Schriften zum Daten-, Informations- und Medienrecht

Boris P. Paal | Tristan Radtke

Training eines Sprachmodells in der Justiz

Urheberrechtliche und datenschutzrechtliche Anforderungen

Nomos

Schriften zum Daten-, Informations- und Medienrecht

Herausgegeben von
Prof. Dr. Boris P. Paal, M.Jur.

Band 84

Boris P. Paal | Tristan Radtke

Training eines Sprachmodells in der Justiz

Urheberrechtliche und datenschutzrechtliche Anforderungen

Nomos

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

1. Auflage 2026

Publiziert von
Nomos Verlagsgesellschaft mbH & Co. KG
Waldseestraße 3–5 | 76530 Baden-Baden
www.nomos.de

Gesamtherstellung:
Nomos Verlagsgesellschaft mbH & Co. KG
Waldseestraße 3–5 | 76530 Baden-Baden

ISBN (Print): 978-3-7560-2041-6
ISBN (ePDF): 978-3-7489-5581-8

DOI: https://doi.org/10.5771/9783748955818

Onlineversion
Nomos eLibrary

Über das Gutachten

Diesem Werk zugrunde liegt ein Rechtsgutachten, das als selbstständige Gutachterleistung durch Herrn Prof. Dr. Boris P. Paal, M.Jur. (Oxford), im Auftrag des Bayerischen Staatsministeriums der Justiz erstellt wurde.

Über die Autoren des Werks

Prof. Dr. Boris P. Paal, M.Jur. (Oxford), ist Inhaber des Lehrstuhls für Law and Regulation of the Digital Transformation an der School of Social Sciences and Technology, Department of Governance, Technische Universität München.

Dr. Tristan Radtke, LL.M. (NYU), ist Akademischer Rat a.Z. und Habilitand ebenda.

Hinweis zu Abkürzungen

Abkürzungen erfolgen nach *Kirchner,* Abkürzungsverzeichnis der Rechtssprache, 11. Aufl. 2024.

Inhaltsverzeichnis

A. Sachverhalt

I. Projekt Generatives Sprachmodell für die deutsche Justiz (GSJ-Projekt)

Gegenstand des Auftrags ist die Erstellung eines Gutachtens über datenschutz- und urheberrechtliche Fragen im Kontext der Entwicklung eines generativen Sprachmodells für die deutsche Justiz (***GSJ-Projekt***) mit einem Fokus auf konkrete Use-Cases.

Von der Begutachtung umfasst sind die folgenden Use-Cases betreffend die Richterschaft bzw. der Richterschaft zuarbeitende Personen[1], wobei die Use-Cases jeweils auf einer Eingabeaufforderung und einem Eingabekontext aus einer konkreten Akte aufbauen:

(1) die Erstellung einer tabellarischen Gegenüberstellung der Sachverhaltsdarstellungen in Zivilsachen (in der Praxis zunächst Miet- und Verkehrssachen) der Parteien mit einer Klassifikation in „strittig/unstrittig". Diese tabellarische Gegenüberstellung soll auf einen entsprechenden Wunsch des Benutzers hin auch relativ zu einer gewählten Anspruchsgrundlage kontextuell erfolgen können;
(2) die Vorformulierung des textuellen Sachverhaltsabschnitts eines Urteils auf der Basis der Streitstrukturierung, gegebenenfalls mit weiterem Input der Richter zur Beweiswürdigung;
(3) die Erstellung eines Zeitstrahls in der Anwendung, der sowohl Ereignisse des Sachverhalts als auch Prozesshandlungen umfasst.

Ein zentrales Forschungsziel des zu begutachtenden GSJ-Projekts ist die Prüfung, wie sich das Training eines Sprachmodells mit großen Mengen deutscher Gerichtsurteile und Aktenauszüge (hier vornehmlich Schriftsätze von Parteien, Gerichtsbeschlüsse) auf die Performanz des Sprachmodells in bestimmten Textgenerierungsaufgaben aus dem Justizalltag auswirkt. Zum Zweck der automatisierten Anonymisierung von Urteilen und Aktenauszügen steht im GSJ-Projekt ein von der Universität Erlangen-Nürnberg entwickeltes Software-Tool zur Verfügung (***Erlanger Tool***).

1 Aus Gründen der besseren Lesbarkeit wird auf die gleichzeitige Verwendung der Sprachformen männlich, weiblich und divers (m/w/d) verzichtet. Die verwendete Sprachform orientiert sich im Folgenden grundsätzlich an der im Gesetz vorgegebenen Sprachform.

Die im Rahmen des GSJ-Projekts verwendeten Gerichtsentscheidungen und Aktenauszüge entstammen insbesondere dem Miet- und Verkehrsrecht. Demgegenüber sind von der Verwendung ausgeschlossen – insbesondere – sensible Verfahrenskategorien (z.B. Strafrecht, Arzthaftungssachen und Verfahren vor den Familiengerichten).

Das GSJ-Projekt wird durch das Bayerische Staatsministerium der Justiz und durch das Ministerium der Justiz des Landes Nordrhein-Westfalen (zusammen ***Ministerien***) geleitet. Die Professur für Legal Tech von Herrn Professor Matthias Grabmair an der TUM School of Computation, Information and Technology führt in Zusammenarbeit mit dem Lehrstuhl für Bürgerliches Recht, Handels- und Gesellschaftsrecht, Arbeitsrecht und Europäische Privatrechtsentwicklung an der Universität Köln von Frau Professorin Barbara Dauner-Lieb (zusammen: ***ausführende Stellen***) im Auftrag der Ministerien das Training des Sprachmodells durch, um den anschließenden Einsatz des trainierten Sprachmodells durch die Ministerien bzw. die Justiz zu ermöglichen. Im Rahmen des GSJ-Projekts übernehmen die ausführenden Stellen eigenständig die technische Umsetzung und führen nach eigenem Ermessen gegebenenfalls das Debugging oder Testläufe durch. Die ausführenden Stellen verwenden die erhaltenen Datensätze nur nach Weisungen der Ministerien und insbesondere nicht für eigene Forschungszwecke.

Im Rahmen des GSJ-Projekts anonymisieren die Ministerien zunächst mittels des Erlanger Tools die Gerichtsentscheidungen sowie Aktenauszüge eigenständig oder durch die ausführenden Stellen als Auftragnehmer. Die (teil-)anonymisierten Gerichtsentscheidungen und Aktenauszüge werden sodann für das weitere KI-Training verwendet.

II. Tatsächliche Annahmen – Thematische Aus- und Eingrenzungen

Der rechtlichen Begutachtung liegen die folgenden tatsächlichen Annahmen sowie Ein- und Ausgrenzungen des Untersuchungsgegenstandes zugrunde.

Nicht zum Gegenstand der Untersuchung werden insbesondere gemacht:

- die technische Prüfung der Anonymisierungsanforderungen, auch und gerade im Hinblick auf den Einsatz des Erlanger Tools. Die Untersuchung dient vielmehr der Herausarbeitung der rechtlichen Maßstäbe und der Überprüfung der Einhaltung rechtlicher Anforderungen auf der

Grundlage technischer Annahmen und zur Verfügung gestellter Sachverhaltsinformationen;
- die Prüfung der Use-Cases anhand der KI-VO.

Das Erlanger Tool erzielt:

- mit Blick auf Gerichtsentscheidungen einen Recall, d.h. einen Anteil erfolgreich anonymisierter Merkmale von allen zu anonymisierenden Merkmalen eines Texts innerhalb einer Kategorie, in einem Spektrum von 94-96 % für eindeutig identifizierende Merkmale bzw. Kennungen, wie etwa Namen und Anschriften natürlicher Personen;[2]
- mit Blick auf die übrigen Informationskategorien (z.B. Daten zum Prozessablauf, Aktenzeichen und Gerichtsort) einen Recall in einem Spektrum von 68-96 %;[3]
- mit Blick auf Aktenauszüge vergleichbare Recall-Werte, wobei sich aufgrund der abweichenden Struktur und aufgrund von Digitalisierungsdefiziten (z.B. betreffend handschriftlicher Anmerkungen) geringere Recall-Werte ergeben können.

Signifikante Abweichungen bei den jeweiligen Recall-Werten, die sich im Laufe der weiteren Evaluation des Erlanger Tools ergeben, könnten gegebenenfalls neue Beurteilungen erfordern.

Es wird unterstellt, dass:

- den Ministerien die Gerichtsentscheidungen und Aktenauszüge rechtmäßig von den Gerichten übermittelt worden sind;
- die Zusammenarbeit der Ministerien im Allgemeinen rechtlich zulässig ist.

Die Verordnung (EU) 2025/327 über den europäischen Gesundheitsdatenraum, die beispielsweise Patientenkurzakten umfasst, bleibt für die Begutachtung außer Betracht.

2 *Adrian et al.*, in: Adrian/Kohlhase/Evert/Zwickel, Manuelle und automatische Anonymisierung von Urteilen, S. 188–189, 194, 196 ff.

3 *Adrian et al.*, in: Adrian/Kohlhase/Evert/Zwickel, Manuelle und automatische Anonymisierung von Urteilen, S. 188–189, 194, 196 ff.

III. Technische Grundzüge des Sprachmodells und späteren KI-Systems

Die technischen Grundzüge eines Sprachmodells (auch: Large Language Model oder KI-Modell) auf Basis neuronaler Netze wurden bereits vielfach dargestellt.[4] Diese Untersuchung beschränkt sich daher zunächst auf eine Kurzdarstellung der wesentlichen Grundlagen und Begrifflichkeiten, die, wo dies erforderlich und angezeigt ist, im Laufe der Begutachtung vertieft wird.

Das Sprachmodell wird auf Grundlage eines Trainingskorpus an vorbearbeiteten Ausgangstexten trainiert.[5] Diese Vorbearbeitung umfasst mehrere Schritte, in deren Rahmen es zur Erstellung weiterer digitaler Kopien der Ausgangstexte und Änderungen an den Ausgangstexten und deren Format kommt.[6] Sodann werden im Rahmen des Trainings im weit verstandenen Sinne[7] allfällige Zeichenketten (sog. Token, ähnlich den Silben eines Wortes)[8] und deren typische, wahrscheinliche Beziehung zueinander ermittelt und in eine mathematische Repräsentation überführt. Mit dieser mathematischen Repräsentation des Modells werden die sog. Gewichte oder Parameter auf mehreren Ebenen (sog. layers) bezeichnet.[9] In einem Modell sind somit wahrscheinlichkeitsbasierte, mathematische Repräsentationen verschiedener Zeichenketten gespeichert.

Der Einsatz eines Sprachmodells[10] erfordert eine Schnittstelle (z.B. in Verbindung mit einer Benutzeroberfläche), die ein Teil des sog. KI-Systems[11] ist. Das KI-System nimmt Eingabeaufforderungen (sog. Prompts) einschließlich eines Eingabekontexts (z.B. einer Verfahrensakte) entgegen, stellt einen einleitenden Befehl voran (den sog. System Prompt), bringt die

4 *Dornis/Stober*, Urheberrecht und Training generativer KI-Modelle, S. 23 ff.; *Pesch/Böhme*, MMR 2023, 917 (918 f.); *T. Radtke*, ZGE 17 (2025), 1 (5 ff.).

5 Im Überblick zum Training *International Working Group on Data Protection in Technology*, Working Paper on Large Language Models (LLMs), S. 11 ff.; etwa *Becher/Berkhin/Freeman*, in: Ramakrishnan/Stolfo/Bayardo/Parsa, S. 424.

6 *Shalev-Shwartz/Ben-David*, Understanding machine learning, S. 228 ff.

7 D.h. einschließlich eines Pre-Trainings und Fine-Tunings, s. z.B. *Devlin et al.*, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

8 Z.B. „B-uch-staben-ket-ten" nach dem Tokenizer für ChatGPT-4o, https://platform.openai.com/tokenizer.

9 *Naveed et al.*, A Comprehensive Overview of Large Language Models, S. 2.

10 S. auch grundlegend zu Mechanismen der sog. Transformer-Architektur *Vaswani et al.*, Attention Is All You Need.

11 S. hierzu auch die Definition eines KI-Systems in Art. 3 Nr. 1 KI-VO sowie eines KI-Modells in Art. 3 Nr. 63 KI-VO. Danach zeichnet ein Modell insbesondere aus, dass es „in eine Vielzahl nachgelagerter Systeme oder Anwendungen integriert werden kann".

gesamte Eingabe in ein geeignetes Format und übergibt die umgewandelten Daten an das Sprachmodell.

Das Sprachmodell führt auf Grundlage der Eingabe und der trainierten Gewichte als Multiplikatoren einzelner, numerisch repräsentierter Aspekte der Eingabe verschiedene Berechnungen durch und liefert sodann über das KI-System eine wahrscheinlichkeitsbasierte Textausgabe zurück. Abhängig von verschiedenen Faktoren einschließlich der sog. Temperatur-Einstellung kann die Ausgabe kreativer ausgestaltet werden, indem für die generierte Ausgabe nicht nur die jeweils wahrscheinlichsten Zeichenketten aneinandergereiht, sondern zufällig Zeichenketten mit geringerer Wahrscheinlichkeit genutzt werden. Diese Ausgaben können teilweise in Abhängigkeit von der Eingabeaufforderung mit Auszügen aus dem Trainingskorpus übereinstimmen. Ferner können die Ausgaben auf Basis der Trainings- und Eingabedaten auch zu unzutreffenden Aussagen führen (sog. Halluzinationen).[12]

Sowohl auf Ebene der Eingabe als auch der Ausgabe können durch das KI-System zahlreiche Beschränkungen vorgesehen werden. Beispielsweise kann die Eingabe derart beschränkt werden, dass der Nutzer des KI-Systems nur eine Auswahl unter mehreren Eingabeaufforderungen treffen kann. Ein- und Ausgabe können außerdem durch einfache oder fortschrittliche Filter (vor-)bearbeitet werden,[13] bevor die Eingabe sodann an das Sprachmodell bzw. die Ausgabe an den Nutzer übergeben wird.

12 Etwa *Wachter/Mittelstadt/Russell*, R Soc Open Sci 11 (2024), 240197; *Seemann*, Künstliche Intelligenz, Large Language Models, ChatGPT und die Arbeitswelt der Zukunft, S. 28.

13 Z.B., indem (reguläre) Ausdrücke ersetzt werden oder die Ausgabe durch ein weiteres KI-System auf das Vorliegen identifizierender Merkmale überprüft wird, hierzu *Moos*, CR 2024, 442 (450).

B. Gutachtenauftrag und Fragestellungen

Gegenstand des Auftrags ist die Erstellung eines Gutachtens über datenschutz- und urheberrechtliche Fragen im Kontext der Entwicklung eines generativen Sprachmodells für die deutsche Justiz mit einem Fokus auf spezifische Use-Cases. Ein Forschungsziel dieses Projekts ist die Prüfung, wie sich das Training eines Sprachmodells mit großen Mengen deutscher Gerichtsurteile und Aktenauszüge (speziell: Schriftsätze von Parteien, Gerichtsbeschlüsse) auf seine Performanz in bestimmten Textgenerierungsaufgaben aus dem Justizalltag auswirkt. Zum Zweck der automatisierten Anonymisierung von Urteilen und Aktenauszügen steht ein von der Universität Erlangen-Nürnberg speziell entwickeltes Software-Tool zur Verfügung, auf das im Rahmen der Begutachtung nach Absprache mit dem Auftraggeber von dem Auftragnehmer zugegriffen werden kann.

Vor diesem Hintergrund soll das Gutachten folgende Fragen beantworten:

F1: Welche Anforderungen sind an die Anonymisierung von unveröffentlichten Urteilen zu stellen, damit sie für die relevanten Akteure als nichtpersonenbezogene Daten eingestuft werden können? Kann dabei auf die von der Rechtsprechung entwickelten Maßstäbe zur Veröffentlichung von Gerichtsentscheidungen abgestellt werden?

F2: Sollte die zweite Frage unter F1 mit nein beantwortet werden: Gelten veröffentlichte gerichtliche Entscheidungen, die vor ihrer Veröffentlichung nach den von der Rechtsprechung entwickelten Maßstäben bearbeitet wurden, als nicht-personenbezogene Daten, wenn sie für das Training eines Sprachmodells verwendet werden? Falls die Frage wiederum mit nein beantwortet wird: Wirkt sich der Umstand der (Urteils-)Anonymisierung nach den gerichtlichen Maßstäben auf die Haftung oder Sanktionierung unter der DSGVO aus?

F3: Ist eine Anonymisierung von unveröffentlichten Urteilen durch das Erlanger Tool ausreichend, um aus diesen nicht-personenbezogene Daten zu machen? Wie ist gegebenenfalls das Restrisiko einer Verletzung des Schutzes personenbezogener Daten bei Einsatz des KI-Systems zu bewerten?

F4: (analog F3) Ist eine Anonymisierung von Aktenauszügen durch das Erlanger Tool ausreichend, um aus diesen nicht-personenbezogene Daten zu machen? Wie ist gegebenenfalls das Restrisiko einer Verletzung des

Schutzes personenbezogener Daten bei Einsatz des KI-Systems zu bewerten?

F5: Enthalten Sprachmodelle, die mit personenbezogenen Daten (hier: unveröffentlichte Urteile, Aktenauszüge) trainiert wurden, personenbezogene Daten und sind daher datenschutzrechtlich von Relevanz? Gibt es Bedingungen, unter denen solche Sprachmodelle veröffentlicht werden können?

F6: Soweit (a) veröffentlichte Entscheidungen (vgl. F2), Urteile und Aktenauszüge nach Anonymisierung durch das Erlanger Tool (vgl. F3 und 4) oder (b) das trainierte Sprachmodell (vgl. F5) personenbezogene Daten enthalten: Besteht für die Verarbeitung dieser personenbezogenen Daten betreffend das Training des Sprachmodells, dessen Einsatz in den einzelnen Use-Cases und die Veröffentlichung des Sprachmodells eine belastbare Rechtsgrundlage nach Art. 6 Abs. 1, Art. 9 Abs. 2 DSGVO, ohne dass ein gesetzgeberisches Tätigwerden erforderlich ist? Welche zusätzlichen datenschutzrechtlichen Verpflichtungen sind in diesem Fall zu beachten (insbesondere Informationspflichten nach der DSGVO sowie etwaige Pflichten aus der nationalen Umsetzung der JI-Richtlinie)?

F7: Wie ist die Verwendung von Aktenauszügen, insbesondere von anwaltlichen Schriftsätzen, zum Modelltraining urheberrechtlich zu bewerten? Gibt es Bedingungen, unter denen ein Modell mit solchen Daten für den justizinternen Gebrauch trainiert werden kann? Gibt es Bedingungen, unter denen es veröffentlicht werden kann?

C. Executive Summary

Das GSJ-Projekt kann **im Einklang mit den datenschutz- und urheberrechtlichen Anforderungen** realisiert werden.

Eine rechtskonforme Realisierung setzt insbesondere voraus: (1) den Einsatz des **Erlanger Tools**, (2) die Prüfung des Anonymisierungsstandards **älterer Gerichtsentscheidungen**, (3) eine grundsätzlich nur **justizinterne Nutzung**, wobei eine **Veröffentlichung** eines KI-Systems (i.e. nicht des Sprachmodells) **möglich** bleibt, (4) den Abschluss einer **Auftragsverarbeitungsvereinbarung** zwischen den Ministerien als gemeinsam Verantwortliche und den ausführenden Stellen, (5) die weitere **Clusterung** von Entscheidungen und Aktenauszügen sowie (6) die Prüfung von **Nutzungsvorbehalten** in den Aktenauszügen oder die Arbeit mit **abgeleiteten Textformaten**. Datenschutzrechtlich **nicht vorausgesetzt ist eine händische Anonymisierung**, sofern aufgrund ergriffener Maßnahmen (s. unter F5.a) das Sprachmodell als anonym zu bewerten ist.

Die **datenschutzrechtliche Zulässigkeit** stützt sich insbesondere auf den **fehlenden Personenbezug des Sprachmodells**, sofern hinreichende technische und organisatorische Maßnahmen einschließlich des Einsatzes des Erlanger Tools implementiert werden. Die **urheberrechtliche Zulässigkeit** ergibt sich insbesondere aus den Schranken der **§§ 44a, 44b UrhG, alternativ aus dem Rückgriff auf abgeleitete Textformate**.

F1.a: Welche Anforderungen sind an die Anonymisierung von unveröffentlichten Urteilen zu stellen, damit sie für die relevanten Akteure als nicht-personenbezogene Daten eingestuft werden können?

- Für die Frage des Vorliegens personenbezogener Daten ist grundsätzlich ein **relatives Konzept der Identifizierbarkeit** anzuwenden. Während **Kennungen** (z.B. der Name oder eine eindeutige Anschrift) stets zu einem Personenbezug führen, sind sonstige identifizierende Merkmale (z.B. Orts- und sonstige Sachverhaltsangaben) darauf zu prüfen, ob die Merkmale nach **allgemeinem Ermessen wahrscheinlich zur Identifizierung** natürlicher Personen genutzt werden. In diese Prognoseentscheidung sind z.B. die Kosten der Identifizierung, der Empfängerkreis und grundsätzlich auch die rechtliche Zulässigkeit einzustellen (s. dazu unter D.I.1).

- **Publikationswürdige Gerichtsentscheidungen** können – gestützt auf Art. 6 Abs. 1 UAbs. 1 lit. e DSGVO i.V.m. Art. 4 Abs. 1 BayDSG bzw. § 3 Abs. 1 DSG NRW, Art. 6 Abs. 1 UAbs. 1 lit. f DSGVO oder Art. 6 Abs. 4 DSGVO sowie wohl auch gestützt auf Art. 9 Abs. 2 lit. f DSGVO – selbst dann rechtskonform veröffentlicht werden, wenn sie **teilweise personenbezogene Daten** enthalten (z.B. mit Blick auf Personen der Zeitgeschichte und identifizierende mediale Berichterstattung). Die Rechtmäßigkeit einer solchen Veröffentlichung setzt insbesondere voraus, dass Kennungen durch geeignete Pseudonyme (i.e. nicht Initialen) ersetzt und sensible Daten möglichst entfernt werden (siehe dazu unter D.II.5.b). Sofern diese Voraussetzungen vorliegen (siehe dazu unter F1.b und F2.a am Beispiel von Nordrhein-Westfalen), können die anonymisierten Urteile **grundsätzlich auch für das Training des GSJ-Sprachmodells weiterverarbeitet** werden (siehe dazu insbesondere unter D.V.4.c.bb).
- Im Übrigen sind unveröffentlichte Gerichtsentscheidungen anhand unterschiedlicher Merkmale zu **clustern** und auf Wort-/Merkmalsebene zuverlässig zu **anonymisieren** (z.B. Überprüfung von Entscheidungen in Arzthaftungssachen oder Familiensachen und Entscheidungen mit einem besonders umfangreichen Tatbestand und einer Vielzahl beteiligter Personen; siehe dazu unter D.I.3). Diese Anforderungen sind **nicht zwingend zu erfüllen**, wenn und soweit das Sprachmodell und das konkrete Einsatzszenario auch ohne vollständige Anonymisierung der Gerichtsentscheidungen zulässig sind (s. hierzu auch unter F3, F5.a).

F1.b und F2.a: Kann dabei auf die von der Rechtsprechung entwickelten Maßstäbe zur Veröffentlichung von Gerichtsentscheidungen abgestellt werden? Sollte die zweite Frage unter F1 mit nein beantwortet werden: Gelten veröffentlichte gerichtliche Entscheidungen, die vor ihrer Veröffentlichung nach den von der Rechtsprechung entwickelten Maßstäben bearbeitet wurden als nicht-personenbezogene Daten, wenn sie für das Training eines Sprachmodells verwendet werden?

- Die Maßstäbe der **deutschen Justiz** betreffend die Entscheidungsanonymisierung sind in Ansehung des Anwendungsvorrangs der DSGVO **nicht maßgeblich**, können allerdings gleichwohl Anhaltspunkte für die Anonymisierung bieten. Denn im Hinblick auf **publikationswürdige Gerichtsentscheidungen** genügt ein geringerer Anonymisierungsstandard entsprechend den Maßstäben der deutschen Rechtsprechung und die Weiterverarbeitung zu Zwecken des GSJ-Projekts ist grundsätzlich **datenschutzrechtlich zulässig** (siehe dazu unter D.II.5.b und D.V.4.c.bb).

- Die **deutsche Praxis** der Entscheidungsanonymisierung wird in den verschiedenen Gerichtssprengeln unterschiedlich vorgenommen. Jüngere Anonymisierungsrichtlinien **genügen** bei sorgfältiger Umsetzung im Einzelfall regelmäßig insgesamt den datenschutzrechtlichen Anforderungen (z.B. die untersuchte Verwaltungsvorschrift der Justiz Nordrhein-Westfalens aus dem Jahr 2021). Demgegenüber bestehen im Hinblick auf **ältere Entscheidungen** und Richtlinien, die insbesondere Kennungen durch Initialen unzureichend anonymisieren, Bedenken (siehe dazu unter D.II.2-4). Im Übrigen besteht insoweit auch **keine Fiktionswirkung** (siehe dazu unter D.II.5).

F2.b: Falls die Frage wiederum mit nein beantwortet wird: Wirkt sich der Umstand der (Urteils-)Anonymisierung nach den gerichtlichen Maßstäben auf die Haftung oder Sanktionierung unter der DSGVO aus?

- Die (Teil-)Anonymisierung kann die **Auswirkungen** auf betroffene Personen **begrenzen**.
- Zugleich kann eine (Teil-)Anonymisierung gegebenenfalls die rechtmäßige Veröffentlichung unter der DSGVO ermöglichen und ist auch im Rahmen der Haftung sowie Sanktionierung **zu berücksichtigen** (siehe dazu unter D.II).

F3: Ist eine Anonymisierung von unveröffentlichten Urteilen durch das Erlanger Tool ausreichend, um aus diesen nicht-personenbezogene Daten zu machen? Wie ist gegebenenfalls das Restrisiko einer Verletzung des Schutzes personenbezogener Daten bei Einsatz des KI-Systems zu bewerten?

- In Ansehung des einen Wert von 99-100 % unterschreitenden Recall-Werts des Erlanger Tools im Hinblick auf Kennungen und eines deutlichen geringeren Werts mit Blick auf sonstige identifizierende Merkmale dürfte das **Erlanger Tool für sich genommen nicht** den **Anforderungen** an eine **Anonymisierung** im Einklang mit der DSGVO genügen (siehe dazu unter D.III.1).
- Es können aber die mittels Erlanger Tool bearbeiteten Gerichtsentscheidungen zum einen nach den Grundsätzen für die **Veröffentlichung** von Entscheidungen im Einzelfall gegebenenfalls **datenschutzrechtlich zulässig weiterverarbeitet** werden, zum anderen im **GSJ-Projekt durch das KI-Training als Anonymisierungsmaßnahme** und weitere technische und organisatorische Maßnahmen als **anonym anzusehen** sein (siehe hierzu und zum Restrisiko unter F5.a zum Sprachmodell).

F4 (analog F3): Ist eine Anonymisierung von Aktenauszügen durch das Erlanger Tool ausreichend, um aus diesen nicht-personenbezogene Daten zu machen? Wie ist gegebenenfalls das Restrisiko einer Verletzung des Schutzes personenbezogener Daten bei Einsatz des KI-Systems zu bewerten?

- In Ansehung des voraussichtlich geringeren Recall-Werts bei der Anonymisierung von Aktenauszügen **genügt das Erlanger Tool** insoweit erst recht **nicht den Anforderungen an eine Anonymisierung** (siehe dazu unter D.VI.1).
- Für Aktenauszüge kann nicht auf die Standards für die Veröffentlichung wie bei gerichtlichen Entscheidungen zurückgegriffen werden. Im **GSJ-Projekt kann im Ergebnis allerdings ebenfalls eine Anonymisierung** der personenbezogenen Daten vorliegen (siehe hierzu und zum Restrisiko unter F5.a zum Sprachmodell). Das setzt voraus, dass ein **Recall-Wert von mind. 90 % mit Blick auf Kennungen** nicht signifikant unterschritten wird, die Aktenauszüge nach **Dokumentenkategorien geclustert** werden (z.B. Aussortierung oder händische Anonymisierung von beigezogenen Akten der Staatsanwaltschaft, psychologischen und gegebenenfalls ärztlichen Gutachten) und weitere technische und organisatorische Maßnahmen ergriffen werden (z.B. auch die Erkennung von OCR-Fehlern; siehe dazu unter D.VI.2).

F5.a: Enthalten Sprachmodelle, die mit personenbezogenen Daten (hier: unveröffentlichte Urteile, Aktenauszüge) trainiert wurden, personenbezogene Daten und sind daher datenschutzrechtlich von Relevanz?

- Ein Sprachmodell, das mit durch das Erlanger Tool bearbeiteten Gerichtsentscheidungen und Aktenauszügen trainiert wurde und im GSJ-Projekt eingesetzt wird, enthält **grundsätzlich keine personenbezogenen Daten** (siehe dazu unter D.IV.1). Insoweit sind datenschutzrechtliche Anforderungen, z.B. aus Art. 9 DSGVO, nicht zu beachten.
- Dieser Befund setzt voraus, dass
  - nicht im Einzelfall **Kennungen** eindeutig im Modell mathematisch repräsentiert sind. Eine solche mathematische Repräsentation wird mit Blick auf die Länge der Kennungen (z.B. eines Namens als mehrere sog. Token) und die Funktionsweise des KI-Modells als unwahrscheinlich zu bewerten sein;
  - in den Trainingsdaten **deutlich überrepräsentierte Textbestandteile** (z.B. Entscheidungsduplikate, Textbausteine) besonders sorgfältig auf enthaltene Kennungen und sonstige identifizierende Merkmale geprüft werden;

  - o im Hinblick auf **sonstige identifizierende Merkmale keine Wahrscheinlichkeit** der Identifizierung einer natürlichen Person besteht, weil die unter D.VIII ausführlich erörterten Maßnahmen ergriffen werden. Hierzu kommen beispielsweise – i.e., nicht zwingend oder abschließend – Maßnahmen wie die folgenden in Betracht:
    - nur ein beschränkter, **justizinterner Personenkreis** den Zugriff auf ein auf dem Sprachmodell aufbauendes KI-System erhält, wobei zudem ein sicherer Login-Mechanismus verwendet wird;
    - der Zugriff auf das KI-System Gegenstand einer separaten **Nutzungsvereinbarung, Dienstanweisung** o.ä. mit Angaben zur (zulässigen) Bedienung ist;
    - ein geeigneter **System Prompt** zum Einsatz kommt oder den Nutzern die Auswahl von Eingabeaufforderung und eines Eingabekontexts (z.B. einer konkreten E-Akte) ermöglicht wird, nicht aber eine individuelle Eingabeaufforderung verlangt wird;
    - eine **Meldefunktion** für etwaige Datenschutzverletzungen vorgesehen ist;
    - **effektive Ausgabefilter** zum Einsatz kommen (z.B. betreffend den Abgleich von Namen und sonstigen Kennungen dahingehend, ob diese auch in den Eingabedaten verwendet wurden);
    - eine **niedrige Temperatur-Einstellung** zur Vermeidung von datenschutzrechtlich relevanten Halluzinationen gewählt wird;
    - **Eingaben nicht** unmittelbar zur **Verbesserung** des Sprachmodells verwendet werden;
- Eine **Verletzung** des Schutzes personenbezogener Daten nach Art. 4 Nr. 12 DSGVO dürfte unter Beachtung dieser Vorgaben als **sehr unwahrscheinlich einzustufen sein**. Um den möglichen Schaden in einem solchen Fall zusätzlich bereits vorsorglich erheblich zu reduzieren, könnten Gerichtsentscheidungen und Aktenauszüge weiter **geclustert** und geprüft werden, z.B. mit Blick auf die Geltendmachung von Schmerzensgeld und typischerweise zu erwartenden Gesundheitsdaten (siehe dazu unter D.VIII.2).

F5.b: Gibt es Bedingungen, unter denen solche Sprachmodelle veröffentlicht werden können?

- Eine **öffentliche Zurverfügungstellung eines entwickelten KI-Systems** dürfte im Fall von beschränkten Eingabemöglichkeiten und effektiven Ausgabefiltern **datenschutzrechtlich vertretbar** sein (siehe dazu unter D.IV.3 und D.VIII.5).

- Die **Open Source-Veröffentlichung des Sprachmodells** einschließlich seiner Gewichte und gegebenenfalls des Trainingskorpus ist **datenschutzrechtlich bedenklich**. Etwaige Sicherungsmechanismen (z.B. Ausgabefilter) greifen nicht mehr ein und es dürfte ein Personenbezug des Sprachmodells anzunehmen sein. Diese Verarbeitung personenbezogener Daten kann allerdings insbesondere dann **gerechtfertigt** sein, wenn das KI-Modell nur auf Gerichtsentscheidungen trainiert wird, die den Anforderungen an eine Anonymisierung für die Veröffentlichung genügen (siehe dazu unter D.IV.3 und D.VIII.5).

F6: Soweit (a) veröffentlichte Entscheidungen (vgl. F2), Urteile und Aktenauszüge nach Anonymisierung durch das Erlanger Tool (vgl. F3 und 4) oder (b) das trainierte Sprachmodell (vgl. F5) personenbezogene Daten enthalten: Besteht für die Verarbeitung dieser personenbezogenen Daten betreffend das Training des Sprachmodells, dessen Einsatz in den einzelnen Use-Cases und die Veröffentlichung des Sprachmodells eine belastbare Rechtsgrundlage nach Art. 6 Abs. 1, Art. 9 Abs. 2 DSGVO, ohne dass ein gesetzgeberisches Tätigwerden erforderlich ist? Welche zusätzlichen datenschutzrechtlichen Verpflichtungen sind in diesem Fall zu beachten (insbesondere Informationspflichten nach der DSGVO sowie etwaige Pflichten aus der nationalen Umsetzung der JI-Richtlinie)?

