MASTER THESIS | MASTER'S THESIS

Titel | Title

Datenschutzrechtliche Herausforderungen beim Training von großen Sprachmodellen - Anforderungen an das berechtigte Interesse als wirksame Rechtsgrundlage und das Potential synthetischer Daten

verfasst von | submitted by

Manuel Rombach

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Laws (LL.M.)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 999 083
Universitätslehrgang lt. Studienblatt | Postgraduate programme as it appears on the student record sheet:	Informations- und Medienrecht
Betreut von | Supervisor:	Univ.-Prof. Mag. Dr. Nikolaus Forgó
Mitbetreut von | Co-Supervisor:	Tobias Haar, LL.M., MBA

Inhaltsverzeichnis

Abkürzungsverzeichnis

Abs. ..... Absatz

AES ..... Advanced Encyption Standard

APIs ..... Application Programming Interfaces

Art. ..... Artikel

BDSG ..... Bundesdatenschutzgesetz

BGB ..... Bürgerliches Gesetzbuch

bspw. ..... beispielsweise

bzw. ..... beziehungsweise

d. h. ..... das heißt

DSGVO ..... Europäische Datenschutz-Grundverordnung

EDSA ..... Europäischer Datenschutzausschuss

ErwGr. ..... Erwägungsgrund

etc. ..... et cetera

EU ..... Europäische Union

EU Kommission ..... Europäische Kommission

EuGH ..... Europäischer Gerichtshof

ff. ..... folgend

gem. ..... gemäß

GG ..... Grundgesetz für die Bundesrepublik Deutschland

ggf. ..... gegebenenfalls

ggü. ..... gegenüber

GRCh ..... Charta der Grundrechte der Europäischen Union

grds. ..... grundsätzlich

HmbBfDI ..... Hamburgische Beauftragte für Datenschutz und Informationsfreiheit

HS. ..... Halbsatz

i. d. R. ..... *in der Regel*

i. S. d. ..... im Sinne der/des

i. V. m. ..... in Verbindung mit

inkl. ..... inklusive

insb. ..... insbesondere

KI ..... künstliche Intelligenz

KI-VO ..... KI-Verordnung

lit. ..... litera

LLMs ..... Large Language Models

NER ..... Named Entity Recogniser

Nr. ..... Nummer

OECD ..... Organization for Economic Cooperation and Development

PATE ..... private aggregation of teacher ensembles

PETs ..... Privacy Enhancing Technologies

s. ..... siehe

SMPC ..... secure multi-party computation

sog. ..... sogenannte

u. a. ..... unter anderem

URL ..... Uniform Ressource Locator

usw. ..... und so weiter

Var. ..... Variante

vgl. ..... vergleiche

z. B. ..... zum Beispiel

Abbildungsverzeichnis

Vorwort

Die vorliegende Masterarbeit entstand im Rahmen des postgradualen Universitätslehrgangs „Informations- und Medienrecht“ an der Universität Wien.
Mein besonderes Interesse galt den Auswirkungen der rasant fortschreitenden Entwicklungen im Bereich der künstlichen Intelligenz auf die Rechte und Freiheiten natürlicher Personen. Die Frage, wie ein angemessener Ausgleich zwischen technologischem Fortschritt und dem Schutz von Grundrechten gelingen kann, wird künftig weiter an Bedeutung gewinnen. Ein verantwortungsvoller Umgang mit künstlicher Intelligenz und eine sachgerechte Regulierung sind essenziell, um demokratische Werte und individuelle Freiheiten zu wahren.

Mein besonderer Dank gilt meinen Betreuern Tobias Haar und Andreas Jagusch für die fachliche Unterstützung und die wertvollen Impulse.
Ebenso danke ich meiner Familie und meinen Freunden, insbesondere meinen Eltern, Angelika Schanz und Jürgen Rombach, für ihre stetige Unterstützung, Ermutigung und Geduld während des gesamten Studiums und der Anfertigung dieser Arbeit.
Darüber hinaus danke ich Vogel & Partner Rechtsanwälte sowie der V-Formation GmbH für das entgegengebrachte Vertrauen und die Flexibilität, die es mir ermöglicht haben, das Masterstudium berufsbegleitend zu absolvieren.

Zur besseren Lesbarkeit wird auf eine geschlechtsspezifische Differenzierung verzichtet. Personenbezeichnungen gelten unabhängig vom Geschlecht.

Wien, Mai 2025
Manuel Rombach

1. Einleitung

Für viele Menschen hat sich künstliche Intelligenz (nachfolgend auch „KI“) als unverzichtbarer Teil des täglichen privaten und beruflichen Lebens etabliert. KI ist eine der strategisch bedeutendsten Technologien, wenn nicht sogar die bedeutendste Technologie des 21. Jahrhunderts, mit dem Potential, Börsenbeben zu verursachen und über den künftigen wirtschaftlichen Erfolg von Nationen zu entscheiden.[1] Insbesondere generative Systeme wie die sog. großen Sprachmodelle (nachfolgend auch „Large Language Models“ oder kurz „LLMs“) ermöglichen bahnbrechende Anwendungen in Bereichen wie der automatisierten Kundenkommunikation, Sprachübersetzung, Datenanalyse oder der medizinischen Forschung und Diagnose.

Die Möglichkeiten, die der Fortschritt künstlicher Intelligenz hervorgebracht hat und bringen wird, werfen jedoch auch ethische und moralische Fragen auf. So birgt die Technologie nicht nur Chancen, sondern auch erhebliche Gefahren für demokratische Gesellschaften.[2] Folglich besteht Regulierungsbedarf, um eine Entwicklung und Nutzung von künstlicher Intelligenz im Einklang mit demokratischen Werten sicherzustellen und die Gesellschaft vor den Gefahren der künstlichen Intelligenz zu schützen. Mit der Verordnung (EU) 2024/168, besser bekannt als KI-Verordnung (nachfolgend auch „KI-VO“), die seit August 2025 in Kraft getreten ist und stufenweise ab Februar 2025 zur Anwendung kommt, ist die Europäische Union (nachfolgend „EU“) weltweit Vorreiter in der horizontalen gesetzlichen Regulierung künstlicher Intelligenz. Die

[1]Vgl. Manager Magazin (2024), Kursrutsch bei Techaktien - Das Deepseek-Beben an der Börse, Artikel v. 27.01.2025, aufrufbar unter: https://www.manager-magazin.de/unternehmen/tech/deepseek-chinesisches-ki-start-up-sorgt-fuer-boersenbeben-us-techwerte-rutschen-ab-a-0178e1d1-f330-4b9e-91a7-c3338cb936fb (letzter Aufruf: 29.01.2025); Hiltscher, Faust (2025) auf Golem.de, Deepseek zwingt die Konkurrenz, sich neu zu erfinden, Artikel v. 28.01.2025 Golem.de, aufrufbar: https://www.golem.de/news/deepseek-der-sputnik-der-ki-branche-2501-192813.html (letzter Aufruf: 28.01.2025); Süddeutsche Zeitung (2024), Eine halbe Billion für die KI, Artikel v. 22.01.2025, aufrufbar unter: https://www.sueddeutsche.de/wirtschaft/donald-trump-investitionen-ki-infrastruktur-usa-li.3187605 (letzter Aufruf: 29.01.2025)

[2] Vgl. Louban, Tahraoui, Aden, Fährmann, Krätzer, & Dittmann (2022) in Friedewald, Roßnagel, Heesen, Krämer, Lamala, Das Phänomen Deepfakes: Künstliche Intelligenz als Element politischer Einflussnahme und Perspektive einer Echtheitsprüfung, S. 265

Regulierungsinitiativen der Europäischen Kommission (nachfolgend „EU Kommission“) im digitalen Bereich (u. a. die Verordnung (EU) 2023/2854 (Data Act) oder die Verordnung (EU) 2024/2847 (Cyber Resilience Act)) werden jedoch nicht von allen Wirtschaftsakteuren positiv wahrgenommen. Auf dem 55. Weltwirtschaftsforum in Davos bewerten Vertreter der internationalen Wirtschaft die Maßnahmen des 47. U.S.-Präsidenten Donald J. Trump zur Deregulierung von KI als positiv und notwendig.[3] Vor dem Hintergrund der bestehenden rechtlichen Anforderungen und Unsicherheiten hinsichtlich der Entwicklung und dem Einsatz von KI-Systemen, die für Akteure innerhalb der EU bestehen, tragen die aktuellen internationalen Entwicklungen vermehrt dazu bei, dass die EU zunehmend an Attraktivität als Wirtschaftsstandort verliert. Es verschärft sich das Spannungsverhältnis zwischen der Angst vor Überregulierung und wirtschaftlicher Stagnation auf der einen Seite und dem steigenden Schutzbedürfnis natürlicher Personen sowie demokratischer Gesellschaften aufgrund der zunehmenden „Datafizierung“ auf der anderen Seite.

Rechtliche Unsicherheiten und Herausforderungen bestehen unter anderem hinsichtlich der Vereinbarkeit von KI-Sachverhalten mit dem europäischen Datenschutzrecht. Der Erfolg künstlicher Intelligenz basiert auf zwei zentralen Aspekten: der Verfügbarkeit von großen Datenmengen und Rechenkapazitäten. Für die Entwicklung von LLMs werden häufig Trainingsdaten im Umfang von vielen Terabytes benötigt.[4] Solche Trainingsdatensätze weisen typischerweise eine hohe Diversität auf und enthalten Daten aus verschiedensten Quellen, darunter üblicherweise auch Daten, die sich auf identifizierbare natürliche Personen beziehen (z. B. Namen und E-Mail-Adressen).[5] Die besondere Funktionsweise und der Datenhunger von KI-Technologien wie LLMs führen

[3] Vgl. Chowdhury, Mattackal (2025) in Reuters, European executives join Trump’s call for action on deregulation, Artikel v. 24.01.2025, aufrufbar unter (https://www.reuters.com/world/davos-european-executives-join-trumps-call-action-deregulation-2025-01-24/ (letzter Aufruf: 29.01.2025); Wendiggensen (2025) in FAZ, Trumps Umkremplung der Digitalpolitik hat begonnen, Artikel v. 22.01.2025, aufrufbar unter: https://www.faz.net/pro/digitalwirtschaft/kuenstliche-intelligenz/donald-trump-will-ki-regeln-abschaffen-was-das-fuer-europa-heisst-110245127.html (letzter Aufruf: 29.01.2025

[4] Vgl. Yan et al. (2024), On Protecting the Data Privacy of Large Language Models (LLMs): A Survey, S. 4

[5] Vgl. Longpre et al. (2023), A pretrainer’s guide to training data: Measuring the effects of data age, domain coverage, quality,& toxicity, S. 5

naturgemäß zu Spannungen mit datenschutzrechtlichen Grundsätzen der Verordnung (EU) 2016/679 (Datenschutz-Grundverordnung, nachfolgend „DSGVO“). Auch wenn die EU Kommission die Bedeutung künstlicher Intelligenz bereits vor Inkrafttreten der DSGVO erkannt und antizipiert hat, scheint die DSGVO trotz ihrer angepriesenen Technologieneutralität mit den Besonderheiten von KI-Sachverhalten an Ihre Grenzen zu stoßen.[6] Forderungen nach einer Revision der DSGVO aus der Praxis, Wissenschaft sowie auch aus der Politik werden immer lauter.[7] Insbesondere zur datenschutzrechtlichen Bewertung von Datenverarbeitungen im Rahmen der Entwicklung der immer relevanter werdenden LLMs bietet die juristische Literatur und Judikatur bisher nur wenig Erkenntnisse, aus denen sich realistische Handlungsempfehlungen für die Praxis ableiten ließen. Praxisrelevante Fragen, die sich in diesem Kontext aktuell stellen sind insbesondere die Folgenden:

- Ist ein Personenbezug von Trainingsdaten und somit die Anwendbarkeit der DSGVO beim Training von LLMs pauschal anzunehmen?
- Wie können verantwortliche Stellen i. S. d. Art. 4 Nr. 7 DSGVO ihre datenschutzrechtlichen Pflichten, insbesondere Informationspflichten sowie die Gewährleistung von Betroffenenrechten und angemessenen Sicherheitsmaßnahmen, wirksam erfüllen?
- Wie können die Anforderungen des Art. 6 Abs. 1 lit. f DSGVO als datenschutzrechtliche Rechtsgrundlage rechtssicher erfüllt werden?

Hier soll die vorliegende Arbeit ansetzen und einen Beitrag zur juristischen Aufarbeitung der datenschutzrechtlichen Aspekte beim Training von LLMs leisten. Insbesondere sollen

[6] Vgl. Europäische Kommission (2018), Künstliche Intelligenz für Europa, COM (2018) 237 final, aufrufbar unter: https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:52018DC0237 (letzter Aufruf: 29.01.2025)

[7] Vgl. u.a. Workshop "Datenschutz neu gedacht – Herausforderungen durch KI und globalen Wettbewerb", veranstaltet von der Universität Wien, 13.–14. Dezember 2024; Wendehorst (2024), Tentative Academic Discussion Draft of a Regulation on the protection of personal data in the context of artificial intelligence and the data economy and amending Regulation (EU) 2024/1689 (‘AI Data Protection Regulation’), Version 1.1, aufrufbar unter: https://zivilrecht.univie.ac.at/fileadmin/user_upload/i_zivilrecht/Wendehorst/Workshop_Datenschutz/Draft_AI_Data_Protection_Regulation_WENDEHORST_24-12-20.pdf (letzter Aufruf: 31.01.2025)

praxistaugliche Lösungsansätze für bestehende datenschutzrechtliche Herausforderungen und Unsicherheiten dargestellt werden. Es wird die These aufgestellt, dass eine innovationshemmende Wirkung der DSGVO vermieden und ein angemessener Ausgleich zwischen den Grundrechtspositionen der Datensubjekte und der verarbeitenden Stellen erreicht werden kann, wenn die DSGVO kontextbasiert unter Berücksichtigung der tatsächlichen Risiken für die Rechte und Freiheiten der Datensubjekte im jeweiligen Einzelfall ausgelegt wird. Im Lichte der immer rasanteren digitalen Entwicklungen ist eine flexible, risikobasierte Anwendung der DSGVO erforderlich und auch mit dem Telos der Verordnung zu vereinbaren.

1.1 Problemstellung

Der sachliche Anwendungsbereich der DSGVO ist insbesondere dann eröffnet, wenn personenbezogene Daten ganz oder teilweise automatisiert verarbeitet werden. Nach der Legaldefinition des Begriffs „personenbezogene Daten“ des Art. 4 Nr. 1 DSGVO, fallen alle Informationen unter den Begriff, die sich auf eine identifizierte oder identifizierbare natürliche Person (sog. „betroffene Person“) beziehen. Als identifizierbar wird eine natürliche Person dabei angesehen, wenn sie direkt oder indirekt, insbesondere mittels Zuordnung zu einer Kennung wie bspw. einem Namen, einer Kennnummer oder zu einem oder mehreren besonderen Merkmalen, die Ausdruck der physischen, physiologischen, genetischen, psychischen, wirtschaftlichen, kulturellen oder sozialen Identität dieser natürlichen Person sind, identifiziert werden kann. Daten, die für das Training von LLMs verwendet werden (s. hierzu Kapitel 2.2 ff.) enthalten in der Regel auch Daten, die grundsätzlich unter den Begriff der personenbezogenen Daten fallen. In der datenschutzrechtlichen Literatur wird bisher häufig pauschal angenommen, dass das Training von LLMs in den sachlichen Anwendungsbereich der DSGVO fällt.[8] Nach der DSGVO gelten u. a. die Grundsätze der Rechtmäßigkeit, der Zweckbindung, der Datenminimierung sowie der Transparenz. Personenbezogene Daten dürfen demnach nur

[8] Vgl. Bartels (2024), A Balancing Act: Data Protection Compliance of Artificial Intelligence, GRUR Int. 2024, 526, 528; Hamburgischer Beauftragter für Datenschutz und Informationsfreiheit (2024), Diskussionspapier: Large Language Models und personenbezogene Daten v. 15. Juli 2024; Rosenthal (2024), Sprachmodelle mit und ohne personenbezogene Daten, Blogbeitrag vom 17.07.2024, aufrufbar unter: https://www.vischer.com/know-how/blog/teil-19-sprachmodelle-mit-und-ohne-personenbezogene-daten/ (letzter Aufruf: 24.11.2024); Moos (2024), Personenbezug von Large Language Models, Rn. 2, CR 7/2024, 442

auf Basis einer konkreten Rechtsgrundlage und nur in dem Umfang verarbeitet werden, der zur Erreichung der vorab zu definierenden Verarbeitungszwecke erforderlich ist. Gleichzeitig müssen den betroffenen Personen transparente Informationen zur Datenverarbeitung bereitgestellt werden. Diese datenschutzrechtlichen Anforderungen sind nur schwer mit den technischen Anforderungen bei der Entwicklung von LLMs (s. Kapitel 2.3 ff.) zu vereinbaren. Dieses Spannungsverhältnis stellt die Entwickler von LLMs vor erhebliche Herausforderungen. Die Stellungnahme 08/2024 des Europäischen Datenschutzausschusses (nachfolgend auch „EDSA“) zu bestimmten datenschutzrechtlichen Aspekten im Zusammenhang mit der Verarbeitung personenbezogener Daten im Rahmen von KI-Modellen vom 17. Dezember 2024 trägt kaum zu mehr Rechtssicherheit bei. Zwar erkennt der EDSA den Art. 6 Abs. 1 lit. f DSGVO als potenziell wirksame Rechtsgrundlage für Datenverarbeitungen beim Training von LLMs an, verweist jedoch auch auf die hohen Anforderungen, die hierfür vorliegen müssen, ohne praxistaugliche Lösungsansätze aufzuzeigen.[9] Entwickler von LLMs stehen weiterhin insbesondere vor der offenen Frage, wie sie ihre datenschutzrechtlichen Informationspflichten erfüllen oder Betroffenenrechte umsetzen können. Aus Sicht der LLM-Entwickler scheinen die hohen datenschutzrechtlichen Anforderungen oft nicht in einem angemessenen Verhältnis zu den Risiken für die Datensubjekte zu stehen. In Frage steht dabei insbesondere, wie sich der Umstand auswirkt, dass es in der Regel nicht zu einer Identifizierung der Datensubjekte durch die LLM-Entwickler kommt. Folglich besteht dringender Bedarf an fundierten Orientierungshilfen und praxistauglichen Lösungsansätzen.

1.2 Zielsetzung

Die vorliegende Arbeit adressiert das Problem der bestehenden Rechtsunsicherheit hinsichtlich der Anwendung der DSGVO auf Datenverarbeitungen beim Training von LLMs. Übergeordnetes Ziel der Arbeit ist eine fundierte Bewertung vermeintlich datenschutzrechtlicher Problemstellungen beim LLM-Training, um auf dieser Basis praxistaugliche Lösungsansätze für bestehende datenschutzrechtliche Herausforderungen und Unsicherheiten aufzeigen zu können. Hierzu wird zunächst ein grober Überblick über die technischen Grundlagen von LLMs geschaffen, um die Problemstellungen bewerten

[9] Vgl. Europäischer Datenschutzausschuss (2024), Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, Rn. 59-108

zu können, die sich bei der Anwendung der DSGVO ergeben. Das Spannungsverhältnis zwischen den technischen Anforderungen für das Training von LLMs und der DSGVO soll differenziert dargestellt werden. Ferner soll eine kritische Reflexion über die Anwendbarkeit der DSGVO auf das LLM-Training initiiert werden. Hierfür erfolgt eine detaillierte Prüfung der Anwendbarkeit der DSGVO unter Berücksichtigung des sog. relativen Begriffs des Personenbezugs. Weiter sollen mögliche Schutzmaßnahmen i. S. d. Art. 32 DSGVO dargestellt werden, die ggf. zu einer weiteren Relativierung der zuvor dargestellten datenschutzrechtlichen Problemstellungen führen können. Insbesondere das Potential synthetischer Daten soll untersucht werden. Schließlich soll eine Hilfestellung zur Durchführung und Dokumentation einer Interessenabwägung nach Art. 6 Abs. 1 lit. f DSGVO gegeben werden. Dazu werden die für eine Interessenabwägung relevanten Aspekte des LLM-Trainings und deren Auswirkung auf die Gewichtung der miteinander abzuwägenden Grundrechtspositionen zusammenfassend dargestellt.

1.3 Methodik

Um die technischen Grundlagen und die datenschutzrechtlichen Anforderungen zu ermitteln und ausführlich darzustellen, wird eine umfassende Literaturrecherche durchgeführt. Die Untersuchung basiert auf einem interdisziplinären Ansatz, der sowohl juristische als auch technische Aspekte berücksichtigt. Zunächst wird die technische Funktionsweise von LLMs, einschließlich ihrer Abhängigkeit von qualitativ hochwertigen Trainingsdaten in großem Umfang, analysiert. Darauf aufbauend erfolgt eine rechtliche Bewertung der relevanten datenschutzrechtlichen Anforderungen der DSGVO. Ein Schwerpunkt liegt dabei auf der Anwendung und Auslegung zentraler Prinzipien wie Zweckbindung, Datenminimierung und Rechtmäßigkeit. Ergänzend werden bestehende technische Lösungsansätze wie Differential Privacy und Daten-Synthetisierung untersucht und auf ihre rechtliche Tragfähigkeit hin bewertet. Die Arbeit stützt sich auf eine Auswertung von juristischer und computerwissenschaftlicher Fachliteratur, Rechtsprechung, Stellungnahmen und Orientierungshilfen europäischer Datenschutzaufsichtsbehörden sowie technischen Studien.

1.4 Aufbau der Arbeit

Die vorliegende Arbeit gliedert sich in acht Kapitel. Im zweiten Kapitel wird der Begriff „künstliche Intelligenz“ definiert und die technischen Grundlagen des maschinellen Lernens erläutert. Dabei sollen vor allem die Notwendigkeit von umfangreichen

Trainingsdatensätzen und das Phänomen des sog. Memorisierens von Trainingsdaten als Risikofaktor für die Rechte und Freiheiten der Datensubjekte beleuchtet werden. Das dritte Kapitel verschafft einen Überblick über das Spannungsverhältnis zwischen den technischen Anforderungen beim LLM-Training und datenschutzrechtlichen Anforderungen der DSGVO. Dabei soll ein Verständnis über die Problematik und die daraus resultierende Rechtsunsicherheit vermittelt werden, während gleichzeitig bereits potenzielle Lösungsansätze aufgezeigt werden. Im vierten Kapitel wird die Anwendbarkeit der DSGVO auf das LLM-Training genauer untersucht. Der thematische Schwerpunkt liegt dabei auf der Auslegung des Begriffs der personenbezogenen Daten nach der sog. relativen Theorie sowie dem risikobasierten Charakter des Art. 9 DSGVO und seiner entsprechenden Auslegung. Im fünften Kapitel werden ausgewählte technische Schutzmaßnahmen für das LLM-Training vorgestellt. Im sechsten Kapitel erfolgt eine Untersuchung des Potentials und der Praxistauglichkeit der Daten-Synthetisierung als potenzielle Anonymisierungsmethode. Besonderes Augenmerk wird dabei auf die Beurteilung der Anonymität, die datenschutzrechtliche Rechtsgrundlage für den Prozess der Daten-Synthetisierung sowie potenzielle Auswirkungen einer Nutzung synthetischer Daten auf die Grundrechte und -freiheiten natürlicher Personen gelegt. Das siebte Kapitel behandelt die Interessenabwägung nach Art. 6 Abs. 1 lit. f DSGVO als Rechtsgrundlage für das LLM-Training und konzentriert sich dabei auf die Gewichtung der widerstreitenden Grundrechtspositionen der betroffenen Personen einerseits und der verantwortlichen Stelle andererseits. In Kapitel acht werden die Erkenntnisse der Arbeit zusammenfassend dargestellt.

2 Begriff des KI-Systems und technische Grundlagen

Spätestens seitdem OpenAI den Chatbot „ChatGPT“ im November 2022 kostenlos für die Öffentlichkeit zugänglich gemacht hat, hat sich der Begriff der künstlichen Intelligenz zu einem populären „buzzword“ entwickelt. Aber was genau ist eigentlich ein KI-System und wie unterscheidet sich ein solches von herkömmlichen Softwareprogrammen, wie bspw. Microsoft Excel? Seitdem die italienische Datenschutzbehörde (Garante per la protezione dei dati personali) im März 2023, nur wenige Monate nach der Veröffentlichung des Chatbots ChatGPT, diesen in Italien vorläufig gesperrt hatte, dürfte zumindest allgemein bekannt sein, dass bestimmte KI-Systeme mit umfangreichen

Datensätze gespeist werden.[10] Jedoch den Wenigsten dürfte bekannt sein, was genau mit diesen Datensätzen tatsächlich passiert. Das übergeordnete Ziel dieses Kapitels ist es, ein grundlegendes Verständnis davon zu vermitteln, was mit den Trainingsdatensätzen beim Training von LLMs geschieht, weshalb die Funktionsweise dieser Technologie sehr umfangreiche Trainingsdatensätze erfordert und welche Rolle die Qualität der Daten spielt.

Für eine nachvollziehbare Darstellung und Untersuchung der datenschutzrechtlichen Problemstellungen sind die Klärung des Begriffs der künstlichen Intelligenz sowie ein zumindest grober Überblick über die technischen Grundlagen zwingend notwendig. Vorliegend beschränken wir uns auf sog. generative KI-Systeme, die auf großen Sprachmodellen (sog. Large Language Models, kurz „LLMs“) und sog. neuronalen Netzen basieren (wie z. B. OpenAI’s ChatGPT). Nachfolgend soll ein grundlegendes Verständnis der wesentlichen Prinzipien und der Funktionsweise der genannten Technologie vermittelt werden, ohne Anspruch auf Vollständigkeit. Für detaillierte und vollständige Ausführungen wird auf die einschlägige computerwissenschaftliche Fachliteratur verwiesen.

2.1 Wesentliche Merkmale und Komponenten von KI-Systemen

Künstliche Intelligenz ist ein sehr heterogenes Konzept, dessen Definition stark vom jeweiligen Kontext abhängt, sei es wissenschaftlich, technologisch oder anwendungsbezogen. Aufgrund der Vielfalt der Teilbereiche innerhalb der KI-Forschung sowie aus der sich weiterentwickelnden Natur der Technologie selbst ist es schwierig, sich auf eine allgemeingültige Definition des Begriffs zu einigen.[11] Als Ursprung des Begriffs „künstliche Intelligenz“ wird häufig der US-amerikanische Informatiker John McCarthy und die von ihm im Jahr 1956 ausgerichtete Dartmouth Summer Research

[10] Vgl. Süddeutsche Zeitung (31. März 2023), ChatGPT in Italien gesperrt – Was hinter der Sperre der Software steckt, https://www.sueddeutsche.de/wirtschaft/chatgpt-italien-sperre-software-1.5779751 (letzter Aufruf am 31.08.2024)

[11] Vgl. König, P. D., Krafft, T. D., Schulz, W., & Zweig, K. A. (2021), Essence of AI: What is AI? in K. Frankish & W. M. Ramsey (Hrsg.), The Cambridge Handbook of Artificial Intelligence, S. 18–34

Project-Konferenz genannt.[12] McCarthy eröffnete die These, dass die grundlegenden Merkmale der menschlichen Intelligenz so konkret beschrieben werden können, dass man eine Maschine zur Simulation der menschlichen Intelligenz konstruieren könne.[13] Ziel der KI-Forschung ist es seit jeher Programme zu entwickeln, die in der Lage sind, Aufgaben zu bewältigen, die bisher nur von Menschen gelöst werden konnten.[14]

Die KI-VO - die weltweit erste gesetzliche Regulierung künstlicher Intelligenz - definiert KI-Systeme in Art. 3 Nr. 1 KI-VO. Demnach ist ein KI-System ein „maschinengestütztes System, das für einen in unterschiedlichem Grade autonomen Betrieb ausgelegt ist, das nach seiner Betriebsaufnahme anpassungsfähig sein kann und das aus den erhaltenen Eingaben für explizite oder implizite Ziele ableitet, wie Ausgaben wie etwa Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen erstellt werden, die physische oder virtuelle Umgebungen beeinflussen können“. Diese Terminologie orientiert sich maßgeblich an der Definition der OECD-Sachverständigengruppe für KI.[15]

Ein KI-System nach klassischem Verständnis (sog. KI-Agent) generiert aus einem Input einen Output, in dem es Informationen verarbeitet. Bei dem generierten Output kann es sich insbesondere um Bilder, Videos, Musik, Texte oder gesprochene Sprache handeln (sog. generative KI-Systeme). Ein KI-Agent besteht typischerweise aus drei Hauptelementen (Sensoren, operative Logik und Aktoren) und kann stark vereinfacht mit folgendem Schaubild dargestellt werden:[16]

[12] Vgl. OECD (2020), Künstliche Intelligenz in der Gesellschaft, S. 20 (letzter Aufruf am 31.08.2024); Vogel (2022), Künstliche Intelligenz und Datenschutz, S. 29; Ertel (2021), Grundkurs Künstliche Intelligenz: Eine praxisorientierte Einführung, S. 1

[13] Vgl. McCarthy et al. (2006), A proposal for the Dartmouth Summer Research Project on Artificial Intelligence, S. 2

[14] Vgl. Vogel (2022), Künstliche Intelligenz und Datenschutz, S. 33

[15] Vgl. OECD (2020), Künstliche Intelligenz in der Gesellschaft, S. 15 (letzter Aufruf am 31.08.2024)

[16] Vgl. OECD (2020), Künstliche Intelligenz in der Gesellschaft, S. 23 (letzter Aufruf am 31.08.2024); Vgl. König, P. D., Krafft, T. D., Schulz, W., & Zweig, K. A. (2021). Essence of AI: What is AI? in K. Frankish & W. M. Ramsey (Hrsg.), The Cambridge Handbook of Artificial Intelligence, S. 25–26

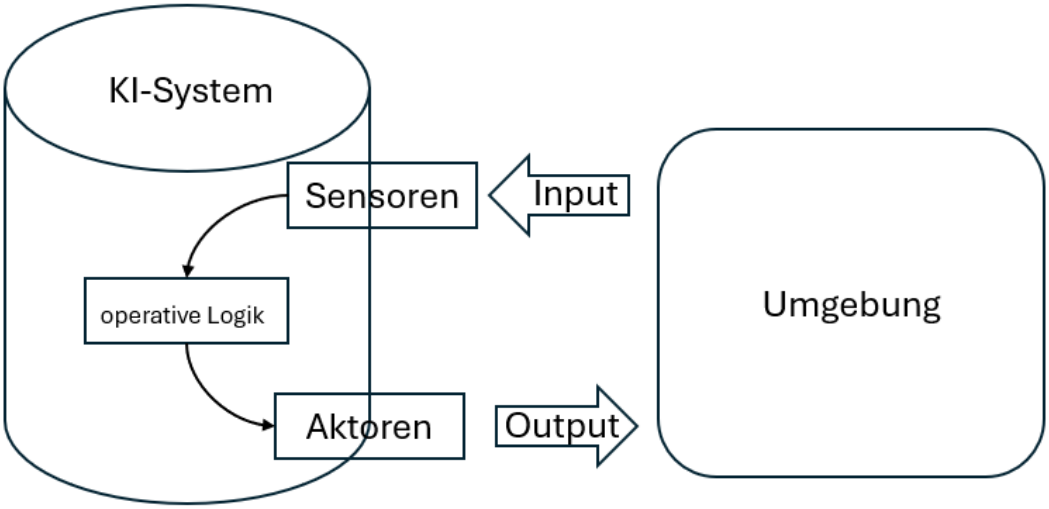

Abbildung 1: stark vereinfachte schematische Darstellung eines KI-Agenten

Die Sensoren erfassen Daten (Input) aus einer realen oder virtuellen Umgebung (bspw. über Sensoren oder manuelle Eingaben), auf deren Basis die operative Logik des KI-Systems einen entsprechenden Output generiert, welcher über die Aktoren an die Umgebung ausgegeben wird.[17]

Das vorangestellte Prinzip ist für die verschiedensten Zwecke und Anwendungsbereiche denkbar. Ein Beispiel sind die KI-Systeme AlphaGo und dessen Nachfolger AlphaGo Zero der Alphabet-Tochtergesellschaft DeepMind, welche das äußerst komplexe Strategiespiel Go besser beherrschen als die weltbesten Go-Spieler. Das Brettspiel Go ist im Rahmen dieses Beispiels eine virtuelle, beobachtbare Umgebung und der Input, mit welchem das KI-System trainiert wurde, sind die Go-Spielregeln sowie Spielzüge aus historischen Spielen, die das System bei Spielen gegen sich selbst optimieren konnte. Auf Basis dieses Inputs kann AlphaGo Zero je nach Spielstand passende Spielzüge vorschlagen und an die Umgebung (hier das jeweilige Go-Spiel) ausgeben.[18] Weitere

[17] Vgl. OECD (2020), Künstliche Intelligenz in der Gesellschaft, S. 23 (letzter Aufruf am 31.08.2024); Vgl. König, P. D., Krafft, T. D., Schulz, W., & Zweig, K. A. (2021). Essence of AI: What is AI? in K. Frankish & W. M. Ramsey (Hrsg.), The Cambridge Handbook of Artificial Intelligence, S. 25–26

[18] Vgl. OECD (2020), Künstliche Intelligenz in der Gesellschaft, S. 25 (letzter Aufruf am 31.08.2024)

Beispiele sind Objekterkennungssysteme in autonomen Fahrzeugen oder textbasierte Chatbots, die mit Ihrer Umgebung interagieren.

Das zentrale Kriterium, anhand welchem sich KI-Systeme von herkömmlichen Softwaresystemen unterscheiden, ist vor allem ein gewisser Grad an Autonomie. Bereits Alan Turing skizzierte in seinem revolutionären Artikel „Computing Machinery and Intelligence“ aus dem Jahr 1950 vorausschauend, wie Maschinen aus vergangenen Ereignissen autonom lernen könnten.[19] Seitdem haben sich die technischen Grundlagen der künstlichen Intelligenz grundlegend verändert. Ausschlaggebend waren vor allem wegweisende Fortschritte beim sog. maschinellen Lernen durch die Zunahme des Umfangs von Rechenkapazitäten und verfügbaren Daten in den 1990er Jahren (der Begriff des maschinellen Lernens wird in den nachfolgenden Absätzen genauer erläutert).[20] Insbesondere Techniken des maschinellen Lernens bieten die Grundlage für wesentliche Charakteristika von heutigen KI-Systemen. In Erwägungsgrund 12 der KI-VO wird als wesentliches Merkmal künstlicher Intelligenz u. a. die Fähigkeit genannt, „Ausgaben wie Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen […] aus Eingaben oder Daten abzuleiten“[21]. Steen beschreibt diesen Ableitungsprozess i. S. d. KI-VO simplifiziert als einen „ein oder mehrschrittigen Vorgang, bei dem ein konkreter Sachverhalt (die Eingabe [z. B. eine Texteingabe in ChatGPT]) durch abstrakte Schlussfolgerungsprinzipien (das Modell [z. B. sog. Large Language Models]) in eine konkrete Schlussfolgerung (die Ausgabe [z. B. Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen]) überführt wird.“[22]

KI-Systeme setzen sich typischerweise aus verschiedenen technischen Komponenten zusammen. So kann man KI-Systeme wie ChatGPT oder Copilot vor allem auf die folgenden drei Komponenten herunterbrechen: Rechenressourcen, Benutzeroberfläche bzw. sog. Application Programming Interfaces (APIs) und KI-Modell (in Abbildung 1 als „operative Logik“ dargestellt). ChatGPT und Copilot, beides sog. generative Sprachmodelle, basieren auf dem Großen Sprachmodell (Large Language Model) GPT

[19] Vgl. A.M. Turing, 1950, Computing Machinery and Intelligence, Mind 49: 433-460, 450 ff.

[20] Vgl. OECD (2020), Künstliche Intelligenz in der Gesellschaft, S. 20 (letzter Aufruf am 31.08.2024)

[21] ErwGr. 12 Verordnung (EU) 2024/1689

[22] Steen (2024), Ableitung als wesentliche Fähigkeit von KI-Systemen nach der KI-VO, KIR 1/2024, S. 8

(Generative Pre-trained Transformer), welches das Kernstück des KI-Systems darstellt. Solche sog. generativen KI-Modelle nutzen maschinelles Lernen und große Datenmengen, um Muster in Sprache, Text oder Bildern zu erkennen und darauf basierend sinnvolle Ausgaben in Form von Bildern, Videos, Musik, gesprochener Sprache oder eben Text zu generieren.[23] Ein KI-System kann demnach als Kombination von Hard- und Softwareelementen verstanden werden, welche mindestens eine lernende oder trainierte Komponente enthält, die in der Lage ist, ihr Verhalten auf Grundlage von Daten und daraus abgeleiteten Mustern anzupassen.[24]

2.2 Large Language Models und neuronale Netze

Der Fokus der vorliegenden wissenschaftliche Arbeit liegt vor allem auf der Komponente KI-Modell und zwar konkret auf den bereits erwähnten großen Sprachmodellen oder LLMs. Bei LLMs handelt es sich um statistische, textbasierte KI-Modelle, die anhand von umfangreichen Datensätzen dazu trainiert werden, Texte zu „verstehen“, zu verarbeiten und zu generieren. LLMs greifen für den oben beschriebenen Ableitungsprozess auf Techniken des maschinellen Lernens zurück, um aus umfangreichen Datensätzen zu lernen.

