Konsultation zum datenschutzkonformen Umgang mit personenbezogenen Daten in KI-Modellen: Stellungnahme von D64 – Zentrum für Digitalen Fortschritt e.V.

Wird das Training der KI mit anonymen Daten durchgeführt, ist die DSGVO auf das Training nicht anwendbar. Allerdings ist bei KI-Modellen eine vollständige Anonymisierung angesichts der zum Training verwendeten Datenmengen in der Regel nicht zuverlässig möglich.

Zu dieser Eingangsthese ist kritisch anzumerken: Die strikte Anonymisierung könnte Large Language Models (LLMs) ihrer zentralen Funktionalität berauben – der Fähigkeit, Wissen über die reale Welt abzubilden und zu verknüpfen. Personenbezogene Daten durchziehen nahezu alle relevanten Trainingstexte: Wissenschaftliche Publikationen nennen ihre Autor:innen, Gerichtsurteile referenzieren Parteien und Richter, Nachrichtenartikel berichten über konkrete Personen und deren Handlungen, historische Texte dokumentieren biografische Zusammenhänge. Diese personenbezogenen Verknüpfungen bilden das Gerüst, an dem sich Faktenwissen, Kausalzusammenhänge und zeitliche Abfolgen orientieren. Ein KI-Modell, das etwa nicht weiß, dass Angela Merkel Bundeskanzlerin war, dass Elon Musk Tesla leitet oder dass bestimmte Rechtsprechung von konkreten Gerichten stammt, verliert seine Orientierung in der Wissenslandschaft. Die verbleibenden anonymisierten Fragmente – abstrakte Konzepte ohne Akteure, Ereignisse ohne Beteiligte, Entwicklungen ohne Urheber – ergäben ein teilweise funktionsunfähiges System, das weder sinnvolle Auskünfte geben noch Zusammenhänge herstellen könnte. Wir fokussieren uns in der Beantwortung der Konsultation daher nicht auf die Frage, ob LLMs der DSGVO unterstehen– was sie offensichtlich tun. Sondern wie mit den inhärenten Spannungen zwischen den Verarbeitungen von personenbezogenen Daten im Kontext der Schaffung und dem Einsatz von LLMs und den Grundsätzen der DSGVO umzugehen ist.

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 EDSA-Stellungnahme 28/2024 definiert Mindeststandards, die ausnahmslos einzuhalten sind, wenn ein LLM als anonym gelten soll. Die gegenwärtige Zurückhaltung der Aufsichtsbehörden mit konkreteren Vorgaben mag der technischen Dynamik geschuldet sein – sie gefährdet jedoch fundamentale Datenschutzprinzipien. Die Datenschutzaufsicht muss zeitnahe handeln und konkrete, durchsetzbare Vorgaben schaffen.

Wir vertreten die Kernthese, dass der in der Praxis übliche massenhafte Einsatz personenbezogener Daten beim Training es als nahezu ausgeschlossen erscheinen lässt, dass Large Language Models mit unbegrenztem Nutzerkreis als anonym gelten.

Erwägungsgrund 26 DSGVO verlangt die Berücksichtigung aller wahrscheinlich genutzten Mittel zur Re-Identifizierung. Diese Vorgabe muss bei LLMs ernst genommen werden:

1. Erweiterter Drittbegriff

Der Begriff „Dritter" umfasst bei solchen LLMs jeden potenziellen Systemzugang durch Nutzende – von Hobbynutzenden bis zu spezialisierten Sicherheitsforschenden, die das Modell auf Schwachstellen überprüfen. Fortgeschrittene Angriffsvektoren (Membership Inference, Model Inversion, Attribute Inference) definieren den Bewertungsmaßstab. Die technische Raffinesse des Angriffs ist irrelevant; entscheidend ist allein seine Durchführbarkeit. Anbieter tragen die volle Beweislast, dass ihr Modell selbst gegen solche Attacken resistent ist.

2. Verpflichtende Bug-Bounty-Programme – Transparente Sicherheitsprüfung

Anbieter von KI-Modellen mit systemischem Risiko und marktbeherrschende Akteure sollten robuste Bug-Bounty-Programme etablieren. Diese Programme erfüllen drei Funktionen:

- **Kontinuierliche Sicherheitsvalidierung** durch die Community
- **Nachweisführung** gemäß Art. 5 Abs. 2 DSGVO durch dokumentierte Prüfprozesse
- **Präventive Schwachstellenerkennung** vor irreversiblen Schäden

Solche Programme erfordern messbare Erfolgskriterien: Welche Datenabrufe gelten als relevant, da sie ein Compliance Problem darstellen? Ab welcher Konfidenz ist eine Re-Identifizierung von Nutzenden wahrscheinlich gegeben? Diese Parameter müssen vorab definiert und veröffentlicht werden.