- Die Verarbeitung eines Restbestands an personenbezogenen Daten im GSJ-Projekt wird **datenschutzrechtlich** – gestützt auf Art. 6 Abs. 1 UAbs. 1 lit. e DSGVO i.V.m. Art. 4 Abs. 1 BayDSG bzw. § 3 Abs. 1 DSG NRW, Art. 6 Abs. 1 UAbs. 1 lit. f DSGVO oder Art. 6 Abs. 4 DSGVO – vorbehaltlich robuster technischer und organisatorischer Maßnahmen als **zulässig** anzusehen sein (siehe dazu unter D.V). Mit Blick auf **besondere Kategorien personenbezogener Daten** empfiehlt sich eine weitgehende **Clusterung** (s.o.), da zwar im Einzelfall rechtfertigende Rechtsgrundlagen aus Art. 9 Abs. 2 DSGVO in Betracht kommen können, aber keine einheitliche Rechtsgrundlage belastbar die Verarbeitung stützen wird (siehe dazu unter D.V.5).
- Darüber hinaus sind weitere datenschutzrechtliche Anforderungen zu beachten:
  - Im GSJ-Projekt dürfte nach dem zugrunde zu legenden Sachverhalt **keine gemeinsame Verantwortlichkeit** der Ministerien und der ausführenden Stellen vorliegen; insbesondere wird für die Annahme einer gemeinsamen Verantwortlichkeit nicht ausreichen, dass die ausführenden Stellen nach eigenem Ermessen über **Debugging-Maßnahmen**

entscheiden. Es bedarf daher insoweit **keiner Rechtsgrundlage für die Übermittlung** personenbezogener Daten an die ausführenden Stellen, wenn diese als **Auftragsverarbeiter** der Ministerien handeln.
  - Die **Ministerien** werden aufgrund der Zusammenführung der Gerichtsentscheidungen als **gemeinsam Verantwortliche** anzusehen sein und daher die Anforderungen aus Art. 26 DSGVO an die Vereinbarung zu beachten haben (siehe dazu unter D.V.3.b).
  - Die **Information** der betroffenen Personen kann nach Art. 14 Abs. 5 lit. b Hs. 1 Var. 2 DSGVO **unterbleiben**. Erforderlich sind allerdings weitergehende **Maßnahmen** (z.B. die Bereitstellung von Informationen für die Öffentlichkeit, siehe dazu unter D.V.6.a).
  - Die **weiteren Betroffenenrechte** und die Einhaltung der **Datenschutzgrundsätze** erweisen sich im Übrigen **mangels Personenbezugs des Sprachmodells** als **handhabbar**, wenn entsprechende Maßnahmen z.B. gegen Halluzinationen getroffen werden (siehe dazu unter D.V.6.b).
  - Es empfiehlt sich die Durchführung einer **Datenschutzfolgenabschätzung**, gegebenenfalls aufbauend auf die Ergebnisse und Empfehlungen dieser Begutachtung. Ferner sind allgemeine **technische und organisatorische Maßnahmen** umzusetzen und regelmäßig zu überprüfen (z.B. Überprüfung des Erlanger Tools und der übrigen Software anhand des Stands der Technik, Vorsehen einer Meldefunktion für etwaige Datenlecks, Wasserzeichen und gegebenenfalls Logs, siehe dazu unter D.V.6.c).

F7.a,b: Wie ist die Verwendung von Aktenauszügen, insbesondere von anwaltlichen Schriftsätzen, zum Modelltraining urheberrechtlich zu bewerten? Gibt es Bedingungen, unter denen ein Modell mit solchen Daten für den justizinternen Gebrauch trainiert werden kann?

- Das **GSJ-Projekt** kann unter Beachtung verschiedener Maßgaben **urheberrechtskonform** im Hinblick auf Aktenauszüge realisiert werden.
- Im Einzelnen ist das Projekt urheberrechtlich wie folgt zu bewerten:
  - Die Aktenauszüge **können urheberrechtlich geschützte Bestandteile enthalten**. Das betrifft insbesondere einige, aber nicht per se alle **Schriftsätze** sowie vor allem verfasste **Gutachten** und sonstige **Anlagen** als Bestandteil der Aktenauszüge (siehe dazu unter E.I).
  - Das KI-Training und der weitere Einsatz im GSJ-Projekt betreffen maßgeblich das Verwertungsrecht der **Vervielfältigung** (§ 16 UrhG) für das Trainingskorpus sowie im Rahmen der Ausgabe und lösen

gegebenenfalls – regelmäßig im Rahmen von Text und Data Mining gerechtfertigte – **Umgestaltungen** aus (§ 23 UrhG). Mangels Reproduzierbarkeit liegt bei Beachtung der auch datenschutzrechtlich gebotenen Maßnahmen nach vorzugswürdiger, aber nicht unbestrittener Ansicht regelmäßig **keine Vervielfältigung im Sprachmodell vor** (siehe dazu unter E.II).

- Diese Nutzungshandlungen sind grundsätzlich einschließlich der Einbindung Dritter (unter E.II.3) zulässig und können:
  - gegebenenfalls teilweise unterbleiben, indem auf **abgeleitete Textformate** und ein Training an der Datenquelle gesetzt wird (siehe dazu unter E.II.1.a);
  - auf **§ 44b Abs. 2 S. 1 UrhG** (**Text und Data Mining**) gestützt werden, soweit nicht im Einzelfall wirksam ein **Vorbehalt** i.S.d. § 44b Abs. 3 UrhG erklärt wurde (siehe dazu unter E.II.2.c);
  - auf **§ 44a Nr. 2 UrhG** gestützt werden, soweit es nur zu **vorübergehenden** Vervielfältigungen kommt (siehe dazu unter E.II.2.d);
  - **nicht auf § 45 UrhG** mangels eines konkreten **Verfahrensbezugs** gestützt werden (siehe dazu unter E.II.2.b);
  - **nicht auf § 60d UrhG** für **wissenschaftliche Forschungszwecke** gestützt werden, da der Forschungszweck nur Nebenzweck ist und im GSJ-Projekt der operative Einsatz in der Justiz im Vordergrund steht (siehe dazu unter E.II.2.c.dd).
- Die **Voraussetzung** für die urheberrechtliche Zulässigkeit ist, vorbehaltlich des Einsatzes abgeleiteter Textformate zur Vermeidung von Vervielfältigungen, **maßgeblich** die Beachtung von **Vorbehalten** im Rahmen des § 44b UrhG (zu einzelnen Vorbehalten siehe unter E.V.1).
- Vereinzelte **Vervielfältigungen** im Rahmen der **Ausgabe** sind durch geeignete Maßnahmen zu **verhindern** (siehe dazu unter E.V), insbesondere durch:
  - **datenschutzrechtlich** gebotene **Anonymisierungsmaßnahmen**, die es insbesondere erschweren, unter Bezugnahme auf Kennungen urheberrechtlich geschützte Ausgaben aus den Trainingsdaten zu extrahieren (z.B. „Schreibe im Stil von x“);
  - **Nutzungsvereinbarungen, Dienstanweisungen** o.ä. über die Benutzung des KI-Systems;
  - gegebenenfalls **System Prompts und Filtermaßnahmen**, soweit diese effektiv die Wahrscheinlichkeit von Wiedergaben aus den Trainingsdaten reduzieren (können).

- Die Wahrscheinlichkeit einer solchen Vervielfältigung ist ferner zu verringern, indem **Dokumentenkategorien aus dem KI-Training ausgeschlossen** oder nachbearbeitet werden, die voraussichtlich urheberrechtlich geschützt sind, aber gegebenenfalls von geringer Bedeutung für den Einsatz im GSJ-Projekt sind (z.B. ausgewählte Gutachtenkategorien, siehe dazu unter E.V.1).

F7.c: Gibt es Bedingungen, unter denen es veröffentlicht werden kann?

- Die **Veröffentlichung** eines auf dem Sprachmodell aufbauenden **KI-Systems** ist **urheberrechtlich grundsätzlich zulässig**. Empfehlenswert sind Einschränkungen bei der Eingabeaufforderung anstelle einer freien Eingabe (siehe dazu unter E.V.3).
- Die **Open Source-Veröffentlichung des Sprachmodells** einschließlich seiner Gewichte und gegebenenfalls des Trainingskorpus ist **urheberrechtlich bedenklich**. In diesen Fällen dürften regelmäßig Vervielfältigungen vorliegen, die nicht auf eine Schrankenbestimmung gestützt werden können (siehe dazu unter E.II.1.a.bb unter E.V.3).

Aus der **Zusammenschau von Datenschutzrecht und Urheberrecht** lassen sich zudem die folgenden Befunde mit Relevanz für das GSJ-Projekt festhalten:

- Der Begriff der **wissenschaftlichen Forschungszwecke** i.S.d. DSGVO und des UrhG ist in seinem jeweiligen Kern unter beiden Rechtsakten ähnlich auszulegen, wobei zugleich Abweichungen in den Details in Betracht kommen. Die Verfolgung wissenschaftlicher Forschungszwecke ist jeweils abzulehnen, wenn Handlungen mit Forschungsbezug (z.B. Verarbeitungen oder Vervielfältigungen zur Entwicklung neuer Produkte) unmittelbar und vorrangig auf die Entwicklung eines Produkts für den operativen Einsatz abzielen. Wenn und soweit **urheberrechtlich** (privilegierte) wissenschaftliche Forschungszwecke i.S.d. § 60d UrhG angenommen werden, würden sich regelmäßig **datenschutzrechtlich** mit Blick auf Art. 89 DSGVO und nationale Vorschriften **strengere Anforderungen** ergeben – insbesondere – für die Implementierung technischer und organisatorischer Maßnahmen einschließlich der Anonymisierung.
- Sowohl im Datenschutzrecht als auch im Urheberrecht ist für die Bewertung des Sprachmodells die **Möglichkeit der Extraktion** von Daten bzw. Werkbestandteilen aus den **Ausgaben** zu **berücksichtigen**, auch wenn sich die Anforderungen im Detail unterscheiden. Daher erfordern beide Rechtsgebiete jeweils Clusterungen und Maßnahmen im Hinblick auf das konkrete KI-System.

- Die **Entfernung von Kennungen aus den Trainingsdaten** ist zwar primär datenschutzrechtlich geboten, kann sich aber auch urheberrechtlich als vorteilhaft erweisen.

Sofern man hypothetisch für eine Zusammenarbeit zwischen Ministerien und ausführenden Stellen die **Verfolgung eigener Forschungszwecke** unterstellt, ergeben sich folgende Besonderheiten:

- In **datenschutzrechtlicher** Hinsicht wären die Ministerien und die ausführenden Stellen **gemeinsam Verantwortliche** für die Übermittlung und Weiterverarbeitung der in den Gerichtsentscheidungen und Aktenauszügen enthaltenen personenbezogenen Daten (Art. 26 DSGVO). Das Ministerium bedürfte daher in Abweichung vom Sachverhalt des GSJ-Projekts einer **Rechtsgrundlage** nach Art. 6, 9 DSGVO für die Übermittlung personenbezogener Daten zu Forschungszwecken der ausführenden Stellen. Eine solche Rechtsgrundlage würde sich angesichts der zahlreichen betroffenen Personen und in Abhängigkeit von den konkret übermittelten Datenkategorien **nicht ohne Weiteres aus den landesrechtlichen oder sonstigen Vorschriften** ergeben. Insbesondere verlangen die landesdatenschutzrechtlichen Forschungsgeneralklauseln regelmäßig eine Anonymisierung, mithin einen Anonymisierungsgrad, der **über den durch das Erlanger Tool gewährleisteten Anonymisierungsgrad** hinausgeht. Darüber hinaus ergäben sich weitere Anforderungen (z.B. mit Blick auf eine Datenschutzfolgenabschätzung; siehe dazu unter D.VII).
- In **urheberrechtlicher** Hinsicht könnte die Notwendigkeit entfallen, allfällige **Vorbehalte** für ein Text und Data Mining nach § 44b Abs. 3 UrhG zu berücksichtigen, wenn einzelne Vervielfältigungen abweichend vom Sachverhalt des GSJ-Projekts primär dem forschungsbezogenen Text und Data Mining der ausführenden Stellen dienten (§ 60d UrhG). Eine solche Ausnahme nach § 60d UrhG von der Verpflichtung zur Beachtung von Text und Data Mining-Vorbehalten gilt auch in diesem hypothetischen Szenario nicht für Vervielfältigungen, die für ein Text und Data Mining unmittelbar zur Entwicklung des Sprachmodells im operativen Justizeinsatz durchgeführt werden (siehe dazu unter E.IV).

Es wird ferner hingewiesen auf **die Kataloge der Handlungsempfehlungen** aus **datenschutzrechtlicher** Perspektive (siehe dazu unter D.VIII) und **urheberrechtlicher Perspektive** (siehe dazu unter E.V).

D. Datenschutzrechtliche Bewertung

Zunächst ist der Sachverhalt aus datenschutzrechtlicher Sicht zu bewerten. Zu diesem Zweck sind in den Blick zu nehmen die Anforderungen an die Anonymisierung unter der DSGVO (unter I.), der Vergleich mit den deutschen Anforderungen an die Anonymisierung von Gerichtsentscheidungen (unter II.), der Einsatz des Erlanger Tools zur Entscheidungsanonymisierung (unter III.), die Anwendbarkeit des Datenschutzrechts auf ein mit Gerichtsentscheidungen trainiertes, veröffentlichtes Sprachmodell (unter IV.), die datenschutzrechtlichen Anforderungen im Fall der Anwendbarkeit (unter V.) sowie etwaige Besonderheiten bei der Bearbeitung der Aktenauszüge durch das Erlanger Tool (unter VI.).

I. Vorgang der Anonymisierung

Für die Bestimmung der Anforderungen an eine Anonymisierung auf der Grundlage und am Maßstab des Datenschutzrechts ist der Blick zunächst auf den Begriff der personenbezogenen Daten zu richten (unter 1.). Sodann werden aus diesem Begriff die Anforderungen an Anonymisierungstechniken im Allgemeinen (unter 2.) und an Anonymisierungstechniken mit Blick auf Gerichtsentscheidungen im Besonderen (unter 3.) abgeleitet.

1. Personenbezug und Identifizierbarkeit

Die DSGVO findet nach Art. 2 Abs. 1 DSGVO Anwendung auf die (teil-)automatisierte Verarbeitung personenbezogener Daten und setzt für ihre Anwendbarkeit nach Art. 4 Nr. 1, 2 DSGVO den Umgang mit und die Verarbeitung von Informationen voraus, „die sich auf eine identifizierte oder identifizierbare natürliche Person […] beziehen". Auf Informationen, die sich nicht auf eine solche zumindest identifizierbare natürliche Person beziehen (i.e. von vornherein sog. anonyme oder gegebenenfalls nachträgli-

che anonymisierte Daten, ErwGr. 26 S. 5, 6 DSGVO)[14] findet die DSGVO keine Anwendung.

a. „Beziehen"

Über das Merkmal des Beziehens[15] werden (geringe) Anforderungen an den Informationsgehalt mit Blick auf die betroffene Person gestellt.[16] Die Rechtsprechung verlangt, dass die Information „auf Grund ihres Inhalts, ihres Zwecks oder ihrer Auswirkungen mit einer bestimmten Person verknüpft ist".[17] Die Variante der Verknüpfung über den Inhalt adressiert Informationen unmittelbar über die Person, die Variante des Zwecks bezieht sich auf die Möglichkeit, Informationen zu verwenden, um eine Person zu bewerten, die Person entsprechend der Informationen zu behandeln oder zu beeinflussen, und die Variante der Auswirkungen erfasst subsidiär Informationen, die sich in anderer Weise auf die betroffene Person auswirken (können).[18]

Diese Anforderungen sind regelmäßig und auch im vorliegend zu begutachtenden Sachverhalt erfüllt,[19] weshalb das Merkmal des Bezugs im Rahmen der weiteren Untersuchung grundsätzlich nicht weiter problematisiert und lediglich mit Blick auf die mathematische Repräsentation von personenbezogenen Daten in Sprachmodellen nachfolgend unter D.IV näher behandelt wird. Vor diesem Hintergrund ist allerdings die häufig verwendete Formulierung eines absoluten oder relativen Personen*bezugs* ungenau. Deshalb wird in dieser Untersuchung auf ein – absolutes oder relatives – Konzept der Identifizierbarkeit abgestellt.

14 Zu beiden Begrifflichkeiten *Mantz/Spittka*, in: Sassenberg/Faber, § 6 Datenschutz und IT-Sicherheit, Rn. 21.

15 Unter der DSRL über das Merkmal „über eine Person", s. *Artikel-29-Datenschutzgruppe*, WP 136, S. 10; keine materiellen Änderungen *Winter/Battis/Halvani*, ZD 2019, 489 (489 f.).

16 *EuG*, ZD 2023, 399 (Rn. 69 ff.); *Krügel*, ZD 2017, 455 (459); *Paal*, ZfDR 2024, 129 (133).

17 *EuGH*, ZD 2018, 113 (Rn. 35) – Nowak; dies wiederum weit auslegend *Generalanwalt Spielmann*, Schlussanträge C-413/23 P, Rn. 33 f.; *Klabunde/Horváth*, in: Ehmann/Selmayr, Art. 4 Rn. 10 f.

18 *Artikel-29-Datenschutzgruppe*, WP 136, S. 11 ff.

19 Die Bedenken des fehlenden Informationsgehalts von in einem Sprachmodell „gespeicherten" Informationen nach *HmbBfDI*, Diskussionspapier: Large Language Models und personenbezogene Daten, S. 6 f. könnten allerdings dahingehend verstanden werden, dass den in einem Sprachmodell gespeicherten Informationen ein Informationsgehalt fehlt und daher das Vorliegen personenbezogener Daten abzulehnen wäre.

b. Identifizierbarkeit

Eng zusammenhängend mit dem Bezug der Information zu einer natürlichen Person muss diese Person auch identifiziert oder identifizierbar sein.[20] Soweit die Person nicht unmittelbar, z.B. über ihren Namen oder andere Kennungen, i.S.d. Art. 4 Nr. 1 DSGVO identifiziert werden kann, kommt es darauf an, ob die Person über weitere Mittel (z.B. Inferenzen und die Zuhilfenahme weiterer Quellen) identifizierbar ist. In diesem Zusammenhang sind insbesondere die Perspektive des jeweiligen Akteurs (relativer Ansatz)[21] und der aktuelle Stand der Technik zu berücksichtigen.

aa. Relatives Konzept der Identifizierbarkeit

Für die Frage der Identifizierbarkeit sind nach ErwGr. 26 S. 3 DSGVO „alle Mittel [zu berücksichtigen], die von dem Verantwortlichen oder einer anderen Person nach *allgemeinem Ermessen wahrscheinlich* genutzt werden, um die natürliche Person direkt oder indirekt zu identifizieren, wie beispielsweise das Aussondern" (Hervorhebung d. d. Verf.). In diesem Sinne legt der EuGH einen relativen Ansatz der Identifizierbarkeit zugrunde[22] und prüft mit Blick auf einen Verantwortlichen, ob dieser mit dem ihm nach vernünftiger Betrachtung[23] zur Verfügung stehenden Mitteln die natürliche

20 Zur Diff. zwischen den Merkmalen „direkt" und „indirekt" *Artikel-29-Datenschutzgruppe*, WP 136, S. 15.

21 Im Überblick zum objektiven bzw. absoluten und dem relativen Ansatz *Keppeler*, CR 2016, 360 (361) m.w.N.; s. auch *Bergt*, ZD 2015, 365.

22 Explizit(er) *Generalanwalt Spielmann*, Schlussanträge C-413/23 P, Rn. 70-74; s. auch *EDPB*, Verbindlicher Beschl. 1/2021, Rn. 147 f.; den EuGH ebenfalls wie hier verstehend *Hüger*, ZfDR 2024, 263 (278 f.); *Paal*, ZfDR 2024, 129 (134 ff.); *Seidel*, DSB 2023, 212; herrschend nach *Karg*, in: Simitis/Hornung/Spiecker gen. Döhmann, Art. 4 Nr. 1 DSGVO Rn. 60 f.; noch offen nach *Assion*, NJW 2023, 2619-2624 (Rn. 8 ff.); offen, aber mit Tendenz zum relativen Personenbezug nach *Klar/Kühling*, in: Kühling/Buchner, Art. 4 Nr. 1 DSGVO Rn. 26; diff. *Franzen*, in: Franzen/Gallner/Oetker, Art. 4 DSGVO Rn. 5; relativ mit Nähe zum absoluten Personenbezug nach *Keppeler/Poncza/Wölke*, CR 2024, 18 (Rn. 7); ähnl. *Moos/Rothkegel*, MMR 2016, 845 (845); *Ziegenhorn*, NVwZ 2017, 213 (217); *Eckhardt*, CR 2016, 786; *Kühling/Klar*, ZD 2017, 27 (28 ff.); dazu steht auch nicht *EDPB*, Guidelines 01/2025 on Pseudonymisation, Rn. 22 in Widerspruch, der betont, dass auch im Fall der Löschung von Informationen zur Auflösung einer Pseudonymisierung zu prüfen ist, ob gegebenenfalls andere Mittel zur Identifizierung zur Verfügung stehen.

23 *EuGH*, ZD 2024, 173 (Rn. 49) – FIN; entsprechend dem Wortlaut der DS-RL *Keppeler/Poncza/Wölke*, CR 2024, 18 (Rn. 13).

Person identifizieren kann.[24] Hierbei kommt es auf einen Willen zur Identifizierung jedenfalls nicht an für solche Quellen, die sich wie Kennungen innerhalb der Sphäre des Verantwortliche befinden und seinem unmittelbaren Zugriff unterliegen.[25] Aus der Möglichkeit zur Identifizierung (und der Anwendbarkeit des Datenschutzrechts) erwächst allerdings nicht die Pflicht, diese Identifizierung tatsächlich durchzuführen, sondern vielmehr bloß eine Pflicht, die Einhaltung der DSGVO zu gewährleisten (Art. 11 Abs. 1 DSGVO).[26]

Aus der jüngeren Rechtsprechung des EuGH im Zusammenhang mit Art. 9 DSGVO lässt sich ableiten, dass eine Identifizierbarkeit nicht die eindeutige Zuordnung zu einer Person erfordert. Im Fall der Bestellung eines Medikaments auf der Website einer Online-Apotheke kommt es nach dem EuGH nicht darauf an, „ob diese Informationen den Nutzer oder eine andere Person betreffen, für die diese Bestellung getätigt wird".[27] Nach dieser Lesart dürfte regelmäßig auch eine wahrscheinliche Zuordnung zu einer Person aus Sicht des Verantwortlichen genügen.[28] Diese Entscheidung des EuGH ist dogmatisch allerdings problematisch, denn eine Wahrscheinlichkeit ist richtigerweise im Rahmen der identifizierenden Mittel zu verorten.

Aus der Sicht eines Verantwortlichen stellt sich nach Maßgabe dieser EuGH-Judikatur die Weitergabe durch ihn pseudonymisierter Daten noch als Übermittlung personenbezogener Daten an einen Empfänger dar,[29] während der Empfänger gegebenenfalls[30] mangels Zugriffs auf die Zuordnungsschlüssel keine personenbezogenen Daten, sondern anonyme Daten erhebt. Diese kleinschrittige, feingranulare Betrachtung der in Rede stehenden Verarbeitungsschritte lag schon der EuGH-Entscheidung in der Rs. Fashion ID zugrunde.[31]

---

24 *EuGH*, NJW 2016, 3579 (Rn. 48 f.) – Breyer; ZD 2024, 173 (Rn. 46) – FIN; s. auch *EuG*, ZD 2023, 399 (Rn. 100); zust. *Baumgartner*, ZD 2023, 402 (402); dies nach der Rspr. für weniger eindeutig haltend *Halim/Marosi*, ZD 2024, 333 (334)

25 *Arning/Forgó/Krügel*, DuD 2006, 700 (703); *Brink/Eckhardt*, ZD 2015, 205 (211).

26 *Hornung/B. Wagner*, CR 2019, 565 (Rn. 34).

27 *EuGH*, NJW 2025, 33 (Rn. 88).

28 Zuvor noch die Singularisierung und eindeutige Zuordnung erfordernd *Karg*, in: Simitis/Hornung/Spiecker gen. Döhmann, Art. 4 Nr. 1 DSGVO Rn. 50; ähnl. auch *Moos*, CR 2024, 442 (Rn. 18); zur Singularisierung *Artikel-29-Datenschutzgruppe*, WP 136, S. 16; *EDPB*, Verbindlicher Beschl. 1/2021, Rn. 146 f.

29 *Generalanwalt Spielmann*, Schlussanträge C-413/23 P, Rn. 74-77; anders zuvor *EuG*, ZD 2023, 399 (Rn. 105).

30 S. insoweit zur Wahrscheinlichkeit nach allgemeinem Ermessen *EDPB*, Guidelines 01/2025 on Pseudonymisation, Rn. 22.

31 *EuGH*, ZD 2019, 455 (Rn. 76 ff.) – Fashion ID.

Das relative Konzept der Identifizierbarkeit lässt Raum für die in Art. 4 Nr. 5 DSGVO definierte und in Art. 6 Abs. 4 lit. e, Art. 25 Abs. 1, Art. 32 Abs. 1 lit. a, Art. 40 Abs. 2 lit. d, Art. 89 Abs. 1 S. 3 DSGVO referenzierte Pseudonymisierung von personenbezogenen Daten (siehe auch ErwGr. 26-29 DSGVO). Im Rahmen der Pseudonymisierung werden die Informationen und Kennungen nebst anderer identifizierender Mittel getrennt.[32] Die DSGVO findet zwar grundsätzlich Anwendung (vgl. ErwGr. 26 S. 2 DSGVO),[33] honoriert aber diesen erhöhten Schutz vor der Identifizierung als technische und organisatorische Maßnahme. Für Personen, die nur Zugriff auf die Informationen, nicht aber auf die identifizierenden Mittel haben, stellen sich die Informationen unter Zugrundelegung des relativen Konzepts der Identifizierbarkeit als anonym dar.[34] Diese Personen unterfallen für den Umgang mit den Informationen daher grundsätzlich nicht den Vorgaben der DSGVO, können aber im Einzelfall aufgrund ihres Einflusses auf die Verarbeitung und der Zusammenarbeit mit einem Verantwortlichen mit Datenzugriff als gemeinsam Verantwortliche nach Art. 4 Nr. 7, Art. 26 DSGVO anzusehen sein.[35]

bb. Anforderungen an die heranzuziehenden Mittel

Im Rahmen der Identifizierung sind nur solche Mittel zu berücksichtigen, die nach „allgemeinem Ermessen wahrscheinlich genutzt werden" (ErwGr. 26 S. 3 DSGVO).[36] Hierfür sind alle „objektiven Faktoren, wie die Kosten der Identifizierung und der dafür erforderliche Zeitaufwand [heranzuziehen], wobei die zum Zeitpunkt der Verarbeitung verfügbare Technologie und technologische Entwicklungen zu berücksichtigen sind" (ErwGr. 26 S. 4 DSGVO). Aus der beispielhaften Nennung („wie") folgt,

32 Hierzu etwa im Wissenschaftskontext *Volodina et al.*, Grandma Karl is 27 years old.
33 *EuGH*, ZD 2024, 209 (Rn. 58); *Generalanwalt Spielmann*, Schlussanträge C-413/23 P, Rn. 51.
34 Vgl. *EuG*, ZD 2023, 399 (Rn. 94).
35 *EuGH*, ZD 2024, 328 (Rn. 58 f.) – TC-String. S. auch nachfolgend unter D.V.3.a.
36 Die geringfügigen Änderungen im Wortlaut ggü. der DSRL sind ohne materielle Auswirkungen, s. *Hubert*, CR 2025, 77 (Rn. 14 f.); *Keppeler/Poncza/Wölke*, CR 2024, 18 (Rn. 13); die rein hypothetische, nicht wahrscheinliche Identifizierung genügt nicht, s. *Artikel-29-Datenschutzgruppe*, WP 136, S. 17.

dass darüber hinaus auch weitere Faktoren Berücksichtigung finden können.[37]

Der EuGH leitet aus ErwGr. 26 DSGVO ab, dass der Einsatz der Mittel nicht mit einem unverhältnismäßigen Aufwand einhergehen darf und rechtlich zulässig sein muss.[38] Ein etwaig bestehendes Restrisiko einer möglichen, aber unwahrscheinlichen Anonymisierung wird insoweit durch den EuGH hingenommen.[39] Die bloß fehlende Verwendungs- bzw. Zusammenführungsabsicht nach einer Pseudonymisierung eines Verantwortlichen soll nach dem EuGH für die Annahme einer Identifizierbarkeit nicht genügen,[40] da andernfalls die Pseudonymisierung der Anonymisierung faktisch gleichstehen würde. Ein solcher Befund stünde im Widerspruch zu den anfangs aufgezählten Inbezugnahmen auf die Pseudonymisierung innerhalb der DSGVO, die gerade die gesetzgeberische Intention einer eigenständigen Behandlung von pseudonymisierten Daten belegen.

Das Merkmal der rechtlichen Zulässigkeit wird in der Literatur zum Teil in Zweifel gezogen,[41] wobei dessen Ablehnung insoweit mit geringfügigen Änderungen in ErwGr. 26 zwischen DSRL und DSGVO begründet wird (zuvor „vernünftigerweise eingesetzt", nun „nach allgemeinem Ermessen wahrscheinlich").[42] Diese Zweifel dürften allerdings im Ergebnis nicht durchgreifen, denn bei tatsächlicher, rechtswidriger Identifizierung im Einzelfall liegt weiterhin eine Verarbeitung vor; eine solche Verarbeitung ist also bloß abstrakt – d.h. z.B. mit Blick auf ein gesamtes Sprachmodell – zu verneinen.[43] Betroffene Personen bleiben daher auch bei Ablehnung eines Personenbezugs eines Sprachmodells geschützt: Dieser Schutz wird zum einen vermittelt durch die weitgehende Verhinderung des Personenbezugs im Rahmen der identifizierenden Mittel, zum anderen durch die Anwendbarkeit der DSGVO, sofern der Personenbezug im Einzelfall nicht verhindert werden kann. Allerdings dürfte der Vorbehalt der rechtlichen Zulässigkeit im Einklang mit der vor allem auf Art. 8 GRCh ausgerichteten Rechtsprechung des EuGH bereits abstrakt nicht eingreifen, wenn von dem

37 *Hubert*, CR 2025, 77 (Rn. 6); *EDPB*, Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, Rn. 41.
38 Vgl. *EuGH*, BeckRS 2024, 3655 (Rn. 51) – OLAF; s. auch schon *LG Berlin*, ZD 2013, 618.
39 *Hubert*, CR 2025, 77 (Rn. 10); vgl. auch *Roßnagel*, DuD 2024, 513 (514).
40 *Brink/Eckhardt*, ZD 2015, 205 (211).
41 *Hacker*, Law, Innovation and Technology 13 (2021), 257 (266 ff.); *Pesch/Böhme*, MMR 2023, 917 (920); zust. *Engeler/Rolfes*, ZD 2024, 423 (427).
42 *Krügel*, ZD 2017, 455 (459).
43 Ähnl. vor der EuGH-Entscheidung *Brink/Eckhardt*, ZD 2015, 205 (211).

rechtlichen Verbot keine effektive Abschreckungswirkung ausgeht.[44] Aus dem rechtlichen Verbot kann also richtigerweise nur eine widerlegbare Vermutung der Unwahrscheinlichkeit der Identifizierung folgen.

Für die Prüfung der Verhältnismäßigkeit sind aus ex-ante-Perspektive insbesondere das (objektive) Interesse an der Identifizierung mit dem Aufwand abzuwägen.[45] Denn selbst wenn und soweit die Zuordnung zu einer unzutreffend identifizierten Person der DSGVO unterfällt, kann eine solche Zuordnung durch einen Verantwortlichen nach allgemeinem Ermessen unwahrscheinlich und daher nicht zu berücksichtigen sein (z.B. über die Heranziehung unglaubwürdiger bzw. unglaubhafter Quellen).[46] Umgekehrt kann die Identifizierung besonders dann wahrscheinlich sein, wenn sich eine Information an ein Publikum richtet, das nach allgemeinem Ermessen weitere Quellen heranziehen kann und typischerweise auch heranziehen wird (z.B. Journalisten).[47]

Es genügt für die Identifizierbarkeit aus Sicht der Verantwortlichen, wenn der Verantwortliche im Rahmen der heranzuziehenden Mittel auf weitere Personen[48] oder Quellen (z.B. eine – gegebenenfalls KI-gestützte – Internetsuche)[49] zurückgreifen kann.[50] Hiervon erfasst sind etwa Konstellationen, in denen die andere Person um identifizierende Mittel ersucht werden kann und (gegebenenfalls vertraglich) zur Beantwortung der ersuchenden Anfrage verpflichtet ist.[51] Insoweit kann es nach dem EuGH sogar ausreichen, wenn rechtliche Mittel zwar zur Verfügung stehen, von diesen

44 *Kühling/Klar*, ZD 2017, 27 (28 ff.); *Purtova*, Law, Innovation and Technology 10 (2018), 40 (64 f.); ähnl. auch *Karg*, in: Simitis/Hornung/Spiecker gen. Döhmann, Art. 4 Nr. 1 DSGVO Rn. 31; *Schwartmann et al.*, Praxisleitfaden zum Anonymisieren personenbezogener Daten, S. 18.

45 Ähnl. *Hubert*, CR 2025, 77 (Rn. 20); *Artikel-29-Datenschutzgruppe*, WP 136, S. 18; *LG Berlin*, ZD 2013, 618 (622 f.); *Kroschwald*, ZD 2014, 75 (76); *Schild*, in: BeckOK Datenschutzrecht, Art. 4 DSGVO Rn. 18; *Klar/Kühling*, in: Kühling/Buchner, Art. 4 Nr. 1 DSGVO Rn. 22; insoweit kann gegebenenfalls auch die Sensibilität von Datenkategorien (und damit auch der potenzielle Schaden für betroffene Personen) berücksichtigt werden, vgl. *Roßnagel*, DuD 2024, 513 (513).

46 *Hubert*, CR 2025, 77 (Rn. 25 f.).

47 *EuGH*, BeckRS 2024, 3655 (Rn. 63) – OLAF.

48 *EuGH*, NJW 2016, 3579 (Rn. 43) – Breyer; ZD 2024, 328 (Rn. 40, 46 ff.) – TC-String; nach *Hubert*, CR 2025, 77 (Rn. 18) sollen auch die individuellen Fähigkeiten des Verantwortlichen mit Blick auf die Mittel berücksichtigt werden.

49 *Hubert*, CR 2025, 77 (Rn. 20 f.).

50 Vgl. *EuGH*, BeckRS 2024, 3655 (Rn. 55 ff.) – OLAF; s. auch *DSK*, Orientierungshilfe Künstliche Intelligenz und Datenschutz, Rn. 4.

