datenquellen von zentraler Bedeutung. Solche Daten liegen typischerweise in natürlicher Sprache und ohne systematische Identifikatoren vor. Ohne strukturierte Kennungen oder konsistente Eigenreferenzen ist es für ein LLM faktisch kaum möglich, Aussagen einzelnen Privatpersonen zuzuordnen oder personenbezogene Daten statistisch zu repräsentieren.

In der Modelltrainings- und Feinabstimmungsphase setzen Unternehmen häufig zusätzliche Verfahren wie Weight Regularization, Gradient Clipping oder Prompt Shielding ein, um eine Überanpassung an sensible Inhalte zu verhindern. In der Bereitstellungsphase ergänzen Ausgabefilter („Output Filters“) und Moderationsmechanismen das Schutzkonzept. Sie dienen dazu, risikobehaftete oder ungewollte Inhalte – etwa sensible personenbezogene Daten – systematisch aus den Modellantworten zu entfernen oder deren Ausgabe zu blockieren.

Differential Privacy kommt in der Praxis vor allem in hochregulierten Anwendungsbereichen, etwa im Gesundheits- oder Finanzsektor, zum Einsatz. Der produktive Einsatz bei großskaligen LLMs ist bislang jedoch aufgrund von Performance-Einbußen und technischer Komplexität begrenzt. Gleichwohl wird Differential Privacy als ergänzendes Instrument in bestimmten Szenarien weiterentwickelt.

Zur weiteren Risikominimierung implementieren Anbieter zunehmend automatisierte Bewertungs- und Metriksysteme, die das Memorisierungsrisiko modellübergreifend analysieren und quantifizieren. Red-Teaming-Ansätze sowie adversarielle Testreihen helfen zusätzlich dabei, mögliche Schwachstellen in der Modellverhalten frühzeitig zu identifizieren, etwa durch gezielte Versuchsprompts zur Extraktion potenziell sensibler Informationen.

Insgesamt zeigt sich in der Anwendungspraxis: Kein einzelnes Verfahren kann eine vollständige Vermeidung der Memorisierung garantieren. Entscheidend ist vielmehr ein mehrstufiger, kontextsensibler Ansatz, der auf die Art des Modells, die Struktur

Deutsches Lobbyregister R000257 **EU-Transparenzregister** 479540331468-69 **USt-IdNr.** DE196415580 **St-Nr.** 27/620/62930

der Daten und den konkreten Einsatzzweck abgestimmt ist. Durch die Kombination technischer Schutzmaßnahmen in allen Entwicklungs- und Betriebsphasen lassen sich Memorisierungsrisiken jedoch signifikant reduzieren, ohne dabei die Leistungsfähigkeit LLMs grundlegend zu beeinträchtigen.

3. **Wie schätzen Sie das Risiko ein, dass personenbezogene Daten aus einem LLM extrahiert werden? Erläutern Sie Ihre Einschätzung möglichst anhand konkreter Beispiele, Einzelfälle oder empirischer Beobachtungen.**

Der BVDW und seine Mitglieder schätzen das Risiko, dass personenbezogene Daten aus einem LLM extrahiert werden können, grundsätzlich als gering ein.

Wichtig ist dabei eine präzise begriffliche Trennung: LLMs speichern keine personenbezogenen Daten im klassischen Sinne, etwa in Form abrufbarer Datenbankeinträge. Stattdessen beruhen mögliche "Wiedergaben" personenbezogener Inhalte auf probabilistischen Rekonstruktionsprozessen innerhalb des Modells. Diese Prozesse basieren auf Wahrscheinlichkeitsverteilungen, die während des Trainings gelernt wurden, und nicht auf einem direkten Zugriff auf die ursprünglichen Trainingsdaten.