2.2.1 Maschinelles Lernen

Beim maschinellen Lernen werden durch Lernalgorithmen komplexe Modelle auf Basis von Beispieldaten entwickelt, die „Wissen“ aus „Erfahrung“ generieren, das auf neue, unbekannte Sachverhalte angewendet werden kann, um mit einem gewissen Grad an Autonomie Vorhersagen, Empfehlungen oder Entscheidungen zu treffen.[25] Im Falle von LLMs werden dabei große Mengen von textbasierten Datensätzen analysiert und bestimmten Wortfolgen Wahrscheinlichkeiten zugwiesen, wodurch die Durchführung von Schreib-, Konversations-, Zusammenfassungs- und anderen sprachlichen Aufgaben

[23] Vgl. Ertel (2021), Grundkurs Künstliche Intelligenz: Eine praxisorientierte Einführung, S. 333-337

[24] Vgl. König, P. D., Krafft, T. D., Schulz, W., & Zweig, K. A. (2021), Essence of AI: What is AI? in K. Frankish & W. M. Ramsey (Hrsg.), S. 25

[25] Vgl. Fraunhofer-Gesellschaft (2018), Maschinelles Lernen – Eine Analyse zu Kompetenzen, Forschung und Anwendung, S. 8

ermöglicht wird.[26] Einen Algorithmus kann man sich grundsätzlich wie ein Kochrezept zur Lösung eines Problems oder zur Durchführung einer bestimmten Aufgabe vorstellen. Bestimmte Algorithmen können sich kontinuierlich auf Basis von Erfahrungswerten optimieren. Als Metapher eignet sich ein Kuchenbäcker, der mit verschiedenen Zutaten und Verhältnissen experimentiert und das Backrezept auf Basis der Resultate anpasst. Ein Algorithmus besteht aus einer Abfolge von präzisen Anweisungen, die so formuliert sein können, dass Sie von einer Maschine bzw. einem Computer oder einem Menschen ausgeführt werden können.[27]

Bei maschinellen Lernverfahren wird grundsätzlich zwischen drei verschiedenen Lernstilen unterschieden: überwachtes Lernen, unüberwachtes Lernen und bestärkendes Lernen.[28]

Beim überwachten Lernen wird der Algorithmus mit Trainingsdaten gespeist, die bereits durch sog. Labels gekennzeichnet sind und ihm wird vorgegeben, was er lernen soll. Der Algorithmus erhält beispielsweise Bilder von Häusern und Autos, die entsprechend gekennzeichnet sind. Durch das jeweilige Label („Haus“ oder „Auto“) kann der Algorithmus das jeweils dargestellte Objekt identifizieren und zwischen Häusern und Autos unterscheiden. Der Algorithmus erhält von dem überwachenden Programmierer stets Rückmeldung, ob sein Ergebnis korrekt ist oder nicht. Er erlernt die wesentlichen Merkmale von Häusern und Autos zu unterscheiden, sodass er später nicht-gelabelte Bilder mit möglichst hoher Trefferquote einer der beiden Kategorien zuordnen kann.[29]

Bei unüberwachtem Lernen erkennt der Algorithmus eigenständig Muster in Trainingsdaten. Dabei erfolgt keine Kennzeichnung der Daten durch Labels, noch wird dem Algorithmus vorgegeben, was er lernen soll. So soll der Algorithmus beispielweise Muster in menschlicher Sprache erkennen und die übliche Rangfolge von Wörtern und Wortpaaren zueinander erlernen. Eine Unterkategorie des unüberwachten Lernens ist das

[26] Vgl. Yan, B., Li, K., Xu, M., Dong, Y., Zhang, Y., Ren, Z. and Cheng, X. (2024), On Protecting the Data Privacy of Large Language Models (LLMs): A Survey, S. 1

[27] Vgl. Ertel (2021), Grundkurs Künstliche Intelligenz: Eine praxisorientierte Einführung, S. 119-120

[28] Vgl. Fraunhofer-Gesellschaft (2018), Maschinelles Lernen – Eine Analyse zu Kompetenzen, Forschung und Anwendung, S. 10

[29] Vgl. Vogel (2022), Künstliche Intelligenz und Datenschutz, S. 39

sog. selbstüberwachte Lernen.[30] Dabei kennzeichnet der Algorithmus die Eingabedaten selbst mit einem entsprechenden Label, mit dem die Parameter bzw. Gewichtungen des Modells während des Trainings aktualisiert werden.[31]

Auch das bestärkende Lernen ist dem unüberwachten Lernen zuzuordnen und zeichnet sich dadurch aus, dass das KI-Modell Feedback aus der Umgebung nutzt, um ihren Output nach dem trial-and-error-Prinzip zu optimieren.[32]

2.2.2 Neuronale Netze

Die dargestellten Stile maschinellen Lernens können auf Grundlage verschiedener Techniken erfolgen. Die wohl prominenteste Technik des maschinellen Lernens sind sog. neuronale Netze. Bei dieser Technologie wird – stark vereinfacht ausgedrückt – die statistische Verteilung numerischer Werte in Eingabedaten mithilfe von mathematischen Funktionen (häufig referenziert als Knotenpunkte oder künstliche Neuronen), welche unterschiedliche Gewichtungen aufweisen, analysiert und bewertet. Aus der Kombination der einzelnen Bewertungen werden Muster abgeleitet, auf deren Basis die Gewichtungen der Neuronen angepasst werden, um die Genauigkeit der abzuleitenden Ausgaben zu verbessern (vgl. Abbildung 2).[33] Diese Technologie lehnt sich dabei an die Architektur des menschlichen Gehirns an, welches ebenso aus etwa 10-100 Milliarden Nervenzellen (oder Neuronen) besteht, die (natürliche) neuronale Netze bilden.[34] Bildlich lässt sich ein neuronales Netzwerk wohl am besten als großes Excel-Arbeitsblatt vorstellen, in welchem die einzelnen Zellen die Neuronen darstellen. Die Eingabeschicht (Input Layer) ist die erste Spalte ganz links und die letzte Spalte ganz rechts ist die Ausgabeschicht

[30] Vgl. Krauss (2023), Künstliche Intelligenz und Hirnforschung, S. 132

[31] Vgl. Krauss (2023), Künstliche Intelligenz und Hirnforschung, S. 132

[32] Vgl. Fraunhofer-Gesellschaft (2018), Maschinelles Lernen – Eine Analyse zu Kompetenzen, Forschung und Anwendung, S. 10; Vgl. Vogel (2022), Künstliche Intelligenz und Datenschutz, S. 42

[33] Vgl. Rosenthal (2024), Was in einem KI-Modell steckt und wie es funktioniert, Blogbeitrag vom 07.05.2024, https://www.vischer.com/know-how/blog/teil-17-was-in-einem-ki-modell-steckt-und-wie-es-funktioniert/, (letzter Aufruf am 31.08.2024); OECD (2020), Künstliche Intelligenz in der Gesellschaft, S. 29 (letzter Aufruf am 31.08.2024); Fraunhofer-Gesellschaft (2018), Maschinelles Lernen: Eine Analyse zu Kompetenzen, Forschung und Anwendung, S. 11

[34] Vgl. Ertel (2021), Grundkurs Künstliche Intelligenz: Eine praxisorientierte Einführung, S. 285

(Output Layer).[35] Dazwischen befinden sich in der Regel weitere Neuronen-Schichten (sog. Zwischenschichten oder Hidden Layer).

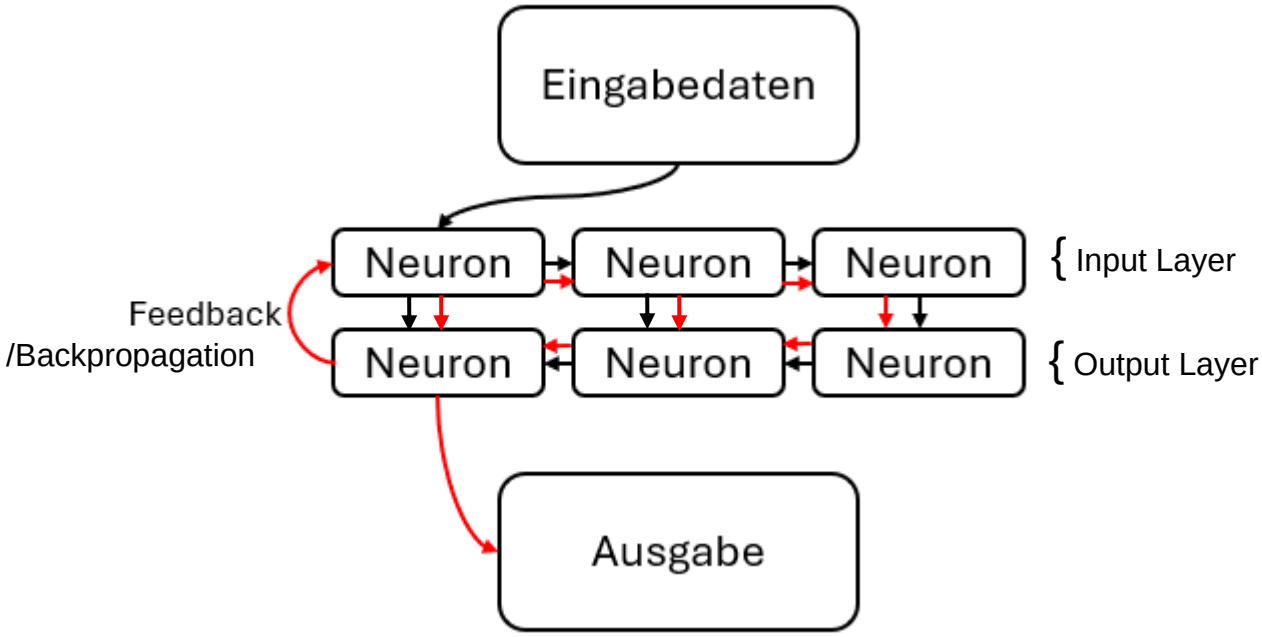

Abbildung 2: stark vereinfachte schematische Darstellung der Funktionsweise eines Neuronalen Netzes

Unter dem Begriff Deep Learning versteht man maschinelles Lernen, bei dem auf besonders große neuronale Netze mit mehreren hundert Neuronen-Schichten bzw. Layers zurückgegriffen wird.[36]

2.3 Training neuronaler Netze

Damit mit neuronalen Netzen KI-Modelle geschaffen werden können, welche qualitativ hochwertigen Output generieren, müssen neuronale Netze zunächst mit großen Datenmengen trainiert werden. GPT-3 wurde beispielsweise anhand von 45 Terabyte

[35] Vgl. Rosenthal (2024), Was in einem KI-Modell steckt und wie es funktioniert, Blogbeitrag vom 07.05.2024, https://www.vischer.com/know-how/blog/teil-17-was-in-einem-ki-modell-steckt-und-wie-es-funktioniert/, (letzter Aufruf am 31.08.2024)

[36] Vgl. OECD (2020), Künstliche Intelligenz in der Gesellschaft, S. 29 (letzter Aufruf am 31.08.2024)

(45.000 Gigabyte) komprimierten Klartexts (vor Filterung) trainiert.[37] Grundsätzlich gilt, je mehr Trainingsdaten ein Algorithmus erhält, desto präziser kann das KI-Modell optimiert und dessen Fehlerquote verringert werden.[38] Dabei spielt auch die Qualität der Trainingsdaten eine Rolle. Wird der Algorithmus mit inhaltlich inkorrekten oder nicht repräsentativen Daten trainiert, wird das KI-Modell anfälliger für unerwünschte Outputs, wie bspw. diskriminierende Inhalte. Bias in Trainingsdaten, also Verzerrungen bzw. ungewollte Korrelationen zwischen Merkmalen, sowie einseitige bzw. unvollständige Trainingsdaten können schwerwiegende Folgen haben. So können bspw. schlecht trainierte KI-Modelle, welche bei autonomem Fahren eingesetzt werden, schwere Verkehrsunfälle verursachen.[39] Die Qualität von Trainingsdatensätzen kann Studien zufolge wohl anhand verschiedener Metriken, wie z. B. dem sog. Label Noise (Fehler bei der Kennzeichnung von Daten), sog. Class Imbalance (Ungleichgewichtsverhältnisse in den Trainingsdaten) oder der sog. Datenhomogenität (Einseitigkeit der Trainingsdaten) bewertet werden.[40] Um ein LLM auf vielfältige und komplexe Fähigkeiten zu trainieren, sind insbesondere heterogene Trainingsdatensätze aus verschiedenen Domänen (z. B. Bücher und Webseiten) erforderlich, die unterschiedliche Wissens- bzw. Fachgebiete abdecken.[41] Trainingsdaten werden in der Regel aus dem Internet über sog. Web Scraping oder Web Crawling oder aus eigenen, nicht-öffentlichen Quellen bezogen. Solche Trainingsdatensätze bestehen z. B. aus Freitext, wie er in Aufsätzen, Briefen usw. vorkommt sowie aus Text, der aus formatierten oder halbstrukturierten Daten, wie bspw.

[37] Vgl. Yan et al. (2024), On Protecting the Data Privacy of Large Language Models (LLMs): A Survey, S. 4

[38] Vgl. Fraunhofer-Gesellschaft (2018), Maschinelles Lernen – Eine Analyse zu Kompetenzen, Forschung und Anwendung, S. 11; Kaplan et al. (2020), Scaling laws for neural language models, S. 3

[39] Vgl. Golson (2016), Tesla's Autopilot was involved in a fatal crash for the first time, The Verge, 30. Juni 2016, https://www.theverge.com/2016/6/30/12072408/tesla-autopilot-car-crash-death-autonomous-model-s (letzter Aufruf: 08.09.2024)

[40] Vgl. Jain et al. (2020), Overview and importance of data quality for machine learning tasks, in Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining, pages 3561–3562

[41] Vgl. Longpre et al. (2023), A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality,& toxicity

Dokumenten, Social Media-Posts, Produktbewertungen oder Webseiten, extrahiert wird.[42] Folglich können Trainingsdatensätze alle möglichen Datenarten enthalten, in der Regel auch personenbezogene Daten im Sinne des Art. 4 Nr. 1 DSGVO. Studien zufolge enthalten hochwertige Datensätze (z. B. Datensätze, die überwiegend Texte aus Büchern und akademische Texte enthalten) oft mehr personenbezogene Daten als minderwertige Datensätze (z. B. Datensätze, die hauptsächlich Webdaten enthalten).[43] Personenbezogene Daten, die in Trainingsdaten enthalten sind, sind in der Regel Namen, E-Mail-Adressen und Telefonnummern.[44] Aus datenschutzrechtlicher Sicht stellen sich insbesondere folgende Fragen:

1. Inwieweit finden datenschutzrechtliche Anforderungen beim Training von LLMs Anwendung und wie können diese erfüllt werden?

2. Welche Risiken bestehen für die Rechte und Freiheiten der betroffenen Personen i. S. d. Art. 4 Nr. 1 DSGVO, wie schwerwiegend sind diese Risiken und welche Abhilfemaßnahmen sind denkbar und ggf. erforderlich?

Für die Darstellung und Bewertung etwaiger datenschutzrechtlicher Problemstellungen beim Training von LLMs, ist es essenziell zu verstehen, was mit den Trainingsdaten passiert und ob dabei eine Verarbeitung i. S. d. Art. 4 Nr. 2 DSGVO stattfindet. Entsprechend ihrer Relevanz und Popularität begrenzen wir uns auf die Technologie der neuronalen Netze. Der Trainingsprozess von auf neuronalen Netzen basierter LLMs umfasst in der Regel die folgenden Phasen, die nachfolgend noch detaillierter erläutert werden: (1.) Trainingsdatensammlung, (2.) Trainingsdatenaufbereitung, (3.) sog. Pre-Training und (4.) sog. Fine Tuning.

[42] Vgl. Gupta et al. (2021), Data quality for machine learning tasks, in Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining, S. 4041

[43] Vgl. Longpre et al. (2023), A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality,& toxicity., S. 5

[44] Vgl. Longpre et al. (2023), A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality,& toxicity., S. 5

2.3.1 Trainingsdatensammlung

Geeignete Trainingsdaten müssen zunächst einmal gesammelt werden. Trainingsdaten können aus nicht-öffentlichen Quellen, wie z. B. Vertriebsdokumentationen und Finanzberichten oder aus öffentlich zugänglichen Quellen, wie z. B. Webseiten und Fachzeitschriften bezogen werden.

Das Internet bietet grundsätzlich riesige Datenmengen, die öffentlich zugänglich sind und Tag für Tag wachsen. Zur Sammlung von öffentlich zugänglichen Daten werden in der Regel sog. Web-Crawler und Web-Scraper verwendet, die unstrukturierte Informationen von Webseiten extrahieren und diese in strukturierte, maschinenlesbare Informationen konvertieren.[45] Solche Programme basieren auf sog. Bots, die das Internet automatisch nach Informationen durchforsten. Die ersten bekannten Web-Crawler wurden von den Suchmaschinenanbietern Google und Microsoft entwickelt. Diese gewährleisten die Funktionen der Suchmaschinen Google und Bing, in dem sie das Internet nach Webseiten durchsuchen, diese „scannen", Informationen extrahieren und einen durchsuchbaren Index erstellen.[46] Dabei extrahiert der Web-Crawler die Informationen aus dem Internet während der Web-Scraper dafür zuständig ist, die relevanten Informationen aus dem gecrawlten Datensatz zu filtern und in ein strukturiertes Format zu konvertieren.[47] Solche Programme basieren in der Regel auf benutzerdefinierten sog. Hypertext Transfer Protocol-Requests (HTTP-Anfragen), eine Technologie zur Kommunikation zwischen Webservern. Der Web-Crawler versendet bspw. einen sog. GET Request an eine Zieladresse (sog. Uniform Ressource Locator, kurz „URL") und erhält eine HTML-Seite als Antwort. Daraufhin extrahiert der Web-Scraper die relevanten Informationen von der HTML-Seite und konvertiert diese in ein strukturiertes Format.[48] Oft sammeln KI-

[45] Vgl. Khder (2021), Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application, International Journal of Advances in Soft Computing and its Applications, S. 146

[46] Vgl. ebd.; Garante per la protezione dei dati personali (2024), Web scraping ed intelligenza artificiale generativa: nota informativa e possibili azioni di contrasto, https://www.gpdp.it/web/guest/home/docweb/-/docweb-display/docweb/10020316 (letzter Aufruf: 12.09.2024)

[47] Vgl. Khder (2021), Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application, International Journal of Advances in Soft Computing and its Applications, S. 152

[48] Vgl. Khder (2021), Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application, International Journal of Advances in Soft Computing and its Applications, S. 152

Hersteller die Trainingsdaten nicht selbst, sondern beziehen diese z. B. über Internetplattformen für den Datenhandel (sog. Datenmarktplätze).[49]

Ferner gibt es auch umfangreiche, teilweise bereits strukturierte Trainingsdatensätze, die öffentlich zur Verfügung gestellt werden, wie z. B. das Projekt Common Crawl oder die Pile-Datenbank. Common Crawl ist ein Non-Profit-Projekt, in dessen Rahmen das Internet monatlich gecrawlt wird und die gesammelten Daten über eine Amazon-Cloud-Umgebung freizugänglich bereitgestellt werden.[50] Pile ist ein 825 Gigabyte großer englischsprachiger Textkorpus, der speziell für das Training von LLMs gedacht ist.[51] Im Gegensatz zu Common Crawl, welches eher unsortierte Rohdaten anbietet, bietet Pile bereits strukturierte Daten, welche aus ausgewählten Quellen stammen (u. a. Fachliteratur, Bücher, Online-Foren etc.)[52]

2.3.2 Aufbereitung der Trainingsdaten

In der Regel müssen die gesammelten Trainingsdatensätze zunächst für die jeweiligen konkreten Zwecke des gewünschten LLM aufbereitet und von unerwünschten, minderwertigen oder anstößigen Inhalten, wie bspw. maschinell generierte oder pornografische Inhalte, bereinigt werden.[53] Die Bereinigung der Trainingsdaten von unerwünschten Inhalten erfolgt oft über automatisierte Filterungs-Prozesse (sog. Content Filtering). Gefiltert wird dabei bspw. anhand von Kriterien, wie Satzlänge, unerwünschte Wörter (sog. Stopwords), Interpunktion und Wiederholungshäufigkeit.[54] Im Rahmen der

[49] Vgl. Trauth/Mayer in Rohde, Bürger, Peneva & Mock (Hrsg.) (2020), Datenwirtschaft und Datentechnologie, S. 19

[50] Vgl. Litwintschik (2017), Analysing Petabytes of Websites, https://tech.marksblogg.com/petabytes-of-website-data-spark-emr.html (letzter Aufruf: 11.09.2024)

[51] Vgl. Gao et al. (2020), The Pile: An 800GB Dataset of Diverse Text for Language Modelling

[52] Vgl. Gao et al. (2020), The Pile: An 800GB Dataset of Diverse Text for Language Modelling

[53] Vgl. Laurençon et al. (2023), The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset, S. 7

[54] Vgl. Longpre et al. (2023), A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality,& toxicity, S. 5; Laurençon et al. (2023), The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset, S. 7

Aufbereitung von Trainingsdaten kann es zudem erforderlich sein, die Trainingsdatensätze einheitlich auf ein bestimmtes Textformat zu konvertieren und von Duplikaten zu bereinigen.[55] Solche Aufbereitungsmaßnahmen können u. a. auch dazu dienen, datenschutzrechtliche Risiken zu mitigieren. So kann mit der Bereinigung von Duplikaten dem sog. Memorisieren von Trainingsdaten (s. hierzu Kapitel 2.4) entgegengewirkt werden oder es werden dediziert personenbezogene Daten aus den Trainingsdaten herausgefiltert.[56]

Nun kann man sich fragen, weshalb man Trainingsdatensätze im Rahmen, der dem Training vorgelagerten Aufbereitung nicht standardmäßig von darin enthaltenen personenbezogenen Daten bereinigt, um etwaige datenschutzrechtliche Problemstellungen zu umgehen. Zum einen kann eine solche Bereinigung der Trainingsdaten von personenbezogenen Daten die Gefahr einer Offenlegung personenbezogener Daten während des Produktiveinsatzes des LLMs (bspw. durch sog. de-anonymization attacks) nicht mit Sicherheit ausschließen und zum anderen haben solche Maßnahmen in der Regel negative Auswirkungen auf die Qualität der Trainingsdaten und verschlechtern die Genauigkeit oder das „Verständnis“ des LLMs.[57] Ferner bestehen technische und finanzielle Hürden für die Bereinigung von so umfangreichen Datensätzen, wie sie für das Training von LLMs benötigt werden. Die KI-Hersteller stehen dabei vor der Herausforderung, den Aufwand einer Datenbereinigung gegen deren Nutzen abzuwägen. In der Regel ergibt eine solche Abwägung im Ergebnis, dass der Aufwand für solche datenschutzfördernden Maßnahmen bei der Trainingsdatenaufbereitung negativ mit der Qualität des LLMs korreliert und eine vollständige Bereinigung der Trainingsdaten von personenbezogenen Daten schlichtweg nicht effizient ist (vgl. hierzu auch Kapitel 5).[58]

[55] Vgl. Rae et al. (2021), Scaling language models: Methods, analysis &insights from training Gopher, S. 39

[56] Vgl. Laurençon et al. (2023), The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset, S. 8

[57] Vgl. Narayanan et al. (2008), Robust De-anonymization of Large Sparse Datasets; Xu et al. (2021), Privacy-Preserving Machine Learning: Methods, Challenges, and Directions, S. 3

[58] Vgl. Xu et al. (2021), Privacy-Preserving Machine Learning: Methods, Challenges, and Directions, S.28; El Mestari et al. (2024), Preserving Data Privacy in Machine Learning Systems, S. 15; Carlini et al. (2021), Extracting Training Data from Large Language Models, S. 3

2.3.3 Pre-Training

Das eigentliche Training eines LLMs findet in der sog. Pre-Training-Phase statt. LLMs generieren einen Output-Text, indem jeweils das wahrscheinlichste nächste Wort in einer Wortfolge ermittelt wird. Dies erfolgt unter Berücksichtigung der Bedeutung der jeweiligen Worte im Kontext mit dem Input-Text, die ein LLM im Rahmen des Trainings erlernt hat.[59] Die jeweiligen Bedeutungen sind im LLM als numerische Werte hinterlegt.[60] Diese Werte werden im Rahmen der ersten Trainingsphase anhand von umfangreichen Trainingsdatensätzen berechnet (sog. Pre-Training). Während des Pre-Trainings lernt das Modell die statistischen Eigenschaften der Trainingsdaten, d. h. wie Wörter und Sätze in einem bestimmten Kontext vorkommen. Es erlernt allgemeine Sprachfähigkeiten wie Syntax, Grammatik und semantische Beziehungen zwischen den Wörtern.[61]

Das Pre-Training ist typischerweise als selbstüberwachtes Lernen zu kategorisieren. In einem ersten Schritt werden dabei die Trainingsdaten in Ihre Bestandteile, d. h. in Wort- bzw. Satzfragmente (sog. Tokens), zerlegt (sog. Tokenisierung). Jeder Eingabesatz wird in mehrere Sequenzen aufgebrochen, welche bspw. aus Silben oder Kombinationen von Silben, Wortfragmenten und manchmal auch vollständigen Wörtern bestehen.[62] Der Satz „Max Schrems wurde 1987 in Salzburg geboren.“ wird vom GPT-4-Tokenizer bspw. wie folgt aufgebrochen: [Max][Sch][re][ms][wurde][ ][198][7][in][Sal][zburg][

[59] Vgl. Rosenthal (2024), Was in einem KI-Modell steckt und wie es funktioniert, Blogbeitrag vom 07.05.2024, https://www.vischer.com/know-how/blog/teil-17-was-in-einem-ki-modell-steckt-und-wie-es-funktioniert/ (letzter Aufruf am 31.08.2024)

[60] Vgl. Rosenthal (2024), Was in einem KI-Modell steckt und wie es funktioniert, Blogbeitrag vom 07.05.2024, https://www.vischer.com/know-how/blog/teil-17-was-in-einem-ki-modell-steckt-und-wie-es-funktioniert/ (letzter Aufruf am 31.08.2024)

[61] Vgl. Rosenthal (2024), Was in einem KI-Modell steckt und wie es funktioniert, Blogbeitrag vom 07.05.2024, https://www.vischer.com/know-how/blog/teil-17-was-in-einem-ki-modell-steckt-und-wie-es-funktioniert/ (letzter Aufruf am 31.08.2024)

[62] Vgl. Rosenthal (2024), Was in einem KI-Modell steckt und wie es funktioniert, Blogbeitrag vom 07.05.2024, https://www.vischer.com/know-how/blog/teil-17-was-in-einem-ki-modell-steckt-und-wie-es-funktioniert/ (letzter Aufruf am 31.08.2024)

geb][oren][.].[63] Diese Tokens werden in GPT-4 durch folgende numerische Werte dargestellt: 6102, 5124, 265, 1026, 27201, 220, 3753, 22, 304, 8375, 79302, 30267, 24568, 13.[64] Den Tokens wird anschließend, je nach ihrer Bedeutung, die ihnen im Trainingsdatensatz durchschnittlich zukommt, eine bestimmte Position in einer Art Koordinatensystem (sog. Embedding-Raum) zugewiesen.[65] Es ist anzumerken, dass das KI-Modell nicht die objektive, allgemeingültige Bedeutung der Wörter bzw. Wortfragmente kennt. Die jeweilige Bedeutung wird vielmehr anhand der durchschnittlichen Häufigkeit des Tokens in Kombination bzw. der Abfolge zu anderen Tokens in den Trainingsdatensätzen definiert.[66] Tokens mit ähnlicher Bedeutung liegen im Embedding-Raum näher beieinander und über Vektoren zwischen den Koordinaten werden Verhältnisse zwischen den Tokens auf Basis der Wahrscheinlichkeit der Wortabfolge abgebildet.[67] So werden verschiedene sog. Cluster gebildet. Bei der Generierung eines Ausgabetextes wählt das Modell das nächstwahrscheinlichste Token basierend auf seiner Position im Embedding-Raum und den zuvor verarbeiteten Tokens. Durch wiederholte Auswahl der wahrscheinlichsten Tokens entsteht eine kohärente Wortfolge als Antwort auf den jeweiligen Input-Text.

Im Rahmen des Pre-Trainings entwickelt das Modell idealerweise ein breites und flexibles Sprachverständnis, welches jedoch oft nicht dem jeweils gewünschten Anwendungsbereich gerecht wird. Vielmehr wird oft eine spezifische Expertise in

[63] Vgl. OpenAI Tokenizer, aufrufbar unter https://platform.openai.com/tokenizer (letzter Aufruf: 07.09.2024)

[64] Vgl. OpenAI Tokenizer, aufrufbar unter https://platform.openai.com/tokenizer (letzter Aufruf: 07.09.2024)

[65] Vgl. Rosenthal (2024), Was in einem KI-Modell steckt und wie es funktioniert, Blogbeitrag vom 07.05.2024, https://www.vischer.com/know-how/blog/teil-17-was-in-einem-ki-modell-steckt-und-wie-es-funktioniert/ (letzter Aufruf am 31.08.2024)

[66] Vgl. Rosenthal (2024), Was in einem KI-Modell steckt und wie es funktioniert, Blogbeitrag vom 07.05.2024, https://www.vischer.com/know-how/blog/teil-17-was-in-einem-ki-modell-steckt-und-wie-es-funktioniert/ (letzter Aufruf am 31.08.2024)

[67] Vgl. Rosenthal (2024), Was in einem KI-Modell steckt und wie es funktioniert, Blogbeitrag vom 07.05.2024, https://www.vischer.com/know-how/blog/teil-17-was-in-einem-ki-modell-steckt-und-wie-es-funktioniert/ (letzter Aufruf am 31.08.2024)

bestimmten Aufgaben oder Domänen gefordert, welche im Rahmen der nächsten Trainingsphase, des sog. Fine Tunings, gewonnen wird.

2.3.4 Fine Tuning

Beim Fine Tuning wird das Model auf bestimmte Aufgaben (bspw. Textklassifikation, maschinelle Übersetzung, Frage-Antwort-Systeme etc.) und Domänen spezialisiert.[68] Das Modell erhält zusätzliche Daten, die als erwünschter oder unerwünschter Output gelabelt sind. So lernt das Modell anhand von Beispielen, welcher Input zu welchem Output führen sollte. Hierdurch werden die Parameter des Models optimiert.

Das Fine Tuning ist dabei oft von Menschen überwacht (überwachtes Lernen), welche dem Modell Feedback geben, um das Modell auf das gewünschte Verhalten zu konditionieren (sog. Alignment). Mit dem Alignment soll sichergestellt werden, dass das Modell nicht nur inhaltlich korrekte, sondern bspw. auch moralisch korrekte und zielführende Antworten generiert.[69]

2.3.5 Inferenz-Phase

Während der sog. Inferenz-Phase befindet sich das fertig trainierte LLM bereits in der Anwendung, um Output-Text auf Basis von Eingaben des jeweiligen Nutzers (sog. Prompts) zu generieren. Dabei werden die während des Trainings erlernten Muster angewendet, ohne dass das Modell weiter trainiert oder verändert wird. In dieser Phase zeigt sich die tatsächliche Leistungsfähigkeit des LLM und insbesondere, wie gut das Modell generalisiert, also wie es auf neue Eingaben reagiert, die nicht explizit in den Trainingsdaten enthalten waren. Ein Modell, das gut generalisiert, kann seine erlernten Muster auch auf neue, unbekannte Eingaben anwenden.

2.4 Memorisieren von Trainingsdaten

Wie in Kapitel 2.3.3 dargestellt, arbeiten LLMs in Wirklichkeit nicht mit ganzen Wörtern, sondern mit Wort- bzw. Satzfragmenten (sog. Tokens). Der Output erfolgt ebenso

[68] Vgl. Wei et al. (2022), Finetuned language models are zero-shot learners, S. 28

[69] Vgl. Rosenthal (2024), Was in einem KI-Modell steckt und wie es funktioniert, Blogbeitrag vom 07.05.2024, https://www.vischer.com/know-how/blog/teil-17-was-in-einem-ki-modell-steckt-und-wie-es-funktioniert/ (letzter Aufruf am 31.08.2024)

zunächst als Folge der Tokens, die dann wieder zurück in die entsprechenden Buchstabenfolgen umgewandelt werden. Der Prozess der Tokenisierung kann nicht mit einer klassischen Verschlüsselung oder Pseudonymisierung verglichen werden, bei welcher jedem Token die jeweils dazugehörige Wortabfolge über einen „Schlüssel" zugeordnet werden kann.[70] Ein LLM ist also nicht als eine Art Datenbank zu verstehen, in welcher die Trainingsdatensätze gespeichert und abrufbar sind. Die beim Training verwendeten Datensätze lassen sich folglich in der Regel auch nicht mehr 1x1 rekonstruieren. Eine Analyse von Texten, die von Transformer-Modellen (darunter GPT-2) generiert wurden, hat gezeigt, dass LLM teilweise sogar Neologismen schaffen (bspw. „swissified" und „IKEA-ness")[71].

Unter bestimmten Umständen kann es dennoch dazu kommen, dass LLMs in ihren Ausgaben umfangreiche Passagen (z. B. bis zu 1.000 Wörter lang)[72] aus Trainingsdatensätzen „kopieren".[73] Es wird vermutet, dass dies insbesondere dann vorkommen kann, wenn die jeweilige Passage wiederholt in den Trainingsdatensätzen vorkommt.[74] Die Skalierung des Modells, Dopplungen in den Trainingsdaten und der Kontext des jeweiligen Prompts wurden als wichtige Faktoren identifiziert, die den Grad der Fähigkeit des Models, Trainingsdaten zu memorisieren, beeinflussen.[75]

[70] Vgl. Hamburgischer Beauftragter für Datenschutz und Informationsfreiheit (2024), Diskussionspapier: Large Language Models und personenbezogene Daten, S. 4

[71] Vgl. McCoy et al. (2023), How much do language models copy from their training data? Evaluating linguistic novelty in text generation using RAVEN, S. 653

[72] Vgl. McCoy et al. (2023), How much do language models copy from their training data? Evaluating linguistic novelty in text generation using RAVEN, S. 658

[73] Vgl. McCoy et al. (2023), How much do language models copy from their training data? Evaluating linguistic novelty in text generation using RAVEN, S. 658

[74] Vgl. McCoy et al. (2023), How much do language models copy from their training data? Evaluating linguistic novelty in text generation using RAVEN, S. 658

[75] Vgl. Carlini et al. (2023), Quantifying Memorization Across Neural Language Models

2.5 Training data extraction attacks

Das Memorisieren von Trainingsdaten bei LLMs kann dazu führen, dass Nutzer – sei es unbeabsichtigt oder vorsätzlich – personenbezogene oder gar sensible Daten durch die Eingabe bestimmter Prompts aus den Trainingsdaten des LLM extrahieren. So ist es bspw. denkbar, dass die Eingabe „Die Adresse von Max Schrems lautet …“ von dem jeweiligen LLM mit einer tatsächlich existierenden Adresse ergänzt wird. Studien zu folge kann von den meisten LLMs rund 1% der Trainingsdaten extrahiert werden.[76]

Nachfolgend sollen die Möglichkeiten der vorsätzlichen Extraktion von Trainingsdaten dargestellt werden. Das Memorisieren von Trainingsdaten kann ausgenutzt werden, um Rückschlüsse auf die konkreten Trainingsdaten zu ziehen oder gar konkrete Textsequenzen aus den verwendeten Trainingsdaten (ggf. inkl. personenbezogener Informationen) zu extrahieren (bspw. mittels sog. membership inference attacks oder sog. model inversion attacks, gemeinsam auch „training data extraction attacks“).[77]

Unter anderem ein übermäßiges Fine Tuning kann die Gefahr einer solchen Extraktion erhöhen, wenn das Modell im Rahmen des Fine Tunings zu stark auf spezifische Details der Trainingsdaten ausgerichtet wird, anstatt auf die Erkennung breiter Muster.[78] In solchen Fällen reagieren LLMs anders auf Eingaben, die Daten beinhalten, die in den Trainingsdatensätzen enthalten waren als auf Daten, die sie zum ersten Mal „sehen“.[79] So erhalten Daten, die in den Trainingsdatensätzen enthalten waren, höhere numerische Vorhersagewerte als solche, die nicht in den Trainingsdatensätzen enthalten waren.[80] Dieses Phänomen wird Overfitting bzw. Überanpassung genannt und bietet die Grundlage für training data extraction attacks.[81] Angreifer können ein überangepasstes LLM zum

[76] Vgl. Müller et al. (2024), LLMs and Memorization: On Quality and Specificity of Copyright Compliance

[77] Vgl. Carlini et al. (2021), Extracting Training Data from Large Language Models, S. 1

[78] Vgl. Mireshghallah et al. (2022), An Empirical Analysis of Memorization in Fine-tuned Autoregressive Language Models

[79] Vgl. Shokri et al. (2017), Membership Inference Attacks Against Machine Learning Models

[80] Vgl. Shokri et al. (2017), Membership Inference Attacks Against Machine Learning Models; Yeom et al. (2018), Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting, S. 2

[81] Vgl. Shokri et al. (2017), Membership Inference Attacks Against Machine Learning Models

Beispiel durch spezielle Eingabeaufforderungen oder Auslösesequenzen (wie Sonderzeichen oder wiederholte Sätze) dazu bringen, konkrete Trainingsdaten zu reproduzieren. So ist es beispielsweise möglich, ein spezifisches Datum aus dem Trainingsdatensatz zu extrahieren, wenn der Kontext, in dem dieses Datum in den Trainingsdaten vorkam, bekannt ist.[82]

Solche training data extraction attacks können auf verschiedenen Ansätzen basieren. Unterschieden wird dabei insbesondere zwischen den folgenden Kategorien:

(1) Sog. membership inference attacks: Solche Attacken zielen darauf ab, festzustellen, ob ein bestimmtes Eingabedatum in den Trainingsdaten enthalten war. Dabei wird davon ausgegangen, dass die Zugehörigkeit eines Datenpunktes zu den Trainingsdaten in der Regel zu höheren numerischen Vorhersagewerten führt, während Datenpunkte, die nicht in den Trainingsdaten enthalten waren, niedrigere Werte liefern.[83] Das bedeutet, dass ein LLM bei einer Abfrage eines Datenpunktes, der in den Trainingsdaten enthalten war, spezifischere, präzisere oder „selbstbewusstere" Antworten bzw. Vorhersagen liefert. Die jeweilige Ausgabe des LLMs wird sodann auf Vorhersagewahrscheinlichkeiten und die Antwortstruktur analysiert.

(2) Sog. reconstruction attacks: Bei Rekonstruktionsangriffen (auch „attribute inference" oder „model inversion" genannt) wird versucht Trainingsdatensätze in Teilen oder vollständig zu rekonstruieren. Dies erfolgt in der Regel durch gezielte Eingaben, bei denen das Modell aufgefordert wird, einen Text, um fehlende Informationen zu ergänzen bzw. das nächste Wort oder die nächste Sequenz in einem unvollständigen Text vorherzusagen.[84] Durch die wiederholte Eingabe zuvor vorhergesagter Wörter und das weitere Fortsetzen dieser Vorhersagen versuchen Angreifer, Textsequenzen zu erzeugen, die wahrscheinlich aus den ursprünglichen Trainingsdaten stammen. Die

[82] Vgl. Thomas et al. (2020), Investigating the Impact of Pre-trained Word Embeddings on Memorization in Neural Networks, S. 3

[83] Vgl. Shokri et al. (2017), Membership Inference Attacks Against Machine Learning Models; Yeom et al. (2018), Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting, S. 2

[84] Vgl. Yeom et al. (2018), Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting, S. 2

Rekonstruktion erfolgt häufig durch statistische Auswertung und gezielte Wiederholungen von Eingaben, um Muster zu erkennen, die auf Trainingsdaten hindeuten.

(3) Sog. property inference attacks: Ziel von property inference attacks ist es, Informationen zu extrahieren, die von dem jeweiligen LLM unbeabsichtigt gelernt wurden und für die konkreten Anwendungsbereich des LLMs nicht relevant sind, wie bspw. die Information über das Verhältnis von männlichen und weiblichen Patienten in einem Patientendatensatz.[85]

(4) Sog. model extraction attacks: Model extraction attacks zielen darauf ab, Informationen über das LLM (insb. Modell-Parameter) zu extrahieren, um das jeweilige Modell zu rekonstruieren und eine Kopie des jeweiligen LLM zu erstellen.

Insbesondere die zuletzt genannten property inference attacks und model extraction attacks erfolgen in der Regel mithilfe sog. shadow models bzw. Schattenmodellen und meta models. Der Angreifer trainiert mehrere Schattenmodelle anhand von sog. Schattendatensätzen, von denen angenommen wird, dass sie dieselbe statistische Verteilung aufweisen wie der Trainingsdatensatz des attackierten LLMs. Nach dem Training der Schattenmodelle erstellt der Angreifer einen sog. Angriffsdatensatz, bei dem jede Eingabe den Ausgaben des Schattenmodells entspricht. Dieser Datensatz wird dann zum Trainieren eines Metamodells verwendet, das lernt, aus den Ausgaben der Schattenmodelle Informationen abzuleiten. Nach dem Training wird dieses Metamodell auf die Ausgaben des Zielmodells angewendet, so dass der Angreifer Rückschlüsse auf die Trainingsdaten und/oder Parameter des Zielmodells ziehen kann.[86]

Dem Memorisieren von Trainingsdaten und dem Erfolg von training data extraction attacks kann durch die Anwendung von verschiedenen datenschutzfördernden Methoden im Rahmen der Trainingsdatenaufbereitung oder des Pre-Trainings entgegengesteuert werden (s. hierzu Kapitel 5). Allerdings sind solche Methoden zeit- und ressourcenaufwendig und führen in der Regel zu einem gewissen Qualitätsverlust, wie bereits in Kapitel 2.3.2 dargestellt.

[85] Vgl. Rigaki, Garcia (2023), A Survey of Privacy Attacks in Machine Learning, S. 9

[86] Vgl. Rigaki, Garcia (2023), A Survey of Privacy Attacks in Machine Learning, S. 12

Die Beobachtung, dass personenbezogene Daten, die in den Trainingsdaten enthalten waren, unbeabsichtigt oder beabsichtigt extrahiert werden können, legt die Vermutung nahe, dass im Rahmen des Trainings von LLMs eine Verarbeitung von personenbezogenen Daten i. S. d. Art. 4 Nr. 2 DSGVO stattfindet und als Folge die Anforderungen der DSGVO zur Anwendung kommen. Als Gegenargumentation ließe sich jedoch insbesondere der Prozess der Tokenisierung von Trainingsdaten und der Umstand, dass die Trainingsdaten-Tokens innerhalb des LLM lediglich als numerische Werte gespeichert werden, anführen (vgl. hierzu Kapitel 2.3.3). Die Fragestellung, ob eine Verarbeitung personenbezogener Daten im Rahmen des Trainings von LLMs anzunehmen ist, wird in Kapitel 4 im Detail erläutert.

