Die Bundesbeauftragte
für den Datenschutz und
die Informationsfreiheit

Bonn, den 22.12.2025

KI in Behörden – Datenschutz von Anfang an mitdenken

Handreichung der BfDI für die Bundesverwaltung

Graurheindorfer Straße 153
53117 Bonn
FON +49 (0)228-997799-0
E-MAIL poststelle@bfdi.bund.de
INTERNET www.bfdi.bund.de

Bonn, den 22.12.2025

Inhaltsverzeichnis

Graurheindorfer Straße 153
53117 Bonn
FON +49 (0)228-997799-0
E-MAIL poststelle@bfdi.bund.de
INTERNET www.bfdi.bund.de

Bonn, den 22.12.2025

1. Einleitung

Technologien der Künstlichen Intelligenz (KI) eröffnen der öffentlichen Verwaltung vielfältige neue Möglichkeiten, Verwaltungsprozesse effizienter, zielgerichteter und bürgernäher zu gestalten. Insbesondere große Sprachmodelle, die die Grundlage für die gängigen Chatbots darstellen, rücken hierbei in den Fokus, da sie im Arbeitsalltag bei der Bewältigung einer Vielzahl von Aufgaben effektive Unterstützung leisten können.

Große Sprachmodelle verarbeiten häufig personenbezogene Daten, sei es im Training oder in konkreten Nutzungsszenarien. Werden KI-Modelle mit personenbezogenen Daten trainiert, können diese in abstrakter Form in den Parametern der Modelle repräsentiert sein und gegebenenfalls wieder ausgegeben werden. Zusätzlich können Daten in den Eingaben und Ausgaben Personenbezug aufweisen.

Der Fokus der vorliegenden Handreichung liegt auf KI-Systemen bestehend aus *großen Sprachmodellen* (im Folgenden *Large Language Models,* kurz: *„LLMs“).*[1]

Diese Handreichung soll öffentlichen Stellen helfen, sich im komplexen Geflecht relevanter Rechtsakte zurechtzufinden. Sie bezieht sich dabei auf die spezifischen datenschutzrechtlichen Fragen, die im Zusammenhang mit LLMs aufgeworfen werden. Die Handreichung soll dabei helfen, zentrale Fragestellungen zu identifizieren und eine strukturierte Herangehensweise an KI-Projekte zu entwickeln. Die Bundesbeauftragte für den Datenschutz und die Informationsfreiheit (BfDI) steht darüber hinaus beratend zur Seite – sei es bei konkreten Vorhaben oder bei weiterführenden Fragen zum datenschutzgerechten Umgang mit KI.

1.1. Bezug zur KI-VO

Die Entwicklung und Nutzung von KI-Systemen unterliegen verschiedenen normativen Vorgaben. Neben datenschutzrechtlichen Bestimmungen[2] sind insbesondere die Vorgaben der KI-Verordnung[3] zu beachten, die das Inverkehrbringen, die Inbetriebnahme und die Verwendung von KI-Systemen regelt. Nach Art. 2 Abs. 7 KI-VO lassen diese Vorgaben datenschutzrechtliche Bestimmungen „unberührt“. Gleichwohl kann es bei KI-bezogenen

[1] Wenn im Folgenden von *KI-System* die Rede ist, bezieht sich dies immer auf *LLM*-basierte *KI-Systeme*.

[2] Die datenschutzrechtlichen Ausführungen, die sich auf die DSGVO beziehen, stehen exemplarisch für datenschutzrechtliche Vorgaben. Auf die Zitierung weiterer einschlägiger Rechtsbestimmungen wird aus Gründen der Übersichtlichkeit verzichtet.

[3] Verordnung über künstliche Intelligenz VO (EU) 2024/1689, nachstehend: KI-VO.

Sachverhalten angezeigt sein, die DSGVO im Lichte der Regelungen und Wertungen der KI-VO auszulegen.[4] KI-VO und DSGVO ergänzen sich zu einem kohärenten unionsrechtlichen Regelungsrahmen für KI-Systeme, indem die KI-VO produktsicherheitsrechtliche Anforderungen an KI festlegt, während die DSGVO die Rechtmäßigkeit und Grenzen der Verarbeitung personenbezogener Daten bestimmt. Die Normen der KI-VO sind also parallel und gegebenenfalls im Wege einer rechtsaktübergreifenden Auslegung der DSGVO zu berücksichtigen. Die rechtsaktübergreifende Auslegung erscheint insbesondere dort sachgerecht, wo die KI-VO spezifischere Anforderungen an KI-Systeme normiert als die DSGVO, sodass sich bei der Erfüllung der Pflichten aus der DSGVO an den Vorgaben der KI-VO orientiert werden kann.[5]

Unabhängig davon, ob die Bestimmungen der KI-VO anzuwenden sind, gilt also weiterhin: Sobald personenbezogene Daten verarbeitet werden, sind die datenschutzrechtlichen Vorgaben zu beachten.

1.2. Datenschutzrechtliche Herausforderungen großer Sprachmodelle

LLMs beruhen auf sogenannten *tiefen neuronalen Netzen.* Zunächst werden diese mit kuratierten Trainingsdaten trainiert, auf Basis derer das Modell lernt, wie es sich zu verhalten hat („Pre-Training“). Technisch spiegelt sich dies in der Anpassung sehr vieler Gewichte („weights“) wider, die die unterschiedlichen Schichten („layers“) eines neuronalen Netzes verknüpfen. Durch das sog. Fine-Tuning kann ein Modell dann entsprechend auf den geplanten Einsatz zugeschnitten werden. Im anschließenden Produktivbetrieb können Eingaben erfolgen, woraufhin das LLM entsprechende Ausgaben liefert. In der datenschutzrechtlichen Bewertung von LLMs sind alle Lebenszyklusabschnitte eines LLMs einzeln zu betrachten, da sie jeweils eigene datenschutzrechtliche Implikationen aufweisen.

In dieser Handreichung wird der Begriff „KI-Modelle“ als Oberbegriff für trainierte neuronale Netze, insbesondere LLMs, verwendet. Werden solche KI-Modelle in eine Softwareanwendung oder Plattformen implementiert, werden sie im Folgenden als KI-Systeme bezeichnet.

---

[4] Vgl. Fuchs/Hansen/Walczak/Wünschelbaum, The Bridge Blueprint - Thesen zur einheitlichen Anwendung von Datenschutz- und KI-Regulierung beim KI-Einsatz, öffentlicher Entwurf v. 0.9.

[5] Bortnikov, DuD 2025, 297 (301) in Bezug auf Transparenzanforderungen.

Bonn, den 22.12.2025

Beispiel:

GPT 4 ist ein KI-Modell, das in die Webanwendung ChatGPT implementiert ist, die ein KI-System darstellt.

In der Anwendung ergeben sich unterschiedliche Möglichkeiten der KI-Bereitstellung. Beispielsweise können Modelle vom Betreiber selbst oder von externen Dienstleistern gehostet werden. LLMs erscheinen hierbei als „Black-Box“[6], da die unterschiedlichen, auch versteckten („hidden“) Schichten mathematisch zu komplex miteinander verstrickt sind, als dass sie für Menschen ihren Entscheidungsweg erkennen ließen. In diesem Zusammenhang werden datenschutzrechtliche Fragen im Hinblick auf die Transparenz aufgeworfen. LLMs generieren Texte auf Basis statistischer Wahrscheinlichkeiten ohne semantische Einsicht oder Faktenkenntnis. Sie vermögen damit nicht zu erkennen, ob ihre Ausgaben korrekt sind. Sie „halluzinieren“ dementsprechend gelegentlich.[7] Personenbezogene Daten sind zwar nicht im Klartext im Datenmodell enthalten. Vielmehr werden sämtliche Texte für das KI-Training in abstrakte Vektorrepräsentationen überführt.[8] Die Trainingsdaten fließen demnach in Form numerischer Vektoren in das Sprachmodell ein. Das Ergebnis findet sich in den Gewichten/Parametern des Modells wieder, die zwar einsehbar sind, jedoch Zahlen darstellen und somit in abstrakter Form Personenbezug aufweisen können[9]. In der Folge kann das Modell Fakten oder Phrasen und damit auch personenbezogene Daten aus den Trainingsdaten (re-)produzieren (Memorisierung[10]). Studien zeigen, dass insbesondere sehr große Modelle bei bestimmten Eingaben (Prompts) oft konkrete Passagen einschließlich personenbezogener Daten aus dem Training ausgeben.[11] Durch den großen Umfang der im

---

[6] Heinze/Sorge/Specht-Riemenschneider, KIR 2024, 11 (12).

[7] https://www.iese.fraunhofer.de/blog/halluzinationen-generative-ki-llm/.

[8] (1) Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. (2) Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

[9] Im Ergebnis a.A. HmbBfDI, Diskussionspapier: Large Language Models und personenbezogene Daten.

[10] Der Begriff der Memorisierung ist für den Zweck der vorliegenden Handreichung bewusst weiter gefasst als das allgemein anerkannte wissenschaftliche Verständnis und umfasst zusätzlich zu wortwörtlich reproduzierten Trainingsdaten auch sinngemäße Reproduktionen, da diese ebenfalls datenschutzrechtlich relevant sein können. (Auch in der Forschung gibt es bereits Empfehlungen, im Sinne des Datenschutzes nicht nur wortwörtliche Reproduktionen als Memorisierung zu betrachten: Ippolito, D., Tramèr, F., Nasr, M., Zhang, C., Jagielski, M., Lee, K., ... & Carlini, N. (2022). Preventing verbatim memorization in language models gives a false sense of privacy. arXiv preprint arXiv:2210.17546).

[11] Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., ... & Raffel, C. (2021). Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)* (pp. 2633-2650); Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., & Zhang, C. (2022, February). Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*; (3) Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., ... & Lee, K. (2023). Scalable extraction of training data from (production)

Training vorhandenen Daten können zudem personenbezogene Daten neu verknüpft werden. Dadurch kann die Verarbeitung mittels eines KI-Systems ein besonderes grundrechtliches Eingriffsgewicht begründen.[12]

Zusammengefasst ergeben sich hieraus unter anderem folgende datenschutzrechtliche Herausforderungen:

- **Black-Box-Charakter**. Durch die besondere Architektur eines tiefen neuronalen Netzes können die Outputs menschlich nicht ohne Weiteres nachvollzogen werden. Dies stellt eine Herausforderung, insbesondere für die Umsetzung der Betroffenenrechte wie dem Recht auf Auskunft, Löschung, Widerspruch oder Berichtigung, dar (→ 2.9 Transparenz und Informationspflichten / 2.10 Berichtigung und Löschung). Da personenbezogene Daten nicht im Klartext im LLM gespeichert sind, sondern in einer Vielzahl von Modellgewichten als statistische Muster und Wahrscheinlichkeiten verteilt vorliegen, ist es technisch schwierig, diese zu identifizieren, zu ändern oder zu entfernen.

- **Halluzinationen**. Ausgaben generativer Sprachmodelle entstehen auf Basis von Wahrscheinlichkeiten, welche das KI-Modell in der Trainingsphase gelernt hat. Das Wissen eines LLMs korreliert stark mit der Häufigkeit der Informationen in den Trainingsdaten. Insbesondere bei Daten, die nur sehr selten im Trainingsdatensatz vorkamen („Long-Tail Knowledge“), steht ein LLM vor Schwierigkeiten, diese Information akkurat auszugeben, ohne jedoch auf diese Einschränkung angepasst zu sein und auf ihre eigene Unkenntnis entsprechend hinzuweisen.[13] Die Ausgaben geben deshalb nicht zwangsläufig Informationen aus den Trainingsdaten wieder. Überdies können die Trainingsdaten falsche Informationen enthalten. Daraus ergibt sich eine Herausforderung für den Grundsatz der Datenrichtigkeit (Art. 5 Abs. 1 Buchst. d DSGVO) (→ 2.11. Datenrichtigkeit).

- **Memorisierung personenbezogener Daten in der KI.** KI-Modelle werden mit sehr vielen, häufig auch personenbezogenen, Daten trainiert. Diese personenbezogenen Daten können, obgleich sie nicht im Klartext im LLM gespeichert sind, unter Umständen im Output ausgegeben werden. Die Ausgabe dieser Daten kann unabsichtlich geschehen oder durch gezielte Attacken hervorgerufen werden. Es

language models. *arXiv preprint arXiv:2311.17035*; (4) Smith, V., Shamsabadi, A. S., Ashurst, C., & Weller, A. (2023). Identifying and mitigating privacy risks stemming from language models: A survey. *arXiv preprint arXiv:2310.01424*.

[12] BVerfG, Urteil vom 16.02.2023, 1 BvR 1547/19, 1 BvR 2634/20, Rn. 100.

[13] Kandpal, N., Deng, H., Roberts, A., Wallace, E., & Raffel, C. (2023). Large language models struggle to learn long-tail knowledge. In *International conference on machine learning* (pp. 15696-15707).

besteht zudem die Möglichkeit, dass durch die Zusammenführung personenbezogener Daten Rückschlüsse auf besondere Kategorien personenbezogener Daten im Sinne von Art. 9 DSGVO (→ 2.5. Besondere Kategorien personenbezogener Daten) gezogen werden können. Da die Verarbeitung solcher Datenarten wegen der damit verbundenen besonderen Risiken für die Rechte und Freiheiten der betroffenen Personen nur in eng begrenzten Fällen zulässig ist, besteht hier ein erhöhtes Risikopotenzial.

- **Bias/Fairness**. LLMs lernen von bestehenden Daten. Diese weisen häufig einen Bias auf, insofern als bestimmte Daten stärker repräsentiert sind als andere. Die Ausgabe von KI-Systemen erfolgt auf Basis dieser Repräsentationen und kann dementsprechend Diskriminierungen hervorrufen (→ 2.12. Fairness). In Bezug auf den Grundsatz von „Treu und Glauben“ (Art. 5 Abs. 1 Buchst. a DSGVO) kann dies zu Herausforderungen führen. Bias ist grundsätzlich kein neu geschaffenes Problem von LLMs, stellt jedoch in seiner Gewichtigkeit durch die Menge der verarbeiteten Daten sowie die häufig fehlende Nachvollziehbarkeit der Ausgaben und den Einfluss moderner KI-Systeme eine große Herausforderung dar.

- **Lebenszyklusphasen.** Der Lebenszyklus eines LLMs gliedert sich in mehrere Phasen, darunter Planung, Entwicklung, Implementierung, Betrieb sowie die kontinuierliche Validierung. Jede Phase des Lebenszyklus eines LLMs begründet eigenständige datenschutzrechtliche Anforderungen. Im Training könnte die hohe Anzahl personenbezogener Daten dazu führen, dass das Modell besonders viele personenbezogene Daten memorisiert. Darüber hinaus ergeben sich datenschutzrechtliche Herausforderungen bezüglich der Speicherbegrenzung (Art. 5 Abs. 1 Buchst. e DSGVO) (→ 2.6. Datenminimierung und Speicherbegrenzung). Im Produktivbetrieb können gezielte Prompts Datenleaks hervorrufen.