51 *EuGH*, ZD 2024, 328 (Rn. 48) – TC-String.

rechtlichen Mitteln aber nur im Ausnahmefall (z.B. für die Auflösung einer IP-Adresse im Zusammenhang mit einer Straftat) Gebrauch gemacht werden kann.[52] Allerdings unterliegt auch der Zugriff auf Dritte zur Identifizierung wie auch die übrigen Mittel dem Vorbehalt der Verhältnismäßigkeit[53] und dürfte entsprechend ErwGr. 26 S. 3, 4 DSGVO eine hinreichende Wahrscheinlichkeit voraussetzen.[54] Die einem Verantwortlichen zugewiesene (öffentlich-rechtliche) Aufgabe ist insoweit zu berücksichtigen (z.B. dürfte eine Identifizierung bei justizinterner Verwendung zur Verbesserung der Arbeitsabläufe unwahrscheinlich sein).[55]

cc. Mittelbarer Personenbezug oder Identifizierbarkeit

In einer Entscheidung betreffend die Einordnung einer Fahrzeugidentifikationsnummer (FIN) gelangt der EuGH zu dem Ergebnis, dass die FIN für den Fahrzeughersteller „mittelbar" ein personenbezogenes Datum darstellen kann, auch wenn der Fahrzeughersteller weder über die notwendigen Mittel zur Identifizierung verfügt noch auf die betreffenden Mittel über andere Wirtschaftsakteure zugreifen kann.[56] Diese Aussage könnte als eine Abkehr von einem relativen Konzept der Identifizierbarkeit verstanden werden, gegebenenfalls indem die spätere Identifizierbarkeit durch einen anderen Verantwortlichen auf einen diese Daten übermittelnden Akteur zurückwirkt, sodass dieser Akteur rückwirkend als Verantwortlicher anzusehen ist.[57]

Die englische Sprachfassung des Urteils („indirectly […] personal data") legt nahe, dass der EuGH sich hier auf Art. 4 Nr. 1 DSGVO beziehen könnte, wonach die direkte oder indirekte Identifizierbarkeit genügt.[58] Allerdings ist die in Art. 4 Nr. 1 DSGVO angelegte Identifizierbarkeit nicht pauschal

52 *EuGH*, NJW 2016, 3579 (Rn. 47) – Breyer; *Hubert*, CR 2025, 77 (Rn. 13); wohl eingeschränkt in der jüngeren Rspr., hierzu *Keppeler/Poncza/Wölke*, CR 2024, 18 (Rn. 15); zu dem Hintergrund der Breyer-Entscheidung auch *Keppeler*, CR 2016, 360 (362 f.); *H. Richter*, EuZW 2016, 912 (913).
53 *Hubert*, CR 2025, 77 (Rn. 12).
54 Krit. insoweit zur EuGH-Entscheidung *Kühling/Klar*, ZD 2017, 27 (28 ff.).
55 *Roßnagel*, DuD 2024, 513 (514).
56 *EuGH*, ZD 2024, 173 (Rn. 49) – FIN.
57 *Hanloser*, ZD 2024, 175 (176); krit. auch mit Blick auf die Unklarheit dieses Konzepts *Keppeler/Poncza/Wölke*, CR 2024, 18 (Rn. 21).
58 S. auch *EuGH*, ZD 2024, 173 (Rn. 45) – FIN; ähnl. *Keppeler/Poncza/Wölke*, CR 2024, 18 (Rn. 19).

gleichzusetzen mit dem vom EuGH verwendeten Begriff der mittelbar personenbezogenen Daten.

Der Gerichtshof nimmt in seinen einschlägigen Ausführungen Bezug auf zwei Ausführungen des EuGH-Generalanwalts Sánchez-Bordona (in der Rs. C-319/22), die bei dem Verständnis dieses Merkmals mit Blick auf den Fahrzeughersteller in diesem Zusammenhang weiterhelfen können: Zum einen wird die Aussage in Bezug genommen, dass die FIN nicht stets überhaupt einer Person zugeordnet werden kann (z.B. wenn das Kraftfahrzeug „keiner natürlichen Person gehört").[59] Zum anderen stellt der Generalanwalt auf die für die Wirtschaftsakteure verfügbaren Mittel ab und hält fest, dass in dem Fall verfügbarer Mittel zur Identifizierung „die FIN für sie (und mittelbar für den Hersteller, der ihnen die FIN zur Verfügung stellt) personenbezogene Daten [wären], deren Verarbeitung der DSGVO unterliegt".[60] Die letztere Aussage erschließt sich nicht ohne Weiteres, dürfte aber mit Blick auf die erste Aussage und den konkret zu beurteilenden Sachverhalt so verstanden werden, dass der Hersteller gegebenenfalls unter Zuhilfenahme der Wirtschaftsakteure selbst natürliche Personen anhand der FIN identifizieren kann. Unter Berücksichtigung der Schlussanträge eines anderen Generalanwalts (i.e. Generalanwalt Spielmann in der Rs. C-413/23 P) könnte diese Aussage jedoch auch dahingehend eingeordnet werden,[61] dass für den Fahrzeughersteller Informationspflichten nach Art. 13, 14 DSGVO eingreifen, wenn der Fahrzeughersteller – für ihn anonyme – Informationen an eine Person oder Stelle weitergibt (sog. Empfänger, Art. 4 Nr. 9 DSGVO), die die Informationen einer natürlichen Person zuordnen kann.[62]

Trotz verbleibender Ungewissheit ist mithin davon auszugehen, dass der EuGH grundsätzlich an einem relativen Konzept der Identifizierbarkeit festhält.

59 *Generalanwalt Sánchez-Bordona*, Schlussanträge C-319/22, Rn. 34.

60 *Generalanwalt Sánchez-Bordona*, Schlussanträge C-319/22, Rn. 41.

61 *Generalanwalt Spielmann*, Schlussanträge C-413/23 P, Rn. 65-75, der die umgekehrte Konstellation der Übermittlung an eine Person, die nicht über die identifizierenden Mittel verfügt, in den Blick nimmt.

62 Hierzu schon zuvor *Gola*, in: Gola/Heckmann, Art. 4 DSGVO Rn. 25; *Mantz/Spittka*, in: Sassenberg/Faber, § 6 Datenschutz und IT-Sicherheit, Rn. 21.

dd. Relatives Konzept der Identifizierbarkeit in zeitlicher Hinsicht

In zeitlicher Hinsicht ist die Anonymisierung, abhängig von dem konkret gewählten Verfahren, ebenfalls relativ und nicht absolut zu bewerten.[63] Durch den fortschreitenden Stand der Technik kann eine zuvor anonymisierte Information zu einem personenbezogenen Datum werden,[64] falls etwa neue Verfahren zur (Wieder-)Herstellung des Personenbezugs zur Verfügung stehen oder sich die Kosten zum Einsatz bestehender Verfahren signifikant reduzieren.[65] Ferner können (potenzielle) Verantwortliche im Laufe der Zeit den Zugriff erhalten auf ergänzende Datenquellen oder auf Dritte mit Zusatzwissen.[66] Eine solche ergänzende Datenquelle kann gegebenenfalls auch durch die betroffene Person selbst eröffnet werden, indem Betroffenenrechte ausgeübt und durch Zusatzinformationen eine Zuordnung der bei dem Verantwortlichen bereits gespeicherten Informationen zu der betroffenen Person ermöglicht werden.[67]

Die vorliegende Untersuchung legt ein weites Begriffsverständnis der Anonymisierung zugrunde; dies gilt unabhängig davon, ob einmal anonymisierte Daten zu einem späteren Zeitpunkt erneut einer natürlichen Person zugeordnet werden können.[68]

c. Zwischenergebnis

Für die Frage, ob personenbezogene Daten nach der DSGVO vorliegen, kommt es sowohl darauf an, ob Informationen Aussagen zu einer natürlichen Person zu entnehmen sind, als auch darauf, ob die Informationen aus Sicht eines Verantwortlichen mit den ihm zur Verfügung stehenden Mitteln tatsächlich auch dieser natürlichen Person zuzuordnen sind. Soweit und solange (*sic!*) diese Identifizierbarkeit ausgeschlossen ist, handelt es sich um anonyme Daten. Diese Annahme wird im Folgenden zugrunde gelegt, womit zugleich auch ein Zielzustand definiert ist.

63 *Hornung/B. Wagner*, CR 2019, 565 (Rn. 3); *Roßnagel*, DuD 2024, 513 (514).
64 *Purtova*, Law, Innovation and Technology 10 (2018), 40; vgl. auch *Artikel-29-Datenschutzgruppe*, WP 216, S. 7; *Krügel*, ZD 2017, 455 (456).
65 *Hornung/B. Wagner*, CR 2019, 565 (Rn. 14).
66 *Hornung/B. Wagner*, CR 2019, 565 (Rn. 10).
67 *Hornung/B. Wagner*, CR 2019, 565 (Rn. 11).
68 S. zur Unumkehrbarkeit und dem Restrisiko *Artikel-29-Datenschutzgruppe*, WP 216, S. 6 f.

2. Überblick zu den Anonymisierungsverfahren

Zur Erreichung dieses – vor allem relativen[69] – Anonymisierungsziels kommen verschiedene Techniken in Betracht, die jeweils eine Verarbeitung personenbezogener Daten erfordern.[70] Diese Techniken werden im Folgenden überblicksartig und stark vereinfacht dargestellt, soweit ihnen für den dieser Untersuchung zugrunde liegenden Sachverhalt eine Relevanz zukommt.

a. Löschung von Merkmalen zur Identifizierung durch Auslassung oder Ersetzung

In Betracht kommen zunächst die Auslassung oder Ersetzung[71] als Löschung von Kennungen und anderen identifizierenden Merkmalen[72] (oder auch: Neutralisierung),[73] wie sie beispielsweise im Rahmen der Anonymisierung von Gerichtsentscheidungen im Wege der Veröffentlichung praktiziert werden (siehe dazu nachfolgend unter D.II). Hierbei werden beispielsweise der Name einer natürlichen Person, E-Mail-Adressen, Internetdomain-Bezeichnungen, Telefonnummern, Kontodaten und sonstige Identifikationsnummern entfernt. Ohne diese Kennungen und sonstigen Merkmale können grundsätzlich auch die übrigen Informationen (z.B. eine getätigte Aussage) nicht mehr einer natürlichen Person zugeordnet werden.

Der Anonymisierungserfolg dieses Verfahrens hängt somit maßgeblich davon ab, ob tatsächlich alle Merkmale identifiziert und entfernt werden, die für sich genommen oder gegebenenfalls mit anderen Merkmalen mit hinreichender Wahrscheinlichkeit[74] eine Identifizierung der natürlichen Person ermöglichen (z.B. ein nicht entfernter Straßenname samt Hausnummer, der aber eindeutig einer Stadt und sodann einem Immobilieneigentümer zugeordnet werden kann). Insoweit kommt es auch darauf an, welche zusätzlichen öffentlichen oder nicht-öffentlichen Quellen zum Zwecke der Identifizierung herangezogen werden können. Insgesamt wird vielfach nicht ausgeschlossen werden können, dass in der Zukunft weitere Quellen

69 *Schwartmann et al.*, Praxisleitfaden zum Anonymisieren personenbezogener Daten, S. 3.
70 S. zur Rechtsgrundlage unter D.V.4.c.aa.
71 *Deuber/Keuchen*, MMR 2023, 338 (342); vgl. auch die Schwärzung *Determann/Paal*, KI-Recht international, S. 57.
72 *Klar/Kühling*, in: Kühling/Buchner, Art. 4 Nr. 1 DSGVO Rn. 34.
73 *BVerwG*, NJW 1997, 2694 (2695).
74 S. zur Inferenz *Artikel-29-Datenschutzgruppe*, WP 216, S. 11.

zur Verfügung stehen, die z.B. eine Anschrift leichter auf eine Person zurückführen lassen.

b. Weitere Anonymisierungstechniken und Maßnahmen i.w.S.

Die Artikel-29-Datenschutzgruppe differenziert im Weiteren zwischen Techniken der Randomisierung als Entfernen der Verbindung zwischen Person und Datum (z.B. Vertauschen von Datenzuordnungen, stochastische Überlagerung, Differential Privacy) und der Generalisierung als Abstrahierung einzelner Worte (z.B. Aggregation durch Entfernung von Straßennahmen, k-Anonymität, l-Diversität und t-Closeness).[75] Diese Techniken eignen sich besonders für die Anwendung auf strukturierte Datensätze. Für die im GSJ-Projekt bearbeiteten Gerichtsentscheidungen und Aktenauszüge liegt die Herausforderung demgegenüber aber in der potenziellen Identifizierung der Merkmale. Die bloße Löschung oder Pseudonymisierung dieser Merkmale schadet der Verständlichkeit nicht, sodass es regelmäßig keiner besonderen Anonymisierungsverfahren über eine Identifizierung und Löschung von Kennungen und sonstigen identifizierenden Merkmalen hinaus bedarf.

Neben den eigentlichen Anonymisierungstechniken kommen zudem auch Maßnahmen in Betracht, die primär an den identifizierenden Mitteln ansetzen und hierdurch zum gleichen Ergebnis gelangen. Der Rückgriff auf Quellen kann gesetzlich eingeschränkt oder einzelne Personen können von diesen Quellen abgeschottet werden.[76] Zu denken ist ferner an eine Selbstverpflichtung des Verantwortlichen, gegebenenfalls als Verhaltensregel i.S.d. Art. 40 DSGVO, sich des Rückgriffs auf verfügbare Mittel zur Identifizierung zu enthalten.[77]

75 *Artikel-29-Datenschutzgruppe*, WP 216, S. 14 ff.; *Dewitte*, in: AI Meets the GDPR, S. 140 f.; *Schwartmann et al.*, Praxisleitfaden zum Anonymisieren personenbezogener Daten, S. 6; *Winter/Battis/Halvani*, ZD 2019, 489 (490 f.); *International Working Group on Data Protection in Technology*, Working Paper on Large Language Models (LLMs), S. 51 ff.; s. auch zur Pseudonymisierung *EDPB*, Guidelines 01/2025 on Pseudonymisation.

76 Zu Letzterem *Hornung/B. Wagner*, CR 2019, 565 (Rn. 42).

77 *Hornung/B. Wagner*, CR 2019, 565 (Rn. 49).

3. Anforderungen an die Anonymisierung unveröffentlichter Gerichtsentscheidungen

Die ermittelten Maßstäbe sind nunmehr auf die Anonymisierung unveröffentlichter Gerichtsentscheidungen und die darin enthaltenen Kennungen sowie indirekten Identifikationsmerkmale zu übertragen.[78]

a. Typische identifizierende Merkmale

Hierzu bedarf es zunächst der Betrachtung einer typischen Gerichtsentscheidung.

Eine Gerichtsentscheidung der Zivilgerichtsbarkeit enthält mit Blick auf Kläger, Beklagte, Prozessbevollmächtigte, sonst im Rubrum genannte Personen (z.B. Nebenintervenienten), Richter und sonstige Beteiligte (z.B. Zeugen) verschiedene Informationen, die sich auf diese Personen beziehen können.

Maßgeblich für die Identifizierbarkeit dieser natürlichen Personen werden insbesondere die folgenden Merkmale sein:

Rubrum	• Vor- und Nachname einer (natürlichen) Person, gegebenenfalls als Vertreter einer juristischen Person • Anrede der Person • Private Wohnanschrift oder gegebenenfalls Anschrift des Arbeitgebers bzw. der Partei • Aktenzeichen (im Zusammenhang mit gegebenenfalls öffentlich einsehbaren Sitzungsplänen) • Nachname des Richters oder der Richterin sowie des Urkundsbeamten oder der Urkundsbeamtin • Ort des Gerichts
Tatbestand und Entscheidungsgründe	• Nach- und gegebenenfalls Vorname einer Person • Ortsangaben • Datumsangaben (im Zusammenhang mit weiteren Informationen, wie etwa dem Ort) • Angaben zu Kommunikationsmitteln (z.B. Telefonnummer oder E-Mail-Adresse) • Angaben zu Zahlungsmitteln (z.B. Kreditkartennummern, Bankverbindungen) • Eigentums- oder vergleichbare Positionen (z.B. Marken, Domains) • Angaben in der Sache (z.B. getätigte Aussagen, Vereins- oder Gesellschaftszugehörigkeit, Geschehen des öffentli-

78 Zum Begriff *Schwartmann et al.*, Praxisleitfaden zum Anonymisieren personenbezogener Daten, S. 10.

	chen Lebens im Zusammenhang mit öffentlich zugänglicher Berichterstattung) • Identifikationsnummern (z.B. Fahrzeugidentifikationsnummer (FIN), persönliche Steueridentifikationsnummer, Sozialversicherungsnummer)
Unterschriften	• Nach- und gegebenenfalls Vorname einer Person

Für Entscheidungen anderer Gerichtsbarkeiten gelten die vorstehenden Ausführungen und Überlegungen entsprechend.

b. Maßstab für Anonymisierungsverfahren

Die zentrale Herausforderung der Anonymisierung von Gerichtsentscheidungen liegt in der (fehlenden) Struktur[79] der Dokumente, wodurch eine Identifizierung der zu anonymisierenden Informationen erheblich erschwert wird. Während Datenbanken nach einzelnen Attributen aufgegliedert sind, enthalten die Gerichtsentscheidungen jeweils einen Text. Die in dem Text enthaltenen Wörter können allein oder in Verbindung mit anderen Wörtern ein solches Attribut darstellen. Besondere Aufmerksamkeit ist daher zu richten auf die Ermittlung eines identifizierenden Personenbezugs unter Berücksichtigung der voraussichtlich verfügbaren (öffentlichen) Quellen im Zusammenhang mit der Anonymisierung von Gerichtsentscheidungen.

aa. Berücksichtigung von weiteren Mitteln zur Identifizierung

Das Gelingen der Anonymisierung hängt maßgeblich davon ab, welche Mittel die Verantwortlichen in rechtmäßiger Weise und mit verhältnismäßigem Aufwand heranziehen könnten, um eine natürliche Person zu identifizieren.

Solche Mittel zur Identifizierung umfassen insbesondere Listen der Orts- und Straßenbezeichnungen,[80] Kalender, identifizierende und nichtidentifizierende mediale Berichterstattung, Geschäftsverteilungspläne sowie weitere Informationsquellen. Dies gilt unabhängig davon, ob diese Informationsquellen über das Internet, über den Handel oder auf individuelle Nachfrage verfügbar sind.

79 Vgl. den Begriff der strukturierten Daten als Schlüssel-Wert-Paar *Schwartmann et al.*, Praxisleitfaden zum Anonymisieren personenbezogener Daten, S. 3.

80 *Deuber/Keuchen*, MMR 2023, 338 (340).

Für einzelne Personen können zudem amtliche Auskünfte (z.B. des Einwohnermeldeamts) und Registerauszüge (z.B. aus dem Handelsregister) heranzuziehen sein. Darüber hinaus kommt in Betracht, dass einzelne Personen, die über ein Insiderwissen mit Blick auf den Prozess verfügen, durch dieses Wissen für weitere Informationen aus der Gerichtsentscheidung einen Personenbezug herstellen können (z.B. Personen, denen Parteien von einem Gerichtsprozess berichtet haben). Vor diesem Hintergrund ist eine (absolute)[81] Anonymisierung allen Personen gegenüber zwar nicht möglich, datenschutzrechtlich aber gerade auch nicht zwingend erforderlich.[82] Nachfolgend wird daher die Anonymisierung mit Ausnahme der Parteien mit Insiderwissen zu der jeweiligen Gerichtsentscheidung im Vordergrund stehen.

bb. Identifizierung der zu anonymisierenden Informationen

Für die Identifizierung der personenbezogenen Daten sind über die typischen identifizierenden Merkmale in Form von Kennungen hinaus sämtliche weiteren Wörter aus dem Text der Gerichtsentscheidung (unter (1)) bzw. typisierend die Entscheidung selbst (unter (2)) darauf zu untersuchen, ob sie mit hinreichend hoher Wahrscheinlichkeit die Identifizierung unter Berücksichtigung der öffentlich verfügbaren Informationsquellen ermöglichen.

(1) Untersuchung auf Wortebene bzw. Zeichenebene

Folgt man einem Ansatz der vollständigen Risikoeliminierung, so sind in einem ersten Schritt sämtliche Informationen als identifizierend zu betrachten und sodann in einem zweiten Schritt typischerweise nicht-identifizierende Wörter auszusondern (z.B. unbestimmte oder bestimmte Artikel als sog. Stoppwörter).[83] Ein solcher Ansatz führt allerdings zur Aussonderung zahlreicher Informationen (z.B. durch Aussonderung verschiedener Substantive und Prädikate) in einem Umfang, dass der Text für Zwecke des Trainings eines Sprachmodells nach einer Veränderung dieser Informationen regelmäßig nicht mehr hinreichend geeignet sein dürfte.

81 *Schwartmann et al.*, Praxisleitfaden zum Anonymisieren personenbezogener Daten, S. 17.
82 S. hierzu insb. unter D.II.5.b sowie unter D.V.4.c.
83 *Brockmeyer*, Text und Data Mining, S. 21.

Es bietet sich daher ein Rekurs auf das der DSGVO zugrundeliegende Konzept eines risikobasierten Ansatzes an, um eine Einordnung der Zeichenfolgen, Wörter, Wortkombinationen in verschiedene Risikokategorien vorzunehmen. Ein solcher Ansatz unter angemessener Berücksichtigung des Risikos einer Identifizierbarkeit differenziert beispielsweise zwischen:

1. mit hoher Wahrscheinlichkeit identifizierenden Wörtern (z.B. Ortsbezeichnungen, Telefonnummern, E-Mail-Adressen und spezifischen Berufsbezeichnungen);
2. potenziell identifizierenden Wörtern oder Zeichenfolgen (z.B. Datumsangaben, besonderen Verhaltensweisen und Eigenschaften einer Person, Angaben in Anführungszeichen, jeweils unter Berücksichtigung länder- und sprachspezifischer Besonderheiten);[84]
3. ohne vernünftigen Zweifel nicht mehr identifizierenden Wörtern (z.B. grundsätzlich identifizierungsneutrale (Stopp-)Wörter wie „zwischen", „durch", „das", „ein", „oder", „und").

(2) Untersuchung auf Entscheidungsebene – Cluster von Gerichtsentscheidungen

Gerichtsentscheidungen mit allenfalls potenziell identifizierenden Wörtern können durch eine Clusterung weitgehend ausgeschlossen werden bzw. für eine händische Anonymisierung vorsortiert werden. So empfiehlt sich in diesem Sinne beispielsweise die Filterung von Gerichtsentscheidungen mit einem besonders umfangreichen Tatbestand sowie anhand ausgewählter Spezialzuständigkeiten (z.B. einer Kammer für Arzthaftungssachen, vgl. § 348 Abs. 1 Nr. 2 lit. e ZPO) und der Anzahl der beteiligten Personen.

In ähnlicher Weise differenziert die Dokumentationsstelle des OLG Frankfurt a. M. und ordnet Gerichtsentscheidungen in vier Anonymisierungsstufen ein:[85]

1. Stufe A – geringer Anonymisierungsaufwand: Ersetzung von Eigennamen durch Kürzel oder Auslassung[86];
2. Stufe B – erhöhter Anonymisierungsaufwand: insbesondere in Familien- und Arzthaftungssachen: Ersetzung bzw. Auslassung auch von „Straßen-

84 S. das anschauliche Beispiel in einem Arzthaftungsfall bei *Nöhre*, MDR 2019, 136 (137); länderspezifische Besonderheiten zu berücksichtigen nach *BayLDA*, Datenschutzkonforme Künstliche Intelligenz, S. 5.

85 *Nöhre*, MDR 2019, 136 (138 ff.).

86 Hierzu etwa *Winter/Battis/Halvani*, ZD 2019, 489 (491) m.w.N.

und Ortsnamen, Hausnummern, Kfz-Kennzeichen, Telefonnummern, Konto- oder Kundennummern, E-Mail- und Internetadressen sowie Identifikationsnummern von Behörden“ sowie gegebenenfalls auch von „Staatsangehörigkeiten, Alters- und Datumsangaben sowie Aktenzeichen anderer Verfahren“;
3. Stufe C – brisant: zusätzliche Auslassung von Textpassagen, z.B. Untersuchungsbefunden;
4. Stufe D – keine Veröffentlichung: z.B. in Staatsschutzprozessen.

Die hierdurch erfolgende Differenzierung kann sich grundsätzlich für eine Vorfilterung eignen. Die Effektivität dieser Vorgehensweise hängt maßgeblich ab von der richtigen Stufenzuordnung und der sorgfältigen Anwendung von Anonymisierungstechniken im Einzelfall.

cc. Training eines Sprachmodells als mögliche Anonymisierungstechnik

Durch das Training eines Sprachmodells (insbesondere eines Large Language Models, LLM) kann die Identifizierbarkeit einer natürlichen Person womöglich ausgeschlossen oder zumindest erschwert werden.[87]

Im Rahmen des Trainings gehen durch die Tokenisierung und Neuberechnung von Informationen auf Grundlage der Trainingsdaten Informationen aus den Ausgangstexten verloren bzw. werden neu angeordnet. Einige zuvor eindeutige Verknüpfungen (z.B. „Herr Müller wohnt in der Rathausstr.“) gehen zugunsten einer wahrscheinlichen Beziehung der Buchstabenketten verloren (z.B. auf die Token „Herr“, „ Müller“, „ woh“, „nt“, „ in“, „ der“ folgt mit einer gewissen Wahrscheinlichkeit das Token „Rathaus“). Diese Wahrscheinlichkeit hängt von den übrigen Trainingsdaten ab, z.B. davon, ob das Token „ Müller“ auch in anderen Zusammenhängen in den Trainingsdaten verwendet wird. Ferner beeinflusst die Eingabeaufforderung an ein KI-System, ob die im Sprachmodell repräsentierten Token in der Ausgabe in Beziehung zueinander gesetzt werden. Denn diese Eingabeaufforderung wird durch das Sprachmodell als eine Sammlung von Token behandelt, deren wahrscheinliche Fortführung durch das Modell ausgegeben wird.

87 S. auch zu Anonymisierungstechniken im Zusammenhang mit KI-Modellen (Privacy Preserving Machine Learning) im rechtswissenschaftlichen Kontext *Winter/Battis/Halvani*, ZD 2019, 489 (492 f.); allg. *Al-Rubaie/J. Morris Chang*, Privacy Preserving Machine Learning: Threats and Solutions.

Nimmt man in einem Gedankenexperiment an, dass die zuvor genannten Token sich nur in exakt dieser Reihenfolge in den Trainingsdaten wiederfinden und in Auszügen im Rahmen der Eingabeaufforderung übergeben werden, wird theoretisch[88] stets der Straßenname ausgegeben. In Ansehung des Umfangs der Trainingsdaten eines LLM ist die genaue Ausgabe dieses identifizierenden Auszugs aus den Trainingsdaten zwar unwahrscheinlich, aber möglich.

Gerade aufgrund dieser Ungewissheit als eine Form der sog. Vertauschung[89] der Merkmale „Person" und „Information" bzw. Verwässerung der Verknüpfung kann das Training eines LLM als eine Anonymisierungstechnik im weiteren Sinne angesehen werden.[90] Das führt zwar nicht dazu, dass das Sprachmodell generell und stets keine personenbezogenen Daten enthält, die in dem Modell gespeichert sind.[91] Die Zuordnung von Informationen aus den Trainingsdaten zu der jeweiligen natürlichen Person wird aber jedenfalls deutlich unwahrscheinlicher. In Abhängigkeit von den in Rede stehenden Trainingsdaten, typischen Eingabeaufforderungen und den dem potenziell Verantwortlichen zur Verfügung stehenden Mitteln kann mit Blick auf einen potenziellen Verantwortlichen die Schwelle der vernünftigerweise bestehenden Identifizierungsmöglichkeit unterschritten werden und es können ihm gegenüber nicht-personenbezogene Daten vorliegen.

c. Nach allgemeinem Ermessen wahrscheinlich genutzte Mittel

Soweit der bearbeitete Entscheidungstext zwar keine identifizierenden Kennungen mehr enthält, aber natürliche Personen durch Hinzuziehung weiterer Quellen identifizierbar bleiben, kommt es darauf an, auf welche Mittel der potenziell Verantwortliche nach vernünftiger Betrachtung[92] wahrscheinlich zurückgreifen wird (vgl. auch ErwGr. 26 S. 3 DSGVO).[93]

88 Soweit hier Faktoren, wie etwa die Temperatur für eine gewisse Zufälligkeit als eine nachgeahmte Kreativität des KI-Systems, ausgeblendet werden.
89 *Artikel-29-Datenschutzgruppe*, WP 216, S. 16 f.
90 Ähnl. i.E. *Hüger*, ZfDR 2024, 263 (278).
91 So aber *HmbBfDI*, Diskussionspapier: Large Language Models und personenbezogene Daten.
92 *EuGH*, ZD 2024, 173 (Rn. 49) – FIN.
93 S. schon unter D.I.1.b.bb.

Das spätere Anwendungsszenario des Sprachmodells, der Verbleib und die Zugänglichkeit des Trainingsdatensatzes können insoweit von Bedeutung für die Anforderungen an die Anonymisierung der Gerichtsentscheidungen sein. Legt man beispielsweise ein Anwendungsszenario zugrunde, bei dem ausschließlich Personen aus der über Art. 1 Abs. 3, Art. 20 Abs. 3 GG besonders verpflichteten Richterschaft und Justizbedienstete einen Zugriff auf das Sprachmodell erhalten, erscheint es unwahrscheinlich(er), dass diese weiteren Mittel tatsächlich verwendet werden, um zufällige Bruchstücke aus den Trainingsdaten einer natürlichen Person zuzuordnen. Bei einer öffentlichen Zugänglichkeit des KI-Systems (z.B. in Form eines öffentlich verfügbaren Chatbots), gegebenenfalls auch des Sprachmodells einschließlich der Gewichte (z.B. als Open Source) sowie des Trainingskorpus werden eine Heranziehung weiterer Quellen und ein Missbrauchspotenzial demgegenüber deutlich wahrscheinlicher sein.[94]

d. Schlussfolgerungen für die Anwendung von Anonymisierungstechniken auf Gerichtsentscheidungen

Für die möglichst weitreichende Entfernung eines Personenbezugs sind ausgewählte Wortkategorien stets zu entfernen, auszutauschen oder mit anderen geeigneten Anonymisierungstechniken zu behandeln (z.B. Straßen- und Ortsbezeichnungen). Gerichtsentscheidungen mit Wortkategorien, die abhängig vom Einzelfall identifizierend wirken können, sind anhand von äußeren Merkmalen (z.B. dem Tatbestandsumfang, der Anzahl mitwirkender Personen oder dem Rechtsgebiet) gegebenenfalls gesondert zu behandeln. Insoweit kann aber bereits das Training eines Sprachmodells die Wahrscheinlichkeit einer Identifizierbarkeit erheblich reduzieren und hierdurch zu einer Anonymisierung beitragen.

Sämtliche Anonymisierungstechniken sind in regelmäßigen Abständen nach dem Stand der Technik zu überprüfen und gegebenenfalls anzupassen.

Für die Anforderungen an die Anonymisierung ist zudem darauf abzustellen, welcher Personenkreis einen Zugriff erhält auf die anonymisierten Gerichtsentscheidungen und gegebenenfalls das hierdurch trainierte Sprachmodell. Bei einer Verwendung innerhalb der Justiz werden insoweit zugunsten einer vereinfachten Annahme der Anonymisierung geringere

94 Ausf. nachfolgend unter D.IV.3.

Anforderungen an die Heranziehung zusätzlicher Quellen zu stellen sein als im Fall einer darüber hinausgehenden Veröffentlichung.

Falls eine Anonymisierung im Sinne der vorbenannten Maßgaben nicht erfolgreich ist und auch nicht aufgrund des konkreten Einsatzes des GSJ-Sprachmodells anzunehmen ist (hierzu nachfolgend unter D.IV.), findet die DSGVO Anwendung und es bedarf für die Datenverarbeitung insbesondere des Vorliegens einer Rechtsgrundlage nach Art. 6 Abs. 1 DSGVO sowie gegebenenfalls nach Art. 9, 10 DSGVO für die Verarbeitung besonderer Kategorien personenbezogener Daten.

II. Vergleich mit den Maßstäben zur Veröffentlichung von Gerichtsentscheidungen

Für die etablierte Anonymisierungspraxis der Rechtspflege betreffend die Veröffentlichung von Gerichtsentscheidungen ist zu untersuchen, ob diese Praxis den vorstehend herausgearbeiteten Vorgaben der DSGVO gerecht werden muss und wird.

1. Anwendbarkeit der Maßstäbe der DSGVO

Die DSGVO als Verordnung i.S.d. Art. 288 Abs. 2 AEUV genießt allgemeine Geltung unmittelbar in jedem Mitgliedstaat und Anwendungsvorrang vor konfligierenden nationalen Rechtsvorschriften (vgl. Art. 23 Abs. 1 GG).

In Ansehung der grundsätzlich vollharmonisierenden[95] Natur der DSGVO findet das DSGVO-Regelungsregime Anwendung auf sämtliche (teil-)automatisierten Vorgänge mit personenbezogenen Daten nach Art. 2, 4 Nr. 1, 2 DSGVO vorbehaltlich der territorialen Anforderungen nach Art. 3 DSGVO. Die DSGVO ist mithin zur Anwendung berufen betreffend die Verarbeitung von in Gerichtsentscheidungen enthaltenen personenbezogenen Daten. Dieser Befund gilt unabhängig davon, ob eine Verwaltungsvorschrift zur Anonymisierung einer Gerichtsentscheidung eingehalten und die Information veröffentlicht wurden[96] oder ob es sich um die Veröffentlichung von Dokumenten durch Behörden handelt. Denn die DSGVO

95 S. insb. ErwGr. 7 ff. DSGVO und die zahlreichen Öffnungsklauseln innerhalb der DSGVO.

96 Vgl. auch *International Working Group on Data Protection in Technology*, Working Paper on Large Language Models (LLMs), S. 43 f.

enthält gerade keine dem § 5 UrhG entsprechende Vorschrift zu amtlichen Werken, die Informationsanordnungen nach Durchlaufen eines vorgesehenen Verfahrens den Schutz unter dem Rechtsakt verwehrt.