Unabhängig von dieser Fragestellung birgt die Funktionsweise von LLMs ein gewisses Konfliktpotential mit datenschutzrechtlichen Anforderungen, insbesondere im Hinblick auf das Memorisieren von Trainingsdaten und die Möglichkeit der Offenlegung von personenbezogenen Daten. Im Rahmen entsprechender Studien konnte festgestellt werden, dass LLMs teilweise memorisierte personenbezogene Daten aus verschiedenen Quellen, die nicht miteinander zusammenhängen, im Ausgabetext vermischen. So hatte ein LLM-basierter Chatbot bspw. einen Zeitungsartikel zu einem Mord generiert, der tatsächlich stattgefunden hat, und den Mord dem Opfer einer Massenschießerei, die sich ebenfalls tatsächlich zugetragen hat, zugeschrieben.[87] Zudem konnten bei training data extraction attacks Trainingsdaten extrahiert werden, die aus dem öffentlichen Internet entfernt wurden und über eine Websuche nicht mehr aufgefunden werden können.[88] Folglich sind insbesondere Konflikte mit dem datenschutzrechtlichen Grundsatz der Vertraulichkeit und dem Grundsatz der Speicherbegrenzung ersichtlich (vgl. hierzu Art. 5 DSGVO). Dieses Spannungsverhältnis zwischen der Funktionsweise von LLMs und datenschutzrechtlichen Anforderungen wird im nachfolgenden Kapitel im Detail beleuchtet.

[87] Vgl. Rigaki, Garcia (2023), A Survey of Privacy Attacks in Machine Learning, S. 12

[88] Vgl. Rigaki, Garcia (2023), A Survey of Privacy Attacks in Machine Learning, S. 12

3 Spannungsverhältnis Künstliche Intelligenz & DSGVO

Datenschutzrechtliche Regulierung, wie die DSGVO, stellt bekanntlich hohe Anforderungen an die Verarbeitung personenbezogener Daten, also an grds. jede Art der Nutzung von Informationen, die sich auf eine identifizierbare natürliche Person beziehen (darunter das Erheben, Speichern, Organisieren, Verändern, Offenlegen und Übermitteln personenbezogener Informationen). Es gilt das Prinzip des sog. Verbots mit Erlaubnisvorbehalt, wenn auch strittig ist, ob dieser Begriff im Rahmen der DSGVO dogmatisch zutreffend ist.[89] Nach diesem Prinzip ist die Verarbeitung personenbezogener Daten als Grundrechtseingriff grundsätzlich verboten, außer sie erfolgt unter der Einhaltung bestimmter datenschutzrechtlicher Anforderungen und Grundsätze. Nach Maßgabe der DSGVO muss eine Verarbeitung personenbezogener Daten insbesondere rechtmäßig, nach Treu und Glauben sowie transparent erfolgen und auf das Maß beschränkt sein, das zur Erreichung der jeweiligen vorab festgelegten Zwecke erforderlich ist. Ferner ist bei der Verarbeitung personenbezogener Daten die Richtigkeit sowie die Integrität und Vertraulichkeit der personenbezogenen Daten zu wahren.[90] Bei diesen Grundsätzen handelt es sich um obligatorische Grundanforderungen an rechtmäßige Datenverarbeitungen.

Die Frage vorweggenommen, ob und welche Prozesse im Rahmen des Lebenszyklus eines LLMs im Anwendungsbereich der DSGVO liegen, werden nachfolgend bestimmte gesetzlichen Anforderungen aus der DSGVO, der im vorherigen Kapitel dargestellten Funktionsweise von LLMs gegenübergestellt. Bei den Problemstellungen, die sich aus der Gegenüberstellung ergeben, handelt es sich um grundlegende Unsicherheiten bei der Anwendung der DSGVO, die außerhalb des KI-Kontextes bereits in der Vergangenheit in der Praxis und der Literatur diskutiert wurden und teilweise Gegenstand von Rechtsprechung sind. Künstliche Intelligenz stellt die Lösung dieser grundlegenden Problemstellungen vor neue Herausforderungen. Auf den ersten Blick scheint es so, als wären grundlegende datenschutzrechtliche Prinzipien unvereinbar mit der Funktionsweise künstlicher Intelligenz. Die Anforderungen und Grundprinzipien der DSGVO sind den

[89] Vgl. Albers/Veit in Wolff/Brink (2020), BeckOK Datenschutzrecht, Art. 6, Rn. 12; Roßnagel (2019), Kein „Verbotsprinzip“ und kein „Verbot mit Erlaubnisvorbehalt“ im Datenschutz-recht: Zur Dogmatik der Datenverarbeitung als Grundrechtseingriff

[90] Vgl. Art. 5 DSGVO

Anforderungen an die Schaffung qualitativ hochwertiger LLMs zunächst konträr. Dies führt dazu, dass berechtigterweise vermehrt die Frage aufgeworfen wird, ob sich die DSGVO zur Regulierung von künstlicher Intelligenz eignet oder deren angepriesene Technologieneutralität an Ihre Grenzen stößt und die Verordnung einer Revision bedarf.[91] Nachfolgend soll die (Un-)Vereinbarkeit der Grundsätze und Anforderungen der DSGVO mit dem Training von LLMs dargestellt werden.

3.1 Grundsatz der Rechtmäßigkeit

Nach dem in Art. 5 Abs. 1 lit. a Var. 1 DSGVO niedergeschriebenen Grundsatz der Rechtmäßigkeit muss für eine rechtmäßige Verarbeitung personenbezogener Daten zumindest einer der Erlaubnistatbestände des Art. 6 Abs. 1 DSGVO und gegebenenfalls mindestens eine der Voraussetzungen des Art. 9 Abs. 2 DSGVO gegeben sein. Die wohl prominentesten und im Kontext mit dem Training von LLMs relevantesten Rechtsgrundlagen für die Verarbeitung personenbezogener Daten sind die Einwilligung der betroffenen Person gem. Art. 6 Abs. 1 lit. a DSGVO sowie die Wahrung berechtigter Interessen der verantwortlichen Stelle und/oder eines Dritten gem. Art. 6 Abs. 1 lit. f DSGVO. Welche Rechtsgrundlage nach Art. 6 Abs. 1 DSGVO und ggf. nach Art. 9 Abs. 2 DSGVO zur Anwendung kommt, ist unter Berücksichtigung der Umstände des jeweiligen Einzelfalls zu bestimmen. Zu berücksichtigende Umstände sind dabei insbesondere die konkreten Zwecke der jeweiligen Datenverarbeitung, die Kategorien der zu verarbeitenden personenbezogenen Daten sowie das Risiko und die Intensität des Eingriffs in die Rechte und Freiheiten der betroffenen Personen. Folglich sollte die datenschutzrechtliche Zulässigkeit und insb. das Vorliegen einer wirksamen datenschutzrechtlichen Rechtsgrundlage für die verschiedenen Prozesse im Rahmen des Lebenszyklus eines LLMs (Trainingsdatensammlung und -aufbereitung, Pre-Training, Fine Tuning und Inferenz) jeweils separat bewertet werden.[92] Für die Zwecke der vorliegenden Arbeit, beschränken sich die nachfolgenden Ausführungen auf die Prozesse im Rahmen des Trainings von LLMs.

[91] Vgl. Spies (2024), Wann passt die DS-GVO auf KI?; Solove (2024), S. 6; Bitkom/DFKI (2017), Entscheidungsunterstützung mit Künstlicher Intelligenz – Wirtschaftliche Bedeutung, gesellschaftliche Herausforderungen, menschliche Verantwortung, S. 10

[92] Vgl. European Data Protection Board (2024), Report of the Work Undertaken by the ChatGPT Taskforce, S. 6

Sofern man den Personenbezug von Trainingsdaten bejaht (vgl. hierzu Kapitel 4.1 ff.), ist bereits für die, dem Training vorgelagerte, Sammlung bzw. Erhebung der Trainingsdaten und deren Aufbereitung (s. hierzu Kapitel 2.3.1 und Kapitel 2.3.2), das Vorhandensein einer wirksamen datenschutzrechtlichen Rechtsgrundlage erforderlich. Je nach Umfang und/oder Quelle des Trainingsdatensatzes, ist denkbar, dass dieser sogar sog. besondere Kategorien personenbezogener Daten gem. Art. 9 Abs. 1 DSGVO (u. a. Informationen über die rassische und ethnische Herkunft, Gesundheitsdaten etc.) beinhalten kann, für welche besonders hohe datenschutzrechtliche Anforderungen gelten (vgl. hierzu Kapitel 4.3). In der Regel ist vorab nicht ersichtlich, ob bzw. welche Arten potenziell personenbezogener Daten in einem Trainingsdatensatz enthalten sind. Allerdings ist wohl insbesondere bei qualitativ hochwertigen und umfangreichen Trainingsdatensätzen davon auszugehen, dass diese Informationen enthalten, die sich auf natürliche, identifizierbare Personen beziehen.[93] Es gilt: je umfangreicher ein Trainingsdatensatz, desto wahrscheinlicher ist es, dass dieser personenbezogene Daten enthält.[94] Wenn man also annimmt, dass die einzelnen Prozesse im Rahmen des LLM-Trainings jeweils Verarbeitungen personenbezogener Daten darstellen, müssen die verarbeitenden Stellen jeweils eine Rechtsgrundlage nach Art. 6 Abs. 1 DSGVO und ggf. Art 9 Abs. 2 DSGVO vorweisen können. Allein der Umstand, dass die Trainingsdatensätze häufig aus öffentlich verfügbaren Daten bestehen, die bspw. von Social Media-Plattformen und Webseiten gecrawlt werden, führt nicht dazu – wie es von datenschutzrechtlichen Laien gerne angenommen wird – dass datenschutzrechtliche Pflichten keine Anwendung finden würden.

3.1.1 Art. 6 Abs. 1 lit. a DSGVO

Rechtsgrundlage für die Datenverarbeitungen im Rahmen des Trainings-Workflows (vgl. hierzu Kapitel 2.3 ff.) könnte jeweils die Einwilligung nach Art. 6 Abs. 1 lit. a DSGVO darstellen. Betrachtet man die Anforderungen an die Einholung wirksamer Einwilligungen, wird jedoch schnell klar, dass diese Rechtsgrundlage in den aller meisten Fällen für die genannten Sachverhalte ungeeignet ist. Eine wirksame Einwilligung muss

[93] Vgl. Longpre et al. (2023), A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality,& toxicity, S. 5

[94] Vgl. Dieker (2024), Datenschutzrechtliche Zulässigkeit der Trainingsdatensammlung: Wie Scraping und Crawling zur KI-Entwicklung eingesetzt werden können, S. 133

insbesondere freiwillig und informiert für konkrete, vorab definierte Zwecke erfolgen.[95] Da die verarbeitenden Stellen zum Zeitpunkt der Trainingsdatenerhebung in der Regel weder Kenntnis darüber haben werden, welche konkreten personenbezogene Daten in den Trainingsdaten enthalten sind, noch wer die betroffenen Personen sind (insbesondere bei einer Erhebung über öffentlich-zugängliche Quellen), scheint das Einholen informierter Einwilligungen nach Art. 6 Abs. 1 lit. a DSGVO unmöglich.

Selbst wenn die betroffenen Personen bekannt sind und die hohen Transparenzanforderungen (s. hierzu Kapitel 3.3) erfüllt werden könnten, wäre die Einwilligung als Rechtsgrundlage aufgrund des Rechts der betroffenen Personen auf jederzeitigen Widerruf ohne die Angabe von Gründen nach Art. 7 Abs. 3 DSGVO, aus Sicht der verarbeitenden Stelle nicht vorzugswürdig. Das Widerrufsrecht wäre zum aktuellen Stand der Technik kaum effektiv zu gewährleisten. Im Falle eines Widerrufs müssten die Informationen über die jeweilige betroffene Person zunächst in den Trainingsdaten ausfindig gemacht werden. Dies wird in der Regel nicht ohne weiteres möglich sein, insbesondere, wenn im Rahmen der Trainingsdatenaufbereitung bereits Schutzmaßnahmen, wie Differential Privacy-Ansätze, umgesetzt wurden (vgl. hierzu Kapitel 5). Sofern keine wirksamen Schutzmaßnahmen während der Trainingsdatenaufbereitung angewendet werden, wäre eine Filterung der betreffenden personenbezogenen Informationen aus dem Trainingsdatensatz theoretisch denkbar, in der Praxis wäre dies jedoch mit einem unverhältnismäßig hohem technischen sowie finanziellen Aufwand verbunden.[96] Selbst wenn eine Identifizierung der betreffenden Informationen und deren Löschung aus den Trainingsdaten möglich wären, hätte dies keine Auswirkung auf das bereits trainierte LLM und würde nichts an den potenziellen Auswirkungen für die Rechte und Freiheiten der Datensubjekte ändern. Üblicherweise erfordert das Entfernen bestimmter Datenpunkte aus einem LLM, dass das jeweilige LLM vollständig neu trainiert werden muss, was vor allem ressourcenintensiv und mühsam ist

[95] Vgl. Art. 7 DSGVO; ErwGr 32, 42 und 43 DSGVO

[96] Vgl. Longpre et al. (2023), A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality,& toxicity, S. 5; Laurençon et al. (2023), The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset, S. 7; Xu et al. (2021), Privacy-Preserving Machine Learning: Methods, Challenges, and Directions, S.28; El Mestari et al. (2024), Preserving Data Privacy in Machine Learning Systems, S. 15; Carlini et al. (2021), Extracting Training Data from Large Language Models, S. 3

(umso mehr, je größer das Modell ist).[97] Zwar kommen aktuelle Studien zu algorithmischen Methoden, die eine effiziente Löschung einzelner Datenpunkte aus ML-Modellen ermöglichen, ohne dass im Anschluss ein erneutes Training des betreffenden Modells erforderlich ist (sog. selective forgetting oder scrubbing), zu vielversprechenden Ergebnissen, allerdings ist die Forschung noch nicht soweit, dass solche Löschalgorithmen effizient auf komplexe Modelle, wie Neuronale Netze und generative Modelle, angewendet werden können.[98]

Vor diesem Hintergrund erscheint die Einwilligung nach Art. 6 Abs. 1 lit. a DSGVO nur in äußersten Ausnahmefällen als geeignete Rechtsgrundlage für Datenverarbeitungen im Rahmen des LLM-Trainings.

3.1.2 Art. 6 Abs. 1 lit. f DSGVO

Für die Datenverarbeitungen im Rahmen des Trainings von LLMs kommt vielmehr die Rechtsgrundlage des Art. 6 Abs. 1 lit. f DSGVO infrage.[99] Nach dieser Norm ist eine Datenverarbeitung rechtmäßig, wenn diese zur Wahrung berechtigter Interessen der verantwortlichen Stelle oder eines Dritten erforderlich ist und die Interessen, Grundrechte und -freiheiten der betroffenen Personen nicht überwiegen. Voraussetzung für eine Datenverarbeitung auf Basis von Art. 6 Abs. 1 lit. f DSGVO ist die sog. Drei-Stufen-Prüfung: (1.) Prüfung, ob berechtigte Interessen der verantwortlichen Stelle oder eines Dritten vorliegen; (2.) Prüfung, ob die Datenverarbeitung zur Wahrung der berechtigten Interessen geeignet und erforderlich ist; und (3.) Durchführung und Dokumentation einer einzelfallbezogenen Abwägung der berechtigten Interessen des Verantwortlichen und/oder Dritten mit den Interessen, Grundrechten und Grundfreiheiten der betroffenen Personen. Um die Datenverarbeitung zulässigerweise auf Art. 6 Abs. 1 lit. f DSGVO stützen zu können, muss man insbesondere im Rahmen der Interessenabwägung zu dem Ergebnis kommen, dass die Interessen der betroffenen Person die berechtigten Interessen

[97] Vgl. Feretzakis et al. (2024), Privacy-Preserving Techniques in Generative AI and Large Language Models: A Narrative Review, S. 12

[98] Vgl. Feretzakis et al. (2024), Privacy-Preserving Techniques in Generative AI and Large Language Models: A Narrative Review, S. 12

[99] Vgl. EDPB (2024), Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models

der verantwortlichen Stelle und/oder des Dritten nicht überwiegen. Der Begriff des berechtigten Interesses ist unbestimmt und weit zu verstehen.[100] Es kommen alle wirtschaftlichen, ideellen oder rechtlichen Interessen in Betracht, die tatsächlich vorliegen und nach dem Unionsrecht und dem Recht des jeweiligen Mitgliedstaats rechtmäßig sind.[101] Als berechtigte Interessen für Datenverarbeitungen zum Zweck des Trainings von LLMs sind bspw. unternehmerische und kommerzielle Interessen an der Entwicklung und Bereitstellung von Diensten, die auf LLMs basieren, denkbar. Ebenso denkbar ist, dass die Verarbeitung von personenbezogenen Daten für die Wahrung der berechtigten Interessen im Rahmen des LLM-Trainings erforderlich ist, sofern angemessene Maßnahmen zur Minimierung personenbezogener Daten ausgeschöpft wurden (vgl. hierzu Kapitel 2.3 sowie Kapitel 5 und 6). Es dürften keine milderen und gleich effektiven Mittel denkbar sein.[102] D. h. Mittel, deren Eingriffsintensität für die Rechte und Freiheiten der betroffenen Personen geringer sind, die keine unverhältnismäßigen Kosten verursachen und dennoch dieselbe Leistungsfähigkeit des LLMs gewährleisten.

Im Rahmen der zu dokumentierenden Interessenabwägung sind alle Umstände des konkreten Einzelfalls sowie die konkreten Auswirkungen auf die Rechte und Freiheiten der betroffenen Personen zu berücksichtigen (vgl. hierzu Kapitel 7).[103] Es ist fraglich, inwieweit alle Umstände sowie die konkreten Auswirkungen einer Verarbeitung personenbezogener Daten für das Training eines LLMs im Rahmen der Interessenabwägung vollständig überblickt und berücksichtigt werden können. Die Schwere der Auswirkungen einer Datenverarbeitung für die Rechte und Freiheiten der betroffenen Personen hängen u. a. von der Kritikalität der zu verarbeitenden personenbezogenen Daten sowie auch von einer etwaigen besonderen Schutzbedürftigkeit der betroffenen Personen ab, z. B. wenn Daten von Kindern oder Senioren verarbeitet werden. Da in der Regel nicht bekannt ist, welche konkreten personenbezogenen Daten in

[100] Vgl. ErwGr 47 S. 2, 6, 7 DGSVO

[101] Vgl. Buchner/Petri in Kühling/Buchner (2020), DS-GVO BDSG, Art. 6, Rn. 146a; BVerwG Urteil v. 27.03.2019 – 6 C 2/18, NJW 2019, 3556, Rn. 47; Artikel-29-Datenschutzruppe (2014), WP 217, S. 31

[102] Vgl. Buchner/Petri in Kühling/Buchner (2020), DS-GVO BDSG, Art. 6, Rn. 147a; EuGH, Urteil v. 11.12.2019 – C/708/18, ZD 2020, 148, Rn. 46

[103] Vgl. DSK (2021), Orientierungshilfe der Aufsichtsbehörden für Anbieter von Telemedien, S. 31

einem Trainingsdatensatz enthalten sind, könnten die Quellen der Trainingsdaten sowie der geplante Anwendungsbereich des jeweiligen LLMs Anhaltspunkte für eine Bewertung des Risikos für die betroffenen Personen im Rahmen einer Interessenabwägung bieten. So ist bspw. bei Trainingsdatensätzen, die für das Training eines LLMs verwendet werden, welches im Gesundheitsbereich eingesetzt werden soll, die Wahrscheinlichkeit höher, dass sensible Gesundheitsdaten in den Trainingsdaten enthalten sind.[104] Trainingsdatensätze, die für das Training von LLMs verwendet werden, die auf einen allgemeineren Anwendungsbereich ausgerichtet sind (bspw. der Chatbot ChatGPT), enthalten in der Regel hauptsächlich weniger sensible personenbezogene Daten, wie Namen und E-Mail-Adressen, die oft aus öffentlich verfügbaren Quellen stammen (z. B. Nachrichten, Webseiten oder Social Media-Plattformen).[105]

Zu den Umständen, die im Rahmen einer Interessenabwägung nach Art. 6 Abs. 1 lit. f DSGVO zu berücksichtigen sind, gehören insbesondere auch die vernünftigen Erwartungen der betroffenen Personen. Das bedeutet, eine Interessenabwägung fällt eher zu Gunsten der verantwortlichen Stelle aus, wenn die betroffenen Personen zum Zeitpunkt der Datenerhebung vernünftigerweise absehen können, dass möglicherweise eine Verarbeitung ihrer personenbezogenen Daten für die konkreten Zwecke erfolgen wird.[106] Dabei sind nicht die Erwartungen der tatsächlich betroffenen Person ausschlaggebend.[107] Vielmehr muss auf die „vernünftigen Erwartungen eines objektiven Dritten in der Rolle der betroffenen Person“[108] abgestellt werden. Sofern diese Absehbarkeit für die betroffenen Personen nicht gegeben ist, könnte dies einen gewichtigen Anhaltspunkt dafür darstellen, dass die Interessen oder Grundrechte und -

[104] Vgl. Thomas et al. (2020), Investigating the Impact of Pre-trained Word Embeddings on Memorization in Neural Networks; Shokri et al. (2017), Membership Inference Attacks Against Machine Learning Models

[105] Vgl. Carlini et al. (2021), Extracting Training Data from Large Language Models, S. 10; Longpre et al. (2023), A pretrainer’s guide to training data: Measuring the effects of data age, domain coverage, quality,& toxicity, S. 5

[106] Vgl. ErwGr 47 DSGVO

[107] Vgl. Robrahn/Bremert (2018), Interessenkonflikte im Datenschutzrecht: Rechtfertigung der Verarbeitung personenbezogener Daten über eine Abwägung nach Art. 6 Abs. 1 lit. f DS-GVO, S. 294; Schulz in Gola (2018), Datenschutz-Grundverordnung – Bundesdatenschutzgesetz, Art. 6, Rn. 63

[108] Herfurth (2018), Interessenabwägung nach Art. 6 Abs. 1 lit. f DS-GVO, S. 518

freiheiten der betroffenen Person den angeführten berechtigten Interessen für die Datenverarbeitung überwiegen.[109] Wenn man davon ausgeht, dass die betroffenen Personen ihre personenbezogenen Daten selbst wissentlich im Internet veröffentlichen, müssen diese Personen aus objektiver Sicht vernünftigerweise davon ausgehen, dass die Möglichkeit besteht, dass die von ihnen veröffentlichten Daten gecrawlt und gescraped werden können. Fraglich ist, inwieweit davon ausgegangen werden kann, dass die vernünftige Erwartungshaltung der betroffenen Personen auch die Nutzung der veröffentlichten Daten für das Training von LLMs umfasst. Man könnte argumentieren, dass aufgrund aktueller öffentlicher Debatten über künstliche Intelligenz und deren Abhängigkeit von großen Datenmengen, eine breite Erwartungshaltung von Internetnutzern angenommen werden kann. Als Gegenargument ließe sich anführen, dass sich die meisten Internetnutzer nicht über den genauen Umfang und die potenziellen Risiken (z. B. im Zusammenhang mit training data extraction attacks) bewusst sind. Für die Gewichtung des Kriteriums der vernünftigen Erwartungen der betroffenen Personen kommt es erneut insbesondere auf die Quellen der personenbezogenen Trainingsdaten sowie die Umstände der Datenerhebung im Einzelfall an.

Ob Art. 6 Abs. 1 lit. f DSGVO wirksam als Rechtsgrundlage für das Training von LLMs herangezogen werden kann, hängt von den Umständen des jeweiligen Einzelfalls ab. Häufig wird dabei ein gewisses Transparenzdefizit vorliegen, welches einen gesteigerten Argumentationsaufwand im Rahmen der Interessenabwägung sowie ggf. die Implementierung zusätzlicher geeigneter Schutzmaßnahmen erforderlich machen wird.[110]

3.1.3 Widerspruchsrecht nach Art. 21 DSGVO

Sofern man im Rahmen der sog. Drei-Stufen-Prüfung und insbesondere bei der erforderlichen Interessenabwägung zu dem Ergebnis kommt, dass die Voraussetzungen des Art. 6 Abs. 1 lit. f DSGVO vorliegen, stellt sich die nächste Herausforderung. Nach Art. 21 Abs. 1 DSGVO steht betroffenen Personen, deren Daten auf Grundlage von Art. 6 Abs. 1 lit. f DSGVO verarbeitet werden, das Recht zu, jederzeit aus Gründen, die sich aus ihrer besonderen Situation ergeben, Widerspruch gegen die Datenverarbeitung

[109] Vgl. Heberlein in Ehmann/Selmayr (2018), Datenschutz-Grundverordnung, Art. 6, Rn. 28

[110] Vgl. Herfurth (2018), Interessenabwägung nach Art. 6 Abs. 1 lit. f DS-GVO: Nachvollziehbare Ergebnisse anhand von 15 Kriterien mit dem sog. „3x5-Modell“, S. 520

einzulegen. Einen Widerspruch muss eine betroffene Person demnach damit begründen, dass für ihre konkrete Situation eine atypische Konstellation gegeben ist, die ihren Interessen ein besonderes Gewicht verleiht.[111] Dabei muss die betroffene Person konkrete, einzelfallbezogene Umstände vortragen, die eine besondere Schutzwürdigkeit der betroffenen Person begründen (z. B. besondere familiäre Umstände).[112] Sofern ein legitimer Widerspruch nach Art. 21 Abs. 1 DSGVO erfolgt, muss die verantwortliche Stelle die Datenverarbeitung stoppen bzw. die jeweiligen Daten löschen, sofern keine zwingenden schutzwürdigen Gründe für die Datenverarbeitung nachgewiesen werden können oder die Datenverarbeitung nicht der Geltendmachung, Ausübung oder Verteidigung von Rechtsansprüchen dient. Als schutzwürdig werden Gründe angesehen, die vom Unionsrecht oder nationalem Recht des entsprechenden Mitgliedstaats anerkannt sind.[113]

Die europäischen Datenschutzaufsichtsbehörden sowie der Europäische Datenschutzausschuss vertreten den Standpunkt, dass das Widerspruchsrecht nach Art. 21 Abs. 1 DSGVO auch beim Training von LLMs mit personenbezogenen Daten zur Anwendung kommt und gewährleistet werden muss, sofern die Datenverarbeitung auf Art. 6 Abs. 1 lit. f DSGVO gestützt wird.[114] In der praktischen Umsetzung stößt man dabei jedoch auf nicht unerhebliche Hürden. Zunächst setzt die Gewährleistung des Widerspruchsrechts voraus, dass die verantwortliche Stelle die betroffenen Personen über das Bestehen des Widerrufrechts unterrichtet hat. Dies wird bei der Verarbeitung von Trainingsdaten aus öffentlich-zugänglichen Quellen i. d. R. nur indirekt, über öffentlich einsehbare Datenschutzinformationen (z. B. auf der Webseite) möglich sein (vgl. hierzu Kapitel 3.4.1). Für den Fall, dass eine natürliche Person davon Kenntnis erlangt, dass ihre personenbezogenen Informationen für das Training eines LLMs verwendet wurden (bspw. wenn diese während der Inferenz-Phase offengelegt werden) und ihr Widerspruchsrecht aus legitimen Gründen ggü. der verantwortlichen Stelle ausübt, steht

[111] Vgl. Herbst in Kühling/Buchner (2020), DS-GVO BDSG Art. 21, Rn. 15

[112] Vgl. Martini in Paal/Pauly (2021), DS-GVO BDSG, Art. 21, Rn. 30

[113] Vgl. Herbst in Kühling/Buchner (2020), DS-GVO BDSG, Art. 21, Rn. 20

[114] Vgl. EDPB (2024), Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, para. 65

die verantwortliche Stelle vor den im Kapitel 3.1.1 bereits beschrieben Herausforderungen hinsichtlich der Identifizierung der jeweiligen Daten in dem Trainingsdatensatz und deren folgenwirksamen Löschung. Zum aktuellen Stand der Technik wird dies kaum unter angemessenem Aufwand zu bewerkstelligen sein.[115]. Hier sind zu erwartende Fortschritte in der Forschung zu Löschalgorithmen und sog. machine unlearning sowie deren Bewährung in der Praxis abzuwarten.[116] Bis dahin ist zweifelhaft, ob das Widerspruchsrecht aktuell einen tatsächlichen Mehrwert für die Datensubjekte bei der Datenverarbeitung im Rahmen des LLM-Trainings bringt und zudem einen angemessenen Ausgleich zwischen den sich gegenüberstehenden Grundrechtspositionen (insb. Art. 16 vs. Art. 7 und 8 Charta der Grundrechte der Europäischen Union (nachfolgend „GRCh“)) schafft.[117] Die Umsetzung technischer Maßnahmen zur Verhinderung unberechtigter Offenlegung von personenbezogenen Informationen während der Inferenz-Phase (vgl. hierzu Kapitel 5 und 6) erscheint effektiver für den Schutz der Rechte und Freiheiten der betroffenen Personen. Eine weite Auslegung des Art. 21 Abs. 1 Satz 2 DSGVO, nach welcher die aktuell bestehenden technischen Hürden und die zweifelhafte Wirksamkeit bei einer Umsetzung eines Widerrufs als schutzwürdige Gründe angesehen werden könnten, die eine besondere Schutzwürdigkeit des jeweiligen Datensubjekts ggf. überwiegen, könnte in diesem Kontext als zielführender Lösungsweg diskutiert werden. Inwiefern die Gewährleistung des Widerspruchsrechts bei Datenverarbeitungen im Rahmen des LLM-Trainings in der Praxis aus Sicht der betroffenen Personen überhaupt für den Schutz ihrer Grundrechte und -freiheiten relevant ist – insbesondere, wenn angemessene und wirksame Schutzmaßnahmen umgesetzt werden – wird künftige Rechtsprechung ggf. zeigen. Für die Hersteller von LLMs verbleibt eine Rechtsunsicherheit, die sich aus der technischen Umsetzbarkeit der rechtlichen Anforderungen ergibt. Insofern wäre eine weite Auslegung

[115] Vgl. Feretzakis et al. (2024), Privacy-Preserving Techniques in Generative AI and Large Language Models: A Narrative Review, Information, S. 12

[116] Vgl. hierzu auch den von Google im Jahr 2023 ausgeschriebenen Wettbewerb zur Entwicklung eines „unlearn-fähigen“ KI-Modells; https://research.google/blog/announcing-the-first-machine-unlearning-challenge/ (letzter Aufruf: 23.01.2025)

[117] Vgl. Moerel (2025), A perspective: Why data subjects’ rights to LLM training data are not relevant, IAPP Privacy Perspectives, aufrufbar unter: https://iapp.org/news/a/perspective-why-data-subjects-rights-to-llm-training-data-are-not-relevant (letzter Aufruf: 23.01.2025)

des Art. 21 Abs. 1 Satz 2 DSGVO unter Berücksichtigung des aktuellen Stands der Technik durch die Gerichte wünschenswert.

3.1.4 Art. 9 Abs. 2 lit. e DSGVO

Vor allem Trainingsdatensätze aus öffentlich zugänglichen Quellen, wie dem Internet, können grundsätzlich alle möglichen Arten von potenziell personenbezogenen Daten enthalten, auch sog. besondere Kategorien personenbezogener Daten i. S. d. Art. 9 Abs. 1 DSGVO. Für Verarbeitungen von solchen Informationen, die ein besonderes Risikopotential für die betroffenen Personen bergen, ist abgesehen von den Anforderungen des Art. 6 Abs. 1 DSGVO zudem das Vorliegen eines Erlaubnistatbestands nach Art. 9 Abs. 2 DSGVO erforderlich.[118]

Im Kontext mit dem Training von LLMs kommt insbesondere der Erlaubnistatbestand des Art. 9 Abs. 2 lit. e DSGVO in Betracht, nach welchem die Verarbeitung von besonderen Kategorien personenbezogener Daten erlaubt ist, sofern diese von der betroffenen Person offensichtlich öffentlich gemacht wurden. Für die Anwendbarkeit dieses Erlaubnistatbestands ist insbesondere erforderlich, dass die betroffene Person die jeweiligen sensiblen Daten selbst und beabsichtigt öffentlich gemacht hat. Die bloße Tatsache, dass die Daten öffentlich zugänglich sind, ist dabei kein eindeutiger Indikator für eine bewusste Veröffentlichung der Daten durch die betroffene Person.[119] Im Rahmen einer Trainingsdatensammlung via Crawling und Scraping wird es für die datensammelnde Stelle regelmäßig nicht möglich sein, abschließend zu bewerten, ob die jeweiligen Daten von den betroffenen Personen selbst beabsichtigt veröffentlicht wurden und die Wirksamkeitsvoraussetzungen des Art. 9 Abs. 2 lit. e DSGVO vorliegen. In solchen Fällen wird der Art. 9 Abs. 2 lit. e DSGVO als wirksame Rechtsgrundlage eher ausscheiden. In Fällen, in denen die datenverarbeitende Stelle darlegen kann, dass die jeweiligen Daten von den Datensubjekten selbst und beabsichtigt veröffentlicht wurden, kommt Art. 9 Abs. 2 lit. e DSGVO als wirksame Rechtsgrundlage in Betracht.

[118] Vgl. EuGH, Urteil v. 21.12.2023 – C-667/21, Rn. 51

[119] Vgl. European Data Protection Board (2024), Report of the Work Undertaken by the ChatGPT Taskforce, S. 7

3.1.5 Art. 9 Abs. 2 lit. j DSGVO

Auch die Rechtsgrundlage des Art. 9 Abs. 2 lit. j DSGVO könnte eine Verarbeitung von in Trainingsdaten enthaltenen besonderen Kategorien personenbezogener Daten rechtfertigen. Nach dieser Norm kann die Verarbeitung besonderer Kategorien personenbezogener Daten insbesondere für wissenschaftliche oder historische Forschungszwecke oder für statistische Zwecke i. S. d. Art. 89 DSGVO auf Basis eines Unionsrechts oder des Rechts eines Mitgliedsstaats erfolgen. Eine entsprechende nationalrechtliche Norm ist bspw. § 27 Abs. 1 Bundesdatenschutzgesetz (BDSG), nach welcher die Verarbeitung sensibler Daten gerechtfertigt sein kann, wenn diese für wissenschaftliche oder historische Forschungszwecke oder für statistische Zwecke erforderlich ist und die Interessen der verantwortlichen Stelle den Interessen der betroffenen Personen an einem Ausschluss der Verarbeitung erheblich überwiegen. Für die Anwendbarkeit des Art. 9 Abs. 2 lit. j DSGVO kommt es u. a. darauf an, ob das Vorliegen wissenschaftlicher Forschungszwecke oder statistischer Zwecke beim Training von LLMs angenommen werden kann.

Was konkret unter wissenschaftlichen Forschungszwecken zu verstehen ist, wird in der DSGVO nicht definiert. Anhaltspunkte können der Rechtsprechung zu Art. 5 Abs. 3 Grundgesetz (GG) entnommen werden, welcher an Art. 13 GRCh angelehnt ist und diesem inhaltlich entspricht.[120] Demnach ist wissenschaftliche Forschung ein Prozess zum Auffinden von Erkenntnissen in methodischer, systematischer und nachprüfbarer Weise.[121] Auch nach ErwGr. 159 Satz 2 DSGVO ist der Begriff der wissenschaftlichen Forschungszwecke weit auszulegen. Demnach schließen wissenschaftliche Forschungszwecke z. B. die technologische Entwicklung, die Grundlagenforschung, die angewandte Forschung und die privat finanzierte Forschung ein. Folglich kann auch Auftragsforschung und zweckgebundene Industrieforschung wissenschaftliche Forschung i. S. d. Art. 89 DSGVO sein.[122] Insofern ist denkbar, dass auch das Training LLMs,

[120] Vgl. BVerfG Urteil v. 29.05.1973 – 1 BvR 424/71, 1 BvR 325/72, NJW 1973, 1176

[121] Vgl. BVerfG Urteil v. 29.05.1973 – 1 BvR 424/71, 1 BvR 325/72, NJW 1973, 1176; Weichert (2020), Die Forschungsprivilegierung in der DS-GVO, ZD 2020, 18, 19

[122] Vgl. Niemann/Kevekordes (2020), Machine Learning und Datenschutz (Teil 2): Sensitive Daten und Betroffenenrechte, S. 180

welches zu wissenschaftlichen Forschungszwecken eingesetzt werden soll und nicht nur rein individuellen, profitorientierten Zwecken dienen soll (bspw. LLMs, die in der medizinischen Forschung eingesetzt werden sollen), unter den Erlaubnistatbestand des Art. 9 Abs. 2 lit. j DSGVO i. V. m. § 27 Abs. 1 BDSG fallen kann.[123]

Auch der Begriff der privilegierten statistischen Zwecke ist nach ErwGr. 162 DSGVO weit zu verstehen. Demnach ist unter dem Begriff „jeder für die Durchführung statistischer Untersuchungen und die Erstellung statistischer Ergebnisse erforderliche Vorgang der Erhebung und Verarbeitung personenbezogener Daten zu verstehen".[124] Das Training von LLMs könnte als für statistische Untersuchungen und die Erstellung statistischer Ergebnisse erforderlicher Vorgang verstanden werden. Das LLM-Training basiert auf statistischen Verfahren, um Muster in großen Datenmengen zu erkennen und darauf basierende Modelle zu generieren, die als statistische Ergebnisse angesehen werden könnten. Allerdings dürfen die Ergebnisse solcher Datenverarbeitungen nach ErwGr. 162 DSGVO keine personenbezogenen Daten, sondern nur aggregierte Daten enthalten und nicht für Maßnahmen oder Entscheidungen gegenüber einzelnen natürlichen Personen verwendet werden.[125] Im Rahmen der sog. Tokenisierung werden die Trainingsdaten während des Trainingsprozesses zwar in „abstrakte mathematische Repräsentationen"[126] überführt, die sich nicht direkt auf Einzelpersonen beziehen, problematisch ist jedoch, dass LLMs unter bestimmten Umständen personenbezogene Daten reproduzieren können (z. B. im Rahmen von training data extraction attacks). Denkbar wäre jedoch, einer Reproduktion personenbezogener Daten durch entsprechende Sicherheitsmaßnahmen entgegenzuwirken (vgl. hierzu Kapitel 5 und 6). Doch auch in Fällen, in denen gewährleistet werden könnte, dass keine personenbezogenen Daten im Rahmen von training data extraction attacks offengelegt bzw. reproduziert werden, lässt sich das LLM-Training nicht ohne Weiteres unter den Begriff der statistischen Zwecke i. S. d. Art. 89 DSGVO subsumieren, da LLMs häufig für Anwendungen genutzt werden,

[123] Vgl. Niemann/Kevekordes (2020), Machine Learning und Datenschutz (Teil 2): Sensitive Daten und Betroffenenrechte, S. 180

[124] ErwGr. 162 S. 3 DSGVO

[125] ErwGr. 162, S. 5 DSGVO

[126] Vgl. Hamburgischer Beauftragter für Datenschutz und Informationsfreiheit (2024), Diskussionspapier: Large Language Models und personenbezogene Daten v. 15. Juli 2024, S. 4

die eine Benutzerinteraktion voraussetzen und aus denen ggf. indirekt oder direkt Maßnahmen oder Entscheidungen gegenüber einzelnen natürlichen Personen abgeleitet werden können.