Das tatsächliche Extraktionsrisiko hängt dabei maßgeblich von drei Faktoren ab: der Art der Trainingsdaten, der technischen Modellarchitektur und den jeweiligen Schutzmaßnahmen. In der Praxis zeigt sich, dass personenbezogene Daten nur dann mit höherer Wahrscheinlichkeit rekonstruiert oder ausgegeben werden, wenn sie im Trainingsdatensatz in besonders hoher Frequenz oder überstrukturierter Form enthalten waren. Dies betrifft typischerweise öffentlich auffindbare Informationen zu Personen des öffentlichen Lebens, nicht jedoch private, nicht öffentlich bekannte Daten über Einzelpersonen. Daten über Privatpersonen, die im Vergleich zur Gesamtmenge des Trainingskorpus nur in geringer Häufigkeit vorkommen, führen in der Regel nicht zu einer statistisch signifikanten Repräsentation im Modell. Die Wahrscheinlichkeit ihrer Memorisierung und späteren Rekonstruktion ist daher äußerst gering.

Die Anbieter von LLMs begegnen den Extraktionsrisiken durch eine Vielzahl präventiver Maßnahmen: Dazu zählen unter anderem die Entfernung

personenbezogener Quellen aus den Trainingsdaten, die Verwendung unstrukturierter Datenformate ohne eindeutige Identifikatoren, deduplizierende Verfahren sowie Fine-Tuning-Strategien, die das Modell gezielt darauf trainieren, keine personenbezogenen Informationen auszugeben. (siehe Antwort auf Frage 2)

Darüber hinaus werden Ausgabefilter eingesetzt, die verhindern, dass sensible Inhalte ausgegeben werden, insbesondere dann, wenn betroffene Personen beispielsweise ihr Recht auf Vergessenwerden nach Art. 17 DSGVO ausgeübt haben. In Fine-Tuning-Phasen werden Modelle zudem oft nicht auf Rohdaten, sondern auf synthetisch generierten Zielantworten trainiert, wodurch die Gefahr einer memorisierten 1:1-Wiedergabe weiter minimiert wird. Auch werden Modelle nicht darauf trainiert, Nutzereingaben zu memorisieren.

Zur Bewertung verbleibender Risiken nutzen Unternehmen zunehmend sogenannte Memorisierungs-Tests oder gezielte Red-Teaming-Szenarien. Dabei werden insbesondere sensible Datenpunkte wie E-Mail-Adressen oder Telefonnummern gezielt abgefragt, um zu überprüfen, ob das Modell diese ausgeben kann. Solche Tests helfen dabei, etwaige Schwachstellen zu identifizieren und Maßnahmen gezielt nachzuschärfen. Probabilistische Methoden, etwa das Discoverable Extraction Framework, ermöglichen darüber hinaus eine differenzierte, quantifizierbare Bewertung des Extraktionsrisikos unter realitätsnahen Bedingungen: Dabei wird nicht nur binär geprüft, ob ein bestimmter Datensatz reproduzierbar ist, sondern mit welcher Wahrscheinlichkeit dieser über eine gegebene Anzahl an Anfragen extrahiert werden könnte. Solche Verfahren bieten eine deutlich praxisnähere Grundlage zur Risikobewertung im industriellen Einsatz.

Ein weiteres Risiko kann sich durch sogenannte Retrieval-Augmented-Generation (RAG)-Architekturen ergeben. Diese Systeme kombinieren LLMs mit externen Vektordatenbanken, um gezielt Informationen zu bestimmten Abfragen nachzuliefern. Wenn hierbei unzureichend klassifizierte oder falsch indizierte Dokumente eingebunden werden, kann es zu unbeabsichtigtem Leakage kommen, auch ohne Memorisierung im Modell selbst. Unternehmen begegnen diesem Risiko durch klar definierte Zugriffskontrollen, isolierte Datenräume und Auditierung im Rahmen des Modellbetriebs.