2. Anonymisierungspraxis in Nordrhein-Westfalen bis zum Jahr 2021

Die Anonymisierungspraxis der Gerichte, hier anhand des Beispiels Nordrhein-Westfalens hat sich im Laufe der Jahrzehnte maßgeblich verändert. So stützte sich die Sicherheit der Anonymisierungsrichtlinien der NRW-Justiz aus dem Jahr 2002 – und wohl auch mit Blick auf die Richtlinie aus dem Jahr 2018 – vornehmlich auf die Geheimhaltung der Anonymisierungstechnik.[97] Sicherheit allein durch Geheimhaltung (z.B. im Hinblick auf eine Substitutionstabelle für Anfangsbuchstaben) ist allerdings kein anerkanntes Konzept nach dem Stand der Technik und wird in der Kryptoanalyse abgelehnt.[98] So konnte nachgewiesen werden, dass sich aufgrund öffentlicher Datenquellen Rückschlüsse auf Straßennamen ziehen lassen, die mit monoalphabetischer Substitution (vermeintlich) anonymisiert worden waren.[99]

3. Jüngere Anonymisierungspraxis in Nordrhein-Westfalen

Jedenfalls die öffentlich zugängliche Verwaltungsvorschrift der Justiz Nordrhein-Westfalen aus dem Jahr 2021 (nachfolgend: Verwaltungsvorschrift)[100] legt geeignetere Techniken der Anonymisierung zugrunde und wird im Folgenden untersucht. Allgemein gilt, dass die effektive Umsetzung mit zunehmendem Fortschritt von KI-basierten Anonymisierungsprojekten leichter fallen dürfte.[101]

Die Verwaltungsvorschrift geht zunächst aus von einer Abwägungsentscheidung gestützt auf nationale Rechtsgrundlagen i.S.d. Art. 6 Abs. 1 lit. c

97 *VG Düsseldorf*, BeckRS 2020, 34887 (Rn. 34); in Abgrenzung hierzu die Fassung von 2021, vgl. *OVG Düsseldorf*, BeckRS 2023, 2497 (Rn. 35).

98 Nur *Deuber/Keuchen*, MMR 2023, 338 (338 f.).

99 *Deuber/Keuchen*, MMR 2023, 338; s. auch zur möglichen Deanonymisierung *Deuber/Keuchen/Christin*, Assessing Anonymity Techniques Employed in German Court Decisions: A De-Anonymization Experiment, 2023.

100 Übermittlung von Entscheidungsabschriften an Dritte und Veröffentlichung in Datenbanken, RV d. JM vom 10. Mai 2021 (1552 - I. 12).

101 S. hierzu z.B. *Biallaß*, MMR 2024, 646 (649 f.) unter Verweis auf JANO und ALeKS.

DSGVO (so konkret etwa § 299 Abs. 2 ZPO, §§ 475, 480 Abs. 1 StPO (Ziff. 1.2.2. Verwaltungsvorschrift)). Zudem legt die Verwaltungsvorschrift in Ziff. 2.3 zugrunde, dass ein „Rückschluss auf die dahinterstehende Person auszuschließen" und gegebenenfalls die Möglichkeit der Identifizierung durch weitere Angaben im Entscheidungstext zu berücksichtigen ist (z.B. Lebensumstände der betroffenen Person und sensible ärztliche Untersuchungsbefunde).

Somit bleiben insbesondere die Namen von absoluten Personen der Zeitgeschichte, literarische Quellenangaben und Aktenzeichen gerichtlicher und staatsanwaltlicher Entscheidungen unberührt (Ziff. 2.4 lit. b, c, d Verwaltungsvorschrift).

Die folgenden Merkmale werden jeweils durch einen für den gesamten Entscheidungstext individuellen Buchstaben oder eine Kombination aus mehreren Buchstaben ersetzt:

- Namen von natürlichen oder juristischen Personen, Städten und Gemeinden, Straßen und Wegen, Grundstücksbezeichnungen (Ziff. 2.5, 2.6, 2.9, 2.10 Verwaltungsvorschrift);
- Telefonnummern, Konto- oder Kundennummern, E-Mail- und (persönliche) Internetadressen, Autokennzeichen sowie Identifikationsnummern von Behörden (z.B. E-Mail-Adresse01; Ziff. 2.8 Verwaltungsvorschrift).

Alters-, Berufs- und Datumsangaben werden entfernt bzw. durch die Ziffer „0" oder Ziffernkombinationen aus „0" ersetzt (Ziff. 2.7, 2.11, 2.12 Verwaltungsvorschrift). Gleiches gilt grundsätzlich für die Benennung der Personen, die an der Entscheidung mitgewirkt haben (z.B. Richter; Ziff. 2.14 Verwaltungsvorschrift).

4. Abgleich im Hinblick auf die Einhaltung der Anforderungen an die Anonymisierung unter der DSGVO

Die unter 3. dargestellte Anonymisierungspraxis der Gerichte ist grundsätzlich geeignet, (zunächst) eine Anonymisierung i.S.d. DSGVO herbeizuführen.

Nicht den Anforderungen an eine Anonymisierung unter der DSGVO im eigentlichen Sinne (unbeschadet der Rechtmäßigkeit[102]) entsprechen dürfte allerdings die Beibehaltung der Namen von absoluten Personen

102 S. unter D.V.4.

der Zeitgeschichte. Denn die DSGVO findet ausnahmslos Anwendung auf jede natürliche Person (Art. 4 Nr. 1 DSGVO) und schließt insoweit auch absolute Personen der Zeitgeschichte ein.

Im Übrigen hängt der Erfolg einer Anonymisierung ab von den im Einzelfall ergriffenen Maßnahmen. Ziff. 2.3 Verwaltungsvorschrift erkennt insoweit ausdrücklich an, dass die dort genannten Anonymisierungsmaßnahmen „in der Regel" und nicht in allen Fällen ausreichend sind. Ob und inwieweit eine Anonymisierung i.S.d. DSGVO erfolgt, hängt daher maßgeblich ab von der Anonymisierung der jeweiligen Entscheidung im Einzelfall.

Eine (erfolgreiche) Anonymisierung i.S.d. DSGVO liegt nahe, wenn es sich um eine einfach gelagerte Entscheidung aus der ordentlichen Gerichtsbarkeit ohne Sonderzuständigkeiten handelt (i.e. keine Arzthaftungstatbestände; keine komplexen Sachverhalte, die Umstände über eine mit einem Pseudonym gekennzeichnete Person offenbaren), die Entscheidung nicht von besonderem öffentlichem Interesse war (i.e. in der Berichterstattung keine Umstände thematisiert wurden, die die Identifizierung einzelner natürlicher Personen ermöglichen) und (nur) wenige Personen beteiligt waren.

Demgegenüber besteht ein nicht nur unerhebliches Risiko der Anwendbarkeit der DSGVO, wenn es sich um komplexe und ungewöhnliche Sachverhalte handelt. Eine dahingehende Vermutung kann sich beispielsweise ergeben aus dem Umfang des Tatbestands, aus einer Spezialzuständigkeit (z.B. einer Kammer für Arzthaftungssachen, vgl. § 348 Abs. 1 Nr. 2 lit. e ZPO) und aus der Anzahl der beteiligten und zu pseudonymisierenden Personen. Ferner können mehrere verknüpfbare Tatsacheninstanzen oder Dokumente mit sich ergänzenden Ausführungen zum Sachverhalt eine Identifizierbarkeit natürlicher Personen erleichtern. Das betrifft beispielsweise zwei aufgrund des Aktenzeichens und Verfahrensverlaufs miteinander verknüpfbare Entscheidungen (z.B. die landgerichtliche Entscheidung in der Berufung gegen eine Entscheidung des Strafrichters, Schöffengerichts oder des Amtsgerichts in Zivilsachen), und auch die jeweilige Entscheidung i.V.m. staatsanwaltlichen Entscheidungen, deren Aktenzeichen nach Ziff. 2.4 lit. d Verwaltungsvorschrift nicht anonymisiert wird. In diesen Fällen kann aufgrund eingeschränkter Verfügbarkeit des Textes der staatsanwaltschaftlichen Entscheidung allerdings mit Blick auf die breite Öffentlichkeit eine Anonymisierung i.S.d. DSGVO vorliegen.

5. Rechtmäßigkeit der Verarbeitung unter der DSGVO

Soweit die Einhaltung der Anonymisierungsanforderungen aus einer Verwaltungsvorschrift nicht zu einer Anonymisierung i.S.d. DSGVO führt, kann unter Beachtung der vorstehend dargestellten Anforderungen die Verarbeitung gegebenenfalls gleichwohl rechtmäßig sein und den DSGVO-Vorgaben genügen.

a. Verwaltungsvorschrift als Rechtsgrundlage

Zunächst ist zu fragen, ob die Verwaltungsvorschrift als eine Rechtsgrundlage unter die Öffnungsklausel des Art. 6 Abs. 1 lit. c, e, Abs. 2, 3 DSGVO fällt und damit grundsätzlich die Verarbeitung erlaubt. Für Art. 6 Abs. 1 lit. c, e, Abs. 2, 3 DSGVO erforderlich ist ein materielles Gesetz,[103] sodass nicht allgemein, sondern nur innerhalb der Verwaltung wirkende (Verwaltungs-)Vorschriften den Anforderungen nicht genügen.[104] Überdies zielt die Verwaltungsvorschrift gerade ab auf die Anonymisierung als eine Verarbeitung, nicht aber auf Verarbeitungen in Folge einer unzureichenden Anonymisierung.

b. Veröffentlichung (teil-)anonymisierter Gerichtsentscheidungen gestützt auf Art. 6, 9 f. DSGVO

Allerdings kommt unter Rekurs auf andere Vorschriften eine Rechtsgrundlage für die Verarbeitung gelegentlich verbliebener personenbezogener Daten[105] in zu veröffentlichenden Gerichtsentscheidungen in Betracht, namentlich betreffend Art. 6 Abs. 1 UAbs. 1 lit. c, e, f DSGVO[106] i.V.m. Art. 4 Abs. 1 BayDSG, § 3 Abs. 1 DSG NRW sowie gegebenenfalls Art. 9 Abs. 2 lit. f, Art. 10 S. 1 Var. 2 DSGVO. Eine derartige Rechtsgrundlage senkt faktisch den Anonymisierungsstandard ab, indem zwar die Gerichtsentscheidungen gegebenenfalls nicht vollständig anonymisiert sind, diese aber aufgrund der einschlägigen Rechtsgrundlage dennoch in der (teil-)anonymisierten Form zulässigerweise veröffentlicht und gegebenenfalls auch weiterverarbeitet werden können i.S.d. Art. 5 Abs. 1 lit. b, Art. 6 Abs. 4 DSGVO.

103 *Frenzel*, in: Paal/Pauly, Art. 6 DSGVO Rn. 16.
104 *Schulz*, in: Gola/Heckmann, Art. 6 DSGVO Rn. 56.
105 Zur Rechtsgrundlage für die Anonymisierung s. unter D.V.4.c.aa.
106 Zur Differenzierung zwischen lit. c, e s. nachfolgend unter D.V.4.b.

Die Veröffentlichung von Gerichtsentscheidungen ist eine öffentliche Aufgabe bzw. gegebenenfalls im Einzelfall sogar Ausfluss einer Pflicht der Gerichte.[107] Zum einen rechtfertigt dieser Befund die Anonymisierung als schonende(re) Maßnahme zum Ausgleich der Interessen der Allgemeinheit an der Veröffentlichung und der Persönlichkeits- und Datenschutzinteressen der betroffenen Personen.[108] Denn die Anonymisierung verwirklicht gerade die Datenschutzgrundsätze der Datenminimierung und Speicherbegrenzung i.S.d. Art. 5 Abs. 1 lit. c, e DSGVO. Zum anderen kann auch die Verarbeitung von personenbezogenen Daten in nicht vollständig anonymisierten Gerichtsentscheidungen der Erfüllung von öffentlichen Aufgaben dienen.[109] Diese Verarbeitung könnte in diesem Sinne (noch) als Kerntätigkeit der Justiz angesehen werden und daher – mit gleichem Ergebnis – auch auf Art. 6 Abs. 1 UAbs. 1 lit. f DSGVO gestützt werden.[110]

Die Brücke zu einer rechtlich niedergelegten Aufgabe i.S.d. Art. 6 Abs. 1 UAbs. 1 lit. e DSGVO schlagen insoweit die Generalklauseln in Art. 4 Abs. 1 BayDSG[111] bzw. § 3 Abs. 1 DSG NRW, die als gesetzliche Regelung grundsätzlich die Zulässigkeit einer erforderlichen Verarbeitung zur Erfüllung einer der jeweiligen öffentlichen Stelle obliegenden Aufgabe vorsehen. Diese Generalklauseln können im Hinblick auf die Anforderungen aus Art. 6 Abs. 2, 3 DSGVO und die grundrechtlich unterlegten Interessen betroffener Personen allerdings keine eingriffsintensiven Verarbeitungen rechtfertigen[112] und setzen insbesondere die Erforderlichkeit, also nicht nur eine bloße Nützlichkeit, für die jeweilige öffentliche Aufgabe voraus.[113]

107 *VG Stuttgart*, ZD 2022, 583 (Rn. 33).

108 *VG Stuttgart*, ZD 2022, 583; allgemein die Zulässigkeit einer Maskierung unterstellend *EDPB*, Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, Rn. 101; s. auch *EDPB*, Guidelines 01/2025 on Pseudonymisation, Rn. 23.

109 *VG Stuttgart*, ZD 2022, 583.

110 S. nämlich unten zu Art. 6 Abs. 1 UAbs. 2 DSGVO unter D.V.4.a.

111 Art. 4 Abs. 2 BayDSG zum Grundsatz der Direkterhebung ist für die Weiterverarbeitung durch die Ministerien gestützt auf Art. 4 Abs. 2 S. 2 Nr. 2, 4, S. 3 BayDSG gewahrt. Insbesondere werden überwiegende schutzwürdige Interessen der betroffenen Personen durch die justizinterne Weiterverarbeitung nicht beeinträchtigt.

112 *Hornung/B. Wagner*, ZD 2020, 223 (225); vgl. *LfDIBW*, Diskussionspapier: Rechtsgrundlagen im Datenschutz beim Einsatz von Künstlicher Intelligenz, S. 31.

113 Zur vergleichbaren Regelung in Baden-Württemberg *LfDIBW*, Diskussionspapier: Rechtsgrundlagen im Datenschutz beim Einsatz von Künstlicher Intelligenz, S. 31.

Die Notwendigkeit der Veröffentlichung von jedenfalls publikationswürdigen,[114] anonymisierten Gerichtsentscheidungen ist anerkannt,[115] gleich ob sie auf verfahrensrechtliche Vorschriften, gegebenenfalls in analoger Anwendung (z.B. § 299 Abs. 2 ZPO zur Akteneinsicht, §§ 474 ff. StPO in strafrechtlichen Verfahren, § 78 Abs. 1 S. 4, 6 SGB X für Sozialdaten, § 30 AO für dem Steuerverhältnis unterfallende Daten; s. Ziff. 1.2.2 Verwaltungsvorschrift), Richterrecht oder sonstige – grundrechtliche – Erwägungen gestützt wird.[116] Hieraus folgt im Ergebnis eine hinreichende Zuweisung der öffentlichen Aufgabe, die i.V.m. der jeweiligen landesrechtlichen Generalklausel für den Bereich der Rechtsprechung den Anforderungen aus der DSGVO genügen kann.

Es ist anerkannt, dass es für einen schonenden Ausgleich zwischen dem öffentlichen und rechtsstaatlich unterlegten Interesse an der Urteilsveröffentlichung und dem Interesse betroffener Personen einer weitreichenden Anonymisierung bedarf.[117] Dieser Ausgleich ist auf Ebene der Erforderlichkeit für die öffentliche Aufgabenerfüllung zu verwirklichen.

Mit Blick auf die für die Anonymisierung zu berücksichtigenden Faktoren, der Öffentlichkeit einer Gerichtsverhandlung und der möglichen Presseberichterstattung, kann eine vollständige Anonymisierung allenfalls selten erreicht werden.[118] Insbesondere erfordert die öffentliche Aufgabe der Urteilsveröffentlichung einen nachvollziehbaren Entscheidungstext, selbst wenn hierdurch ein Personenbezug erhalten bleibt (z.B. durch die Wahl von einheitlichen Kürzeln für jede auftauchende Person anstelle eines einheitlichen Kürzels für alle Personen oder die Beibehaltung von Aktenzeichen[119]). Weiterhin sind der Bekanntheitsgrad der betroffenen Person,[120] die

114 Für dieses in der Rechtsprechung verwendete Merkmal fehlen klare Maßstäbe. Das *BVerwG*, NJW 1997, 2694 (2695) grenzt etwa zur grundsätzlichen Bedeutung i.S.d. Revisionsrechts ab und geht davon aus, dass sich die Veröffentlichungswürdigkeit „aus der Sicht derjenigen, die mit der Publikation erreicht werden sollen", beurteilt.

115 *BVerfG*, ZD 2016, 77; *BVerwG*, NJW 1997, 2694; *BGH*, NJW 2018, 3123 (Rn. 14); *OLG München*, ZD 2021, 379; *Schild*, in: BeckOK Datenschutzrecht, Syst. E. Rn. 56; *OLG Celle*, NStZ 1990, 553; *OVG Lüneburg*, NJW 1996, 1489; *VG Hannover*, NJW 1993, 3282; *Walker*, JurPC 1998, 34; *Hirte*, NJW 1988, 1698; *OVG Bremen*, NJW 1989, 926; *Hoffman-Riem*, JZ 1989, 637; *Kockler*, JurPC 1996, 46.

116 Überblick etwa bei *OLG Celle*, NStZ 1990, 553 (553); *BGH*, NJW 2018, 3123 (Rn. 14); *Ludyga*, ZUM 2021, 887 (887 f.).

117 *BVerwG*, NJW 1997, 2694; *Nöhre*, MDR 2019, 136 (136) m.w.N.

118 Dementsprechend spricht bspw. das *BVerwG*, NJW 1997, 2694 (2695) auch von einer „neutralisierten Fassung".

119 S. hierzu Ziff. 2.4 Verwaltungsvorschrift.

120 S. hierzu Ziff. 2.4 Verwaltungsvorschrift.

(zulässigen) übrigen Quellen sowie der Sensibilitätsgrad der Daten auch unterhalb der Schwellen von Art. 9, 10 DSGVO zu berücksichtigen.

Vor diesem Hintergrund fällt die Abwägung regelmäßig zugunsten der Veröffentlichung aus, wenn:

- Kennungen durch geeignete Pseudonyme ersetzt werden, z.B.:
  - o Namen werden nicht durch Initialen, sondern durch andere, eindeutige Kürzel ersetzt (z.B. „Person 1“);
  - o Ortsangaben werden nicht bloß gekürzt, sondern – soweit nicht für das Verständnis erforderlich – ebenfalls durch andere, eindeutige Kürzel ersetzt (z.B. „Anschrift 1“);
  - o E-Mail-Adressen, Telefonnummern werden ebenfalls in geeigneter Weise ersetzt;
- sensible personenbezogene Daten auch unter Inkaufnahme von Verständlichkeitseinbußen möglichst nicht mehr einer natürlichen Person zugeordnet werden können;
- gegebenenfalls einzelne Datumsangaben entfernt werden, die sich besonders zur Identifizierung einer natürlichen Person eignen, jedoch nicht verständnisfördernd sind;
- sonstige identifizierende Merkmale überprüft und gegebenenfalls bei hoher Identifizierungswahrscheinlichkeit und sehr geringem Beitrag zur Verständlichkeit der Gerichtsentscheidung entfernt bzw. neutralisiert werden.

Die unterschiedliche Anonymisierungspraxis der Gerichte dürfte, wie es das Beispiel Nordrhein-Westfalen zeigt, regelmäßig diesen Anforderungen genügen. Das gilt insbesondere für die jüngeren, strengeren Anonymisierungsrichtlinien.

Die Verarbeitung eines Restbestands an besonderen Kategorien personenbezogener Daten i.S.d. Art. 9, 10 DSGVO kann verfassungs- und grundrechtlich geboten sein, da ansonsten Gerichtsentscheidungen aus beispielsweise dem Strafrecht nie veröffentlicht werden könnten, was eine rechtstaatlich bedenkliche Lücke in die Veröffentlichungspraxis reißen würde. Insbesondere Art. 9 DSGVO hält nur wenige Ausnahmetatbestände bereit. Die Veröffentlichung einer weitgehend anonymisierten Gerichtsentscheidung kann – mit einem Restrisiko einer abweichenden Auffassung des EuGH – allerdings noch als Handlung des Gerichts im Rahmen der justiziellen Tätigkeit im konkreten Verfahren angesehen werden und daher gestützt auf Art. 9 Abs. 2 lit. f DSGVO zulässig sein. Für personenbezogene Daten über Straftaten mag das deutsche Verfassungsrecht i.V.m. den ein-

schlägigen Anknüpfungs- und Konkretisierungstatbeständen in den Verfahrens- und Prozessordnungen[121] als mitgliedstaatliches Recht herangezogen werden.

Die Veröffentlichung weitgehend anonymisierter Gerichtsentscheidungen ist somit datenschutzrechtlich grundsätzlich zulässig. Voraussetzung hierfür ist regelmäßig ein hoher Anonymisierungsgrad (mind. 99 %) mit Blick auf Kennungen, wobei ein deutlich geringerer Anonymisierungsgrad sonstiger identifizierender Merkmale (u.U. sogar deutlich unter 50 %) unter Berücksichtigung der Verständlichkeitsanforderungen genügen kann. Veröffentlichte Gerichtsentscheidungen, die – wie es für Gerichtsentscheidungen aus der jüngeren Vergangenheit oft der Fall sein dürfte – diese Voraussetzungen erfüllen, oder unveröffentlichte, aber nach diesem Standard anonymisierte Urteile können daher in Ansehung einer zulässigen Veröffentlichung auch risikoarm für justizinterne Zwecke weiterverarbeitet werden. Es besteht zwar insoweit keine echte Zulässigkeitsfiktion, wohl aber die Möglichkeit der zulässigen Weiterverarbeitung für justizinterne Zwecke.

c. Keine Haftungsprivilegierung nach Art. 4 ff. Digital Services Act

Der Digital Services Act (DSA) sieht in Art. 4 ff. für die Übermittlung, die Zwischenspeicherung („Caching") und die Speicherung („Hosting") Haftungsprivilegien für Diensteanbieter im Internet im Hinblick auf fremde Inhalte vor. Zu fragen ist, ob diese Haftungsprivilegien für den Rückgriff eines Drittanbieters auf veröffentlichte Gerichtsentscheidungen Anwendung finden und damit die weitere Verwertbarkeit gegebenenfalls nicht vollständig anonymisierter Gerichtsentscheidungen erleichtern können. An dieser Stelle ist darauf zu verweisen, dass Art. 2 Abs. 4 lit. g DSA einen (eingeschränkten) Vorrang der DSGVO vorsieht. Die Haftungsprivilegierungen entfalten insoweit gegenüber einer Haftung nach der DSGVO für die Verarbeitung personenbezogener Daten grundsätzlich keine Wirkung.[122] Überdies lägen ohnehin mangels durch einen Nutzer bereitgestellter Informationen die Tatbestandsmerkmale der Art. 4 ff. DSA nicht vor.

121 S. z.B. unter D.II.3.
122 *T. Radtke*, in: Gersdorf/Paal, Art. 6 DSA Rn. 15.

d. Sonderregelung für die Aufsicht über die justizielle Tätigkeit

Die Aufsichtsbehörden i.S.d. DSGVO sind nach Art. 55 Abs. 3 DSGVO nicht zuständig[123] für „die Aufsicht über die von Gerichten im Rahmen ihrer justiziellen Tätigkeit vorgenommenen Verarbeitungen". Das betrifft die Tätigkeit der Rechtsprechung, die gerade Ausdruck der verfassungsrechtlich garantierten richterlichen Unabhängigkeit ist.[124] Nur die einzelne Veröffentlichung durch ein Gericht kann dieser Ausnahme unterfallen, nicht dagegen die Anonymisierung zahlreicher Urteile durch eine andere Stelle zur Optimierung von Arbeitsabläufen der Justiz.

e. Verschulden, Art. 82; Aufsichtsmaßnahmen, Art. 58 und 83 DSGVO

Sofern es im Einzelfall zu einem Verstoß kommt, sind gegebenenfalls etwaig getroffene Maßnahmen im Rahmen einer Schadensersatzhaftung nach Art. 82 DSGVO sowie im Rahmen von Aufsichtsmaßnahmen nach Art. 58, 83 DSGVO zu berücksichtigen.

Für die Begründung eines Schadensersatzanspruchs nach Art. 82 DSGVO genügt ein weitgefasstes Vertretenmüssen, das nur dann abzulehnen sein wird, wenn der Verantwortliche „nachweist, dass er in keinerlei Hinsicht für den Umstand, durch den der Schaden eingetreten ist, verantwortlich ist" (Art. 82 Abs. 3 DSGVO). In Ansehung der hohen Anforderungen, die an den Entlastungsbeweis zu stellen sind,[125] dürfte die Berufung auf eine gegebenenfalls unzureichende Anonymisierungsrichtlinie regelmäßig nicht für eine Haftungsbefreiung i.S.d. Art. 82 Abs. 3 DSGVO genügen. Dabei ist insbesondere zu berücksichtigen, dass dem Verantwortlichen selbst (z.B. einem Oberlandesgericht) regelmäßig die Anonymisierungsrichtlinien zugerechnet werden können und der Verantwortliche sich insoweit nicht für Unzulänglichkeiten in der von ihm zugrunde gelegten Anonymisierungsrichtlinie entlasten kann.

Ein datenschutzrechtlich induziertes Bußgeld kann deutsche öffentliche Stellen nicht treffen (vgl. § 43 Abs. 3 BDSG, Art. 22 BayDSG, § 32 DSG NRW, Art. 83 Abs. 7 DSGVO). In das Auswahlermessen bzgl. etwaiger Aufsichtsmaßnahmen dürfte (unter anderem) einzustellen sein die Schwere des Verstoßes (Art. 83 Abs. 2 S. 2 lit. a DSGVO), die „Vorsätzlichkeit oder

123 S. zur Zuständigkeit ErwGr. 20 DSGVO.
124 Vgl. *Nguyen/Stroh*, in: Gola/Heckmann, Art. 55 DSGVO Rn. 10.
125 *EuGH*, NJW 2024, 1561 (Rn. 44 ff.).

Fahrlässigkeit des Verstoßes“ (Art. 83 Abs. 2 S. 2 lit. b DSGVO) sowie der „Umfang der Zusammenarbeit mit der Aufsichtsbehörde, um dem Verstoß abzuhelfen und seine möglichen nachteiligen Auswirkungen zu mindern“ (Art. 83 Abs. 2 S. 2 lit. f DSGVO).

6. Zwischenergebnis

Die DSGVO wird auf bereits (vermeintlich) anonymisierte und veröffentlichte Gerichtsentscheidungen anzuwenden sein. Insbesondere ältere Gerichtsentscheidungen sind gegebenenfalls nicht nach dem Stand der Technik anonymisiert und erfordern Nachbesserungen.

Am Beispiel der Anonymisierungsrichtlinien der Justiz in Nordrhein-Westfalen zeigt sich, dass der Begriff der Anonymisierung nicht stets im technischen Sinne verstanden wird. So unterfallen beispielsweise Informationen über absolute Personen der Zeitgeschichte ebenfalls der DSGVO. Selbst wenn eine Gerichtsentscheidung unter Berücksichtigung dieser Aspekte gegebenenfalls als nicht vollständig anonymisiert anzusehen ist, wird insoweit allerdings vielfach eine Rechtfertigung nach Art. 6, 9 f. DSGVO in Betracht kommen können. Neuere Anonymisierungsrichtlinien wie die der Justiz in Nordrhein-Westfalen können vor diesem Hintergrund die Grundlage für eine datenschutzkonforme Teilanonymisierung und zulässige Veröffentlichung der Gerichtsentscheidungen legen.

Im Einklang mit den Anonymisierungsrichtlinien ist ein besonderes Augenmerk auf die Anonymisierung von Entscheidungen zu legen, die nach äußeren Merkmalen (z.B. Umfang des Tatbestands, Anzahl der beteiligten Personen oder gerichtliche Spezialzuständigkeit) verschiedene Informationen enthalten können, weshalb zusammengenommen und unter Berücksichtigung weiterer Umstände natürliche Personen identifiziert können.

III. (Teil-)Anonymisierung unveröffentlichter Gerichtsentscheidungen durch das Erlanger Tool

Zur automatisierten Durchführung der Anonymisierung kommt im vorliegend zu begutachtenden Sachverhalt das sog. Erlanger Tool zum Einsatz. Das Erlanger Tool wird nachfolgend daraufhin untersucht, ob es die zuvor dargelegten datenschutzrechtlichen Anforderungen an eine Anonymisierung umsetzt und welche datenschutzrechtlichen Restrisiken gegebenenfalls bestehen.

1. Einhaltung der Anonymisierungsanforderungen im Allgemeinen

Das Erlanger Tool ist grundsätzlich geeignet, eine wirksame Anonymisierung von Gerichtsentscheidungen durchzuführen, kann aber zugleich keine pauschale Anonymisierung sämtlicher Gerichtsentscheidungen gewährleisten.

Das Erlanger Tool erzielt nach den dieser Untersuchung zugrunde gelegten Annahmen insbesondere einen Recall, d.h. einen Anteil erfolgreicher anonymisierter Merkmale an allen zu anonymisierenden Merkmalen eines Text innerhalb einer Kategorie, von 94-96 % für eindeutig identifizierende Merkmale bzw. Kennungen, wie etwa Namen und Anschriften natürlicher Personen.[126] Bei den übrigen Informationskategorien (z.B. Daten zum Prozessablauf, Aktenzeichen und Gerichtsort) wird demgegenüber nur ein Recall in einem (weiten) Spektrum von 68-96 % erzielt.[127] Daten aus diesen übrigen Informationskategorien können die Identifizierung natürlicher Personen durch die Heranziehung weiterer Quellen ermöglichen.

In Ansehung der Recall-Prozentwerte, die für eindeutig identifizierende Informationskategorien (i.e. Kennungen) hoch erscheinen, aber noch einen Spielraum in Richtung auf einen Recall-Wert von mind. 99 %[128] erkennen lassen, sind nicht alle bearbeiteten Gerichtsentscheidungen allen in Betracht kommenden Akteuren gegenüber anonym. Mit diesem Befund für sich genommen ist allerdings gerade noch keine abschließende Bewertung der datenschutzrechtlichen Zulässigkeit verbunden, denn die DSGVO kann auch die Veröffentlichung und Weiterverarbeitung von (teil-)anonymisierten Entscheidungen zulassen.[129]

126 *Adrian et al.*, in: Adrian/Kohlhase/Evert/Zwickel, Manuelle und automatische Anonymisierung von Urteilen, S. 188–189, 194, 196 ff.

127 *Adrian et al.*, in: Adrian/Kohlhase/Evert/Zwickel, Manuelle und automatische Anonymisierung von Urteilen, S. 188–189, 194, 196 ff.

128 Eine 100%ige Recall-Wahrscheinlichkeit dürfte technisch bei mehreren Millionen Urteilen in Ansehung der Vielgestaltigkeit des Aufbaus der Gerichtsentscheidungen sowie möglicher unterschiedlicher Formulierung und etwaiger Tippfehler nicht erreichbar sein. Einer solchen Sicherheit bedarf es mit Blick auf die Weiterverarbeitung im Sprachmodell auch nicht.

129 S. zuvor ausf. unter D.II.5.b.

2. Wahrscheinlichkeit der Identifizierung nach allgemeinem Ermessen anhand bearbeiteter Gerichtsentscheidungen

Für die datenschutzrechtskonforme Umsetzung des GSJ-Projekts ist daher zunächst im Rahmen der Frage nach dem Vorliegen personenbezogener Daten mit Blick auf die geringen Recall-Werte für sonstige identifizierende Merkmale die Wahrscheinlichkeit der Identifizierung nach allgemeinem Ermessen zu berücksichtigen.[130] Wenn und soweit eine solche Wahrscheinlichkeit im Einzelfall bejaht wird, kann gegebenenfalls eine Rechtfertigung nach Art. 6 Abs. 1 DSGVO in Betracht kommen.[131]

Es ist daher im Hinblick auf die Anonymisierung zu fragen, ob die Bearbeitung der Gerichtsentscheidungen durch das Erlanger Tool eine Identifizierung natürlicher Personen mittels der Entfernung von Kennungen verhindern kann oder eine solche Identifizierung jedenfalls nach allgemeinem Ermessen hinreichend unwahrscheinlich ist im Fall einer Veröffentlichung der bearbeiteten Gerichtsentscheidung (unter a.) und im Fall einer Verwendung der bearbeiteten Gerichtsentscheidung für ein KI-Training und einen Einsatz eines trainierten KI-Systems für die Use-Cases im GSJ-Projekt (unter b.).

a. Veröffentlichung der bearbeiteten Gerichtsentscheidung

Im Fall der Veröffentlichung steht die bearbeitete Gerichtsentscheidung in vollem Umfang einem unbeschränkten Adressatenkreis zur Verfügung. Diese Adressaten werden mindestens den Zugriff auf sämtliche öffentlich verfügbaren Quellen haben (z.B. Kartendienste, Presseberichterstattung und Registerauskünfte). Das Recalldefizit des Erlanger Tools i.H.v. 4-6 % für eindeutig identifizierende Merkmale (z.B. die Anschrift einer natürlichen Person) lässt eine Identifizierung hierbei möglich erscheinen[132] bzw. führt gegebenenfalls bei eindeutigen Kennungen (z.B. dem Namen einer natürlichen Person) unmittelbar zur Identifizierung und damit jeweils zur Annahme eines Personenbezugs i.S.d. DSGVO. Die (teilweise deutlich) schlechteren Recall-Prozentwerte bei den übrigen Merkmalskategorien in

130 S. oben unter D.I.1.b.

131 S. zur allgemeinen Veröffentlichung teilanonymisierter Gerichtsentscheidungen unter D.II.5.b sowie zur Weiterverarbeitung teilanonymisierter Gerichtsentscheidungen unter D.V.4.

132 S. beispielhaft einen solchen Versuch von *Deuber/Keuchen*, MMR 2023, 338.

Verbindung mit dem unbeschränkten Adressatenkreis erhöhen die Wahrscheinlichkeit eines Personenbezugs weiter.

Unter Verwendung des Erlanger Tools bearbeitete Gerichtsentscheidungen sind daher in dieser Konstellation regelmäßig nicht als anonym anzusehen.