3. Open-Source/Open-Weight-Modelle – Irreversible Risikoexposition

Die Veröffentlichung von Modellgewichten ist datenschutzrechtlich problematisch: Einmal freigegebene Modelle entziehen sich dauerhaft jeder Kontrolle. Nachträgliche Löschbegehren, Berichtigungen oder Einschränkungen – die auch bei proprietären Modellen derzeit kaum umsetzbar erscheinen – werden faktisch unmöglich.

4. Sofortmaßnahmen für Rechtssicherheit

KI-Anbieter sollten verschärften Vorab-Prüfpflichten unterliegen:

- Nachweis der Differential Privacy mit garantierten Privacy-Budgets,
- Zertifizierte Resistenz gegen alle bekannten Extraktionsangriffe,
- Öffentliches Red-Teaming vor Release.

Die Aufsicht sollte schnellstmöglich:

- Technische Mindeststandards für Anonymisierung in LLMs definieren,
- Standardisierte Testverfahren für Re-Identifizierungsrisiken etablieren.

Eine Zusammenarbeit europäischer Datenschutzbehörden im Kontext des EDPB zur Erarbeitung und Ansätzen ist wünschenswert. Der Schutzbedarf natürlicher Personen duldet keine experimentellen Grauzonen. Da die Nicht-Identifizierbarkeit von den

Anbietern nicht nachvollziehbar belegt werden kann, müssen ihre Modelle als Verarbeitung personenbezogener Daten behandelt werden – mit allen rechtlichen Konsequenzen.

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?

Als zivilgesellschaftlicher Verein adressiert diese Frage primär nicht uns. Unsere Beobachtungen zeigen jedoch: Betreiber und Entwickler von LLMs vernachlässigen die genannten technischen Schutzmaßnahmen systematisch.

Konkrete Belege dieser Vernachlässigung:

Das Meta-Verfahren zum Training von KI-Modellen mit Nutzerdaten (OLG Köln, Urteil vom 23.05.2025 - 15 UKI 2/25) bestätigt diese Einschätzung eindrücklich. Meta argumentierte insbesondere mit einer De-Identifizierung ausgewählter personenbezogener Daten als Schutzmaßnahme, um die Verarbeitung teils sensibler Daten auf Art. 6 Abs. 1 lit. f DSGVO stützen zu können. Offenkundig nehmen Anbieter wie Meta die dokumentierten Risiken von Membership Inference und Data Extraction Attacks nicht hinreichend ernst.

Es ist auch offenkundig, dass Large Language Models während des Trainings zwangsläufig Teile ihrer Trainingsdaten memorisieren – dies ist kein Fehler, sondern fundamentaler Bestandteil ihrer Funktionsweise, da sie durch die Optimierung von Milliarden von Parametern statistische Muster und konkrete Informationssequenzen in ihren Gewichtungen kodieren. Die Transformer-Architektur balanciert dabei zwischen Generalisierung (abstrakten Mustern) und direkter Memorisierung (konkreten Fakten), wobei besonders häufige oder einzigartige Datenkombinationen sich dauerhaft in die Modellparameter einprägen. Diese inhärente Memorisierung macht eine vollständige Trennung zwischen „gelerntem Wissen“ und „gespeicherten personenbezogenen Daten“ technisch unmöglich, da beides untrennbar in denselben Parametern verschmolzen ist.

Die aufgeführten technischen Maßnahmen greifen bei Large Language Models strukturell ins Leere: Deduplikation reduziert lediglich Mehrfachvorkommen, verhindert aber nicht die Memorisierung einmaliger Daten; anonymisierte Trainingsdaten würden die Funktionsfähigkeit des Modells beeinträchtigen; Fine-Tuning ohne personenbezogene Daten ändert nichts an der bereits im Basismodell eingebrannten Memorisierung; und Differential Privacy degradiert nachweislich die Modellperformance so erheblich, dass kommerzielle Anbieter darauf verzichten. Diese Maßnahmen sind keine wirksamen Lösungen, sondern bestenfalls marginale Abschwächungen eines systemimmanenten Problems.

Auch gegenüber dem Einsatz von Filtern bei dem Output der LLMs sind wir ausgesprochen skeptisch. Output-Filter adressieren lediglich die Symptome, nicht die Ursache der Memorisierung: Die personenbezogenen Daten bleiben vollständig in den Modellparametern kodiert und können durch geschicktes Prompting, Jailbreaking-Techniken oder indirekte Abfragen extrahiert werden – der Filter blockiert nur offensichtliche Ausgaben, nicht das zugrundeliegende Wissen. Zudem operieren Filter regelbasiert oder über Blacklists, während kreative Umformulierungen, Umschreibungen oder kontextuelle Einbettungen die gespeicherten Informationen dennoch preisgeben können. Das fundamentale Problem bleibt bestehen: Ein Modell, das personenbezogene Daten memorisiert hat, trägt diese permanent in sich – Output-Filter sind lediglich ein löchriger Zaun um einen bereits kontaminierten Datenspeicher.

