Friedrich-Alexander-Universität
Fachbereich Rechtswissenschaft

Institut für Recht und Technik (IRuT)

Juniorprofur für Bürgerliches Recht, Recht der Digitalisierung, des Datenschutzes und der künstlichen Intelligenz

Prof. Dr. Paulina Jo Pesch

Hindenburgstr. 34, 91054 Erlangen
paulina.pesch@fau.de
https://www.digit.rw.fau.de/

Stellungnahme zur Konsultation der BfDI zum datenschutzkonformen Umgang mit personenbezogenen Daten in KI-Modellen

Hinweis: Die Stellungnahme beschränkt sich auf Large Language Models (LLMs).

I. Grundlege Einschätzung

Im Bereich von LLMs sind **einerseits vortrainierte, andererseits feinabgestimmte Modelle** zu unterscheiden. Vortrainierte LLMs lassen sich für eine Vielzahl von Einsatzzwecken einsetzen und sind daher als KI-Modelle mit allgemeinem Verwendungszweck i. S. v. Art. 3 Nr. 63 (*general purpose AI models*) der KI-Verordnung einzuordnen. Sie werden auf diversen, i. d. R. sehr großen Trainingsdatensets trainiert, die eine Vielzahl von Domänen umfassen. Vortrainierte LLM bilden als Basismodelle (*foundation models*) die Grundlage für domänenspezifische Modelle, die durch das Training mit domänenspezifischen Daten für bestimmte Aufgaben oder Aufgabenbereiche feinabgestimmt, also einem Fine-Tuning unterzogen werden.

Existierende vortrainierte LLMs sind sämtlich auch **mit personenbezogenen Daten trainiert** worden, z. B. enthalten die Trainingsdaten Informationen über Autorinnen von Fachartikeln oder Büchern, in Wikipedia-Artikeln in Bezug genommene Personen oder Nutzerinnen von Foren. Zumindest Llama (Meta) und Grok (xAI) werden auch mit Social-Media-Daten trainiert, die auf Inhaberinnen von Accounts natürlicher Personen, darunter auch Minderjährige, und Dritte beziehbar sind. Bei domänenspezifischen Modellen hängen der Personenbezug und die Sensibilität personenbezogener Trainingsdaten vom Anwendungsfall ab.

Bei allen relevanten vortrainierten LLMs ist es Wissenschaftlerinnen gelungen, Trainingsdaten aus den Modellen zu extrahieren, darunter auch personenbezogene Daten (*privacy leakage*).[1] Im Vergleich zu anderen KI-Modellen im Bereich des maschinellen Lernens erlauben generative KI-Modelle die **Extrahierung von Trainingsdaten** in weitaus größerem Umfang. Sie sind konzeptionell auf die

[1] *Carlini et al.*, 30th USENIX Security Symposium 2021, S. 2633, S. 2637 ff.; *Carlini et al.*, ICLR 2023, S. 5; *Nasr et al.*, arXiv abs/2311.17035, 2023, S. 3 ff, S. 8 ff.; *Sivashanmugam*, arXiv:2507.04478; *Zharmagambetov et al.*, arXiv:2503.09780.

Ausgabe von Teilen von Trainingsdaten angelegt, auch wenn die wörtliche Reproduktion von Trainingstexten nur begrenzt erwünscht ist (z. B. für gängige Begriffe). Das Phänomen der „Memorisierung" von Trainingsdaten geht über das erwünschte Maß jedoch weit hinaus. Es ist noch wenig erforscht und es fehlt bislang weitestgehend an generalisierbaren Erkenntnissen dazu, wie es zur „Memorisierung" kommt und inwieweit und durch welche Maßnahmen sie sich wirksam vermeiden lässt. Obwohl sich Trainingsdaten bislang nicht effizient systematisch aus LLMs extrahieren lassen, werden im Bereich der Trainingsdatenextrahierung stetig Fortschritte erzielt. Wie viele personenbezogene Daten sich aus existierenden LLMs extrahieren lassen, ist nicht absehbar. U. a. in juristischen Publikationen zu findende[2] generalisierende Behauptungen, die Extrahierung personenbezogener Daten aus LLMs sei besonders unwahrscheinlich oder schwierig, sind schon nach jetzigem Forschungsstand falsch und lassen künftige Extrahierungsmethoden außer Betracht. Da die Identifizierung zumindest einiger Betroffener für vortrainierte LLMs nach allgemeinem Ermessen wahrscheinlich ist, sind die Modelle als personenbezogen einzuordnen.[3]

