

NUTZERDATEN ALS TRAININGSDATEN VON KI

Informationen zum Datenschutz | September 2025

Einleitung

KI-Modelle, insbesondere die Sprachmodelle wie ChatGPT (OpenAI) oder Gemini (Google), sind auf Trainingsdaten angewiesen. Bereits das Erheben und Sammeln von Trainingsdaten ist eine Datenverarbeitung. Je umfangreicher und vielfältiger diese Daten sind, desto leistungsfähiger werden die Modelle. Die Anbieter greifen dabei auf unterschiedliche Mittel zurück, um an die Daten zu gelangen. Unproblematisch ist der Rückgriff auf eigene Daten oder von Dritten lizenzierte Daten. Ein besonders umstrittener Weg ist hingegen die Verwendung von Inhalten, die Nutzer von Online-Diensten selbst online teilen. Dabei kann das Datenmaterial regelmäßig auch personenbezogene Daten umfassen, so dass dann die Anforderungen der Datenschutzgrundverordnung (DSGVO) zu beachten sind. Aus diesem Grund stellen sich bei der Verwendung von Inhalten, die Nutzer von Online-Diensten veröffentlichen, vor allem Fragen nach den datenschutzrechtlichen Rahmenbedingungen, insbesondere der einschlägigen Rechtsgrundlage und Betroffenenrechten.

Die Rahmenbedingungen für das KI-Training

Bei dem Training von KI-Modellen mit Inhalten, die Nutzer von Online-Diensten veröffentlicht haben, ist auf den ersten Blick die Einhaltung der einzelnen Grundsätze des Art. 5 DSGVO problematisch.

Der Transparenzgrundsatz nach Art. 5 Abs. 1 lit. a) Var. 3 DSGVO verlangt, dass die Datenverarbeitungen in einer für die betroffene Person nachvollziehbaren Weise stattfinden müssen. Allerdings sind die daraus entstehenden Informationspflichten nach Art. 14 Abs. 5 lit. b) DSGVO entbehrlich, wenn diese unmöglich oder nur mit einem unverhältnismäßig hohem Aufwand erfüllt werden können. Das KI-Training ist von autonomen Verarbeitungsprozessen und selbstlernenden Vorgängen geprägt, sodass die Datenverarbeitung durch die KI für die Anbieter selbst oft nicht nachvollziehbar ist und diese ihren Informationspflichten nur schwer und unter unverhältnismäßig hohem Aufwand nachkommen können. Folglich kann der Transparenzgrundsatz bei der Verarbeitung personenbezogener Daten zum Zwecke des KI-Trainings eingehalten werden.

Nach dem Grundsatz der Datenminimierung nach Art. 5 Abs. 1 lit. c) DSGVO muss die Datenverarbeitung auf das für die Zwecke notwendige Maß beschränkt sein. Dass die KI-Modelle für ihre Qualität und Funktionalität auf große Datenmengen angewiesen sind, steht der Einhaltung dieses Grundsatzes nicht entgegen. Der Grundsatz der Datenminimierung beinhaltet nicht die Verpflichtung, grundsätzlich eine kleine Anzahl an Daten zu verarbeiten. Vielmehr soll sicherge-

stellt werden, dass die Datenverarbeitung zweckgebunden stattfindet. Anbieter sollen also keine Daten verarbeiten, die den Zweck – die Entwicklung und Verbesserung des KI-Modells – nicht erreichen. Hier kann bei einer Betrachtung jeder einzelnen personenbezogenen Information hervorgebracht werden, dass genau diese Information nicht für die Erreichung des Zwecks notwendig ist. Es ist jedoch eine Gesamtbetrachtung vorzunehmen, das heißt, ob die Daten gerade in ihrer Gesamtheit für das Training eines qualitativ hochwertigen KI-Modells als notwendig zu betrachten sind. Dies kann bei der Entwicklung und Verbesserung von KI-Modellen mit dem „Je mehr, desto besser“-Argument bejaht werden.

Mit dem Grundsatz der Datenrichtigkeit nach Art. 5 Abs. 1 lit. d) DSGVO soll gewährleistet werden, dass die Daten sachlich richtig und erforderlichenfalls auf dem neusten Stand sind. Unrichtige Daten sollen mithilfe angemessener Mittel unverzüglich gelöscht oder berichtigt werden. Die Qualität des Outputs eines KI-Modells ist von der Qualität des vorherigen Inputs abhängig. Die Erfüllung dieses Grundsatzes liegt ebenfalls in dem eigenen Interesse des Anbieters an einem qualitativ hochwertigen KI-Modell, sodass sich dieser grundsätzlich im Rahmen technischer Möglichkeiten um die Richtigkeit der Daten bemühen wird.