- **Inferenz.** LLMs werden immer besser darin, personenbezogene Informationen aus den Eingaben zu inferieren (abzuleiten).[14] Vermeintlich harmlose Sätze können dazu führen, dass ein LLM aus einem Prompt personenbezogene Informationen wie Herkunft, Wohnort oder Geschlecht ableiten kann. Beispielsweise wird das Wort „Kiez“ vor allem in Berlin als Bezeichnung für einen Stadtteil verwendet. Eine Anfrage, die das Wort „Kiez“ beinhaltet, ließe sich also recht einfach durch das

[14] Staab, Robin, et al. "Beyond memorization: Violating privacy via inference with large language models." *arXiv preprint arXiv:2310.07298* (2023).

Kontextwissen des LLMs auf Berlin als mutmaßlichen Ort der sprachlichen Sozialisation der eingebenden Person zurückführen (→ 2.1. Personenbezug).

- **KI-Bereitstellung**. KI-Modelle können entweder über die Cloud zur Verfügung gestellt werden und von einem externen Dienstleister gehostet werden („AI-as-a-Service“, kurz: „*AIaaS*“). Dieses Modell kann dann in der vorgefertigten Fassung verwendet werden („Off-the-shelf“) oder an die eigenen Bedürfnisse des Betreibers durch Fine-Tuning angepasst werden. Alternativ können Modelle auch lokal auf eigenen Servern gehostet werden. Je nach Ausgestaltung der technischen Architektur und der organisatorischen Einbindung eines KI-Systems können sich unterschiedliche Zuordnungen der datenschutzrechtlichen Verantwortlichkeit (→ 2.2. Verantwortlichkeit) ergeben.
- **Datenübermittlung in Drittländer**. Wie viele andere IT-Anwendungen werden auch LLMs oft in Drittländern oder durch Unternehmen aus Drittländern gehostet. Dieser Umstand birgt ein Risiko staatlicher Zugriffe auf die übermittelten und gespeicherten Daten oder gerichtliche Untersagung von Löschungen.[15] Im Falle einer Datenübertragung an Drittstaaten sind die Voraussetzungen der DSGVO einzuhalten.[16]

2. Datenschutzkonforme KI

2.1. Personenbezug

In jeder Lebenszyklusphase eines LLMs stellt sich die Frage, ob, in welchem Umfang und gegebenenfalls welche personenbezogenen Daten verarbeitet werden. Es ist wichtig, diesen Aspekt bereits in der Planungsphase zu berücksichtigen. Denn nur, wenn Verantwortliche wissen, welche konkreten Verarbeitungsvorgänge vorliegen und in Betracht kommen, können die Rechtmäßigkeit der Verarbeitungsvorgänge sichergestellt und **geeignete Schutzmaßnahmen** eingeplant werden.

Folgende Daten im Zusammenhang mit KI-Modellen und -Systemen können Personenbezug aufweisen:

[15] Beispiel: Löschungsuntersagung eines US-amerikanischen Gerichts gegen OpenAI, vgl. https://dsb.ruhr-uni-bochum.de/beitrag/nytvsopenai/.

[16] EDSA – Übersicht über die verabschiedeten Leitlinien: https://www.edpb.europa.eu/our-work-tools/general-guidance/guidelines-recommendations-best-practices_en.

- **Trainingsdaten**. Die für die Entwicklung eines LLMs gesammelten Daten enthalten in der Regel personenbezogene Daten, insbesondere bei einer Sammlung von Trainingsdaten mittels Web Scraping[17]. Auch nach einer Aufbereitung und Filterung dieser Datensätze kann oft weiterhin Personenbezug bestehen, eine vollständige Anonymisierung ist in diesen Fällen praktisch kaum durchführbar.[18] Dementsprechend wird in vielen Fällen die DSGVO auf das KI-Modell selbst anzuwenden sein. Für die Prüfung der Identifizierungsrisiken kommen alle Mittel in Betracht, die von dem Verantwortlichen oder einer anderen Person nach allgemeinem Ermessen wahrscheinlich genutzt werden, um eine natürliche Person direkt oder indirekt zu identifizieren.[19]

- **Memorisierte Daten** (Personenbezug des KI-Modells). Die personenbezogenen Trainingsdaten finden auf abstrakte Weise Einzug in die Sprachmodelle. Dies geschieht dadurch, dass Informationen aus den Trainingsdaten insbesondere in den Gewichten „memorisiert“ werden (siehe Beschreibung unter 1.2) und dadurch unter bestimmten Bedingungen wieder reproduziert werden können. Die Folge ist, dass auch das eigentliche KI-Modell (über Ein- und Ausgaben hinaus) Personenbezug aufweisen kann.[20] Diese memorisierten Daten können gegebenenfalls in den Ausgaben des KI-Systems reproduziert werden.

- **Eingaben und Interaktionsdaten**. Die Eingaben der einzelnen Nutzer (Prompts) können personenbezogene Daten enthalten. Außerdem können in der Interaktion mit dem KI-System weitere personenbezogene Daten entstehen: die Login-Daten der Nutzenden sowie Metadaten über die Nutzenden wie Tippgeschwindigkeit, Uhrzeit der Nutzung etc. Zudem besteht die Möglichkeit, dass anhand der Eingabetexte personenbezogene Daten über die Nutzenden ableitbar sind (zum

[17] Beim Web Scraping werden häufig unstrukturierte Daten automatisiert aus dem Internet zusammengetragen, um sie anschließend für das Training von KI-Modellen zu verwenden.

[18] Lison, P., Pilán, I., Sanchez, D., Batet, M., & Øvrelid, L. (2021). Anonymisation models for text data: State of the art, challenges and future directions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 4188-4203).

[19] Beispielhafte Auflistung solcher objektiven Faktoren in EDSA, Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, Rn. 41.

[20] EDSA, Opinion 28/2024, Rn 38; im Ergebnis a.A. HmbBfDI, Diskussionspapier: Large Language Models und personenbezogene Daten.

Beispiel Geschlecht, Einkommen etc.) (→ 1.2. Datenschutzrechtliche Herausforderungen/ Inferenz).[21]

- **Ausgaben.** Die Ausgaben von KI-Systemen können aus verschiedenen Gründen Personenbezug aufweisen. In der Ausgabe können personenbezogene Daten aus der Eingabe, aus einer anderen verbundenen Quelle oder aus dem KI-Modell selbst reproduziert werden. Zudem können personenbezogene Daten „halluziniert" werden (→ 1.2. Datenschutzrechtliche Herausforderungen/ Halluzination).

2.2. Verantwortlichkeit

Verantwortlicher im Sinne der DSGVO ist diejenige Stelle, die über Zwecke und Mittel der Verarbeitung personenbezogener Daten entscheidet (Art. 4 Nr. 7 DSGVO) – und damit die rechtliche Hauptverantwortung für die Einhaltung der datenschutzrechtlichen Pflichten trägt.

Für jede Verarbeitung personenbezogener Daten bei Entwicklung und Einsatz von LLMs sind die Rollen der beteiligten Akteure vorab festzulegen und die Verantwortlichkeit im Sinne von Art. 4 Nr. 7 DSGVO zu prüfen. Diese Beurteilung muss differenziert für die einzelnen Verarbeitungsvorgänge und unter Berücksichtigung aller maßgeblichen Umstände des Einzelfalls erfolgen.[22]

Für öffentliche Stellen ist zu beachten, dass sich die Verantwortlichkeit aus Rechtsvorschriften ergeben kann (Art. 4 Nr. 7 Hs. 2 DSGVO). Im Übrigen sind die Rollen der Akteure nach ihrem faktischen Einfluss zu bestimmen.[23]

2.2.1. Entwicklung

Wer an der Entwicklung eines KI-Modells beteiligt ist, kommt als Verantwortlicher hinsichtlich der Sammlung und Aufbereitung personenbezogener Daten zum Training sowie der Verwendung dieser Daten zum Training des LLMs in Betracht. Das gleiche gilt für die Sammlung und Aufbereitung von Daten zum Fine-Tuning und das Fine-Tuning selbst, sowie für ein späteres Fine-Tuning und weiteres Training eines Basismodells.

---

[21] Staab, R., Vero, M., Balunović, M., & Vechev, M. (2023). Beyond memorization: Violating privacy via inference with large language models. *arXiv preprint arXiv:2310.07298*. https://arxiv.org/abs/2310.07298v1.

[22] EuGH, Urteil vom 29.07.2019, C-40/17, Rn. 70.

[23] EDSA, Leitlinien 07/2020 zu den Begriffen „Verantwortlicher" und „Auftragsverarbeiter" in der DSGVO, Rn. 12.

Beispiel:

Eine Behörde möchte ein LLM entwickeln, um es selbst zum Zwecke ihrer Aufgabenerfüllung zu verwenden. Hierzu sammelt die Behörde Daten, erstellt daraus Datensätze zum Training und bereitet diese auf. Anschließend trainiert sie mit diesen Datensätzen ein LLM und führt ein Fine-Tuning durch. Sofern diese Datensätze personenbezogene Daten enthalten, werden diese durch die einzelnen Schritte verarbeitet und die Behörde ist hierfür verantwortlich.

2.2.2. Produktivbetrieb

2.2.2.1. Personenbezogene Daten in Eingaben und Ausgaben

Die Verantwortlichkeit für die Verarbeitung personenbezogener Daten in Eingaben und Ausgaben eines KI-Modells unterscheidet sich nicht grundsätzlich von der automatisierten Datenverarbeitung mithilfe anderer technischer Mittel. Wenn eine öffentliche Stelle entscheidet, ein KI-Modell zu eigenen Zwecken einzusetzen und personenbezogene Daten in Ein- und Ausgaben zu verwenden, handelt sie insofern als Verantwortlicher. Das umfasst die Verantwortlichkeit für memorisierte Daten, sofern diese in den Ausgaben reproduziert werden. Wenn das KI-Modell bei einer externen Stelle für die öffentliche Stelle gehostet wird, ist diese externe Stelle hinsichtlich personenbezogener Daten in Ein- und Ausgaben in der Regel Auftragsverarbeiter, die Behörde bleibt verantwortlich.[24] Sofern die externe Stelle diese Daten außerdem zu eigenen Zwecken wie beispielsweise dem Training des KI-Modells verwendet, ist sie für diese Verarbeitungstätigkeiten verantwortlich.

2.2.2.2. Memorisierte Daten

Hinsichtlich der Verarbeitung memorisierter Daten beim Einsatz eines LLMs hängt die Zuordnung der Verantwortlichkeit von der konkreten Weise der Implementierung des KI-Modells und den Einflussmöglichkeiten auf die Entscheidung über Zwecke und Mittel der Verarbeitung ab. Hierbei ist zu berücksichtigen, dass Verantwortlichkeit gemäß Art. 4 Nr. 7 DSGVO sich nicht per se auf bestimmte Daten bezieht, sondern an Verarbeitungsvorgänge anknüpft. Die DSGVO kennt entsprechend Szenarien, in denen ein Akteur personenbezogene Daten von Dritten erhält und dann selbst für Verarbeitungen dieser Daten verantwortlich ist, sobald und soweit er dann selbst über Zwecke und Mittel der Verarbeitung entscheidet. Eine etwaige Rechtswidrigkeit früherer Verarbeitungsvorgänge

[24] Vgl. DSK, Orientierungshilfe „Künstliche Intelligenz und Datenschutz“ vom 06.05.2024, Rn. 32.

strahlt nicht pauschal auf spätere Verarbeitungen – zumal durch einen anderen Verantwortlichen – aus, sodass auch ein rechtswidrig trainiertes KI-Modell unter Umständen rechtmäßig eingesetzt werden kann[25] (→ 2.14.1. Rechenschaftspflicht hinsichtlich des Trainings des KI-Modells durch den Hersteller).

Maßgebliche Faktoren zur Bestimmung der Verantwortlichkeit für Verarbeitungen memorisierter Daten beim Einsatz von LLMs sind beispielsweise die Art des Zugriffs auf das LLM, ob es sich um eine exklusiv für die Behörde betriebene Modellinstanz handelt und inwieweit das Modell durch die Behörde bzw. auf Veranlassung der Behörde anpassbar ist.

- **Nutzung eigener Infrastruktur.** Wenn eine Behörde ein LLM auf eigener Infrastruktur (eigenes Rechenzentrum oder eigener Server) betreibt (*on premise*), hat sie direkten Zugriff auf das Modell und entscheidet über Speicherung und Verwendung der memorisierten Daten; sie handelt damit als Verantwortlicher.

- **Nutzung fremder Infrastruktur**. Wenn ein LLM auf externer Infrastruktur eines Dritten betrieben wird, hängt die Verantwortlichkeit für die Verarbeitung der memorisierten Daten von den konkreten Einflussmöglichkeiten der Behörde ab. Im Folgenden werden beispielhaft drei Szenarien angesprochen, maßgeblich sind aber immer die konkreten Umstände des Einzelfalls.

    - Stellt der Dritte lediglich Rechenkapazität oder Speicher zur Verfügung (beispielsweise als Cloud-Dienstleister) und die Behörde entscheidet, dass auf dieser Infrastruktur ein LLM betrieben werden soll, wählt das LLM aus und speichert es auf der Infrastruktur, hat direkten Zugriff auf das Modell (open weight model) und verwendet es zu eigenen Zwecken, dann ist die Behörde – wie bei der Nutzung eigener Infrastruktur – für die Verarbeitung personenbezogener Daten in dem LLM verantwortlich. Der Dritte handelt als Auftragsverarbeiter für die Behörde.

    - Ferner tritt die Konstellation auf, dass ein Dritter der Behörde die (Mit-)Nutzung eines geschlossenen KI-Modells über eine Schnittstelle ermöglicht, dabei aber die alleinige Kontrolle über die Parameter (closed weights) behält. Die Behörde hat selbst keine Einwirkungsmöglichkeit auf das Modell und kann auch nicht dahingehend auf den Dritten einwirken, dass dieser das Modell für sie verändert oder die Modellinstanz löscht. In dieser Konstellation ist die Behörde in der Regel nicht für die Verarbeitung der personenbezogenen Daten im LLM

[25] Specht-Riemenschneider/Bortnikov/Steffens, KIR 2025, 213 (214 f.).

verantwortlich.[26] Dasselbe gilt in der Regel für die Verarbeitung memorisierter Daten, die bei der Nutzung eines Chatbots über den Webbrowser oder eine App erfolgt.

- Zudem ist das Szenario anzutreffen, dass die Behörde bei einem Dritten ein Basismodell (AI-off-the-shelf) auswählt und über eine Schnittstelle Zugriff auf eine eigene Instanz des Modells erhält, das sie durch Fine-Tuning und weiteres Training anpassen kann. In diesem Fall ist die Behörde für die Verarbeitung der memorisierten Daten in dieser Modellinstanz verantwortlich, *soweit und solange* sie durch ihre Möglichkeiten, das Modell anzupassen, erheblichen Einfluss auf die memorisierten Daten nehmen und zudem die Löschung der Modellinstanz veranlassen kann. Ein fehlender Zugang zu den memorisierten Daten steht einer Verantwortlichkeit nicht entgegen.[27] Der Dritte ist regelmäßig Auftragsverarbeiter.