Die datenschutzrechtliche Debatte konzentriert sich bislang zu einseitig auf das Phänomen der Trainingsdatenextrahierung. Dringend verstärkt in den Blick zu nehmen sind sog. „**Halluzinationen"**. Der uneinheitlich verwendete Begriff kann im Sinne einer Generierung von in den Trainingsdaten nicht vorhandenen falschen Tatsachen durch LLMs verstanden werden. Halluzinationen lassen sich nicht nur (zumindest derzeit) nicht wirksam vermeiden, sondern es fehlt bereits an Methoden zur zuverlässigen Erkennung von Halluzinationen. Deren Ausmaß ist daher kaum quantifizierbar. Halluzinationen verstärken Datenschutzrisiken erheblich, weil LLMs nicht nur personenbezogene Daten aus den Trainingsdaten ausgeben, sondern über Betroffene auch abseits bloßer Verkürzungen Falschinformationen erzeugen können. Darüber hinaus können LLMs auch neue unrichtige Informationen über Personen erzeugen, die nicht in den Trainingsdaten in Bezug genommen werden, z. B. aufgrund personenbezogener Prompts. Enthalten Ausgaben von LLMs unrichtige Tatsachen über natürliche Personen, läuft dies dem Datenrichtigkeitsgebot (Art. 5 I lit. c DSGVO) zuwider.[4]

In der stark von Lobbyvertreterinnen beeinflusste Datenschutzdebatte werden die **Risiken für Betroffene** zum Teil vorschnell und pauschal als niedrig eingeordnet. Die mit Verarbeitungen im Kontext von LLMs verbundenen Risiken lassen sich jedoch nur für bestimmte Modelle, Implementierungen, Vertriebsmodelle, Anwendungsszenarien und Nutzerinnengruppen substantiiert bestimmen. So besteht etwa ein erheblicher Risikounterschied zwischen funktional begrenzten Implementierungen und öffentlich verfügbaren LLMs. So ist die Ausgabe personenbezogener Trainingsdaten bei einem Transkriptionsfeature in einem Messenger weitaus unwahrscheinlicher als bei einem öffentlich

---

2 S. nur *Bartels*, GRUR Int. 2024, 526 (529); *Franke*, RDi 2023, 565 (566).

3 Grundlegend *Pesch/Böhme*, MMR 2023, 917. So auch EDSA Opinion 28/2024, Rn. 36 ff.; *Engeler/Rolfes*, ZD 2024, 423; *Nolte/Finck/Meding*, arXiv:2503.01630. In der mündlichen Verhandlung des OLG Köln zu Metas KI-Training (15 UKl 2/25) hat der *HmbBfDI* laut zwei unabhängigen Quellen seine gegenteilige noch im Diskussionspapier: Large Language Models und personenbezogene Daten, S.5 ff., vertretene Auffassung explizit aufgegeben, vgl. auch *HmbBfDI*, Meta starts AI training with personal data (english version), https://datenschutz-hamburg.de/news/meta-starts-ai-training-with-personal-data (Abruf: 31.8.2025).

4 Ausführlich *Pesch/Böhme*, MMR 2023, 917 (921f.).

verfügbaren und in der Datenschutzforschung gezielt Extrahierungsmethoden ausgesetzten vortrainierten Modells.

Im Hinblick auf Risiken für Betroffene mehr in den Fokus der datenschutzrechtlichen Debatte muss die **Nutzung von LLMs im Bereich der Entscheidungsunterstützung** rücken. LLMs werden etwa im Bereich der Bewerberinnenauswahl[5] schon eingesetzt. Außerdem lassen sich LLMs zur automatisierten Generierung von Entscheidungsentwürfen einsetzen, die anschließend nur noch auf sachliche und rechtliche Fehler überprüft werden könnten.[6] Angesichts des Einsatzes von LLMs zur Entscheidungsunterstützung sind die Vereinbarkeit mit dem Verbot automatisierter Entscheidungen (Art. 22 DSGVO) sowie die Realisierbarkeit einer hinreichenden menschlichen Aufsicht (Art. 12 KI-Verordnung) dringend zu erforschen. Näher und für konkrete Entscheidungsunterstützungssysteme bzw. -prozesse ist zu klären, wie sich Bias in personenbezogenen Trainingsdaten auf Entscheidungen auswirkt. Solche Bias liegen insb. in Social-Media-Daten, wie sie Meta verwendet, in besonderem Umfang vor.