Wegen der Unterschreitung eines Recall-Werts i.H.v. mind. 99 % mit Blick auf Kennungen genügt eine solche Anonymisierung regelmäßig zudem auch nicht den Anforderungen an eine Anonymisierung der Gerichtsentscheidung vor Veröffentlichung.[133]

b. Einsatz zum KI-Training im GSJ-Projekt

Das konkrete Verwendungsszenario der bearbeiteten Gerichtsentscheidungen im GSJ-Projekt hat erheblichen Einfluss auf die Wahrscheinlichkeit einer Identifizierung.

aa. KI-Training als Anonymisierungsmaßnahme

Das KI-Training wirkt bereits wie eine (Teil-)Anonymisierungsmaßnahme, indem die bearbeitete Gerichtsentscheidung nicht einfach gespeichert wird, sondern – trotz möglicher Memorising-Effekte (d.h. der Möglichkeit zur Ausgabe von Artefakten der Trainingsdaten durch ein KI-System) – lediglich die Ausgangsbasis für eine Berechnung darstellt. In Ansehung der zu großen Teilen herausgefilterten Kennungen und identifizierenden Merkmalen sinkt die Wahrscheinlichkeit, dass sich die noch verbleibenden Trainingsdaten eindeutig[134] in dem Sprachmodell oder in den Ausgaben des KI-Systems wiederfinden. Gerade diese Ausgabe des KI-Systems ist aber maßgeblich,[135] wenn man eine mögliche Veröffentlichung auch des Sprachmodells an dieser Stelle außer Betracht lässt.

Sofern sich – stark vereinfachend – der Name „Petra Musterfrau" aufgrund des Einsatzes des Erlanger Tools in dem gesamten Trainingskorpus nur noch in 40-60 Fällen statt zuvor in 1.000 Fällen wiederfindet, sinkt der Anteil dieser Information an den übrigen Trainingsdaten und damit grundsätzlich auch die Wahrscheinlichkeit, dass sich diese Information in

133 S. oben unter D.II.5.b.
134 S. zu Kennungen im Sprachmodell nachfolgend unter D.IV.1.
135 Vgl. oben unter D.I.1.b.bb.

einer Ausgabe wiederfindet. Durch die Tokenisierung, sprich die Aufteilung eines Wortes in zusammenhängende Zeichenfolgen, sinkt ferner die Wahrscheinlichkeit einer eindeutigen Personenzuordnung beim „Zusammenführen" der Trainingsdaten, indem nun auf das Token „Pet" sowohl „ra" (für „Petra") als auch „er" (für „Peter") und verschiedene andere mögliche Zeichen aus einem ähnlichen Kontext folgen könnten.

Die Wahrscheinlichkeit der Identifizierung hängt für das KI-Training im GSJ-Projekt somit maßgeblich ab von der konkreten Eingabeaufforderung an das trainierte KI-System. Ohne eine solche Eingabeaufforderung erhalten – ungeachtet einer Veröffentlichung der Gerichtsentscheidung – keine neuen Personengruppen den Zugriff auf möglicherweise personenbezogene Informationen aus den Gerichtsentscheidungen.

bb. Versehentliche Ausgabe von Trainingsdaten

In den vorbenannten Use-Cases 1-3 sind Eingabeaufforderungen zu erwarten, die auf einen Entwurf von Sachverhalts- und chronologischen Darstellungen und Gegenüberstellungen abzielen. Hierzu werden als Eingabekontext allfällige Auszüge aus der jeweils zu bearbeitenden Akte übergeben. Diese Eingabeaufforderungen lassen vorbehaltlich einer empirischen Untersuchung vereinzelte Ausgaben aus den Trainingsdaten möglich erscheinen. Eine belastbare prozentuale Wahrscheinlichkeitsprognose kann in dieser rechtlichen Untersuchung insoweit allerdings nicht vorgenommen werden.

Unter Berücksichtigung der Eingabeaufforderungen sowie der Bearbeitung der Gerichtsentscheidungen durch das Erlanger Tool dürfte eine solche beiläufige Ausgabe von – dann gegenüber den Nutzern aufgedrängten Trainingsdaten[136] – jedenfalls sehr unwahrscheinlich sein und zudem nicht signifikant abweichen von dem stets verbleibenden Risiko eines Datenlecks trotz effektiver technischer und organisatorischer Maßnahmen nach dem Stand der Technik. Die Wahrscheinlichkeit einer solchen versehentlichen Ausgabe von personenbezogenen Daten kann ferner durch die Verwendung eines entsprechenden System Prompts und durch Ausgabefilter weiter reduziert werden.

136 S. unter D.V.3.b.dd.

cc. Gezielte Extraktion von Trainingsdaten

Neben der versehentlichen ist auch die gezielte Extraktion identifizierender Merkmale und damit letztlich personenbezogener Daten im Rahmen der Wahrscheinlichkeit der Anonymisierung zu betrachten. Die gezielte Extraktion aus den Trainingsdaten ist in den Use-Cases zwar nicht unmittelbar vorgesehen, könnte aber in Abweichung von den Use-Cases durch eine missbräuchliche Benutzung im Einzelfall erfolgen. Eine solche missbräuchliche Nutzung könnte dadurch erleichtert werden, dass die Struktur der Trainingsdaten bekannt ist und gegebenenfalls sogar die Trainingsdaten in Form von veröffentlichen Urteilen öffentlich einsehbar sind. Die Wahrscheinlichkeit einer solchen missbräuchlichen (erfolgreichen) Eingabeaufforderung wird nach allgemeinem Ermessen bei Richtern, Beamten und sonstigen Justizangestellten aufgrund der institutionellen, organisatorischen und gegebenenfalls vertraglichen Bindung an die Rechtmäßigkeit der Nutzung unter Berücksichtigung von Nutzungsanweisungen in einem äußerst niedrigen Bereich anzusiedeln sein.

In dem GSJ-Projekt mit den konkreten Use-Cases führt eine Bearbeitung der Gerichtsentscheidungstexte durch das Erlanger Tool somit grundsätzlich zu einer Anonymisierung der Gerichtsentscheidungen. Die hiermit bereits angedeuteten datenschutzrechtlichen Implikationen eines Sprachmodells sind sogleich unter D.IV näher zu untersuchen. Das verbleibende Restrisiko der Identifizierung natürlicher Personen ist im Hinblick auf eine datenschutzrechtliche Rechtfertigung nach Art. 6, 9 f. DSGVO nachfolgend unter D.V.4 und D.V.5 ausführlich zu prüfen.

IV. Anwendbarkeit des Datenschutzrechts auf ein (veröffentlichtes) Sprachmodells

Von dem Training eines Sprachmodells mit Gerichtsentscheidungen, die durch das Erlanger Tool behandelt wurden, können zusätzliche datenschutzrechtliche Implikationen ausgehen.

1. Sprachmodell einschließlich seiner Gewichte als Speicherung personenbezogener Daten

Ein Sprachmodell erfordert stark vereinfacht – jedenfalls – drei (Verarbeitungs-)Schritte mit Blick auf die Verarbeitung personenbezogener Trai-

ningsdaten: (a) das Training i.w.S. des Sprachmodells, (b) die Speicherung des Sprachmodells und (c) den Einsatz des Sprachmodells als Teil eines KI-Systems durch eine Eingabeaufforderung (sog. Prompt).

a. Training i.w.S.

Das Training erfolgt in mehreren Verarbeitungsschritten (so insbesondere Pre-Training für ein Basismodell, Fine-Tuning für die Anpassung z.B. an einen Einsatz als Chatbot),[137] in denen die Texte verschiedene Formatänderungen durchlaufen und die Texte nur die Grundlage für eine in dem Modell präsentierte Wahrscheinlichkeit der Aufeinanderfolge von Buchstabenketten unter Berücksichtigung von wünschenswerten, hilfreichen Ausgaben bilden.[138] Soweit die Trainingsdaten nicht vollständig anonymisiert sind, werden diese personenbezogenen Daten im Laufe des Trainingsprozesses ausgelesen, verändert, gespeichert und sonst verwendet i.S.d. Art. 4 Nr. 2 DSGVO. Diese Prozesse stellen mithin eine Verarbeitung personenbezogener Daten dar.[139] Hierbei schließt der Verarbeitungsbegriff der DSGVO insbesondere nicht die wahrscheinlichkeitsbasierte Zusammenführung von personenbezogenen Informationen aus.[140]

Zugleich kann diese Verarbeitung als eine Anonymisierungstechnik wirken (s.o. unter D.I.3.b.cc). Es ist allerdings zu berücksichtigen, dass das Training abhängig von der Zusammensetzung des Trainingsdatensatzes einzelne, oft in ähnlicher Weise vorkommende Textabschnitte verstärkt berücksichtigen und daher gegebenenfalls eher in die Ausgabe einfließen lassen dürfte. Textbausteine, die aus einem Verfahren stammen und in andere Verfahren übernommen wurden, sind daher besonders sorgfältig darauf zu prüfen, ob sich hierin Kennungen oder sonstige identifizierende Merkmale finden.

137 S. etwa *Dornis/Stober*, Urheberrecht und Training generativer KI-Modelle, S. 48.
138 Zu den einzelnen Trainingsphasen *International Working Group on Data Protection in Technology*, Working Paper on Large Language Models (LLMs), S. 11 ff.
139 *Hüger*, ZfDR 2024, 263 (277); *Schäfer*, ZD 2025, 12 (14); *Golland*, EuZW 2024, 846 (847); *Werry*, MMR 2023, 911 (912); wohl auch *Dewitte*, in: AI Meets the GDPR, S. 140; a.A. aufgrund der Tokenisierung *A. Spies*, MMR 2024, 289 (290).
140 *Schwartmann/Köhler*, RDV 2024, 316 (316).

b. Speicherung des Sprachmodells

Mit Abschluss des Trainings könnte ein etwaiger Personenbezug (zunächst) vollständig verloren gegangen sein und erst wieder (neu) im Rahmen des Einsatzes des Sprachmodells hergestellt werden. In diesem Fall wäre die fortlaufende Speicherung des Sprachmodells keine Speicherung personenbezogener Daten i.S.d. Art. 4 Nr. 2 DSGVO und die DSGVO wäre insoweit nicht zur Anwendung berufen.

Dieser Befund betrifft insbesondere die Betroffenenrechte auf Berichtigung (Art. 16 DSGVO) und Löschung (Art. 17 DSGVO), die nur Verantwortliche aufgrund der Konzeption eines Sprachmodells vor Herausforderungen stellen, wenn in dem Sprachmodell gegebenenfalls zu berichtigende oder löschende Informationen gespeichert sind. Die weniger umstrittene Einstufung der Ausgabe personenbezogener Daten als eine Verarbeitung[141] bliebe demgegenüber grundsätzlich ohne Auswirkungen auf das Sprachmodell insgesamt. In diesem Sinne wäre für die Berichtigung oder Löschung einer Ausgabe nur die konkrete Ausgabe, nicht aber das gesamte Sprachmodell in den Blick zu nehmen.

Die Diskussion um die Speicherung personenbezogener Daten in einem Sprachmodell wurde maßgeblich durch ein Diskussionspapier des HmbBfDI angestoßen.[142] Hiernach seien die in einem Sprachmodell gespeicherten Parameter als bloße Wahrscheinlichkeitswerte nicht mit kodierten oder verschlüsselten Informationen vergleichbar.[143] Die Parameter wiesen insbesondere keinen „individuellen Informationsgehalt […] ‚über' natürliche Personen" auf.[144] Verfahren zur gezielten Extraktion von personenbezogenen Daten durch eine Eingabeaufforderung seien aufgrund des unverhältnismäßigen und gegebenenfalls rechtswidrigen Einsatzes nicht als Mittel eines Verantwortlichen zur Identifizierung natürlicher Personen zu berücksichtigen.[145]

141 Hierzu nachfolgend unter D.IV.2.

142 *HmbBfDI*, Diskussionspapier: Large Language Models und personenbezogene Daten; s. auch *Fuchs/Wünschelbaum*, RDV 2024, 314.

143 *HmbBfDI*, Diskussionspapier: Large Language Models und personenbezogene Daten, S. 4; *Fuchs/Wünschelbaum*, RDV 2024, 314 (314).

144 *HmbBfDI*, Diskussionspapier: Large Language Models und personenbezogene Daten, S. 6 f.

145 *HmbBfDI*, Diskussionspapier: Large Language Models und personenbezogene Daten, S. 8.

Die übrigen Aufsichtsbehörden[146] und das Schrifttum zeigen sich in dieser Frage gespalten:[147] Es zeichnet sich allerdings eine Tendenz gegen die pauschale Annahme eines Personenbezugs ab. Das EDPB scheint maßgeblich und vorzugswürdig darauf abzustellen, ob nach allgemeinem Ermessen (bzw. im Rahmen von „reasonable means") personenbezogene Trainingsdaten aus den Parametern abgeleitet oder über die Ausgabe im Einzelfall aus dem Modell extrahiert werden können.[148] Es verbleibt aber das (Rest-)Risiko, dass der EuGH gegebenenfalls einen strengen Maßstab im Rahmen des allgemeinen Ermessens anlegt.

Ohnehin gelangt das Datenschutzrecht jedenfalls dann isoliert zur Anwendung, wenn und soweit konkrete personenbezogene Daten aus dem Trainingskorpus wiedergegeben werden.

Zur weiteren Auseinandersetzung mit diesem Meinungsstreit gilt es, zunächst die Frage nach einer notwendigen Differenzierung zwischen Kennungen und sonstigen identifizierenden Merkmalen (unter aa.), der Reichweite des Personenbezugs im Lichte von sog. Mischdatensätzen (unter bb.), die Frage der Identifizierbarkeit (unter cc.), des Informationsgehalts (unter dd.) und der Anforderungen an die Speicherung (unter ee.) zu adressieren. Hierbei wird herausgearbeitet werden, dass richtigerweise zwischen Kennungen und sonstigen identifizierenden Merkmalen zu differenzieren ist,

146 Personenbezug des Sprachmodells entgegen HmbBfDI möglich nach *LfDIBW*, Diskussionspapier: Rechtsgrundlagen im Datenschutz beim Einsatz von Künstlicher Intelligenz, S. 9; wohl auch *BayLDA*, Datenschutzkonforme Künstliche Intelligenz, S. 5; womöglich auch *EDPB*, Report of the work undertaken by the ChatGPT Taskforce, Rn. 25; ähnl. dem HmbBfDI *Dänische Datatilsynet*, Offentlige myndigheders brug af kunstig intelligens, S. 7.

147 Dem HmbBfDI folgend *Hüger*, ZfDR 2024, 263 (277 f.); *Moos*, CR 2024, 442 (Rn. 15); *Wrobel/Pentzien*, DSB 2024, 200 (203); in der Regel kein Personenbezug nach *Golland*, EuZW 2024, 846 (847); ähnl. *Mühlhoff/Ruschemeier*, ZfDR 2024, 337 (341); gegen HmbBfDI *Hansen/Walczak*, KIR 2024, 82 (85 f.); Personenbezug ebenfalls denkbar nach *Pesch/Böhme*, MMR 2023, 917 (920); *Engeler/Rolfes*, ZD 2024, 423 (426 f.); *Kaulartz*, in: Kaulartz/Braegelmann, Kap. 8.9, Rn. 13, 16; abhängig von den Ausgaben *Schwartmann/Köhler*, in: Schwartmann/Keber/Zenner, 3. Kapitel, Rn. 10; schon vor einigen Jahren einen Personenbezug des Modells in Abhängigkeit vom möglichen Output annehmend *Veale/Binns/Edwards*, Phil. Trans. R. Soc. A 376, 20180083 (6 ff.).

148 *EDPB*, Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, Rn. 38, 43: „that, with reasonable means: (i) personal data, related to the training data, cannot be extracted out of the model; and (ii) any output produced when querying the model does not relate to the data subjects whose personal data was used to train the model"; *EDPB*, Report of the work undertaken by the ChatGPT Taskforce, Rn. 25.

wobei nur erstere im untersuchungsgegenständlichen Projekt vereinzelt zur Annahme eines Personenbezugs des Modells führen können. Für sonstige identifizierende Merkmale wird ein Personenbezug in Ansehung der im GSJ-Projekt wahrscheinlichen KI-Ausgaben grundsätzlich ausgeschlossen sein.

aa. Differenzierung zwischen Kennungen und sonstigen identifizierenden Merkmalen

Kennungen, wie etwa der Name einer natürlichen Person (vgl. Art. 4 Nr. 1 DSGVO), ermöglichen die direkte Identifizierbarkeit einer natürlichen Person und unterliegen nicht dem Vorbehalt der wahrscheinlichen Zuordnung. Soweit also Kennungen im KI-Modell gespeichert sind, ist ein Personenbezug anzunehmen. Zwar schadet insoweit nicht die Darstellung der Kennung in einem anderen Format, es dürften aber die Kennungen regelmäßig nur in zerlegte Bestandteile als wahrscheinliche Verbindung gespeichert sein. Eine direkte Identifizierung ist in diesen Fällen nicht möglich, sodass das Sprachmodell mit Blick auf Kennungen grundsätzlich nicht per se personenbezogen ist bzw. sich der Personenbezug auch für die Kennungen nur aus der indirekten Identifizierung auf Grundlage einer Wahrscheinlichkeitsprognose ergeben kann.

Daher ist nachfolgend für Kennungen als auch für sonstige identifizierende Merkmale vor allem nach verfügbaren Mitteln für die Annahme einer Identifizierbarkeit und nach den übrigen Anforderungen an eine Speicherung zu fragen.

bb. Reichweite des Personenbezugs – Mischdatensätze

Es könnte erwogen werden, dass ein Personenbezug und die Anwendbarkeit der DSGVO bereits deshalb ausscheiden, weil nur ganz vereinzelt personenbezogene Daten enthalten sein sollen.

Für sog. Mischdatensätze aus besonderen Kategorien personenbezogener Daten (i.e. sensiblen Daten) und sonstigen Kategorien personenbezogener Daten hat der EuGH allerdings bereits ausgeurteilt, dass „wenn ein Datensatz, der sowohl sensible als auch nicht sensible Daten enthält, Gegenstand solcher Vorgänge ist und insbesondere als Ganzes erhoben wird, ohne dass die Daten zum Zeitpunkt dieser Erhebung voneinander getrennt werden

können, ist die Verarbeitung dieses Datensatzes aber [grundsätzlich] als im Sinne von Art. 9 Abs. 1 DS-GVO untersagt anzusehen".[149] Bei untrennbaren Datenverbindungen ist die Datenkategorie maßgeblich, an die die strengsten Anforderungen zu stellen sind.

Entsprechendes gilt auch für die (Vor-)Frage, ob ein Datensatz überhaupt personenbezogen ist. Sofern die Trainingsdaten sich nach dem Training des Sprachmodells in Form der Parameter nicht mehr (sinnvoll) trennen lassen, führt der Personenbezug von einzelnen Bestandteilen gegebenenfalls insgesamt zu einer Verarbeitung personenbezogener Daten im Sinne der DSGVO-Vorgaben. Denn die DSGVO sieht gerade keine Schwelle oder das Erfordernis eines überwiegenden Anteils personenbezogener Daten vor.

cc. (Wahrscheinliche) Identifizierbarkeit

Die in dem Sprachmodell gespeicherten Parameter könnten nicht eindeutig und hinreichend wahrscheinlich die Identifizierung natürlicher Personen zulassen und daher als anonym im Sinne der DSGVO zu bewerten sein.

Unterstellt man eine weitgehende Anonymisierung der Trainingsdaten und das Training als zufällig effektive Anonymisierungstechnik (z.B. aufgrund der Zusammensetzung der verwendeten Trainingsdaten),[150] kann eine Person gegebenenfalls nicht (mehr) anhand der Parameter unter Zuhilfenahme des Einsatzes des Sprachmodells durch eine Eingabeaufforderung identifizierbar sein. Für die Frage der wahrscheinlichen Identifizierung nach allgemeinem Ermessen kommt im GSJ-Projekt grundsätzlich[151] nur die Ausgabe des KI-Systems in Betracht (vgl. zur Wiedergabe von Trainingsdaten in der Ausgabe die sog. Memorisierung[152] und diese aus-

149 *EuGH*, ZD 2023, 664 (Rn. 89) – Meta Platforms u.a.; hierzu auch *Nabulsi*, ZD 2025, 3.

150 *Carlini et al.*, The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks, S. 2 f., 10, 12 ff.

151 S. aber nachfolgend zur möglichen Veröffentlichung des Sprachmodells unter D.I.3.

152 Hierzu etwa *Carlini et al.*, Extracting Training Data from Large Language Models; *Karamolegkou et al.*, Copyright Violations and Large Language Models; *Nasr et al.*, Scalable Extraction of Training Data from (Production) Language Models; *Kaulartz*, in: Kaulartz/Braegelmann, Kap. 8.9, Rn. 12; *Nolte/Finck/Meding*, Machine Learners Should Acknowledge the Legal Implications of Large Language Models as Personal Data, S. 2 m.w.N.; *Mueller et al.*, LLMs and Memorization:

nutzend sog. Model Inversion Attacks[153]). Denn die justizinternen Nutzer können im Ausgangsfall nur auf das KI-System mit seiner Benutzeroberfläche zugreifen, nicht aber selbst das Sprachmodell mitsamt der enthaltenen Gewichte einsehen. Extraktionen außerhalb der Ausgabe[154] sind daher vorliegend außer Betracht zu lassen.

Nachfolgend sind nach einer grundlegenden dogmatischen Betrachtung (unter (1)) Faktoren für die Wahrscheinlichkeit der Identifizierung im GSJ-Projekt herauszuarbeiten (unter (2) und (3)).

(1) Dogmatische Bedenken mit Blick auf das Zusammenfallen einer Verarbeitung und der Heranziehung von Identifizierungsmitteln

Der Schritt von der Speicherung des Modells zum Einsatz im Einzelfall kann womöglich auch nur als ein dem Verantwortlichen zur Verfügung stehendes und vernünftigerweise eingesetztes Mittel zur Identifizierung natürlicher Personen angesehen werden.[155] In der Folge wären die in einem Sprachmodell gespeicherten Informationen auf eine natürliche Person zurückzuführen (und daher gegebenenfalls personenbezogen), womit bereits die Speicherung des Sprachmodells als eine Speicherung personenbezogener Daten anzusehen wäre.

Dem lässt sich entgegenhalten, dass die Mittel nicht gleichbedeutend sein dürfen mit einer Verarbeitung personenbezogener Daten. Mit anderen Worten gilt, dass erst der Einsatz des Sprachmodells (als Verarbeitung) mittels einer entsprechenden Eingabeaufforderung (als ein „Mittel") die Verarbeitung personenbezogener Daten darstellt. Diese Verarbeitung wirkt nicht zurück auf vorherige Vorgänge im Zusammenhang mit den Daten, über die sich noch keine natürliche Person identifizieren ließ.

On Quality and Specificity of Copyright Compliance; hierzu auch im datenschutzrechtlichen Kontext *Maltzan/Käde*, DSRITB 2020, 505 (516).

153 *Fredrikson/Jha/Ristenpart*, in: Ray/N. Li/Kruegel, 1322; *Veale/Binns/Edwards*, Phil. Trans. R. Soc. A 376, 20180083; speziell zu sog. Membership Inference Attacks, die im vorliegenden Fall Rückschlüsse darauf zulassen könnten, ob eine natürliche Person an einem Gerichtsprozess beteiligt war, *Shokri et al.*, Membership Inference Attacks against Machine Learning Models; *Kaulartz*, in: Kaulartz/Braegelmann, Kap. 8.9, Rn. 11.

154 *EDPB*, Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, Rn. 38.

155 Vgl. oben unter D.I.1.b.bb.

Im Ergebnis lässt sich jedoch der einschlägigen EuGH-Judikatur die Grundtendenz entnehmen, dass die spätere Zusammenführung der identifizierenden Merkmale und übriger Informationen einen Personenbezug der zuvor getrennt gespeicherten Merkmale und Informationen begründet.[156]

(2) Begünstigende Faktoren für eine Identifizierung

Die Möglichkeit zur Identifizierung einzelner natürlicher Personen besteht insbesondere dann fort, wenn und wo eine Vorstellung vom Inhalt und der Struktur der Trainingsdaten gegeben ist.[157] Für ein in der Justiz eingesetztes Sprachmodell dürfte eine solche Vorstellung regelmäßig anzunehmen sein, in der (anonymisierte) Gerichtsentscheidungen zumeist frei zugänglich und die Grundzüge der Struktur der Trainingsdaten grundsätzlich bekannt sind.

(3) Vorbehalt der wahrscheinlichen Identifizierung nach allgemeinem Ermessen

Dieser Befund der grundsätzlich möglichen Identifizierung steht allerdings unter dem Vorbehalt, dass verfügbare Mittel (wie z.B. entsprechende Eingabeaufforderungen) nach allgemeinem Ermessen wahrscheinlich zur Identifizierung genutzt werden.[158] Insoweit ist zwischen der zufälligen Ausgabe von personenbezogenen Trainingsdaten und der gezielten Extraktion solcher Trainingsdaten zu unterscheiden.

Die zufällige Ausgabe von personenbezogenen Trainingsdaten ist unter Berücksichtigung des Einsatzes des Erlanger Tools, einer erweiterten Anonymisierung einzelner Gerichtsentscheidungen nach Clusterung sowie des Einsatzes von System Prompts und Ausgabefiltern sehr unwahrscheinlich. Diese Prognose kann durch Beschränkungen der Eingabemöglichkeiten weiter abgesichert werden, indem z.B. die Eingabeaufforderung nicht selbst

156 S. z.B. *EuGH*, NJW 2016, 3579 – Breyer.

157 S. zur erhöhten Memorisierungswahrscheinlichkeit durch längeren Kontext *Carlini et al.*, Quantifying Memorization Across Neural Language Models, S. 5 f.; vgl. *Moos*, CR 2024, 442 (Rn. 48 ff.); *Kaulartz*, in: Kaulartz/Braegelmann, Kap. 8.9, Rn. 13.

158 Ebenfalls die Ausgabe in diesem Zusammenhang hervorhebend *EDPB*, Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, Rn. 38.

durch den Nutzer in einem Freitextfeld festgelegt werden, sondern demgegenüber nur unter festgelegten Möglichkeiten gewählt werden kann. Selbst wenn es zu einer Ausgabe kommen sollte, die eine Identifizierung ermöglichen könnte, ist es bei Verwendung der Ausgabe innerhalb der Justiz nach allgemeinem Ermessen ebenfalls als sehr unwahrscheinlich zu bewerten, dass weitere Mittel genutzt werden, um die Identifizierung tatsächlich durchzuführen.[159] Das Interesse an einer Identifizierung wird innerhalb der justiziellen Tätigkeit und Befassung mit einem anderen Verfahren gering sein, während zugleich dürfte der Aufwand für diese Identifizierung regelmäßig als hoch bis sehr hoch einzustufen sein dürfte.[160]

Entsprechende Überlegungen gelten für die Möglichkeit, durch gezielte Eingabeaufforderungen Informationen zu extrahieren, die gegebenenfalls unter Heranziehung weiterer Quellen auf eine natürliche Person zurückgeführt werden. Denn auch insoweit besteht nach allgemeinem Ermessen justizintern kein Interesse an einer solchen Identifizierung.[161] Dieser Befund wird durch zwei weitere Argumente gestützt: Zum einen kann innerhalb der Justiz gegebenenfalls schon anderweitig die – rechtlich zulässige – Möglichkeit bestehen, an eine Gerichtsentscheidung zu gelangen. Das Interesse an der aufwändigen Extrahierung von personenbezogenen Daten ist dementsprechend als noch geringer einzuschätzen. Zum anderen sind nach der Rechtsprechung des EuGH grundsätzlich[162] nur rechtlich zulässige Mittel einer Identifizierung heranzuziehen. Die gezielte Extraktion durch die Richterschaft und die Justiz im Übrigen kann eine Pflichtverletzung darstellen oder sogar dem § 202a StGB oder sonstigen Strafvorschriften unterfallen.[163] Entsprechende Dienstanweisungen oder separate vertragliche Vereinbarungen für den Umgang mit einem KI-System können gegebenenfalls (soweit zulässig) diesen Maßstab weiter ausgestalten und folglich die Schwelle zu einer wirksamen Anonymisierung weiter herabsetzen. Flankiert werden sollten diese Vorgaben auch und gerade durch Log-Mechanismen, die eine missbräuchliche Nutzung im Nachhinein nachvollziehbar machen und insoweit auch präventiv wirken können.

159 S. allgemein zu diesem Interesse am Beispiel von Unternehmen *Moos*, CR 2024, 442 (Rn. 44).
160 *Moos*, CR 2024, 442 (Rn. 54 ff.).
161 Vgl. auch *Moos*, CR 2024, 442 (Rn. 52 f.).
162 S. oben unter D.I.1.b.bb.
163 Dies hervorhebend *HmbBfDI*, Diskussionspapier: Large Language Models und personenbezogene Daten, S. 8.

Eine andere Bewertung wird angezeigt sein (und erfolgt nachfolgend unter D.IV.3), sofern das KI-System oder sogar die Gewichte des Sprachmodells und gegebenenfalls der Trainingskorpus öffentlich verfügbar sind. Denn in diesem Fall sind vor allem auch sog. White-Box Attacks durch Dritte möglich, sprich Angriffe unter Ausnutzung der bekannten Struktur des Modells.[164]

dd. Informationsgehalt

Allerdings könnte über das Merkmal des „Beziehens" den im Sprachmodell gespeicherten Parametern kein hinreichend sicherer Informationsgehalt über eine natürliche Person zuzuordnen sein. Die Rechtsprechung verlangt, dass die Information „auf Grund ihres Inhalts, ihres Zwecks oder ihrer Auswirkungen mit einer bestimmten Person verknüpft ist".[165] Die im Modell gespeicherten Parameter als Repräsentation einer Wahrscheinlichkeit des Zusammenhangs von Buchstaben(schnipseln) könnten als zu abstrakt oder zu prognostisch angesehen werden, um ihnen einen solchen Informationsgehalt über eine natürliche Person zu entnehmen.

(1) Abstraktheit

Diese Ansicht scheint der HmbBfDI zu teilen und verweist auf einen mangelnden Informationsgehalt der Parameter betreffend eine Person.[166] Mit Blick auf das weite Verständnis des Merkmals „beziehen" dürfte allerdings die Rekonstruierbarkeit aussagekräftiger Informationen für ein „Beziehen" der Informationen auf eine natürliche Person genügen.[167] In Abhängigkeit von den jeweiligen Informationen ist daher die abstrakte Ausgestaltung der Token bzw. ihrer wahrscheinlichkeitsbasierten Repräsentation im Modell regelmäßig unschädlich.

164 *Fredrikson/Jha/Ristenpart*, in: Ray/N. Li/Kruegel, 1322 (1322 f.); *C. Song/Ristenpart/Shmatikov*, in: Thuraisingham/Evans/Malkin/Xu, 587 (587 f.).

165 *EuGH*, ZD 2018, 113 (Rn. 35) – Nowak; dies wiederum weit auslegend *Generalanwalt Spielmann*, Schlussanträge C-413/23 P, Rn. 33 f.; *Klabunde/Horváth*, in: Ehmann/Selmayr, Art. 4 Rn. 10 f. S. hierzu auch schon unter D.I.1.a.

166 *HmbBfDI*, Diskussionspapier: Large Language Models und personenbezogene Daten, S. 6 f.

167 *Veale/Binns/Edwards*, Phil. Trans. R. Soc. A 376, 20180083 (7).

(2) Unsicherheit unter Berücksichtigung von Halluzinationen

Die Unsicherheit über den Wahrheitsgehalt einer Information aufgrund möglicher Halluzinationen schadet nicht per se.[168] Die EuGH-Judikatur zu besonderen Kategorien personenbezogener Daten lässt sich entsprechend auf sämtliche Kategorien personenbezogener Daten übertragen. So soll es bereits genügen, wenn aus einem Medikament auf mögliche Diagnosen geschlossen werden kann, auch wenn etwa bei einem rezeptfreien Medikament die Möglichkeit einer Bestellung einer natürlichen Person für eine andere Person besteht.[169] Auf eine Richtigkeit der Information wird es gerade nicht ankommen (arg. ex Art. 16 DSGVO).[170] Bei einer Verdachtsberichterstattung dürfte ebenfalls kein Zweifel an der Annahme einer Verarbeitung personenbezogener Daten bestehen. Offen bleibt, ob bei einer gänzlich unwahrscheinlichen und als solcher erkennbaren Annahme einer Verarbeitung ein „Beziehen" auf eine natürliche Person abzulehnen ist. In Ansehung der weiten EuGH-Rechtsprechung wäre eine solche Ausnahme jedenfalls sehr restriktiv zu handhaben und auf Extremfälle zu beschränken.

ee. Speicherung

Die notwendige Heranziehung weiterer Mittel in Form der Eingabeaufforderung an das KI-System kann sich nicht nur auf Ebene der Identifizierbarkeit auswirken, sondern betrifft auch die Frage des Vorliegens einer Speicherung i.S.d. Art. 4 Nr. 2 DSGVO.

Die Speicherung wird definiert als die „Aufbewahrung personenbezogener Daten in verkörperter Form auf einem Datenträger […] mit dem Ziel, die Daten zu einem späteren Zeitpunkt weiterverarbeiten zu können".[171]

168 A.A. und u.a. hierauf die Ablehnung des Personenbezugs eines Sprachmodells stützend *Moos*, CR 2024, 442 (Rn. 18 ff.); s. hierzu auch *Artikel-29-Datenschutzgruppe*, WP 136, S. 17.

169 *EuGH*, NJW 2025, 33 (Rn. 83, 88).

170 *EuGH*, NJW 2025, 33 (Rn. 87); *Artikel-29-Datenschutzgruppe*, WP 136, S. 7; *Hubert*, CR 2025, 77 (Rn. 24).

171 *Arning/Rothkegel*, in: Taeger/Gabel, Art. 4 DSGVO Rn. 76.

(1) Identifizierung im Zusammenhang mit der Speicherung

Die Verantwortlichkeit für eine Speicherung setzt grundsätzlich insbesondere auch eine Zugriffsmöglichkeit voraus.[172] Eine solche Zugriffsmöglichkeit besteht auch und gerade durch die für das Sprachmodell verantwortlichen Stellen.

Das OVG Hamburg verlangt für das Speichern insoweit einen menschlichen, aktiven Umgang mit den entsprechenden personenbezogenen Daten und verneint das Vorliegen eines solchen Umgangs im Fall des Eintritts in das Eigentum an einem Gebäude mit Blick auf darin gelagerte Akten.[173] Diese Überlegung führt in der vorliegend zu begutachtenden Konstellation allerdings nicht zu einer Ablehnung einer Speicherung. Denn das Sprachmodell wurde gerade auch durch die verantwortliche Stelle trainiert.