Verarbeitet ein Dritter personenbezogene Daten (sowohl in Ein- und Ausgaben als auch memorisierte Daten) im Auftrag einer verantwortlichen Behörde, hat diese mit dem Dritten einen Auftragsverarbeitungsvertrag nach Art. 28 Abs. 3 DSGVO zu schließen.

2.3. Zweckbestimmung

In jeder Phase des KI-Lebenszyklus muss bei der Verarbeitung personenbezogener Daten vorab der Zweck der Verarbeitungen festgelegt werden. Im Folgenden werden die Phasen der Entwicklung und des Einsatzes beleuchtet.

2.3.1. Entwicklung

Für Verarbeitungen im Rahmen der KI-Entwicklung, zum Beispiel im Zusammenhang mit dem KI-Training, darf bei der Zweckfestlegung nicht lediglich „KI-Training“ als Zweck aufgeführt werden, sondern es bedarf weiterer Konkretisierungen. So müssen bei der Zweckbeschreibung die Art von KI-Modell (zum Beispiel LLM), die absehbaren Funktionalitäten und (zu diesem Zeitpunkt so konkret wie möglich) der spätere Einsatzzweck des Modells angegeben werden.[28]

Beispiel:

[26] Vgl. EDSA, Leitlinien 07/2020, Rn. 27 Beispiel Telekommunikationsdienstleister.

[27] EuGH, Urteil vom 10.07.2018, C-25/17, Rn. 68 f.

[28] EDSA, Opinion 28/2024, Rn. 64.

Die Daten aus dem E-Mailverkehr der Behörde werden benutzt, um ein LLM zu trainieren. Dieses soll dazu benutzt werden, die späteren E-Mail-Eingänge automatisiert an die zuständigen Bearbeiterinnen und Bearbeiter innerhalb der Behörde weiterzuleiten.

Die Zweckbestimmung muss unter anderem im Verzeichnis der Verarbeitungstätigkeiten (Art. 30 Abs. 1 Buchst. b DSGVO) dokumentiert werden.

Anbieter von KI-Modellen mit allgemeinem Verwendungszweck im Sinne des Art. 3 Nr. 63 KI-VO können außerdem einer Dokumentationspflicht gemäß Art. 53 KI-VO unterliegen. Das KI-Büro hat eine Vorlage zu dieser Dokumentation veröffentlicht.[29] Die dortigen Angaben zu „Intended uses“, „Model architecture“, „Input modalities“ und „Output modalities“ entsprechen der datenschutzrechtlichen erforderlichen Dokumentation der Zweckbestimmung für KI-Training. Diese Angaben können in das Verzeichnis der Verarbeitungstätigkeiten übertragen werden, um den Zweck zu beschreiben.

Ein weiterer Zweck im Rahmen der KI-Entwicklung kann auch die Erstellung und Speicherung eines Datensatzes sein, der für weitere geplante Fortentwicklung von KI-Modellen oder zur Erfüllung von Betroffenenrechten vorgehalten werden muss. Dies ist relevant auch für die Prinzipien der Datenminimierung und der Speicherbegrenzung (→ 2.6. Datenminimierung und Speicherbegrenzung).

2.3.2. Produktivbetrieb

Beim Einsatz von KI ist es nicht erforderlich, den Einsatz der Technologie ausdrücklich im Rahmen der Zweckfestlegung zu benennen. Es genügt, wenn der konkrete Zweck der Verarbeitung technologieneutral beschrieben wird.

Beispiel:

Wird ein KI-System für die automatisierte Zuordnung von eingehenden E-Mails zu den einzelnen Sachbearbeiterinnen und Sachbearbeiter eingesetzt, bedarf es keiner gesonderten Nennung der KI in der Zweckfestlegung. Der Einsatz der KI ist lediglich ein technisches Mittel zur Erreichung des festgelegten Zwecks[30], der mit der Verarbeitung von E-Mail-Eingängen verfolgt wird.

[29] https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai (dort unter „Model Documentation Form“).

[30] Vgl. Bortnikov/Dukart, ZD 2024, 558 (559).

Bonn, den 22.12.2025

2.4. Rechtsgrundlagen

Die Verarbeitung personenbezogener Daten im Zusammenhang mit KI-Modellen und KI-Systemen erfordert in jeder Lebenszyklusphase der KI eine belastbare datenschutzrechtliche Grundlage gemäß Art. 6 Abs. 1 DSGVO. Hier erfolgt eine Darstellung für die Phasen der Entwicklung und des Einsatzes von KI.

2.4.1. Entwicklung

2.4.1.1. Spezifische Rechtsgrundlagen

Es liegen vereinzelt spezifische Rechtsgrundlagen für die Verarbeitung personenbezogener Daten zur Entwicklung von KI-Systemen oder -Modellen vor.[31]

2.4.1.2. Erfüllung einer öffentlichen Aufgabe

Jenseits dieser spezifischen Rechtsgrundlagen ist für die öffentlichen Stellen des Bundes die Generalklausel der Verarbeitung zur Erfüllung einer öffentlichen Aufgabe gemäß Art. 6 Abs. 1 Buchst. e DSGVO i.V.m. § 3 BDSG i.V.m. der jeweiligen Aufgabenzuweisungsnorm zentral.

Diese Rechtsgrundlage kommt zur Verarbeitung personenbezogener Daten für die KI-Entwicklung in Betracht, wenn die personenbezogenen Daten von der Behörde rechtmäßig zur Erfüllung dieser ihr zugeordneten öffentlichen Aufgabe erhoben wurden und die KI-Entwicklung genau der Erfüllung dieser Aufgabe dienen soll. Dabei muss bei der Prüfung der Erforderlichkeit der Rechtsgrundlage eine Abwägung zwischen den aufgabenbezogenen Belangen der Behörden mit den Grundrechten der Betroffenen vorgenommen werden.

Bei der Abwägung ist insbesondere zu berücksichtigen, ob die gesetzliche Aufgabe der Behörde ohne die intendierte Datenverarbeitung nicht oder nicht in gleicher Weise erfüllt werden kann. Effizienz- und Effektivitätserwägungen sind relevante Abwägungsfaktoren, können aber nicht allein maßgeblich sein.

Bei der Weiterverarbeitung von bereits erhobenen personenbezogenen Daten ist zu beachten, dass auch bei Zweckvereinbarkeit im Sinne des Art. 5 Abs. 1 Buchst. b i.V.m. Art. 6 Abs. 4 Hs. 2 DSGVO der Verantwortliche zu prüfen hat, ob für diese Weiterverarbeitung eine Rechtsgrundlage im Sinne des Art. 6 Abs. 1 DSGVO vorliegt.[32] Allein ausschlaggebend für die Weiterverarbeitung von personenbezogenen Daten zur KI-Entwicklung ist, dass eine

[31] Vgl. z.B. § 37a PolDVG Hamburg.

[32] Vgl. jüngst VG Hannover, Urteil vom 05.06.2025 – 10 A 4017/23, BeckRS 2025, 28017 Rn. 168 mit weiteren Nachweisen aus Rechtsprechung und Literatur.

Rechtsgrundlage (zum Beispiel Art. 6 Abs. 1 Buchst. e DSGVO i.V.m. § 3 BDSG i.V.m. der jeweiligen Aufgabenzuweisungsnorm) für die Weiterverarbeitung vorliegt.

Aufgrund des Charakters von § 3 BDSG als Generalklausel können damit nur Datenverarbeitungen mit geringfügiger Eingriffsintensität gerechtfertigt werden.[33]

Folgende Aspekte sprechen zum Beispiel für eine Überschreitung der Geringfügigkeitsschwelle:

- Entwicklung eines KI-Modells mit allgemeinem Verwendungszweck mit systemischen Risiken (insbesondere große generative Modelle) im Sinne des Art. 51 KI-VO.
- Zusammenführung einer großen Anzahl personenbezogener Daten, insbesondere aus unterschiedlichen internen Datenquellen zu einer zusammenhängenden Trainingsdatenbank.

Die Eingriffsintensität kann durch geeignete Mitigationsmaßnahmen durch die verantwortliche Stelle gesenkt werden. Oft vermögen verschiedene Mitigationsmaßnahmen erst in ihrer Kombination eine signifikante Reduktion der Eingriffsintensität zu bewirken.

Beispiele für mitigierende Maßnahmen:

- Pseudonymisierung der Trainingsdaten vor dem Training,
- bestmögliches Entfernen personenbezogener Daten, wie Namen, Telefon- und Steuernummern vor dem Training,
- Differential Privacy, anhand derer der Datensatz vor dem Training möglichst anonymisiert wird,
- Filter im KI-System, die eine Extraktion personenbezogener Daten aus dem KI-Modell möglichst mindern,
- Maßnahmen, die den Zugriff auf die Trainingsdaten regeln und ihren Missbrauch mindern.

Einzelheiten zu den verschiedenen mitigierenden Maßnahmen finden sich am Ende dieser Handreichung (→ 3. Übersicht und Erläuterung mitigierender Maßnahmen anhand des KI-Lebenszyklus).

[33] BT-Drs. 18/11325, 81 (zu § 3); BVerwG, Urteil vom 20.03.2024 – 6 C 8.22, NVwZ 2024, 1163 Rn. 44; vgl. BayVGH, Beschluss vom 15.02.2024 – 4 CE 23.2267, ZD 2024, 293 Rn. 19 f.; Gola/Heckmann/Starnecker, 3. Aufl. 2022, BDSG § 3 Rn. 13; Kühling/Buchner/Petri, 4. Aufl. 2024, BDSG § 3 Rn. 9.

Für das Training, das die Geringfügigkeitsschwelle überschreitet, ist die Schaffung spezifischer Rechtsgrundlagen durch den Gesetzgeber erforderlich. Hier sollte die Behörde aktuelle Entwicklungen im Blick behalten.

2.4.2. Produktivbetrieb

2.4.2.1. Spezifische Rechtsgrundlagen

Wenn die Verarbeitung von personenbezogenen Daten mittels KI einen schwerwiegenden Eingriff darstellt, ist eine spezifische Rechtsgrundlage erforderlich.[34] Pauschal stellt die Verarbeitung von personenbezogenen Daten mittels KI-Systemen aufgrund der Technologieoffenheit der DSGVO aber keinen schwerwiegenden Eingriff dar.[35] Die Schwere des Eingriffs ist in jedem Einzelfall unter Abwägung aller Umstände zu ermitteln.

2.4.2.2. Erfüllung einer öffentlichen Aufgabe

Auch für den Produktivbetrieb kann Art. 6 Abs. 1 Buchst. e DSGVO i.V.m. § 3 BDSG i.V.m. mit der jeweiligen Aufgabenzuweisungsnorm eine Rechtsgrundlage darstellen. Wie oben ausgeführt, können durch § 3 BDSG aber nur geringfügige Eingriffe in das Grundrecht auf informationelle Selbstbestimmung gerechtfertigt werden. Die Eingriffstiefe kann durch mitigierende Maßnahmen gemindert werden.

Folgende Aspekte sprechen für eine Überschreitung der Geringfügigkeitsschwelle:

- Das KI-System führt zu einer automatisierten Entscheidung im Sinne des Art. 22 DSGVO.
- Es werden sensible Daten oder besonders viele Daten über eine konkrete Person verarbeitet, um neues personenrelevantes Wissen zu erzeugen.[36]

Folgende mitigierende Maßnahmen können die Eingriffstiefe abmildern:

- Benutzung durch geschultes Personal,

[34] Vgl. BVerfG, Urteil vom 16.02.2023, 1 BvR 1547/19, 1 BvR 2634/20, Rn. 71.

[35] Vgl. Reichert/Radtke/Eske, ZD 2024, 483 (487).

[36] BVerfG, Urteil vom 16.02.2023, 1 BvR 1547/19, 1 BvR 2634/20, Rn. 77.

- Zugriffs-, Rechte- und Rollenkonzept, beispielsweise ermöglicht durch eine RAG-Architektur,
- Technische Beschränkungen der Herausgabe personenbezogener Daten, die eventuell in den KI-Modellen enthalten sein können, beispielsweise Guardrails, Output-Filter, ein System-Prompt, der die Ausgabe personenbezogener Daten verbietet etc.[37],
- Anweisungen und Leitfäden hinsichtlich der Arbeitsabläufe mit dem jeweiligen KI-System.

Memorisierte Daten:

Wie oben geschildert finden personenbezogene Trainingsdaten auf abstrakte Weise Einzug in die Gewichte des Sprachmodelle, wodurch auch das KI-Modell (über Ein- und Ausgaben hinaus) Personenbezug aufweisen kann (→ 1.2. Datenschutzrechtliche Herausforderungen großer Sprachmodelle).[38] Der Betrieb eines solchen KI-Modells, zum Beispiel auf der eigenen Serverinfrastruktur, stellt somit eine Verarbeitung personenbezogener Daten dar, die ebenfalls einer Rechtsgrundlage bedarf. Der Betrieb personenbezogener LLMs kann auf die Rechtsgrundlage der Erfüllung einer öffentlichen Aufgabe gestützt werden, wenn insbesondere das Risiko der Extraktion von Trainingsdaten aus den Modellgewichten entsprechend dem Anwendungsfall gemindert ist. Dafür müssen die Extraktionswahrscheinlichkeit betrachtet und gegebenenfalls Maßnahmen zur Senkung dieser Wahrscheinlichkeit auf Systemebene ergriffen werden (→ 3. Übersicht und Erläuterung mitigierender Maßnahmen anhand des KI-Lebenszyklus).

2.4.2.3. Erfüllung einer gesetzlichen Verpflichtung

Ebenfalls relevant ist die Rechtsgrundlage der Verarbeitung zur Erfüllung einer gesetzlichen Verpflichtung gemäß Art. 6 Abs. 1 Buchst. c DSGVO i.V.m. der jeweiligen gesetzlichen Norm. Aufgrund der Technologieneutralität der DSGVO ist eine Verarbeitung mittels KI grundsätzlich nicht anders zu bewerten als herkömmliche technische Verfahren.

[37] Siehe → 3. Übersicht und Erläuterung mitigierender Maßnahmen anhand des KI-Lebenszyklus.

[38] Im Ergebnis a.A. HmbBfDI, Diskussionspapier: Large Language Models und personenbezogene Daten.

Bonn, den 22.12.2025

2.5. Besondere Kategorien personenbezogener Daten

Die Verarbeitung besonderer Kategorien personenbezogener Daten – wie beispielsweise biometrischer Daten zur eindeutigen Identifizierung einer natürlichen Person, Gesundheitsdaten oder Daten zum Sexualleben – unterliegt besonderen Anforderungen. Eine solche Verarbeitung ist nur erlaubt, wenn einer der in Art. 9 Abs. 2 DSGVO aufgeführten Tatbestände erfüllt ist oder wenn eine Verarbeitung auf Grundlage eines nationalen Erlaubnistatbestands erfolgt, der auf einer Öffnungsklausel des Art. 9 Abs. 2 DSGVO basiert. Für die Verarbeitung von den besonderen Kategorien personenbezogener Daten im KI-Kontext ist Folgendes zu beachten.