Im Zusammenhang mit Modellbias kommt **besonderen Kategorien personenbezogener Daten** (Art. 9 I DSGVO) in den Trainingsdaten besondere Relevanz zu. Wie etwa der Fall von COMPAS[7] deutlich macht, können im Bereich des maschinellen Lernens reine Proxy-Daten, also Daten, die lediglich indirekt auf besondere Kategorien personenbezogener Daten schließen lassen, zu erheblichen Modellbias und intensiven Grundrechtseingriffen führen. Der EuGH hat solche Proxy-Daten ausdrücklich in den Anwendungsbereich von Art. 9 I DSGVO einbezogen. Obwohl Einschränkungen der Anwendung der Vorschrift in bestimmten Fällen geboten sind, sind diese im Bereich von KI-Modellen im Bereich des maschinellen Lernens und damit auch LLMs allenfalls nach einer sorgfältigen Risikoanalyse begründbar. Eine solche erfordert in der Regel ausgiebige Tests.

Zusammenfassend ist festzustellen, dass es an generalisierbaren Erkenntnissen zu Datenschutzrisiken und der Wirksamkeit von Datenschutzmaßnahmen bei LLMs noch weitgehend fehlt. Eine Bewertung der Datenschutzkonformität erfordert eine sorgfältige Risikoanalyse unter Einbeziehung konkreter LLMs, Implementierungen, Vertriebsmodelle, Anwendungsszenarien und Nutzerinnengruppen. Die Intransparenz von Entwicklungen und Komplexität der Problematik erschwert oder verunmöglicht Gerichten und Aufsichtsbehörden substantiierte Einschätzungen. Die effektive Durchsetzung des Datenschutzrechts ggü. Anbieterinnen und Anwenderinnen gelingt nur, wenn das **Rechenschaftsprinzip** (Art. 5 II DSGVO) voll zur Geltung gebracht wird. Hierzu sind Anbieterinnen und Anwenderinnen darauf zu verweisen, zunächst eigene Studien durchzuführen und – statt bloßer Behauptungen – nachvollziehbare substantiierte Forschungsergebnisse zu Risiken für Betroffene und zur Realisierbarkeit und Wirksamkeit von Datenschutzmaßnahmen vorzulegen. Die damit verbundenen Datenverarbeitungen lassen sich auf berechtigte Interessen (Art. 6 I lit. f DSGVO) – ggf. i. V. m.

---

5 Z. B. durch die Lieferando-Mutter Just Eat Takeaway, https://web.archive.org/web/20230321060806/https://www.hirevue.com/case-studies/justeat (Abruf: 31.8.2025), zu Datenschutz- und KI-Regulierungsfragen *Kätscher/Pesch*, KIR 2024, 46.

6 *Pesch*, EJRR 2025, 76.

7 *Angwin et al.*, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (Abruf: 31.8.2025).

den in nationalen Datenschutzgesetzen niedergelegten Forschungsprivilegien – stützen, soweit Risiken für Betroffene hinreichend reduziert werden.[8] Nur wenn Anbieterinnen und Anwenderinnen– der Beweislastverteilung nach der DSGVO entsprechend – in die Pflicht genommen werden, Risiken substantiiert darzulegen, kann Regulierung wirksame Anreize zur Entwicklung datenschutzfreundlicher LLMs und LLM-basierter Systeme und Verfahren setzen.