Dennoch bestehen Zweifel, ob für die Speicherung als rein passiver Vorgang (vgl. Art. 18 Abs. 2 DSGVO)[174] die gleichen Anforderungen an die Einbeziehung von identifizierenden Mitteln gelten wie für andere, aktive(re) Verarbeitungsvorgänge (wie z.B. die Verwendung). Denn das Risiko der Identifizierbarkeit realisiert sich erst dann, wenn über die Speicherung hinaus weitere Daten erhoben und gegebenenfalls abgeglichen werden. Insoweit ist auf Art. 4 Nr. 2, Art. 18 DSGVO zu verweisen, die zeigen, dass grundsätzlich auch die Speicherung auf denselben Begriff der personenbezogenen Daten abstellt.[175]

Es kann für den Personenbezug nach den Schlussanträgen des EuGH-Generalanwalts Spielmann in der Rs. C-41/23 P also das jeweilige Verhältnis und die Verpflichtung zu berücksichtigen sein,[176] sodass die Annahme personenbezogener Daten in dem konkreten Verhältnis mit Blick auf die Speicherung gegebenenfalls einer anderen Bewertung unterfallen mag.

Weiterhin kommt in Betracht, dass die „Speicherung" des Sprachmodells aufgrund des Begriffs der Speicherung i.S.d. Art. 4 Nr. 2 DSGVO keine Verarbeitung personenbezogener Daten darstellt. Die für Ende 2025 zu

172 *Herbst*, in: Kühling/Buchner, Art. 4 Nr. 2 DSGVO Rn. 24; eine Herrschaft über die Aufbewahrung erfordernd *Roßnagel*, in: Simitis/Hornung/Spiecker gen. Döhmann, Art. 4 Nr. 2 DSGVO Rn. 19.

173 *OVG Hamburg*, BeckRS 2020, 30248 (Rn. 21).

174 Danach hat die betroffene Person bspw. bei Unklarheiten betreffend die Richtigkeit der personenbezogenen Daten das Recht auf Einschränkung der Verarbeitung. In diesem Fall bleibt die Speicherung als eingriffsarme Verarbeitung nach Art. 18 Abs. 2 DSGVO zulässig.

175 *Hubert*, CR 2025, 77 (Rn. 31).

176 *Generalanwalt Spielmann*, Schlussanträge C-413/23 P, Rn. 72.

erwartende EuGH-Entscheidung in der Rs. C-413/23 P könnte weitere Anhaltspunkte hierfür bieten.

(2) Aufbewahrung als Ausgangspunkt für eine Reproduktion

Gegen die Annahme einer Speicherung lässt sich ferner anführen, dass die Trainingsdaten in einem Sprachmodell nicht aufbewahrt werden, sondern lediglich die Ausgangsbasis für die Produktion von gänzlich neuen auszugebenden Informationen basierend auf Wahrscheinlichkeiten darstellen. Allerdings sind in dem Sprachmodell gerade wahrscheinliche Ableitungen aus den Trainingsdaten gespeichert, die eine Zuordnung von in den Trainingsdaten enthaltenen Informationen zu einer Person deutlich wahrscheinlicher machen als dies beispielsweise bei einem von Grund auf neu geschriebenen Text der Fall ist. Mit anderen Worten gilt: Wenn die Trainingsdaten einen engen Zusammenhang zwischen „Herrn Müller" und der „Rathausstraße" nahelegen, besteht eine hohe Wahrscheinlichkeit, dass sich durch eine entsprechende Eingabeaufforderung dieser Zusammenhang wiederherstellen lässt. Dieser Zusammenhang wurde aber nicht neu erdacht, sondern ist gerade aufgrund der aus den Trainingsdaten abgeleiteten Zusammenhänge wahrscheinlich für eine Ausgabe. Ebendiese Zusammenhänge sind in dem Sprachmodell repräsentiert und insoweit aufbewahrt.

Mit Blick auf die bisherige Judikatur des EuGH[177] sprechen gewichtige Argumente dafür, dass der EuGH unter diesen Umständen von einer Speicherung ausgehen wird; gleichwohl ist auch eine abweichende Entscheidung mit Blick auf das Merkmal der Aufbewahrung denkbar.

ff. Zwischenergebnis

Die Speicherung eines Sprachmodells kann als die Verarbeitung personenbezogener Daten zu bewerten sein, wenn in den Trainingsdaten (noch) personenbezogene Daten enthalten waren.[178] Hierfür genügt bereits das

177 S. auch die Verarbeitung personenbezogener Daten im Zusammenhang mit einem Wahrscheinlichkeitswert (Score) in *EuGH*, MMR 2024, 153 – Schufa; s. aber früher zur Differenzierung zwischen Wahrscheinlichkeitswert und personenbezogenen Daten *BGH*, ZD 2014, 306.

178 *EDPB*, Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, Rn. 27-34.

vereinzelte Vorkommen personenbezogener Daten, die sich nicht von nicht-personenbezogenen Daten trennen lassen. Die wahrscheinlichkeitsorientierte Architektur genügt den Anforderungen an einen Informationsgehalt über eine natürliche Person. Zudem dürfte eine Speicherung trotz Reproduktion anzunehmen sein, wobei diese Frage allerdings als (noch) weitgehend ungeklärt einzustufen ist.

Die Bewertung des Vorliegens von Anonymität hängt – im Einklang mit der Auffassung des EDPB – maßgeblich davon ab, ob entweder Kennungen eindeutig gespeichert sind oder die Nutzer des Sprachmodells und eines hierauf aufbauenden KI-Systems über die tatsächlichen und rechtlichen Möglichkeiten verfügen, über Eingabeaufforderungen personenbezogene Daten zu extrahieren. Eindeutige Kennungen werden regelmäßig nicht nach herkömmlichem Verständnis in dem Modell „gespeichert". Vielmehr wird es auch und gerade darauf ankommen, ob sich Kennungen und sonstige identifizierende Merkmale über die Ausgabe des KI-Systems extrahieren lassen.

Ein justizintern eingesetztes KI-System (i.e. ohne Möglichkeit, die Gewichte des Sprachmodells einzusehen), dessen Einsatz zur Extrahierung personenbezogener Daten rechtswidrig ist und für das durch technische Mittel (z.B. Ausgabefilter und System Prompt) die Extrahierung verhindert wird, kann daher als anonym zu bewerten sein. Eine solche Bewertung setzt die Beachtung weiterer technischer und organisatorischer Maßgaben voraus. So sollten etwa behörden- bzw. justizinterne Leitlinien zum Umgang mit dem KI-System den Einsatz des KI-Systems absichern, um das Restrisiko einer abweichenden Bewertung durch die Rechtsprechung zu minimieren.

2. Verarbeitung durch Einsatz des Sprachmodells im Einzelfall

Durch den Einsatz des Sprachmodells als KI-System mit einer Eingabeaufforderung im Einzelfall kann es zu einer Verarbeitung personenbezogener Daten kommen. Eine solche Verarbeitung wirkt über die Identifizierungsmittel nicht nur zurück auf die zuvor untersuchte Fragestellung des Personenbezugs eines Sprachmodells, sondern ist darüber hinaus auch eigenständig an der DSGVO zu messen.

Zur Wahrung der Anonymität des Sprachmodells kann zunächst auf die oben unter D.IV.1.b.cc(3) erörterten Sicherungsmechanismen verwiesen werden. Soweit diese Sicherungsmechanismen einen Personenbezug effek-

tiv verhindern, stellen weder interne Vorgänge innerhalb des Sprachmodells noch die Ausgabe eine Verarbeitung personenbezogener Daten im Sinne der DSGVO dar; im Übrigen kommt eine rechtfertigende Rechtsgrundlage nach Art. 6 DSGVO in Betracht.[179]

Für den unwahrscheinlichen Fall einer Offenlegung personenbezogener Daten in der Ausgabe handelt es sich um eine Verletzung des Schutzes personenbezogener Daten (Art. 4 Nr. 12 DSGVO), die gegebenenfalls Meldepflichten nach Art. 33 f. DSGVO auslösen kann. Hieraus folgt keinesfalls zwangsläufig rückwirkend ein Personenbezug des gesamten Sprachmodells, sofern eine Identifizierung im Einzelfall nach allgemeinem Ermessen sehr unwahrscheinlich war und bleibt.

Die sog. Halluzination, d.h. die Ausgabe unrichtiger personenbezogener Daten, stellt ebenfalls eine Verarbeitung dar.[180] Eine Halluzination, die maßgeblich auf den Eingabedaten beruht (z.B. einem eingegebenen Namen und einer hierzu „halluzinierten" Information), wirkt allerdings nicht auf die Identifizierbarkeit anhand des trainierten Sprachmodells zurück, sondern ist als Verarbeitung der Eingabedaten eigenständig zu betrachten.

3. Anforderungen an die Veröffentlichung des Sprachmodells mit Blick auf die Anonymisierung

Die Veröffentlichung des Sprachmodells kann sich mit Blick auf die Mittel zur Identifizierung natürlicher Personen auf die Frage nach deren Anonymität und der Anwendbarkeit der DSGVO auswirken. Dieser Befund gilt unabhängig davon, ob eine solche Veröffentlichung z.B. mit Blick auf den fair trial-Grundsatz[181] oder aus Gründen einer transparenten Verwaltung gewünscht ist.

Für die Veröffentlichung ist zwischen – jedenfalls – drei Stufen zu unterscheiden: Zurverfügungstellung eines KI-Systems (unter a.), Veröffentlichung des Sprachmodells (unter b.) und des Trainingskorpus (unter c.). Mit fortschreitender Stufe nimmt die Wahrscheinlichkeit einer tatsächlich erfolgreichen Anonymisierung jeweils ab.[182]

179 S. unter D.V.4.

180 Arg. Art. 16 DSGVO und s. schon oben unter D.IV.1.b.dd(2) sowie *noyb*, Complaint against OpenAI.

181 *Biallaß*, in: Ory/Weth, Kap. 8, Rn. 372.

182 Vgl. *Roßnagel*, DuD 2024, 513 (519); *Kaulartz*, in: Kaulartz/Braegelmann, Kap. 8.9, Rn. 7.

a. Zurverfügungstellung des KI-Systems

Bei einer Zurverfügungstellung des KI-Systems in Form eines Chatbots erhält zwar ein gegebenenfalls unbeschränkter Personenkreis den Zugriff, dies beschränkt sich aber auf die Eingabeaufforderung und im System implementierte Filter. Anders als bei einem Einsatz innerhalb der Justiz bestehen vielgestaltige Interessenlagen, die gezielte Attacken zur Extrahierung personenbezogener Daten wahrscheinlicher werden lassen. Durch die zuvor durchgeführte Anonymisierung der Trainingsdaten sowie System Prompts und Ausgabefilter besteht weiterhin ein erheblicher Aufwand für Extrahierungsangriffe. In dieser Konstellation kommt daher zwar grundsätzlich eine Nichtanwendbarkeit des Datenschutzrechts in Betracht, ist aber von Faktoren wie der Urteilsanonymisierung, den öffentlich zur Verfügung stehenden Quellen und der Effektivität von Filtern im KI-System abhängig.

Sofern das Datenschutzrecht zur Anwendung berufen ist, kann gegebenenfalls eine Rechtsgrundlage aus Art. 6 DSGVO eingreifen (siehe nachfolgend unter D.V.4). Ferner sind im Hinblick auf die weltweite Abrufbarkeit des KI-Systems etwaige Drittstaaten-Übermittlungen zu beachten und gegebenenfalls zu beschränken.[183]

b. Veröffentlichung des trainierten Sprachmodells (Open Source)

Mit einer Veröffentlichung des Sprachmodells und seiner Parameter entfallen sämtliche Filter- und sonstigen Maßnahmen des KI-Systems. In diesem Sinne werden sog. White Box Attacks,[184] abzugrenzen von den Black Box Attacks im vorgenannten Szenario,[185] durch die Kenntnis der Modellparameter durch jedermann möglich. Vor diesem Hintergrund steigt die Wahrscheinlichkeit einer Anwendbarkeit der DSGVO signifikant und ermittelt sich (fast) nur noch in Abhängigkeit davon, wie erfolgreich die

183 *Ashkar*, ZD 2023, 523 (530); *DSK*, Orientierungshilfe Künstliche Intelligenz und Datenschutz, Rn. 17.

184 *Fredrikson/Jha/Ristenpart*, in: Ray/N. Li/Kruegel, 1322 (1322 f.); *C. Song/Ristenpart/Shmatikov*, in: Thuraisingham/Evans/Malkin/Xu, 587 (587 f.).

185 *Ico*, https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/how-should-we-assess-security-and-data-minimisation-in-ai/; *Kaulartz*, in: Kaulartz/Braegelmann, Kap. 8.9, Rn. 7.

Trainingsdaten anonymisiert wurden. Eine datenschutzrechtliche Rechtsgrundlage für die Verarbeitung aus Art. 6 DSGVO kommt vor allem dann in Betracht,[186] wenn enthaltene Gerichtsentscheidungen in dieser Form ohnehin hätten veröffentlicht werden dürfen.[187]

c. Veröffentlichung des Trainingskorpus (Open Source)

Sofern darüber hinaus auch der gesamte Trainingskorpus veröffentlicht wird, hängt die Anwendbarkeit der DSGVO (nur) noch davon ab, ob die Trainingsdaten selbst als anonym zu betrachten sind. Da das Erlanger Tool keine 100%ige Anonymisierung gewährleisten kann und zugleich die gesamte Öffentlichkeit den Zugriff auf die gerichtlichen Entscheidungen hat, sprechen gewichtige Gründe für die Annahme einer Anwendbarkeit der DSGVO. Gleichwohl ist auch in diesem Fall eine Rechtfertigung der betreffenden Datenverarbeitung nach Art. 6 DSGVO (theoretisch) denkbar.[188] Das betrifft insbesondere Gerichtsentscheidungen, die den allgemeinen (hohen) Anforderungen an die Veröffentlichung entsprechen.[189]

V. Datenschutzrechtliche Anforderungen im Fall des Vorliegens personenbezogener Daten

Wenn und soweit personenbezogene Daten verarbeitet werden, was im GSJ-Projekt mit Blick auf das trainierte Sprachmodell bei Umsetzung hinreichender technischer und organisatorischer Maßnahmen nicht der Fall ist (s. hierzu unter D.IV.1.b.ff.), ergeben sich aus der DSGVO entsprechende Anforderungen an diese Verarbeitung.

1. Relevante Verarbeitungen

Die datenschutzrechtlichen Anforderungen knüpfen an die Verarbeitung als „jeden mit oder ohne Hilfe automatisierter Verfahren ausgeführten Vorgang oder jede solche Vorgangsreihe im Zusammenhang mit personen-

186 S. nachfolgend unter D.V.4.c.cc.
187 S. hierzu unter D.II.5.b.
188 S. nachfolgend unter D.V.4.c.cc.
189 S. hierzu unter D.II.5.b.

bezogenen Daten“ nach Art. 4 Nr. 2 DSGVO. Der nicht-abschließende Katalog der Vorschrift nennt beispielhaft Vorgänge wie das Erheben und Erfassen personenbezogener Daten, wobei die konkrete Einordnung vielfach offenbleiben kann.

Für das untersuchungsgegenständliche Projekt in Betracht kommen insbesondere Verarbeitungen im Zusammenhang mit der (teilweisen) Anonymisierung, dem Training, einer etwaigen Speicherung im Sprachmodell und dem Einsatz des Sprachmodells und der Ausgabe.

2. Verhältnis zum mitgliedstaatlichen Recht

Von der DSGVO abweichende Anforderungen nach mitgliedstaatlichem Recht können Bedeutung erlangen, wenn und soweit sich das mitgliedstaatliche Recht auf eine Öffnungsklausel innerhalb der DSGVO stützen kann.

Hierzu kommen neben den nachfolgend unter D.V.4 zu erörternden Art. 6 Abs. 1 lit. c, e, Abs. 2, 3, 4 DSGVO insbesondere in Betracht: Art. 23 Abs. 1 lit. f DSGVO mit Blick auf Einschränkungen von Betroffenenrechten zum Schutz der Unabhängigkeit der Justiz und des Schutzes der Gerichtsverfahren, Art. 85 Abs. 1, 2 DSGVO mit Blick auf weitgehende Abweichungen verschiedener DSGVO-Kapitel zur Wahrung der Meinungs- und Wissenschaftsfreiheit, Art. 86 DSGVO mit Blick auf Einschränkungen der DSGVO unter Wahrung des Wesenskerns[190] für den Zugang zu amtlichen Dokumenten[191] sowie Art. 89 DSGVO mit Blick auf weitgehende Einschränkungen der DSGVO-Vorschriften für wissenschaftliche Forschungszwecke.

Allerdings setzen diese Vorschriften jeweils Unions- oder mitgliedstaatliches Recht[192] voraus, das von einer dieser Öffnungsklauseln für die Verarbeitung personenbezogener Daten im Zusammenhang mit dem Training eines Sprachmodells Gebrauch macht. Eine solche (einschlägige) unionsrechtliche oder mitgliedstaatliche spezifische Regelung ist vorliegend aber gerade nicht ersichtlich.

Die KI-VO lässt die DSGVO ebenfalls nach Art. 2 Abs. 7 KI-VO unberührt (siehe auch ErwGr. 10 KI-VO).

190 *Schnabel*, in: Simitis/Hornung/Spiecker gen. Döhmann, Art. 86 DSGVO Rn. 30 ff.
191 Zu Gerichten als Behörden etwa *Schnabel*, in: Simitis/Hornung/Spiecker gen. Döhmann, Art. 86 DSGVO Rn. 20.
192 *Buchner/Petri*, in: Kühling/Buchner, Art. 6 DSGVO Rn. 84.

3. Verantwortlichkeit

Die Adressaten des Pflichtenkanons der DSGVO richten sich maßgeblich nach der datenschutzrechtlichen Verantwortlichkeit.

a. Anforderungen

Soweit die DSGVO zur Anwendung berufen ist, trifft der umfangreiche Pflichtenkatalog grundsätzlich den für die Verarbeitung – allein oder gemeinsam – Verantwortlichen. Der Verantwortliche ist nach Art. 4 Nr. 7 Hs. 1 DSGVO „die natürliche oder juristische Person, Behörde, Einrichtung oder andere Stelle, die allein oder gemeinsam mit anderen über die Zwecke und Mittel der Verarbeitung von personenbezogenen Daten entscheidet".

Die Entscheidung oder Festlegung setzt regelmäßig die tatsächliche[193] Einflussnahme und Steuerung der Verarbeitung aus Eigeninteresse voraus.[194] Die Festlegung erfordert hierbei ausgeübte Entscheidungsgewalt mit Blick auf die wesentlichen Umstände der Verarbeitung,[195] wobei diese Festlegung zu einem gewissen Grad abstrakt bleiben kann.[196] Bezugspunkt für die Entscheidung müssen die Zwecke und die wesentlichen Mittel der Verarbeitung sein, sprich die Kategorien personenbezogener Daten und Kategorien betroffener Personen sowie die Speicherdauer und etwaige Empfänger.[197] Eines sicheren Wissens um das Vorliegen personenbezogener Daten bedarf es hierbei nicht,[198] ebenso wenig wie eines tatsächlichen Zugriffs auf personenbezogene Daten durch einen Verantwortlichen.[199]

Wenn und soweit mehrere Verantwortliche gemeinsam entscheiden, liegt eine gemeinsame Verantwortlichkeit (nur) für den konkreten Verarbeitungsvorgang vor,[200] wobei diese gemeinsame Verantwortlichkeit zusätz-

193 Abzugrenzen von der bloßen formalen Stellung *EDPB*, Leitlinien 07/2020 zu den Begriffen „Verantwortlicher" und „Auftragsverarbeiter" in der DSGVO, Rn. 12.
194 *EuGH*, ZD 2018, 469 (Rn. 68) – Zeugen Jehovas.
195 *EDPB*, Leitlinien 07/2020 zu den Begriffen „Verantwortlicher" und „Auftragsverarbeiter" in der DSGVO, Rn. 40.
196 *T. Radtke*, Gemeinsame Verantwortlichkeit unter der DSGVO, S. 133 ff.
197 *EDPB*, Leitlinien 07/2020 zu den Begriffen „Verantwortlicher" und „Auftragsverarbeiter" in der DSGVO, Rn. 40.
198 Vielmehr genügt die abstrakte Kenntnis um stattfindende Verarbeitungen, s. *T. Radtke*, Gemeinsame Verantwortlichkeit unter der DSGVO, S. 135 ff.
199 *EuGH*, NJW 2018, 2537 (Rn. 38) – Wirtschaftsakademie.
200 *EuGH*, ZD 2019, 455 (Rn. 74 ff.) – Fashion ID.

liche Pflichten und eine gesamtschuldnerische Haftung nach Art. 26 Abs. 3, Art. 82 Abs. 4 DSGVO auslöst. Indizien für die Annahme einer gemeinsamen Verantwortlichkeit sind nach der weiten Auslegung der maßgeblichen Bestimmungen durch den EuGH für betroffene Personen nicht erwartbare Übermittlungen sowie sich ergänzende[201] Verarbeitungsbeiträge und Zwecke bezogen auf gemeinsame Datensätze bei gleichzeitigen Elementen der Kooperation.[202]

Die Annahme einer Verantwortlichkeit ist insbesondere dann zu rechtfertigen, wenn sich aufgrund der Beteiligung des betreffenden Akteurs das bestehende datenschutzrechtliche Risiko für betroffene Personen erhöht oder ein neues Risiko entsteht (z.B. durch Zweckfestlegungen oder die Änderung wesentlicher Mittel, wie etwa der Speicherdauer oder der Datenempfänger). Vor diesem Hintergrund stuft das EDPB etwa die Verbesserung eines bereits existierenden Dienstes aus einem Nutzungsverhältnis und die Betrugsprävention als eigenständige Zwecke ein,[203] deren Festlegung eine (gemeinsame) Verantwortlichkeit indizieren kann.

In den Fällen des Art. 6 Abs. 1 UAbs. 1 lit. c, e DSGVO kann die Verantwortlichkeit durch Rechtsvorschrift nach Art. 4 Nr. 7 Hs. 2 DSGVO festgelegt werden.[204]

Eine datenschutzrechtliche Stelle, die nur punktuell Verarbeitungen für einen Verantwortlichen und nach dessen Weisungen ausführt, unterliegt als Auftragsverarbeiter nach Art. 4 Nr. 8, Art. 26 DSGVO einem nur eingeschränkten Pflichtenkatalog. Insbesondere bedarf die Übermittlung personenbezogener Daten zwischen Verantwortlichem und Auftragsverarbeiter keiner datenschutzrechtlichen Rechtsgrundlage nach Art. 6, 9 DSGVO.[205]

b. Einordnung der Rollen der Ministerien, der ausführenden Stellen sowie der anwendenden Justizbediensteten und Richter

In dem untersuchungsgegenständlichen GSJ-Projekt bedarf es insbesondere einer Einordnung der Rollen der Ministerien, der ausführenden Stellen sowie der ein trainiertes KI-System anwendenden Gerichte. Hierbei ist

201 Vgl. *EuGH*, ZD 2024, 209 (Rn. 43); *EDPB*, Leitlinien 07/2020 zu den Begriffen „Verantwortlicher" und „Auftragsverarbeiter" in der DSGVO, Rn. 55.
202 *EuGH*, NJW 2018, 2537 (Rn. 31 ff.) – Wirtschaftsakademie; ZD 2024, 328 (Rn. 59 ff.) – IAB Europe.
203 *EDPB*, Leitlinien 02/2019, Rn. 48-50.
204 *T. Radtke*, Gemeinsame Verantwortlichkeit unter der DSGVO, S. 105 f.
205 Ausf. *Gabel/Lutz*, in: Taeger/Gabel, Art. 28 DSGVO Rn. 11 m.w.N.

zwischen mehreren Projektphasen mit unterschiedlichen Verarbeitungsvorgängen zu differenzieren.

Anders als in der Konstellation von sog. aufgedrängten personenbezogenen Daten[206] wird es vorliegend in Ansehung des abstrakten Wissens der Ministerien um stattfindende Verarbeitungen unschädlich sein, dass die Verarbeitung personenbezogener (Trainings-)Daten gerade nicht gewollt ist.

aa. (Teilweise) Anonymisierung

Eine (teilweise) Anonymisierung stellt ebenfalls eine Verarbeitung dar.[207] Die Entscheidung über die Anonymisierung steht in engem Zusammenhang mit der späteren Verwendung, die durch die Anonymisierung ermöglicht werden soll. Für eine Anonymisierung, die aus der ex ante-Perspektive ausschließlich die Verwendung zum KI-Training und dem späteren Einsatz in der Justiz ermöglichen soll und dementsprechend durch ein Ministerium im Interesse der eigenen Justiz beauftragt wird, ist das jeweilige Ministerium der datenschutzrechtlich Verantwortliche. Denn das Ministerium entscheidet in diesen Fällen nicht nur über den Zweck, sondern auch über die heranzuziehende Datengrundlage in Form der Entscheidung, mithin auch über Kategorien der personenbezogenen Daten (z.B. über den Ausschluss von Arzthaftungssachen mit möglichen Gesundheitsdaten), sowie den Kreis und die Kategorien betroffener Personen.

Die ausführenden Stellen können vertraglich an die Weisungen der verantwortlichen Ministerien gebunden und sodann als Auftragsverarbeiter nach Art. 4 Nr. 8, Art. 28 DSGVO eingesetzt werden. Soweit die ausführenden Stellen allerdings personenbezogene Daten in Abweichung von den Weisungen oder aus Eigeninteresse an einer weiteren Forschungsnutzung personenbezogene Daten verarbeiten, werden sie selbst als Verantwortliche anzusehen sein bzw. gelten als solche (vgl. Art. 28 Abs. 10 DSGVO). Die ausführenden Stellen müssen in der Folge insbesondere entsprechend der jeweiligen Auftragsverarbeitungsvereinbarung hinreichende technische und organisatorische Maßnahmen zur Datensicherheit treffen.

206 Hierzu etwa *L. Meyer*, RDi 2025, 125.

207 S. schon oben unter D.II.5.b zur Veröffentlichung teilanonymisierter Gerichtsentscheidungen sowie zur Anonymisierung als Verarbeitung nachfolgend unter D.V.4.c.aa.

bb. KI-Training

Ähnliche Erwägungen gelten mit Blick auf das KI-Training einschließlich der vorbereitenden Zusammenstellung der Trainingsdaten. Wenn und soweit das Training einschließlich des notwendigen Debuggings allein für den späteren Einsatz in der Justiz erfolgt, werden die Ministerien als Verantwortliche und die ausführenden Stellen als Auftragsverarbeiter zu qualifizieren sein. Ein nach eigenem Ermessen der ausführenden Stellen durchgeführtes Debugging ändert nichts an dem Risiko der Verarbeitungsvorgänge für die betroffenen Personen und bewegt sich als Entscheidung über unwesentliche Mittel der Verarbeitung zudem im Rahmen des Spielraums eines Auftragsverarbeiters, soweit und solange die Ministerien den Zweck des KI-Trainings für das GSJ-Sprachmodell sowie die Speicherdauer, Datenempfänger und die Kategorien der betroffenen Personen selbst vorgeben.

Da beide Ministerien allfällige Gerichtsentscheidungen übermitteln und den Einsatz des trainierten KI-Modells in ihrem Verantwortungsbereich beabsichtigen, kommt mit Blick auf das Training eine gemeinsame Verantwortlichkeit in Betracht. Sobald sich die personenbezogenen Daten aus unterschiedlichen Gerichtsentscheidungen im Rahmen des Trainings des Sprachmodells nicht (mehr) klar voneinander trennen lassen, sondern ein gemeinsamer Datenpool im Interesse beider Ministerien entsteht, wird eine gemeinsame Verantwortlichkeit der Ministerien anzunehmen sein. Somit bedarf es einer Verantwortlichkeitszuteilung zwischen beiden Ministerien, die den Anforderungen des Art. 26 Abs. 1 S. 2, 3, Abs. 2 S. 1 DSGVO entspricht und deren wesentlicher Inhalt den betroffenen Personen nach Art. 26 Abs. 2 S. 2 DSGVO zur Verfügung gestellt wird. Außerdem resultiert hieraus eine gesamtschuldnerische (zivilrechtliche) Haftung (Art. 26 Abs. 3, Art. 82 Abs. 4 DSGVO).

Die gemeinsame Beauftragung eines Auftragsverarbeiters durch die Ministerien ist aufgrund der gemeinsamen Verantwortlichkeit mit zusammenhängenden Interessen möglich,[208] wenn und soweit sich die Ministerien auf einheitliche Weisungen verständigen.

208 *T. Radtke*, Gemeinsame Verantwortlichkeit unter der DSGVO, S. 416.

cc. Speicherung des KI-Modells

Die Speicherung des KI-Modells ist datenschutzrechtlich (nur) relevant, wenn und soweit Daten nach allgemeinem Ermessen einer natürlichen Person zugeordnet werden können.[209] Selbst wenn sich in den Trainingsdaten noch personenbezogene Daten befinden, kann das Training als Anonymisierung dergestalt wirken, dass die betreffenden Daten nicht (mehr) mit hinreichender Wahrscheinlichkeit zusammengeführt und einer natürlichen Person zugeordnet werden können. Hierzu trägt nicht nur das Training selbst bei, sondern auch allfällige Maßnahmen auf Ebene der Ausgabe (z.B. Ausgabefilter).

Falls eine solche Zuordnung möglich ist, enthält das Sprachmodell – eingebettet in das KI-System – personenbezogene Daten und es stellt sich die Frage nach der Verantwortlichkeit für die Speicherung des Modells als Verarbeitung personenbezogener Daten. Die Speicherung dürfte regelmäßig durch ein Ministerium als Verantwortlichen[210] unter Einbindung eines Rechenzentrumsanbieters als Auftragsverarbeiter gesteuert werden. Ein gemeinsamer Betrieb eines KI-Systems durch beide Ministerien, die zuvor bereits die Trainingsdaten zur Verfügung gestellt haben und das KI-System nun zu eigenen Zwecken einsetzen, wird wiederum eine gemeinsame Verantwortlichkeit darstellen.

dd. Einsatz des KI-Systems

Durch den Einsatz des KI-Systems kann es sowohl zu einer Verarbeitung personenbezogener (Trainings-)Daten kommen, wenn bereits das Sprachmodell selbst einen Personenbezug aufweist, als auch wenn das Sprachmodell mangels Wahrscheinlichkeit der Identifizierungen keinen Personenbezug aufweist, ein solcher unwahrscheinlicher Fall allerdings aufgrund der konkreten Eingabeaufforderung im Einzelfall auftritt.

In der ersten Konstellation setzt sich die – gegebenenfalls gemeinsame – Verantwortlichkeit des Ministeriums fort. Denn das Ministerium hat hier maßgeblich über den Einsatz des Sprachmodells mit Personenbezug sowie die Zusammensetzung der (personenbezogenen) Trainingsdaten entschieden. Hinzutreten können weitere Verantwortliche aus dem Nutzerkreis

209 ErwGr. 26 S. 3 DSGVO sowie s. oben unter D.I.1.b.
210 Zugleich als Anbieter des KI-System i.S.d. KI-VO.

(z.B. das jeweilige Gericht[211]). Soweit diese weiteren Verantwortlichen allerdings eine Eingabeaufforderung übergeben, die gerade keine Ausgabe personenbezogener Trainingsdaten erwarten lässt, wird die Bewertung für die Verarbeitung dieser sog. aufgedrängten personenbezogenen Daten[212] anders auszufallen haben. Denn den Nutzern fehlt in diesem Fall das kognitive Element für eine Entscheidung über die Verarbeitung und die Verantwortlichkeit verbleibt deshalb bei den Ministerien.

In der zweiten Konstellation gleicht die Bewertung grundsätzlich den vorherigen Ausführungen. In dieser Konstellation liegt eine Verantwortlichkeit des Nutzers allerdings näher. Dieser Befund betrifft zumindest die aufgrund der Eingabeaufforderung erwartbare oder gezielte Extraktion von personenbezogenen Trainingsdaten durch eine entsprechende Eingabeaufforderung. Für eine erwartbare Ausgabe personenbezogener Trainingsdaten ist regelmäßig das Vorliegen einer gemeinsamen Verantwortlichkeit anzunehmen, denn es wirkt sich weiterhin die Entscheidung zur Zusammenstellung der Trainingsdaten durch die Ministerien im Zusammenspiel mit der konkreten Eingabeaufforderung aus.

Für eine gezielte, missbräuchliche Extraktion personenbezogener Trainingsdaten durch eine Eingabeaufforderung kann darüber hinaus sogar eine alleinige Verantwortlichkeit des Nutzers anzunehmen sein. Denn wenn der Nutzer auf die Extraktion konkreter Datenkategorien und Kategorien betroffener Personen unter Umgehung etwaiger Schutzmaßnahmen abzielt, die er zu eigenen Zwecken aus dem Sprachmodell extrahiert, legt der Nutzer nicht nur einen eigenen Zweck fest, sondern verengt auch die wesentlichen Mittel der Verarbeitung erheblich. Zusammengenommen ist die Verarbeitung daher nur noch auf die Entscheidung des Nutzers zurückzuführen und dient nicht (mehr) dem Zweck des Betriebs des Sprachmodells.

ee. Veröffentlichung des KI-Modells oder KI-Systems

Im Fall der Veröffentlichung des KI-Modells oder KI-Systems[213] verbleibt es mit Blick auf die Trainingsdaten bei einer (gemeinsamen) Verantwortlichkeit der veröffentlichenden Ministerien. Mit Blick auf die Nutzer gelten

211 *Schild*, in: BeckOK Datenschutzrecht, Syst. E. Rn. 9.

212 Ausf. zu den Lösungsansätzen in diesem Fall *L. Meyer*, RDi 2025, 125; *T. Radtke*, in: Gersdorf/Paal, Art. 6 DSA Rn. 14 f.

213 S. zu den drei verschiedenen Stufen unter D.IV.3.

die zuvor gemachten Ausführungen entsprechend. Soweit sich die Nutzer im Rahmen der Privathaushaltsausnahme bewegen (Art. 2 Abs. 2 lit. c DSGVO) oder Nutzer aus deren Sicht nicht zu erwartende Ausgaben personenbezogener Trainingsdaten als sog. aufgedrängte personenbezogene Daten erhalten, kommen diese Nutzer schon gar nicht als Verantwortliche in Betracht.