Zusammenfassend lässt sich sagen: Das Risiko der Extraktion personenbezogener Daten aus einem LLM kann als niedrig bewertet werden. Eine pauschale Gefährdungslage lässt sich aus empirischer Sicht nicht ableiten. Vielmehr erfordert eine valide Risikoeinschätzung stets eine einzelfallbezogene Betrachtung, unter Berücksichtigung des Trainingsdatenprofils, der eingesetzten Schutzmaßnahmen und des beabsichtigten Verwendungsrahmens. Modelle, die unter Beachtung deduplizierender, entpersonalisierender und kontrollierter Verfahren trainiert und betrieben werden, zeigen bislang keine systematische oder realweltlich dokumentierte Gefährdungslage im Sinne einer massenhaften oder gezielten Extraktion personenbezogener Daten über Privatpersonen.

4. **Datenschutzrecht knüpft an die Verarbeitung personenbezogener Daten an. Jede Eingabe eines Prompts löst eine Berechnung im KI-Modell aus, bei der die in Form von Parametern repräsentierten (personenbezogenen) Daten Einfluss auf das Berechnungsergebnis nehmen. Stellt diese Berechnung eine Verarbeitung dieser Daten im Sinne von Artikel 4 Nr. 2 DSGVO dar, selbst wenn das Berechnungsergebnis, also die Ausgabe des KI-Modells, nicht personenbezogen ist?**

Nach Auffassung des BVDW und seiner Mitglieder sollte die bloße Berechnung innerhalb eines trainierten KI-Modells, insbesondere eines LLMs, nicht pauschal als Verarbeitung personenbezogener Daten im Sinne von Art. 4 Nr. 2 DSGVO eingestuft werden. Eine solche Sichtweise würde nicht nur mit dem Systemverständnis der DSGVO kollidieren, sondern auch zu einer rechtlichen Überdehnung des Verarbeitungsbegriffs führen, die eine Vielzahl statistischer Verfahren grundlos unter Datenschutzrecht subsumieren würde.

Trainierte LLMs bestehen aus numerischen Vektoren und Gewichtungen, die abstrakte statistische Muster kodieren, nicht jedoch aus direkt oder mittelbar identifizierbaren Informationen über natürliche Personen. Die Parameter eines Modells sind nicht als personenbezogene Daten zu qualifizieren, solange sie nicht im konkreten Anwendungsfall eine Rückführbarkeit auf identifizierbare Personen ermöglichen. Dies entspricht auch der Klarstellung in Erwägungsgrund 26 DSGVO sowie der funktionalen Betrachtung der EDSA-Leitlinie 28/2024, wonach eine Identifizierbarkeit im Kontext geprüft werden muss.

Position

Die durch einen Prompt ausgelöste Inferenz eines LLMs basiert auf mathematischen Operationen über trainierte Wahrscheinlichkeitsverteilungen. Die Tatsache, dass ein LLM generalisierte Muster „anwendet“, um auf eine Nutzereingabe zu reagieren, bedeutet nicht, dass es damit personenbezogene Daten „verarbeitet“. Es greift nicht auf originalgetreue oder rekonstruktive Speicherinhalte zurück, sondern generiert neue Ausgaben auf Basis von Wahrscheinlichkeitsmatrizen.

Eine pauschale Anwendung des Verarbeitungsbegriffs auf jede mathematische Operation innerhalb eines KI-Modells hätte weitreichende Konsequenzen: Sie würde faktisch jedes statistische System, das auf großen Datensätzen trainiert wurde (z. B. Rechtschreibkorrekturen, Übersetzungstools oder Empfehlungs-Systeme), unter Datenschutzrecht stellen, selbst wenn die Modelle keinerlei Rückschlüsse auf Einzelpersonen zulassen. Dies würde zu erheblichen Rechtsunsicherheiten führen.