II. Zu einzelnen der Konsultationsfragen

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?*

Zur Anonymität ist zunächst anzumerken, dass die DSGVO begrifflich einem binären Verständnis von Anonymität (anonym oder personenbezogen) folgt, während Anonymität auf technischer Ebene tatsächlich graduell ist (Identifizierbarkeit mit mehr oder weniger Aufwand). Diese Gradualität wird aber in der Abwägung, ob die Identifizierung nach allgemeinem Ermessen wahrscheinlich ist, gerade berücksichtigt. Ob eine Identifizierung von Betroffenen, mit deren Daten ein LLM trainiert wurde, möglich ist, lässt sich nur unter Berücksichtigung aller objektiven Faktoren, insb. dem Kosten- und Zeitaufwand für eine Identifizierung sowie verfügbaren und absehbaren Technologien, beantworten (vgl. Erwägungsgrund 26 Satz 4 DSGVO). Auch LLMs, aus denen personenbezogene Trainingsdaten extrahierbar sind, können also als anonym zu bewerten sein, wenn eine Identifizierung nach allgemeinem Ermessen nicht wahrscheinlich ist. Dabei lassen sich zu diesem Zeitpunkt angesichts des Forschungsstands nur begrenzt generalisierbare Aussagen zu den in Rn. 41 der Stellungnahme 28/2024 des EDSA aufgelisteten Faktoren treffen. Anbieterinnen und Anwenderinnen sind insoweit rechenschaftspflichtig und müssen die Risiken substantiiert. Insb. bei öffentlich verfügbaren LLMs für allgemeine Verwendungszwecke ist eine Identifizierung einiger Betroffener nach allgemeinem Ermessen höchstwahrscheinlich oder sogar sicher, da allein in der Datenschutzforschung ständig und mit Erfolg Versuche unternommen werden, auch personenbezogene Daten aus den Modellen zu extrahieren. Bei Chatbots können Nutzerinnen bereits durch einfache Prompts wie „Wer ist…“ Trainingsdaten extrahieren.

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.*

[8] *Pesch/Magnussen*, DuD 2025, 367.

anonymisierter Trainingsdaten, Fine-Tuning ohne personenbezogene Daten, Differential Privacy, etc.)? Welche Erfahrungen haben Sie damit gemacht?

Derzeit ist die Wirksamkeit von vorgeschlagenen Datenschutzmaßnahmen für LLMs noch nicht hinreichend abgesichert und verbleiben zahlreiche offene Forschungsfragen. Ich koordiniere das vom Bundesministerium für Forschung, Technologie und Raumfahrt (BMFTR) geförderte Projekt SMARD-GOV[9], in dessen Rahmen generalisierbare Erkenntnisse über die Wirksamkeit von Datenschutzmaßnahmen gewonnen werden sollen. Geplant ist etwa die Untersuchung der Auswirkungen unterschiedlicher Modellgrößen und Techniken für das Fine-Tuning (volles Fine-Tuning, LoRA, QLoRA).

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.

Wie zu Frage 1. ausgeführt, ist jedenfalls derzeit die Extrahierung personenbezogener Trainingdaten aus öffentlich verfügbaren LLMs für allgemeine Verwendungszwecke höchstwahrscheinlich bzw. schon gelungen. Bei der Bewertung der Datenschutzkonformität des Einsatzes solcher LLMs durch Beaufsichtigte kommt es aber auf das allgemeine Risiko der Extrahierung von Trainingsdaten aus solchen Modellen nicht entscheidend an. Vielmehr sind die mit den spezifischen Datenverarbeitungsvorgängen verbundenen Risiken entscheidend. Setzt ein Unternehmen ein LLM für bestimmte Aufgaben ein, wird sich dadurch das Identifikationsrisiko in Bezug auf das vortrainierte Modell regelmäßig nicht in relevantem Umfang erhöhen. Bei einem Fine-Tuning, z. B. mit unternehmensinternen Daten oder behördeneigenen Akten, muss sich die Risikobewertung auf das domänenspezifische Modell und seine konkrete Anwendung konzentrieren. Wenn Verantwortliche domänenspezifische Modelle entwickeln und einsetzen, rückt die Extrahierbarkeit von Trainingsdaten des vortrainierten Modells jedenfalls dann in den Hintergrund, wenn das Fine-Tuning die Extrahierbarkeit von Daten aus dem Vortraining nicht begünstigt. Der Schwerpunkt der Betrachtung liegt dann auf den für das Fine-Tuning verwendeten domänenspezifischen Daten und deren Extrahierbarkeit und den für Betroffene entstehenden Risiken. Je nach Anwendungsfall können Trainingsdaten von vornherein nicht-personenbezogen oder mit verhältnismäßigem Aufwand anonymisierbar sein und kann das Risiko der Extrahierung personenbezogener Fine-Tuning-Daten gering sein. Entwicklerinnen müssen die Risiken insoweit hinreichend substantiiert darlegen.

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

Es trifft zu, dass in der Extrahierbarkeit personenbezogener Trainingsdaten ein Hauptproblem von LLMs liegt. Datenschutzrechtliche Probleme ergeben sich hieraus insb. in Bezug auf Betroffenenrechte. So können LLMs das Recht auf Vergessenwerden aushöhlen, wenn etwa nach

[9] SMARD-GOV wird vom BMFTR i. R. d. Bekanntmachung „Plattform Privatheit" unter den Kennzeichen 16KIS2303K, 16KIS2304 und 16K1S2305 gefördert, https://www.digit.rw.fau.de/forschung/forschungsprojekt-smard-gov/ (Abruf: 31.8.2025).