2.5.1. Entwicklung (z.B. Erheben von Daten für das Training und das Training selbst)

Für die Entwicklung einiger leistungsfähiger KI-Modelle sind in der Regel sehr große unstrukturierte Datenmengen erforderlich. Bei einigen großen Datensätzen sind besondere Kategorien personenbezogener Daten oft nicht praktikabel vor dem KI-Training herauszufiltern.[39] Um die Betroffenen dennoch effektiv zu schützen und gleichzeitig die Entwicklung jedenfalls solcher KI-Modelle und -Systeme zu ermöglichen, die der Verbesserung von Verwaltungshandeln und damit Gemeinwohlinteressen dienen, ist es aufgrund des risikobasierten Ansatzes der DSGVO sowie im Wege der rechtsaktübergreifenden Auslegung (→ 1.1. Bezug zur KI-VO) zwecks Verwirklichung der Ziele der KI-VO zulässig und auch geboten, hinsichtlich der KI-Entwicklung zwischen zielgerichteter und nicht-zielgerichteter Verarbeitung von besonderen Kategorien personenbezogener Daten zu differenzieren.[40]

Zielgerichtet ist eine Verarbeitung dann, wenn sie den Zweck verfolgt, die spezifischen Merkmale, die personenbezogene Daten zu besonderen Kategorien von Daten im Sinne von Art. 9 Abs. 1 DSGVO machen, einer bestimmten Person zuordenbar auszuwerten.

Eine nicht-zielgerichtete Verarbeitung von sensiblen Daten liegt bei der Entwicklung von KI-Modellen vor, wenn sensible Daten verarbeitet werden, um dem KI-Modell allgemeine Informationsmuster beizubringen, ohne die Eigenschaften der sensiblen Daten einer bestimmten Person zuordbar zu machen. Eine nicht-zielgerichtete Verarbeitung besonderer Kategorien personenbezogener Daten ist insbesondere dann weniger risikogeneigt als eine zielgerichtete Verarbeitung besonderer Kategorien personenbezogener Daten, wenn der

[39] Lison, P., Pilán, I., Sanchez, D., Batet, M., & Øvrelid, L. (2021). Anonymisation models for text data: State of the art, challenges and future directions. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 4188-4203).

[40] Vgl. OLG Köln, Urteil vom 23.05.2025, 15 UKI 2/25, Rn. 120.

Verantwortliche erhöhte Anforderungen an mitigierende Maßnahmen erfüllt[41] und damit auch tatsächlich das Risiko sinkt, dass besondere Kategorien personenbezogener Daten memorisiert werden.

Indizien, die im Rahmen einer wertenden Betrachtung für eine nicht-zielgerichtete Verarbeitung von sensiblen personenbezogenen Daten beim KI-Training sprechen, liegen zum Beispiel in den folgenden Fällen vor:

- Die Trainingsdaten werden um personenbezogene Daten, wie Namen, Adressen und Telefonnummern bereinigt,
- Entfernung von doppelten Einträgen aus den Trainingsdaten (Deduplikation), um die Überanpassung (Overfitting) der Modelle an die Trainingsdaten zu vermeiden,
- Fine-Tuning des Modells, das das Modell lehrt, personenbezogene Daten nicht herauszugeben,
- Datenquellen, die besonders wahrscheinlich sensible Daten enthalten, werden nicht für die Sammlung der Trainingsdaten verwendet.

Bei nicht-zielgerichteten Verarbeitungen besonderer Kategorien personenbezogener Daten ist das in Art. 9 Abs. 1 DSGVO vorgesehene Verarbeitungsverbot nur auf der Grundlage eines Antrags der betroffenen Person anwendbar. (→ vgl. zur Löschung 2.10.2 Löschung).[42] Das heißt: macht eine betroffene Person geltend, dass besondere Kategorien personenbezogener Daten von ihr memorisiert sind, muss geprüft werden, ob für diese Daten einer der Ausnahmetatbestände des Art. 9 Abs. 2 DSGVO vorliegt. Ist dies nicht der Fall, erfolgt die Verarbeitung rechtswidrig und der Verantwortliche muss entsprechend reagieren. Ein beanstandungswürdiger Verstoß gegen Art. 9 Abs. 1 DSGVO kommt aber erst ab diesem Zeitpunkt in Betracht.

Eine Differenzierung zwischen zielgerichteter und nicht-zielgerichteter Verarbeitung besonderer Kategorien personenbezogener Daten findet sich bereits in der Rechtsprechung[43] und ergibt sich im Zusammenhang mit der KI-Entwicklung aus dem Ziel der KI-VO, bei der „*Entwicklung einer sicheren, vertrauenswürdigen und ethisch vertretbaren KI weltweit eine Führungsrolle einzunehmen*“[44]. Der Unionsgesetzgeber war sich dabei der

[41] Vgl. DSK, Kurzpapier Nr. 17: Besondere Kategorien personenbezogener Daten, S. 2.

[42] OLG Köln, Urteil vom 23.05.2025, 15 UKl 2/25, Rn. 121; zu Suchmaschinen EuGH, Urteil vom 24.09.2019, C-136/17, Rn. 45 ff.; Schlussanträge des Generalanwalts vom 10.01.2019, C-136/17, Rn. 56 f.; zustimmend auch Schumacher/Stegemann ZD 2025, 516 (524); Wybitul/Kramer/Maturana/Ziegler, NZKart 2025, 423 (429); Schwartmann/Köhler, EuDIR 2025, 287 (292); a.A. Lindner, KIR 2025, 332, (335); von Welser, NJW 2025, 3123 (3126); Glocker, RDi 2025, 427 (431).

[43] Vgl. OLG Köln, Urteil vom 23.05.2025, 15 UKl 2/25, Rn. 120.

[44] EG 8 KI-VO.

gängigen Praxis der Verarbeitung erheblicher Datenmengen bewusst.[45] Dieser gesetzgeberische Wille würde bei einer undifferenzierten Anwendung des Art. 9 Abs. 1 DSGVO unterlaufen, weil die verarbeiteten Daten in der Praxis regelmäßig auch besondere Kategorien personenbezogener Daten enthalten. Für die Differenzierung spricht ferner, dass die KI-VO ungeachtet ihrer oben aufgeführten Zielsetzung nur zwei datenschutzrechtliche Rechtsgrundlagen enthält, insbesondere den Art. 10 Abs. 5 KI-VO. Aus der Tatsache, dass nur die sehr eingeschränkte Erlaubnisnorm des Art. 10 Abs. 5 KI-VO für die besonderen Kategorien personenbezogener Daten Eingang in die KI-VO gefunden hat, ist abzuleiten, dass der Gesetzgeber nur für den zielgerichteten Einsatz von besonderen Kategorien personenbezogener Daten einen Regelungsbedarf gesehen hat.[46] Bei der nicht-zielgerichteten Verarbeitung ergibt sich eine teleologische Reduktion des Umfangs der Verantwortlichkeit nach Art. 4 Nr. 7 DSGVO im Rahmen des KI-Trainings in Anlehnung an die Rechtsprechung des EuGH zu Suchmaschinen.[47] Entscheidendes Argument des EuGH ist eine praktische Unmöglichkeit einer ex-ante-Prüfung der Rechtmäßigkeit der Datenverarbeitung für den Suchmaschinenbetreiber. Ebendies gilt auch für die Prüfung der Sensibilität personenbezogener Daten gem. Art. 9 DSGVO.

2.5.2. Produktivbetrieb

Auch im Produktivbetrieb sind die Anforderungen des Art. 9 Abs. 2 DSGVO vollumfänglich einzuhalten. Dies gilt insbesondere für Eingaben und Ausgaben von KI-Systemen.

Memorisierte Daten

Insbesondere bei generativen KI-Modellen wie LLMs ist es nicht ausgeschlossen, dass sie besondere Kategorien personenbezogener Daten in abstrakter Form enthalten und dadurch diese reproduzieren könnten (→ 1.2. Datenschutzrechtliche Herausforderungen großer Sprachmodelle). Der Betrieb eines solchen KI-Modells, zum Beispiel auf der eigenen Serverinfrastruktur als Teil eines KI-Systems, stellt somit eine Verarbeitung dieser sensiblen personenbezogenen Daten dar. Folglich bedarf es neben einer Rechtsgrundlage auch eines Ausnahmetatbestandes im Sinne des Art. 9 Abs. 2 DSGVO.

Die oben aufgeführte Unterscheidung zwischen zielgerichteter und nicht-zielgerichteter Verarbeitung greift auch hinsichtlich gegebenenfalls im KI-Modell memorisierter besonderer Kategorien personenbezogener Daten.

---

[45] Vgl. EG 105 S. 2 KI-VO.

[46] OLG Köln, Urteil vom 23.05.2025, 15 UKl-2/25, Rn. 120.

[47] EuGH, Urteil vom 24.09.2019, C-136/17, Rn. 45.

Eine nicht-zielgerichtete Verarbeitung von besonderen Kategorien personenbezogener Daten beim Einsatz von KI-Modellen liegt vor, wenn KI-Modelle nicht dafür benutzt werden, um eventuell in ihnen enthaltene besondere Kategorien personenbezogener Daten zu entnehmen.

Indizien für eine nicht-zielgerichtete Verarbeitung von sensiblen personenbezogenen Daten beim Betrieb von KI-Modellen sind zum Beispiel:

- Es wurden Output-Filter auf Systemebene eingerichtet, die die Herausgabe von besonderen Kategorien personenbezogenen Daten aus dem KI-Modell verhindern sollen.
- Es wurden Input-Filter auf Systemebene eingerichtet, die Prompts, welche die Herausgabe von besonderen Kategorien personenbezogenen Daten zum Ziel haben, sperren.
- Es wurde ein System-Prompt eingerichtet, der dem System verbietet, sensible personenbezogene Daten herauszugeben.
- Das KI-Modell wird derart in eine RAG-Architektur eingebettet, dass keine gelernten Informationen aus dem Modell für die Beantwortung des Prompts verwendet werden, sondern die Informationen aus der Vektordatenbank.
- Es wurden verbindliche Richtlinien und Abläufe für die Benutzer des KI-Systems seitens des Verantwortlichen definiert, die eine Extraktion von besonderen Kategorien personenbezogener Daten untersagen und diese verpflichten, besondere Kategorien personenbezogener Daten in den Outputs zu löschen sowie einen solchen Vorfall an die interne IT-Stelle / den behördlichen Datenschutzbeauftragten zu melden.

2.6. Datenminimierung und Speicherbegrenzung

Für die Entwicklung von LLMs sind regelmäßig große Datenmengen notwendig. Das ist mit dem Grundsatz der Datenminimierung gemäß Art. 5 Abs. 1 Buchst. c DSGVO nicht per se unvereinbar, weil dieser Grundsatz nicht zwangsläufig eine absolute Reduzierung der verarbeiteten Datenmenge verlangt.[48] Der Verantwortliche für die Planung und Konzeption des KI-Modells muss allerdings sicherstellen und dokumentieren, dass die geplanten Datenverarbeitungen angemessen und auf das notwendige Maß beschränkt sind. Der konkrete Verarbeitungszweck (→ 2.3. Zweckbestimmung) darf also nicht in zumutbarer Weise ohne oder mit weniger personenbezogenen Daten zu erreichen sein. Die

[48] Vgl. OLG Hamburg, Urteil vom 27.02.2025, 5 U 30/24, GRUR-RS 2025, 3406, Rn. 39.

Erforderlichkeit ist nicht für jedes einzelne personenbezogene Datum zu prüfen, sondern für den Gesamtumfang der verwendeten Daten.[49]

Zunächst ist die Verwendung synthetischer Daten in Erwägung zu ziehen.[50] Eine weitere Möglichkeit besteht in der Anonymisierung personenbezogener Trainings-, Validierungs- und Testdaten, hilfsweise ist die Pseudonymisierung zu prüfen (→ 3. Übersicht und Erläuterung mitigierender Maßnahmen). Sowohl die Anonymisierung als auch die Pseudonymisierung stellen Verarbeitungsschritte dar, die einer Rechtsgrundlage bedürfen. Diese Rechtsgrundlage ergibt sich aus der Rechtsgrundlage für die zugrundeliegende KI-Entwicklung. Während die Anonymisierung aus dem Anwendungsbereich der DSGVO hinaus führt, handelt es sich bei der Pseudonymisierung um eine technische und organisatorische Maßnahme zum Schutz der Betroffenen.[51]

Der Grundsatz der Speicherbegrenzung nach Art. 5 Abs. 1 Buchst. e DSGVO verlangt, die Verarbeitung personenbezogener Daten in der Entwicklungsphase auch in zeitlicher Hinsicht auf das Erforderliche zu beschränken. Deshalb sollten grundsätzlich vorab Löschfristen implementiert werden für den Fall, dass das mit den Daten beabsichtigte Training abgeschlossen ist.[52] Dies hat keine Auswirkung auf die Dokumentationspflichten gemäß Art. 11, 53 KI-VO, die unter anderem eine Beschreibung der Trainingsdaten fordern. Diese Beschreibungen sind gegebenenfalls vor der Löschung zu erstellen.

Eine Aufbewahrung der Daten nach Abschluss der Entwicklung ist rechtfertigungsbedürftig. Als Rechtfertigung kommen gegebenenfalls die spätere Überprüfung hinsichtlich möglicher Biases (→ 2.12. Fairness), weiteres Training sowie die Gewährleistung von Betroffenenrechten (→ 2.9. Transparenz und Informationspflichten) in Betracht. Aus dem Grundsatz der Speicherbegrenzung folgt auch, dass eine nicht erforderliche Memorisierung von Trainingsdaten (→ 2.1. Personenbezug) durch geeignete Maßnahmen zu verhindern ist.

Außerdem kann bereits in der Planungsphase darauf hingewirkt werden, dass im Produktivbetrieb möglichst wenig personenbezogene Daten in Eingaben verarbeitet werden[53] (→ 2.1. Personenbezug). Entsprechend können von vornherein Eingabefilter oder

[49] Vgl. OLG Köln, Urteil vom 23.05.2025, 15 UKl 2/2025, Rn. 73 mwN.

[50] Vgl. BfDI, Kurzposition: Personenbezogene Daten bei Software-Entwicklung und -Tests, Juni 2025.

[51] Art. 25 Abs. 1, Art. 32 Abs. 1 Buchst. a DSGVO.

[52] Vgl. ICO, How should we assess security and data minimisation in AI?, abrufbar unter: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/how-should-we-assess-security-and-data-minimisation-in-ai/.

[53] Datenschutz durch Technikgestaltung und durch datenschutzfreundliche Voreinstellungen, Art. 25 DSGVO.

eine Beschränkung der Eingabe auf standardisierte Prompts im Entwicklungskonzept vorgesehen werden.

Eine Speicherung personenbezogener Prompts und Ausgaben zu Analysezwecken ist nur bei klarer Zweckbindung und angemessener Aufbewahrungsfrist zulässig. Sollte eine solche Speicherung nicht erforderlich sein, empfehlen sich standardisierte Löschroutinen, beispielsweise nach Abschluss jeder Verwendung.

