VERBAND DER INTERNETWIRTSCHAFT E.V.

STELLUNGNAHME

zur Konsultation zum datenschutzkonformen Umgang mit personenbezogenen Daten in KI-Modellen der Bundesbeauftragten für den Datenschutz und die Informationsfreiheit (BfDI)

Berlin, 29.08.2025

Zusammenfassung

Die Entwicklung und Nutzung von Künstlicher Intelligenz, insbesondere von Large Language Models (LLMs), ist für die Wettbewerbsfähigkeit der Digitalwirtschaft in Deutschland und Europa von zentraler Bedeutung. Zugleich gilt es, die Grundrechte der Bürgerinnen und Bürger zu achten und den Datenschutz nach der DSGVO wirksam umzusetzen. Eine Regulierung sollte dabei in erster Linie Rechtssicherheit **und** praktische Anwendbarkeit schaffen und zugleich sicherstellen, dass neue Technologien effizient entwickelt und eingesetzt werden können.

eco befürwortet Regelungen, die risikobasiert, kontextbezogen und innovationsfördernd sind. Eine vollständige Anonymisierung ist in der Praxis kaum zu erreichen, weshalb nach Auffassung der Internetwirtschaft die tatsächliche Wahrscheinlichkeit einer Re-Identifizierung im jeweiligen Kontext entscheidend ist. So ist etwa in kontrollierten internen Einsatzumgebungen ein niedrigerer Anonymitätsschwellenwert sachgerecht, da die Wahrscheinlichkeit einer Re-Identifizierung wesentlich reduziert ist. Ergänzende Maßnahmen wie Differential Privacy, Loss Masking, synthetische Daten, Output-Filter und strenge Zugriffskontrollen gewährleisten bereits ein hohes Datenschutzniveau und ihre Implementierung durch die Modell Anbieter muss ebenfalls mit in die Bewertung eines Modells einfließen. Da Modelle nach dem Training zudem in der Regel keinen Zugriff mehr auf Rohdaten haben und ihre Ausgaben keine personenbezogenen Informationen enthalten, ist ihre Verarbeitung im Ergebnis mit der Nutzung anonymisierter Daten vergleichbar. Zusätzlich setzen Unternehmen auf sichere Infrastrukturen, Verschlüsselung, Zugriffsbeschränkungen, umfassende Dokumentation und Nutzerkontrollen wie Opt-out-Optionen und Werkzeuge zum Recht auf Vergessenwerden.

Aus Perspektive der Internetwirtschaft ist es als positiv zu erachten, dass die Bundesbeauftragte für den Datenschutz und die Informationsfreiheit (BfDI) sich in der vorliegenden Konsultation mit dem Spannungsfeld zwischen innovativen Rahmenbedingungen für KI und dem Schutz personenbezogener Daten auseinandersetzt. Die Internetwirtschaft vertritt die Ansicht, dass an dieser Stelle zusätzliche Klarstellungen erforderlich sind, um Sicherheit für Anbieter von KI-Modellen in Deutschland und Europa zu gewährleisten. Die Verwirklichung eines solchen Anspruches kann nur durch die Implementierung eines pragmatischen, risikobasierten Ansatzes erfolgen, welcher den verantwortungsvollen Einsatz von KI in Deutschland und Europa stärkt.

Zu den Fragen der Konsultation

Frage 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?

Aus Sicht der Internetwirtschaft ist die derzeitige Fokussierung auf vollständige Anonymität von Sprachmodellen zu einseitig. Grundsätzlich müssen nach Ansicht von eco stattdessen verschiedene Faktoren beachtet werden, um das tatsächliche Risiko einer De-Anonymisierung mit zumutbaren Mitteln praxisnah erfassen zu können. So ist es etwa erforderlich den Kontext, in dem ein Modell genutzt wird, sowie den Kreis der Zugriffsberechtigten mit zu berücksichtigen. In streng kontrollierten, internen Umgebungen, in denen ein Modell ausschließlich autorisierten Nutzerinnen und Nutzern zur Verfügung steht und umfassende technische sowie organisatorische Maßnahmen gelten, ist dieses Risiko faktisch minimal. Daher sollte dort ein niedrigerer Anonymitätsschwellenwert angesetzt werden.