Die Behauptung, durch die genannten Techniken ließe sich datenschutzkonformes Training realisieren, ignoriert die fundamentale Architektur von Transformer-Modellen: Sie sind Memorisierungsmaschinen, deren Leistungsfähigkeit gerade auf der Speicherung und Rekombination konkreter Trainingsdaten beruht.

Aktive Umgehung bestehender Standards:

Unsere Skepsis gegenüber Maßnahmen zum Schutz vor Memorisierung vorausgeschickt beobachten wir obendrein mit Sorge, dass LLM-Entwickler gezielt etablierte Mechanismen des Selbstschutzes missachten. Perplexity setzt beispielsweise getarnte, nicht deklarierte Crawler ein, um robots.txt-Direktiven zu umgehen (dokumentiert durch Cloudflare: https://blog.cloudflare.com/perplexity-is-using-stealth-undeclared-crawlers-to-evade-website-no-crawl-directives/.

Zentrale Herausforderung:

Im Konfliktfeld zwischen Content-Anbietern und Hosting-/Netzwerk-Dienstleistern bleibt die entscheidende Frage: Welche technischen und regulatorischen Instrumente schützen personenbezogene Daten effektiv vor unerlaubter Verwendung in Trainingsdatensätzen? Die bisherigen Marktmechanismen versagen offensichtlich.

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.

Wir sehen die Wahrscheinlichkeit als sehr hoch an. Dies zeigt bereits die Eingabe der Frage „Wer ist Louisa Specht-Riemenschneider" in marktübliche Angebote wie Claude Opus 4.1, die ein Ergebnis zu dieser Person bereitstellen und somit Daten zu ihr memorisieren. Dies gilt auch, wenn die Recherchefunktion – also das Nachladen von Informationen aus dem Internet – ausgeschaltet ist.

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?

Zur rechtlichen Einordnung der Inferenzphase

Die Ausgangslage ist eindeutig: Wurden beim Training des KI-Modells personenbezogene Daten verwendet und sind diese in den Modellparametern repräsentiert, liegt zunächst eine Speicherung personenbezogener Daten vor. Der entscheidende Punkt betrifft jedoch die

Inferenzphase – also den Moment, in dem das trainierte Modell zur Generierung von Ausgaben verwendet wird.

Nach Art. 4 Nr. 2 DSGVO erfasst der Verarbeitungsbegriff jeden mit oder ohne Hilfe automatisierter Verfahren ausgeführten Vorgang im Zusammenhang mit personenbezogenen Daten, einschließlich des „Verwendens" und „Abfragens". Die mathematischen Operationen während der Inferenz, bei denen die in den Parametern encodierten Informationen das Berechnungsergebnis beeinflussen, erfüllen diese Merkmale. Der Umstand, dass der Output selbst keine erkennbaren personenbezogenen Daten enthält, ändert nichts an der Verarbeitungsqualität des zugrundeliegenden Rechenvorgangs.

Diese Auslegung entspricht dem Schutzzweck der DSGVO: Die gesetzgebenden Institutionen wollte gerade auch intransparente und für Betroffene nicht nachvollziehbare Datennutzungen erfassen. Die technische Komplexität der Verarbeitung von Daten darf nicht zu einer Schutzlücke führen.

Es gilt, diese Problemlage anzuerkennen und nach angemessenen Lösungen zu suchen, statt die Anwendbarkeit der DSGVO – und damit zugleich die Grundrechtssphären der Betroffenen – in Frage zu stellen.

Das Transparenzdilemma und seine praktischen Konsequenzen

Die aufgeworfene Problematik der Intransparenz trifft den Kern der datenschutzrechtlichen Herausforderungen bei KI-Systemen. Nutzende eines vortrainierten Modells können tatsächlich nicht erkennen, ob und welche personenbezogenen Daten in die Modellparameter eingeflossen sind. Diese strukturelle Informationsasymmetrie führt zu einem faktischen Compliance-Problem: Ohne Kenntnis der Datengrundlage des Modells kann der Nutzer weder die erforderlichen Rechtsgrundlagen prüfen noch seinen Informationspflichten nach Art. 13 und 14 DSGVO nachkommen.

Die derzeitige Rechtslage führt zu einer unbefriedigenden Situation, in der technische Innovation und Datenschutz-Compliance in einem Spannungsverhältnis stehen.