Personenbezogene Ausgaben können beispielsweise durch einen entsprechenden System-Prompt, Fine-Tuning und Ausgabefilter reduziert werden.

2.7. Menschliche Aufsicht

Mit KI-Systemen lassen sich Entscheidungsprozesse ohne oder mit minimaler menschlicher Beteiligung datenbasiert und algorithmisch steuern. Das Datenschutzrecht und auch die KI-VO machen an verschiedenen Stellen allerdings Vorgaben zur Einbindung eines Menschen.

2.7.1. Automatisierte Entscheidungen einschließlich Profiling

Bei automatisierten Entscheidungen auf der Grundlage personenbezogener Daten, die gegenüber den Betroffenen rechtliche Wirkung entfalten oder sie in ähnlicher Weise beeinträchtigen, ist Vorsicht geboten. Fällt eine automatisierte Datenverarbeitung unter Art. 22 Abs. 1 DSGVO, ist sie im Falle der öffentlichen Verwaltung nur zulässig, wenn eine nationale Rechtsvorschrift den Erlass der automatisierten Entscheidung im Einzelfall erlaubt und angemessene Maßnahmen zur Wahrung der Rechte und Freiheiten sowie der berechtigten Interessen der betroffenen Person enthält.[54]

Wird allerdings ein menschlicher Entscheider in den Prozess eingebunden, der die Ergebnisse der automatisierten Datenverarbeitung eigenständig bewertet, ist das Verbot des Art. 22 Abs. 1 DSGVO nicht einschlägig. Voraussetzung dafür ist jedoch, dass das Handeln des menschlichen Entscheiders vom Ergebnis der automatisierten Verarbeitung nicht „maßgeblich geleitet“ wird.[55] Wichtig ist also, dass der menschliche Entscheider über echte Befugnisse und einschlägige Kompetenzen (→ 2.13. KI-Kompetenz) verfügt, das Ergebnis der automatisierten Datenverarbeitung gegebenenfalls zu ändern.

[54] Pilniok, DÖV 2024, 581 (591); vgl. Art. 22 Abs. 2 Buchst. b DSGVO.

[55] Vgl. EuGH, Urteil vom 07.12.2023, C-634/21, Rn. 62.

Um gewährleisten und nachweisen zu können, dass der menschliche Entscheider nicht nur pro forma eingesetzt wird, sollten klare Prozesse definiert werden, wie mit Ausgaben eines KI-Systems umzugehen ist, insbesondere welche anderen Parameter bei der menschlichen Aufsicht in welchem Ausmaß berücksichtigt werden können.

Bedienstete der öffentlichen Verwaltung sind angehalten, die Ergebnisse automatisierter Datenverarbeitung stets kritisch zu prüfen und nicht ungeprüft zu übernehmen. Öffentliche Stellen sollten entsprechende Handlungsanleitungen zur Verfügung stellen und Schulungen anbieten.

2.7.2. Menschliche Aufsicht bei Hochrisiko-KI-Systemen

Hochrisiko-KI-Systeme werden gemäß Art. 14 Abs. 1 KI-VO so konzipiert und entwickelt, dass sie während der Dauer ihrer Verwendung von natürlichen Personen wirksam beaufsichtigt werden können. Personen, denen die menschliche Aufsicht übertragen wurde, müssen unter anderem in die Lage versetzt werden,

- die einschlägigen Fähigkeiten und Grenzen des Hochrisiko-KI-Systems angemessen zu verstehen und seinen Betrieb ordnungsgemäß zu überwachen, einschließlich in Bezug auf das Erkennen und Beheben von Anomalien, Fehlfunktionen und unerwarteter Leistung;
- die Ausgabe des Hochrisiko-KI-Systems richtig zu interpretieren, in einer bestimmten Situation zu beschließen, das Hochrisiko-KI-System nicht zu verwenden oder die Ausgabe des Hochrisiko-KI-Systems außer Acht zu lassen, außer Kraft zu setzen oder rückgängig zu machen.[56]

2.8. Datenschutz-Folgenabschätzung

Die Durchführung einer Datenschutz-Folgenabschätzung (DSFA) ist für viele KI-Systeme nicht nur ein rechtliches Erfordernis gemäß Art. 35 DSGVO, sondern ein zentrales Instrument zur Risikoerkennung und -begrenzung.

2.8.1. Rechtlicher Rahmen und Anwendungsvoraussetzungen

Gemäß Art. 35 Abs. 1 DSGVO ist eine DSFA durchzuführen, wenn eine Form der Verarbeitung – insbesondere bei Verwendung neuer Technologien – voraussichtlich ein hohes Risiko für die Rechte und Freiheiten natürlicher Personen mit sich bringt. Die Artikel-29-Datenschutzgruppe nennt in ihren Leitlinien (WP 248 rev.01, S. 10 ff.) neun Kriterien, bei

[56] Vgl. Art. 14 Abs. 4 KI-VO.

deren Vorliegen eine DSFA regelmäßig erforderlich ist – darunter „automatisierte Entscheidungsfindung mit Rechtswirkung oder ähnlich bedeutsamer Wirkung“, „systematische Überwachung“, „Datenverarbeitung in großem Umfang“ und „Innovative Nutzung oder Anwendung neuer technologischer oder organisatorischer Lösungen“.

Beim aktuellen Stand der technischen Entwicklung im Bereich KI dürfte neben dem Kriterium „Datenverarbeitung in großem Umfang“ derzeit noch regelmäßig das Kriterium „Innovative Nutzung oder Anwendung neuer technologischer Lösungen“ erfüllt sein. Eine DSFA wird deshalb in der Regel erforderlich sein.

2.8.2. Besonderheiten bei KI-Systemen

Die DSFA muss die besonderen Risiken bei KI (→ 1.2. Datenschutzrechtliche Herausforderungen großer Sprachmodelle) nicht nur systematisch beschreiben, sondern auch konkrete Maßnahmen zu deren Minderung vorsehen – von technischen Schutzmaßnahmen (→ 3. Übersicht und Erläuterung mitigierender Maßnahmen) bis hin zu Governance-Strukturen (z. B. internes AI Risk Board).

2.8.3. Zusammenspiel mit der Grundrechte-Folgenabschätzung nach der KI-VO

Die KI-VO sieht in Art. 27 die Durchführung einer Grundrechte-Folgenabschätzung (GRFA) für Betreiber von einigen Hochrisiko-KI-Systemen vor. Während sich die DSFA auf Risiken im Sinne des Datenschutzrechts konzentriert, umfasst die GRFA ein umfassenderes Grundrechtsverständnis – einschließlich des Diskriminierungsverbots, des Rechts auf Bildung, auf freie Meinungsäußerung oder soziale Teilhabe.

Die DSFA und die GRFA unterscheiden sich somit hinsichtlich ihrer inhaltlichen Reichweite. Gleichzeitig weisen sie erhebliche Schnittmengen auf, insbesondere bei der Prüfung automatisierter Entscheidungen, Transparenzanforderungen und Auswirkungen auf besonders schutzbedürftige Gruppen.

Die GRFA ist nicht als Ersatz, sondern als Ergänzung zur DSFA zu verstehen. Für öffentliche Stellen bedeutet das konkret:

- Die Notwendigkeit einer DSFA nach Art. 35 DSGVO kommt für alle personenbezogenen Datenverarbeitungen in Betracht.
- Bei einigen Hochrisiko-KI-Systemen ist ergänzend die GRFA gemäß Art. 27 KI-VO durchzuführen.
- Beide Verfahren sollten gemeinsam geplant, koordiniert, dokumentiert und mit konsistenten Ergebnissen abgeschlossen werden.

Bonn, den 22.12.2025

Praxisempfehlungen für die Durchführung

1. **Rollenklärung:** Die DSFA muss vom Verantwortlichen durchgeführt werden. Vor der DSFA muss geklärt und sodann festgehalten werden, ob die Behörde Verantwortliche ist (→ 2.2. Verantwortlichkeit).
2. **Frühzeitige Integration**: Die DSFA darf nicht nachgelagert durchgeführt werden, sondern muss integraler Bestandteil der KI-Beschaffung und -Entwicklung sein. Art. 35 Abs. 1 DSGVO fordert die Durchführung der DSFA „vorab“, das heißt vor Beginn der Verarbeitung.
3. **Lebenszyklusbezug**: Die DSFA sollte Trainings-, Test-, Implementierungs- und Nutzungsphase erfassen.
4. **Kontextanalyse**: Öffentliche Stellen müssen ihren spezifischen gesetzlichen Auftrag einbeziehen. Beispielsweise sind im Sozial- oder Polizeibereich die spezifischen Risiken zu betrachten und bewerten.
5. **Verknüpfung von DSFA und GRFA**: Öffentliche Stellen sollten im Rahmen der Governance-Verantwortung klare Schnittstellen und Zuständigkeiten zwischen DSFA und GRFA definieren.
6. **Dokumentation und Beteiligung**: Die behördliche Datenschutzbeauftragte ist frühzeitig einzubeziehen (Art. 35 Abs. 2 DSGVO). Die betroffene Öffentlichkeit kann gegebenenfalls im Rahmen einer Konsultation angehört werden (Art. 35 Abs. 9 DSGVO).

2.9. Transparenz und Informationspflichten

Der Transparenzgedanke durchdringt sowohl das gesamte Datenschutzrecht als auch die KI-Verordnung als ein grundlegendes Prinzip.

2.9.1. Datenschutzrechtliche Transparenz

Die DSGVO verankert die Transparenz in Art. 5 Abs. 1 Buchst. a als einen Grundsatz für die Verarbeitung personenbezogener Daten. Die Informationspflichten (Art. 13, 14 DSGVO) und das Auskunftsrecht (Art. 15 DSGVO) konkretisieren diesen.

2.9.1.1. Informationspflichten

Art. 13 und 14 DSGVO enthalten jeweils einen Katalog an Informationen, die der Verantwortliche den betroffenen Personen bereitstellen muss. Dies betrifft insbesondere die Informationen darüber, wer und zu welchem Zweck und auf welcher Rechtsgrundlage die Daten verarbeitet. Eine allgemeine Pflicht, über den Einsatz von KI als solchen zu

informieren, lässt sich aus der DSGVO nicht herleiten.[57] Besondere Transparenzanforderungen gelten allerdings bei automatisierten Entscheidungen im Sinne des Art. 22 Abs. 1 DSGVO. Nach Art. 13 Abs. 2 Buchst. f und Art. 14 Abs. 2 Buchst. g DSGVO muss über das Bestehen einer automatisierten Entscheidungsfindung einschließlich Profiling informiert werden. Gegenstand der Informationspflicht sind „aussagekräftige Informationen über die involvierte Logik sowie die Tragweite und die angestrebten Auswirkungen einer derartigen Verarbeitung für die betroffene Person." Der Begriff „Logik" impliziert zumindest, dass die Methode der Datenverarbeitung bezogen auf die Funktionsweise des Programmablaufs im Zusammenhang mit der konkreten Anwendung zu erläutern ist.[58]

Zeitpunkt der Information

Erhebt ein Verantwortlicher die personenbezogenen Daten bei den betroffenen Personen (sog. Direkterhebung), müssen die notwendigen Informationen im Vorfeld oder spätestens bei der Erhebung mitgeteilt werden. Bei indirekter Datenerhebung – zum Beispiel bei Erhebung von Daten aus öffentlich zugänglichen Quellen – richtet sich der Zeitpunkt der Informationserteilung nach Art. 14 Abs. 3 DSGVO. In der Regel ist also spätestens innerhalb eines Monats zu informieren.

In Fällen einer indirekten Datenerhebung entfällt die Informationspflicht allerdings, wenn die Erteilung dieser Informationen sich als unmöglich erweist oder einen unverhältnismäßigen Aufwand erfordern würde.[59] Die EU-Kommission hat ein Template für General-Purpose-KI-Systeme, worunter LLMs i.d.R. fallen dürften, veröffentlicht, in dem KI-Anbieter (freiwillig) mehr Informationen bezüglich der Herkunft ihrer Trainingsdaten teilen, wie beispielsweise Größe des Trainingsdatensets, Beschreibung der Nutzung öffentlicher sowie privater Datensets, gescrapte Datensets etc.[60] Es wird empfohlen, dass Behörden diese Informationen nutzen, um sich über das LLM zu informieren und Betroffene auf die Datenquellen hinzuweisen.

[57] Bortnikov, DuD 2025, 297 (298); Vogel, Künstliche Intelligenz und Datenschutz, 2022, S. 149 ff.

[58] DSK, Orientierungshilfe „Künstliche Intelligenz und Datenschutz" vom 06.05.2024, Rn. 23.

[59] Art. 14 Abs. 5 Buchst. b DSGVO.

[60] https://digital-strategy.ec.europa.eu/en/library/explanatory-notice-and-template-public-summary-training-content-general-purpose-ai-models.

Weitergehende Transparenzinformationen könnten in Form von sog. Model Cards erfolgen, die Informationen zum jeweiligen Modell in Form von Benchmarks knapp und übersichtlich zusammenfassen.[61]

Modalitäten der Informationsübermittlung

Die Pflichtinformationen sind gemäß Art. 12 Abs. 1 S. 1 DSGVO „in präziser, transparenter, verständlicher und leicht zugänglicher Form in einer klaren und einfachen Sprache zu übermitteln". Maßstab für die Verständlichkeit sollte eine Person mit durchschnittlichen Fähigkeiten und Erkenntnismöglichkeiten sein,[62] es sei denn, die Informationen richten sich im Einzelfall an ein spezifisches Publikum.

2.9.1.2. Auskunftsrecht und Recht auf Erläuterung

Über die Informationspflichten hinaus verbürgt die DSGVO in Art. 15 Abs. 1 ein Auskunftsrecht der betroffenen Person.[63] Auch wenn der Katalog in Art. 15 Abs. 1 DSGVO ähnlich formuliert ist wie der in Art. 13 und 14 DSGVO, sollten bei der Beantwortung eines Auskunftsersuchens spezifischere Informationen im Hinblick auf die Situation des Antragstellers bereitgestellt werden. Das bedeutet zum Beispiel, dass die tatsächlichen Empfänger personenbezogener Daten offenzulegen sind.[64]

Zu beachten ist dabei, dass personenbezogene Daten in verschiedenen Phasen verarbeitet werden können (→2.1. Personenbezug). Im Hinblick auf memorisierte Daten kann es schwierig sein festzustellen, ob und gegebenenfalls welche personenbezogenen Daten ein KI-Modell über die auskunftsbegehrende Person enthält. Zunächst ist das KI-Modell selbst als mögliche Quelle von anderen Datenquellen (beispielsweise RAG-System oder Internet) zu unterscheiden und gegebenenfalls zu trennen. Der Verantwortliche muss gemäß Art. 11 Abs. 1 DSGVO keine zusätzlichen Informationen allein zur Identifizierung aufbewahren.

Liegen die Trainingsdaten vor, können sie mit den Ausgaben des KI-Modells abgeglichen werden.[65] Es ist ungenügend, lediglich die bisherigen Ausgaben zu überprüfen. In Bezug auf LLMs kann es im Einzelfall in Ermangelung der Trainingsdaten und auf Grund der

---

[61] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B. & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220-229).