Es sind Szenarien denkbar, für welche der Art. 9 Abs. 2 lit. j DSGVO i. V. m. § 27 Abs. 1 BDSG als wirksame Rechtsgrundlage für eine Datenverarbeitung im Rahmen des LLM-Trainings herangezogen werden könnte. Zu berücksichtigen ist in solchen Fällen jedoch, dass im Rahmen der erforderlichen Interessenabwägung ein noch strengerer Maßstab als bei der Interessenabwägung nach Art. 6 Abs. 1 lit. f DSGVO anzuwenden ist.[127] Anzumerken ist an dieser Stelle auch, dass das Training von LLMs, die zu rein kommerziellen Zwecken eingesetzt werden sollen, nicht auf den Privilegierungstatbestand des Art. 9 Abs. 2 lit. j DSGVO gestützt werden können.[128]

3.1.6 Zwischenfazit

Auch wenn die Erfüllung der Anforderungen der Art. 6 Abs. 1 lit. f sowie Art. 9 Abs. 2 lit. e und lit. j DSGVO bei der Verarbeitung personenbezogener Daten im Rahmen des Trainings von LLMs unter bestimmten Umständen grundsätzlich denkbar ist, bleiben dennoch datenschutzrechtliche Unsicherheiten bestehen.[129] Ob eine Datenverarbeitung im Rahmen des Trainings eines LLMs auf Basis von Art. 6 Abs. 1 lit. f und ggf. Art. 9 Abs. 2 lit. e oder lit. j DSGVO rechtmäßig erfolgen kann, ist von den konkreten Umständen des jeweiligen Einzelfalls abhängig (s. hierzu auch Kapitel 7) und kann nicht pauschal beantwortet werden.

3.2 Grundsatz der Zweckbindung

Eines der Kernprinzipien der DSGVO ist der in Art. 5 Abs. 1 lit. b DSGVO verankerte Grundsatz der Zweckbindung. Demnach dürfen personenbezogene Daten nur für „festgelegte, eindeutige und legitime Zwecke“ verarbeitet werden. Es dürfen nur solche Daten verarbeitet werden, die für die Erreichung des jeweiligen Zwecks erforderlich sind. Auch der zulässige Umfang der Datenverarbeitung (insb. die Speicherdauer) bemisst sich

[127] Vgl. § 27 Abs. 1 S. 1 BDSG

[128] Vgl. Vogel (2022), Künstliche Intelligenz und Datenschutz, S. 156

[129] Vgl. European Data Protection Board (2024), Report of the Work Undertaken by the ChatGPT Taskforce, S. 7

nach der Erforderlichkeit zur Erreichung des jeweiligen Zwecks. Ferner dürfen personenbezogene Daten gemäß Art. 5 Abs. 1 lit. b DSGVO nicht ohne weiteres zu anderen Zwecken als die zuvor festgelegten Zwecke weiterverarbeitet werden.[130] Hierdurch soll verhindert werden, dass zu bestimmten Zwecken erhobene Daten willkürlich weiterverarbeitet werden, ohne dass die betroffenen Personen darauf Einfluss nehmen können. Die Zweckbindung ist ein grundlegendes Instrument zum Schutz der Rechte und Freiheiten natürlicher Personen. Die Beurteilung der Zulässigkeit einer Datenverarbeitung erfordert insbesondere eine Beurteilung der Erforderlichkeit der Datenverarbeitung für die Erreichung des jeweiligen festgelegten Zweckes. Folglich ist die Definition der Verarbeitungszwecke der Dreh- und Angelpunkt einer zulässigen Datenverarbeitung.

Eine gewisse rechtliche Unsicherheit für Datenverarbeitungen im Rahmen des Trainings von LLMs ergibt sich aus der Fragestellung, wie eindeutig bzw. wie präzise die Zwecke der Datenverarbeitung definiert und benannt werden müssen. Häufig wird beim Training von LLMs noch nicht abschließend klar sein, für welche genauen Zwecke das LLM künftig eingesetzt werden soll, insbesondere bei sog. General Purpose AI. Reicht in solchen Fällen die Benennung eines übergeordneten Zwecks, wie bspw. „Training / Entwicklung eines Large Language Models“? In der DSGVO finden sich keine konkreten Aussagen dazu, wie präzise die Zwecke einer Datenverarbeitung definiert und kommuniziert werden müssen.[131]

Auch wenn generische Zweckbeschreibungen, wie z. B. „Optimierung der Nutzererfahrung“, „Marketingzwecke“ oder „Gewährleistung der IT-Sicherheit“ in der Praxis gang und gäbe sind, erfüllen solche generischen Zweckbeschreibungen nach Auffassung des Europäischen Datenschutzausschusses regelmäßig nicht die datenschutzrechtlichen Anforderungen.[132] Wie detailliert ein Verarbeitungszweck angegeben werden muss, hängt jedoch vom jeweiligen Kontext der Datenerhebung und

[130] Vgl. Art. 5 Abs. 1 lit. b DSGVO

[131] Vgl. Schantz in Wolff/Brink/v. Ungern-Sternberg, BeckOK Datenschutzrecht (2021), 51. Edition, Art. 5, Rn. 15

[132] Vgl. Artikel-29-Datenschutzgruppe (2013), Opinion 03/2013 (WP 203), S. 16

der Art der jeweiligen Daten ab.[133] Grundsätzlich müssen die Zwecke einer Datenverarbeitung so präzise definiert werden, dass die betroffenen Personen abschätzen können, welche Auswirkungen und Risiken für sie mit der jeweiligen Datenverarbeitung verbunden sind.[134] Die Zwecke müssen also umso präziser definiert werden je größer die Risiken für die Rechte und Freiheiten der betroffenen Personen sind. Folglich kommt es für die Beurteilung, ob die Verarbeitungszwecke ausreichend präzise definiert sind, auf den jeweiligen Einzelfall an.[135]

Bei der Verarbeitung personenbezogener Daten im Rahmen des Trainings von LLMs kommt es folglich insbesondere auf die Art der in den Trainingsdaten enthaltenen personenbezogenen Daten sowie auf die umgesetzten Sicherheitsmaßnahmen an.[136] In Sachverhalten, in denen mit ausreichender Wahrscheinlichkeit davon ausgegangen werden kann, dass die verwendeten Trainingsdaten keine besonders risikobehafteten Arten personenbezogener Daten beinhalten (bspw. weil die Trainingsdaten aus entsprechenden Quellen stammen) und angemessene Sicherheitsmaßnahmen umgesetzt sind (vgl. hierzu Kapitel 5 und 6), ist also denkbar, dass der Rückgriff auf den übergeordneten Zweck des Trainings oder der Entwicklung eines LLMs zulässig sein könnte.

Der Zweck „Training / Entwicklung von LLMs“ wird in der Regel jedoch nicht dem initialen Zweck der Datenerhebung bzw. Datenveröffentlichung entsprechen. Wenn die Datenverarbeitung zum Sekundärzweck des LLM-Trainings auf eine separate Rechtsgrundlage gestützt wird (bspw. Art. 6 Abs. 1 lit. f DSGVO), ist der Umstand, dass die in den Trainingsdatensätzen enthaltenen personenbezogenen Daten nicht zum Zweck des LLM-Trainings erhoben bzw. veröffentlicht wurden, unschädlich. Ein Rückgriff auf die ursprüngliche Rechtsgrundlage der Datenerhebung bzw. -veröffentlichung wäre hingegen nur möglich, wenn

[133] Vgl. Artikel-29-Datenschutzgruppe (2013), Opinion 03/2013 (WP 203), S. 16

[134] Vgl. Schantz in Wolff/Brink/v. Ungern-Sternberg, BeckOK Datenschutzrecht (2021), 51. Edition, Art. 5, Rn. 15

[135] Vgl. Artikel-29-Datenschutzgruppe (2013), Opinion 03/2013 (WP 203), S. 16

[136] Vgl. Niemann/Kevekordes (2020), Machine Learning und Datenschutz (Teil 1): Grundsätzliche datenschutzrechtliche Zulässigkeit, S. 21

a. es sich bei den Sekundärzwecken um im öffentlichen Interesse liegende Archivzwecke, wissenschaftliche oder historische Forschungszwecke oder statistische Zwecke handelt; oder

b. die Weiterverarbeitung mit den ursprünglichen Zwecken vereinbar ist;

c. die betroffenen Personen hierin eingewilligt haben; oder

d. eine entsprechende unionsrechtliche oder nationale gesetzliche Erlaubnis vorliegt.[137]

Für die Verarbeitung von Trainingsdaten im Rahmen des Trainings von kommerziellen LLMs wird in der Praxis regelmäßig Variante b. relevant sein.[138] Art. 6 Abs. 4 DSGVO fordert dabei die Dokumentation eines Kompatibilitätstest, in dessen Rahmen insbesondere

- die Verbindung der Sekundärzwecke mit den ursprünglichen Zwecken;
- die Umstände der Datenerhebung;
- das Verhältnis zwischen verantwortlicher Stelle und betroffener Personen;
- die Art der personenbezogenen Daten;
- potenzielle Folgen der Weiterverarbeitung für die betroffenen Personen;
- sowie das Vorhandensein geeigneter Garantien zum Schutz der Rechte und Freiheiten der betroffenen Personen (z. B. Verschlüsselung oder Pseudonymisierung) zu berücksichtigen sind.

Ergibt der einzelfallbezogene Kompatibilitätstest, dass die Sekundärzwecke mit den Zwecken der initialen Datenerhebung bzw. -veröffentlichung zu vereinbaren sind und sich aus der Weiterverarbeitung keine besonderen Risiken für die Rechte und Freiheiten der betroffenen Personen ergeben, lässt sich die Weiterverarbeitung auf Art. 6 Abs. 4 DSGVO stützen.[139]

[137] Vgl. Art. 5 Abs. 1 lit. b DSGVO; Art. 6 Abs. 4 DSGVO; ErwGr. 50 DSGVO

[138] Vgl. Vogel (2022), Künstliche Intelligenz und Datenschutz, S. 156

[139] Vgl. Vogel (2022), Künstliche Intelligenz und Datenschutz, S. 157

Auch wenn man bei der Vereinbarung des Zweckbindungsgrundsatz gemäß Art. 5 Abs. 1 lit. b DSGVO mit den Charakteristika von KI-Systemen auf gewisse Hürden stößt, sollte die datenschutzrechtliche Zulässigkeit des Trainings eines LLMs nur in solchen Fällen an dem Zweckbindungsgrundsatz scheitern, in denen besonders risikobehaftete personenbezogene Daten betroffen sind.

3.3 Grundsatz der Verarbeitung nach Treu und Glauben

Der Grundsatz von Treu und Glauben gemäß Art. 5 Abs. 1 lit. a Var. 2 DSGVO verbietet, dass eine Verarbeitung personenbezogener Daten in einer Weise erfolgt, die für die betroffenen Personen ungerechtfertigt nachteilig, unrechtmäßig, diskriminierend, unerwartet oder irreführend ist.[140] Dieser Grundsatz basiert auf der gleichlautenden Formulierung des Grundrechts auf Schutz personenbezogener Daten gemäß Art. 8 Abs. 2 Satz 1 Var. 1 GRCh. Der Begriff „Treu und Glauben" ist dabei nicht nach dem deutschen Rechtsverständnis bzw. im Sinne des § 242 Bürgerliches Gesetzbuch (BGB) zu verstehen.[141] Bei dem deutschen Rechtsverständnis des Grundsatzes von Treu und Glauben handelt es sich um ein Instrument zur Rechtsfortbildung, um gesetzlich nicht geregelte Streitfälle zu einem sachgerechten Ausgleich zu bringen.[142] Demgegenüber regelt der Grundsatz von Treu und Glauben nach dem unionsrechtlichen Rechtsverständnis eine „faire" Verarbeitung personenbezogener Daten. Insofern kann die deutsche Sprachfassung des Art. 5 Abs. 1 lit. a Var. 2 DSGVO missverstanden werden. Zutreffender wäre vielmehr die Anwendung des Begriffs nach der englischen Sprachfassung („Personal data shall be processed […] fairly […]").[143] Ziel dieses Verarbeitungsgrundsatzes ist es u. a., Datenverarbeitungen zu verhindern, die das Vertrauen oder Fehlvorstellungen der betroffenen Personen ausnutzen und nicht

[140] Vgl. Schubert in Münchner Kommentar zum Bürgerlichen Gesetzbuch (2022), BGB § 242 Leistung nach Treu und Glauben, Rn. 103; Vgl. Vogel (2022), Künstliche Intelligenz und Datenschutz, S. 104,

[141] Vgl. Heberlein in Ehmann/Selmayr (2024), Datenschutz-Grundverordnung, Verarbeitung nach Treu und Glauben, Rn. 13

[142] Vgl. Schubert in Münchner Kommentar zum Bürgerlichen Gesetzbuch (2022), BGB § 242 Leistung nach Treu und Glauben, Rn. 103; Vgl. Vogel (2022), Künstliche Intelligenz und Datenschutz, S. 104

[143] Vgl. Schubert in Münchner Kommentar zum Bürgerlichen Gesetzbuch (2022), BGB § 242 Leistung nach Treu und Glauben, Rn. 103; Vgl. Vogel (2022), Künstliche Intelligenz und Datenschutz, S. 104

unmittelbar gegen andere datenschutzrechtliche Pflichten verstoßen.[144] Der Grundsatz von Treu und Glauben soll die Interessen und Rechte der betroffenen Personen gewährleisten und sie vor intransparenten und diskriminierenden Datenverarbeitungen schützen. Der Grundsatz von Treu und Glauben kann ferner auch als „Scharniernorm" verstanden werden, über welche Wertungen und Fairnessgebote anderer Rechtsgebiete in der DSGVO berücksichtigt werden können, sofern diese in unmittelbarem Zusammenhang mit der jeweiligen Datenverarbeitung stehen.[145]

Aus dem Grundsatz von Treu und Glauben könnte ein grundsätzliches Verbot von heimlichen Verarbeitungen personenbezogener Daten abgeleitet werden.[146] Da betroffene Personen bei einer Verarbeitung ihrer Daten für das LLM-Training in der Regel nicht wissen, dass eine Verarbeitung ihrer Daten zum vorgenannten Zweck erfolgt, ist fraglich, ob eine solche Verarbeitung mit dem Grundsatz von Treu und Glauben zu vereinbaren ist. Eine „heimliche" Datenverarbeitung bedarf einer besonderen Rechtfertigung, welche sich ggf. aus Art. 14 Abs. 5 lit. b DSGVO ergeben kann, wenn die Bereitstellung transparenter Informationen für die betroffenen Personen unmöglich oder mit einem unverhältnismäßigen Aufwand verbunden ist. Demnach ist die Zulässigkeit einer heimlichen Verarbeitung von Trainingsdaten unter bestimmten Umständen denkbar (vgl. hierzu Kapitel 3.4.1). Abgesehen von der im nachfolgenden Unterkapitel dargestellten Problematik des Transparenzdefizits, ergeben sich in der Praxis jedoch keine offensichtlichen Problemstellungen hinsichtlich der Vereinbarkeit einer Datenverarbeitung im Rahmen des LLM-Trainings und dem Grundsatz von Treu und Glauben. Solche Problemstellungen ergeben sich vielmehr im Rahmen von Datenverarbeitungen während der Inferenz-Phase. So ist z. B. denkbar, dass eine Datenverarbeitung im Rahmen des Betriebs des KI-Systems COMPAS, das US-amerikanische Richter bei der Bewertung der Rückfallwahrscheinlichkeit von Straftätern

[144] Vgl. Schantz in Wolff/Brink/v. Ungern-Sternberg (2021), BeckOK Datenschutzrecht, Art. 5, Rn. 8; Hacker (2020), Datenprivatrecht: Neue Technologien im Spannungsfeld von Datenschutzrecht und BGB, S. 153

[145] Vgl. Hacker (2020), Datenprivatrecht: Neue Technologien im Spannungsfeld von Datenschutzrecht und BGB, S. 153

[146] Vgl. Vogel (2022), Künstliche Intelligenz und Datenschutz, S. 104; Schantz in Wolff/Brink/v. Ungern-Sternberg (2021), BeckOK Datenschutzrecht, Art. 5, Rn. 9

unterstützt, gegen den datenschutzrechtlichen Grundsatz von Treu und Glauben verstoßen würde, wenn das System tatsächlich ein Bias hinsichtlich dunkelhäutiger Menschen aufweist.[147]

3.4 Transparenzgrundsatz

Nach dem Transparenzgrundsatz gemäß Art. 5 Abs. 1 lit. a Var. 3 DSGVO muss die Verarbeitung personenbezogener Daten in einer für die betroffenen Personen nachvollziehbaren Weise erfolgen. Dieses Prinzip wird in Erwägungsgrund 39 zur DSGVO konkretisiert. Demnach muss bei einer Datenverarbeitung Transparenz dahingehend bestehen, ob und in welchem Umfang personenbezogene Daten verarbeitet werden und künftig noch verarbeitet werden sollen. Dies setzt voraus, dass alle Informationen zur Datenverarbeitung leicht zugänglich und in verständlicher, klarer und einfacher Sprache zur Verfügung gestellt werden (vgl. auch Art. 12 ff. DSGVO). Die Transparenz stellt einen fundamentalen Grundsatz des Datenschutzrechts dar. Ohne die Kenntnis über das „Ob“ und „Wie“ der Datenverarbeitung, kann das Grundrecht auf Schutz personenbezogener Daten kaum effektiv wahrgenommen werden. Dies spiegelt sich in den Art. 12 ff. DSGVO.

3.4.1 Art. 14 DSGVO

In Art. 13 und Art. 14 DSGVO finden sich die konkreten Informationspflichten der verantwortlichen Stelle ggü. den betroffenen Personen. Die Pflichten des Art. 13 DSGVO kommen zur Anwendung, wenn personenbezogene Daten unmittelbar bei der betroffenen Person erhoben werden (direkte Datenerhebung). Art. 14 DGSVO hingegen ist für Situationen maßgeblich, in denen personenbezogene Daten nicht direkt bei den betroffenen Personen erhoben werden. Diese Normen legen fest, welche Informationen den betroffenen Personen bereitgestellt werden müssen und wann die Informationsbereitstellung zu erfolgen hat, um eine transparente und faire Datenverarbeitung zu gewährleisten. Informationen, die bereitgestellt werden müssen, sind insb. der Name und die Kontaktdaten der verantwortlichen Stelle, die konkreten Zwecke der jeweiligen Datenverarbeitung sowie die zur Anwendung kommende Rechtsgrundlage für die Datenverarbeitung. Art. 13 und 14 DSGVO unterscheiden sich insbesondere in den Vorgaben zum Zeitpunkt der Informationsbereitstellung. Nach

[147] Vgl. Vogel (2022), Künstliche Intelligenz und Datenschutz, S. 105,

Art. 13 Abs. 1 Satz 1 DSGVO muss die Informationsbereitstellung grundsätzlich zum Zeitpunkt der Datenerhebung erfolgen. Werden die personenbezogenen Daten nicht direkt bei den betroffenen Personen erhoben, sodass Art. 14 DSGVO zur Anwendung kommt, müssen die betroffenen Personen innerhalb einer angemessenen Frist, die sich nach dem konkreten Einzelfall richtet, spätestens jedoch innerhalb eines Monats informiert werden (vgl. Art. 14 Abs. 3 lit. a DSGVO, ErwGr. 61 DSGVO).

Im Hinblick auf Datenverarbeitungen bei der Entwicklung und dem Betrieb von LLMs würde Art. 13 DSGVO für die Verarbeitung von personenbezogenen Daten der jeweiligen Nutzer des LLMs im Rahmen der Inferenz-Phase zur Anwendung kommen, während Art. 14 DSGVO bei der Verarbeitung personenbezogener Daten maßgeblich ist, die in Trainingsdaten enthalten sind und nicht direkt von den betroffenen Personen erhoben werden.

Es stellt sich die Frage, wie die Bereitstellung von Informationen über eine Datenverarbeitung im Rahmen eines LLM-Trainings gemäß Art. 14 DSGVO gewährleistet werden kann. Der Umstand, dass die personenbezogenen Trainingsdaten aus öffentlich zugänglichen Quellen stammen, führt nicht automatisch dazu, dass die Informationspflichten keine Anwendung finden.[148]

In diesem Zusammenhang könnte jedoch der Ausnahmetatbestand des Art. 14 Abs. 5 lit. b Satz 1 HS. 1 DSGVO zur Anwendung kommen. Nach dieser Norm ist die Anwendung der Informationspflichten gemäß Art. 14 Abs. 1-4 DSGVO ausgeschlossen, wenn die Erteilung der relevanten Informationen unmöglich oder mit einem unverhältnismäßigen Aufwand verbunden ist. Dieser Ausnahmetatbestand ist eng auszulegen.[149] Demnach muss die jeweils verantwortliche Stelle eine etwaige Unmöglichkeit der Informationserteilung belegen können[150] In der Praxis wird sich die

[148] Vgl. Art. 14 Abs. 2 lit. f DSGVO

[149] Vgl. Artikel-29-Datenschutzgruppe (2018), Leitlinien für Transparenz gemäß der Verordnung 2016/679, WP 260 rev.01, S. 35; Knyrim in Ehmann/Selmayr (2024), Datenschutz-Grundverordnung, Art. 14, Rn. 42-48

[150] Vgl. Artikel-29-Datenschutzgruppe (2018), Leitlinien für Transparenz gemäß der Verordnung 2016/679, WP 260 rev.01, S. 35; Knyrim in Ehmann/Selmayr (2024), Datenschutz-Grundverordnung, Art. 14, Rn. 42-48

Darlegung der Unmöglichkeit schwierig gestalten. Auch wenn die jeweilige verantwortliche Stelle bei einer Datenverarbeitung im Rahmen des LMM-Trainings i. d. R. keine Kenntnis über die Identität der betroffenen Personen hat, ist es grundsätzlich denkbar, dass betroffene Personen, deren Daten ggf. in Trainingsdaten enthalten sind, theoretisch ausfindig gemacht werden könnten (vgl. hierzu Kapitel 2.3.2).

Vielmehr kommt die Anwendung des Art. 14 Abs. 5 lit. b Satz 1 HS. 1 Var. 2 DSGVO infrage, der eine Ausnahme der Informationspflichten nach Art. 14 Abs. 1-4 DSGVO vorsieht, wenn die Informationsbereitstellung einen unverhältnismäßigen Aufwand begründen würde. Aufgrund des in der Regel großen Umfangs und der Diversität von Datensätzen, die für das Training von LLMs genutzt werden, würde das Ausfindigmachen der betroffenen Personen objektiv betrachtet regelmäßig einen hohen Aufwand erfordern. Erwägungsgrund 62 DSGVO nennt die Zahl der betroffenen Personen, das Alter der Daten oder etwaige geeignete Garantien als Anhaltspunkte, die für die Begründung eines unverhältnismäßigen Aufwands herangezogen werden sollten. Für die wirksame Anwendung des Art. 14 Abs. 5 lit. b Satz 1 HS. 1 Var. 2 DSGVO ist die Dokumentation einer Beurteilung des mit der Informationsbereitstellung verbundenen Aufwands sowie eine dokumentierte Abwägung dieses Aufwands mit den Auswirkungen einer unterbliebenen Informationsbereitstellung für die Rechte und Freiheiten der betroffenen Personen notwendig.[151] Ob die Voraussetzungen des Art. 14 Abs. 5 lit. b Satz 1 HS. 1 Var. 2 DSGVO vorliegen, hängt folglich von den Umständen des jeweiligen Einzelfalls ab. So kann der Umstand, dass die personenbezogenen Trainingsdaten aus öffentlich zugänglichen Quellen stammen, ein geringeres Informationsinteresse der betroffenen Personen begründen.[152] Auf der anderen Seite überwiegen die potenziellen Auswirkungen einer unterbliebenen Information auf die Rechte und Freiheiten der betroffenen Personen den Aufwand der Informationsbereitstellung ggf., wenn besondere Kategorien personenbezogener Daten verarbeitet werden.

Wenn die verantwortliche Stelle im Rahmen der Abwägung darlegen kann, dass der Aufwand der Informationsbereitstellung den Auswirkungen auf die Rechte und Freiheiten

[151] Vgl. Artikel-29-Datenschutzgruppe (2018), Leitlinien für Transparenz gemäß der Verordnung 2016/679, WP 260 rev.01, S. 38

[152] Vgl. Werry (2023), Generative KI-Modelle im Visier der Datenschutzbehörden, S. 914

der betroffenen Personen überwiegt, fordert Art. 14 Abs. 5 lit. b Satz 2 DSGVO ferner, dass die verantwortliche Stelle geeignete Maßnahmen zum Schutz der Rechte und Freiheiten sowie der berechtigten Interessen der betroffenen Personen ergreift. Als geeignete Maßnahme wird insbesondere die Bereitstellung von Informationen über die Datenverarbeitung für die Öffentlichkeit genannt. Eine solche Informationsbereitstellung für die Öffentlichkeit ist bspw. über die Webseite der verantwortlichen Stelle denkbar. Der Europäische Datenschutzausschuss nennt u. a. die Durchführung einer Datenschutz-Folgenabschätzung und die Anwendung von Pseudonymisierungstechniken als weitere Maßnahmen, die in diesem Kontext ggf. ergriffen werden sollten.[153] Folglich ist die Informationsbereitstellung auch dann erforderlich, wenn der Ausnahmetatbestand des Art. 14 Abs. 5 lit. b DSGVO zur Anwendung kommt. Aufgrund der Komplexität der Funktionsweise von LLMs dürfte es in der Regel eine Herausforderung für die jeweilige verarbeitende Stelle darstellen, leicht verständliche Informationen über die Datenverarbeitung zu formulieren.

3.4.2 Art. 15 DSGVO

Der in Art. 5 Abs. 1 lit. a Var. 3 DSGVO verankerte Transparenzgrundsatz kommt auch durch das Auskunftsrecht der betroffenen Personen nach Art. 15 DSGVO zum Ausdruck. Nach dieser Norm haben betroffene Personen das Recht, von der verantwortlichen Stelle Auskunft darüber zu erhalten, ob sie betreffende personenbezogene Daten verarbeitet werden sowie eine Kopie der gegenständlichen personenbezogenen Daten in gängigem maschinenlesbarem Format zu erhalten. Da die verantwortliche Stelle bei der Verarbeitung personenbezogener Daten im Rahmen des LLM-Trainings jedoch in der Regel keine Kenntnis von der Identität der betroffenen Personen hat ist fraglich, wie das Auskunftsrecht nach Art. 15 DSGVO in der Praxis umgesetzt werden kann.

In solchen Situationen könnte die verantwortliche Stelle ggf. gemäß Art. 11 Abs. 2 DSGVO von der Pflicht zur Erfüllung der Betroffenenrechte nach Art. 15-20 DSGVO (Recht auf Auskunft, Recht auf Berichtigung, Recht auf Löschung, Recht auf Einschränkung der Verarbeitung, Mitteilungspflicht und Recht auf Datenübertragbarkeit) befreit sein, wenn sie nachweisen kann, dass ihr die Identifizierung der betroffenen

[153] Vgl. Artikel-29-Datenschutzgruppe (2018), Leitlinien für Transparenz gemäß der Verordnung 2016/679, WP 260 rev.01, S. 38

Personen nicht möglich ist. Art. 11 Abs. 2 DSGVO adressiert Situationen, in denen ein Personenbezug der Daten objektiv gegeben ist, die verantwortliche Stelle selbst jedoch nicht in der Lage ist, die betroffenen Personen ohne zusätzliche Informationen zu identifizieren. Also Situationen, in denen sich personenbezogene Daten auf identifizierbare, aber nicht auf identifizierte Personen beziehen. Da der Wegfall der Betroffenenrechte nach Art. 15-20 DSGVO eine schwerwiegende Einschränkung der Rechte und Freiheiten der betroffenen Personen darstellen kann, muss die verantwortliche Stelle die für sie gegebene Unmöglichkeit der Identifizierung nachweisen können. Ferner haben die betroffenen Personen nach Art. 11 Abs. 2 Satz 1 DSGVO das Recht, durch Bereitstellung weiterer Daten ihre Identifizierbarkeit herzustellen, sodass die verantwortliche Stelle in die Lage versetzt wird, den Anfragen nach Art. 15-20 nachzukommen.

Für den Sachverhalt der Datenverarbeitung im Rahmen des LLM-Trainings kommt es für die Beurteilung, ob die Identifizierung der betroffenen Personen für die verantwortliche Stelle möglich ist oder nicht, maßgeblich auf den Zeitpunkt der Betroffenenanfrage bzw. auf die Phase des LLM-Trainings an. Theoretisch sind Methoden denkbar, die es unter bestimmten Bedingungen erlauben, personenbezogene Daten einer Person aus dem Trainingsdatensatz herauszufiltern (vgl. hierzu Kapitel 2.3.2 sowie Kapitel 5.2.1). In der Praxis wäre die Umsetzung solcher Methoden jedoch aufgrund des enormen Umfangs der Trainingsdaten nur unter erheblichem Aufwand möglich. Vor diesem Hintergrund erscheint es auch für die Beurteilung der Unmöglichkeit i. S. d. Art. 11 Abs. 2 DSGVO als angemessen, eine Abwägung des Aufwands mit den Auswirkungen einer unterbliebenen Informationsbereitstellung für die Rechte und Freiheiten der betroffenen Personen anzuwenden, anstatt von der Erforderlichkeit einer absoluten Unmöglichkeit auszugehen. Einen Anhaltspunkt für eine solche Auslegung gibt der Europäische Datenschutzausschuss, welcher als Positivbeispiel für einen Sachverhalt, für den die Regelung des Art. 11 Abs. 2 DSGVO zur Anwendung kommen kann, ein Auskunftsersuchen bezüglich der Datenverarbeitung im Rahmen der Videoüberwachung eines Gebäudes nennt, welches sich auf einen langen Zeitraum der Videoaufzeichnungen bezieht. In diesem Fall könnte sich die verantwortliche Stelle auf Art. 11 Abs. 2 DSGVO berufen, wenn sie nicht in der Lage ist, eine so große Datenmenge zu verarbeiten.[154]

[154] Vgl. European Data Protection Board (2023), Leitlinien 01/2022, Version 2.1, Beispiel 10, Rn. 61

3.4.3 Zwischenfazit

Zusammenfassend lässt sich feststellen, dass ein gewisses Spannungsverhältnis zwischen dem Training von LLMs und der Einhaltung der Transparenzanforderungen der DSGVO besteht. Die Anforderung einer transparenten, verständlichen und nachvollziehbaren Verarbeitung personenbezogener Daten, stößt bei der Entwicklung und dem Betrieb von LLMs auf praktische Grenzen.

Art. 14 Abs. 5 lit. b sowie Art. 11 Abs. 2 DSGVO sehen in Fällen, in denen die Informationsbereitstellung mit einem unverhältnismäßigen Aufwand verbunden oder gar unmöglich ist, zwar Ausnahmen von den Informationspflichten vor, die Anwendung dieser Ausnahmen ist traditionell jedoch eng auszulegen und setzt eine Abwägung des Aufwands der Informationsbereitstellung gegen die möglichen Auswirkungen auf die Rechte und Freiheiten der betroffenen Personen sowie eine strenge Nachweispflicht voraus.

In der Praxis könnte ein risikobasierter Ansatz dazu beitragen, das Spannungsverhältnis zwischen Transparenzanforderungen und den praktischen Herausforderungen des LLM-Trainings zu entschärfen. Durch eine pragmatische Auslegung der Transparenzpflichten, die sich an den tatsächlichen Risiken für die Rechte und Freiheiten der betroffenen Personen orientiert, könnte ein angemessener Ausgleich geschaffen werden. Es besteht jedoch weiterhin rechtliche Unsicherheit darüber, inwieweit bestimmte Transparenzanforderungen im Kontext von LLMs anwendbar sind. Künftige Rechtsprechung wird hier von entscheidender Bedeutung sein, um eine klare und praxisgerechte Linie für die Einhaltung datenschutzrechtlicher Anforderungen bei der Entwicklung und dem Betrieb von LLMs zu schaffen.

3.5 Grundsatz der Datenminimierung

Nach dem Grundsatz der Datenminimierung (Art. 5 Abs. 1 lit. c DSGVO) muss die Verarbeitung personenbezogener Daten „dem Zweck angemessen und erheblich sowie auf das für die Zwecke der Verarbeitung notwendige Maß beschränkt sein“.[155] Daraus folgt insbesondere, dass nicht mehr personenbezogene Daten verarbeitet werden dürfen, wie für die Erreichung des jeweiligen Zwecks unbedingt erforderlich sind.

[155] Art. 5 Abs. 1 lit. c DSGVO

Aufgrund ihrer Funktionsweise sind LLMs auf sehr umfangreiche Datensätze angewiesen. Während nach der DSGVO personenbezogene Daten nur in dem Umfang verarbeitet werden sollten, wie es für die Erreichung der jeweiligen Zwecke erforderlich ist, gilt für das Training von LLMs: je mehr Daten desto besser. Folglich zeichnet sich hier ein Konflikt zwischen der Funktionsweise von LLMs und dem datenschutzrechtlichen Grundsatz der Datenminimierung ab.

Unbestimmte Rechtsbegriffe in der Formulierung des Art. 5 Abs. 1 lit. c DSGVO wie „angemessen", „erheblich" und „notwendig" bieten jedoch eine gewisse Flexibilität. Folglich verstößt das Training eines LLMs mit umfangreichen Datensätzen nicht per se gegen den Grundsatz der Datenminimierung. Vielmehr kommt es darauf an, ob eine Datenverarbeitung in dem konkreten Ausmaß für die Erreichung des Verarbeitungszwecks (bspw. Entwicklung eines LLMs) erforderlich ist. Insoweit steht der Grundsatz der Datenminimierung in Abhängigkeit zum Zweckbindungsgrundsatz. Allerdings ist die Verarbeitung personenbezogener Daten für das erfolgreiche Training eines LLMs nicht direkt notwendig. Beim Training von LLMs kommt es vielmehr auf die statistischen Eigenschaften der Trainingstexte und weniger auf die Identifizierbarkeit der Datensubjekte an (vgl. hierzu Kapitel 2.3.3). Dennoch besteht eine gewisse indirekte Erforderlichkeit der Verarbeitung personenbezogener Daten beim Training von LLMs, zumindest in den Fällen, in denen eine konsequente Anonymisierung der Trainingsdatensätze mit einem unverhältnismäßigen Aufwand und ggf. einem Qualitätsverlust einhergeht (vgl. hierzu Kapitel 2.3.2). Folglich ist ein Einklang zwischen dem Training von LLMs und dem Grundsatz der Datenminimierung denkbar, wenn man den Grundsatz der Datenminimierung nach einem relativen Verständnis und nicht nach einem absoluten Verständnis teleologisch auslegt.[156] ErwGr. 39 Satz 9 DSGVO bietet einen Anhaltspunkt für ein solches relatives Verständnis. Demnach sollen personenbezogene Daten nur verarbeitet werden dürfen, „wenn der Zweck der Verarbeitung nicht in zumutbarer Weise durch andere Mittel erreicht werden kann".[157]. Dennoch müssen alle möglichen und zumutbaren Maßnahmen umgesetzt werden, welche die Datenverarbeitung auf das unbedingt notwendige Maß beschränken und keinen

[156] Vgl. Paal (2024), KI-Training mit öffentlichen frei zugänglichen Daten im Lichte der DS-GVO-Vorgaben, S. 141; Vogel (2022), Künstliche Intelligenz und Datenschutz, S. 162

[157] Vgl. ErwGr. 39 S. 9 DSGVO

unverhältnismäßigen Aufwand begründen, um das Training eines LLMs in Einklang mit dem Grundsatz der Datenminimierung zu bringen. Denkbare Maßnahmen wären z. B. sog. Differential Privacy-Ansätze oder die Nutzung von synthetischen Daten (vgl. hierzu Kapitel 5 und 6).

3.6 Zwischenfazit

Das dritte Kapitel hat verdeutlicht, dass die Anwendung der DSGVO auf das Training von LLMs zahlreiche Problemstellungen aufwirft, die sich maßgeblich aus der spezifischen Funktionsweise des maschinellen Lernens ergeben. Man könnte jedoch auch sagen, dass maschinelles Lernen die Zukunft ist, die schon vor langer Zeit – u. a. von Alan Turing – vorhergesagt wurde und nun die Versäumnisse und Schwächen der DSGVO eklatant offenlegt.[158] Insbesondere die Notwendigkeit umfangreicher und diverser Trainingsdatensätze steht in einem Spannungsverhältnis zu grundlegenden Datenschutzprinzipien wie Zweckbindung, Datenminimierung, Transparenz und der Verarbeitung nach Treu und Glauben. Dies führt zu erheblichen Rechtsunsicherheiten, die sowohl die Rechtsgrundlagen als auch die praktische Umsetzung datenschutzrechtlicher Anforderungen betreffen.

Trotz der genannten Spannungen und Unsicherheiten ist die Vereinbarkeit von Datenverarbeitung im Rahmen des LLM-Trainings mit der DSGVO unter bestimmten Umständen denkbar. Ein risikobasierter Ansatz, der die datenschutzrechtlichen Grundsätze in ihrer relativen Auslegung berücksichtigt und sich an den tatsächlichen Auswirkungen der Datenverarbeitung auf die Rechte und Freiheiten der betroffenen Personen orientiert, bietet dabei einen möglichen Lösungsweg. Dabei ist die Umsetzung geeigneter technischer und organisatorischer Maßnahmen (vgl. hierzu Kapitel 5 und 6) sowie eine detaillierte Einzelfallprüfung essenziell.

4 Anwendbarkeit der DSGVO auf das Training von LLMs

Die im vorangegangenen Kapitel dargestellten datenschutzrechtlichen Herausforderungen werden jedoch nur relevant, sofern das Training von LLMs in den Anwendungsbereich der DSGVO fällt. Die Anwendung der DSGVO auf das Training von LLMs steht und fällt mit der Bewertung des Personenbezugs der jeweiligen Trainingsdaten. Insbesondere

[158] Vgl. Solove (2024), Artificial Intelligence and Privacy, S. 7

der Hamburgische Beauftragte für Datenschutz und Informationsfreiheit (HmbBfDI) hat mit der Veröffentlichung des Diskussionspapiers „Large Language Models und personenbezogene Daten“ vom 15. Juli 2024 eine rege Diskussion darüber angestoßen, inwieweit Datenverarbeitungen im Rahmen eines LLMs in den Anwendungsbereich der DSGVO fallen. In genanntem Diskussionspapier vertritt der HmbBfDI die Meinung, dass Trainingsdaten im Rahmen des Trainingsprozesses – genauer: im Rahmen der sog. Tokenisierung (vgl. hierzu Kapitel 2.3.3) – in „abstrakte mathematische Repräsentationen überführt werden“ und dieser Abstraktionsprozess dazu führt, dass ein etwaig vorhandener Personenbezug verloren geht.[159] Dies hätte zur Folge, dass zumindest ab dem Zeitpunkt der Tokenisierung keine Datenverarbeitung i. S. d. DSGVO mehr vorliegt, sodass der Betrieb oder die Speicherung eines LLMs an sich keine datenschutzrechtliche Relevanz hätte.[160] Nach Auffassung des HmbBfDI sind nur der Input (das Training und die Eingabe von Prompts) sowie ggf. der Output von LLMs datenschutzrechtlich relevant.[161]

Die bisherige Literatur beschäftigt sich hauptsächlich mit der Frage, ob ein LLM personenbezogene Daten enthält. Dabei wird häufig pauschal angenommen, dass das Training von LLMs in den Anwendungsbereich der DSGVO fällt, da die Trainingsdaten i. d. R. Daten enthalten, die sich grds. zur Identifikation einzelner, natürlicher Personen eignen.[162] Nachfolgend soll ein Beitrag zu dieser Diskussion geleistet werden, in dem

[159] Vgl. Hamburgischer Beauftragter für Datenschutz und Informationsfreiheit (2024), Diskussionspapier: Large Language Models und personenbezogene Daten v. 15. Juli 2024, S. 4

[160] Vgl. Hamburgischer Beauftragter für Datenschutz und Informationsfreiheit (2024), Diskussionspapier: Large Language Models und personenbezogene Daten v. 15. Juli 2024, S. 4

[161] Vgl. Hamburgischer Beauftragter für Datenschutz und Informationsfreiheit (2024), Diskussionspapier: Large Language Models und personenbezogene Daten v. 15. Juli 2024, S. 11

[162] Vgl. Bartels (2024), A Balancing Act: Data Protection Compliance of Artificial Intelligence, S. 528; Hamburgischer Beauftragter für Datenschutz und Informationsfreiheit (2024), Diskussionspapier: Large Language Models und personenbezogene Daten v. 15. Juli 2024; Rosenthal (2024), Sprachmodelle mit und ohne personenbezogene Daten, Blogbeitrag vom 17.072024, aufrufbar unter: https://www.vischer.com/know-how/blog/teil-19-sprachmodelle-mit-und-ohne-personenbezogene-daten/ (letzter Aufruf: 24.11.2024); Moos (2024), Personenbezug von Large Language Models, Rn. 2

insbesondere die datenschutzrechtliche Relevanz des Trainings von LLMs hinterfragt wird.