Im Fall der gezielten Extraktion personenbezogener Trainingsdaten durch entsprechende Eingabeaufforderungen kommt allerdings eine (gemeinsame) Verantwortlichkeit der Nutzer in Betracht.

4. Rechtsgrundlage nach Art. 6 DSGVO

Für die vorgenannten Verarbeitungsvorgänge bedarf es nach Art. 6 Abs. 1 DSGVO jeweils einer Rechtsgrundlage.

a. Einschränkungen aufgrund des Vorbehalts des Gesetzes (Art. 6 Abs. 1 UAbs. 2 DSGVO)

Nach Art. 6 Abs. 1 UAbs. 2 DSGVO können sich Behörden „in Erfüllung ihrer Aufgaben" nicht auf berechtigte Interessen nach Art. 6 Abs. 1 UAbs. 1 lit. f DSGVO berufen, sondern vielmehr nur auf den Rechtfertigungskatalog des Art. 6 Abs. 1 UAbs. 1 DSGVO im Übrigen. Der europäische Gesetzgeber begründet dies damit, dass es über Art. 6 Abs. 1 lit. c, e DSGVO „dem Gesetzgeber obliegt, per Rechtsvorschrift die Rechtsgrundlage für die Verarbeitung personenbezogener Daten durch die Behörden zu schaffen" (ErwGr. 47 S. 5 DSGVO). Die Vorschrift soll einfachgesetzlich den Vorbehalt des Gesetzes sichern[214] und versperrt insoweit auch die Berücksichtigung der behördlichen Interessen durch Dritte im Rahmen des Art. 6 Abs. 1 lit. f DSGVO[215] sowie der Interessen Dritter durch Behörden.[216]

Der Begriff der öffentlichen Aufgaben ist unionsautonom auszulegen und muss sich eindeutig, insbesondere aufgrund hoheitlicher Befugnisse, von

214 *Albers/Veit*, in: BeckOK Datenschutzrecht, Art. 6 DSGVO Rn. 65; *Schulz*, in: Gola/Heckmann, Art. 6 DSGVO Rn. 60; *Buchner/Petri*, in: Kühling/Buchner, Art. 6 DSGVO Rn. 157; *Schantz*, in: Simitis/Hornung/Spiecker gen. Döhmann, Art. 6 Abs. 1 DSGVO Rn. 97.

215 *Heberlein*, in: Ehmann/Selmayr, Art. 6 DSGVO Rn. 38.

216 *Schulz*, in: Gola/Heckmann, Art. 6 DSGVO Rn. 60.

einem Verhältnis zwischen Privaten abgrenzen.[217] Dieses Verhältnis umfasst neben der Eingriffsverwaltung auch Aufgaben der Leistungsverwaltung sowie die Öffentlichkeitsarbeit,[218] nicht dagegen ein privatrechtliches Handeln der Behörde.[219] Öffentliche Stellen, die keine Behörde sind, werden von der Vorschrift nicht adressiert.[220]

Es ist nicht abschließend geklärt, ob auch staatliche Hochschulen[221] und Stellen mit judikativer Funktion[222] dem Begriff der Behörde unterfallen. Für ein weites Verständnis mit Blick auf die hier besonders relevanten judikativen Stellen lässt sich in systematischer Hinsicht anführen, dass Art. 6 Abs. 1 UAbs. 2 DSGVO solche Konstellationen erfassen soll, in denen Art. 6 Abs. 1 UAbs. 1 lit. e DSGVO einschlägig ist; das betrifft auch die Gerichte.[223] Allerdings differenziert die DSGVO an anderen Stellen gerade zwischen Behörden und Gerichten (ErwGr. 97 S. 1 DSGVO)[224] und sieht Sonderregelungen für Gerichte vor, wie etwa mit Blick auf die Rechtsgrundlage (Art. 9 Abs. 2 lit. f DSGVO) und die Zuständigkeit der Aufsichtsbehörden (Art. 55 Abs. 3, ErwGr. 20 DSGVO).[225]

Diese systematischen Argumente und die auch durch die DSGVO anerkannte Unabhängigkeit der Justiz legen nahe, dass für die Kerntätigkeit der Justiz in Form der unabhängigen Rechtsprechung entweder bereits regelmäßig eine spezifische mitgliedstaatliche Rechtsgrundlage vorliegen dürfte oder sich das Gericht jedenfalls gegebenenfalls (auch) auf berechtigte Interessen nach Art. 6 Abs. 1 UAbs. 1 lit. f DSGVO stützen kann. Diese Feststellung wird allerdings gerade nur die Kerntätigkeit einer unabhängigen Justiz umfassen, sprich eine Verarbeitung im Rahmen der Rechtsprechung.[226]

Ein durch ein Ministerium organisiertes Projekt zur Arbeitserleichterung wirkt sich zwar auf die (vorbereitende) Rechtsprechungstätigkeit aus, be-

217 *Heberlein*, in: Ehmann/Selmayr, Art. 6 DSGVO Rn. 39.
218 *BVerwG*, BeckRS 2018, 35172 (Rn. 26); *Heberlein*, in: Ehmann/Selmayr, Art. 6 DSGVO Rn. 39; *Schulz*, in: Gola/Heckmann, Art. 6 DSGVO Rn. 60.
219 *Schulz*, in: Gola/Heckmann, Art. 6 DSGVO Rn. 60; einschränkend *Reimer*, in: Sydow/Marsch, Art. 6 DSGVO Rn. 91; *Taeger*, in: Taeger/Gabel, Art. 6 DSGVO Rn. 122.
220 *Heberlein*, in: Ehmann/Selmayr, Art. 6 DSGVO Rn. 39.
221 Bejahend *Reimer*, in: Sydow/Marsch, Art. 6 DSGVO Rn. 59, 90 ff. m.w.N.
222 Bejahend *Reimer*, in: Sydow/Marsch, Art. 6 DSGVO Rn. 90; implizit die andere Ansicht zugrunde legend *Freye/Schnebbe*, ZD 2020, 502 (504); ebenfalls a.A. *Labusga/Petit*, NJW 2022, 300 (Rn. 21 ff.).
223 *Reimer*, in: Sydow/Marsch, Art. 6 DSGVO Rn. 90.
224 *Labusga/Petit*, NJW 2022, 300 (Rn. 22).
225 *Labusga/Petit*, NJW 2022, 300 (Rn. 22).
226 Vgl. *Labusga/Petit*, NJW 2022, 300 (Rn. 23).

trifft aber nicht den Kernbereich unabhängiger Rechtsprechung in einem konkreten Verfahren – und kann daher als möglicher Bestandteil der allgemeinen öffentlichen Aufgabenerfüllung wegen Art. 6 Abs. 1 UAbs. 2 DSGVO nicht auf Art. 6 Abs. 1 UAbs. 1 lit. f DSGVO gestützt werden.

Die Veröffentlichung des trainierten Sprachmodells fällt erst recht nicht in den Kerntätigkeitsbereich einer unabhängigen Justiz, stellt aber zugleich auch noch kein privatrechtliches, den Art. 6 Abs. 1 UAbs. 2 DSGVO ausschließendes Handeln dar. Denn der Zugriff auf Trainingsdaten in der vorliegend in Rede stehenden Quantität ist gerade Ausfluss der öffentlichrechtlichen Stellung der Justiz und der hiermit verbundenen Möglichkeit, einen Zugriff auf unveröffentlichte Gerichtsentscheidungen zu nehmen.

b. Überblick über die Rechtsgrundlagen aus Art. 6 Abs. 1 UAbs. 1 DSGVO

Für die öffentlichen Stellen scheiden im hier untersuchungsgegenständlichen Szenario neben Art. 6 Abs. 1 UAbs. 1 lit. f DSGVO (berechtigte Interessen)[227] auch die Rechtsgrundlagen aus lit. a (Einwilligung)[228] aus praktischen Gründen bzw. aus lit. b (Vertragserfüllung) mangels eines abgeschlossenen Vertrags mit der betroffenen Person aus. In Betracht kommen also vor allem lit. c (Erfüllung einer rechtlichen Verpflichtung) und lit. e (Erfüllung einer öffentlichen Aufgabe), die insbesondere über das Merkmal der Berechtigung bzw. Verpflichtung voneinander abzugrenzen sind.[229]

c. Art. 6 Abs. 1 UAbs. 1 lit. e DSGVO (öffentliches Interesse)

Für das dieser Untersuchung zugrundeliegende GSJ-Projekt ist vor allem an die Erfüllung einer Aufgabe im öffentlichen Interesse i.S.d. lit. e zu denken, konkret einerseits betreffend die (Teil-)Anonymisierung, die auch der öffentlichen Aufgabe an der Veröffentlichung der Gerichtsentscheidungen dient, und andererseits betreffend das KI-Training und den Einsatz

227 Hierzu etwa *Paal*, ZfDR 2024, 129 (147 ff.); *Dewitte*, in: AI Meets the GDPR, S. 146 ff.; *Dieker*, ZD 2024, 132 (134 f.); *Schäfer*, ZD 2025, 12 (14); *Golland*, EuZW 2024, 846 (848 f.); *Ashkar*, ZD 2023, 523 (525 f.); *BayLDA*, Datenschutzkonforme Künstliche Intelligenz, S. 9; *Engeler/Rolfes*, ZD 2024, 423 (425 f.); *Mertens/D. Meyer*, K&R 2023, 563 (565 ff.); *Hüger*, Künstliche Intelligenz und Diskriminierung, S. 329 ff.

228 *Ashkar*, ZD 2023, 523 (525).

229 Vgl. *Albers/Veit*, in: BeckOK Datenschutzrecht, Art. 6 DSGVO Rn. 49.

des KI-Systems, die der effektiven Ausübung der Rechtsprechungstätigkeit dienen könnten.

Grundsätzlich kann es sich bei diesen Betätigungen um Aufgaben im öffentlichen Interesse[230] i.S.d. lit. e (siehe auch Ziff. 1.2.1 Verwaltungsvorschrift) handeln. Die Veröffentlichungspflicht[231] im Hinblick auf Gerichtsentscheidungen könnte insoweit eine Pflicht zur Verarbeitung personenbezogener Daten im Wege der Anonymisierung und Veröffentlichung i.S.d. lit. c implizieren;[232] nimmt man eine solche Verarbeitungspflicht an, gelten die nachfolgend dargestellten Maßgaben entsprechend. In beiden Konstellationen sind die ergänzenden Anforderungen an die Bestimmtheit der Rechtsgrundlage aus Art. 6 Abs. 3 DSGVO zu beachten.[233] Danach bedarf es einer Rechtsvorschrift mit festgelegtem Datenverarbeitungsbezug, die „klar und präzise und [deren] Anwendung […] für die Rechtsunterworfenen […] vorhersehbar" ist (ErwGr. 41 S. 2 DSGVO).

In Anwendung des Merkmals der Erforderlichkeit und der Anforderungen an den Zusammenhang mit einer Aufgabe im öffentlichen Interesse letztlich eine Interessenabwägung durchzuführen (hierfür spricht auch die Gleichstellung von lit. e und f in Art. 21 Abs. 1 S. 1 DSGVO). Je geringer die Auswirkungen auf die betroffenen Personen sind, desto geringer fallen die Anforderungen an den Zusammenhang zur öffentlichen Aufgabe aus.

aa. (Teil-)Anonymisierung der Gerichtsentscheidungen

Im GSJ-Projekt stellt sich insoweit die Frage, ob die Anonymisierung auch dann für eine Aufgabe im öffentlichen Interesse erforderlich ist, wenn die Anonymisierung zwar auch der späteren Veröffentlichung dienen kann, aber im Zuge der Vorbereitung für ein KI-Training zur späteren Optimierung der Arbeitsabläufe der Justiz erfolgt.

An dieser Stelle gilt es, sich zunächst in Erinnerung zu rufen, dass es sich auch bei einer Anonymisierung um eine grundsätzlich rechtfertigungs-

230 *BVerfG*, NJW 1997, 2694 (2694); *OVG Lüneburg*, NJW 1996, 1489 (1489).

231 *BVerfG*, NJW 1997, 2694; *BVerwG*, NJW 1997, 2694; *BGH*, NJW 2018, 3123 (Rn. 14): „die aus dem Rechtsstaatsgebot einschließlich der Justizgewährungspflicht, dem Demokratiegebot und dem Grundsatz der Gewaltenteilung folgt"; *OLG Celle*, NStZ 1990, 553; Überblick auch bei *Ludyga*, ZUM 2021, 887 (887 f.); *Köhnlein*, in: Graf, § 23 EGGVG Rn. 181.

232 *VG Stuttgart*, ZD 2022, 583 (Rn. 33).

233 Zum Verhältnis von Abs. 2 und 3 *Roßnagel*, in: Simitis/Hornung/Spiecker gen. Döhmann, Art. 6 Abs. 2 DSGVO Rn. 7 ff.

bedürftige Verarbeitung handelt.[234] In Ansehung der datenschutzrechtlich gerade erwünschten Folgen einer Anonymisierung dürften die Anforderungen an eine solche Rechtfertigung (z.B. nach Art. 6 Abs. 1 UAbs. 1 lit. c, e, f DSGVO) sehr gering ausfallen, unabhängig davon, welcher Lösungsweg konkret gewählt wird.[235]

Im GSJ-Projekt ist für die Anonymisierung zunächst der konkrete Zweck der Verarbeitung zur Anonymisierung zu identifizieren, um zu prüfen, ob auch dieser Zweck der Aufgabenerfüllung i.S.d. Art. 6 Abs. 1 UAbs. 1 lit. e DSGVO zuzuordnen ist.

Die eindeutige Zweckfestlegung nach Art. 5 Abs. 1 lit. b DSGVO erfordert eine möglichst kleinschrittige Betrachtung,[236] losgelöst von Fernzielen,[237] wie etwa dem späteren Einsatz als Trainingsdatum für ein Sprachmodell. Der vorliegend in Rede stehende Zweck ist vor allem die Anonymisierung als Aufbereitung der Entscheidung zur Weiterverarbeitung. Die weiteren Verarbeitungen wären – im Fall einer erfolgreichen Anonymisierung – sodann gerade keine Verarbeitungen i.S.d. DSGVO (mehr) und daher nicht zu berücksichtigen. Diese Entscheidungsaufbereitung als Voraussetzung der Weiterverarbeitung der amtlichen Entscheidungen in jeglicher Form (z.B. in der Rechtsanwendung) steht in unmittelbarem Zusammenhang mit der öffentlichen Aufgabe. Einzelne Gerichte sehen dementsprechend auch bei einer nur teilweisen Anonymisierung von Entscheidungen einen Zusammenhang mit einer öffentlichen Aufgabe.[238]

Eine abweichende Auffassung lässt sich darauf stützen, dass vorliegend eben gerade nicht mehr zwangsläufig die Justiz selbst, sondern gegebenenfalls die ausführenden Stellen die Anonymisierung vornehmen. Eine solche Auslagerung ist zwar im Rahmen der Auftragsverarbeitung und Zurechnung zu dem jeweiligen Verantwortlichen unschädlich, wäre aber anders zu bewerten im Fall einer gemeinsamen Verantwortlichkeit zwischen den Ministerien und den ausführenden Stellen.

---

234 *Artikel-29-Datenschutzgruppe*, WP 216, S. 7 f.; *Paal*, ZfDR 2024, 129 (136) m.w.N.; bejahend *Burghoff*, ZD 2023, 658 (660); *BfDI*, Positionspapier zur Anonymisierung unter der DSGVO unter besonderer Berücksichtigung der TK-Branche, S. 5 ff.; *Hornung/B. Wagner*, ZD 2020, 223; *Roßnagel/Geminn*, ZD 2021, 487; *Stürmer*, ZD 2020, 626 (629); ablehnend hingegen *Thüsing/Rombey*, ZD 2021, 548. S. auch ErwGr. 61 a.E. KI-VO.

235 Zu den Lösungswegen im Einzelnen *Paal*, ZfDR 2024, 129 (136) m.w.N.

236 Ausf. *Jansen/T. Radtke*, PinG 2024, 61 (63) m.w.N.

237 Selbst das Fernziel ist zumindest von der DSGVO anerkannt, etwa indem Art. 23 Abs. 1 lit. f DSGVO auch auf „eine ordnungsgemäße Rechtspflege abzielt", s. *EuGH*, ZD 2023, 396 (Rn. 38) – Norra Stockholm Bygg AB.

238 *VG Stuttgart*, ZD 2022, 583.

Die teilweise Anonymisierung in Vorbereitung einer Weiterverarbeitung der Gerichtsentscheidung ist also im Rahmen des Art. 6 Abs. 1 UAbs. 1 lit. e DSGVO grundsätzlich zulässig. Hieraus lässt sich aber noch nicht bereits ableiten, dass auch eine Veröffentlichung der mittels Erlanger Tools anonymisierten Entscheidungen mit gegebenenfalls geringerem Anonymisierungsgrad gegenüber einer händischen Anonymisierung zulässig ist.[239]

bb. Weiterverarbeitung teilanonymisierter Entscheidungen: Training i.w.S., Speicherung und Einsatz des Sprachmodells

Für die folgenden Schritte in Form des Trainings i.w.S., der Speicherung des Sprachmodells und dem Einsatz des auf dem Sprachmodell basierenden KI-Systems stellt sich die Frage, ob diese Verarbeitungstätigkeiten (noch) zur Wahrnehmung der öffentlichen Aufgabe(n)[240] der Gerichte und Ministerien erforderlich sind. Insoweit ist zu untersuchen, ob die Weiterverarbeitung von teilanonymisierten Gerichtsentscheidungen zur Optimierung von Arbeitsabläufen in der Justiz (noch) den der Justiz zugewiesenen Aufgaben im öffentlichen Interesse dient.

Im Ausgangspunkt dürfte weitgehend Einigkeit darüber bestehen, dass die Veröffentlichung teilanonymisierter Gerichtsentscheidungen zur Berücksichtigung in der Fachwelt[241] ebenso wie die Vorlage von Gerichtsakten als Arbeitsnachweis in richterlichen Bewerbungsverfahren oder die Umwandlung von in Gerichtsakten enthaltenen Informationen zu Statistikzwecken öffentlichen Aufgaben der Justiz zuzuordnen ist.[242] Gerade die Statistikzwecke dienen zwar noch (jedenfalls teilweise) der Transparenz gegenüber der Öffentlichkeit, ermöglichen aber vor allem die Prüfung und Anpassung von Arbeitsabläufen und der Arbeitsorganisation in der Justiz.

239 S. zu den Anforderungen oben unter D.II.5.b.

240 Für die Übermittlung zwischen den Ministerien zu Zwecken der (zulässigen) Zusammenarbeit ist ebenfalls ein Aufgabenzusammenhang anzunehmen, vgl. auch Art. 5 Abs. 1 S. 1 Nr. 1 BayDSG und § 6 DSG NRW.

241 Z.B. eine zwar bei Betrachtung des Texts vollständige Anonymisierung, die aber die Herstellung eines Personenbezugs unter Zuhilfenahme von Presseberichterstattung zulässt. S. zuvor unter D.II.5.b.

242 S. bspw. die Anordnung über die Erhebung von statistischen Daten in der Verwaltungsgerichtsbarkeit (VwG-Statistik), https://media.frag-den-staat.de/files/foi/570849/VwG-Anordnung_2020.pdf: Danach sind u.a. das Aktenzeichen einer Sache, die Verfahrensart, die Art der Verfahrensbeendigung sowie Informationen zur Prozesskostenhilfe zu erheben. Diese Informationen sind (zunächst) aufgrund des Bezugs zu einer konkreten Akte personenbezogen.

Die Aufgaben der Justiz mit Datenverarbeitungsbezug im öffentlichen Interesse beschränken sich also nicht nur auf die eigentliche Rechtsprechung, sondern umfassen darüber hinaus auch Verarbeitungsvorgänge in engem Zusammenhang mit der Rechtsprechung sowie personellen und organisatorischen Angelegenheiten der Justiz. Eine Zusammenarbeit der Gerichte, gegebenenfalls unter Einbindung oder ausgehend von den jeweiligen Justizministerien, ist – wie in Statistiksachen – unschädlich. Denn ungeachtet der richterlichen Unabhängigkeit im Rahmen der Rechtsprechung i.e.S. werden sich insoweit stets organisatorische und institutionelle Anforderungen (schon insbesondere aus dem Beamtenrecht) ergeben, die zwar über das einzelne Gericht hinausreichen, aber einen hinreichenden Bezug zu öffentlichen Aufgaben aufweisen.

Die hier untersuchungsgegenständliche Verarbeitung zum Training eines Justiz-Sprachmodells als Arbeitshilfe zur Sachverhaltsdurchdringung in neuen Verfahren (siehe die Use-Cases 1-3) wird daher ebenso wie die Führung einer digitalen Akte oder einer aktenübergreifenden Statistik in einem hinreichenden Zusammenhang mit der öffentlichen Aufgabenerfüllung stehen und zur Aufgabenerfüllung erforderlich sein.[243]

Die Erforderlichkeit i.S.d. Art. 6 Abs. 1 UAbs. 1 lit. e DSGVO zielt hierbei nicht bloß ab auf den möglichen Einsatz eines milderen Mittels,[244] sondern darüber hinaus auch auf einen in Ansehung der mit der Datenverarbeitung verbundenen Risiken hinreichenden Funktionszusammenhang zwischen Verarbeitung und Aufgabenwahrnehmung. Daher wird es für das untersuchungsgegenständliche Projekt unschädlich sein, dass bereits ein geringer Anteil an Gerichtsentscheidungen in anonymisierter Form öffentlich zugänglich ist. Selbst wenn man an dieser Stelle eine Erforderlichkeit im Sinne des deutschen Verfassungsrechts verlangen würde (i.e. die Prüfung eines milderen, gleich effektiven Mittels), dürfte dem Erfordernis bei einer erheblich gesteigerten Datengrundlage (z.B. einer Vervielfachung der zur Verfügung stehenden Urteile um einen Faktor deutlich > 5) zum effektiveren Training eines Sprachmodells Genüge getan sein.

In diesem Sinne erkennt auch der europäische Gesetzgeber selbst in der KI-VO den Bedarf nach großen Datenmengen zum KI-Training an (z.B. in ErwGr. 97 S. 3, 98, 105 S. 2 KI-VO: „riesigen Mengen“ sowie vgl. Art. 3

243 Wohl krit. im Hinblick auf die Aufgabenerfüllung durch ein KI-Training, allerdings ohne Berücksichtigung der hier vorgenommenen Teilanonymisierung und weiterer Maßnahmen, *LfDIBW*, Diskussionspapier: Rechtsgrundlagen im Datenschutz beim Einsatz von Künstlicher Intelligenz, S. 31.

244 *Albers/Veit*, in: BeckOK Datenschutzrecht, Art. 6 DSGVO Rn. 59.

Nr. 29-32 und Art. 10 KI-VO), da sich aus der Verwendung dieser großen Datenmengen ein signifikanter Vorteil gegenüber kleineren Datenmengen ergeben kann (z.B. entsprechend der wahrscheinlichkeitsbasierten Ausgabe im Hinblick auf die Genauigkeit der Ausgabe, einen sog. Bias oder die Unterbindung von Ausgaben von – personenbezogenen – Daten betreffend einzelne Entscheidungen aus dem Trainingsdatensatz).[245]

Weiterhin ist zu beachten, dass die hier grundsätzlich einschlägigen Generalklauseln aus Art. 4 Abs. 1 BayDSG bzw. § 3 Abs. 1 DSG NRW nur eine eingriffsarme Verarbeitung tragen können. Die untersuchungsgegenständliche Verarbeitung umfasst nur einen Restbestand an überwiegend nicht-sensiblen personenbezogenen Daten,[246] die innerhalb der Einflusssphäre der für die Justiz Verantwortlichen verbleiben. Die Grundrechte (insbesondere Art. 8 GRCh) und Interessen betroffener Personen sind also allenfalls geringfügig betroffen und Schäden in Form finanzieller Nachteile oder drohender Diskriminierungen (vgl. ErwGr. 75 DSGVO) sind unwahrscheinlich.

Diesen Befund bestätigen weitere Umstände des Einzelfalls, die sich bereits auf Ebene der Annahme personenbezogener Daten[247] aus- und an dieser Stelle fortwirken:[248] Umfasst sind nach einer Bearbeitung durch das Erlanger Tool vor allem Informationen, die nicht als Kennungen unmittelbar auf eine Person hinweisen (d.h. insbesondere nur selten Namen oder Anschriften, sondern überwiegend Sachverhaltsangaben, Aktenzeichen und Daten). Diese Informationen erfordern einen erhöhten Aufwand als Ausgangspunkt für eine Identifizierung und schränken schon hierdurch die Eingriffsintensität ein.

Vor allem aber ist der Kreis der Empfänger erheblich begrenzt: Für das Training verbleiben die Daten im Einflussbereich der Ministerien unter Einschaltung der ausführenden Stellen als Auftragsverarbeiter. Erst im Fall eines Datenlecks im Rahmen der Ausgabe des eingesetzten KI-Systems würde der Empfängerkreis erweitert, indem beispielsweise Gerichte verse-

245 Vgl. die Forschung zu Duplikaten im Trainingsdatensatz *Kandpal/Wallace/Raffel*, Deduplicating Training Data Mitigates Privacy Risks in Language Models; sowie im Überblick zur Datenqualität *Albalak et al.*, A Survey on Data Selection for Language Models.

246 Vgl. auch *EDPB*, Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, Rn. 73.

247 S. oben unter D.I.1.b.bb.

248 Vgl. auch *Hornung/B. Wagner*, CR 2019, 565 (Rn. 50).

hentlich und ganz vereinzelt[249] Zugriff auf verbliebene personenbezogene Daten aus dem Trainingsdatensatz erhielten. Selbst dieser erweiterte Empfängerkreis wird in Ansehung der besonderen Stellung der Gerichte (vgl. Art. 1 Abs. 3, Art. 20 Abs. 3 GG), der Richterschaft und Beamtenschaft, des dienstlichen Kontexts und des großen Interesses an der effizienten Bearbeitung der jeweils aktuellen Akte allerdings nicht mit einer signifikanten Eingriffstiefe für betroffene Personen verbunden sein. Überdies wird dieser (justizinterne) Empfängerkreis auch im Rahmen der berechtigten Erwartungen betroffener Personen liegen (ErwGr. 47 S. 3 DSGVO).

Nach alledem wird die Verarbeitung vereinzelter personenbezogener Daten für das KI-Training, die Speicherung und der justizinterne Einsatz des Sprachmodells in einem KI-System – gestützt auf Art. 6 Abs. 1 UAbs. 1 lit. e, Abs. 2, 3 DSGVO i.V.m. Art. 4 Abs. 1 BayDSG bzw. § 3 Abs. 1 DSG NRW – als grundsätzlich zulässig anzusehen sein. Das gilt in Ansehung des eng begrenzten Empfängerkreises, der sehr geringen Quantität personenbezogener Daten aufgrund des Einsatzes von Anonymisierungstools und weiterer Maßnahmen sowie des engen Zusammenhangs mit dem Einsatz innerhalb der Justiz zur Unterstützung bei Arbeitsabläufen auf Basis der Erkenntnisse aus vergangenen Gerichtsverfahren.

cc. Veröffentlichung des Sprachmodells

Die Veröffentlichung des Sprachmodells als KI-System, des eigentlichen Modells mitsamt der Gewichte und der vollständigen Trainingsdaten unterfällt jeweils nicht mehr der Aufgabenerfüllung i.S.d. Art. 6 Abs. 1 UAbs. 1 lit. e DSGVO.

Die Veröffentlichung, gleich in welcher Form, dient zwar als Arbeitshilfe für die Praxis der Rechtsanwendung, aber auch der Transparenz der Arbeitsweise der Justiz sowie gegebenenfalls wissenschaftlichen Forschungszwecken.

Die Transparenz öffentlichen Handelns ist beispielsweise in den Informationsfreiheits- und Transparenzgesetzen der Länder als legitimes Ziel angelegt. Wissenschaftliche Forschungszwecke sind an verschiedenen Stel-

249 Das erfordert erneut entsprechende Maßnahmen auf Ausgabeebene (z.B. Ausgabefilter).

len in der DSGVO privilegiert.[250] Es müssten aber gerade diese Zwecke im Zusammenhang mit der Erfüllung einer Aufgabe der Justiz bzw. Ministerien im öffentlichen Interesse stehen, für die die Verarbeitung personenbezogener Daten erforderlich ist. Unabhängig davon, ob überhaupt noch ein Aufgabenbezug angenommen werden kann, dürfte jedenfalls die maßgebliche Erforderlichkeit ausscheiden. Denn die personenbezogenen Daten werden gegebenenfalls einem unbeschränkten Empfängerkreis zugänglich gemacht, wovon eine erhebliche Erhöhung des Risikos für die betroffenen Personen ausgeht. Nur in der niedrigsten Veröffentlichungsstufe, sprich dem bloß eingeschränkten Zugriff der Öffentlichkeit auf ein KI-System, können diese Auswirkungen zwar erheblich begrenzt werden, räumen aber gleichwohl nicht die Zweifel im Hinblick auf den Zusammenhang zur öffentlichen Aufgabenerfüllung aus.

Je näher das Erlanger Tool oder ein anderes Instrument zur automatisierten Anonymisierung an einen Anonymisierungsstandard heranreicht, der auch bei sorgfältiger händischer Anonymisierung erzielt worden wäre,[251] desto eher wird tendenziell auch die Zulässigkeit der Veröffentlichung des Sprachmodells hinzunehmen sein. Denn wenn die Veröffentlichung der – gegebenenfalls nur weitgehend – anonymisierten Gerichtsentscheidung zulässig wäre, birgt die Zusammenführung dieser ohnehin veröffentlichungsfähigen Gerichtsentscheidungen im Rahmen des KI-Trainings kaum (mehr) erhöhte Risiken für betroffene Personen und kann daher zugleich einen geringen Zusammenhang mit der öffentlichen Aufgabenerfüllung ausreichen lassen.

d. Art. 6 Abs. 1 UAbs. 1 lit. c DSGVO (rechtliche Pflicht)

Wie gezeigt werden konnte, dürfte für das untersuchungsgegenständliche Projekt regelmäßig Art. 6 Abs. 1 UAbs. 1 lit. e DSGVO anstelle von lit. c in Betracht kommen.[252] Falls im Einzelfall eine vollständige Anonymisierung bewirkt werden können sollte, kann die Anonymisierung einer Löschung

250 S. z.B. Art. 5 Abs. 1 lit. b, e, Art. 9 Abs. 2 lit. j, Art. 14 Abs. 5 lit. b, Art. 17 Abs. 3 lit. b, Art. 21 Abs. 6, Art. 85, 89 DSGVO sowie ErwGr. 33, 50, 52, 53, 62, 65, 133, 156 ff. DSGVO.

251 S. oben unter D.II.5.b.

252 S. oben unter D.V.4.c.

gleichstehen und auf Art. 6 Abs. 1 UAbs. 1 lit. c i.V.m. Art. 17 DSGVO gestützt werden.[253]

e. Zweckänderung (Art. 5 Abs. 1 lit. b, Art. 6 Abs. 4 DSGVO)

Weiterhin – oder nach einer Ansicht alternativ zu den Voraussetzungen aus Art. 6 Abs. 1 DSGVO – sind als Ausprägung aus dem in Art. 5 Abs. 1 lit. b DSGVO niedergelegten Zweckbindungsgrundsatz die Anforderungen an eine Zweckänderung aus Art. 6 Abs. 4 DSGVO[254] zu beachten, wenn für die weitere Verarbeitung von dem Erhebungszweck abgewichen wird. Das betrifft die untersuchungsgegenständliche Konstellation der Weiterverarbeitung personenbezogener Daten durch das Ministerium zu Zwecken des KI-Trainings.

aa. Verhältnis zu Art. 6 Abs. 1 DSGVO

Es sprechen durchaus beachtliche Argumente dafür, Art. 6 Abs. 4 DSGVO als eine Spezialvorschrift zu Art. 6 Abs. 1 DSGVO anzusehen. In diesem Fall würde die zuvor anhand des Art. 6 Abs. 1 DSGVO durchgeführte Prüfung vollständig durch den nach Art. 6 Abs. 4 DSGVO erforderlichen Kompatibilitätstest verdrängt (siehe hierzu nachfolgend unter D.V.4.e.bb(2)). Die Rechtfertigung einer Verarbeitung im GSJ-Projekt wäre mithin an einem weniger strengen Maßstab zu messen. In Literatur[255] und Rechtsprechung ist allerdings (noch) nicht abschließend geklärt, ob ein solches Spezialitäts-

253 *Stürmer*, ZD 2020, 626 (630).

254 Wenn und soweit man die landesrechtlichen Vorschriften Art. 5 Abs. 1 S. 1 Nr. 2 BayDSG bzw. § 8 Abs. 2 DSG NRW als geeignete Rechtsgrundlage entgegen *BVerwG*, BeckRS 2018, 35172 (Rn. 25 f.) ansieht, treten diese – i.E. ohne Auswirkungen – an die Stelle des Art. 6 Abs. 4 DSGVO.

255 Für die Notwendigkeit einer separaten Rechtsgrundlage *Albers/Veit*, in: BeckOK Datenschutzrecht, Art. 6 DSGVO Rn. 108; und die Rechtsprechung des EuGH hierzu für „unmissverständlich" haltend *Buchner/Petri*, in: Kühling/Buchner, Art. 6 DSGVO Rn. 184; *Heberlein*, in: Ehmann/Selmayr, Art. 6 DSGVO Rn. 26; *DSK*, Erfahrungsbericht der unabhängigen Datenschutzaufsichtsbehörden des Bundes und der Länder zur Anwendung der DS-GVO, S. 13 f.; *Albrecht*, CR 2016, 88 (92); *Schantz*, NJW 2016, 1841 (1844); *Pauly/Nabulsi*, ZD 2023, 519 (522); a.A. *Roßnagel*, in: Simitis/Hornung/Spiecker gen. Döhmann, Art. 6 Abs. 4 DSGVO Rn. 12; *Schulz*, in: Gola/Heckmann, Art. 6 DSGVO Rn. 142; *Taeger*, in: Taeger/Gabel, Art. 6 DSGVO Rn. 173; *Culik/Döpke*, ZD 2017, 226 (230); *Stürmer*, ZD 2020, 626

verhältnis anzunehmen ist oder vielmehr die Anforderungen aus Art. 6 Abs. 4 DSGVO *zusätzlich* zu den Anforderungen aus Art. 6 Abs. 1 DSGVO zu beachten sind, also insgesamt höhere Anforderungen an die zweckändernde Weiterverarbeitung zu stellen sind.