[62] Bortnikov/Dukart, ZD 2024, 558 (560).

[63] Vgl. ausführlich BfDI, https://www.bfdi.bund.de/DE/Buerger/Inhalte/Allgemein/Betroffenenrechte/Betroffenenrechte_Auskunftsrecht.html.

[64] Art. 15 Abs. 1 Buchst. c DSGVO.

[65] DSK, Orientierungshilfe zu empfohlenen technischen und organisatorischen Maßnahmen bei der Entwicklung und beim Betrieb von KI-Systemen, Version 1.0, Juni 2025, S. 23.

probabilistischen Art der Ausgaben notwendig sein, das Modell auf Grundlage der vorhandenen Daten mehrfach und mit variierenden Prompts nach der auskunftsbegehrenden Person zu fragen. Anschließend sind die Ausgaben darauf zu überprüfen, ob sie personenbezogene Daten über die Person enthalten, die über die Eingabe hinausgehen. Hierbei ist sicherzustellen, dass die Eingaben nicht anderweitig (beispielsweise zum Training) verwendet werden.

Besonderheiten bei automatisierten Entscheidungen

Nach Auffassung des EuGH beinhaltet Art. 15 Abs. 1 Buchst. h DSGVO ein „echtes Recht der betroffenen Personen auf Erläuterung der Funktionsweise des Mechanismus der automatisierten Entscheidungsfindung, der diese Personen unterworfen worden sind, und des Ergebnisses, zu dem diese Entscheidung geführt hat“.[66] Die „aussagekräftigen Informationen über die involvierte Logik“ einer automatisierten Entscheidungsfindung müssen das Verfahren und die Grundsätze, die konkret zur Anwendung kommen, so beschreiben, dass die betroffene Person nachvollziehen kann, welche ihrer personenbezogenen Daten im Rahmen der in Rede stehenden automatisierten Entscheidungsfindung auf welche Art verwendet wurden.“[67]

2.9.2. Vorgaben der KI-Verordnung

Neben der DSGVO sind auch die Vorgaben der KI-Verordnung zu beachten. Art. 50 KI-VO zum Beispiel regelt Transparenzpflichten für Anbieter und Betreiber bestimmter KI-Systeme.[68]

Art. 53 Abs. 1 Buchst. d KI-VO verpflichtet Anbieter von KI-Modellen mit allgemeinem Verwendungszweck, eine hinreichend detaillierte Zusammenfassung der für das Training des KI-Modells mit allgemeinem Verwendungszweck verwendeten Inhalte nach einer vom Büro für Künstliche Intelligenz bereitgestellten Vorlage zu veröffentlichten. Diese Zusammenfassung kann im Einzelfall geeignet sein, der Informationspflicht gemäß Art. 14 DSGVO zu genügen, falls die Erteilung der Information im Einzelfall unmöglich oder unverhältnismäßig im Sinne von Art. 14 Abs. 5 Buchst. b DSGVO ist.

Nach Art. 26 Abs. 7 KI-VO müssen Betreiber, die Arbeitgeber sind, vor der Inbetriebnahme oder Verwendung eines Hochrisiko-KI-Systems am Arbeitsplatz die Arbeitnehmervertreter

[66] EuGH, Urteil vom 27.02.2025, C-203/22, Rn. 57.

[67] EuGH, Urteil vom 27.02.2025, C-203/22, Rn. 61.

[68] So müssen Betreiber eines KI-Systems, das Bild-, Ton- oder Videoinhalte erzeugt oder manipuliert, die ein Deepfake sind, gemäß Art. 50 Abs. 4 UAbs. 1 Satz 1 KI-VO offenlegen, dass die Inhalte künstlich erzeugt oder manipuliert wurden.

und die betroffenen Arbeitnehmer darüber informieren, dass sie der Verwendung des Hochrisiko-KI-Systems unterliegen werden.

Zudem sieht Art. 86 KI-VO ein Recht auf Erläuterung für die Betroffenen bestimmter Ausgaben von Hochrisiko-KI-Systemen gegenüber dem Betreiber vor. Dieses Recht ist subsidiär gegenüber anderweitigen vergleichbaren Rechten (beispielsweise Art. 15 Abs. 1 Buchst. h DSGVO).

2.9.3. Empfehlungen im Zusammenhang mit Transparenz

- Etablieren Sie ein Transparenzkonzept je KI-System.
- Setzen Sie nach Möglichkeit auf KI-Modelle, die ihr Reasoning schrittweise erklären (→ 3. Übersicht und Erläuterung mitigierender Maßnahmen anhand des KI-Lebenszyklus/ Explainable AI (XAI)).
- Führen Sie ein aktuelles Verzeichnis der Verarbeitungstätigkeiten gemäß Art. 30 DSGVO.
- Prüfen Sie im Einzelfall, ob eine automatisierte Entscheidungsfindung vorliegt.
- Nutzen Sie neben textlichen Erläuterungen geeignete Visualisierungstechniken.
- Schulen Sie die Mitarbeitenden, die mit KI-Systemen arbeiten.

2.10. Berichtigung und Löschung

Die Informationspflichten gemäß Art. 13, 14 DSGVO und der Auskunftsanspruch gemäß Art. 15 DSGVO können Grundlage für die Geltendmachung von Ansprüchen auf Berichtigung oder Löschung nach Art. 16, 17 DSGVO sein. Verantwortliche sollten im Vorhinein prüfen und Abläufe dafür festlegen, wie mit entsprechenden Anträgen Betroffener umzugehen ist. Gegenstand dieser Ansprüche können grundsätzlich alle personenbezogenen Daten in den Trainingsdaten, in Ein- und Ausgaben sowie im KI-Modell selbst sein (→ 2.1. Personenbezug).

2.10.1. Berichtigung

Unrichtige personenbezogene Daten in Ausgaben sind zu verhindern, beispielsweise durch Filter, und gegebenenfalls vor ihrer Weiterverwendung zu berichtigen.[69] Werden Ein- bzw. Ausgaben zu Analysezwecken gespeichert (→ 2.6. Datenminimierung und Speicherbegrenzung), kann in der Regel keine Berichtigung darin enthaltener falscher Daten verlangt werden. Denn die Richtigkeit und Vollständigkeit ist nach Art. 5 Abs. 1 Buchst.

[69] Vgl. DSK, Orientierungshilfe zu empfohlenen technischen und organisatorischen Maßnahmen bei der Entwicklung und beim Betrieb von KI-Systemen, Version 1.0, Juni 2025, S. 22.

d DSGVO im Hinblick auf den Verarbeitungszweck zu beurteilen, der insofern nur bei Aufbewahrung der originalen Eingaben erreicht werden kann.[70]

2.10.2. Löschung

Art. 17 DSGVO gewährt Betroffenen das Recht, vom Verantwortlichen die Löschung ihrer personenbezogenen Daten zu verlangen. Dieser Anspruch setzt einen der in Art. 17 Abs. 1 DSGVO aufgeführten Fälle voraus, was im Einzelfall zu prüfen ist. Der Verantwortliche kann sich gegebenenfalls auf eine Ausnahme von der Löschpflicht berufen. Eine solche Ausnahme kann sich aus Art. 17 Abs. 3 DSGVO ergeben. Für öffentliche Stellen kommt insbesondere Art. 17 Abs. 3 Buchst. b Alt. 2 oder 3 DSGVO in Betracht, wenn die Verarbeitung auf Art. 6 Abs. 1 Buchst. c oder e DSGVO gestützt wird. Bei Datenverarbeitung im Anwendungsbereich der JI-Richtlinie sieht § 58 Abs. 3 S. 1 Nr. 3 BDSG eine (eng auszulegende) Ausnahme für den Fall vor, dass die Löschung nur mit unverhältnismäßigem Aufwand möglich wäre.

Besteht ein Löschanspruch gemäß Art. 17 DSGVO, muss die Löschung im Ergebnis dazu führen, dass die Daten irreversibel derart unkenntlich gemacht werden, dass sie nicht bzw. nicht mit vertretbarem Aufwand weiterverarbeitet werden können.[71]

Die sicherste Lösung wäre das regelmäßige Ersetzen des entsprechenden KI-Modells durch ein neues KI-Modell, welches ohne die zu löschenden Trainingsdaten trainiert wurde. Alternativ kommen auch Methoden des Machine Unlearnings[72] in Betracht. Ebenso sind die zu löschenden Daten auch aus den Trainingsdaten für zukünftige Updates zu entfernen.

Wie in der Orientierungshilfe der DSK zu empfohlenen technischen und organisatorischen Maßnahmen bei der Entwicklung und beim Betrieb von KI-Systemen ausgeführt, sind geeignete technische und organisatorische Maßnahmen, z. B. Ein- und Ausgabefilter, als Übergangslösung einzusetzen.[73] Allerdings stellen nach Auffassung der DSK solche mitigierenden Maßnahmen, insbesondere auch Filter, an sich keine Löschung dar,[74] sodass der Gesetzgeber angehalten ist, vorzugeben, nach welchen Maßgaben sich eine Löschung von Daten aus einem KI-Modell richten soll. Nur so kann die dringend benötigte

[70] Vgl. EuGH, Urteil vom 20.12.2017, C-434/16, Rn. 53.

[71] In Abgrenzung zum „Vernichten“, vgl. Sydow/Marsch DSGVO/BDSG/Peuker, 3. Aufl. 2022, DSGVO Art. 17 Rn. 32.; s.a. Kühling/Buchner/Herbst, 4. Aufl. 2024, DSGVO Art. 17 Rn. 37.

[72] Vgl. Ziffer 3. Übersicht und Erläuterung mitigierender Maßnahmen.

[74] Vgl. DSK, Orientierungshilfe zu empfohlenen technischen und organisatorischen Maßnahmen bei der Entwicklung und beim Betrieb von KI-Systemen, Version 1.0, Juni 2025, S. 23.

[74] DSK, Orientierungshilfe zu empfohlenen technischen und organisatorischen Maßnahmen bei der Entwicklung und beim Betrieb von KI-Systemen, Version 1.0, Juni 2025, S. 23.

Rechtssicherheit bei den Anforderungen an eine Löschung von Daten in einem KI-Modell erreicht werden.

2.11. Datenrichtigkeit

Personenbezogene Ausgaben von KI-Systemen müssen dem Grundsatz der Datenrichtigkeit gemäß Art. 5 Abs. 1 Buchst. d DSGVO genügen.[75] Die Richtigkeit der Ausgaben ist unabhängig von einem Antrag Betroffener auf Berichtigung (→ 2.10 Berichtigung und Löschung) zu gewährleisten. Dabei ist die Unrichtigkeit eines Datums im Hinblick auf den Verarbeitungszweck zu bestimmen und erfordert vom Verantwortlichen die Ergreifung aller „angemessenen“ Maßnahmen.[76]

Auch insofern realisieren sich KI-spezifische Besonderheiten. Die erste Besonderheit hängt mit der Verwendung riesiger Datenmengen als Trainingsmaterial zusammen. Diese Trainingsdaten können zum einen inhaltliche Fehler enthalten und bilden zum anderen notwendigerweise nur den Informationsstand bis zu einem bestimmten Zeitpunkt ab. Sofern diese Daten nach dem Training im Modell verbleiben, sind sie gegebenenfalls unrichtig oder nicht auf dem neuesten Stand. Dies setzt sich in den Ausgaben fort, wenn diese Daten darin reproduziert werden. Die zweite Besonderheit folgt aus dem Phänomen der Halluzination. Danach kann eine Ausgabe zwar technisch im formalen Sinne fehlerfrei berechnet worden sein, ohne inhaltlich durch die Trainingsdaten gedeckt zu sein. Im Hinblick auf den Grundsatz der Datenrichtigkeit ist es unerheblich, aus welchem dieser Gründe die Ausgabe eines KI-Systems unrichtig ist.

Ob personenbezogene Ausgabedaten „unrichtig“ sind, ist unter Berücksichtigung der Erwartungshaltung des Durchschnittsnutzers im konkreten Kontext zu ermitteln.[77] Entsprechend macht es einen Unterschied, ob eine Ausgabe als Tatsache dargestellt wird oder ob sie durch einen deutlichen Hinweis als Ergebnis einer Wahrscheinlichkeitsberechnung, das vor der weiteren Verwendung zu überprüfen ist, gekennzeichnet wird. Ein solcher Hinweis an die Nutzer kann ergänzend in allgemeinen Handlungsanweisungen erfolgen und kann im Einzelfall also bereits eine „angemessene“ Abhilfemaßnahme im Sinne von Art. 5 Abs. 1 Buchst. d DSGVO darstellen.

[75] Pesch/Böhme, MMR 2023, 917 (921).

[76] Reichert/Radtke/Eske, ZD 2024, 483 (485).

[77] Vgl. BGH, Urteil vom 14.05.2013, VI ZR 269/12, GRUR 2013, 751, Rn. 16.

2.12. Fairness

Nach Art. 5 Abs. 1 Buchst. a DSGVO müssen personenbezogene Daten „nach Treu und Glauben“, also „fair“, verarbeitet werden. Gegen diesen Grundsatz der Fairness verstößt eine Verarbeitung personenbezogener Daten, die Betroffene ungerechtfertigt diskriminiert. Bei der Datenverarbeitung durch KI-Systeme kann eine solche ungerechtfertigte Diskriminierung durch KI-inhärente Biases potenziell schwerwiegende Auswirkungen auf die Betroffenen haben.

Unter einem Bias versteht man eine kognitive Verzerrung.[78] Bei KI-Systemen können Biases auf vielfältige Weise Eingang finden, insbesondere durch die Trainingsdaten. Im Folgenden wird eine beispielhafte Übersicht möglicher Biases[79], die für die Fairness-Bewertung einer KI relevant sind:

Gesellschaftlicher Bias. Datensätze beinhalten in der Regel die gesellschaftlichen Sichten ihrer Zeit, die heute nicht mehr zeitgemäß sind („Historical Bias“). Darüber hinaus hatten bestimmte Personengruppen einen erschwerten Zugang zur Gesellschaft, was sich in einer geringeren Repräsentation dieser Gruppen in den Datensätzen widerspiegelt („Representation Bias“). Gleichzeitig kann auch die Art der Datenerhebung dazu führen, dass bestimmte Gruppen unterrepräsentiert sind. Dies kann wiederum dazu führen, dass bei der Datenerhebung bestimmte Gruppen stärker berücksichtigt werden als andere („Sampling Bias“). All diese Biases können im Ergebnis zu einer ungerechtfertigten Diskriminierung bestimmter Personengruppen führen. Ein KI-Recruitment-Tool wurde beispielsweise schnell wieder vom Markt genommen, da es Lebensläufe, die das Wort „women“ enthielten, systematisch schlechter bewertete.[80] In einem anderen Fall wurde eine Gesichtserkennungs-KI zu 80 Prozent mit Gesichtern weißer Personen trainiert und wies daraufhin einen Bias gegenüber Menschen mit dunkler Hautfarbe auf.[81] KI-Diskriminierung kann dementsprechend eine Vielzahl marginalisierter Gruppen treffen.