Zudem ist zu berücksichtigen, dass moderne Datenschutztechniken wie Differential Privacy, Loss Masking, die Nutzung synthetischer Daten, Output-Filter oder strenge Zugriffskontrollen das Risiko einer Re-Identifizierung erheblich verringern. Modelle von Anbietern, die nachweisen können, diese Techniken implementiert zu haben, sollten dementsprechend als anonym gelten, auch wenn ein theoretisches Restrisiko zur De-Anonymisierung besteht. Die Extraktion personenbezogener Daten aus den betreffenden Modellen ist insgesamt mit einer äußerst geringen Wahrscheinlichkeit verbunden. Darüber hinaus ist es von entscheidender Bedeutung, dass nach Abschluss des Trainings keine weiteren Zugriffe auf die ursprünglichen Rohdaten durch LLMs erfolgen. Es erfolgt ausschließlich der Verwendung statistischer Parameter, die allgemeine Muster repräsentieren, jedoch keine identifizierbaren Datensätze enthalten. Unter der Voraussetzung, dass die Ausgaben des Modells keine personenbezogenen Informationen enthalten und das Modell nicht gezielt auf die Rekonstruktion individueller Inhalte ausgelegt ist,, ist die Verarbeitung im Ergebnis der Nutzung anonymisierter Daten gleichzusetzen.

Nach Ansicht von eco wäre eine pragmatische, risikobasierte Auslegung der Anonymität im Einklang mit Erwägungsgrund 26 der DSGVO wünschenswert und sachgemäß. Dieser Ansatz unterstützt verantwortungsvolle Innovationen und stellt sicher, dass der Datenschutz sinnvoll und den tatsächlichen Risiken angemessen ist.

Frage 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?

Zum Schutz vor ungewollter Wiedergabe personenbezogener Daten setzen viele Unternehmen der Internetwirtschaft auf ein durchgängiges Datenschutzkonzept über den gesamten KI-Lebenszyklus hinweg. Bereits vor dem Training werden Daten reduziert und anonymisiert, riskante Quellen vermieden und, wo sinnvoll und möglich, synthetische Daten eingesetzt. Die Modellarchitektur wird so gestaltet, dass sie Muster erkennt, aber keine einzelnen Datenpunkte speichert. Zudem verhindern nachgelagerte Schutzmechanismen wie Filter und Prompt-Shields zusätzlich, dass sensible Inhalte ausgegeben werden. Die Maßnahmen werden in der Regel kontinuierlich, etwa durch Red-Teaming und automatisierte Risikomessung, auf ihre Wirksamkeit überprüft.

Frage 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.

Die Bewertung des Risikos, dass ein Sprachmodell personenbezogene Daten preisgibt, erfolgt meistens in einem mehrstufigen Prozess, der alle Phasen der Modellentwicklung umfasst. Bereits vor dem Training werden bekannte oder leicht identifizierbare Quellen personenbezogener Daten, etwa Personen-Suchmaschinen, Kontaktinformationen oder Foren mit Klarnamen, aus den Datensätzen entfernt. So wird das Risiko sensibler Inhalte an der Quelle reduziert. Während des Trainings kommen Techniken wie „Loss Masking“ zum Einsatz, bei denen sensible Teile von Texten gezielt ausgeblendet oder durch neutrale Platzhalter beziehungsweise synthetische Inhalte ersetzt werden. Dadurch lernt das Modell zwar Strukturen und Muster, nicht aber den konkreten sensiblen Inhalt. Nach Abschluss des Trainings haben Modelle keinen Zugriff mehr auf die ursprünglichen Rohdaten, sondern arbeiten ausschließlich mit statistischen Parametern. Das macht eine direkte Rekonstruktion von Originaltexten praktisch unmöglich. Dies stellt sicher, dass das Modell Eingaben nicht wortwörtlich wiedergibt. Im laufenden Betrieb verhindern zusätzliche Filtersysteme, dass sensible Informationen ausgegeben werden, etwa über Personen, die ihr Recht auf Löschung nach der DSGVO geltend gemacht haben.

Frage 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?