Bei Betrachtung der einzelnen Grundsätze ist darüber hinaus zu beachten, dass der DSGVO eine innovationsfördernde Auslegung zugrunde gelegt wird. Die Einhaltung der einzelnen Grundsätze aus Art. 5 DSGVO soll also auch bei innovativen Entwicklungen – wie KI-Modellen – möglich sein.

Rechtsgrundlagen für das KI-Training

Die Verarbeitung personenbezogener Daten steht nach Art. 6 Abs. 1 DSGVO allgemein unter einem präventiven Verbot mit Erlaubnisvorbehalt. Das heißt, eine Datenverarbeitung ist grundsätzlich als rechtswidrig anzusehen und kann nur durch eine Rechtsgrundlage nach Art. 6 Abs. 1 lit. a) bis f) DSGVO gerechtfertigt werden. Für die Verarbeitung von Daten, die Nutzer von Online-Diensten veröffentlicht haben, kommen vor allem die Rechtsgrundlagen der Einwilligung nach Art. 6 Abs. 1 lit. a) DSGVO sowie das berechnete Interesse gem. Art. 6 Abs. 1 lit. f) DSGVO in Betracht.

Einwilligung nach Art. 6 Abs. 1 lit. a) DSGVO

Für die Einwilligung nach Art. 6 Abs. 1 lit. a) DSGVO stellt insbesondere der fehlende Kontakt zu den Nutzern von Online-Diensten ein Hindernis dar. Der Anbieter müsste die Personen, deren veröffentlichte Daten für das KI-Training genutzt werden, vorab identifizieren

und eine Einwilligung einholen. Zudem ist die Einwilligung nach Art. 4 Nr. 11 DSGVO in „informierter Weise“ abzugeben, was aufgrund der komplexen und autonomen Verarbeitungsprozesse einer KI kaum umsetzbar ist. Sofern Personen später ihre zunächst erteilte Einwilligung widerrufen, müssten die Daten gelöscht werden. Dies ist bei bereits erfolgtem Training der KI kaum technisch umsetzbar. Zudem besteht für die Anbieter die Gefahr, gegen das Kopplungsverbot nach Art. 7 Abs. 4 DSGVO zu verstoßen. Danach wird die für eine Einwilligung erforderliche Freiwilligkeit ausgeschlossen, sofern die Erfüllung eines Vertrages von der Einwilligung abhängig gemacht wird und die konkrete Datenverarbeitung für die Erfüllung des Vertrages nicht erforderlich ist. Die Erforderlichkeit der Erhebung und Verwendung durch den Nutzer veröffentlichter Daten im Rahmen eines KI-Trainings ist für den Betrieb und die Bereitstellung eines Online-Dienstes grundsätzlich zu verneinen. Somit darf die Bereitstellung des Online-Dienstes nicht von einer Einwilligung abhängig gemacht werden – auch nicht faktisch durch die Einschränkung bestimmter Funktionen des Online-Dienstes. Schließlich würde die Einwilligung möglicherweise durch fehlende faktische Auswahlmöglichkeiten für den Nutzer und mangelhafte Informationsbereitstellung an Wert verlieren und zu einer bloßen „faktischen“ Einwilligung degradiert werden. Konsequenz der sogenannten einwilligungsbezogenen Ermüdung ist insbesondere eine Desensibilisierung der Nutzer. Folglich scheint die Einwilligung nach Art. 6 Abs. 2 lit. a) DSGVO nicht als Rechtsgrundlage für das KI-Training mit öffentlichen Informationen geeignet.

Berechtigtes Interesse nach Art. 6 Abs. 1 lit. f) DSGVO

Das berechtigte Interesse nach Art. 6 Abs. 1 lit. f) DSGVO ist aufgrund eines Abwägungsbedarfs einerseits mit einer gewissen Rechtsunsicherheit verbunden, andererseits können so die technologischen Besonderheiten des KI-Trainings berücksichtigt werden. Zunächst muss ein berechtigtes Interesse für den Anbieter an der Datenverarbeitung bestehen. Hier rücken wirtschaftliche Interessen in den Vordergrund, bspw. um das KI-Modell auf dem Markt anzubieten. Im Medizin- und Gesundheitswesen können gesellschaftliche und soziale Interessen bestehen. Im Bereich der Forschung kann von der Wissenschafts- und Informationsfreiheit ein ideelles Interesse abgeleitet werden.