4.1 Personenbezug von Trainingsdaten

Trainingsdaten für LLMs enthalten häufig Texte von öffentlich verfügbaren Webseiten, Zeitungsartikeln und Social Media-Beiträgen, die aus dem Internet gecrawlt und gescraped werden (vgl. hierzu Kapitel 2.3.1). Solche Trainingsdatensätze können alle möglichen Datenarten enthalten, unter anderem auch Daten, die üblicherweise personenbezogene Daten darstellen, wie z. B. Namen, E-Mail-Adressen und Telefonnummern.[163] Auch Trainingsdaten aus nicht-öffentlich verfügbaren Quellen (z. B. Vertriebsdokumentationen und Finanzberichte von Unternehmen) können personenbezogene Daten enthalten.

Der sachliche Anwendungsbereich der DSGVO ist eröffnet, wenn personenbezogene Daten ganz oder teilweise automatisiert verarbeitet oder nichtautomatisiert in einem Dateisystem gespeichert werden.[164] Nach Art. 4 Nr. 1 DSGVO sind personenbezogene Daten alle Informationen, die sich auf eine identifizierte oder identifizierbare natürliche Person beziehen. Dabei fallen nicht nur solche Informationen unter den Begriff der personenbezogenen Daten, die eine unmittelbare Identifikation ermöglichen, sondern auch solche, die eine Identifikation nur in Zusammenhang mit entsprechenden Zusatzinformationen ermöglichen, wie bspw. eine Personalnummer. Informationen wie Namen, E-Mail-Adressen und Telefonnummern, die regelmäßig in Trainingsdatensätzen aus öffentlichen Quellen enthalten sind, eignen sich grundsätzlich dazu, einzelne natürliche Personen zu identifizieren.

Die bloße Tatsache, dass eine Information von sich aus potenziell zur Identifizierung natürlicher Personen geeignet sein kann, reicht nach der Rechtsprechung des Europäischen Gerichtshofs (EuGH) jedoch nicht aus, um die Information automatisch als personenbezogenes Datum zu qualifizieren.[165] Damit bestätigt der EuGH den sog.

[163] Vgl. Longpre et al. (2023), A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality,& toxicity, S. 5

[164] Vgl. Art. 2 Abs. 1 DSGVO

[165] Vgl. EuGH (2. Kammer), Urt. v. 19.10.2016 – C-582/14 (Breyer/Deutschland), Rn. 38-39

relativen Begriff des Personenbezugs, nach welchem für die Bewertung des Personenbezugs auf die Sicht der datenverarbeitenden Stelle abzustellen ist. Wenn dieser die relevanten Zusatzinformationen bzw. die Mittel zur Identifizierung des Datensubjekts nicht vorliegen, dann fehlt es der jeweiligen Information aus Sicht der datenverarbeitenden Stelle an einem Personenbezug.[166]. Bei der Bewertung müssen alle Mittel berücksichtigt werden, die von der verarbeitenden Stelle nach allgemeinem Ermessen wahrscheinlich genutzt werden, um eine natürliche Person direkt oder indirekt zu identifizieren.[167] Es muss sich also um Mittel handeln, die „vernünftigerweise" zur Bestimmung der Datensubjekte eingesetzt werden können.[168]

Hinsichtlich der Trainingsdaten für LLMs sollte hinterfragt werden, welche Mittel der datenverarbeitenden Stelle zur Verfügung stehen und ob eine Identifikation über die vorliegenden Mittel „vernünftig" ist. Wann ein Mittel vernünftigerweise zur Identifizierung der jeweiligen Personen eingesetzt werden kann, lässt der EuGH bis dato leider offen. Unklar ist, ob ein Mittel bereits vernünftigerweise eingesetzt werden kann, wenn der Einsatz theoretisch möglich ist, oder ob vielmehr auf die praktische Wahrscheinlichkeit abzustellen ist. Nach ErwGr. 26 Satz 4 DSGVO ist auf alle objektiven Faktoren, wie die Kosten der Identifizierung und der dafür erforderliche Zeitaufwand sowie die, zum Zeitpunkt der Verarbeitung verfügbare, Technologie und technologische Entwicklungen abzustellen. Nach Ansicht des Generalanwalts Sánchez-Bordona im Rahmen seiner Schlussanträge zum referenzierten Breyer-Urteil des EuGH, kommt es insbesondere auf die Mittel an, „[…] die im Rahmen des Gesetzes realisierbar und deshalb „vernünftig" [sind]".[169]

Abgesehen von Ausnahmefällen, in denen es der verarbeitenden Stelle objektiv betrachtet ohne unverhältnismäßigen Aufwand möglich sein sollte, die Datensubjekte direkt über die Trainingsdaten zu identifizieren – etwa, wenn es sich um unternehmensinterne

[166] Vgl. EuGH (2. Kammer), Urt. v. 19.10.2016 – C-582/14 (Breyer/Deutschland), Rn. 45; EuGH, Urteil v. 9.11.2023 – C-319/22, Rn. 46

[167] Vgl. ErwGr. 26 S. 3 DSGVO

[168] Vgl. EuGH (2. Kammer), Urt. v. 19.10.2016 – C-582/14 (Breyer/Deutschland), Rn. 45; EuGH, Urteil v. 9.11.2023 – C-319/22, Rn. 46

[169] Vgl. Generalanwalt Sánchez-Bordona, Schlussanträge zu EuGH-Breyer Rn. 68 ff.

Informationen zu Mitarbeitenden, Kunden oder sonstigen Geschäftspartnern handelt –, ist fraglich, ob ihr im Allgemeinen vernünftige Mittel zur Verfügung stehen, um die Datensubjekte über die Trainingsdaten zu identifizieren. Dies gilt insbesondere, wenn die Trainingsdatensätze aus öffentlich verfügbaren Informationen bestehen, die aus dem Internet gescraped und gecrawlt werden. Insbesondere in diesen Fällen hat die datenverarbeitende Stelle i. d. R. weder Kenntnis von den konkreten personenbezogenen Daten in einem Trainingsdatensatz noch von der Identität der Datensubjekte.

Unter Berücksichtigung verfügbarer Technologien, wäre eine Filterung der Trainingsdaten nach potenziell personenbezogenen Daten grundsätzlich möglich.[170] Abhängig vom Umfang des jeweiligen Trainingsdatensatzes, ist dies mit mehr oder weniger hohen technischen und finanziellen Hürden verbunden und aufgrund des damit einhergehenden Qualitätsverlusts für das LLM i. d. R. nicht effizient.[171] Angenommen, die potenziell personenbezogenen Daten könnten im Einzelfall mit einem verhältnismäßigen Aufwand aus den Trainingsdaten gefiltert werden, würden sich darunter auch Informationen finden, die sich auf fiktive Personen beziehen (z. B. aus Romanen) sowie ggf. synthetische Daten (vgl. Kapitel 6). Davon abgesehen, würde es zudem vom Kontext und der Art der konkreten Daten abhängen, ob diese eine direkte Identifizierung der Datensubjekte zulassen. So genügt der Name einer Person ohne weitere Zusatzinformationen, wie bspw. einer Wohnadresse, regelmäßig nicht, eine Person eindeutig zu identifizieren. Zumindest in solchen Fällen könnte eine Identifikation der Datensubjekte und folglich ein Personenbezug unter Anwendung des relativen Begriffs des Personenbezugs verneint werden (vgl. hierzu auch Kapitel 3.4.1 zur Anwendbarkeit des Art. 14 Abs. 5 lit. b DSGVO).

Da die Trainingsdatensätze i. d. R. einen großen Umfang und eine große Diversität aufweisen, ist jedoch grds. denkbar, dass darin alle möglichen Arten potenziell personenbezogener Daten vorhanden sind, auch solche, die eine direkte Identifikation der

[170] Vgl. Longpre et al. (2023), A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality,& toxicity. S. 5; Laurençon et al. (2023), The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset, S. 7-8

[171] Vgl. Xu et al. (2021), Privacy-Preserving Machine Learning: Methods, Challenges, and Directions, S.28; El Mestari et al. (2024), Preserving Data Privacy in Machine Learning Systems, S. 15; Carlini et al. (2021), Extracting Training Data from Large Language Models, S. 3

Datensubjekte zulassen würden. Aufgrund der Unverhältnismäßigkeit gegenüber dem Nutzen, werden die Trainingsdaten in der Praxis jedoch regelmäßig nicht nach potenziell personenbezogenen Daten gefiltert. Fraglich ist insoweit, wie sich der Umstand auswirkt, dass die datenverarbeitende Stelle üblicherweise kein Interesse an der Identifikation der Datensubjekte hat. Im Hinblick auf die Rechtsprechung des EuGH zu sensiblen personenbezogenen Daten, aus der für bestimmten Sachverhalte abgeleitet werden kann, dass personenbezogene Daten nicht als sensibel i. S. d. Art. 9 DSGVO gelten, wenn es an einer subjektiven Auswertungsabsicht der potenziell sensiblen Daten fehlt, sollte auch bei der Bewertung, ob Trainingsdaten personenbezogen sind, berücksichtigt werden, ob eine Absicht zur Identifikation besteht oder nicht.[172] Ebenso wenig, wie es einem Juwelier im Rahmen der Videoüberwachung seines Juweliergeschäfts auf einen etwaigen sensiblen Informationsgehalt der Videoaufnahmen ankommt (bspw. wenn eine Person mit Brille aufgezeichnet wird), kommt es einem KI-Entwickler auf einen etwaigen personenbezogenen Informationsgehalt der Trainingsdaten an. Die Ansicht des Generalanwalts Sánchez-Bordona, dass bereits die Möglichkeit der Identifikation im Rahmen der Legalität genügt, um einen Personenbezug anzunehmen, scheint zu weit gegriffen. Vielmehr könnte die Anwendung der Formel des Schweizer Bundesgerichts angemessen sein. Nach dieser scheidet ein Personenbezug aus, wenn der „Aufwand derart gross [ist], dass nach der allgemeinen Lebenserfahrung nicht damit gerechnet werden muss, dass ein Interessent diesen auf sich nehmen wird […]“.[173] Auch nach Ansicht des Schweizer Bundesgerichts kommt es nicht allein darauf an, „welcher Aufwand objektiv erforderlich ist, um eine bestimmte Information einer Person zuordnen zu können, sondern auch [darauf], welches Interesse der Datenbearbeiter oder ein Dritter an der Identifizierung hat.“[174] Wie Rosenthal erkannt hat, steht die Berücksichtigung des subjektiven Interesses an einer Identifikation, dem objektiven Bewertungsmaßstab der

[172] Vgl. Kohn/Schleper (2023), Die (zufällige) Erhebung sensibler Daten, S. 723; Schneider/Schindler (2018), Videoüberwachung als Verarbeitung besonderer Kategorien personenbezogener Daten, S. 463; vgl. hierzu ebenfalls EDSA (2020), Leitlinien 3/2019 zur Verarbeitung personenbezogener Daten durch Videogeräte, Version 2.0, Rn. 62-72; vgl. hierzu auch Rosenthal (2024), Sprachmodelle mit und ohne personenbezogene Daten, Blogbeitrag vom 17.072024, aufrufbar unter: https://www.vischer.com/know-how/blog/teil-19-sprachmodelle-mit-und-ohne-personenbezogene-daten/ (letzter Aufruf: 24.11.2024)

[173] Schweizer Bundesgericht, Urteil v. 08.09.2010 BGE 136 II 508, E. 3.2 ff.

[174] Schweizer Bundesgericht, Urteil v. 08.09.2010 BGE 136 II 508, E. 3.2 ff.

DSGVO nicht entgegen, da bei einer objektiven Betrachtung eben dieses subjektive Interesse entscheidet, welchen Aufwand die jeweilige Stelle für eine Identifikation aufbringen würde. „Wer kein Interesse an einer Identifikation hat, betreibt dafür – vernünftigerweise – auch keinen Aufwand, weil der Nutzen den Aufwand nicht rechtfertigt.“[175] Rosenthal nutzt für seine Argumentation eine Analogie zu einem großen, ungeordneten Stapel an Unterlagen, welcher ggf. personenbezogene Daten enthält: wenn der Besitzer des Stapels nicht weiß, ob bzw. welche und geschweige denn an welcher Stelle im Stapel sich die personenbezogenen Daten befinden, wird er vernünftigerweise nur dann den Aufwand aufbringen, den Stapel danach zu durchsuchen, wenn sein Interesse daran entsprechend groß ist.[176]

Schließlich hängt die Bewertung, wie so oft, von den Umständen des konkreten Einzelfalls ab. Unabhängig vom konkreten Einzelfall, kann jedoch festgehalten werden, dass eine Identifikation von Datensubjekten durch den Entwickler eines LLMs über Informationen, die in den Trainingsdaten enthalten sind, von einer Mehrzahl von Variablen abhängt. Es zeigt sich, dass unter Anwendung und Auslegung des relativen Begriffs des Personenbezugs, ein Personenbezug von Trainingsdaten nicht pauschal angenommen werden kann.

Auch wenn es KI-Herstellern nicht auf den potenziell personenbezogenen Informationsgehalt der Trainingsdaten ankommt und ein Personenbezug aus Sicht der KI-Hersteller verneint werden würde, kann das LLM-Training zumindest indirekt Auswirkungen auf die Rechte und Freiheiten der Datensubjekte haben. Betrachtet man den Prozess der Trainingsdatensammlung sowie den Prozess des Pre-Trainings an sich, werden diese i. d. R. keine Auswirkungen auf die Rechte und Freiheiten der Datensubjekte haben. Die Daten sind regelmäßig bereits im Internet öffentlich zugänglich und werden im Rahmen des Trainings keinem neuen, größeren Personenkreis

[175] Rosenthal (2024), Sprachmodelle mit und ohne personenbezogene Daten, Blogbeitrag vom 17.072024, aufrufbar unter: https://www.vischer.com/know-how/blog/teil-19-sprachmodelle-mit-und-ohne-personenbezogene-daten/ (letzter Aufruf: 24.11.2024)

[176] Vgl. Rosenthal (2024), Sprachmodelle mit und ohne personenbezogene Daten, Blogbeitrag vom 17.072024, aufrufbar unter: https://www.vischer.com/know-how/blog/teil-19-sprachmodelle-mit-und-ohne-personenbezogene-daten/ (letzter Aufruf: 24.11.2024)

offengelegt.[177] Allerdings können sich, je nach Anwendungsbereich des LLM, mehr oder minder schwere Auswirkungen für die Rechte und Freiheiten der betroffenen Personen in der nachgelagerten Anwendungsphase des LLM ergeben. Studien haben bspw. gezeigt, dass im Rahmen von training data extraction attacks in Trainingsdaten enthaltene Daten zu einzelnen Personen, wie der Name, die E-Mail-Adresse, die Telefonnummer sowie die physische Adresse, offengelegt werden können.[178] In solchen Fällen kommt es aus technischer Sicht strenggenommen nicht zu einer Offenlegung der Trainingsdaten, sondern zu einer Rekonstruktion von Texten auf Basis der erlernten Wahrscheinlichkeitswerte.[179] Die Schwere der Auswirkungen einer Rekonstruktion von Daten, die sich auf eine identifizierbare Person beziehen, auf die Rechte und Freiheiten des jeweiligen Datensubjekts, hängt insbesondere von der Art bzw. dem jeweiligen Informationsgehalt der rekonstruierten Daten ab. Folglich hängt das Risiko maßgeblich von den Trainingsdaten und etwaigen im Rahmen des Trainings umgesetzten Schutzmaßnahmen ab. Der Umstand, dass es der datenverarbeitenden Stelle nicht auf einen etwaigen personenbezogenen Informationsgehalt der Trainingsdaten ankommt, neutralisiert die potenziellen Risiken für die Rechte und Freiheiten der Datensubjekte dabei nicht vollständig. Vielmehr sind wirksame Sicherheitsmaßnahmen erforderlich, um potenziellen negativen Auswirkungen für Datensubjekte entgegenzuwirken. Insofern könnte auch argumentiert werden, dass die Bewertung des Personenbezugs von Trainingsdaten ausschließlich aus der Perspektive der KI-Hersteller bzw. nach der relativen Theorie, dem Schutzziel der DSGVO in diesem Fall entgegenläuft und eine weniger differenzierte Betrachtung erforderlich wäre.

4.2 Zwischenfazit

Die Frage nach der Anwendbarkeit der DSGVO auf das Training von LLMs hängt maßgeblich davon ab, ob die für das Training verwendeten Datensätze einen Personenbezug i. S. d. Art. 4 Nr. 1 DSGVO aufweisen. Man muss davon ausgehen, dass

[177] Vgl. Bartels (2024), A Balancing Act: Data Protection Compliance of Artificial Intelligence, S. 532

[178] Vgl. Carlini et al. (2021), Extracting Training Data from Large Language Models

[179] Vgl. Hamburgischer Beauftragter für Datenschutz und Informationsfreiheit (2024), Diskussionspapier: Large Language Models und personenbezogene Daten v. 15. Juli 2024; Moos (2024), Personenbezug von Large Language Models, Rn. 2

typische Datensätze für das Training von LLMs Informationen enthalten, die sich auf identifizierbare natürliche Personen beziehen. Auch wenn die Möglichkeit einer Identifikation von Datensubjekten durch KI-Hersteller grundsätzlich bestünde, lassen sich unter Anwendung des relativen Begriffs des Personenbezugs und unter Berücksichtigung der technischen Herausforderungen in der Praxis sowie der subjektiven Interessen von KI-Herstellern Argumente für die Verneinung eines Personenbezugs von Trainingsdaten finden.

Letztendlich ist der Trainingsprozess dennoch für die Schwere der potenziellen Risiken für die Rechte und Freiheiten der Datensubjekte maßgeblich. Dies ist auch dann der Fall, wenn nicht zu erwarten ist, dass eine Identifikation der Datensubjekte durch den KI-Hersteller erfolgt. Insofern erscheint ein Ergebnis, nach welchem die DSGVO auf das Training von LLMs keine Anwendung finden würde, unbefriedigend und wirft die Frage auf, ob die starre Abhängigkeit der Anwendbarkeit der DSGVO von dem Merkmal des Personenbezugs noch zeitgemäß ist. Abgesehen vom Grundrecht zum Schutz personenbezogener Daten (Art. 8 GRCh), ist auch das Grundrecht zum Schutz des Privat- und Familienlebens (Art. 7 GRCh) Schutzgut der DSGVO. Der Sachverhalt des Trainings von LLMs stellt ein Beispiel dafür dar, dass sich Risiken für die Privatsphäre natürlicher Personen auch ergeben können, wenn Daten verarbeitet werden, die aus Sicht der datenverarbeitenden Stelle keinen Personenbezug aufweisen. Insofern Bedarf es ggf. einer Neuausrichtung der DSGVO, weg von einer starren Ausrichtung auf das Schutzgut der personenbezogenen Daten und hin zu einer verstärkten Berücksichtigung der tatsächlichen Risiken für die Privatsphäre natürlicher Personen. Gegebenenfalls sollte der Datenschutz weniger als restriktiver Selbstzweck und eher als innovationsoffenes Instrument zum Schutz der Privatsphäre natürlicher Personen verstanden und ausgelegt werden.

4.3 Besondere Kategorien personenbezogener Daten in Trainingsdaten

Wenn man den Personenbezug von Trainingsdaten bejahen sollte, stellt sich die Folgefrage, inwiefern die verschärften Anforderungen des Art. 9 DSGVO für Verarbeitungen von sog. besonderen Kategorien personenbezogener Daten beachtet werden müssen. Welche Arten von potenziell personenbezogenen Daten in einem Trainingsdatensatz enthalten sind, hängt von den Quellen ab, aus denen die Trainingsdaten bezogen werden. Qualitativ hochwertige Trainingsdaten sind divers und

enthalten idealerweise Daten aus verschiedenen Quellen, wie z. B. Webseiten, Social Media, Wikipedia-Artikeln, Büchern sowie akademischer Literatur aus verschiedenen Fachbereichen (z. B. Rechtswissenschaften oder Medizin).[180] Aufgrund des typischerweise hohen Grads an Diversität und des großen Umfangs von Trainingsdatensätzen, ist grundsätzlich denkbar, dass auch besondere Kategorien personenbezogener Daten, wie Informationen über die Gesundheit, die sexuelle Orientierung oder die rassische und ethnische Herkunft identifizierbarer Personen, in den Trainingsdate enthalten sind. Selbst, wenn in einem Trainingsdatensatz keine unmittelbar sensiblen Informationen i. S. d. Art. 9 Abs. 1 DSGVO enthalten sein sollten, können Studien zu Folge von einem LLM ggf. memorisierte demografische Informationen über identifizierbare Personen, die in den Trainingsdaten enthalten waren und im Rahmen der Nutzung des LLM regeneriert werden, genutzt werden, um Informationen über politische Ansichten oder die wirtschaftliche Situation der Datensubjekte abzuleiten.[181] Nach Rechtsprechung des EuGH, kann der Anwendungsbereich des Art. 9 DSGVO bereits dann eröffnet sein, wenn Daten verarbeitet werden, die dazu geeignet sind, sensible Informationen indirekt zu offenbaren.[182]

Wenn man diese Rechtsprechung des EuGH jedoch undifferenziert anwendet, würde man den Anwendungsbereich des Art. 9 DSGVO unverhältnismäßig weit ausdehnen. So wäre Art. 9 DSGVO bereits bei der Verarbeitung eines Namens anzunehmen, aus dem sich ggf. die ethnische Herkunft des Datensubjekts erschließen lässt. Eine solche Auslegung würde dazu führen, dass die verschärften Anforderungen des Art. 9 DSGVO, welche ein besonderes Risikopotential einer Datenverarbeitung für die Rechte und Freiheiten einer betroffenen Person mitigieren sollen, auch bei Datenverarbeitungen zur Anwendung kommen könnten, die faktisch kein hohes Risikopotential aufweisen.

[180] Vgl. Longpre et al. (2023), A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality,& toxicity, S. 4

[181] Vgl. Plant et al. (2024), You Are What You Write: Author Re-identification Privacy Attacks in the Era of Pre-trained Language Models. Computer Speech & Language

[182] Vgl. EuGH, Urteil v. 01.08.2022 – C-184/20; EuGH, Urteil v. 04.07.2023 – C-252/21

Dies erscheint im Hinblick auf den risikobasierten Ansatz der DSGVO nicht sachgerecht.[183] Aus der Rechtsprechung des EuGH geht auch hervor, dass die Bewertung der Anwendbarkeit des Art. 9 DSGVO maßgeblich von den Risiken für die Rechte und Freiheiten der betroffenen Personen abhängt, welche sich wiederum aus dem konkreten Verarbeitungstext ergeben.[184] Die DSGVO ist auch im Hinblick auf die primärrechtlichen Regelungen auszulegen, insbesondere unter Berücksichtigung der GRCh. Demnach stehen insbesondere das Grundrecht auf Achtung des Privat- und Familienlebens (Art. 7 GRCh) und das Grundrecht auf Schutz personenbezogener Daten (Art. 8 GRCh) der unternehmerischen Freiheit (Art. 16 GRCh) entgegen. Folglich ist unter dem Grundsatz der Verhältnismäßigkeit ein Ausgleich zwischen den sich gegenüberstehenden Rechtspositionen zu finden. Eine Auslegung des Art. 9 DSGVO streng nach dem Wortlaut der Norm ohne Berücksichtigung des konkreten Verarbeitungskontexts kann zu einer unverhältnismäßigen Intensität des Eingriffs in die unternehmerische Freiheit führen.[185] Einige Stimmen in der Literatur halten deshalb die Anwendung eines kontextorientierten Ansatzes als sachgerecht.[186] Dabei ist unter Berücksichtigung des konkreten Verarbeitungszwecks, der subjektiven Auswertungsabsicht der verarbeitenden Stelle sowie dem Umstand, ob sich die verarbeitende Stelle den sensiblen Informationsgehalt tatsächlich zu Nutze macht, zu prüfen, ob der potenziell sensible Informationsgehalt derart in den Verarbeitungsprozess einfließt, dass sich daraus ein gesteigertes Risiko für die Rechte und Freiheiten der betroffenen Personen ergibt.[187] Ein

[183] Vgl. Kohn/Schleper (2023), Die (zufällige) Erhebung sensibler Daten, S. 725

[184] Vgl. EuGH, Urteil v. 04.07.2023 – C-252/21, Rn. 73; Vgl. EuGH, Urteil v. 01.08.2022 – C-184/20, Rn. 126; Vgl. EuGH, Urteil v. 24.09.2019 – C-136/17, Rn. 44

[185] Vgl. Britz/Indenhuck/Langerhans (2021), Die Verarbeitung „zufällig" sensibler Daten, S. 560

[186] Vgl. Kohn/Schleper (2023), Die (zufällige) Erhebung sensibler Daten, S. 723; Britz/Indenhuck/Langerhans (2021), Die Verarbeitung „zufällig" sensibler Daten, S. 559; EDSA (2020), Leitlinien 3/2019 zur Verarbeitung personenbezogener Daten durch Videogeräte, Version 2.0, Rn. 62-72; Matejek/Kremer (2023), Sensible Daten soweit das Auge reicht: Der EuGH zapft den „hidden layer" an, S. 218

[187] Vgl. Kohn/Schleper (2023), Die (zufällige) Erhebung sensibler Daten, S. 723; Britz/Indenhuck/Langerhans (2021), Die Verarbeitung „zufällig" sensibler Daten, S. 559; EDSA (2020), Leitlinien 3/2019 zur Verarbeitung personenbezogener Daten durch Videogeräte, Version 2.0, Rn. 62-72;

gesteigertes Risiko kann regelmäßig angenommen werden, wenn sich die verarbeitende Stelle den sensiblen Informationsgehalt gezielt zu Nutze macht und/oder eine hinreichende Wahrscheinlichkeit dafür besteht, dass der sensible Informationsgehalt offengelegt wird. So ist ein gesteigertes Risiko eher anzunehmen, wenn unmittelbar sensible Informationen verarbeitet werden als bei einer Verarbeitung von Daten, aus denen ein sensibler Informationsgehalt nur mittelbar abgeleitet werden könnte.[188] Fehlt es an einer zielgerichteten Nutzung sensibler Informationsinhalte und besteht keine hinreichende Wahrscheinlichkeit für eine Offenlegung solcher Informationsgehalte (z. B. weil angemessene technische und organisatorische Maßnahmen umgesetzt werden), sodass kein gesteigertes Risiko für die Rechte und Freiheiten der betroffenen Personen anzunehmen ist, ist die Anwendung des Verarbeitungsverbots des Art. 9 Abs. 1 DSGVO nach der kontextorientierten Auslegung der EuGH-Rechtsprechung zu verneinen.[189] Ein solcher risikobasierter und kontextorientierter Ansatz erscheint geeignet, um einen primärrechtskonformen, angemessenen Ausgleich zwischen den sich gegenüberstehenden Interessen der verarbeitenden Stelle und der betroffenen Personen zu schaffen.

Demnach kommt es für die Bewertung der Anwendbarkeit des Art. 9 DSGVO beim Training von LLMs insbesondere auf den konkreten Anwendungszweck des jeweiligen LLMs an. Soll bspw. ein LLM für einen Chatbot zur Transkription von Patientengesprächen für Arztpraxen entwickelt werden, kommt es dem Hersteller im Rahmen des Trainings ggf. gerade auf den sensiblen Informationsgehalt der Trainingsdaten an, sodass sich ggf. besondere Risiken für die Rechte und Freiheiten der betroffenen Person ergeben können. In anderen Fällen hingegen wird weder eine Auswertungsabsicht des KI-Herstellers hinsichtlich des potenziell sensiblen Informationsgehalts vorliegen noch wird eine hinreichende Wahrscheinlichkeit vorliegen,

---

Matejek/Kremer (2023), Sensible Daten soweit das Auge reicht: Der EuGH zapft den „hidden layer“ an, S. 218

[188] Vgl. Britz/Indenhuck/Langerhans (2021), Die Verarbeitung „zufällig“ sensibler Daten, S. 562

[189] Vgl. Kohn/Schleper (2023), Die (zufällige) Erhebung sensibler Daten, S. 723;
Britz/Indenhuck/Langerhans (2021), Die Verarbeitung „zufällig“ sensibler Daten, S., 559; EDSA (2020), Leitlinien 3/2019 zur Verarbeitung personenbezogener Daten durch Videogeräte, Version 2.0, Rn. 62-72, Matejek/Kremer (2023), Sensible Daten soweit das Auge reicht: Der EuGH zapft den „hidden layer“ an, S. 218

dass über die in den Trainingsdaten enthaltenen personenbezogenen Daten ein Rückschluss auf sensible Informationen erfolgt. Auch wenn es denkbar ist, dass demografische Informationen, die in Trainingsdaten enthalten waren, im Rahmen von training data extraction attacks regeneriert werden und daraus Informationen über politische Ansichten der jeweiligen Datensubjekte abgeleitet werden könnten, ist in den allermeisten Fällen nicht von einer hinreichenden Wahrscheinlichkeit auszugehen, dass sich dieses Szenario tatsächlich verwirklicht, insbesondere, wenn angemessene Schutzmaßnahmen umgesetzt wurden.[190] Vermutungen und Spekulationen, über die Möglichkeit der Ableitung sensibler Informationsinhalte i. S. d. Art. 9 DSGVO ohne Bezug zu dem konkreten Verarbeitungskontext, reichen für die Annahme einer Verarbeitung besondere Kategorien personenbezogener Daten nicht aus.[191] Auch wenn das Verarbeitungsverbot des Art. 9 Abs. 1 DSGVO nicht zur Anwendung kommt kann angenommen werden, dass die Rechte und Freiheiten der betroffenen Personen hinreichend durch die Anwendung der allgemeinen Grundsätze der DSGVO geschützt sind. Die datenverarbeitende Stelle hat demnach ohnehin technische und organisatorische Schutzmaßnahmen umzusetzen, die im Hinblick auf die konkrete Datenverarbeitung angemessen und wirksam sind.[192] Demnach sollten in Sachverhalten, in denen sich die verarbeitende Stelle potenziell sensible Informationsgehalte der Trainingsdaten nicht zu Nutze machen möchte, Maßnahmen umgesetzt werden, die eine (ggf. zufällige) Verarbeitung oder Offenlegung sensibler Informationsgehalte verhindern.[193]

Die Anwendbarkeit des Verarbeitungsverbots nach Art. 9 Abs. 1 DSGVO auf das Training von LLMs erfordert eine differenzierte Betrachtung, die den konkreten Verarbeitungskontext berücksichtigt. Während aufgrund der Diversität und des Umfangs von Trainingsdatensätzen grundsätzlich denkbar ist, dass besondere Kategorien personenbezogener Daten enthalten sind, zeigt sich, dass eine pauschale Anwendung der verschärften Anforderungen des Art. 9 DSGVO häufig nicht sachgerecht wäre. Eine

[190] Vgl. Kohn/Schleper (2023), Die (zufällige) Erhebung sensibler Daten, S. 725

[191] Vgl. Matejek/Kremer (2023), Sensible Daten soweit das Auge reicht: Der EuGH zapft den „hidden layer“ an, S. 223

[192] Vgl. Art. 32 DSGVO

[193] Vgl. Britz/Indenhuck/Langerhans (2021), Die Verarbeitung „zufällig“ sensibler Daten, S. 562

pauschale Annahme der Anwendbarkeit des Art. 9 DSGVO im Kontext mit dem Training von LLMs wird weder der Rechtsprechung des EuGH noch den Prinzipien der DSGVO gerecht. Vielmehr ist eine differenzierte Bewertung unter Berücksichtigung des Verarbeitungskontexts erforderlich, um eine ausgewogene, primärrechtskonforme Lösung zu gewährleisten, die sowohl den Schutz der Rechte der betroffenen Personen als auch die unternehmerische Freiheit wahrt.

5 Angemessene technische und organisatorische Maßnahmen beim Training von Large Language Models

Wie in den vorangegangenen Kapiteln bereits angemerkt, hängt die Bewertung der datenschutzrechtlichen Zulässigkeit des Trainings von LLMs maßgeblich von den Umständen des jeweiligen Einzelfalls und der Umsetzung angemessener Schutzmaßnahmen ab. Insbesondere technische und organisatorische Maßnahmen spielen eine entscheidende Rolle, um die Risiken für die Rechte und Freiheiten betroffener Personen zu minimieren und zugleich die Anforderungen der DSGVO zu erfüllen. Art. 32 DSGVO verpflichtet die verantwortliche Stelle ein Schutzniveau zu gewährleisten, dass den Risiken des jeweiligen Sachverhalts angemessen ist. Ferner erfordern der Grundsatz der Datenminimierung sowie der Zweckbindungsgrundsatz, dass eine Verarbeitung personenbezogener Daten, auf das unbedingt notwendige Maß beschränkt wird (vgl. Kapitel 3.5 und ErwGr. 39 Satz 1 DSGVO).

Vor diesem Hintergrund sollen im vorliegenden Kapitel ein paar ausgewählte technische und organisatorische Maßnahmen für das Training von LLMs nach dem aktuellen Stand der Technik vorgestellt werden. Dabei begrenzen wir uns auf Maßnahmen, die Risiken hinsichtlich der Vertraulichkeit personenbezogener Daten und insb. die sog. training data extraction attacks adressieren und verzichten auf einen vollständigen Überblick möglicher Sicherheitsbedrohungen für LLMs (u. a. membership inference attacks, model inversion attacks, backdoor attacks, model poisoning attacks, contextual privacy attacks etc.)[194] und der jeweiligen Sicherheitsmaßnahmen, um den Rahmen dieser Arbeit nicht zu sprengen. Insbesondere Risiken hinsichtlich der Integrität, Verfügbarkeit und Belastbarkeit eines LLMs sowie Schutzmaßnahmen im Rahmen der Inferenz-Phase (bspw. Output-Filter)

[194] Vgl. Feretzakis et al. (2024), Privacy-Preserving Techniques in Generative AI and Large Language Models: A Narrative Review, S. 2

werden außen vorgelassen. Für detaillierte und vollständige Ausführungen wird auf die einschlägige computerwissenschaftliche Fachliteratur verwiesen.

5.1 Risiken für die Rechte und Freiheiten betroffener Personen

Ausgangspunkt für die Bewertung der Angemessenheit und Geeignetheit technischer und organisatorischer Maßnahmen i. S. d. DSGVO sind die potenziellen Risiken, die sich aus der jeweiligen Datenverarbeitung für die Rechte und Freiheiten der betroffenen Personen ergeben. Das Training von LLMs kann vor allem dann zu einer Verletzung der Rechte und Freiheiten der Datensubjekte führen, wenn personenbezogene Informationen unberechtigterweise, während der Inferenz-Phase offengelegt bzw. rekonstruiert werden (vgl. hierzu Kapitel 2.4 und 2.5). Das Risiko einer unberechtigten Offenlegung kann zudem auch während des Trainingsprozesses bestehen, z. B. wenn dieser an einen Dienstleister ausgelagert wird. Die Schwere eines solchen Eingriffs in die Rechte und Freiheiten der betroffenen Personen hängt maßgeblich vom Informationsgehalt ab, der offengelegt wird. Wenn sensible Informationen zum Gesundheitszustand, zur sexuellen Orientierung oder zu politischen Ansichten einer Person offengelegt werden, kann neben dem Risiko einer Verletzung des Grundrechts auf Schutz personenbezogener Daten (vgl. Art. 8 GRCh) auch das Risiko einer Verletzung des Grundrechts auf Achtung des Privat- und Familienlebens (vgl. Art. 7 GRCh) oder das Risiko einer Verletzung des Grundrechts auf Nichtdiskriminierung (vgl. Art. 21 GRCh) bestehen.

Um die Angemessenheit bestimmter Sicherheitsmaßnahmen zu bewerten, sind vor allem die Schwere der potenziellen Risiken sowie deren Eintrittswahrscheinlichkeit zu berücksichtigen. Die Schwere der potenziellen Risiken hängt wiederum von Art, Umfang, Zweck sowie den Umständen der jeweiligen Datenverarbeitung ab.[195] Studien zu Folge hängt die Eintrittswahrscheinlichkeit der vorgenannten Risiken beim Training von LLMs wesentlich von Maßnahmen im Rahmen der Trainingsdatenaufbereitung sowie Maßnahmen im Rahmen des Fine Tunings ab. So kann ein übermäßiges Fine Tuning bspw. dazu führen, dass ein Modell zu stark auf spezifische Details der Trainingsdaten ausgerichtet wird und somit die Gefahr einer Veröffentlichung bzw. Rekonstruktion von

[195] Vgl. ErwGr. 76 DSGVO

Trainingsdaten erhöht wird (sog. Overfitting).[196] Dem Memorisieren von Trainingsdaten sowie dem Overfitting kann durch Anwendung verschiedener sog. Privacy Enhancing Technologies (PETs), z. B. sog. Differential Privacy-Ansätze, entgegengewirkt und die Eintrittswahrscheinlichkeit des Risikos einer Offenlegung personenbezogener Informationen verringert werden.