Geltendmachung des Anspruchs aus Art. 17 DSGVO gegenüber einer dritten Verantwortlichen gelöschte oder nicht mehr Suchmaschinen-indexierte Daten aus LLMs extrahierbar sind. In Bezug auf Modelle selbst ist die Umsetzung des Rechts auf Vergessenwerden hochproblematisch. Zwar ist die Entfernung von Daten (*machine unlearning*) aus einem LLM entgegen Behauptungen von Anbieterinnen regelmäßig nicht unmöglich, sondern lediglich aufwändig. Allerdings kann die Entfernung personenbezogener Daten aus einem LLM den Datenschutz der darin verbleibenden Daten wahrscheinlicher machen, deren Extrahierung also erleichtern.[10] Den Betroffeneninteressen könnte insoweit im Einzelfall durch Filterlösungen angemessen Rechnung zu tragen sein.

Wie schon unter I. ausgeführt, ergeben sich überdies erhebliche Datenschutzrisiken aus Halluzinationen. Darüber hinaus kommt es auch jenseits von Halluzinationen zu Verstößen gegen das Datenrichtigkeitsgebot mit potenziell hohen Risiken für Betroffene, wenn unrichtige Trainingsdaten verwendet werden. Insb. Trainingsdaten aus sozialen Medien enthalten in großem Ausmaß unrichtige Informationen über natürliche Personen.

Die mit LLMs verbundenen Risiken sowie potenziellen Risiken sind insb. bei der Interessenabwägung i. R. v. Art. 6 I lit. f (berechtigtes Interesse) sorgfältig zu ermitteln und einzubeziehen. Auch in Eilverfahren verbietet sich eine Abwägung aufgrund bloßer Behauptungen und unsubstantiierter Annahmen.[11] In die Abwägung sind die vernünftigen Erwartungen von Betroffenen einzubeziehen (Erwägungsgrund 47 Satz 1 Halbsatz 2 DSGVO). In diesem Zusammenhang ist pauschalen Hinweisen auf die Erwartbarkeit des LLM-Trainings mit Online-Daten entgegenzutreten. Dass massenweise auch rechtswidrig oder möglicherweise rechtswidrig Daten für das Training von LLMs genutzt werden, lässt sich nicht ohne Wertungswidersprüche rechtfertigend für eine Datenverarbeitung heranziehen. Entsprechendes gilt auch im Rahmen der Zweckvereinbarkeitsprüfung nach Art. 6 IV DSGVO, wenn – wie bis auf wenige Ausnahmen üblich – zu anderen Zwecken erhobene Daten für das Training verwendet werden.

Außerdem werden LLMs bereits zur –ganzen oder teilweisen – Automatisierung von Entscheidungen eingesetzt und sind hier die Anforderungen von Art. 22 DSGVO, Art. 14 KI-VO zu konkretisieren. Die Anforderungen an die menschliche Aufsicht für LLM-basierte Entscheidungsunterstützung werden in SMARD-GOV[12] näher untersucht werden.

Bei LLM kann es datenschutzrechtlich auch relevant werden, dass Anbieterinnen LLM-basierter Systeme durch die Ausgestaltung des Systems und des Trainings sowie die Auswahl der Trainingsdaten erheblichen Einfluss auf die Ausgaben von LLM haben, während Nutzerinnen oft kaum Einblicke in die Trainingsdaten und Einfluss auf die Modelle haben. Hieraus kann sich eine gemeinsame Verantwortlichkeit der Anbieterinnen und Nutzerinnen für von Nutzerinnen durchgeführte Verarbeitungsvorgänge ergeben.[13]

---

[10] *Hayes et al.*, arXiv:2403.01218, 2024.

[11] Ein extremes Negativbeispiel bildet *OLG Köln*, Urt. v. 23.5.2025 – 15 UKl 2/25, wobei die fehlenden Ausführungen Urteil zum trainierten Modell und seinen Einsatzzwecken nahelegen, dass das Gericht nicht verstanden hat, dass Meta Llama, also ein Modell mit allgemeinem Verwendungszweck trainiert.

[12] Fn. 9.

[13] *Kätscher/Pesch*, KIR 2024, 46 (53).