Ein trainiertes Modell verarbeitet bei der Beantwortung neuer Eingaben keine personenbezogenen Daten im Sinne von Artikel 4 Absatz 2 DSGVO, solange es nicht gezielt zur Rekonstruktion individueller Informationen eingesetzt wird. Die internen Parameter sind abstrakte statistische Repräsentationen und keine identifizierbaren Datensätze. Solange die Ausgaben keine personenbezogenen Daten enthalten und das Modell nicht dafür konzipiert ist, solche zu generieren, entspricht die Berechnung funktional der Verarbeitung anonymisierter Daten. Dieser Umstand wird auch in der Stellungnahme 28/2024 des EDSA betont. Grundsätzlich geht der EDSA davon aus, dass personenbezogene Informationen in die Parameter eines Modells einfließen können. Gleichzeitig wird in der Stellungnahme jedoch klargestellt, dass KI-Modelle in der Regel nicht dafür konzipiert sind, solche Daten aus den Trainingsbeständen bereitzustellen. Zudem enthalten ihre Parameter keine Datensätze, die unmittelbar isoliert oder eindeutig einer Person zugeordnet werden könnten.

Frage 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?

Modellanbieter nutzten verschiedene Test- und Schutzmethoden, um das Risiko der Memorisation zu erkennen und zu minimieren. Dazu gehören intensive Red-Teaming-Tests, die gezielt versuchen, sensible Trainingsinhalte zu provozieren, sowie der Einsatz von Loss Masking während der Feinabstimmung, um zu verhindern, dass sensible Eingaben erlernt oder wiedergegeben werden. Modelle werden außerdem so konstruiert, dass sie keinen Zugriff auf die ursprünglichen Trainingsdaten während der Nutzung haben. Diese Verfahren haben sich als wirksam erwiesen, sodass das verbleibende Restrisiko als gering einzustufen ist und sich überwiegend im theoretischen Bereich bewegt.

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

Nach dem Training enthalten Modelle in der Regel keine direkten Kopien der Trainingsdaten, sondern lediglich angepasste Parameter, die allgemeine Muster repräsentieren. Eine exakte Quantifizierung, wie viel personenbezogenes Material in dieser Form eventuell noch enthalten ist, ist technisch nicht möglich und auch nicht sinnvoll. Aus Sicht von eco sollte der Fokus auf der Prävention und der nachweislichen Umsetzung von wirkungsvollen Maßnahmen zur Risikominimierung liegen, nicht auf schwer zu quantifizierenden Mengenangaben.

Frage 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?

In der Regel werden die für das Training verwendeten Daten bei der Entwicklung von KI-Modellen in Modellgewichte und -parameter umgewandelt, so dass es technisch nicht möglich ist, bestimmte persönliche Daten nach dem Training zu isolieren oder zu entfernen. Wenn Betroffene ihre Rechte nach den Artikeln 15 bis 17 DSGVO geltend machen, wird versucht, die Anfragen im Rahmen des technisch Machbaren umzusetzen. Der Schwerpunkt liegt dabei technisch bedingt auf Abhilfemaßnahmen nach dem Training, wie z. B. Red Teaming und Filter. Grundsätzlich ist es für Anbieter von KI-Modellen oft nicht möglich, Daten aus dem Modell zu entfernen. Lediglich der Ausschluss der Daten auf künftigen Trainings und die Blockierung entsprechender Ausgaben, wodurch sich die Chance der Reproduktion in den Ausgaben deutlich reduziert, ist technisch möglich und umsetzbar.

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

Neben den beschriebenen Maßnahmen spielen hohe Sicherheitsstandards wie Verschlüsselung, Zugriffsbeschränkungen und sichere Infrastrukturen eine zentrale Rolle, um unbefugten Zugriff oder Datenabfluss zu verhindern. Ebenso wichtig ist die Dokumentation der Datenquellen, der Verarbeitungsschritte und der Datenschutz-Folgenabschätzungen. Ergänzend haben Nutzerinnen und Nutzer die Möglichkeit, einer Nutzung ihrer Daten zu widersprechen, Löschungen zu beantragen und unangemessene Inhalte zu melden, um den Datenschutz in KI-Systemen ganzheitlich zu sichern.