Gleichwohl kann in spezifischen Fällen, etwa bei Memorisierung hochfrequenter personenbezogener Informationen oder der Verwendung sensibler Trainingsdaten, eine datenschutzrechtliche Relevanz entstehen. Dies betrifft jedoch nicht die Inferenzberechnung an sich, sondern die potenzielle Wiedergabe personenbezogener Inhalte im Output oder deren gezielte Rekonstruktion. Nur in diesen Fällen ist eine Verarbeitung im Sinne der DSGVO anzunehmen. Die bloße Tatsache, dass Modellparameter ursprünglich aus personenbezogenen Daten abgeleitet wurden, begründet keine fortbestehende Verarbeitungsqualität bei der Inferenz.

Die Berechnung innerhalb eines KI-Modells stellt nicht per se eine Verarbeitung personenbezogener Daten i. S. v. Art. 4 Nr. 2 DSGVO dar, wenn das Ergebnis nicht personenbezogen ist und keine funktionale Rückführbarkeit auf Einzelpersonen besteht. Eine pauschale Gleichsetzung von Modellinferenz und Datenverarbeitung widerspricht sowohl der Systematik der DSGVO als auch dem technischen Verständnis moderner KI-Modelle. Stattdessen ist eine risikobasierte, kontextabhängige Betrachtung geboten, die auf das Ergebnis der Modellverarbeitung und die konkreten Nutzungsumstände abstellt.

5. Haben Sie bereits Erfahrung gemacht mit Methoden, die die Menge und Art der personenbezogenen memorisierten Daten abschätzen, bzw. ob das verwendete KI-Modell personenbezogene Daten einer bestimmten Person enthält (z.B. Privacy Attacks/PII Extraction Attacks, etc.)? Wenn ja, wie bewerten Sie deren Aussagekraft und mögliche Einschränkungen?

Die Unternehmen in der digitalen Wirtschaft verfügen über umfangreiche Erfahrungen mit Methoden zur Risikobewertung der Memorisierung personenbezogener Daten in KI-Modellen. Dabei kommen sowohl empirische Tests zur Modellausgabe als auch gezielte Angriffssimulationen (z. B. Regurgitation Attacks, Membership Inference, Model Inversion) zum Einsatz. Diese Methoden haben sich als wichtige Instrumente zur technischen Schwachstellenanalyse etabliert.

Erprobte Methoden zur Abschätzung von Memorization und personenbezogenen Leaks

a. Red Teaming & Privacy-specific Prompting
   Ein zentrales Verfahren ist das strukturierte Red Teaming, bei dem Teams adversarial agieren und versuchen, durch gezielte Prompts personenbezogene Inhalte aus dem Modell abzuleiten. Dabei werden unterschiedliche Szenarien getestet – etwa die Rekonstruktion typischer E-Mail-Signaturen, wiederkehrender Phrasen oder seltener Kombinationen von PII (z. B. Name + Ort + Arbeitgeber).

b. Empirical Memorization Testing
   Diese Tests evaluieren, ob ein Modell bei Eingabe bestimmter Prompts Inhalte aus dem Trainingsdatensatz unverhältnismäßig wahrscheinlich oder wörtlich reproduziert. Besonders im Fokus stehen sensible oder identifizierende Daten wie Telefonnummern, Namen mit Kontext oder individuelle Texte (z. B. Bewerbungen, Briefe, Chatverläufe). Die Tests werden meist nach dem Pretraining oder Finetuning durchgeführt.

c. Privacy Attack Simulations

Sie werden je nach Use-Case implementiert. Sie dienen der gezielten Analyse von Risikokonstellationen, insbesondere bei Modellen mit spezifischem Finetuning auf unternehmensspezifischen Daten.

d. Probabilistic Discoverable Extraction
   Ein neuartiger Ansatz erlaubt es, das Risiko extrahierbarer Inhalte zu quantifizieren, indem abgeschätzt wird, mit welcher Wahrscheinlichkeit ein Zieltext in einer bestimmten Zahl von Prompt-Versuchen reproduziert werden kann. Dies bietet eine skalenfähige, industriekompatible Metrik und adressiert die Limitierungen rein binärer Tests (Leak oder kein Leak).