5.2 Maßnahmen während der Trainingsdatenaufbereitung

Die Aufbereitung der Trainingsdaten, bevor Sie für das eigentliche Training des LLMs genutzt werden, wird regelmäßig der erste Prozessschritt für die Anwendung von Sicherheitsmaßnahmen bzw. von PETs sein (vgl. Kap. 2.3.2). Ansätze, um das Risiko einer Offenlegung bzw. Rekonstruktion von personenbezogenen Daten zu mitigieren sind in dieser Phase des LLM-Trainings häufig (1.) die Identifizierung sensibler Attribute in den Trainingsdatensätzen und deren teilweise oder vollständige Unkenntlichmachung sowie (2.) das Hinzufügen von sog. Rauschen oder „Noise“ über sog. Differential Privacy-Methoden.[197]

5.2.1 Unkenntlichmachung von potenziell personenbezogenen Attributen

Sensible oder potenziell personenbezogene Attribute in Trainingsdatensätzen können bspw. über sog. Named Entity Recogniser (NER) identifiziert werden. NERs sind in der Regel LLMs, die darauf spezialisiert sind, bestimmte Entitäten in einem Text zu erkennen und zu klassifizieren (z. B. Personennamen, Ortsnamen, Organisationen, Zeitangaben etc.; vgl. hierzu Abbildung 3).[198] Den besten NER wird eine Genauigkeit von etwa 90% nachgesagt. Dabei spielen die grammatikalische Struktur der Textdaten sowie darin etwaig enthaltene Typos eine wichtige Rolle für die Genauigkeit eines NERs.[199] Auf

[196] Vgl. Carlini et al. (2023), Quantifying Memorization Across Neural Language Models; Vgl. Mireshghallah et al. (2022), An Empirical Analysis of Memorization in Fine-tuned Autoregressive Language Models, S.1816–1826

[197] Vgl. El Mestari et al. (2024), Preserving Data Privacy in Machine Learning Systems, Computers & Security, S. 8

[198] Vgl. El Mestari et al. (2024), Preserving Data Privacy in Machine Learning Systems, Computers & Security, S. 9

[199] Vgl. El Mestari et al. (2024), Preserving Data Privacy in Machine Learning Systems, Computers & Security, S. 9

identifizierte potenziell personenbezogene Attribute können dann Methoden der sog. k-Anonymität angewendet werden, um diese unkenntlich zu machen. Methoden der k-Anonymität zielen darauf ab, dass Attribute, die sich auf eine bestimmte Person beziehen, nicht von einer Mindestanzahl (k) an Attributen, die sich auf andere Personen beziehen, unterschieden werden können.[200] Hierfür werden die identifizierten potenziell personenbezogenen Attribute bspw. entfernt oder durch andere, generische Attribute ersetzt. Die k-Anonymität kann um weitere Konzepte der Anonymisierung, wie der sog. I-Diversität und der sog. t-Geschlossenheit, ergänzt werden.[201] Ein prominentes Open Source-Tool, welches in den Workflow der Trainingsdatenaufbereitung implementiert werden kann, um sensible oder personenbezogene Informationen in unstrukturierten Textdaten zu identifizieren und zu ersetzen, ist Microsoft Presidio.[202] Die Identifizierung und Unkenntlichmachung potenziell personenbezogener Attribute haben das Potential einer Anonymisierungsmethode, zumindest nach der relativen Theorie des Personenbezugs (vgl. hierzu Kap. 4.1.2).[203] Ein erheblicher Nachteil ist jedoch, dass mit dem Entfernen oder Ersetzen der relevanten Attribute ein Informationsverlust (sog. utility loss) einhergeht, der sich i. d. R. negativ auf die Performanz des zu entwickelnden LLMs auswirkt.[204]

[200] Vgl. Schwartmann, Jaspers, Lepperhoff, Weiß, Meier (2022), Praxisleitfaden zum Anonymisieren personenbezogener Daten: Anforderungen, Einsatzklassen und Vorgehensmodell, S. 33, 34

[201] Vgl. El Mestari et al. (2024), Preserving Data Privacy in Machine Learning Systems, S. 8

[202] Vgl. Vgl. Feretzakis et al. (2024), Privacy-Preserving Techniques in Generative AI and Large Language Models: A Narrative Review, S. 9; s. hierzu auch Chen (2022), Building a Customized PII Anonymizer with Microsoft Presidio

[203] Vgl. El Mestari et al. (2024), Preserving Data Privacy in Machine Learning Systems, S. 9

[204] Vgl. El Mestari et al. (2024), Preserving Data Privacy in Machine Learning Systems, S. 8

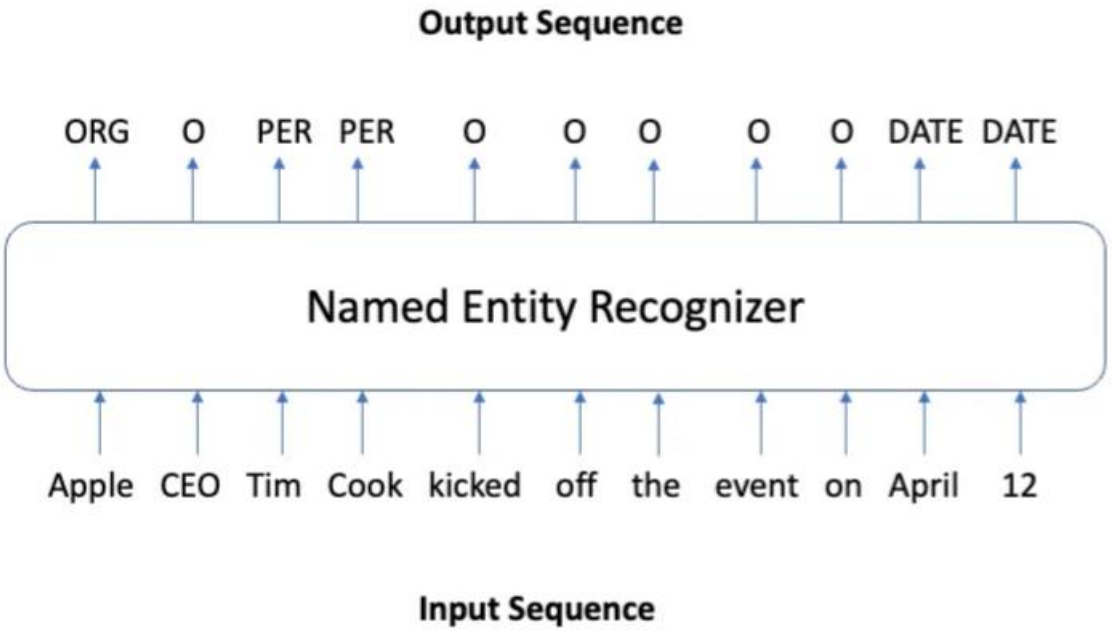

Abbildung 3: stark vereinfachte schematische Darstellung der Funktionsweise eines Named Entity Recognizer

5.2.2 Differential Privacy

Differential Privacy-Maßnahmen sollen gewährleisten, dass ein Datensatz nicht mehr Informationen über eine bestimmte Person preisgibt, als wenn keine Informationen über diese Person in dem Datensatz enthalten wären.[205] Dabei wird entweder dem Trainingsdatensatz im Rahmen der Trainingsdatenaufbereitung oder dem Output sog. Rauschen in gewissem Maß hinzugefügt, um zu verhindern, dass im Rahmen von training data extraction attacks Rückschlüsse auf die Zugehörigkeit einzelner Datenpunkte gemacht werden können. Dies kann z. B. durch das Entfernen oder durch Randomisierung von Datenpunkten - sog. Subsampling (ein zufälliger Anteil des Trainingsdatensatzes wird mit einer bestimmten Wahrscheinlichkeit für das Training verwendet) - oder das Einbringen von sorgfältig kalibriertem Rauschen erfolgen.[206] Auch mit der Anwendung von Differential Privacy-Ansätzen geht allerdings ein entsprechender

[205] Vgl. Royal Society (2019), Protecting Privacy in Practice: The Current Use, Development and Limits of Privacy Enhancing Technologies in Data Analysis, S. 41

[206] Vgl. Royal Society (2019), Protecting Privacy in Practice: The Current Use, Development and Limits of Privacy Enhancing Technologies in Data Analysis, S. 43

utility loss einher, welcher umso drastischere Auswirkungen hat, je kleiner der jeweilige Trainingsdatensatz ist. Je größer ein Datensatz ist, desto weniger Einfluss hat das Rauschen auf das Ergebnis.[207]

5.3 Maßnahmen während des Pre-Trainings

Sicherheitsmaßnahmen, die während der Phase des Pre-Trainings angewendet werden können, sind bspw.

(i) Störungstechniken, die dem Trainingsalgorithmus selbst hinzugefügt werden, um Differential Privacy umzusetzen;

(ii) Training auf verschlüsselten Daten (z. B. sog. homomorphic encryption oder secure multi-party computation (SMPC));

(iii) datenschutzfreundliche Architekturen, wie sog. Federated Learning oder sog. PATE (private aggregation of teacher ensembles); und

(iv) Training auf datenschutzfreundlichen Hardwarelösungen, wie sog. trusted operating environments.[208]

Kryptografische Verfahren, wie die sog. homomorphe Verschlüsselung und secure multi-party computation, ermöglichen Rechenoperationen auf verschlüsselten Daten und adressieren damit das Risiko einer Offenlegung von Trainingsdaten während des Trainingsprozesses (bspw. im Rahmen von dezentralen Trainingsansätzen).[209] Herkömmliche Verschlüsselungstechniken, wie z. B. der sog. Advanced Encyption Standard (AES), gewährleisten zwar eine Verschlüsselung während der Datenübertragung und -speicherung, eignen sich jedoch nicht für Rechenoperationen, wie bspw. das Training von LLMs. Wenn man mit AES verschlüsselte Daten für das LLM-Training nutzen würde, wäre das Ergebnis unbrauchbar. Man müsste die verschlüsselten Daten vor

[207] Vgl. Royal Society (2019), Protecting Privacy in Practice: The Current Use, Development and Limits of Privacy Enhancing Technologies in Data Analysis, S. 43

[208] Vgl. Royal Society (2019), Protecting Privacy in Practice: The Current Use, Development and Limits of Privacy Enhancing Technologies in Data Analysis, S. 10

[209] Vgl. Royal Society (2019), Protecting Privacy in Practice: The Current Use, Development and Limits of Privacy Enhancing Technologies in Data Analysis, S. 10

dem Training entschlüsseln.[210] Die Homomorphe Verschlüsselung basiert – wie der Name schon sagt – auf einem Homomorphismus. Über Homomorphismen lassen sich die (mathematischen) Strukturen der unverschlüsselten Originaldaten in den verschlüsselten Daten beibehalten bzw. „strukturtreu übersetzen", sodass sich die verschlüsselten Daten hinsichtlich derer statistischen Struktur ebenso verhalten, wie die unverschlüsselten Originaldaten.[211] Für Rechenoperationen mit Daten, die über homomorphe Verschlüsselung verschlüsselt wurden bedeutet dies, dass das Ergebnis der Rechenoperation verschlüsselt bleibt und eine verschlüsselte Version des Ergebnisses der gleichen Rechenoperation mit den unverschlüsselten Daten darstellt.[212]

Eine Ergänzung der homomorphen Verschlüsselung ist die sog. secure multi-party computation (SMPC). SMPC-Protokolle ermöglichen sog. private distributed computations. Dabei können mehrere Parteien Berechnungen und Analysen anhand von zusammengeführten Daten im Verbund vornehmen, ohne die jeweils eigenen Eingabedaten gegenüber den anderen Parteien offenzulegen. Die beteiligten Parteien teilen miteinander nur die Ergebnisse ihrer Berechnungen und Analysen, jedoch keine Informationen über die Eingabedaten der anderen Parteien, da die Berechnungen auf verschlüsselten Eingabedaten erfolgen.[213] SMPC erlaubt es somit mehreren Parteien ein gemeinsames LLM dezentral zu trainieren (sog. Federated Learning), ohne die Trainingsdaten direkt miteinander zu teilen (sog. Secured Federated Learning). Der Begriff Federated Learning beschreibt grds. eine dezentralisierte Trainingsarchitektur, die darauf basiert, dass Kopien eines LLMs auf separaten Instanzen lokal trainiert und deren Parameter sodann über gezielte Updates des Zielmodells zusammengeführt werden.[214] Ein weiterer dezentraler Trainingsansatz ist die sog. private aggregation of teacher

[210] Vgl. El Mestari et al. (2024), Preserving Data Privacy in Machine Learning Systems, S. 10

[211] Wikipedia (2024), Homomorphismus, aufrufbar unter: https://de.wikipedia.org/wiki/Homomorphismus (letzter Aufruf: 17.12.2024)

[212] Vgl. El Mestari et al. (2024), Preserving Data Privacy in Machine Learning Systems, S. 10

[213] Vgl. Royal Society (2019), Protecting Privacy in Practice: The Current Use, Development and Limits of Privacy Enhancing Technologies in Data Analysis, S. 29; Bierbauer/Helminger (2023), Offenlegung von Daten unter Wahrung der Privatsphäre mittels SMPC, S. 7-9

[214] Vgl. Liu et al. (2021), When Machine Learning Meets Privacy: A Survey and Outlook, S. 5

ensembles (PATE), wobei mehrere sog. Teacher Models anhand von sensiblen Datensätzen auf verschiedenen Instanzen trainiert werden, um ein sog. Student Model auf Basis einer „verrauschten“ Aggregation aller Outputs der Teacher Models zu trainieren.[215]

Ein erheblicher Nachteil der beschriebenen Maßnahmen ist, dass diese noch mehr Rechenkapazitäten erfordern, als das herkömmliche Training ohnehin schon erfordert. Insbesondere das Training mit verschlüsselten Daten macht den Trainingsprozess entsprechend langsamer und das Modell weniger präzise.[216] Folglich besteht die Herausforderung darin, ein Schutzniveau zu gewährleisten, welches in einem angemessenen Verhältnis zu den Berechnungsressourcen (Zeit und Rechenleistung) sowie der Aussagekraft der Trainingsdaten steht.

5.4 Zwischenfazit

Nach Art. 25 Abs. 1 DGSVO sowie nach Art. 32 Abs. 1 DSGVO ist die Angemessenheit einer Schutzmaßnahme insbesondere unter Berücksichtigung des Stands der Technik, der Implementierungskosten und der Eintrittswahrscheinlichkeit und Schwere der mit der Verarbeitung verbundenen Risiken für die Rechte und Freiheiten natürlicher Personen zu bewerten. Die Anwendung von PETs beim Training von LLMs erfordert stets einen Kompromiss zwischen dem Schutz personenbezogener Informationen und der utility, sprich der Performanz des LLMs.[217] Im Hinblick auf die verschiedenen PETs können die utility-trade offs verschiedener Natur sein. So muss man bei der Anwendung von Differential Privacy-Ansätzen i. d. R. Einbußen bei der Genauigkeit des LLMs hinnehmen, während der Nachteil beim Training mit verschlüsselten Daten und beim Federated Learning in den gesteigerten Ressourcen (Zeit und Rechenleistung) liegt.[218]

[215] Vgl. Xue et al. (2020), Machine Learning Security: Threats, Countermeasures and Evaluations, S. 74720–74742

[216] Vgl. Xue et al. (2020), Machine Learning Security: Threats, Countermeasures and Evaluations, S. 74736; Liu et al. (2021), When Machine Learning Meets Privacy: A Survey and Outlook, S. 12; El Mestari et al. (2024), Preserving Data Privacy in Machine Learning Systems, S. 106

[217] Vgl. Xue et al. (2020), Machine Learning Security: Threats, Countermeasures and Evaluations, S. 74736

[218] Vgl. Royal Society (2019), Protecting Privacy in Practice: The Current Use, Development and Limits of Privacy Enhancing Technologies in Data Analysis

Folglich besteht die Herausforderung darin, ein Schutzniveau zu gewährleisten, welches in einem angemessenen Verhältnis zur Schwere und Eintrittswahrscheinlichkeit der mit der Verarbeitung verbundenen Risiken sowie den erforderlichen Ressourcen (Zeit und Rechenleistung ≙ Kosten) und der Performanz des Modells steht.[219]

Zu berücksichtigen ist ferner, dass es keine allgemeinwirksamen „off-the-shelf"-PETs gibt, die ein angemessenes Schutzniveau gewährleisten. Angemessene Sicherheitsmaßnahmen müssen kontextbezogen für den jeweiligen Einzelfall evaluiert und ausgewählt werden. Die Angemessenheit einer Sicherheitsmaßnahme hängt aus technischer Sicht insb. davon ab, welches Risikoszenario adressiert werden soll, in welcher Trainingsphase die Maßnahme implementiert werden soll sowie, ob ein akzeptabler Kompromiss zwischen Schutzniveau und utility erzielt werden kann. So bestehen die Sicherheitsgarantien der verschiedenen PETs nur unter der Annahme einer Reihe von Prämissen (sog. trust assumptions), welche sich bspw. auf die Fähigkeiten der potenziellen Angreifer beziehen. Folglich kann eine Anwendung von State-of-the-Art-PETs unwirksam sein, wenn bestimmte Schlüsselfaktoren außer Acht gelassen werden.[220]

Die Forschung und Entwicklung von PETs im Bereich Machine Learning steckt aktuell noch in den Kinderschuhen, wobei rasche Entwicklungen zu erwarten sind.[221] Im Hinblick auf die sich stetig weiterentwickelnden generativen KI-Systeme – mit steigender Leistungsfähigkeit von KI-Systemen nehmen auch die verschiedenen Risikoszenarien zu – werden Innovation und kontinuierliche Optimierung erforderlich sein, um der Praxistauglichkeit von PETs und regulatorischen Anforderungen gerecht zu werden.[222] Dabei wird es vor allem um die Suche nach dem optimalen Gleichgewicht zwischen Effizienz, Performanz und Datenschutz gehen.[223]

[219] Vgl. Feretzakis et al. (2024), Privacy-Preserving Techniques in Generative AI and Large Language Models: A Narrative Review, Information, 15(11), Article 697, S. 1, doi:10.3390/info15110697

[220] Vgl. El Mestari et al. (2024), Preserving Data Privacy in Machine Learning Systems, S. 1

[221] Vgl. El Mestari et al. (2024), Preserving Data Privacy in Machine Learning Systems, S. 1

[222] Vgl. Feretzakis et al. (2024), Privacy-Preserving Techniques in Generative AI and Large Language Models: A Narrative Review, S. 1-16

[223] Vgl. Feretzakis et al. (2024), Privacy-Preserving Techniques in Generative AI and Large Language Models: A Narrative Review, S. 1-16

6 LLM-Training mit synthetischen Daten

Eine weitere Maßnahme, die das Potential hat, das Risiko einer Offenlegung personenbezogener Informationen auszuschließen, ist die Nutzung synthetischer Daten für das Training von LLMs. Nachfolgend soll das Potential synthetischer Daten und deren Praxistauglichkeit untersucht werden. Synthetische Datensätze sind künstliche Datensätze, die die statistischen Eigenschaften von Originaldatensätzen aufweisen und über sog. Daten-Synthetisierung generiert werden. Synthetische Daten stellen eine Repräsentation von Originaldaten dar, welche weiterhin die Zusammenhänge der Originaldaten aufweisen.[224] Zur Veranschaulichung des Prinzips synthetischer Daten eignet sich folgende stark vereinfachte, bildliche Vorstellung: ein Excel-Sheet enthält in jeder Spalte ein Merkmal einer natürlichen Person (Spalte 1: Vorname; Spalte 2: Nachname; Spalte 3: Geburtsdatum; Spalte 4: Adresse; Spalte 5: Einkommen etc.). Die Spalten 1, 2 und 3 werden unabhängig voneinander willkürlich vertikal verschoben. Dadurch können die Merkmale in den nachfolgenden Spalten nicht mehr der ursprünglichen Person zugeordnet werden. Die statistischen Eigenschaften des Datensatzes bleiben erhalten.

Bei der Daten-Synthetisierung lassen sich zwei Ansätze unterscheiden: (1.) die sog. vollständige Synthetisierung (alle Daten des Originaldatensatzes werden durch künstliche Werte ersetzt) und (2.) die sog. teilweise Synthetisierung (nur eine Teilmenge der Daten des Originaldatensatzes wird durch synthetische Daten ersetzt).[225] Die oben dargestellte Analogie beschreibt eine teilweise Synthetisierung.

Abgesehen davon, dass das Training eines LLMs mit synthetischen Daten das Risiko der Offenlegung personenbezogener Daten im Rahmen der Inferenz-Phase reduziert oder gar gänzlich ausschließt, kann die Verwendung synthetischer Daten für das Training von LLMs noch weitere Vorteile mit sich bringen. So können synthetische Daten in beliebiger Menge generiert und gezielt von Verzerrungen bereinigt werden, die in einem Originaldatensatz vorkommen können. Unter der Annahme, dass ein synthetischer

[224] Vgl. Drechsler/Jentzsch (2018), Synthetische Daten: Innovationspotential und gesellschaftliche Herausforderungen, S. 2

[225] Vgl. Drechsler/Jentzsch (2018), Synthetische Daten: Innovationspotential und gesellschaftliche Herausforderungen, S. 9

Datensatz anonym ist, können die hohen datenschutzrechtlichen Anforderungen der DSGVO umgangen werden, ohne dass der wesentliche Informationsgehalt des Originaldatensatzes verloren geht.[226]

6.1 Anonymität synthetischer Daten

Die Daten-Synthetisierung bietet nach allgemeiner Auffassung das Potential einer Anonymisierungsmethode.[227] Bereits eine Handvoll von Unternehmen haben sich dies zum Geschäftsmodell gemacht, indem Sie „Daten-Synthetisierung-as-a-Service" für Unternehmen aus verschiedenen Branchen anbieten (u. a. Finanzdienstleister, Telekommunikationsdienstleister und Forschungsinstitutionen). Dabei werben sie mit einer angeblichen DSGVO-Compliance aufgrund einer Anonymisierung im Rahmen der Daten-Synthetisierung.[228] Tatsächlich ist die datenschutzrechtliche Einordnung der Nutzung synthetischer Daten jedoch noch nicht abschließend geklärt und geht mit einer gewissen Rechtsunsicherheit einher.[229]

Die Bewertung, ob ein synthetisches Datum ein anonymes Datum i. S. d. DSGVO darstellt, hängt unter Anwendung der sog. relativen Theorie (vgl. hierzu Kap. 4.1.2) – die aktuell in der Literatur sowie teilw. auch in der Rechtsprechung überwiegend Zustimmung findet[230] – insbesondere von der Zusammensetzung des Originaldatensatzes,

---

[226] Vgl. Datenethikkommission (2019), Gutachten der Datenethikkommission: Ethik und Digitalisierung – Daten, Algorithmen, Künstliche Intelligenz, S. 59; Raji (2021), Rechtliche Bewertung synthetischer Daten für KI-Systeme, S. 305

[227] Vgl. Lettieri/Kipker (2024), Synthetische Daten in der medizinischen Forschung – Eine datenschutzrechtliche Bewertung, S. 285

[228] Vgl. https://mostly.ai/?_gl=1*neyywl*_up*MQ..*_ga*MTM4NzMzMjIyMy4xNzM2Mjc2NTA2*_ga_8NGESMV97J*MTczNjI3NjUwNi4xLjEuMTczNjI3NjUxMC4wLjAuMA (letzter Aufruf: 09.01.2025); https://hazy.com/ (letzter Aufruf: 09.01.2025)

[229] Vgl. Vgl. Lettieri/Kipker (2024), Synthetische Daten in der medizinischen Forschung – Eine datenschutzrechtliche Bewertung, S. 285

[230] Vgl. u.a. Vgl. EuGH (2. Kammer), Urt. v. 19.10.2016 – C-582/14 (Breyer/Deutschland); EuGH, Urteil v. 9.11.2023 – C-319/22; Ziebarth in: Sydow (2018), Europäische Datenschutzgrundverordnung, Art. 4 Rn. 29 f.

dem verfolgten Synthetisierungs-Ansatz (vollständige oder teilweise Synthetisierung) sowie dem Aufwand der Re-Identifikation im konkreten Einzelfall ab. Grundsätzlich kann bei teilweise synthetischen Daten von einem höheren Re-Identifikationsrisiko ausgegangen werden als bei vollständig synthetischen Daten.[231] Die abschließende Bewertung der Anonymität synthetischer Daten ist in der Praxis jedoch aufwendig und wenig praktikabel. Die Frage nach dem Aufwand bzw. dem Risiko der Re-Identifizierung hängt vor allem von den konkreten technischen Maßnahmen und Möglichkeiten sowie den potenziellen Angriffsszenarien im Einzelfall ab.[232] Um den Aufwand einer Re-Identifizierung bzw. die Höhe des Risikos einer Re-Identifizierung aus technischer Sicht belastbar zu bewerten, muss i. d. R. eine sehr umfangreiche Anzahl an potenziellen Szenarien berücksichtigt werden, die je nach Einzelfall und Akteuren variieren können.[233] Erschwerend hinzu kommen die zunehmend schnelllebigen technischen Entwicklungen, die dazu führen, dass eine Risikobewertung theoretisch regelmäßig angepasst werden sollte.[234] Eine belastbare Risikobewertung erfordert folglich erhebliche zeitliche und finanzielle Ressourcen. Insofern ist es faktisch kaum möglich, rechtssichere Standards für eine entsprechende Bewertung der Anonymität synthetischer Daten zu entwickeln. Es kann jedoch angenommen werden, dass das Risiko einer Re-Identifizierung eher gering ist, solange die jeweiligen synthetischen Daten und das jeweils angewendete Synthetisierungsverfahren nicht gegenüber potenziellen Angreifern offengelegt werden.[235] Folglich sind im Kontext mit dem Training von LLMs durchaus Szenarien denkbar, in welchen die Annahme einer Anonymität i. S. d. DSGVO von synthetischen Daten juristisch vertretbar wäre.

[231] Vgl. Drechsler/Jentzsch (2018), Synthetische Daten: Innovationspotential und gesellschaftliche Herausforderungen, S. 10

[232] Vgl. Lettieri/Kipker (2024), Synthetische Daten in der medizinischen Forschung – Eine datenschutzrechtliche Bewertung, S. 286

[233] Vgl. Lettieri/Kipker (2024), Synthetische Daten in der medizinischen Forschung – Eine datenschutzrechtliche Bewertung, S. 286

[234] Vgl. Lettieri/Kipker (2024), Synthetische Daten in der medizinischen Forschung – Eine datenschutzrechtliche Bewertung, S. 287

[235] Vgl. Lettieri/Kipker (2024), Synthetische Daten in der medizinischen Forschung – Eine datenschutzrechtliche Bewertung, S. 287

6.2 Datenschutzrechtliche Rechtsgrundlage für die Daten-Synthetisierung

Auch unter der Annahme, dass ein bestimmter synthetischer Datensatz anonym ist, finden datenschutzrechtliche Anforderungen auf den Prozess der Daten-Synthetisierung Anwendung, da hierfür i. d. R. eine Verarbeitung der jeweiligen Originaldaten erfolgt. Insoweit bedarf es für die Daten-Synthetisierung eine datenschutzrechtliche Rechtsgrundlage.[236] Sofern der Originaldatensatz „normale" personenbezogene Daten enthält, kommt insbesondere Art. 6 Abs. 1 lit. f DSGVO als Rechtsgrundlage in Betracht, wobei im Kontext mit dem Training von LLMs ggf. die unter Kapitel 3.1.2 dargestellte Problematik im Zusammenhang mit den vernünftigen Erwartungen der betroffenen Personen zu berücksichtigen ist. Ferner kann eine Daten-Synthetisierung zulässig sein, wenn die Anforderungen des Art. 6 Abs. 4 DSGVO erfüllt sind.[237] In diesem Fall ist keine andere gesonderte Rechtsgrundlage erforderlich als diejenige für die ursprüngliche Erhebung der personenbezogenen Daten.[238]

Sofern jedoch von einer Verarbeitung besonderer Kategorien personenbezogener Daten i. S. d. Art. 9 Abs. 1 DSGVO auszugehen ist (vgl. hierzu Kap. 4.3), sind zudem die Anforderungen des Art. 9 Abs. 2 DSGVO zu beachten. Für die Verarbeitung besonderer Kategorien personenbezogener Daten im Rahmen einer Daten-Synthetisierung für das Training von LLMs werden die Erlaubnistatbestände des Art. 9 Abs. 2 DSGVO jedoch nur in Ausnahmefällen wirksam Anwendung finden (vgl. hierzu Kapitel 3.1.4 und 3.1.5).[239] Dies führt zu der absurden Problematik, dass eine vermeintlich datenschutzfördernde Daten-Synthetisierung, die ggf. sogar eine anonymisierende Wirkung hat, bei „normalen" personenbezogenen Daten zulässig sein kann, bei besonderen Kategorien personenbezogener Daten jedoch nicht. Wenn man berücksichtigt,

[236] Vgl. BfDI (2020), Positionspapier zur Anonymisierung unter der DSGVO unter besonderer Berücksichtigung der TK-Branche, S. 5; Hornung/Wagner (2020), Anonymisierung als datenschutzrelevante Verarbeitung? Rechtliche Anforderungen und Grenzen für die Anonymisierung personenbezogener Daten, S. 223

[237] Vgl. BfDI (2020), Positionspapier zur Anonymisierung unter der DSGVO unter besonderer Berücksichtigung der TK-Branche, S. 7

[238] Vgl ErwGr. 50 S. 3 DSGVO

[239] Vgl. Hornung/Wagner (2020), Anonymisierung als datenschutzrelevante Verarbeitung? Rechtliche Anforderungen und Grenzen für die Anonymisierung personenbezogener Daten, ZD 2020, S. 226

dass eine Daten-Synthetisierung die Risiken für die Rechte und Freiheiten der betroffenen Personen vermeintlich mindert, erscheint dies im Hinblick auf einen angemessenen Ausgleich zwischen den sich gegenüberstehenden Grundrechtspositionen der verarbeitenden Stelle und der betroffenen Personen nicht sachgerecht. Sinn und Zweck der verschärften Anforderungen des Art. 9 Abs. 1 DSGVO ist es, einen angemessenen zusätzlichen Schutz für besonders risikobehaftete Datenverarbeitungen zu schaffen. Nach Hornung und Wagner würde ein Beharren auf das Verarbeitungsverbot nach Art. 9 Abs. 1 DSGVO im vorliegenden Kontext vielmehr zu einem erhöhten Risiko für die Rechte und Freiheiten der betroffenen Personen führen, da die Daten-Synthetisierung als Schutzmaßnahme verwehrt bleiben würde.[240] Man kann davon ausgehen, dass der Gesetzgeber diese Problematik schlichtweg übersehen hat, weshalb eine entsprechende teleologische Reduktion des Verarbeitungsverbots in Art. 9 Abs. 1 DSGVO mit dem Schutzziel der DSGVO ggf. zu vereinbaren wäre und als Lösungsweg unter Inkaufnahme einer gewissen Rechtsunsicherheit vertretbar sein könnte.[241] Die teleologische Reduktion des Art. 9 Abs. 1 DSGVO ist in der Literatur jedoch berechtigterweise umstritten, da in der Praxis nicht eindeutig ist, wann die Schwelle zur Anonymität erreicht ist und die tatsächliche Schutzwirkung vom konkreten Einzelfall abhängt.[242]

Ein sachgerechter Ausgleich könnte hier wiederum die Anwendung eines risikobasierten und kontextorientierten Ansatzes schaffen (vgl. hierzu Kap. 4.3). Dabei bedarf es einer differenzierten Bewertung unter Berücksichtigung des jeweiligen Verarbeitungskontexts und der tatsächlichen Risiken für die Rechte und Freiheiten der betroffenen Personen, um sowohl den Schutz der Rechte der betroffenen Personen als auch die unternehmerische Freiheit angemessen zu wahren. Besteht kein gesteigertes Risiko für die Rechte und Freiheiten der betroffenen Personen, bspw. weil sich die verarbeitende Stelle den sensiblen Informationsgehalt nicht gezielt zu Nutze macht und/oder keine hinreichende Wahrscheinlichkeit dafür besteht, dass der sensible Informationsgehalt offengelegt wird, wäre die Anwendung des Verarbeitungsverbots des Art. 9 Abs. 1 DSGVO nach der

[240] Vgl. Hornung/Wagner (2020), Anonymisierung als datenschutzrelevante Verarbeitung? Rechtliche Anforderungen und Grenzen für die Anonymisierung personenbezogener Daten, ZD 2020, S. 228

[241] Vgl. Hornung/Wagner (2020), Anonymisierung als datenschutzrelevante Verarbeitung? Rechtliche Anforderungen und Grenzen für die Anonymisierung personenbezogener Daten, ZD 2020, S. 228

[242] Vgl. Raji (2021), Rechtliche Bewertung synthetischer Daten für KI-Systeme, S. 308

kontextorientierten Auslegung der EuGH-Rechtsprechung zu verneinen.[243] Dieses Ergebnis hätte zur Folge, dass auch die Verarbeitung von potenziell besonderen Kategorien personenbezogener Daten im Rahmen einer Daten-Synthetisierung unter Berücksichtigung der Anforderungen des Art. 6 Abs. 1 lit. f DSGVO oder der Anforderungen des Art. 6 Abs. 4 DSGVO zulässig wäre.

6.3 Auswirkungen synthetischer Daten auf die Rechte und Freiheiten natürlicher Personen

Im Rahmen der Diskussion um die Rechtsgrundlage für die Daten-Synthetisierung wird auch hinterfragt, inwiefern die Daten-Synthetisierung und die damit ggf. einhergehende Anonymisierung personenbezogener Daten tatsächlich eine Schutzmaßnahme zu Gunsten der Grundrechte und -freiheiten natürlicher Personen darstellt. Raji merkt dabei an, dass auch in Fällen, in denen eine Anonymität der synthetischen Daten angenommen werden kann, die Nutzung solcher Daten Auswirkungen auf die Rechte und Freiheiten natürlicher Personen haben kann, da der relevante Informationsgehalt erhalten bleibt.[244] Erkenntnisse, die im Rahmen der Verarbeitung anonymer, synthetischer Daten gewonnen werden – bspw. über bisher unbekannte Muster in menschlichem Verhalten – können genutzt werden, um Entscheidungen mit Auswirkungen auf die Rechte und Freiheiten natürlicher Personen (insbesondere Art. 21 GRCh) zu treffen.[245] Kritisch hinterfragt werden sollte in diesem Zusammenhang, ob die Nichtanwendbarkeit des Datenschutzrechts auf die Nutzung anonymer, synthetischer Daten (insb. im Hinblick auf die Betroffenenrechte) und die damit ggf. einhergehende Unwissenheit der „Datensubjekte“ auch negative Effekte auf den Grundrechtsschutz haben kann.[246] Raji weist auf das Risiko hin, dass eine vermehrte Nutzung synthetischer Daten außerhalb des

[243] Vgl. Kohn/Schleper (2023), Die (zufällige) Erhebung sensibler Daten, S. 723; Britz/Indenhuck/Langerhans (2021), Die Verarbeitung „zufällig“ sensibler Daten, S. 559; EDSA (2020), Leitlinien 3/2019 zur Verarbeitung personenbezogener Daten durch Videogeräte, Version 2.0, Rn. 62-72; Matejek/Kremer (2023), Sensible Daten soweit das Auge reicht: Der EuGH zapft den „hidden layer“ an, S. 218; EuGH, Urteil v. 04.07.2023 – C-252/21, Rn. 73; Vgl. EuGH, Urteil v. 01.08.2022 – C-184/20, Rn. 126; Vgl. EuGH, Urteil v. 24.09.2019 – C-136/17, Rn. 44

[244] Vgl. Raji (2021), Rechtliche Bewertung synthetischer Daten für KI-Systeme, S. 308

[245] Vgl. Raji (2021), Rechtliche Bewertung synthetischer Daten für KI-Systeme, S. 308

[246] Vgl. Raji (2021), Rechtliche Bewertung synthetischer Daten für KI-Systeme, S. 308

Anwendungsbereichs der DSGVO zu einer Schwächung des Grundrechtsschutzes beitragen kann.[247] Auch der Europäischen Datenschutzbeauftragte erkennt in seinem Tech Sonar Report 2021-2022 das Risiko an, dass auch die Auswertung synthetischer Daten negative Auswirkungen für die Privatsphäre und andere Menschenrechte realer Personen haben kann.[248] Potenzielle Risiken für Grundrechte und -freiheiten natürlicher Personen bei der Verarbeitung von anonymen, synthetischen Daten, sind ggf. ein weiteres Argument dafür, dass die starre Abhängigkeit der Anwendbarkeit der DSGVO von dem Merkmal des Personenbezugs nicht mehr zeitgemäß ist (vgl. Kapitel 4.2). Welche Auswirkungen die Nutzung synthetischer Daten auf die Grundrechte und -freiheiten natürlicher Personen zukünftig haben wird, bleibt abzuwarten. Die Entwicklungen Rund um die Nutzung synthetischer Daten sollten nicht nur im Kontext mit dem Training von LLMs, sondern insgesamt (insb. auch im Hinblick auf die Generierung synthetischer Stimmen und sog. Deepfakes) kritisch beobachtet werden.[249] Die Abwesenheit des Datenschutzrechts darf nicht zu einer Abkehr von grundrechtlichen Maßstäben führen. Insofern bedarf es ggf. einer Neuausrichtung des „Rechts zum Schutz der Privatsphäre“ oder zumindest der Definition von klaren Richtlinien für eine ethische Erzeugung und Nutzung synthetischer Daten.

6.4 Zwischenfazit

Wie oben dargestellt, bietet die Daten-Synthetisierung in der Theorie eine vielversprechende Möglichkeit, anonyme Trainingsdaten für das Training von LLMs in großem Umfang zu generieren und datenschutzrechtliche Hürden beim Training von LLMs zu umgehen. Ob sich diese Technologie auch in der Praxis für das Training von LLMs eignet, hängt maßgeblich davon ab, ob die synthetischen Daten tatsächlich die erforderliche Aussagekraft und Qualität gewährleisten können und sich die Aussage des Head of Technology and Privacy des Europäischen Datenschutzbeauftragten, die Daten-Synthetisierung würde im KI-Kontext immer mehr an Bedeutung gewinnen und biete

[247] Vgl. Raji (2021), Rechtliche Bewertung synthetischer Daten für KI-Systeme, S. 308-309

[248] Vgl. European Data Protection Supervisor (2021), TechSonar Report 2021-2022, S. 11

[249] Vgl. Wagner (2021), Privacy Enhancing Technologies and Synthetic Data, S. 5-6

einen robusten Schutz der Privatsphäre sowie die Möglichkeit, nützliche Daten zu generieren, bestätigt.[250]

Zumindest empirische Studien zur utility von synthetischen Daten kommen teilweise zu dem Ergebnis, dass mit der Nutzung synthetischer Daten für das Training von Machine Learning-Modellen keine signifikanten Leistungseinbußen im Vergleich zur Nutzung von Originaldaten einhergehen.[251] Eine Studie beschäftigte sich mit der Performanz von komplexen, synthetischen Mikrobiom-Daten in unterschiedlichen Aufgaben des maschinellen Lernens. Dabei haben die synthetischen Datensätze gut bei den jeweiligen maschinellen Lernaufgaben abgeschnitten, während die Performanz im Vergleich zu denselben Aufgaben unter Verwendung der Originaldaten nur geringfügig abgewichen ist.[252]

Anzumerken ist jedoch, dass es aktuell keine allgemeingültige Metrik für die Evaluierung der utility von synthetischen Daten gibt, und die Ergebnisse solcher Studien von einer Vielzahl von Parametern abhängen können.[253] Welche konkreten Aspekte und wie synthetische Daten bei deren Nutzung für das Training von LLMs die Leistung von LLMs beeinflussen ist bis dato noch weitgehend unklar.[254] Der aktuelle Hype und die umfangreiche Forschung zur Daten-Synthetisierung im Zusammenhang mit der Entwicklung von KI-Systemen ist jedoch vielversprechend. Eine Schlagwortsuche nach „data synthesis“ über Google Scholar ergibt, dass allein im Januar 2025 bereits 4.370 Publikationen zu dem Thema veröffentlicht wurden.

---

[250] Vgl. Zerdick (2021), Is the future of privacy synthetic?, aufrufbar unter: https://edps.europa.eu/press-publications/press-news/blog/future-privacy-synthetic_en (letzter Zugriff: 14.01.2025)

[251] Vgl. u.a. Hittmeir, Mayer, Ekelhart, (2019), On the Utility of Synthetic Data: An Empirical Evaluation on Machine Learning Tasks, S. 6; Hittmeir, Mayer, Ekelhart (2022), Utility and Privacy Assessment of Synthetic Microbiome Data

[252] Vgl. Hittmeir, Mayer, Ekelhart (2022), Utility and Privacy Assessment of Synthetic Microbiome Data

[253] Vgl. Wang et al. (2024), A Survey on Data Synthesis and Augmentation for Large Language Models, S. 2

[254] Vgl. Chen et al. (2024), On the diversity of synthetic data and its impact on training large language models, S. 1

Trotz der aufgezeigten Herausforderungen weist die Technologie der Daten-Synthetisierung erhebliche potenzielle Chancen für die Entwicklung von LLMs auf. Die verantwortungsvolle und effektive Nutzung anonymer, synthetischer Daten für das Training von LLMs könnte die Lösung für datenschutzrechtliche Herausforderungen und Unsicherheiten sein und gleichzeitig den quantitativ steigenden Bedarf an qualitativ hochwertigen Trainingsdaten befriedigen. Dieses Potential erkennt auch der Europäische Datenschutzbeauftragte an.[255]

Allerdings sollte angemerkt werden, dass die Daten-Synthetisierung die datenschutzrechtlichen Problemstellungen im Kontext mit dem Training von LLMs nicht gänzlich und für jedermann aus der Welt schaffen kann, da auch diejenigen Modelle, die für die Daten-Synthetisierung genutzt werden, zunächst mit Originaldaten trainiert werden müssen.[256]

7 Interessenabwägung nach Art. 6 Abs. 1 lit. f DSGVO

Im nachfolgenden Kapitel soll die Abwägung der berechtigten Interessen der verantwortlichen Stelle und/oder Dritter mit den Grundrechten und Grundfreiheiten der betroffenen Personen als zwingende Anforderung des Art. 6 Abs. 1 lit. f DSGVO für eine zulässige Verarbeitung personenbezogener Daten für Zwecke der Entwicklung von LLMs genauer betrachtet werden. Der Fokus soll dabei auf den Aspekten des Sachverhalts des Trainings von LLMs liegen, die für die Abwägung relevant sind und das Gewicht der sich gegenüberstehenden Grundrechtspositionen beeinflussen. Das Kapitel zielt darauf ab, einen Überblick über die relevanten Aspekte und deren Auswirkung auf die Interessenabwägung nach Art. 6 Abs. 1 lit. f DSGVO zu geben.