Darüber hinaus können auch Design-Entscheidungen bei der Konzeption von KI-Modellen sowie die Auswahl der Datenquellen zu Bias führen. Für eine technische

[78] Beyer, R., & Gerlach, R. (2011). *Sprache und Denken*. Wiesbaden: VS Verlag für Sozialwissenschaften.

[79] Shrishak, K. (2024). AI: Complex Algorithms and effective Data Protection Supervision – Bias Evaluation. *Support Pool of Experts*, EDPB; (2) BSI (2025), Whitepaper zu Bias in der künstlichen Intelligenz.

[80] Dastin, J. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. Reuters, 9 October 2018.

[81] Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR.

Auseinandersetzung mit Bias, beispielsweise zu Evaluationsbias, Messungsbias etc. wird das BSI Whitepaper zu Bias empfohlen.[82]

2.12.1. Bias im Training mitigieren

Als Verantwortlicher sollte man die Trainingsdaten auf mögliche Biases hin überprüfen. Methoden wie Data Provenance[83], Data Statements[84] und Kausalanalysen[85] können helfen, herauszufinden, wie sich die Trainingsdaten zusammensetzen und wie sie miteinander in Beziehung stehen. So können Proxy-Variablen verhindert werden, die nach Bereinigung der direkten personenbezogenen Daten trotzdem noch indirekt Erkenntnisse über die Personen liefern. Im SPE-Report von Shrishak[86] und BSI Whitepaper[87] findet sich eine umfassendere Auseinandersetzung mit den technischen Möglichkeiten.

Im Training sollte möglichen historischen Biases gegengesteuert werden. Im Fine-Tuning könnte beispielsweise ein kuratiertes Datenset verwendet werden, das bestimmte Eigenschaften erfüllt, beispielsweise Geschlechter-Neutralität.[88] Im Reinforcement Learning mit menschlichem Feedback schätzen menschliche Annotatoren den Bias einer Ausgabe ein. Diese Einschätzungen fließen wiederum in das Modell ein.[89]

All diese Methoden werden aktuell weiter wissenschaftlich erforscht. Vollumfängliche Benchmarks müssen sich hier noch etablieren. Die gelisteten Methoden sollen als Hinweise dienen, wie mit Bias im KI-Modell umgegangen werden kann.

---

[82] BSI (2025), Whitepaper zu Bias in der künstlichen Intelligenz.

[83]Cheney, J., Chiticariu, L., & Tan, W. C. (2009). Provenance in Databases: Why, How, and Where. Found.

Trends Databases, 1(4), 379–474.

[84] Bender, E. M., & Friedman, B. (2018). Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. Transactions of the Association for Computational Linguistics, 6, 587–604. https://doi.org/10.1162/tacl_a_00041.

[85] Glymour, B., & Herington, J. (2019). Measuring the Biases that Matter: The Ethical and Casual Foundations for Measures of Fairness in Algorithms. FAT, 269–278.

[86] Shrishak, K. (2024). *AI: Complex Algorithms and effective Data Protection Supervision – Bias Evaluation*. Support Pool of Experts, EDPB.

[87] BSI (2025), Whitepaper zu Bias in der künstlichen Intelligenz.

[88] Solaiman, I., & Dennison, C. (2021). Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets. NeurIPS, 5861–5873.

[89] Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., & Irving, G.(2020). Fine-Tuning Language Models from Human Preferences (arXiv:1909.08593)., http://arxiv.org/abs/1909.08593.

2.12.2. Prüfung eines LLMs hinsichtlich Bias

Um ein LLM auf Bias in den Ausgaben hin zu überprüfen, kann die Reaktion des LLMs auf eine Abfrage mit unterschiedlichen personenbezogenen Daten hinsichtlich unterschiedlicher Ausgaben für einzelne Gruppen analysiert werden, beispielsweise hinsichtlich Geschlecht, Ethnie, Alter etc. Je nach Einsatzfall sollten Gruppen betrachtet werden, die potenziell eine Diskriminierung erfahren könnten.

Beispiel:

Eine Behörde möchte ein Sprachmodell einsetzen, um automatisch die eingehenden Anfragen zu bearbeiten und bei der Beantwortung zu unterstützen. Eine Vorab-Einschätzung ergibt, dass ein mögliches Risiko des LLMs darin bestehen könnte, dass die Antworten je nach Name unterschiedlich ausfallen, da dieser beispielsweise Rückschlüsse auf Geschlecht oder Ethnie erlaubt. Die Trainingsdaten könnten einen Bias enthalten, dass Anträge von Personen mit ausländisch erscheinendem Namen historisch mit mehr Ablehnung beantwortet wurden. In der Prüfung kann das LLM nun mit mehreren Anfragen konfrontiert werden, die deckungsgleich sind und sich nur im Namen unterscheiden. Sollten diese Anfragen ohne sachlich gerechtfertigten Grund zu unterschiedlichen Ausgaben führen, sollte ein Einsatz dieses Sprachmodells noch einmal überdacht werden. Alternativ/zusätzlich sollte von vornherein auf Daten, die unter Umständen eine solche Diskriminierung hervorrufen könnten, in der Eingabe verzichtet werden.

2.12.3. Umgang mit Bias in LLMs

Die Nutzenden sollten über mögliche Verzerrungen des KI-Modells informiert sein und wissen, wie sie mit möglicherweise verzerrten Ausgaben umzugehen haben. Hier könnte beispielsweise eine Dienstvorschrift die Nutzendenden anleiten, verzerrte Ausgaben nicht weiter zu verwenden und die Eingabe erneut ohne Personenbezug durchzuführen. Mithilfe von Prompt Engineering, bei dem man dem LLM mitgibt, Ausgaben ohne geschlechts- oder kulturspezifische Stereotypen zu erzeugen, kann ebenfalls versucht werden, ein „neutrales“ Ergebnis zu erzielen. Dies ändert jedoch nichts an den Modell-inhärenten Biases.

2.13. KI-Kompetenz

Die Anbieter und Betreiber von KI-Systemen müssen gemäß Art. 4 KI-VO dafür Sorge tragen, dass die mit dem Einsatz von KI befassten Personen über ausreichende KI-Kompetenz

verfügen. Die Bundesnetzagentur stellt in einem Whitepaper[90] sowie in ihrem „KI-Service Desk“[91] weitere Informationen zur Verfügung.

Die Sicherstellung ausreichender KI-Kompetenz ist auch für eine datenschutzkonforme Verarbeitung personenbezogener Daten durch KI-Systeme relevant. So kann unzureichende KI-Kompetenz zur Folge haben, dass die Datenverarbeitung durch ein KI-System als automatisierte Entscheidung im Einzelfall gemäß Art. 22 DSGVO zu bewerten ist, wenn das Handeln des menschlichen Entscheiders mangels KI-Kompetenz maßgeblich durch die Ausgabe des KI-Systems geleitet wird[92] (→ 2.7. Menschliche Aufsicht).

Darüber hinaus stellt KI-Kompetenz aufgrund der KI-spezifischen Risiken für den Datenschutz (→ 1.2. Datenschutzrechtliche Herausforderungen großer Sprachmodelle) regelmäßig eine erforderliche organisatorische Maßnahme zur Gewährleistung einer rechtmäßigen Verarbeitung dar. Diese KI-Kompetenz hat der Verantwortliche sowohl bei sich selbst (Art. 24 Abs. 1 DSGVO) als auch bei einem etwaigen Auftragsverarbeiter sicherzustellen (Art. 28 Abs. 1 DSGVO).

Öffentliche Stellen können zur Gewährleistung der KI-Kompetenz einen KI-Beauftragten benennen. Diese Aufgabe kann grundsätzlich dem oder der behördlichen Datenschutzbeauftragten übertragen werden. Hierbei müssen potenzielle Interessenskonflikte vermieden werden, beispielsweise indem dieser als KI-Beauftragter keine Entscheidungen über die Verarbeitung personenbezogener Daten mithilfe von KI-Anwendungen trifft, die er als Datenschutzbeauftragter gleichzeitig überwachen soll. Der/die Datenschutzbeauftragte ist zumindest in die Frage einzubeziehen, welche KI-spezifischen Risiken im konkreten Kontext zu berücksichtigen sind und welchen Inhalt KI-Kompetenz im Einzelfall haben sollte.

2.14. Rechenschaftspflicht, Monitoring, Datenschutzverträge

Die gesetzlichen Anforderungen an KI-Systeme müssen fortlaufend während des gesamten Betriebs eines KI-Systems eingehalten werden.[93] Technische oder rechtliche Entwicklungen

[90] BNetzA, Hinweispapier KI-Kompetenzen nach Artikel 4 KI-Verordnung, Juni 2025.

[91] https://www.bundesnetzagentur.de/DE/Fachthemen/Digitales/KI/7_Kompetenz/start.html.

[92] Vgl. EuGH, Urteil vom 07.12.2023, C-634/21, Rn. 62.

[93] Vgl. statt aller Almada, Training curriculum on AI and data protection Law & Compliance in AI Security&Data Protection, Seite 11, abrufbar unter: https://www.edpb.europa.eu/system/files/2025-06/spe-training-on-ai-and-data-protection-legal_en.pdf.

können Änderungen, zum Beispiel am System, am Modell oder an internen Regelwerken erforderlich machen.[94]

Ein fortlaufendes Monitoring ist deshalb unabdingbar, um die tatsächlichen Auswirkungen des KI-Systems im praktischen Einsatz zu bewerten, neu aufkommende Risiken festzustellen und Risikomanagementmaßnahmen anzupassen.[95]

Gerade im Bereich der leistungsfähigen LLMs ist es möglich, dass die öffentlichen Stellen des Bundes keine vollständig eigenen Entwicklungen verwenden, sondern auf die Technologie anderer Hersteller zurückgreifen. Dabei ist sowohl eine Nutzung von KI als Service (AI-as-a-Service) als auch der Betrieb eines KI-Modells eines anderen Herstellers auf eigener Hardware denkbar.

2.14.1. Rechenschaftspflicht hinsichtlich des Trainings des KI-Modells durch den Hersteller

Setzt eine Bundesbehörde ein KI-Modell oder KI-System eines anderen Herstellers als datenschutzrechtlich Verantwortliche (→ 2.2. Verantwortlichkeit) ein, das personenbezogene Daten aus den Trainingsdatensätzen memorisiert hat, muss sie im Rahmen ihrer Rechenschaftspflicht gemäß Art. 5 Abs. 2 i.V.m. Art. 5 Abs. 1 Buchst. a DSGVO darlegen, dass sie eine Rechtsgrundlage hinsichtlich der eigenen Verarbeitungen memorisierter Daten vorliegt (→ 2.4. Rechtsgrundlagen).

In dieser Konstellation ist nicht pauschal davon auszugehen, dass ein rechtswidrig entwickeltes KI-Modell nie rechtmäßig eingesetzt werden kann.[96] Der für den Einsatz Verantwortliche muss die Rechtmäßigkeit seiner Verarbeitungen gewährleisten und nachweisen können.[97] Er muss sich zudem in geeigneter Weise vergewissern, dass die Entwicklung rechtmäßig war.[98] Dabei unterliegt die verantwortliche Behörde einer abgestuften Nachforschungspflicht (s. Beispiel unten). Der Umfang der Nachforschungspflicht richtet sich nach dem konkreten Risiko für die von der Memorisierung betroffenen Personen.

---

[94] Almada, Training curriculum on AI and data protection Law & Compliance in AI Security&Data Protection, Seite 11, abrufbar unter: https://www.edpb.europa.eu/system/files/2025-06/spe-training-on-ai-and-data-protection-legal_en.pdf.

[95] DSK, Orientierungshilfe „Künstliche Intelligenz und Datenschutz“ vom 06.05.2024, Rn. 47; Barberá, AI Privacy Risks & Mitigations Large Language Models (LLMs), S. 5, abrufbar unter: https://www.edpb.europa.eu/system/files/2025-04/ai-privacy-risks-and-mitigations-in-llms.pdf.

[96] Specht-Riemenschneider/Bortnikov/Steffens, KIR 2025, 213 (214).

[97] EDSA, Opinion 28/2024, Rn. 126.

[98] EDSA, Opinion 28/2024, Rn. 129.

Beispiel:

Der Nachweispflicht kann zum Beispiel bei Implementierung von LLMs zu einfachen Verarbeitungen, wie dem Zusammenfassen von Eingabetexten, Genüge getan sein, wenn (1) der Hersteller die DSGVO-Konformität des Trainings vertraglich zusichert und/oder in der Konformitätserklärung im Sinne der KI-VO[99] darstellt und (2) keine Entscheidungen von Gerichten oder Datenschutzbehörden die DSGVO-widrigkeit des Trainings festgestellt haben.

Diese Nachforschungspflicht bezieht sich nur auf KI-Modelle, die personenbezogene Daten aus den Trainingsdaten memorisiert haben. Enthält ein KI-Modell keine personenbezogenen Daten und ist dementsprechend anonym, oder wurde das Modell erfolgreich anonymisiert, sodass es keinen Personenbezug mehr aufweist, unterfällt der Einsatz des Modells nicht den Anforderungen der DSGVO. Damit entfällt auch eine Nachforschungspflicht. Es wird dennoch empfohlen, die Einschätzung des fehlenden Personenbezugs und ihre Gründe zu dokumentieren und gegebenenfalls nachweisen zu können.

2.14.2. Ausschluss des Trainings mit personenbezogenen Eingaben und Ausgaben

Gerade in AI-as-a-Service-Konstellationen behalten sich einige KI-Anbieter vor, dass sie die Eingaben und Ausgaben der Benutzer für das weitere Training ihrer KI-Modelle benutzen dürfen.

Aufgrund der Möglichkeit der Memorisierung von Trainingsdaten und den potenziell sensiblen und personenbezogenen Eingaben durch die Mitarbeitenden von Behörden (interne Dokumente, Zeugnisse von Bewerbenden, Krankschreibungen etc.) muss in einer solchen Konstellation mit dem KI-System- oder Modell-Anbieter ausgeschlossen werden, dass dieser die Ein- und Ausgaben für das weitere Training seiner Modelle nutzen darf.[100] Ein solcher Ausschluss ist vertraglich mit dem KI-Anbieter zu vereinbaren.

[99] Art. 16 Buchst. g und Art. 47 KI-VO i.V.m. Anhang V Nr. 5 KI-VO.

[100] Ähnlich zu ChatGPT: HmbBfDI, Checkliste zum Einsatz LLM-basierter Chatbots, S. 3, Link: https://datenschutz-hamburg.de/fileadmin/user_upload/HmbBfDI/Datenschutz/Informationen/20231113_Checkliste_LLM_Chatbots_DE.pdf; BayLDA, Datenschutzkonforme Künstliche Intelligenz Checkliste mit Prüfkriterien nach DS-GVO, S. 11, https://www.lda.bayern.de/media/ki_checkliste.pdf.