Diese Methoden leisten wertvolle Dienste, insbesondere bei der Erkennung von überrepräsentierten oder duplizierten Trainingsdaten, beim Aufzeigen technischer Schwächen in Legacy-Daten und bei der Analyse feinabgestimmter Modelle. Auch wenn einzelne Tests aufgrund der probabilistischen Natur von LLMs nicht immer reproduzierbare Ergebnisse liefern oder bei großskaligen Modellen an technische Grenzen stoßen, ermöglichen sie dennoch wichtige Erkenntnisse über potenzielle Risikokonstellationen. Dabei zeigt sich, dass selbst im Fall erfolgreicher Rekonstruktionsversuche die Ergebnisse meist fragmentarisch, uneindeutig und ohne externes Kontextwissen nicht personenbezogen im Sinne der DSGVO sind. Die Kombination dieser Testverfahren mit präventiven Maßnahmen wie Deduplikation, Datenmaskierung und kontrollierten Finetuning-Strategien hat sich in der Praxis als effektiver Ansatz zur Reduktion verbleibender Memorisierungsrisiken bewährt.

Zudem bestehen keine einheitlichen Benchmarks oder Standardmetriken, um die Erfolgswahrscheinlichkeit solcher Angriffe systematisch zu bewerten oder regulatorisch zu validieren. Forschungsarbeiten zeigen auch, dass erfolgreiche *Membership Inference Attacks* nicht zwangsläufig bedeutet, dass personenbezogene Daten extrahierbar sind. Regulierungsbehörden sollten berücksichtigen, dass viele dieser Methoden in akademischen Kontexten entwickelt wurden und daher möglicherweise Einschränkungen aufweisen, die ihre Skalierung oder Anwendung in realen industriellen Kontexten erschweren, sowohl für KI-Entwickler als auch für böswillige Akteure.

Deutsches Lobbyregister R000257 **EU-Transparenzregister** 479540331468-69 **USt-IdNr.** DE196415580 **St-Nr.** 27/620/62930

Parallel zur Anwendung dieser Testverfahren setzt die digitale Wirtschaft auf eine Kombination aus:

- Technischen Präventivmaßnahmen (vgl. Frage 2: Deduplikation, Preprocessing, Output Filter),
- Transparenzinstrumenten (z. B. Modellkarten, Dokumentation von Finetuningdaten),
- Privacy Impact Assessments
- Policy-gesteuerten Prompt- und Output-Filtern, insbesondere bei öffentlich zugänglichen Schnittstellen oder in RAGs.

6. Wie hoch ist die Menge personenbezogener memorisierter Daten in Ihnen bekannten KI-Modellen (in Prozent sowie Gesamtmenge Trainingsdaten)?

Als BVDW warnen wir vor der Annahme, dass sich die Menge personenbezogener Daten in einem KI-Modell überhaupt sinnvoll quantifizieren lässt, weder absolut noch prozentual. Anders als bei klassischen Datenbanksystemen werden in modernen KI-Modellen keine identifizierbaren Einzelinformationen gespeichert. Vielmehr werden statistische Muster in numerischen Modellparametern kodiert. Eine Rückverfolgbarkeit einzelner personenbezogener Datensätze in diesen Parametern ist in der Praxis grundsätzlich nicht gegeben. Trainingsdaten großer Sprachmodelle umfassen häufig Milliarden bis Billionen Tokens, die zu großen Teilen aus öffentlich zugänglichen, unstrukturierten Quellen stammen.

Zudem zeigt die aktuelle Forschung, dass selbst bei gezielten Angriffsszenarien der tatsächliche Anteil memorisierter Ausgaben verschwindend gering ist. Im technischen Bericht zur Gemini 2.X-Modellfamilie beispielsweise wurde die Memorisierungs Rate auf rund 0,2 Prozent geschätzt und selbst diese Ausgaben enthielten laut zusätzlicher Analyse keine personenbezogenen Informationen, sondern vor allem generische Textbausteine wie Quellcode oder allgemein verfügbare Webinhalte.