Sodann muss die Datenverarbeitung zur Verwirklichung des berechtigten Interesses erforderlich sein. Die Erhebung und Sammlung der Daten ist für das Training der KI in der Entwicklungsphase unerlässlich. Sofern die Daten hingegen nicht mehr für die Entwicklung der KI, sondern für die (minimale) Qualitätssteigerung einer bereits entwickelten KI verwendet werden, erscheint die Erforderlichkeit fraglich. Insbesondere im Rahmen eines wirtschaftlichen Interesses sind die KI-Modelle jedoch dem Wettbewerb ausgesetzt und auch eine minimale Qualitätssteigerung ist von großer Bedeutung. Zudem ist die Erforderlichkeit an der Verwirklichung des Interesses mit gleich effektiven Mitteln zu messen. Solange eine Qualitätssteigerung von einem größeren Datensatz abhängt, sind andere Mittel nicht gleich effektiv und die Datenverarbeitung bleibt erforderlich. Schließlich müssen die Interessen umfassend mit den entgegenstehenden Interessen, Grundfreiheiten und Grundrechten der Nutzer abgewogen werden. Hier ist auf Seiten der Nutzer zu beachten, dass diese regelmäßig keine positive Kenntnis über die Verarbeitung ihrer veröffentlichten Daten haben und somit auch keine Betroffenenrechte geltend machen können. Zudem gibt es gewisse Reverse-Engineering-Techniken, mit denen auf der Output-Ebene der KI die Trainingsdaten rekonstruiert und offengelegt werden können. Auf Seiten der Anbieter ist zu beachten, dass diese von verschiedenen organisatorischen und technischen Sicherheitsmaßnahmen Gebrauch machen sollten. Dadurch können insbesondere Risiken, die durch Reverse-Engineering-Techniken entstehen, verhindert oder zumindest hinreichend gemindert werden. Sofern sensible Datenkategorien nach Art. 9 Abs. 1 DSGVO verarbeitet wer-

den, kann die Datenverarbeitung auf Art. 9 Abs. 2 lit. e) DSGVO gestützt werden. Wichtig bei sensiblen Datenkategorien ist, dass die Daten eigenständig von den Nutzer veröffentlicht worden sein müssen und die bloße Öffentlichkeit der Daten nicht ausreicht. Auch spricht für die Anbieter, dass Nutzer möglicherweise mit der Verwendung der von ihnen veröffentlichten Daten rechnen müssen. Ob und wann Einzelpersonen die Verwendung ihrer öffentlichen personenbezogenen Daten vernünftigerweise erwarten können, hängt maßgeblich von der Art der Beziehung zwischen der Person und dem Anbieter, der Art des Online-Dienstes, dem Kontext für die Erhebung der personenbezogenen Daten und der tatsächlichen Kenntnis der Person ab, dass diese personenbezogenen Daten online sind.

Zusammengefasst stellt das berechtigte Interesse nach Art. 6 Abs. 1 lit. f) DSGVO eine praxistaugliche Rechtsgrundlage dar.

OLG Köln zur Verarbeitung personenbezogener Daten für KI-Systeme

Auch die Gerichte haben die Verarbeitung von Nutzerdaten zum Zwecke des KI-Trainings auf Basis des berechtigten Interesses nach Art. 6 Abs. 1 lit. f) DSGVO akzeptiert. In einem Eilverfahren des OLG Kölns standen sich die Verbraucherzentrale NRW und eine Tochtergesellschaft der Meta Platforms Inc. gegenüber ([OLG Köln, Urteil v. 23.05.2025 - 15 UKI 2/25](#)). Bei der Tochtergesellschaft handelt es sich um die Meta Platforms Ireland Limited, welche die Online-Dienste Instagram und Facebook betreibt. Die Verbraucherzentrale klagte gegen Meta Platforms Ireland Limited auf Unterlassung der Datenverarbeitung wegen Verstößen gegen Art. 6 und 9 DSGVO, nachdem Meta angekündigt hatte, öffentliche Inhalte ihrer volljährigen Nutzer auf Facebook und Instagram zum Training seiner KI zu verwenden.

Das Gericht sah die beabsichtigte Datenverarbeitung als rechtmäßig an. Nach Auffassung des Gerichts könne die Verarbeitung auf Art. 6 Abs. 1 lit. f) DSGVO gestützt werden. Danach ist die Datenverarbeitung zulässig, wenn diese zur Wahrung der berechtigten Interessen des Verantwortlichen erforderlich ist und nicht die Interessen der betroffenen Person überwiegen. Das Training des KI-Modells stellt ein solches berechtigtes Interesse dar. Das Gericht stellte zudem fest, dass die Verarbeitung der Daten für den Zweck des KI-Trainings auch erforderlich war.