Sobald im Rahmen der Drei-Stufen-Prüfung (vgl. Kapitel 3.1.2) berechtigte Interesse der verantwortlichen Stelle oder eines Dritten (Stufe 1) sowie die Erforderlichkeit der Datenverarbeitung für die Wahrung dieser berechtigten Interessen (Stufe 2) bejaht werden können, ist auf der dritten Stufe die Ausgewogenheit zwischen den berechtigten

[255] Vgl. European Data Protection Supervisor (2021), TechSonar Report 2021-2022, S. 11; Zerdick (2021), Is the future of privacy synthetic?, aufrufbar unter: https://edps.europa.eu/press-publications/press-news/blog/future-privacy-synthetic_en, (letzter Zugriff: 14.01.2025)

[256] Vgl. El Mestari et al. (2024), Preserving Data Privacy in Machine Learning Systems, S. 9

Interessen des Verantwortlichen und den Interessen oder Grundrechten und -freiheiten der Datensubjekte zu prüfen. Da alle Umstände des konkreten Einzelfalls zu berücksichtigen sind, handelt es sich bei dieser Prüfung i. d. R. nicht um eine Abwägung zweier leicht qualifizierbarer und einfach zu vergleichender Positionen.[257] Die sich gegenüberstehenden Interessen sind im Rahmen der Abwägung zu gewichten, wobei die Ausgestaltung der Verarbeitung im Einzelfall sowie konkrete, potenzielle Auswirkungen der Verarbeitung auf die betroffenen Personen eine Rolle spielen.[258] Grundsätzlich sollten die folgenden Faktoren bei einer ordnungsgemäßen Interessenabwägung berücksichtigt werden:[259]

- der Charakter und Ursprung des berechtigten Interesses;
- die Erforderlichkeit der Datenverarbeitung zur Wahrung eines Grundrechts oder eines öffentlichen Interesses;
- der Charakter und Ursprung der betroffenen Interessen der Datensubjekte;
- die begründeten Erwartungen der betroffenen Personen betreffend die Art und Weise der Datenverarbeitung;
- die potenziellen Folgen für die betroffenen Personen;
- die Art, Menge und Quelle der personenbezogenen Daten;
- der Grad der Schutzbedürftigkeit der betroffenen Personen;
- die Menge der datenverarbeitenden Akteure und der betroffenen Personen sowie das (Macht-)Verhältnis zwischen diesen;
- die Art und der Kontext der Datenerhebung und -verarbeitung;

[257] Vgl. Artikel-29-Datenschutzgruppe (2014), WP 217, S. 3

[258] Vgl. EDPB (2024), Guidelines 1/2024 on the processing of personal data based on Article 6(1)(f) GDPR, Rn. 32

[259] Vgl. Artikel-29-Datenschutzgruppe (2014), WP 217, S. 3, 49, 50; EDPB (2024), Guidelines 1/2024 on the processing of personal data based on Article 6(1)(f) GDPR, Rn. 32

- zusätzliche Sicherheitsmaßnahmen, die unangemessene Folgen der Datenverarbeitung für die betroffenen Personen begrenzen könnten.

7.1 Berechtigte Interessen der verantwortlichen Stelle

Auf Seiten der verantwortlichen Stelle muss auf den unterschiedlichen Charakter berechtigter Interessen eingegangen werden. Ein berechtigtes Interesse könnte, wie die Artikel-29-Datenschutzgruppe erläutert, von „unbedeutend" über eine „gewisse Bedeutung" besitzend bis hin zu „zwingend" reichen.[260] Maßgebend für die Beurteilung des Charakters eines berechtigten Interesses ist, ob dieses durch ein konkretes Grundrecht geschützt ist.[261] Grundsätzlich ist einem berechtigten Interesse ein höheres Gewicht zuzusprechen, wenn es verfassungsrechtlich anerkannt ist.[262] Berechtigte Interessen im Kontext mit dem Training von LLMs können bspw. durch die Freiheit der Meinungsäußerung und Informationsfreiheit (Art. 11 GRCh), die Wissenschafts- und Kunstfreiheit (Art. 13 GRCh) oder die unternehmerische Freiheit (Art. 16 GRCh) geschützt sein. Ferner kann einem berechtigten Interesse ein höheres Gewicht zugesprochen werden, wenn es in anderen Rechtsinstrumenten, insbesondere unionsrechtlichen Rechtsvorschriften oder nationalen Rechtsvorschriften des jeweiligen Mitgliedstaats, anerkannt wird.[263] Für die Gewichtung berechtigter Interessen zu berücksichtigen ist auch, ob diese nur für die verantwortliche Stelle respektive die jeweiligen Dritten oder auch für Außenstehende wie bspw. der breiten Öffentlichkeit relevant sind.[264]

Für die Gewichtung von berechtigten Interessen im Rahmen der Entwicklung von LLMs könnte demnach die europäischen KI-Strategie berücksichtigt werden. Die Europäische

[260] Vgl. Artikel-29-Datenschutzgruppe (2014), WP 217, S. 39

[261] Vgl. Robrahn/Bremert (2018), Interessenskonflikte im Datenschutzrecht: Rechtfertigung der Verarbeitung personenbezogener Daten über eine Abwägung nach Art. 6 Abs. 1 lit. f DS-GVO, S. 291, 294

[262] Vgl. Herfurth (2018), Interessenabwägung nach Art. 6 Abs. 1 lit. f DS-GVO: Nachvollziehbare Ergebnisse anhand von 15 Kriterien mit dem sog. „3x5-Modell", S. 514, 515

[263] Vgl. Herfurth (2018), Interessenabwägung nach Art. 6 Abs. 1 lit. f DS-GVO: Nachvollziehbare Ergebnisse anhand von 15 Kriterien mit dem sog. „3x5-Modell", S. 514, 515

[264] Vgl. Herfurth (2018), Interessenabwägung nach Art. 6 Abs. 1 lit. f DS-GVO: Nachvollziehbare Ergebnisse anhand von 15 Kriterien mit dem sog. „3x5-Modell", S. 514, 515

Kommission verfolgt mit ihrer KI-Strategie insbesondere die Wettbewerbsfähigkeit von Europa im Bereich der KI zu sichern und Forschung, Innovation und Entwicklung von KI-Technologien zu unterstützen.[265] Diese Ziele sind auch in der KI-VO verankert.[266] Folglich entsprechen berechtigte Interessen an der Entwicklung „vertrauenswürdiger" und grundrechtskonformer KI-Modelle, grundsätzlich der Wirtschaftsstrategie der EU Kommission und werden durch deren Gesetzgebung anerkannt. Dies könnte berechtigten Interessen für das Training von LLMs zusätzliches Gewicht verleihen. Je offensichtlicher in der allgemeinen Öffentlichkeit anerkannt und erwartet wird, dass die verantwortliche Stelle zur Verfolgung der jeweiligen berechtigten Interessen personenbezogene Daten verarbeiten kann, umso mehr Gewicht kann diesen berechtigten Interessen bei der Abwägung zugeschrieben werden.[267] Allerdings darf sich die verantwortliche Stelle nicht als „selbsternannter Sachverwalter" von eigentlich staatlich zu schützenden Allgemeininteressen verstehen.[268] Es besteht die Gefahr, dass gewichtige Allgemeininteressen die Abwägung immer zu Gunsten der verantwortlichen Stelle beeinflussen. Insofern ist eine positive Relevanz für die Allgemeinheit zwar zu berücksichtigen, allerdings sollte sich die verantwortliche Stelle nicht direkt und alleinig auf die Wahrnehmung von Allgemeininteressen berufen.[269]

Für den Fall, dass berechtigte Interessen der verantwortlichen Stelle als nicht zwingend zu beurteilen sind, können diese in der Regel nur dann Vorrang vor den Interessen der betroffenen Person haben, wenn die potenziellen Folgen für die betroffenen Personen

[265] Vgl. European Kommission (2018), Mitteilung der Kommission an das Europäische Parlament, den Europäischen Rat, den Rat den Europäischen Wirtschafts- und Sozialausschuss und den Ausschuss der Regionen – Künstliche Intelligenz für Europa, COM(2018) 237 final; Europäische Kommission (2021), Koordinierter Plan für künstliche Intelligenz, Überarbeitung 2021, COM(2021) 205 final

[266] Vgl. ErwGr. 1 KI-VO

[267] Vgl. Artikel-29-Datenschutzgruppe (2014), WP 217, S. 46

[268] Vgl. Robrahn/Bremert (2018), Interessenskonflikte im Datenschutzrecht: Rechtfertigung der Verarbeitung personenbezogener Daten über eine Abwägung nach Art. 6 Abs. 1 lit. f DS-GVO, S. 291, 292

[269] Vgl. Robrahn/Bremert (2018), Interessenskonflikte im Datenschutzrecht: Rechtfertigung der Verarbeitung personenbezogener Daten über eine Abwägung nach Art. 6 Abs. 1 lit. f DS-GVO, S. 291, 292

noch geringer sind.[270] Dabei können zusätzliche Schutzmaßnahmen, die eine potenzielle Beeinträchtigung durch die jeweilige Datenverarbeitung reduzieren, im Rahmen der Interessenabwägung zu Gunsten der verantwortlichen Stelle wirken.[271] Solche zusätzlichen Maßnahmen sind jedoch von den Schutzmaßnahmen zu differenzieren, die die verantwortliche Stelle ohnehin anwendet, um seine Pflichten nach den Art. 25 und 32 DSGVO zu erfüllen.[272]

7.2 Interessen, Grundrechte und -freiheiten der betroffenen Personen

In der anderen Waagschale liegen die Interessen, Grundrechte und Grundfreiheiten der betroffenen Personen. Zu berücksichtigen sind nicht nur Grundrechte und -freiheiten, wie z. B. das Grundrecht auf Achtung des Privat- und Familienlebens (Art. 7 GRCh), das Grundrecht auf Schutz personenbezogener Daten (Art. 8 GRCh) oder das Grundrecht auf Nichtdiskriminierung (Art. 21 GRCh), sondern auch alle sonstigen einschlägigen Interessen der betroffenen Personen.[273] Einschlägige Interessen sind solche, die im Zusammenhang mit dem Schutz der betroffenen personenbezogenen Daten stehen.[274] Im Zusammenhang mit Datenverarbeitungen im Rahmen der Entwicklung von LLMs können z. B. das Interesse, die Kontrolle über die eigenen personenbezogenen Daten zu behalten (u. a. gestützt durch das Grundrecht auf informationelle Selbstbestimmung)[275], oder Interessen an der Vermeidung von Reputationsschäden, Identitätsdiebstahl oder Betrug, relevant sein. Dabei ist zu beachten, dass die durch die DSGVO geschützten Interessen der betroffenen Personen nicht den subjektiven Interessen der einzelnen Person entsprechen, sondern vielmehr den objektivierten Interessen betroffener Personen im

[270] Vgl. Artikel-29-Datenschutzgruppe (2014), WP 217, S. 39

[271] Vgl. EDPB (2024), Guidelines 1/2024 on the processing of personal data based on Article 6(1)(f) GDPR, Rn. 56

[272] Vgl. EDPB (2024), Guidelines 1/2024 on the processing of personal data based on Article 6(1)(f) GDPR, Rn. 57

[273] Vgl. Artikel-29-Datenschutzgruppe (2014), WP 217, S. 38

[274] Vgl. Schwartmann/Klein in Schwartmann/Jaspers/Thüsing/Kugelmann (2020), DS-GVO/BDSG: Datenschutz-Grundverordnung Bundesdatenschutzgesetz (Heidelberger Kommentar), Art. 6, Rn. 153

[275] Vgl. BVerfG, Urteil des Ersten Senats vom 15. Dezember 1983 – 1 BvR 209/83, Rn. 1-215

Kontext der konkreten Datenverarbeitung.[276] Ebenso wie die berechtigten Interessen der verantwortlichen Stelle oder eines Dritten, sind auch die Interessen der betroffenen Personen im Zuge der Abwägung zu gewichten. Es gilt, je mehr Grundrechte durch die Datenverarbeitung betroffen werden, desto schwerer ist das Interesse der betroffenen Personen zu gewichten.[277]

7.2.1 Potenzielle Folgen der Datenverarbeitung für die betroffenen Personen

Für die Gewichtung sind auch die potenziellen negativen Folgen der jeweiligen Datenverarbeitung für die Rechte und Freiheiten der betroffenen Personen abzuschätzen und ihrem Schweregrad entsprechend zu berücksichtigen.[278] Anhaltspunkte für die Abschätzung der Folgen können die Art sowie die Quelle der zu verarbeitenden Daten aber auch die Art und Weise der Verarbeitung sein.[279]

Betreffend die Art der Daten ist zwischen sensiblen und sonstigen personenbezogenen Daten zu differenzieren.[280] Es gilt, je sensibler die Informationen, umso höher ist das Risiko potenzieller negativer Folgen für die betroffenen Personen. In diesem Kontext sind mit sensiblen Daten nicht nur die besonderen Kategorien personenbezogener Daten i. S. d. Art. 9 Abs. 1 DSGVO gemeint, sondern auch solche Daten, die besonders risikobehaftet sind, wie bspw. Informationen über die wirtschaftlichen und finanziellen Verhältnisse der betroffenen Person oder Kontodaten.[281] Auch bestimmte Daten deren Aussagegehalt besonders hoch ist, wie bspw. Standortdaten, können eine höhere

[276] Vgl. Schwartmann/Klein in Schwartmann/Jaspers/Thüsing/Kugelmann (2020), DS-GVO/BDSG: Datenschutz-Grundverordnung Bundesdatenschutzgesetz (Heidelberger Kommentar), Art. 6, Rn. 153

[277] Vgl. Herfurth (2018), Interessenabwägung nach Art. 6 Abs. 1 lit. f DS-GVO: Nachvollziehbare Ergebnisse anhand von 15 Kriterien mit dem sog. „3x5-Modell“, S. 514, 515

[278] Vgl. Artikel-29-Datenschutzgruppe (2014), WP 217, S. 39

[279] Vgl. Artikel-29-Datenschutzgruppe (2014), WP 217, S. 47

[280] Vgl. Herfurth (2018), Interessenabwägung nach Art. 6 Abs. 1 lit. f DS-GVO: Nachvollziehbare Ergebnisse anhand von 15 Kriterien mit dem sog. „3x5-Modell“, S. 514, 515

[281] Vgl. Gierschmann (2018), Gestaltungsmöglichkeiten bei Verwendung von personenbezogenen Daten in der Werbung: Auslegung des Art. 6 Abs. 1 lit. f DS-GVO und Lösungsvorschläge, S. 7, 11

Gewichtung begründen.[282] In diesem Zusammenhang ist auch der Umfang der erhobenen Daten über eine bestimmte Person zu berücksichtigen, da eine große Menge von Daten die Aussagekraft über die betroffene Person erhöhen kann.[283]

In Bezug auf die Quelle der Daten kann es für die Abschätzung relevant sein, ob die zu verarbeitenden Daten bereits von der betroffenen Person oder einem Dritten öffentlich zugänglich gemacht wurden, bspw. auf einer Social Media-Plattform.[284] Sofern die personenbezogenen Daten von den betroffenen Personen offensichtlich, freiwillig selbst veröffentlicht wurden, kann angenommen werden, dass die Verarbeitung dieser personenbezogenen Daten den Interessen der betroffenen Personen nicht zuwiderläuft.[285]

Ausschlaggebend für Beurteilung der Art und Weise der Datenverarbeitung ist vor allem die Dauer der Datenspeicherung, die Frequenz der Datenverarbeitung sowie die Menge der personenbezogenen Daten, die verarbeitet werden. Eine Datenverarbeitung ist umso belastender, je länger die Daten gespeichert werden, je höher die Frequenz der Verarbeitungsvorgänge ist und je mehr Datensätze verarbeitet werden.[286] Für die Abschätzung der Folgen ist auch die Menge der datenverarbeitenden Akteure zu berücksichtigen. Je höher die Anzahl der an der Datenverarbeitung beteiligten Akteure (bspw. Verantwortliche, Auftragsverarbeiter, Empfänger) desto höher ist auch die Anzahl der potenziellen Risikoquellen für eine Datenschutzverletzung und desto mehr Personen erhalten Einblicke in die Privatsphäre der betroffenen Person.[287]

---

[282] Vgl. Schantz in Simitis/Hornung/Spiecker (2019), Datenschutzrecht: DSGVO mit BDSG, Art. 6, Rn. 105

[283] Vgl. Schantz in Simitis/Hornung/Spiecker (2019), Datenschutzrecht: DSGVO mit BDSG, Art. 6, Rn. 105

[284] Vgl. Artikel-29-Datenschutzgruppe (2014), WP 217, S. 50

[285] Vgl. Herfurth (2018), Interessenabwägung nach Art. 6 Abs. 1 lit. f DS-GVO: Nachvollziehbare Ergebnisse anhand von 15 Kriterien mit dem sog. „3x5-Modell“, S. 514, 517

[286] Vgl. Herfurth (2018), Interessenabwägung nach Art. 6 Abs. 1 lit. f DS-GVO: Nachvollziehbare Ergebnisse anhand von 15 Kriterien mit dem sog. „3x5-Modell“, S. 519

[287] Vgl. Herfurth (2018), Interessenabwägung nach Art. 6 Abs. 1 lit. f DS-GVO: Nachvollziehbare Ergebnisse anhand von 15 Kriterien mit dem sog. „3x5-Modell“, S. 514, 517

Letztendlich muss die Wahrscheinlichkeit der Verwirklichung potenzieller negativer Folgen für die betroffenen Personen betrachtet werden.[288] Aus dem Zusammenspiel der aufgeführten Schlüsselfaktoren ergibt sich die Gesamtabschätzung des Schweregrads der potenziellen Folgen der Datenverarbeitung.[289]

Bei Datensätzen für das Training von LLMs wird nicht immer eindeutig sein, welche Arten personenbezogener Daten tatsächlich enthalten sind. Da Datensätze, die für das Training von LLMs verwendet werden, typischerweise sehr umfangreich und heterogen sind, können darin grundsätzlich alle möglichen Arten personenbezogener Daten enthalten sein.[290] Grundsätzlich sollte eine Interessenabwägung unter Zuhilfenahme praktischer Annahmen erfolgen, welche auf dem beruhen, „[…] was ein vernünftiger Mensch unter den gegebenen Umständen annehmbar finden würde […]".[291] Bei der Beurteilung der Kritikalität der zu verarbeitenden Daten können auch technische Maßnahmen berücksichtigt werden, die im Rahmen der Trainingsdatenaufbereitung angewendet werden, um den potenziellen negativen Folgen der Datenverarbeitung entgegen zu wirken (z. B. teilweise oder vollständige Unkenntlichmachung von Datenpunkten oder Differential Privacy-Methoden; vgl. hierzu Kapitel 5.2). Ferner sind die Spezifika hinsichtlich der Art und Weise der Datenverarbeitung im Rahmen des LLM-Trainings zu berücksichtigen. Auch wenn Datensätze für das Training von LLMs typischerweise sehr umfangreich sind, enthalten diese i. d. R. personenbezogene Informationen zu einer Vielzahl von natürlichen Personen, sodass die Aussagekraft der enthaltenen Informationen über die einzelnen Datensubjekte relativ gering ist. Der Umstand, dass personenbezogene Daten zu einer Vielzahl von Datensubjekten verarbeitet werden, kann vielmehr zu einer gewissen Relativierung der Betroffenheit führen. In diesem Zusammenhang sollte auch berücksichtigt werden, dass beim Training von LLMs nicht darauf abgezielt wird, Aussagen oder Analysen über einzelne Personen zu treffen, sondern Erkenntnisse über die statistischen Eigenschaften der Trainingsdaten, d. h. wie

[288] Vgl. Artikel-29-Datenschutzgruppe (2014), WP 217, S. 48

[289] Vgl. Artikel-29-Datenschutzgruppe (2014), WP 217, S. 49

[290] Vgl. Longpre et al. (2023), A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality,& toxicity, S. 4

[291] Vgl. Artikel-29-Datenschutzgruppe (2014), WP 217, S. 39

Wörter und Sätze in einem bestimmten Kontext vorkommen, zu gewinnen (vgl. hierzu Kapitel 2.3.3).[292] Betrachtet man den Prozess des LLM-Trainings an sich, wird dieser i. d. R. keine Auswirkungen auf die Rechte und Freiheiten der Datensubjekte haben, da die Daten regelmäßig bereits im Internet öffentlich zugänglich sind und im Rahmen des Trainings keinem neuen, größeren Personenkreis offengelegt werden.[293] Die Wahrscheinlichkeit, dass im Trainingsdatensatz enthaltene personenbezogene Daten während der Inferenz-Phase unberechtigt offengelegt bzw. rekonstruiert werden, hängt Studien zufolge zudem maßgeblich davon ab, wie häufig die jeweiligen Daten im Trainingsdatensatz vorkommen.[294] Textdaten, die wiederholt in einem Trainingsdatensatz vorkommen, stammen oft aus nationalen und internationalen Nachrichten und beziehen sich folglich häufig auf Personen, die in der Öffentlichkeit stehen, bspw. Politiker.[295] Betrachtet man die Quellen der Trainingsdaten, sind dies meist Webseiten, Social Media-Plattformen, Wikipedia-Artikel, Bücher sowie akademische Literatur.[296] Für eine überwiegende Mehrzahl der darin enthaltenen personenbezogenen Informationen wird man annehmen können, dass diese von den Datensubjekten selbst und/oder in einem professionellen Kontext veröffentlicht wurden. Insofern wird man regelmäßig davon ausgehen können, dass für die von einer Datenverarbeitung im Rahmen des Trainings eines LLMs betroffenen Personen zumindest keine besondere Schutzbedürftigkeit vorliegt. Grundsätzlich ist die Wahrscheinlichkeit, dass sich das Risiko einer unberechtigten Offenlegung bzw. Rekonstruktion von personenbezogenen Informationen in der Inferenz-Phase verwirklicht, relativ gering, insbesondere wenn angemessene Schutzmaßnahmen umgesetzt wurden (vgl. hierzu Kapitel 5 und 6).[297] Unter

[292] Vgl. Rosenthal (2024), Was in einem KI-Modell steckt und wie es funktioniert, Blogbeitrag vom 07.05.2024, https://www.vischer.com/know-how/blog/teil-17-was-in-einem-ki-modell-steckt-und-wie-es-funktioniert/ (letzter Aufruf am 31.08.2024)

[293] Vgl. Bartels (2024), A Balancing Act: Data Protection Compliance of Artificial Intelligence, S. 532

[294] Vgl. Carlini et al. (2023), Quantifying Memorization Across Neural Language Models

[295] Vgl. Carlini et al. (2021), Extracting Training Data from Large Language Models, S. 10

[296] Vgl. Longpre et al. (2023), A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality,& toxicity, S. 4

[297] Vgl. Müller et al. (2024), LLMs and Memorization: On Quality and Specificity of Copyright Compliance

Berücksichtigung der üblichen Spezifika und Umstände von Datenverarbeitungen im Rahmen des LLM-Trainings, wird man häufig zu dem Ergebnis kommen, dass die potenziellen negativen Folgen, die sich für die Datensubjekte daraus ergeben, eher weniger schwer ins Gewicht fallen.

7.2.2 Vernünftige Erwartungen der betroffenen Personen

Wie sich bereits aus dem Wortlaut des ErwGr. 47 Satz 1 DSGVO ergibt, sind ebenfalls die vernünftigen Erwartungen der betroffenen Personen für die Gewichtung deren Interessen zu berücksichtigen. Nach allgemeiner Auffassung wird davon ausgegangen, dass eine Datenverarbeitung für die betroffene Person weniger belastend ist, je mehr sie den Erwartungen der betroffenen Personen entspricht. Dementsprechend ist die Datenverarbeitung für die betroffenen Personen umso belastender, wenn sie diese nicht vermuten.[298] Dabei sind etwaige Beziehungen der verantwortlichen Stelle gegenüber den betroffenen Personen sowie die damit zusammenhängenden rechtlichen oder vertraglichen Verpflichtungen zu betrachten.[299] Je enger die Beziehung zwischen der verantwortlichen Stelle und den betroffenen Personen ist, desto eher kann eine Verarbeitung von personenbezogenen Daten erwartet werden.[300] Die vernünftigen Erwartungen ergeben sich zunächst aus den Informationen, die sich für die betroffene Person im Rahmen der Informationspflichten (Art. 13 u. 14 DSGVO) des Verantwortlichen ergeben.[301]

Bei Datenverarbeitungen im Rahmen des LLM-Trainings wird häufig keine direkte vertragliche Beziehung zwischen der verantwortlichen Stelle und den betroffenen Personen vorliegen und eine Informationsbereitstellung nach Art. 13 bzw. 14 DSGVO

[298] Vgl. Herfurth, (2018), Interessenabwägung nach Art. 6 Abs. 1 lit. f DS-GVO: Nachvollziehbare Ergebnisse anhand von 15 Kriterien mit dem sog. „3x5-Modell“, S. 518

[299] Vgl. Artikel-29-Datenschutzgruppe (2014), WP 217, S. 51; Schulz in Gola (2018), Datenschutz-Grundverordnung: VO (EU) 2016/679, Art. 6, Rn. 64 f.

[300] Vgl. Herfurth (2018), Interessenabwägung nach Art. 6 Abs. 1 lit. f DS-GVO: Nachvollziehbare Ergebnisse anhand von 15 Kriterien mit dem sog. „3x5-Modell“, S. 518

[301] Vgl. Robrahn/Bremert (2018), Interessenskonflikte im Datenschutzrecht: Rechtfertigung der Verarbeitung personenbezogener Daten über eine Abwägung nach Art. 6 Abs. 1 lit. f DS-GVO, S. 294; Schulz in Gola (2018), Datenschutz-Grundverordnung: VO (EU) 2016/679, Art. 6, Rn. 63

wird häufig nicht möglich sein. Auch wenn man davon ausgehen kann, dass die personenbezogenen Daten von den Datensubjekten in vielen Fällen selbst, wissentlich veröffentlicht wurden und diese Personen aus objektiver Sicht vernünftigerweise davon ausgehen müssen, dass die von ihnen veröffentlichten Daten gecrawlt und gescraped werden, kann für Sachverhalte des LLM-Trainings regelmäßig ein gewisses Informationsdefizit angenommen werden. Ein solches Informationsdefizit würde sich negativ auf die Beurteilung der vernünftigen Erwartungen der betroffenen Personen auswirken und zu einer schwereren Gewichtung der die Interessen, Grundrechte und -freiheiten der betroffenen Personen führen.

7.3 Zwischenfazit

Insbesondere bei Sachverhalten des LLM-Trainings stellen sich besondere Herausforderungen hinsichtlich der angemessenen Berücksichtigung aller relevanten Faktoren für die Gewichtung und Abwägung der sich gegenüberstehenden Grundrechtspositionen. Häufig – insbesondere, wenn es um die Verarbeitung von Trainingsdaten aus öffentlich zugänglichen Quellen geht – wird man nicht alle relevanten Faktoren abschließend abschätzen können, sodass die Zuhilfenahme objektiv plausibler Annahmen erforderlich sein wird. Dies führt in der Praxis zu einer gewissen Rechtsunsicherheit für die verantwortlichen Stellen.

Auch wenn, beim Training von LLMs regelmäßig ein Transparenzdefizit anzunehmen sein wird, wird die Datenverarbeitung in den meisten Fällen keine schwerwiegenden Risiken für die Rechte und Freiheiten der betroffenen Personen mit sich bringen. Demnach ist grundsätzlich denkbar, dass eine Ausgewogenheit, der sich gegenüberstehenden Positionen erzielt werden kann. Maßgeblich für die Bewertung der Risiken und das Ergebnis der Interessenabwägung im Einzelfall, sind die konkreten Zwecke der Anwendung des LLMs (z. B. allgemeiner Chatbot oder Anwendung im Gesundheitsbereich) sowie die umgesetzten Schutzmaßnahmen (vgl. Kapitel 5 und 6). Aufgrund der verbleibenden rechtlichen Unsicherheiten empfiehlt es sich, zusätzliche, angemessene Schutzmaßnahmen zu implementieren.

8 Zusammenfassung und Fazit

Die zuvor dargestellte Abhängigkeit der datenschutzrechtlichen Zulässigkeit von der Anwendung angemessener und wirksamer Schutzmaßnahmen betont die Relevanz des

sog. risikobasierten Ansatzes, welcher die Grundlage für die abschließende Bewertung der Fragestellungen dieser Arbeit bildet. Nachfolgend werden die zentralen Erkenntnisse zusammengefasst und in einen größeren Zusammenhang gestellt.

8.1 Training von Large Language Models

Im zweiten Kapitel wurde die grundlegende Funktionsweise von maschinellem Lernen und neuronalen Netzen dargestellt. Ferner wurden die einzelnen Trainingsphasen während des Trainings von LLMs beleuchtet. Es kann festgehalten werden, dass für die Entwicklung von qualitativ hochwertigen KI-Modellen Trainingsdaten in großem Umfang benötigt werden. Grundsätzlich gilt dabei, je mehr Trainingsdaten ein Algorithmus erhält, desto präziser kann das KI-Modell optimiert und dessen Fehlerquote verringert werden. Beim LLM-Training werden große Mengen von textbasierten Datensätzen analysiert und bestimmten Wahrscheinlichkeiten von Wortfolgen zugwiesen, wodurch die Durchführung von Schreib-, Konversations-, Zusammenfassungs- und anderen sprachlichen Aufgaben ermöglicht wird. Dabei spielt auch die Qualität der Trainingsdaten eine Rolle. Wird der Algorithmus mit falschen oder nicht repräsentativen Daten trainiert, wird das Modell anfälliger für unerwünschte Outputs, wie bspw. diskriminierende Inhalte. Oft sind heterogene Trainingsdatensätzen aus verschiedenen Domänen (z. B. Bücher und Webseiten) erforderlich, die unterschiedliche Wissens- bzw. Fachgebiete abdecken. Trainingsdaten werden in der Regel aus dem Internet über sog. web-scraping oder web-crawling oder aus eigenen, nicht-öffentlichen Quellen bezogen. Solche Trainingsdatensätze bestehen z. B. aus Freitext, wie er in Aufsätzen, Briefen usw. vorkommt sowie aus Text, der aus formatierten oder halbstrukturierten Daten, wie bspw. Dokumenten, Social Media-Posts, Produktbewertungen oder Webseiten, extrahiert wird. Trainingsdaten enthalten i. d. R. auch Informationen, die sich auf identifizierbare natürliche Personen beziehen. Während des Pre-Trainings lernt das Modell die statistischen Eigenschaften der Trainingsdaten, d. h. wie Wörter und Sätze in einem bestimmten Kontext vorkommen. Der Output-Text wird generiert, indem jeweils das wahrscheinlichste nächste Wort in einer Wortfolge ermittelt wird. Die beim Training verwendeten Datensätze lassen sich aufgrund des Prozesses der sog. Tokenisierung normalerweise nicht mehr 1x1 rekonstruieren. Unter bestimmten Umständen (insb. wenn bestimmte Passagen wiederholt im Trainingsdatensatz vorkommen) kann es dennoch vorkommen, dass LLMs in ihren Ausgaben ganze Passagen aus Trainingsdatensätzen rekonstruieren (sog. Memorisieren). Das Memorisieren von Trainingsdaten kann dazu

führen, dass personenbezogene Daten durch die Eingabe bestimmter Prompts in der Inferenz-Phase offengelegt werden. Dem Memorisieren von Trainingsdaten und dem Erfolg von sog. training data extraction attacks kann durch verschiedene Maßnahmen im Rahmen der Trainingsdatenaufbereitung oder des Pre-Trainings entgegengewirkt werden (z. B. Differential Privacy-Methoden). Allerdings sind solche Maßnahmen oft zeit- und ressourcenaufwendig und wirken sich i. d. R. negativ auf die Performanz des jeweiligen Modells aus.

8.2 Spannungsverhältnis zwischen LLM-Training und DSGVO

Das dritte Kapitel untersucht potenzielle Konflikte zwischen ausgewählten datenschutzrechtlichen Anforderungen der DSGVO und den technischen Anforderungen beim Training von LLMs. Im Fokus stehen die Grundsätze der Rechtmäßigkeit, Transparenz, Fairness (Art. 5 Abs. 1 lit. a DSGVO), der Grundsatz der Zweckbindung (Art. 5 Abs. 1 lit. b DSGVO) und der Grundsatz der Datenminimierung (Art. 5 Abs. 1 lit. c DSGVO). Die Untersuchung in Bezug auf den Grundsatz der Rechtmäßigkeit hat ergeben, dass die wirksame Anwendung datenschutzrechtlicher Rechtsgrundlagen auf das LLM-Training unter bestimmten Umständen grundsätzlich denkbar ist. Dabei kommt es insbesondere auf die Umstände der Datenerhebung, den Anwendungszweck des LLMs sowie die Umsetzung wirksamer Schutzmaßnahmen an. Ob sich eine Datenverarbeitung im Rahmen des LLM-Trainings zulässigerweise auf Art. 6 Abs. 1 lit. f DSGVO oder Art. 9 Abs. 2 lit. e und lit. j DSGVO stützen lässt, kann nicht pauschal beantwortet werden. Insbesondere, dass bei solchen Datenverarbeitungen typischerweise ein gewisser Grad an Intransparenz ggü. den Datensubjekten anzunehmen ist, stellt sich als Herausforderung für die Erfüllung der Wirksamkeitsvoraussetzungen dar. Dieser Umstand wird regelmäßig einen gesteigerten Argumentationsaufwand sowie die Implementierung zusätzlicher Schutzmaßnahmen erforderlich machen. Wenn sich die Datenverarbeitung im Rahmen des LLM-Trainings zulässigerweise auf Art. 6 Abs. 1 lit. f DSGVO stützen lässt, steht die verantwortliche Stelle vor der Herausforderung, das Widerspruchsrecht der betroffenen Personen nach Art. 21 DSGVO zu gewährleisten. Die Umsetzung eines erfolgten Widerrufs ist zum aktuellen Stand der Technik kaum effektiv zu gewährleisten und mit einem unverhältnismäßigem technischen sowie finanziellem Aufwand verbunden. Eine weite Auslegung des Art. 21 Abs. 1 Satz 2 DSGVO unter Berücksichtigung des aktuellen Stands der Technik, nach welcher die aktuell bestehenden technischen Hürden und die zweifelhafte Wirksamkeit bei einer Umsetzung eines

Widerrufs als schutzwürdige Gründe angesehen werden können, die eine besondere Schutzwürdigkeit des jeweiligen Datensubjekts ggf. überwiegen, könnte in diesem Kontext als zielführender Lösungsweg diskutiert werden. Es verbleibt eine gewisse Rechtsunsicherheit, die sich aus der technischen Umsetzbarkeit der rechtlichen Anforderungen ergibt.

Eine gewisse rechtliche Unsicherheit ergibt sich zudem aus der Frage, wie eindeutig bzw. wie präzise die Zwecke der Datenverarbeitung definiert und benannt werden müssen. Die datenschutzrechtliche Zulässigkeit des Trainings eines LLMs sollte jedoch nur in solchen Fällen an dem Zweckbindungsgrundsatz scheitern, in denen besondere Risiken für die Datensubjekte drohen und keine wirksamen Schutzmaßnahmen vorhanden sind.

Problemstellungen im Zusammenhang mit dem Grundsatz von Treu und Glauben bzw. dem Farinessgrundsatz sind strenggenommen auf die Intransparenz zurückzuführen, die beim Training von LLMs regelmäßig anzunehmen ist. Folglich liegt die Herausforderung eher in der Erfüllung des Transparenzgrundsatzes. Eine Einhaltung der Vorschriften in Zusammenhang mit dem Transparenzgrundsatz könnte über die Ausnahmetatbestände der Art. 14 Abs. 5 lit. b Satz 1 HS. 1 Var. 2 DSGVO und Art. 11 Abs. 2 DSGVO gewährleistet werden. Art. 14 Abs. 5 lit. b Satz 1 HS. 1 Var. 2 DSGVO sieht eine Ausnahme der Informationspflichten nach Art. 14 Abs. 1-4 DSGVO vor, wenn die Informationsbereitstellung einen unverhältnismäßigen Aufwand begründet. Aufgrund des in der Regel großen Umfangs und der Diversität der Trainingsdatensätze, würde das Ausfindigmachen der betroffenen Personen objektiv betrachtet regelmäßig einen hohen Aufwand erfordern. Der Ausnahmetatbestand erfordert, dass die verantwortliche Stelle einen unverhältnismäßigen Aufwand darlegen kann, der den negativen Auswirkungen auf die Rechte und Freiheiten der betroffenen Personen überwiegt und das geeignete Schutzmaßnahmen umgesetzt werden. Erforderlich ist zudem die Bereitstellung von Informationen über die Datenverarbeitung für die Öffentlichkeit (bspw. über die Webseite der verantwortlichen Stelle). Nach Art. 11 Abs. 2 DSGVO kann sich die verantwortliche Stelle von der Pflicht zur Erfüllung der Betroffenenrechte nach Art. 15-20 DSGVO (u. a. Recht auf Auskunft) befreien, wenn sie nachweisen kann, dass ihr die Identifizierung der betroffenen Personen nicht möglich ist. Der Ausnahmetatbestand nach Art. 11 Abs. 2 DSGVO ist traditionell jedoch eng auszulegen. Für die Anwendung des Art. 11 Abs. 2 DSGVO auf das LLM-Training wäre eine teleologische Reduktion der Ausnahmeregelung erforderlich. Dabei sollte die Beurteilung der Unmöglichkeit i. S. d.

Art. 11 Abs. 2 DSGVO nicht auf eine absolute Unmöglichkeit beschränkt sein. Stattdessen sollte auf eine Abwägung zwischen dem Aufwand der Informationsbereitstellung und den Auswirkungen ihres Unterbleibens auf die Rechte und Freiheiten der betroffenen Personen abgestellt werden. In der Praxis könnte eine risikobasierte Auslegung, die sich an den tatsächlichen Risiken für die Rechte und Freiheiten der betroffenen Personen orientiert, dazu beitragen, das Spannungsverhältnis zwischen Transparenzanforderungen und den technischen Hürden im Rahmen des LLM-Trainings zu entschärfen.

Da für das Training von LLMs typischerweise sehr umfangreiche Trainingsdatensätze verwendet werden, personenbezogene Daten aus technischer Sicht jedoch nicht direkt notwendig sind, ist ein gewisser Konflikt mit dem Grundsatz der Datenminimierung absehbar. Beim Training von LLMs kommt es mehr auf die statistischen Eigenschaften der Trainingstexte und weniger auf die Identifizierbarkeit der Datensubjekte an. Es kann jedoch eine gewisse indirekte Erforderlichkeit der Verarbeitung personenbezogener Daten angenommen werden. Der Konflikt mit dem Grundsatz der Datenminimierung ist lösbar, wenn man den Grundsatz der Datenminimierung nach einem relativen Verständnis auslegt. Verantwortliche Stellen müssen dabei alle möglichen und zumutbaren Maßnahmen umsetzen, um die Verarbeitung personenbezogener Daten auf das unbedingt notwendige Maß zu beschränken.

Die technischen und qualitativen Erfordernisse beim LLM-Training sind den datenschutzrechtlichen Anforderungen der DSGVO zunächst konträr. Unter Anwendung eines risikobasierten Ansatzes, der die datenschutzrechtlichen Grundsätze in ihrer relativen Auslegung berücksichtigt und sich an den tatsächlichen Auswirkungen der Datenverarbeitung auf die Rechte und Freiheiten der betroffenen Personen orientiert, ist die Vereinbarkeit von Datenverarbeitung im Rahmen des LLM-Trainings mit den Anforderungen der DSGVO dennoch denkbar. Dabei ist die Umsetzung geeigneter technischer und organisatorischer Maßnahmen sowie eine detaillierte Einzelfallprüfung essenziell.