Vor diesem Hintergrund halten wir eine pauschale Schätzung zum Anteil personenbezogener Daten im Modell oder in memorisiertem Output für weder sachgerecht noch sinnvoll. Die bloße Tatsache, dass ein Modell auf

personenbezogenen Daten trainiert wurde, lässt keine Rückschlüsse auf deren spätere Repräsentation im Modell zu. Stattdessen ist eine risikobasierte, kontextabhängige Bewertung geboten, bei der technische Schutzmaßnahmen, Modellarchitektur, Trainingsdatenqualität und Einsatzkontext gemeinsam betrachtet werden. Die Forderung nach prozentualen oder absoluten Mengenangaben ist demgegenüber irreführend und steht im Widerspruch zu datenschutzrechtlichen Grundprinzipien.

7. Wie gehen Sie vor, wenn eine Person ihren Anspruch auf Auskunft über personenbezogene Daten, Berichtigung oder Löschung ihrer personenbezogenen Daten im KI-Modell geltend macht?

Als BVDW erkennen wir die Bedeutung der Betroffenenrechte nach Kapitel III der DSGVO auch im Kontext von KI-Modellen ausdrücklich an. Gleichzeitig bestehen bei den KI-Modellen selbst im Hinblick auf deren praktische Umsetzung erhebliche technische und rechtliche Herausforderungen. Diese ergeben sich insbesondere aus der strukturellen Besonderheit, dass KI-Modelle keine personenbezogenen Daten in einer abfragbaren oder editierbaren Form enthalten, sondern Trainingsinhalte in Form von nicht direkt rückführbaren statistischen Parametern abstrahieren.

Aus Sicht der DSGVO ist entscheidend, dass sich die Verpflichtung zur Umsetzung von Auskunfts-, Berichtigungs- oder Löschungsrechten stets auf die Verarbeitung personenbezogener Daten bezieht. Die Trainingsdatenphase und die Eingabe- bzw. Ausgabeschicht eines generativen KI-Systems sind dabei datenschutzrechtlich relevant, nicht jedoch die bloße Existenz des Modells als mathematische Repräsentation bereits verarbeiteter Daten. Ein KI-Modell speichert keine konkreten Dateninstanzen, sondern codiert Wahrscheinlichkeitsverteilungen. Das Modell ist somit kein Container personenbezogener Daten im Sinne der DSGVO.

Für Auskunftsersuchen bedeutet dies, dass eine individuelle Identifikation personenbezogener Informationen im Modell in der Regel technisch nicht möglich und rechtlich nicht geschuldet ist. Artikel 11 Absatz 1 DSGVO stellt klar, dass eine nachträgliche Identifizierung einer betroffenen Person allein zum Zweck der Betroffenenrechte nicht erforderlich ist, sofern der Verantwortliche die betroffene

Person ohne zusätzliche Informationen nicht identifizieren kann. Eine darüberhinausgehende Verpflichtung zur Strukturierung und Zuordnung von Trainingsdaten würde im Gegenteil mit dem Grundsatz der Datenminimierung kollidieren.

Gleichwohl bestehen für betroffene Personen effektive Möglichkeiten, auf die Modellverwendung Einfluss zu nehmen. Sofern eine Person glaubhaft macht, dass ihre personenbezogenen Daten in Form eines bestimmten Outputs durch das Modell reproduziert werden, kann dieser spezifische Output durch technische Maßnahmen auf der Anwendungsebene unterdrückt werden. Damit kann einem Löschungs- oder Widerspruchsbegehren faktisch entsprochen werden, ohne dass dies eine nachträgliche Veränderung des Modells erforderlich macht. Vergleichbar mit der etablierten Rechtsprechung des EuGH zu Suchmaschinen ist auch hier eine ex-post-Kontrolle im Einzelfall der angemessene datenschutzrechtliche Ausgleich, da auf der Ausgabe-/Anwendungsebene sowohl das größere Schadenspotenzial als auch die Möglichkeit wirksamer Schutzmaßnahmen besteht.