Betroffenenrechte

Personen, die von der Verarbeitung personenbezogener Daten betroffen sind, stehen in der DSGVO umfassende Betroffenenrechte zu. Bei der Verarbeitung von Daten, die Nutzer von Online-Diensten veröffentlicht haben, ist das Widerspruchsrecht nach Art. 21 DSGVO von besonderer Bedeutung. Durch sogenannte Opt-Out-Funktionen der Online-Dienste können Nutzer der Datenverarbeitung widersprechen. Der Widerspruch greift jedoch nur für die Inhalte, die auf dem jeweiligen Nutzerkonto veröffentlicht werden. Sofern auf anderen Nutzerkonten personenbezogene Daten über den widersprechenden Nutzer veröffentlicht werden, bleibt eine Datenverarbeitung weiterhin möglich. Eine Ausnahme besteht ausschließlich für sensible Datenkategorien nach Art. 9 Abs. 2 lit. e) DSGVO. Ein verspäteter Widerspruch macht die zuvor durchgeführte Datenverarbeitung nicht rückgängig und die Löschung ist durch die autonomen Verarbeitungsprozesse und selbstlernenden Vorgänge nach den gegenwärtigen technischen Möglichkeiten nicht umsetzbar. Infolgedessen verlangt ein effektiver Schutz durch das Widerspruchsrecht, dass alle Nutzer in einem Umfeld widersprechen und dies auch rechtzeitig tun. Eine weitere Möglichkeit, um seine veröffentlichten personenbezogenen Daten zu schützen, ist eine Beschwerde bei der zuständigen Datenschutzbehörde. Dabei kann es jedoch zu längeren Bearbeitungszeiten kommen, in denen eine Datenverarbeitung weiterhin stattfindet.

Fazit

Der Zugang zu Trainingsdaten, sowie die Qualität dieser Daten, sind für den Erfolg von KI-Modellen entscheidend. Die Verarbeitung von Inhalten, die Nutzer von Online-Diensten veröffentlicht haben und die stets personenbezogene Daten umfassen, ist für das Training von KI-Modellen daher von großer Bedeutung. Es ist zu erwarten, dass zukünftig immer mehr Anbieter auf diese Inhalte zurückgreifen werden.

Zunächst kann festgehalten werden, dass scheinbar bestehende Konfliktverhältnisse mit den Grundsätzen der DSGVO bei näherer Betrachtung der Verwendung von öffentlichen Informationen für das Training von KI-Modellen nicht entgegenstehen. Vielmehr wird ein Ausgleich zwischen dem Schutz personenbezogener Daten und der Entwicklung und Verbesserung von KI-Modellen angestrebt.

Die Datenverarbeitung kann auf das berechtigte Interesse an der Entwicklung und Verbesserung des KI-Modells nach Art. 6 Abs. 1

lit. f) DSGVO gestützt werden. Die eigene oder durch andere Nutzer getätigte Veröffentlichung von Inhalten mindert die Schutzbedürftigkeit der personenbezogenen Daten erheblich, sodass die Interessen der Nutzer an einem Schutz ihrer Daten regelmäßig nicht überwiegen und der Datenverarbeitung nicht entgegenstehen. Im Gegensatz zu der Einwilligung nach Art. 6 Abs. 1 lit. a) DSGVO ist durch das berechtigte Interesse keine Mitwirkungspflicht der Nutzer von Online-Diensten erforderlich. Vielmehr müssen Nutzer eigeninitiativ handeln, also beispielsweise von bereitgestellten Opt-Out-Funktionen der Online-Dienste Gebrauch machen und gegebenenfalls ihr Umfeld zu demselben Verhalten animieren. Infolgedessen sollte das Bewusstsein gestärkt werden, dass das eigene Verhalten und die Nutzung von Online-Diensten eine Grundlage für die umfangreiche Verarbeitung der eigenen personenbezogenen Daten begründen kann.

Marc-Levin Joppek / Mira Husemann



Kontakt:

BRANDI Rechtsanwälte
Partnerschaft mbB
Adenauerplatz 1
33602 Bielefeld

Marc-Levin Joppek

Wissenschaftlicher Mitarbeiter

T +49 521 96535 - 890
F +49 521 96535 - 113
E levin.joppek@brandi.net

Mira Husemann

Wissenschaftliche Mitarbeiterin

T +49 521 96535 - 890
F +49 521 96535 - 113
E mira.husemann@brandi.net