8.3 Personenbezug von Trainingsdaten

Die Anwendbarkeit der DSGVO auf das Training von LLMs wurde in der datenschutzrechtlichen Literatur bisher häufig pauschal angenommen. Im vierten Kapitel der vorliegenden Arbeit erfolgt eine differenziertere Betrachtung der Anwendbarkeit der

DSGVO. Dabei wird der Personenbezug von Trainingsdaten unter Berücksichtigung des relativen Begriffs des Personenbezugs bewertet. Es zeigt sich, dass unter Anwendung des relativen Begriffs des Personenbezugs, ein Personenbezug von Trainingsdaten nicht pauschal angenommen werden kann. Auch wenn die Möglichkeit einer Identifikation von Datensubjekten durch KI-Hersteller grundsätzlich bestünde, lassen sich unter Anwendung des relativen Begriffs des Personenbezugs und unter Berücksichtigung der technischen Herausforderungen in der Praxis sowie der subjektiven Interessen von KI-Herstellern, Argumente für die Verneinung eines Personenbezugs von Trainingsdaten aus öffentlich zugänglichen Quellen finden.

Auch wenn man annimmt, dass aus der Perspektive der KI-Hersteller kein Personenbezug vorliegt, neutralisiert dies potenzielle negative Auswirkungen auf die Rechte und Freiheiten der Datensubjekte jedoch nicht gänzlich. Dies wirft die Frage auf, ob die starre Abhängigkeit der Anwendbarkeit der DSGVO von dem Merkmal des Personenbezugs noch zeitgemäß ist. Im Sinne des Telos der Technologieneutralität sollte ggf. eine Neuausrichtung der DSGVO diskutiert werden, die weniger strikt auf das Schutzgut der personenbezogenen Daten fokussiert ist und stattdessen verstärkt die tatsächlichen Risiken für die Privatsphäre natürlicher Personen berücksichtigt.

Ebenso zeigt sich, dass eine pauschale Anwendung der verschärften Anforderungen des Art. 9 DSGVO häufig nicht sachgerecht wäre. Eine Auslegung des Art. 9 DSGVO streng nach dem Wortlaut der Norm ohne Berücksichtigung des konkreten Verarbeitungskontexts kann zu einem unverhältnismäßigen Eingriff in die unternehmerische Freiheit führen. Für die Bewertung der Anwendbarkeit des Verarbeitungsverbots nach Art. 9 Abs. 1 DSGVO auf das Training von LLMs wäre eine differenzierte Betrachtung angemessen, die den konkreten Verarbeitungskontext, die subjektive Absicht der verarbeitenden Stelle hinsichtlich der Auswertung sensibler Informationen sowie ein mögliches besonderes Risiko für die Rechte und Freiheiten der betroffenen Personen berücksichtigt. Eine solche Auslegung stünde im Einklang mit der Rechtsprechung des EuGH und dem risikobasierten Charakter der DSGVO.

8.4 Schutzmaßnahmen und Daten-Synthetisierung

Es zeichnet sich ab, dass die datenschutzrechtliche Zulässigkeit des LLM-Trainings maßgeblich von der Umsetzung wirksamer Schutzmaßnahmen abhängt, welche die Eintrittswahrscheinlichkeit möglicher negativer Folgen für die Datensubjekte so weit wie

möglich verringern. In Kapitel 5 und 6 werden einige Schutzmaßnahmen vorgestellt, die nach dem aktuellen Stand der Technik infrage kommen. Die Anwendung von Schutzmaßnahmen erfordert dabei stets einen Kompromiss zwischen deren Wirksamkeit und der utility, sprich der Performanz des LLMs. Die Herausforderung besteht darin, ein Schutzniveau zu gewährleisten, welches in einem angemessenen Verhältnis zur Schwere und zur Eintrittswahrscheinlichkeit der mit der Verarbeitung verbundenen Risiken sowie den erforderlichen Ressourcen (Zeit und Rechenleistung ≙ Kosten) und der Performanz des Modells steht. Im Hinblick auf die sich stetig weiterentwickelnden generativen KI-Systeme und sich daraus ergebende neue Risikoszenarien, werden technische Innovation und kontinuierliche Optimierung erforderlich sein, um die Wirksamkeit und Praxistauglichkeit sog. PETs gewährleisten zu können. Die Forschung im Bereich des privacy-preserving machine learning befindet sich zwar noch in einem frühen Entwicklungsstadium, macht jedoch schnelle Fortschritte. Die Angemessenheit und Wirksamkeit von Schutzmaßnahmen für das LLM-Training sollte folglich regelmäßig und ggf. kurzfristig überprüft werden. In diesem Zusammenhang sollten auch Entwicklungen in der Technologie der Daten-Synthetisierung kontinuierlich beobachtet werden. Die Daten-Synthetisierung bietet in der Theorie eine vielversprechende Möglichkeit, anonyme Trainingsdaten für LLMs zu generieren. Aus der Perspektive der KI-Hersteller könnte eine verantwortungsvolle und effektive Nutzung anonymer, synthetischer Daten für das Training von LLMs eine Lösung für datenschutzrechtliche Herausforderungen darstellen und zugleich den steigenden Bedarf an qualitativ hochwertigen Trainingsdaten decken.

8.5 Interessenabwägung nach Art. 6 Abs. 1 lit. f DSGVO

In Kapitel sieben werden die Aspekte des Sachverhalts des LLM-Trainings, die für die Abwägung der berechtigten Interessen der verantwortlichen Stelle und/oder Dritter mit den Grundrechten und Grundfreiheiten der betroffenen Personen nach Art. 6 Abs. 1 lit. f DSGVO relevant sind, genauer betrachtet. Es stellt sich heraus, dass bei Sachverhalten des LLM-Trainings besondere Herausforderungen hinsichtlich der angemessenen Berücksichtigung aller relevanten Faktoren für die Gewichtung und Abwägung der sich gegenüberstehenden Grundrechtspositionen bestehen. Häufig wird für die Abwägung die Heranziehung objektiv plausibler Annahmen erforderlich sein. Dies führt in der Praxis zu einer gewissen Rechtsunsicherheit für die verantwortlichen Stellen. Grundsätzlich ist jedoch denkbar, dass eine Ausgewogenheit, der sich gegenüberstehenden Positionen

erzielt werden kann. Maßgeblich für die Bewertung der Risiken und das Ergebnis der Interessenabwägung im Einzelfall, sind die konkreten Zwecke der Anwendung des LLMs (z. B. allgemeiner Chatbot oder Anwendung im Gesundheitsbereich) sowie die Umsetzung wirksamer Schutzmaßnahmen.

8.6 Fazit

Die vorangehenden Ausführungen verdeutlichen die Herausforderungen, denen sich KI-Entwickler unter Anwendung des Datenschutzregimes der DSGVO ausgesetzt sehen. Komplexe KI-basierte Technologien, wie LLMs, bringen die DSGVO an ihre Grenzen und lassen an der angepriesenen Technologieneutralität der Verordnung zweifeln. Die Kritik, dass die DSGVO Innovationen und Fortschritt hemme, ist nachvollziehbar, resultiert jedoch häufig aus einer restriktiven Anwendung der Vorschriften. Die zunehmende Verbreitung künstlicher Intelligenz bringt neue Gefahren für demokratische Gesellschaften sowie für die Grundrechte und -freiheiten natürlicher Personen mit sich.[302] Vor diesem Hintergrund wäre eine pauschale Deregulierung zu kurz gegriffen. Vielmehr zeigt sich ein verstärkter Bedarf an einem robusten Datenschutzrecht.[303]

Die meisten Problemstellungen, die sich bei der Anwendung der DSGVO auf neue Technologien, wie LLMs, ergeben, lassen sich durch eine kontext- und risikobasierte Auslegung der DSGVO lösen. Eine solche Auslegung lässt sich in den meisten Fällen mit der gesetzgeberischen Intention vereinbaren und erscheint notwendig, um den Grundrechtsschutz im Hinblick auf immer rasantere technologische Entwicklungen gewährleisten zu können. Um eine innovations- und fortschrittshemmende Wirkung der DSGVO zu vermeiden, müssen die Risiken und Auswirkungen auf Grundrechte natürlicher Personen in der Praxis sorgfältig differenziert und der risikobasierte Ansatz als wichtige Grundlage berücksichtigt werden. Der Datenschutz sollte nicht als Selbstzweck betrachtet werden, sondern als Mittel zum Schutz der Privatsphäre. Dabei

[302] Vgl. Louban et al. in Friedewald, Roßnagel, Heesen, Krämer, Lamala (2022), Das Phänomen Deepfakes: Künstliche Intelligenz als Element politischer Einflussnahme und Perspektive einer Echtheitsprüfung, S. 265

[303] Vgl. Jordan in Friedewald, Roßnagel, Heesen, Krämer, Lamala (2022), Nothing personal? Der Personenbezug von Daten in der DSGVO im Licht von künstlicher Inteliligenz und Big Data, S. 64

muss auch berücksichtigt werden, dass es keinen generellen Vorrang des „Grundrechts auf Datenschutz und Privatsphäre“ vor anderen Grundrechten und -freiheiten gibt.

Ein solcher Lösungsansatz zeugt jedoch gerade nicht von einem robusten Datenschutzrecht und stellt lediglich eine Symptombekämpfung dar, die zu einem relativ instabilen Gleichgewicht führt, das – so Rita Jordan – in ein „Leerlaufen des Rechtsschutzes aufgrund einer unübersichtlichen, schwer zu durchdringenden Rechtslage zu kippen droht.“[304] Um langfristige Rechtssicherheit zu schaffen, bedarf es vielmehr einer Ursachenbehebung. Nach Jordan gibt es deutliche Hinweise darauf, dass die derzeitige Ausgestaltung der DSGVO dem steigenden Schutzbedürfnis nicht gerecht wird.[305] Im Rahmen dieser Arbeit hat sich ergeben, dass insbesondere die Fokussierung der DSGVO auf das Schutzgut der personenbezogenen Daten im Lichte neuer Technologien und der zunehmenden „Datafizierung“ nicht mehr zeitgerecht erscheint. Insofern Bedarf es ggf. einer grundlegenden Neuausrichtung der DSGVO, weg von einer starren Ausrichtung auf das Schutzgut der personenbezogenen Daten und hin zu einer verstärkten Berücksichtigung der tatsächlichen Risiken für die Privatsphäre natürlicher Personen. Datenschutz und Wettbewerbsfähigkeit in Europa müssen nicht miteinander im Widerspruch stehen und sollten in Einklang gebracht werden.

[304] Vgl. Jordan in Friedewald, Roßnagel, Heesen, Krämer, Lamala (2022), Nothing personal? Der Personenbezug von Daten in der DSGVO im Licht von künstlicher Inteliligenz und Big Data, S. 67

[305] Vgl. Jordan in Friedewald, Roßnagel, Heesen, Krämer, Lamala (2022), Nothing personal? Der Personenbezug von Daten in der DSGVO im Licht von künstlicher Inteliligenz und Big Data, S. 64

9 Literaturverzeichnis

Artikel-29-Datenschutzgruppe. (2. April 2013). Stellungnahme 03/2013 zu Zweckbindung.

Artikel-29-Datenschutzgruppe. (2014). Stellungnahme 06/2014 zum Begriff des berechtigten Interesses des für die Verarbeitung Verantwortlichen gemäß Artikel 7 der Richtlinie 95/46/EG. *Working Papar 217*.

Artikel-29-Datenschutzgruppe. (11. April 2018). Leitlinien für Transparenz gemäß der Verordnung 2016/679.

Bartels, M. (2024). A Balancing Act: Data Protection Compliance of Artificial Intelligence. *GRUR Int.*, S. 526.

Bierbauer, D., & Helminger, L. (2023). Offenlegung von Daten unter Wahrung der Privatsphäre mittels SMPC (Secure Multiparty Computation). *Austrian Law Journal*(10), 1-34.

Bitkom e.V.; Deutsches Forschungszentrum für Künstliche Intelligenz GmbH. (2024). *Entscheidungsunterstützung mit Künstlicher Intelligenz - Wirtschaftliche Bedeutung, gesellschaftliche Herausforderungen, menschliche Verantwortung.* Von https://www.bitkom.org/sites/main/files/file/import/FirstSpirit-1496912702488Bitkom-DFKI-Positionspapier-Digital-Gipfel-AI-und-Entscheidungen-13062017-2.pdf abgerufen

Bremert, B., & Robrahn, R. (2018). Interessenskonflikte im Datenschutzrecht: Rechtfertigung der Verarbeitung personenbezogener Daten über eine Abwägung nach Art. 6 Abs. 1 lit. f DS-GVO. *ZD*(7), S. 291–297.

Brink, S., & Wolff, H. (2020). *BeckOK Datenschutzrecht* (34 Ausg.). München: C.H. Beck.

Britz, T., Indenhuck, M., & Langerhans, T. (2021). Die Verarbeitung „zufällig“ sensibler Daten. *ZD*, S. 559.

Bundesbeauftragte für den Datenschutuz und die Informationsfreiheit. (2020). *Positionspapier zur Anonymisierung unter der DSGVO unter besonderer*

Berücksichtigung der TK-Branche. Von https://www.bfdi.bund.de/SharedDocs/Downloads/DE/Konsultationsverfahren/1_Anonymisierung/Positionspapier-Anonymisierung.pdf?__blob=publicationFile&v=4 abgerufen

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramèr, F., & Zhang, C. (2023). Quantifying Memorization Across Neural Language Models. *ICLR 2023.*

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., . . . Raffel, C. (15. Juni 2021). Extracting Training Data from Large Language Models.

Chen, H., Waheed, A., Li, X., Wang, Y., Wang, J., Raj, B., & Abdin, M. (2024). On the Diversity of Synthetic Data and its Impact on Training Large Language Models. doi:https://doi.org/10.48550/arXiv.2410.15226

Chen, L. (29. August 2022). Building a Customized PII Anonymizer with Microsoft Presidio. *Towards Data Science*. Abgerufen am 1. Januar 2025 von https://towardsdatascience.com/building-a-customized-pii-anonymizer-with-microsoft-presidio-b5c2ddfe523b/

Chowdhury, D., & Mattackal, L. P. (24. Januar 2025). European executives join Trump's call for action on deregulation. *Reuters*. Abgerufen am 29. Januar 2025 von https://www.reuters.com/world/davos-european-executives-join-trumps-call-action-deregulation-2025-01-24/

Datenethikkommission - Bundesministerium des Inneren, für Bau und Heimat. (2019). Gutachten der Datenethikkommission: Ethik und Digitalisierung – Daten, Algorithmen, Künstliche Intelligenz. Berlin.

Dieker, A. (2024). Datenschutzrechtliche Zulässigkeit der Trainingsdatensammlung. *ZD*, S. 132.

Drechsler, J., & Jentzsch, N. (2018). *Synthetische Daten: Innovationspotential und gesellschaftliche Herausforderungen.* Berlin: Stiftung Neue Verantwortung.

Ehmann, E., & Selmayr, M. (2018). *DS-GVO: Datenschutz-Grundverordnung* (2 Ausg.). München: C.H.Beck.

Ehmann, E., & Selmayr, M. (2024). *Datenschutz-Grundverordnung.* C.H. Beck.

Ertel, W. (2021). Grundkurs Künstliche Intelligenz - Eine praxisorientierte Einführung. In G. K.-I. Barbara Hammer (Hrsg.), *Computational Intelligence* (Bd. 5). Wiesbaden, Deutschland: Springer Vieweg.

Europäische Kommission. (2018). Künstliche Intelligenz für Europa, COM (2018) 237 final. Abgerufen am 29. Januar 2025 von https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:52018DC0237

Europäische Kommission. (4. April 2018). Mitteilung der Kommission an das Europäische Parlament, den Europäischen Rat, den Rat, den Europäischen Wirtschafts- und Sozialausschuss und den Ausschuss der Regionen - Künstliche Intelligenz für Europa. Abgerufen am 1. Januar 2025 von https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:52018DC0237

Europäische Kommission. (2021). Koordinierter Plan für künstliche Intelligenz, Überarbeitung 2021.

Europäischer Datenschutzausschuss . (8. Oktober 2024). Guidelines 1/2024 on processing of personal data based on Article 6(1)(f) GDPR.

Europäischer Datenschutzausschuss. (29. Januar 2020). Leitlinien 3/2019 zur Verarbeitung personenbezogener Daten durch Videogeräte.

Europäischer Datenschutzausschuss. (28. März 2023). Leitlinien 01/2022 zu den Rechten der betroffenen Person - Auskunftsrecht.

Europäischer Datenschutzausschuss. (17. Dezember 2024). Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models. 19-30.

Europäischer Datenschutzausschuss. (2024). Report of the Work Undertaken by the ChatGPT Taskforce.

Europäischer Datenschutzbeauftragter. (kein Datum). TechSonar Report 2021-2022. Von https://www.edps.europa.eu/system/files/2021-12/techsonar_2021-2022_report_en.pdf abgerufen

Feretzakis, G., Papaspyridis, K., Gkoulalas-Divanis, A., & Verykios, V. (2024). Privacy-Preserving Techniques in Generative AI and Large Language Models: A Narrative Review. *Information*(15), S. 697. doi:10.3390/info15110697

Fraunhofer-Gesellschaft. (2018). *Maschinelles Lernen - Eine Analyse zu Kompetenzen Forschung und Anwendung.*

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., . . . Leahy, C. (31. Dezember 2020). The Pile: An 800GB Dataset of Diverse Text for Language Modeling.

Garante per la protezione dei dati personali. (30. März 2024). Web scraping ed intelligenza artificiale generativa nota informativa e possibili azioni di contrasto. Von https://www.gpdp.it/web/guest/home/docweb/-/docweb-display/docweb/10020316 abgerufen

Gierschmann, S. (2018). Gestaltungsmöglichkeiten bei Verwendung von personenbezogenen Daten in der Werbung: Auslegung des Art. 6 Abs. 1 lit. f DS-GVO und Lösungsvorschläge. *MMR*, S. 7-12.

Gola, P. (2018). *Datenschutz-Grundverordnung: VO (EU) 2016/679* (2 Ausg.). München: C.H.Beck.

Golson. (30. Juni 2016). Tesla's Autopilot was involved in a fatal crash for the first time. *The Verge*. Abgerufen am 8. September 2024 von https://www.theverge.com/2016/6/30/12072408/tesla-autopilot-car-crash-death-autonomous-model-s

Hacker, P. (2020). *Datenprivatrecht: Neue Technologien im Spannungsfeld von Datenschutzrecht und BGB.* Thübingen: Mohr Siebeck. doi:10.1628/978-3-16-159618-6

Hamburgischer Beauftragter für Datenschutz und Informationsfreiheit. (15.. Juli 2024). Diskussionspapier: Large Language Models und personenbezogene Daten. Von https://datenschutz-hamburg.de/fileadmin/user_upload/HmbBfDI/Datenschutz/Informationen/240715_Diskussionspapier_HmbBfDI_KI_Modelle.pdf abgerufen

Hazy. (kein Datum). Abgerufen am 9. Januar 2025 von https://hazy.com/

Herfurth, C. (2018). Interessenabwägung nach Art. 6 Abs. 1 lit. f DS-GVO: Nachvollziehbare Ergebnisse anhand von 15 Kriterien mit dem sog. „3x5-Modell“. *ZD*(11), S. 514–520.

Hiltscher, J., & Faust, M. (28. Januar 2025). Deepseek zwingt die Konkurrenz, sich neu zu erfinden. *golem.de*. Abgerufen am 29.. Januar 2025 von https://www.golem.de/news/deepseek-der-sputnik-der-ki-branche-2501-192813.html

Hittmeir, M., Ekelhart, A., & Mayer, R. (2019). On the Utility of Synthetic Data: An Emperical Evaluation on Machine Learning Tasks. *Proceedings of the 14th International Conference on Availabililty, Reliability and Security (ARES 2019).* Canterbury: ACM. doi:10.1145/3339252.3339281

Hittmeir, M., Mayer, R., & Ekelhart, A. (2022). Utility and Privacy Assessment of Synthetic Microbiome Data. *Data and Applications Security and Privacy XXXVI: Proceedings of the 36th Annual IFIP WG 11.3 Conference*, (S. 15-27). Newark. doi:10.1007/978-3-031-10684-2_2

Hornung, G., & Wagner, B. (2020). Anonymisierung als datenschutzrelevante Verarbeitung? Rechtliche Anforderungen und Grenzen für die Anonymisierung personenbezogener Daten. *ZD*, S. 223-228.

Jordan, R. (2022). Nothing personal? Der Personenbezug von Daten in der DSGVO im Licht von künstlicher Intelligenz und Big Data. In M. Friedewald, A. Roßnagel, J. Heesen, N. Krämer, & J. Lamla, *Künstliche Intelligenz, Demokratie und Privatheit.* Baden-Baden: Nomos.

Khder, M. (Dezember 2021). Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. *International Journal of Advances in Soft Computing and its Applications*, 145-168. doi:10.15849/IJASCA.211128.11

Kipker, D.-K., & Lettieri, V. (2024). Synthetische Daten in der medizinischen Forschung – Eine datenschutzrechtliche Bewertung. *Datenschutz und Datensicherheit*(5), S. 284.

Kohn, M., & Schleper, J. (2023). Die (zufällige) Erhebung sensibler Daten. *ZD* , S. 723.

Konferenz der unabhängigen Datenschutzaufsichtsbehörden des Bundes und der Länder (DSK). (2021). Orientierungshilfe der Aufsichtsbehörden für Anbieter:innen von Telemedien ab dem 1. Dezember 2021 (OH Telemedien 2021).

König, P., Krafft, T., Schulz, W., & Zweig, K. (2021). Essence of AI: Whats AI? In K. F. Ramsey (Hrsg.), *The Cambridge Handbook of Artificial Intelligence.* Cambridge University Press. doi:https://doi.org/10.1017/9781009072168.005

Krauss, P. (2023). *Künstliche Intelligenz und Hirnforschung - Neuronale Netze, Deep Learning und die Zukunft der Kognition.* Berlin: Springer. doi:https://doi.org/10.1007/978-3-662-67179-5

Kühling , J., & Buchner, B. (2020). *Datenschutz-Grundverordnung, BDSG* (3 Ausg.). München: C.H.Beck.

Laurencon, H., Saulnier, L., Wang, T., Akiki, C., Villanova del Moral, A., Le Scao, T., . . . Jernite, Y. (7. März 2023). The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset. *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks*.

Litwintschik, M. (2017). *Analysing Petabytes of Websites*. Abgerufen am 11. September 2024 von https://tech.marksblogg.com/petabytes-of-website-data-spark-emr.html

Liu, B., Ding, M., Shaham, S., Rahayu, W., Farokhi, F., & Lin, Z. (2021). When Machine Learning Meets Privacy: A Survey and Outlook. *ACM Comput. Surveys*(54). doi:10.1145/3436755

Longpre, S., Yauney, G., Reif, E., Lee, K., Roberts, A., Zoph, B., . . . Ippolito, D. (13.. November 2023). A Pretrainer’s Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity. S. 5. doi:https://doi.org/10.48550/arXiv.2305.13169

Louban, A., Tahraoui, M., Aden, H., Fährmann, J., Krätzer, C., & Dittmann, J. (2022). Das Phänomen Deepfakes. Künstliche Intelligenz als Element politischer Einflussnahme und Perspektive einer Echtheitsprüfung. In M. Friedewald, A. Roßnagel, J. Heesen, N. Krämer, & J. Lamla, *Künstliche Intelligenz, Demokratie*

und Privatheit . Baden-Baden: Nomos. doi:https://doi.org/10.5771/9783748913344-265,

Louban, Tahraoui, Aden, Fährmann, Krätzer, & Dittmann. (2022). *Das Phänomen Deepfakes: Künstliche Intelligenz als Element politischer Einflussnahme und Perspektive einer Echtheitsprüfung* . (R. H. Friedewald, Hrsg.) Baden-Baden: Nomos. doi:https://doi.or/10.5771/9783748913344-1

Manager Magazin. (27. Januar 2024). Kursrutsch bei Techaktien - Das Deepseek-Beben an der Börse. *Manager Magazin*. Abgerufen am 29. Januar 2025 von https://www.manager-magazin.de/unternehmen/tech/deepseek-chinesisches-ki-start-up-sorgt-fuer-boersenbeben-us-techwerte-rutschen-ab-a-0178e1d1-f330-4b9e-91a7-c3338cb936fb

Matejek, M., & Kremer, S. (2023). Sensible Daten soweit das Auge reicht: Der EuGH zapft den „hidden layer“ an. *CR*(4), S. 218.

Mayer, J., & Trauth, D. (2020). Grenzkostenfreie IoT-Services in den Datenmarktplätzen der Zukunft. In M. Rohde, M. Bürger, K. Peneva, J. Mock, & I. f. GmbH (Hrsg.), *Datenwirtschaft und Datentechnologie* (S. 11-30). Berlin: Springer Vieweg. doi:https://doi.org/10.1007/978-3-662-65232-9

McCandlish, S., & Kaplan, J. (23. Janaur 2020). Scaling Laws for Neural Language Models.

McCarthy, P. (15. December 2006). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. *AI Magazine*, S. 27(4), 12. doi:https://doi.org/10.1609/aimag.v27i4.1904

McCoy, R., Smolensky, P., Linzen, T., Gao, J., & Celikyilmaz, A. (2023). How much do language models copy from their training data? Evaluating linguistic novelty in text generation using RAVEN. *Transactions of the Association for Computational Linguistics*(11), S. 652-670.

Mireshghallah, F., Uniyal, A., Wang, T., Evans, D., & Berg-Kirkpatrick, T. (2022). An Empirical Analysis of Memorization in Fine-tuned Autoregressive Language Models. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, (S. 1816-1826). Abu Dhabi.

Moerel, L. (15. Januar 2025). *Perspective: Why data subjects' rights to LLM training data are not relevant*. Abgerufen am 23. Januar 2025 von IAPP Privacy Perspectives: https://iapp.org/news/a/perspective-why-data-subjects-rights-to-llm-training-data-are-not-relevant

Moos, F. (Juli 2024). Personenbezug von Large Language Models. *CR*(7/2024), S. 442.

MOSTLY AI Solutions MP GmbH. (kein Datum). *Data Access and Data Insights for Everyone*. Abgerufen am 9. Januar 2025 von https://mostly.ai/?_gl=1*neyywl*_up*MQ..*_ga*MTM4NzMzMjIyMy4xNzM2Mjc2NTA2*_ga_8NGESMV97J*MTczNjI3NjUwNi4xLjEuMTczNjI3NjUxMC4wLjAuMA

Mueller, F., Görge, R., Bernzen, A., Pirk, J., & Poretschkin, M. (28. Juni 2024). LLMs and Memorization: On Quality and Specificity of Copyright Compliance.

Müller, F., Görge, R., Bernzen, A., Pirk, J., & Poretschkin, M. (28. Juni 2024). LLMs and Memorization: On Quality and Specificity of Copyright Compliance.

Narayanan, A., & Shmatikov, V. (2008). Robust De-anonymization of Large Sparse Datasets. *2008 IEEE Symposium on Security and Privacy*, (S. 111-125). doi:10.1109/SP.2008/SP.33

Niemann, F., & Kevekordes, J. (2020). Machine Learning und Datenschutz (Teil 1): Grundsätzliche Zulässigkeit. *CR*(1), S. 17-25.

Niemann, F., & Kevekordes, J. (2020). Machine Learning und Datenschutz (Teil 2). *CR*(3), S. 179.

OECD. (2020). *Künstliche Intelligenz in der Gesellschaft.* Paris: OECD Publishing. doi:https://doi.org/10.1787/6b89dea3-de

OpenAI. (kein Datum). *OpenAI Tokenizer*. Abgerufen am 7. September 2024 von https://platform.openai.com/tokenizer

Paal , B., & Pauly, D. (2018). *Datenschutz-Grundverordnung Bundesdatenschutzgesetz* (2 Ausg.). München: C.H.Beck.

Paal, B. (2024). KI-Training mit öffentlichen frei zugänglichen Daten im Lichte der DS-GVO-Vorgaben. *ZfDR*, S. 129.

Patel, H., Nagalapatti, L., Gupta, N., Mehta, S., Guttula, S., Sharma Mittal, R., . . . Jain, A. (2020). Overview and Importance of Data Quality for Machine Learning Tasks. *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining.* Virtual Event, USA. doi:https://doi.org/10.1145/3394486.3406477

Plant, R., Giuffrida, V., & Gkatzia, D. (2024). You Are What You Write: Author re-identification privacy attacks in the era of pre-trained language models. *Computer Speech & Language*, 1001746. doi:10.1016/j.csl.2024.101746

Rae, J., & et al. (21. Januar 2022). Scaling language models: Methods, analysis &insights from training Gopher. *Deep Mind*.

Raji, B. (2021). Rechtliche Bewertung synthetischer Daten für KI-Systeme. *Datenschutz und Datensicherheit*(5).

Rigaki, M., & Garcia, S. (kein Datum). A Survey of Privacy Attacks in Machine Learning. *ACM Comput. Surveys*(56, 4), S. 101. doi:https://doi.org/10.1145/3624010

Rosenthal, D. (7. Mai 2024). *Teil 17: Was in einem KI-Modell steckt und wie es funktioniert*. Abgerufen am 31. August 2024 von https://www.vischer.com/know-how/blog/teil-17-was-in-einem-ki-modell-steckt-und-wie-es-funktioniert/

Rosenthal, D. (2024. Juli 2024). *Teil 19: Sprachmodelle mit und ohne personenbezogene Daten*. Abgerufen am 24. November 2024 von https://www.vischer.com/know-how/blog/teil-19-sprachmodelle-mit-und-ohne-personenbezogene-daten/

Roßnagel , A. (2019). Kein „Verbotsprinzip“ und kein „Verbot mit Erlaubnisvor-behalt“ im Datenschutzrecht: Zur Dogmatik der Datenverarbeitung als Grundrechtseingriff. *NJW*(1), S. 1–5.

Royal Society. (2019). *Protecting Privacy in Practice: The Current Use, Development and Limits of Privacy Enhancing Technologies in Data Analysis.* London.

Säcker, F., Rixecker, R., Oetker, H., Limperg, B., & Schubert, C. (2022). *Münchner Kommentar zum Bürgerlichen Gesetzbuch.* München: C.H. Beck.

Schneider, J., & Schindler, S. (2018). Videoüberwachung als Verarbeitung besonderer Kategorien personenbezogener Daten. *ZD*, S. 463.

Schwartmann, R., Jaspers, A., Lepperhoff, N., & Weiß, S. (2022). Praxisleitfaden zum Anonymisieren personenbezogener Daten - Anforderungen, Einsatzklassen und Vorgehensmodell. Stiftung Datenschutz. Abgerufen am 15. 12 2024 von https://stiftungdatenschutz.org/fileadmin/Redaktion/Dokumente/Anonymisierung_personenbezogener_Daten/SDS_Studie_Praxisleitfaden-Anonymisieren-Web_01.pdf

Schwartmann, R., Jaspers, A., Thüsing, G., & Kugelmann, D. (2020). *DS-GVO/BDSG: Datenschutz-Grundverordnung Bundesdatenschutzgesetz (Heidelberger Kommentar).* Heidelberg: C.F. Müller.

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (kein Datum). Membership Inference Attacks Against Machine Learning Models. *2017 IEEE Symposium on Security and Privacy*, (S. 3-18). doi:10.1109/SP.2017.41

Solove, D. (2024). *Artificial Intelligence and Privacy.* (F. L. Review, Hrsg.)

Spies, A. (2024). Wann passt die DS-GVO auf KI? *MMR*, S. 289.

Spiros, S., Gerrit, H., & Spiecker, I. (2019). *Datenschutzrecht: DSGVO mit BDSG.* Baden-Baden: Nomos.

Steen, A. (August 2024). Ableitung als wesentliche Fähigkeit von KI-Systemen. *Künstliche Intelligenz und Recht*(1/2024), S. 7-11.

Süddeutsche Zeitung. (31. März 2023). ChatGPT in Italien gesperrt - Was hinter der Sperre der Software steckt. *Süddeutsche Zeitung*. Abgerufen am 31. August 2024 von https://sueddeutsche.de/wirtschaft/chatgpt-italien-sperre-software-1.5779751

Süddeutsche Zeitung. (22. Januar 2025). Eine halbe Billion für die KI. *Sddeutsche Zeitung*. Abgerufen am 29. Januar 2025 von https://www.sueddeutsche.de/wirtschaft/donald-trump-investitionen-ki-infrastruktur-usa-li.3187605

Sydow, G. (2018). *Europäische Datenschutzgrundverordnung* . Nomos.

Thomas, A., Ifeoluwa Adelani, D., Davody, A., Mogadala, A., & Klakow, D. (2020). Investigating the Impact of Pre-trained Word Embeddings on Memorization in Neural Networks. *23rd International Conference on Text, Speech and Dialogue.* brno.

Turing, A. (1950). Computing Machinery and Intelligence. *Mind*, S. 433-460.

Vogel, P. (2022). Künstliche Intelligenz und Datenschutz. In P. D. Prof. Dr. Dr. Eric Hilgendorf (Hrsg.), *Robotik und Recht* (Bd. 26). Baden-Baden, Deutschland: Nomos. doi:https://doi.org/10.5771/9783748930952

Wagner, P. (2021). Privacy Enhancing Technologies and Synthetic Data. *SSRN Electronic Journal*. doi:http://dx.doi.org/10.2139/ssrn.3762686

Wang, K., Zhu, J., Ren, M., Liu, Z., Li, S., Zhang, Z., . . . Wang, Y. (2024). A Survey on Data Synthesis and Augmentation for Large Language Models. doi:https://doi.org/10.48550/arXiv.2410.12896

Wei, J., Bosma, M., Zhao, V., Guu, K., Wei Yu, A., Lester , B., . . . Le, Q. (2022). FINETUNED LANGUAGE MODELS ARE ZERO-SHOT LEARNERS. *ICLR 2022.*

Wendehorst, C. (20.. Dezember 2024). Tentative Academic Discussion Draft of a Regulation on the protection of personal data in the context of artificial intelligence and the data economy and amending Regulation (EU) 2024/1689 ('AI Data Protection Regulation') Version 1.1. Abgerufen am 31.. Januar 2025 von https://zivilrecht.univie.ac.at/fileadmin/user_upload/i_zivilrecht/Wendehorst/Workshop_Datenschutz/Draft_AI_Data_Protection_Regulation_WENDEHORST_24-12-20.pdf

Wendiggensen, M. (22.. Januar 2025). Trumps Umkremplung der Digitalpolitik hat begonnen. *Frankfurter Allgemeine Zeitung*. Abgerufen am 29.. Janauar 2025 von https://www.faz.net/pro/digitalwirtschaft/kuenstliche-intelligenz/donald-trump-will-ki-regeln-abschaffen-was-das-fuer-europa-heisst-110245127.html

Werry, S. (2023). Generative KI-Modelle im Visier der Datenschutzbehörden. *MMR*, S. 911.

Wikipedia. (kein Datum). *Homomorphismus.* Abgerufen am 17. Dezember 2024 von https://de.wikipedia.org/wiki/Homomorphismus

Wolff, H., Brink, S., & v. Ungern-Sternberg, A. (2021). *BeckOK Datenschutzrecht.* C.H. Beck.

Xu, R., Baracaldo, N., & Joshi, J. (22. September 2021). Privacy-Preserving Machine Learning: Methods, Challenges, and Directions.

Xue, M., Yuan, C., Wu, H., Zhang, Y., & Liu, W. (kein Datum). Machine Learning Security: Threats, Countermeasures and Evaluations. *IEEE Access*, 74720-74742.

Yan, B., Li, K., Xu, M., Dong, Y., Zhang, Y., Ren, Z., & Cheng, X. (14.. März 2024). On Protecting the Data Privacy of Large Language Models (LLMs): A Survey. S. 4. doi:https://doi.org/10.48550/arXiv.2403.05156

Yeom, S., Giacomelli, I., Fredrikson, M., & Jha, S. (2018). Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. *IEEE Computer Security Foundation Symposium (CSF).*

Zeitung, S. (31. März 2023). ChatGPT in Italien gesperrt - Was hinter der Sperre der Softare steckt. *Süddeutsche Zeitung*.

Zerdick, T. (14. Juli 2021). *Is the future of privacy synthetic?* Abgerufen am 14. Januar 2025 von https://www.edps.europa.eu/press-publications/press-news/blog/future-privacy-synthetic_en

Zohra El Mestari, S., Lenzini, G., & Demirci, H. (2024). Preserving data privacy in machine learning systems. *Computers & Security*(137 (2024) 103605), 103605. doi:https://doi.org/10.1016/j.cose.2023.103605

10 Urteilsverzeichnis

BVerfG, Urteil v. 15.12.1983 – 1 BvR 209/83

BVerfG, Urteil v. 29.05.1973 – 1 BvR 424/71, 1 BvR 325/72, NJW 1973, 1176

BVerwG, Urteil v. 27.03.2019 – 6 C 2/18, NJW 2019, 3556

EuGH (2. Kammer), Urteil v. 19.10.2016 – C-582/14 (Breyer/Deutschland)

EuGH, Urteil v. 01.08.2022 – C-184/20

EuGH, Urteil v. 04.07.2023 – C-252/21

EuGH, Urteil v. 09.11.2023 – C-319/22

EuGH, Urteil v. 11.12.2019 – C-708/18, ZD 2020, 148

EuGH, Urteil v. 21.12.2023 – C-667/21

EuGH, Urteil v. 24.09.2019 – C-136/17

Schweizer Bundesgericht, Urteil v. 08.09.2010 – BGE 136 II 508

Anhang A: Abstract (Deutsch)

Diese Masterarbeit untersucht die datenschutzrechtlichen Herausforderungen beim Training großer Sprachmodelle (Large Language Models, LLMs) unter besonderer Berücksichtigung des berechtigten Interesses gemäß Art. 6 Abs. 1 lit. f DSGVO als Rechtsgrundlage sowie dem Potenzial synthetischer Daten für ein datenschutzfreundliches Training großer Sprachmodelle. Ausgangspunkt ist die zunehmende Nutzung LLM-basierter KI-Systeme, der steigende Bedarf an Trainingsdaten und die damit verbundenen Spannungen mit zentralen Grundsätzen der DSGVO wie Zweckbindung, Datenminimierung und Transparenz. Die Arbeit beleuchtet die Funktionsweise von LLMs sowie die rechtlichen Unsicherheiten bei der Anwendung der DSGVO. Im Zentrum steht die Frage, ob typische Trainingsdaten einen Personenbezug i. S. d. DSGVO aufweisen und das Training von LLMs in den Anwendungsbereich der DSGVO fällt. Untersucht wird zudem, unter welchen Voraussetzungen das berechtigte Interesse eine tragfähige Rechtsgrundlage für das LLM-Training darstellen kann. Dabei wird auch das Potenzial synthetischer Daten als möglicher Ausweg aus dem Spannungsverhältnis zwischen DSGVO und LLM-Training geprüft. Basierend auf einer interdisziplinären Methodik werden technische Grundlagen, rechtliche Rahmenbedingungen sowie Schutzmaßnahmen wie Differential Privacy evaluiert. Ziel ist es, praxisnahe Empfehlungen für datenschutzkonformes LLM-Training zu entwickeln. Die Arbeit kommt zu dem Ergebnis, dass ein risikobasierter und kontextabhängiger Umgang mit den Anforderungen der DSGVO erforderlich ist, um voranschreitende technische Innovationen und Grundrechtsschutz in Einklang zu bringen.

Anhang B: Abstract (English)

This master’s thesis examines the data protection challenges involved in training large language models (LLMs), with a particular focus on the legitimate interest under Article 6(1)(f) GDPR as a legal basis, as well as the potential of synthetic data for privacy-friendly LLM training. The analysis begins with the increasing use of LLM-based AI systems, the growing demand for training data, and the resulting tensions with key GDPR principles such as purpose limitation, data minimization, and transparency. The thesis explains the technical functioning of LLMs and highlights the legal uncertainties regarding the GDPR's applicability. At its core, the study investigates whether typical training data constitutes personal data under the GDPR and whether LLM training falls within the scope of the regulation. It further explores the conditions under which the legitimate interest may serve as a valid legal basis for LLM training. The potential of synthetic data is assessed as a possible way to ease the tension between GDPR requirements and LLM development. Using an interdisciplinary approach, the thesis evaluates technical fundamentals, legal frameworks, and safeguards such as differential privacy. The objective is to provide practical guidance for GDPR-compliant LLM training. The thesis concludes that a risk-based and context-sensitive interpretation of the GDPR is essential to align ongoing technological innovation with the protection of fundamental rights.