Eine Berichtigung personenbezogener Daten im Modell selbst ist hingegen nach dem aktuellen Stand der Technik nicht möglich, da dies ein vollständiges Retraining unter Ausschluss der betroffenen Daten erfordern würde. Solche Verfahren wie „Machine Unlearning“ befinden sich derzeit noch in der experimentellen Entwicklung und sind für produktive Anwendungen nicht belastbar. In der Zwischenzeit ist es sachgerecht, die Qualität der Ausgangsdaten (etwa bei First-Party-Daten in strukturierten Trainingspipelines) zu sichern und neue Trainingsläufe nur auf bereinigten Datensätzen durchzuführen.

Insgesamt plädiert die digitale Wirtschaft für einen pragmatischen und verhältnismäßigen Umgang mit Betroffenenrechten im KI-Bereich, der sowohl den technischen Gegebenheiten als auch den Schutzzielen der DSGVO gerecht wird. So können datenschutzrechtliche Rechte effektiv durchgesetzt werden, ohne die Innovationsfähigkeit von KI-Systemen durch technisch oder wirtschaftlich unzumutbare Anforderungen zu behindern.

8. Gibt es andere Aspekte, die aus Ihrer Perspektive beim Schutz der personenbezogenen Daten in KI-Modellen eine Rolle spielen?

Aus Sicht des BVDW und seiner Mitglieder ist der Schutz personenbezogener Daten im Kontext KI-gestützter Systeme eine vielschichtige Herausforderung, die weit über Fragen der Trainingsdatensätze hinausgeht. Ein effektiver und innovationsfreundlicher Datenschutzrahmen muss auf einem ganzheitlichen Lebenszyklusansatz basieren und technische, organisatorische sowie rechtliche Perspektiven integrieren.

Derzeit bestehen keine einheitlichen Standards für „Privacy by Design“ in generativen KI-Systemen. Begriffe wie Anonymisierung, Memorisierung, Machine Unlearning oder RAG sind weder technisch einheitlich definiert noch regulatorisch geklärt. Dies erschwert die Risikobewertung, Auditierung und Rechtsdurchsetzung. Der BVDW und seine Mitglieder plädieren daher für die Entwicklung interoperabler Mindeststandards auf technischer und semantischer Ebene, idealerweise im Einklang mit EU-weit harmonisierten Rahmenwerken.

Eine effektive Regulierung sollte sich am Schutzziel orientieren – nicht an der Modellarchitektur. Wie auch bei anderen systemischen Komponenten (Server, Speicher, Netze) ist nicht deren bloße Existenz, sondern deren Nutzung entscheidend. Regulierungen, die generative Modelle abstrakt und modellzentriert adressieren, bergen die Gefahr ineffektiver Vorgaben und innovationshemmender Pflichten. Stattdessen sollte der Fokus auf konkreten Datenverarbeitungsprozessen und deren Auswirkungen liegen.

Die digitale Wirtschaft setzt sich für einen risikobasierten, technologieoffenen und praxistauglichen Datenschutzrahmen für KI-Modelle ein. Dabei müssen bestehende Datenschutzprinzipien konsequent auf moderne Systemarchitekturen übertragen werden, ohne Innovation und Rechtsschutz gegeneinander auszuspielen. Entscheidend ist nicht, ob ein KI-Modell personenbezogene Daten enthält, sondern ob der Einsatz der Technologie Risiken für Rechte und Freiheiten betroffener Personen mit sich bringt – und wie diese durch geeignete Schutzmaßnahmen wirksam adressiert werden können.