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

Als zivilgesellschaftlicher Verein sehen wir uns nicht als primären Adressaten der Frage. Wir sehen aber mit Sorge, dass auch die Modellentwickler nicht hinreichend transparent mit entsprechenden Fragen umgehen.

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 zivilgesellschaftlicher Verein sehen wir uns nicht als primären Adressaten der Frage. Wir teilen aber gerne unsere Positionen zu den regulatorischen Anforderungen:

Auskunftsrecht: Modellbetreiber müssen präzise nachweisen, an welchen Stellen personenbezogene Daten der anfragenden Person in den Trainingsdaten auftreten. Sie müssen Betroffenen ebenfalls eine vollständige Kopie der identifizierten Datensegmente zur Verfügung stellen. Für Betroffene muss erkennbar sein, aus welchem Kontext die Daten stammen und wie häufig sie im Rahmen des Modelltrainings verwendet wurden.

Löschungsrechte: Modellentwickler müssen betroffenen Datensätze aus den Trainingsdaten entfernen und bei künftigen Trainingszyklen durch technische Sperrmechanismen vom Lernprozess ausschließen. Auch ein de-learning ist in der DSGVO verbürgt. Parallel implementieren Betreiber ein Unterlassungsrecht: Betroffene können verlangen, dass das Sprachmodell keinerlei Informationen über ihre Person ausgibt – gleichgültig, ob diese Informationen zutreffend oder fehlerhaft sind.

Berichtigungsrecht: Anstelle komplexer Korrekturverfahren, die erhebliche technische Herausforderungen bergen und Manipulationsrisiken schaffen, müssen Betreiber ersatzweise die vollständige Löschung der betroffenen Daten anbieten. Diese pragmatische Lösung vermeidet die Gefahr inkonsistenter Teilkorrekturen und gewährleistet einen sauberen Datenbestand.

Technische Umsetzung: Zur praktischen Durchsetzung dieser Rechte sollten Betreiber auf mehrschichtige Kontrollmechanismen einsetzen müssen: Reinforcement Learning trainiert das Modell darauf, bestimmte Informationen systematisch zu unterdrücken. Stichwortbasierte Filter blockieren zusätzlich die Ausgabe sensibler Daten. Obwohl absolute Compliance technisch unerreichbar bleiben dürfte, sollten sich Betreiber zur bestmöglichen Umsetzung der Betroffenenrechte durch kontinuierliche Weiterentwicklung ihrer Schutzmechanismen verpflichten.

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

Systematische Qualitätskontrolle bei faktenbezogenen Ausgaben: KI-Modelle produzieren bei der Informationsrecherche und -zusammenfassung regelmäßig faktisch falsche Aussagen über reale Personen. Da jede Ausgabe durch probabilistische Berechnungen aus den Trainingsdaten generiert wird, entstehen zwangsläufig Fehlerquoten. Der bloße Hinweis der Anbieter auf mögliche Ungenauigkeiten genügt angesichts der tatsächlichen Nutzungspraxis und Produktvermarktung nicht den datenschutzrechtlichen Anforderungen.

Verbindliche Standards und Benchmarks: Die Branche benötigt standardisierte Testverfahren mit klar definierten Grenzwerten für akzeptable Fehlerquoten bei personenbezogenen Informationen. Diese Benchmarks müssen:

- Einheitliche Messverfahren für die Faktentreue etablieren,
- Verbindliche Maximalwerte für Falschaussagen über natürliche Personen festlegen,
- Regelmäßige Überprüfungen und Zertifizierungen vorsehen,
- Korrekturmechanismen für systematisch auftretende Falschinformationen implementieren,
- Beschwerdemechanismen für betroffene Nutzenden.

Konkrete Gefährdungslagen: Die Tragweite fehlerhafter KI-Ausgaben zeigt sich in dokumentierten Fällen schwerwiegender Falschbehauptungen. ChatGPT unterstellte

beispielsweise einem unbescholtenen Bürger einen Doppelmord an Kindern – eine Falschaussage mit verheerenden Folgen für die betroffene Person (Der Standard, 2024: https://www.derstandard.at/consent/tcf/story/3000000262017/chatgpt-dichtete-einem-mann-einen-doppelten-kindsmord-an). Solche Vorfälle verdeutlichen die Dringlichkeit rechtlicher Schutzmechanismen.

Technische Schutzmaßnahmen: Ergänzend zu rechtlichen Vorgaben müssen technische Sicherheitsvorkehrungen implementiert werden:

- Konfidenzwerte für personenbezogene Aussagen mit automatischer Unterdrückung bei Unsicherheit,
- Verifizierungssysteme für Faktenchecks bei Personennennungen,
- Audit-Logs zur Nachvollziehbarkeit der Datenquellen,
- Sperrverzeichnisse für besonders schutzbedürftige Personen.