3. Übersicht und Erläuterung mitigierender Maßnahmen anhand des KI-Lebenszyklus[101]

3.1. Entwicklung des LLMs

3.1.1. Planung

Auswahl von Anbietern. Bei der Auswahl der Anbieter sollte ggfs. berücksichtigt werden, dass bzw. ob diese europäische Rechenzentren nutzen.

3.1.2. Preprocessing

Gegensteuerung gegen Bias in den Trainingsdaten. Methoden wie Data Provenance oder Data Statements unterstützen dabei, die Zusammensetzung der Trainingsdaten besser zu verstehen.[102] Im Zweifelsfall könnten kuratierte Datensätze verwendet werden, die hinsichtlich Biases geprüft sind.

Umgang mit Ausreißern. Statistische Ausreißer (= Outlier) können dazu führen, dass das Sprachmodell auf diese einzelnen Datensätze überangepasst ist ("Overfitting"). Das vereinfacht es Angreifenden, anhand der Vorhersagewerte Rückschlüsse auf die Gruppen der betroffenen Personen zu treffen. Im Preprocessing sollte der Datensatz auf Ausreißer geprüft und diese entsprechend entfernt werden oder mittels Aggregationen die statistische Verteilung geglättet werden.[103]

Anonymisierung. Die Verarbeitung personenbezogener Daten mit der Folge, dass die betroffene Person nicht mehr oder nur mit unverhältnismäßig hohem Aufwand identifiziert werden kann. Verwendet werden Techniken wie Randomisierung, insbesondere Differential Privacy (als Form der Randomisierung) und Generalisierung.[104]

Pseudonymisierung. Ersetzen von Identifikationsmerkmalen durch ein Pseudonym, um Identifizierung einer Person auszuschließen oder zu erschweren. Im Gegensatz zur Anonymisierung besteht ein möglicher Personenbezug für die Besitzenden des entsprechenden Keys fort. Die Transformation der personenbezogenen Daten kann mithilfe kryptographischer Algorithmen oder Look-up Tabellen geschehen. Eine Veränderung der

[101] Es handelt sich hierbei nicht um eine erschöpfende Liste an Maßnahmen.

[102] Shrishak, K. (2024). *AI: Complex Algorithms and effective Data Protection Supervision – Bias Evaluation*. Support Pool of Experts, EDPB.

[103] CNIL (2025), IA : Analyser le statut d'un modèle d'IA au regard du RGPD. https://www.cnil.fr/fr/ia-analyser-le-statut-dun-modele-dia-au-regard-du-rgpd.

[104] Stellungnahme 5/2014 der Artikel-29-Datenschutzgruppe zu Anonymisierungstechniken.

Rohdaten mag notwendig sein, um eine effektive Pseudonymisierung zu erreichen. Tiefgreifendere Informationen finden Sie in den EDSA-Leitlinien zur Pseudonymisierung.[105]

Deduplikation. Daten, die besonders häufig in gleicher Form im Training vorkommen, werden besser memorisiert.[106] Beispielsweise werden Sequenzen, die zehn Mal in den Trainingsdaten vorkommen, im Durchschnitt 1000-fach so häufig reproduziert wie einfach vorhandene Daten. Abhilfe kann Deduplikation schaffen als Vorgang, Doppelungen zu reduzieren und jeden Datensatz nur noch einfach vorkommen zu lassen.

3.1.3. Pre-Training & Fine-Tuning

Kleinere Sprachmodelle. Forschungsergebnisse[107] legen nahe, dass größere Modelle – also Modelle, die mit sehr vielen Parametern trainiert wurden – mehr Daten memorisieren. Kleinere Modelle mit weniger Parametern können dieses Risiko verringern. Unter zusätzlichem Einbezug eines RAG-Modells dürfte die Nutzung kleinerer Sprachmodelle in einigen Anwendungsfällen eine datenschutzfreundlichere Alternative zu LLMs darstellen.[108]

Differential Privacy. Mathematisches Rahmenwerk, bei dem durch gezielt zugesetzten Noise basierend auf statistischer Wahrscheinlichkeit eine Identifikation einzelner Personen oder Datenpunkte innerhalb eines Datensatzes nicht mit vollständiger Sicherheit möglich ist, die Gesamtaussage eines Datensatzes jedoch statistisch aussagekräftig bleibt.[109]

Fine-Tuning ohne personenbezogene Daten. Forschungsergebnisse[110] legen nahe, dass Daten, die zuletzt hinzutrainiert wurden, insbesondere im Fine-Tuning, häufiger memorisiert werden. Eine mitigierende Maßnahme wäre entsprechend, ohne personenbezogene Daten im Fine-Tuning zu arbeiten.

3.1.4. Ergänzende Systeme

RAG. Bei Retrieval Augmented Generation (= RAG)-Systemen wird ein Sprachmodell in der Regel um eine Datenbank erweitert. Ein Prompt wird mit Inhalten aus der Datenbank angereichert, die die größte semantische Ähnlichkeit aufweisen. Für eine technische

---

[105] EDSA, Guidelines 01/2025 on Pseudonymisation.

[106] Kandpal, N., Wallace, E., & Raffel, C. (2022). Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning (pp. 10697-10707)*. PMLR; (2) Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., & Carlini, N. (2021). Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499.*

[107] Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., & Zhang, C. (2022). Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

[108] Zusätzlich muss das RAG-System selber auch auf Datenschutzkonformität überprüft werden. Siehe Eintrag „RAG“.

[109] Desfontaines, D. (2018). *Differential privacy in (a bit) more detail.* Ted is writing things (personal blog). https://desfontain.es/blog/differential-privacy-in-more-detail.html.

[110] Smith, Victoria, et al. "Identifying and mitigating privacy risks stemming from language models: A survey." *arXiv preprint arXiv:2310.01424 (2023)*. https://arxiv.org/html/2310.01424v2.

Auseinandersetzung wird hier die DSK-Orientierungshilfe[111] zu RAG-Systemen empfohlen. Daten in einem RAG-System können wesentlich leichter gelöscht oder verändert werden als Daten in einem LLM. Zudem ermöglicht es RAG, kleinere Sprachmodelle zu verwenden, da die relevanten Informationen durch die Datenbank zugeliefert werden und lediglich die Spracherzeugungskomponenten des Sprachmodells benötigt werden. Potenziell kann durch die gezielte Zugabe der relevanten Texte die Richtigkeit der Ausgaben des LLMs erhöht werden. Die Qualität eines RAG-Systems hängt allerdings stark von der Qualität der Referenzdokumente und des Embeddings ab. Zudem ändert das RAG-System nichts an den Eigenschaften seiner LLM-Komponente, deren datenschutzrechtliche Risiken weiterhin zu analysieren und mitigieren sind.[112]

3.1.5. Überprüfen der Schwachstellen

Das KI-System sollte auf gängige Angriffe hin untersucht werden, darunter insbesondere folgende:[113]

Red Teaming. Umfassender Prozess, in dem die Rolle eines potenziellen Angreifers eingenommen und gezielt versucht wird, Schwachstellen, Sicherheitslücken, ethische Probleme und Risiken in KI-Systemen aufzudecken.[114]

Membership Inference Attacks (MIAs). Diese Attacken zielen darauf ab, herauszufinden, ob bestimmte Daten im Trainingsdatensatz vorhanden waren. Daraus können Angreifer Rückschlüsse auf personenbezogene Daten ziehen, z.B. wenn nur Daten bestimmter Personengruppen in ein Modell eingeflossen sind. Verantwortliche eines LLMs sollten überprüfen, wie sich ihr Modell bei MIAs verhält. Technisch werden Modelle genutzt (bzw. extra erzeugt), die dem Ziel-Modell ähnlich sind bzw. sein sollen, sogenannte Shadow-Modelle, und eine dem Ziel-LLM ähnliche Ausgabe erreichen sollen. Die Ausgaben der zwei Modelle werden verglichen und ein Angriffsklassifikator erstellt. Anhand deren Ähnlichkeit lässt sich ermitteln, ob bestimmte Daten im Modell enthalten sein müssen.[115]

Prompt Injection Attacks. Laut OWASP[116] gehören Prompt-Injection Angriffe zu den größten Sicherheitsrisiken, die mit LLMs einhergehen. Hier versuchen Angreifende, mit

---

[111] DSK, Orientierungshilfe zu datenschutzrechtlichen Besonderheiten generativer KI-Systeme mit RAG-Methode, Version 1.0, Oktober 2025.

[112] DSK, Orientierungshilfe zu datenschutzrechtlichen Besonderheiten generativer KI-Systeme mit RAG-Methode, Version 1.0, Oktober 2025.

[113] Vgl. CNIL (2025). IA : Analyser le statut d'un modèle d'IA au regard du RGPD. https://www.cnil.fr/fr/ia-analyser-le-statut-dun-modele-dia-au-regard-du-rgpd.

[114] Feffer, M., Sinha, A., Deng, W. H., Lipton, Z. C., & Heidari, H. (2024). Red-teaming for generative AI: Silver bullet or security theater?. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (Vol. 7, pp. 421-437).

[115] Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)* (pp. 3-18). IEEE.

[116] https://genai.owasp.org/llmrisk/llm01-prompt-injection/.

gezielten Prompts die eigentliche Funktionsweise des LLMs auszuhebeln. Verantwortliche sollten überlegen, auf welche Art von Angriff das LLM möglicherweise vulnerabel reagiert und potenziell personenbezogene Daten ausgibt. Dies sollte getestet und entsprechend eingeschränkt werden, damit das Modell für solche Angriffe nicht anfällig ist. Solche Angriffe werden stetig weiterentwickelt.

3.2. Produktivbetrieb

System-Prompt. Anwendungsspezifische Anweisungen, die dem generativen KI-System vom Systementwickler oder Anwendungsdesigner im Kontext bereitgestellt werden. Sie werden in der Regel anderen Inputs vorangestellt und genießen möglicherweise ein höheres Vertrauen als andere Inputformen, wie zum Beispiel der Nutzer-Prompt.[117] Dem LLM kann so mitgegeben werden, dass es keine personenbezogenen Daten ausgeben soll.

Filter. Dienen der Erkennung personenbezogener Daten im Eingabeprompt (Prompt-Filter) oder in der Ausgabe des LLMs (Output-Filter). Die Filter können entweder hard-coded sein, also mit bestimmten Regeln und Mustererkennung arbeiten und explizit definierte Datenformate und Schlüsselwörter herausfiltern.[118] Oder die Filter sind ebenfalls KI-basiert und darauf trainiert, sensible oder personenbezogene Daten zu erkennen und durch anderen Text zu ersetzen. Die Wirksamkeit eines Filters muss im Einzelfall geprüft und sichergestellt werden.

Guardrails. Anforderungen an das LLM, bestimmte Informationen gar nicht erst zu generieren. Werden im Prompt mitgeliefert und wirken teils effektiver als Output-Filter. Sollten nicht im System-Prompt hinterlegt sein.[119] Bei langen Anfragen wirken sich Guardrails jedoch stark auf die Komplexität der Anfrage und entsprechende Dauer/Ressourcen zur Berechnung aus.[120]

Chain-of-Thought-Prompting/Few-Shot Prompting bezeichnet Prompting-Techniken, die typischerweise verwendet werden, um „Black-Box“-KI-Modelle verständlicher zu gestalten. In Bezug auf LLMs und Datenschutz prägt sich dies insbesondere darin aus, die Ausgaben für Nutzende verständlich zu machen. Dies kann geschehen durch „Chain-of-Thought“-Prompting, bei dem das LLM explizit angewiesen wird, die Antwort in Zwischenschritten zu formulieren. Eine Alternative ist „Few-Shot Prompting“, bei welchem dem Modell eine Beispielaufgabe gegeben wird, die in mehreren Schritten argumentativ

[117] Vassilev, A., Oprea, A., Fordyce, A., & Anderson, H. (2024). Adversarial machine learning. *Gaithersburg, MD*. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2025.pdf

[118] Wie beispielsweise das vom HBDI veröffentliche Tool zur Erkennung personenbezogener Daten, insbesondere E-Mail-Adressen, IPv4-Adressen etc., https://gitlab.opencode.de/hbdi/pbd-toolkit.

[119] https://genai.owasp.org/llmrisk/llm072025-system-prompt-leakage/; zuletzt abgerufen: 13. August 2025.

[120] Thaker, P., Maurya, Y., Hu, S., Wu, Z. S., & Smith, V. (2024). Guardrail baselines for unlearning in llms. *arXiv preprint arXiv:2403.03329*.

beantwortet wird. Das Modell soll diese Herangehensweise dann auf neue Aufgaben übertragen. Diese Methoden tragen auch dazu bei, dass das Modell lernt, deutlich komplexere Aufgaben zu lösen.[121]

Machine Unlearning bezeichnet Verfahren zur nachträglichen Anpassung eines LLM, beispielsweise bei einer Löschanfrage, wenn ein komplettes Neu-Trainieren des Modells nicht praktikabel ist. Machine Unlearning gilt noch als sehr neues Forschungsfeld. Die Methoden sind meist noch nicht effektiv bzw. ressourcenintensiv mit Einbußen zulasten der Genauigkeit. Ausprägung in *Exact Unlearning* (beispielsweise SISA – Sharded, Isolated, Sliced, and Aggregated)[122] und *Approximate Unlearning,* bei dem die Daten nicht direkt gelöscht werden, sondern das Modell entsprechend dem Einfluss der Daten angepasst wird.[123]

Zugriffs-, Rechte-, und Rollenkonzept. Beispielsweise könnten bei RAG-Systemen durch differenzierte Zugriffsrechte auf die zugeschaltete Datenbank sichergestellt werden, dass bestimmte personenbezogene Daten nur von bestimmten Personen verarbeitet/eingesehen werden können.

Löschfristen (der Eingaben). Löschfristen für Eingaben sollten implementiert werden, um diese nicht länger als notwendig zu speichern und somit anfällig für Angriffe und Datenschutzrechtsverstöße zu machen.

Schulungen des Personals / Leitfäden. Wenn Nutzende des KI-Systems wissen, wie damit umzugehen ist, - wie beispielsweise Halluzinationen des Modells oder Verzerrungen des KI-Systems erkannt werden können – können bereits einige wesentliche Risiken mitigiert werden. Hier können Schulungen und Trainingsunterlagen hilfreich sein.[124] (→ 2.13. KI-Kompetenz)

Model Cards bieten eine Übersicht relevanter Benchmarks in Bezug auf den Einsatz eines spezifischen LLMs.

---

[121] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E. & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, 24824-24837.

[122] Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., & Papernot, N. (2021). Machine Unlearning. 2021 *IEEE Symposium on Security and Privacy (SP)*, 141– 59. https://doi.org/10.1109/SP40001.2021.00019.

[123].Shrishak, K. (2024). AI: Complex Algorithms and effective Data Protection Supervision – Effective implementation of data subject's rights. *Support Pool of Experts*, EDPB.

[124] Almada, M. (2025). Training curriculum on AI and data protection – Law & Compliance in AI Security & Data Protection. *Support Pool of Experts, EDPB;* (2) Glerean, E. (2025). Training curriculum on AI and data protection – Fundamentals of Secure AI Systems with Personal Data. *Support Pool of Experts*, EDPB.