(1) Art. 6 Abs. 4 DSGVO als ergänzende Regelung

Die ein Spezialitätsverhältnis ablehnenden Stimmen verweisen darauf, Art. 6 Abs. 4 DSGVO sei gerade nicht als ein gesonderter Rechtfertigungstatbestand konzipiert, weshalb die Fragen nach der Rechtsgrundlage und der Zweckbindung wie schon unter der DSRL voneinander zu trennen seien.[256] Ferner ergebe sich der Befund einer stets notwendigen Rechtsgrundlage auch aus Art. 8 Abs. 2 GRCh, dessen Erfordernisse der europäische Gesetzgeber in Art. 6 Abs. 4 DSGVO achten wollte.[257] Zudem wird eine dann möglicherweise konsequente Anwendung des Art. 6 Abs. 4 DSGVO auch auf besondere Kategorien personenbezogener Daten i.S.d. Art. 9 DSGVO kritisiert.[258] In historischer Hinsicht wird darauf hingewiesen, dass ein Vorschlag, in Art. 6 Abs. 4 DSGVO auf einzelne Rechtsgrundlagen zu verweisen („Ist der Zweck der Weiterverarbeitung mit dem Zweck, für den die personenbezogenen Daten erhoben wurden, nicht vereinbar, muss auf die Verarbeitung mindestens einer der in Absatz 1 Buchstaben a bis e genannten Gründe zutreffen"), durch das Europäische Parlament im Gesetzgebungsverfahren zurückgewiesen wurde.[259]

(2) Art. 6 Abs. 4 DSGVO als eigenständige Rechtsgrundlage

Allerdings indiziert die Formulierung in Art. 83 Abs. 5 lit. a DSGVO („Bei Verstößen gegen die folgenden Bestimmungen …: die Grundsätze für die Verarbeitung, einschließlich der Bedingungen für die Einwilligung, gemäß

(629 f.); *P. Richter*, DuD 2015, 735 (736); Überblick zum Streitstand auch bei *Ziegenhorn/Schulz-Große*, ZD 2023, 581 (586) m.w.N.

256 *Buchner/Petri*, in: Kühling/Buchner, Art. 6 DSGVO Rn. 184.

257 *Buchner/Petri*, in: Kühling/Buchner, Art. 6 DSGVO Rn. 184; *Artikel-29-Datenschutzgruppe*, WP 203, S. 12.

258 *Pauly/Nabulsi*, ZD 2023, 519 (522).

259 *Pauly/Nabulsi*, ZD 2023, 519 (522); *Albrecht*, CR 2016, 88 (92). Vgl. auch die DSGVO in der Entwurfsfassung 2017/C 378/55, abrufbar unter: https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=OJ:C:2017:378:FULL&from=DE.

den Artikeln 5, 6, 7 und 9;"), dass die jeweiligen Absätze des Art. 6 DSGVO zusammen zu betrachten sind, sprich Art. 6 Abs. 4 DSGVO das Erfordernis einer Rechtsgrundlage nach Art. 6 Abs. 1 DSGVO gegebenenfalls verdrängen kann.

Das Erfordernis eines zusätzlichen Erlaubnistatbestands würde außerdem geringfügige Zweckänderungen durch einen Verantwortlichen benachteiligen (z.B. die justizinterne Weiterverarbeitung von personenbezogenen Daten in Gerichtsentscheidungen zu Dokumentations- oder Statistikzwecken), da für diese Verarbeitung sowohl die Voraussetzungen aus Art. 6 Abs. 4 DSGVO als auch aus Art. 6 Abs. 1 DSGVO mit Blick auf die Weitervereinbarung einzuhalten wären.[260] Eine Neuerhebung durch einen anderen Verantwortlichen zu einem gänzlich anderen Zweck bedürfte demgegenüber bloß einer Rechtsgrundlage nach Art. 6 Abs. 1 DSGVO. Diese Benachteiligung geringfügiger Zweckänderungen durch einen Verantwortlichen könnte aber aufgrund der eigenständigen Bedeutung der Zweckbindung (Art. 5 Abs. 1 lit. DSGVO) gerechtfertigt sein. Insoweit wäre die Erwartung betroffener Personen geschützt, die weitere Verarbeitung durch einen Verantwortlichen erfolge entsprechend dem ursprünglichen Erhebungszweck.

Hervorzuheben ist allerdings zugunsten eines Spezialitätsverhältnisses, dass ErwGr. 50 S. 2 DSGVO im Einklang mit der Systematik des Art. 6 DSGVO eindeutig formuliert, dass in diesen Fällen „keine andere gesonderte Rechtsgrundlage […] als diejenige für die Erhebung der personenbezogenen Daten" erforderlich ist.[261] Daran vermag auch der *allgemeine* Verweis auf die Einhaltung der Datenschutzgrundsätze in ErwGr. 50 S. 8 DSGVO nichts zu ändern.[262] Der Gesetzgeber hat sich zudem gerade gegen einen Verweis auf Art. 6 Abs. 1 DSGVO für die Weiterverarbeitung und damit auch gegen das Erfordernis eines gesonderten Rechtmäßigkeitstatbestands entschieden. Andernfalls, sprich wenn gegebenenfalls erneut eine (zusätzliche) Einwilligung nach Art. 6 Abs. 1 lit. a DSGVO zu prüfen wäre, wäre zudem die gesonderte Nennung der Einwilligung und der Rechtsvorschriften in Art. 6 Abs. 4 Hs. 1 DSGVO anstelle des Kompatibilitätstests nach Art. 6 Abs. 4 Hs. 2 DSGVO nicht erforderlich und liefe faktisch weitgehend leer.

260 *Roßnagel*, in: Simitis/Hornung/Spiecker gen. Döhmann, Art. 6 Abs. 4 DSGVO Rn. 12.

261 *Culik/Döpke*, ZD 2017, 226 (230); *P. Richter*, DuD 2015, 735 (736).

262 So aber etwa *DSK*, Erfahrungsbericht der unabhängigen Datenschutzaufsichtsbehörden des Bundes und der Länder zur Anwendung der DS-GVO, S. 13.

(3) Berücksichtigung der EuGH-Rechtsprechung

Der EuGH hat sich bisher, soweit ersichtlich, noch nicht eindeutig positioniert, scheint aber zu einem Spezialitätsverhältnis zu tendieren: Soweit Stimmen in der Literatur[263] auf die EuGH-Entscheidung in der Rs. Bara verweisen, stellt der EuGH dort im Kontext einer Übermittlung bloß fest: „Gemäß den Bestimmungen des Kap. II („Allgemeine Bedingungen für die Rechtmäßigkeit der Verarbeitung personenbezogener Daten") der RL 95/46 muss jede Verarbeitung personenbezogener Daten […] den in Art. 6 der RL [i.e. Art. 5 DSGVO] aufgestellten Grundsätzen in Bezug auf die Qualität der Daten und einem der in Art. 7 der RL [i.e. Art. 6 DSGVO] angeführten Grundsätze in Bezug auf die Zulässigkeit der Verarbeitung von Daten genügen".[264] Damit ist allerdings gerade nichts zu dem Verhältnis von Art. 6 Abs. 1 DSGVO und Art. 6 Abs. 4 DSGVO (i.V.m. Art. 5 Abs. 1 lit. b DSGVO) gesagt: Einerseits liegt im Fall der Übermittlung an einen neuen Verantwortlichen ohnehin eine neue Erhebung i.S.d. Art. 6 Abs. 4 vor, sodass in jedem Fall Art. 6 Abs. 1 DSGVO für den neuen Verantwortlichen zur Anwendung berufen wäre.[265] Andererseits verhält sich der EuGH an der vorbenannten Stelle gerade nicht ausdrücklich zu Art. 6 Abs. 4 DSGVO, der in dieser Form als Ausprägung des Zweckbindungsgrundsatzes in die DSGVO aufgenommen wurde.

Die vorzugswürdige Ansicht, die ein Spezialitätsverhältnis annimmt, dürfte mittlerweile jedenfalls eine Stütze in der EuGH-Rechtsprechung finden.[266] In einer Entscheidung aus dem Jahr 2022 befasste sich der EuGH mit der Frage, ob die separate Speicherung zu Testzwecken mit dem Grundsatz der Zweckbindung (Art. 5 Abs. 1 lit. b DSGVO) vereinbar ist.[267] In diesem Zusammenhang ging der EuGH nur auf die Rechtmäßigkeit nach Art. 6 Abs. 1 DSGVO für die Erhebung ein[268] und erwähnte mit keinem Wort das Erfordernis einer Rechtsgrundlage aus Art. 6 Abs. 1 DSGVO für die zweckändernde Weiterverarbeitung – wie es sich bei einem anderen Normverständnis ansonsten aufgedrängt hätte –, weshalb der Gerichtshof schließlich zu dem Ergebnis kommt, „dass Art. 5 Abs. 1 Buchst. b der Verordnung 2016/679 dahin auszulegen ist, dass der darin vorgesehene

263 *Buchner/Petri*, in: Kühling/Buchner, Art. 6 DSGVO Rn. 184.
264 *EuGH*, ZD 2015, 577 (Rn. 30) – Bara.
265 So auch *P. Richter*, DuD 2015, 735 (736).
266 S. auch *EuGH*, ZD 2023, 396 (Rn. 43, 46 ff.) – Norra Stockholm Bygg AB.
267 *EuGH*, ZD 2023, 31.
268 *EuGH*, ZD 2023, 31 (Rn. 28).

Grundsatz der ‚Zweckbindung' es dem Verantwortlichen *nicht verwehrt*, in einer zu Testzwecken und zur Behebung von Fehlern eingerichteten Datenbank personenbezogene Daten zu erfassen und zu speichern, die zuvor erhoben und in einer anderen Datenbank gespeichert wurden, *wenn* diese Weiterverarbeitung mit den konkreten Zwecken *vereinbar ist*, für die die personenbezogenen Daten ursprünglich erhoben wurden, was anhand der in Art. 6 Abs. 4 dieser Verordnung genannten Kriterien und sämtlicher Umstände des Einzelfalls zu beurteilen ist"[269] (Hervorhebung d. d. Verf.).

(4) Zwischenergebnis

Insgesamt sprechen also durchaus gewichtige Argumente für ein Spezialitätsverhältnis, sodass die Weiterverarbeitung durch die Ministerien bzw. ihre Auftragsverarbeiter ausschließlich an Art. 6 Abs. 4 DSGVO (und nicht an Art. 6 Abs. 1 DSGVO) zu messen wäre. Allerdings besteht ein nicht unerhebliches Risiko, dass die Rechtsprechung sich in der Zukunft gegen ein solches Spezialitätsverhältnis ausspricht. Eine Verarbeitung im GSJ-Projekt sollte daher nicht nur auf Art. 6 Abs. 4 DSGVO gestützt werden, sondern darüber hinaus die Anforderungen aus Art. 6 Abs. 1 DSGVO beachten.

bb. Kompatibilität im Einzelfall

Entweder als maßgebliche Zulässigkeitsvoraussetzung oder ergänzend zu Art. 6 Abs. 1 DSGVO für den Fall der Zweckänderung durch einen Verantwortlichen, wie im untersuchungsgegenständlichen GSJ-Projekt, sind stets die Kompatibilitätsanforderungen aus Art. 6 Abs. 4 DSGVO zu beachten. Für die Zulässigkeit nach Art. 6 Abs. 4 DSGVO kommt entweder (Hs. 1) eine Rechtsvorschrift in Betracht, „die in einer demokratischen Gesellschaft eine notwendige und verhältnismäßige Maßnahme zum Schutz der in Artikel 23 Absatz 1 genannten Ziele darstellt" oder (Hs. 2) die Durchführung des Kompatibilitätstests anhand der in lit. a-e nicht abschließend aufgeführten Kriterien.

269 *EuGH*, ZD 2023, 31 (Rn. 45).

(1) Rechtsvorschrift (Hs. 1)

Die Anforderungen an das Vorliegen einer Rechtsvorschrift i.S.d. Art. 6 Abs. 4 DSGVO sind im Einzelnen umstritten.[270] Mit dem EuGH ist allerdings davon auszugehen, dass Art. 6 Abs. 4 DSGVO über Art. 6 Abs. 1 UAbs. 1 lit. c, e DSGVO hinausgeht, indem die Rechtsvorschrift die Zweckänderung berücksichtigen muss und den in Art. 23 Abs. 1 DSGVO genannten Zielen entsprechen muss.[271] Art. 23 Abs. 1 lit. f DSGVO nennt unter anderem „den Schutz der Unabhängigkeit der Justiz und den Schutz von Gerichtsverfahren", der die Justiz vor „internen oder externen Eingriffen [schützt], aber auch auf eine ordnungsgemäße Rechtspflege abzielt".[272]

Es bestehen Zweifel, ob die landesrechtlichen Generalklauseln i.V.m Aufgabenzuweisungen an die Rechtspflege[273] hinreichend konkret i.S.d. Art. 6 Abs. 4 DSGVO eine Zweckänderung zur Optimierung der Arbeitsabläufe in der Rechtspflege zulassen. Im Ergebnis wäre das vorliegend insbesondere dann zu bejahen, wenn die Auswirkungen auf die betroffenen Personen sich in Grenzen hielten und die Zwecke eng miteinander verknüpft wären. Eine Aussage über diesen Umfang der Auswirkungen und die Zwecknähe enthält bereits der Art. 6 Abs. 4 Hs. 2 DSGVO.

Im Ergebnis ergeben sich somit für die Frage der Reichweite der Rechtsvorschrift i.S.d. Art. 6 Abs. 4 Hs. 1 DSGVO ähnliche Erwägungen wie innerhalb des Kompatibilitätstests nach Art. 6 Abs. 4 Hs. 2 DSGVO. Vor diesem Hintergrund soll nachfolgend dieser Kompatibilitätstest durchgeführt werden.

(2) Kompatibilitätstest (Hs. 2)

Der Kompatibilitätstest als eine der beiden alternativen Zulässigkeitsvoraussetzungen nach Art. 6 Abs. 4 DSGVO erfordert eine normative[274] Bewertung der Vereinbarkeit des ursprünglichen und des neuen Zwecks unter Berücksichtigung der in lit. a-e genannten Maßgaben sowie der berechtigten Erwartungen betroffener Personen (ErwGr. 50 S. 6 DSGVO). Die

270 Überblick bei *Albers/Veit*, in: BeckOK Datenschutzrecht, Art. 6 DSGVO Rn. 112.
271 *EuGH*, ZD 2023, 396 (Rn. 34 ff.) – Norra Stockholm Bygg AB.
272 *EuGH*, ZD 2023, 396 (Rn. 38) – Norra Stockholm Bygg AB.
273 S. oben unter D.V.4.c.
274 So auch *Ziegenhorn/Schulz-Große*, ZD 2023, 581 (585); zu den Kriterien auch *Ashkar*, ZD 2023, 523 (526 f.).

Weiterverarbeitung zur Teilanonymisierung und zum KI-Training ist daher insbesondere, jedoch nicht nur an den Kriterien der lit. a-e zu prüfen.

Die Verbindung der Zwecke (lit. a) spricht im Hinblick auf den Einsatz in der Justiz und im Zusammenhang mit (anderen) Gerichtsverfahren zwar zunächst für die Kompatibilität. Zugleich ist aber der ursprüngliche Zweck auf ein Gerichtsverfahren, der neue Zweck auf sämtliche zukünftige Gerichtsverfahren und allgemeine Arbeitsabläufe gerichtet.

Im Rahmen des Erhebungszusammenhangs und des Verhältnisses zwischen Verantwortlichem und betroffener Person (lit. b) sind insbesondere die berechtigten Erwartungen betroffener Personen aus diesem Verhältnis heranzuziehen. Abhängig von der Erhebung im Einzelfall und den in Rede stehenden Kategorien betroffener Personen (z.B. Partei, Zeuge oder – nur – im Rahmen eines Schriftsatzes genannte Person) besteht im Regelfall eine Erwartung der Verarbeitung für justizinterne Zwecke, im Einzelfall eine nur geringfügige Erwartung bzw. ein Wissen, überhaupt im Rahmen des Gerichtsverfahrens erwähnt worden zu sein. Für die erstgenannten Konstellationen wird die Erwartung betroffener Personen die Verarbeitung durch die Richterschaft und die Geschäftsstellen sowie gegebenenfalls auch das Justizministerium umfassen, wenn und soweit ein enger Zusammenhang mit der Rechtspflege besteht und keine Weitergabe an Dritte erfolgt (z.B. zu Statistikzwecken). Die untersuchungsgegenständliche Zweckänderung zu Zwecken der verfahrensübergreifenden Optimierung von Arbeitsabläufen durch Anonymisierung, ein KI-Training und den KI-Einsatz dürfte daher vielfach im Rahmen des Erwartbaren liegen; dies gilt unabhängig davon, ob der Einsatz für die konkrete Software (in Form einer KI-Anwendung) erwartet wurde.

Die Art der personenbezogenen Daten (lit. c) hängt insbesondere davon ab, ob besondere Kategorien personenbezogener Daten i.S.d. Art. 9 DSGVO verarbeitet werden. Sofern eine mit Blick auf Art. 9 DSGVO sinnvolle Clusterung der Urteile erfolgt,[275] wird auch an dieser Stelle davon auszugehen sein, dass regelmäßig keine besonderen Kategorien personenbezogener Daten verarbeitet werden. Die Sensibilität der verarbeiteten Daten wird dann im Übrigen als eher gering einzuschätzen sein, zumal die Informationen unter Umständen zumeist bereits aus anderen Quellen bekannt sein können (z.B. aus einer medialen Berichterstattung).

Die Folgen der beabsichtigten Weiterverarbeitung (lit. d) weisen Überschneidungen auf mit dem Vorhandensein geeigneter Garantien (lit. e).

275 S. hierzu nachfolgend unter D.V.5.

Durch die anzuwendenden Garantien in Form der Teilanonymisierung, der weiteren Gewährleistung der Zweckbindung durch technische und organisatorische Maßnahmen sowie einen auf die Justiz beschränkten Empfängerkreis sind keine schwerwiegenden Folgen für betroffene Personen zu erwarten, soweit überhaupt (vereinzelt) personenbezogene Daten vorliegen.

Über die genannten Merkmale hinaus können auch die bereits im Rahmen von Art. 6 Abs. 1 UAbs. 1 lit. e DSGVO genannten Abwägungsaspekte fruchtbar gemacht werden.[276] Diese Abwägungsaspekte stützen ebenfalls den Befund der Kompatibilität des neuen Zwecks mit dem ursprünglichen Zweck der Verarbeitung.

cc. Schlussfolgerungen für das untersuchungsgegenständliche Projekt

Die Anforderungen des Art. 6 Abs. 4 Hs. 2 DSGVO sind also bei Vorliegen hinreichender Garantien (siehe nachfolgend unter D.VIII) und der sorgfältigen Clusterung mit Blick auf besondere Kategorien personenbezogener Daten grundsätzlich gewahrt. In Ansehung der im Rahmen des Kompatibilitätstests ermittelten Zwecknähe erscheint alternativ auch eine Rechtfertigung nach Art. 6 Abs. 4 Hs. 1 DSGVO i.V.m. landesrechtlichen Vorschriften[277] denkbar.

Die Weiterverarbeitung im untersuchungsgegenständlichen Projekt zu KI-Trainingszwecken und dem späteren Einsatz des Sprachmodells in der Justiz kann also entweder – nach bestrittener Auffassung – bereits auf Art. 6 Abs. 4 DSGVO gestützt werden oder aber die Verarbeitung erfüllt zumindest die zusätzlichen Voraussetzungen aus Art. 6 Abs. 4 DSGVO und ist zulässig, soweit jeweils auch eine Rechtsgrundlage aus Art. 6 Abs. 1 DSGVO einschlägig ist.

5. Rechtsgrundlagen für besondere Datenkategorien nach Art. 9, 10 DSGVO

Art. 9 DSGVO sieht das Erfordernis einer gesonderten Rechtsgrundlage vor für die „Verarbeitung personenbezogener Daten, aus denen die rassische und ethnische Herkunft, politische Meinungen, religiöse oder weltanschau-

276 S. oben unter D.V.4.c.bb.
277 S. oben unter D.V.4.c.

liche Überzeugungen oder die Gewerkschaftszugehörigkeit hervorgehen, sowie [für] die Verarbeitung von genetischen Daten, biometrischen Daten zur eindeutigen Identifizierung einer natürlichen Person, Gesundheitsdaten oder Daten zum Sexualleben oder der sexuellen Orientierung einer natürlichen Person". Art. 10 DSGVO modifiziert diese Anforderungen mit Blick auf Daten über strafrechtliche Verurteilungen und Straftaten.

Art. 9, 10 DSGVO sehen für eine Abwägung, wie etwa unter Art. 6 Abs. 1 UAbs. 1 lit. f DSGVO, nur einen sehr eingeschränkten Spielraum vor. Vor diesem Hintergrund werden für eingriffsarme Verarbeitungsvorgänge (z.B. eine Anonymisierung) eine teleologische Reduktion des Art. 9 Abs. 1 DSGVO[278] bzw. vergleichbare Lösungsansätze für zufällig verarbeitete sensible Daten vorgeschlagen.[279] Solche teleologisch orientierten Überlegungen erhalten Aufwind durch ein weites Verständnis des Begriffs der Gesundheitsdaten durch den EuGH, der bereits Daten über eine Bestellung nicht verschreibungspflichtiger Medikamente als Gesundheitsdaten einordnet.[280]

a. Relevanz für das untersuchungsgegenständliche Projekt

Für das GSJ-Projekt ist mit Blick auf nach einer Teilanonymisierung gegebenenfalls weiterhin vorhandene personenbezogene Daten zu fragen, ob eine Rechtsgrundlage aus dem Katalog des Art. 9 Abs. 2 DSGVO einschlägig ist. In Ansehung des Ausschlusses von Gerichtsentscheidungen und Aktenauszügen in Strafsachen (vgl. Art. 10 DSGVO) sowie insbesondere auch in Arzthaftungssachen (i.e. betreffend Gesundheitsdaten i.S.d. Art. 9 Abs. 1 DSGVO) und Familiensachen ist bereits in konzeptioneller Hinsicht die Wahrscheinlichkeit des Vorliegens besonderer Kategorien personenbezogenen Daten erheblich reduziert.

Weiterhin in Betracht kommen allerdings gleichwohl das Vorhandensein von vereinzelt enthaltenen Gesundheitsdaten und Daten zum Sexualleben oder der sexuellen Orientierung (z.B. im Hinblick auf deliktsrechtliche Anspruchsgrundlagen, gegebenenfalls i.V.m. §§ 823 ff. BGB oder im Rahmen des Schadensersatzumfangs nach §§ 249 ff. BGB, insbesondere mit Blick auf ein Schmerzensgeld, § 253 Abs. 1, 2 BGB) sowie von Daten über politische Meinungen oder Gewerkschaftszugehörigkeiten (z.B. im Arbeitsrecht mit

278 *Stürmer*, ZD 2020, 626 (630); *Hornung/B. Wagner*, ZD 2020, 223 (228).
279 Etwa *Britz/Indenhuck/Langerhans*, ZD 2021, 559; *Kohn/Schleper*, ZD 2023, 723.
280 *EuGH*, NJW 2025, 33 (Rn. 88).

Gewerkschaftsbezug sowie geltend gemachte Unterlassungsansprüche im Deliktsrecht im Hinblick auf getätigte Aussagen). Darüber hinaus können auch und gerade im Zusammenhang mit deliktsrechtlichen Ansprüchen vereinzelt Verweise auf Strafverfahren und strafrechtliche Verurteilungen enthalten sein, die Art. 10 DSGVO unterfallen. Insoweit bietet sich eine weitergehende Clusterung an, wie sie im Nachfolgenden vorgeschlagen wird.

b. Personenbezogene Daten über politische Meinungen, Gesundheit und sexuelle Orientierung (Art. 9 DSGVO)

Zunächst sind die besonderen Kategorien personenbezogener Daten i.S.d. Art. 9 DSGVO in den Blick zu nehmen.

aa. Art. 9 Abs. 2 lit. e DSGVO (offensichtliche Öffentlichmachung)

Eine Verarbeitung ist nach Art. 9 Abs. 2 lit. e DSGVO zulässig, wenn sie sich auf personenbezogene Daten bezieht, „die die betroffene Person offensichtlich öffentlich gemacht hat". Eine solche „offensichtliche Öffentlichmachung" setzt die Bereitstellung der Daten an einen unbestimmten Personenkreis voraus, wofür die Unterbreitung von Informationen an die Gerichtsöffentlichkeit bzw. oftmals nur Schriftsatzöffentlichkeit als bestimmten Personenkreis nicht genügen wird. Ferner dürfte vorliegend die entsprechend dem Zweck der Regelung vorausgesetzte freiwillige[281] Öffentlichmachung bei Informationen im Rahmen eines Gerichtsverfahrens zumindest zweifelhaft sein.

Der EuGH berücksichtigt insoweit zudem einschränkend den Kontext der Öffentlichmachung, z.B. im Fall der im Rahmen einer Podiumsdiskussion öffentlich gemachten sexuellen Orientierung mit Blick auf die Verarbeitung der Daten über die sexuelle Orientierung im Kontext eines sozialen Netzwerks.[282]

Vor diesem Hintergrund werden im GSJ-Projekt verarbeitete besondere Kategorien personenbezogener Daten regelmäßig nicht bereits deswegen

281 Am Beispiel der Impressumspflicht *DSK*, Orientierungshilfe der Aufsichtsbehörden zur Verarbeitung von personenbezogenen Daten für Zwecke der Direktwerbung unter Geltung der Datenschutz-Grundverordnung, S. 15; *Reichert/K. Radtke/Eske*, ZD 2024, 483 (485).

282 *EuGH*, NJW 2025, 207 (Rn. 76 ff.).

als öffentlich gemacht anzusehen sein, weil sie der Gerichtsöffentlichkeit präsentiert wurden. Im Einzelfall, ohne belastbar alle in dem GSJ-Projekt in Betracht kommenden Konstellationen erfassen zu können, ist allerdings denkbar, dass die Information bereits auf anderem Wege (z.B. medial oder über das Internet) öffentlich gemacht wurde und daher Art. 9 Abs. 2 lit. e DSGVO zur Anwendung gelangt.

bb. Art. 9 Abs. 2 lit. f DSGVO (Rechtsverteidigung und justizielle Tätigkeit der Gerichte)

Nach Art. 9 Abs. 2 lit. f DSGVO ist eine Verarbeitung besonderer Kategorien personenbezogener Daten zulässig, soweit diese erforderlich ist „zur Geltendmachung, Ausübung oder Verteidigung von Rechtsansprüchen oder bei Handlungen der Gerichte im Rahmen ihrer justiziellen Tätigkeit". Die Vorschrift stellt eine spezielle Ausprägung der in Art. 23 Abs. 1 lit. f DSGVO niedergelegten und weitergefassten Öffnungsklausel zum „Schutz der Unabhängigkeit der Justiz und den Schutz von Gerichtsverfahren" dar. Art. 9 Abs. 2 lit. f DSGVO hat vor allem Bedeutung mit Blick auf die Tatbestandsalternative der Rechtsverfolgung und den individuellen Justizgewährleistungsanspruch.[283] Das Merkmal der justiziellen Tätigkeit knüpft an diese Tatbestandsalternative an und stellt sicher, dass die Justiz die für die Entscheidungsfindung im Einzelfall notwendigen Informationen verarbeiten kann.[284]

Die Verwaltungstätigkeit der Gerichte ist nach dem Wortlaut des Art. 9 Abs. 2 lit. f DSGVO („justizielle Tätigkeit") allerdings gerade nicht adressiert,[285] mithin auch nicht die Organisation, Entwicklung und Beschaffung von Hilfsmitteln für die justizielle Tätigkeit. Wegen der erhöhten Schutzbedürftigkeit der betroffenen Personen, wie sie in Art. 9 Abs. 1 DSGVO zum Ausdruck kommt, kann das untersuchungsgegenständliche KI-Training somit nicht auf diesen Ausnahmetatbestand gestützt werden.

Etwas anderes mag allerdings gelten, soweit der Anonymisierungsstandard einer zulässigen Veröffentlichung erreicht wird:[286] In diesem Fall stellt sich die durch jedermann mögliche Weiterverarbeitung als eingriffsintensiv

283 *Weichert*, in: Kühling/Buchner, Art. 9 DSGVO Rn. 83; *Kampert*, in: Sydow/Marsch, Art. 9 DSGVO Rn. 33.
284 *Mester*, in: Taeger/Gabel, Art. 9 DSGVO Rn. 27.
285 *Kampert*, in: Sydow/Marsch, Art. 9 DSGVO Rn. 33.
286 S. unter D.II.5.b.

dar und ein hinreichender Zusammenhang mit der justiziellen Tätigkeit der Gerichte bleibt gewahrt.

cc. Art. 9 Abs. 2 lit. g DSGVO (Öffnungsklausel für ein erhebliches öffentliches Interesse)

Art. 9 Abs. 2 lit. g DSGVO enthält eine Öffnungsklausel, zu deren Wahrnehmung die nationale Rechtsvorschrift einem erheblichen öffentlichen Interesse dienen muss. Die landesrechtlichen Generalklauseln (z.B. aus Art. 4 Abs. 1 BayDSG) bedienen allerdings nicht ein solches erhebliches Interesse. Das zeigt sich in systematischer Hinsicht etwa an Art. 8 BayDSG, der die Verarbeitung besonderer Kategorien personenbezogener Daten adressiert.

Entsprechendes ergibt sich für das DSG NRW, das in § 16 DSG NRW Regelungen für die Verarbeitung besonderer Kategorien personenbezogener Daten vorsieht. In beiden Vorschriften finden sich im Ergebnis keine Tatbestände, auf die sich die untersuchungsgegenständliche Verarbeitung belastbar stützen lässt.

Allerdings ist eine Verarbeitung beispielsweise zu Zwecken des Rechts der sozialen Sicherheit (Art. 8 Abs. 1 S. 1 Nr. 1 BayDSG, § 16 Abs. 1 Nr. 4 DSG NRW), der Arbeitsmedizin (Art. 8 Abs. 1 S. 1 Nr. 3 BayDSG), der öffentlichen Gesundheit (Art. 8 Abs. 1 S. 1 Nr. 4 BayDSG, § 16 Abs. 1 Nr. 3 DSG NRW) sowie zur Abwehr erheblicher Nachteile für die öffentliche Sicherheit (Art. 8 Abs. 1 S. 1 Nr. 5 BayDSG i.V.m. Art. 6 Abs. 2 Nr. 3 lit. a BayDSG, § 16 Abs. 1 Nr. 1 DSG NRW) zugelassen.[287]

Die unionsrechtlich in Art. 10 Abs. 5 KI-VO vorgesehene Rechtsgrundlage[288] für die Verarbeitung besonderer Kategorien personenbezogener Daten erlaubt nur die Verarbeitung „für die Erkennung und Korrektur von Verzerrungen im Zusammenhang mit Hochrisiko-KI-Systemen". Die vorliegend für das GSJ-Projekt in Rede stehende Verarbeitung erfolgt allerdings gerade nicht gezielt zum Zwecke einer Verzerrungserkennung.

287 Ähnl. auf Bundesebene nach § 22 BDSG.
288 *Steinrötter/Markert*, RDi 2024, 400 (Rn. 24).

dd. Art. 9 Abs. 2 lit. j DSGVO (wissenschaftliche Forschungszwecke)

Art. 9 Abs. 2 lit. j DSGVO enthält eine Öffnungsklausel für wissenschaftliche Forschungszwecke, von der die nationalen und Landesgesetzgeber in § 27 BDSG[289] sowie in Art. 8 Abs. 1 S. 1 Nr. 5 BayDSG i.V.m. Art. 6 Abs. 2 Nr. 3 lit. c BayDSG und § 17 DSG NRW Gebrauch gemacht haben. Die hier maßgeblichen landesrechtlichen Vorschriften verlangen jeweils im Einklang mit Art. 89 Abs. 1 DSGVO[290] eine Interessenabwägung sowie gegebenenfalls weitere technische und organisatorische Maßnahmen.

Für die Anwendbarkeit von Art. 9 Abs. 2 lit. j DSGVO kommt es entscheidend darauf an, ob die betreffenden Verarbeitungen wissenschaftlichen Forschungszwecken dienen. Die Regelung ist unionsautonom auszulegen, wobei der Begriff der wissenschaftlichen Forschungszwecke wegen der unterschiedlichen Konzeption beider Rechtsgebiete nicht zwangsläufig in allen Einzelheiten mit dem nachfolgend unter E.II.2.c.dd erörterten urheberrechtlichen Verständnis gleichläuft.[291] Im Rahmen sowohl des Datenschutzrechts als auch des Urheberrechts wird die Verfolgung wissenschaftlicher Forschungszwecke allerdings jedenfalls dann abzulehnen sein, wenn Handlungen mit Forschungsbezug (z.B. Verarbeitungen oder Vervielfältigungen zur Entwicklung „innovativer" Produkte) unmittelbar und vorrangig auf die Entwicklung eines Produkts für den operativen Einsatz abzielen.

Der Begriff der wissenschaftlichen Forschungszwecke ist (auch) im Kontext der DSGVO unter Berücksichtigung der grundrechtlich garantierten akademischen Freiheit (Art. 13 S. 2 GRCh) auszulegen; umfasst ist hiervon „alles, was nach Inhalt und Form als ernsthafter, planmäßiger Versuch zur Ermittlung der Wahrheit anzusehen ist".[292] Der europäische Gesetzgeber geht in ErwGr. 159 S. 2 DSGVO ausdrücklich von einer weiten Auslegung aus,[293] wonach der Begriff beispielsweise „die technologische Entwicklung und die Demonstration, die Grundlagenforschung, die angewandte Forschung und die privat finanzierte Forschung" umfassen soll. Die Inbezug-

289 S. insb. zur Unsicherheit, ob die Regelung eine belastbare Rechtsgrundlage darstellt, *LfDIBW*, Diskussionspapier: Rechtsgrundlagen im Datenschutz beim Einsatz von Künstlicher Intelligenz, S. 35.

290 Zu der Rechtsnatur des Art. 89 Abs. 1 DSGVO *Matzke*, in: BeckOK Datenschutzrecht, Art. 89 DSGVO Rn. 11.

291 Hierzu ausf. unter E.II.2.c.dd.

292 Vgl. *BVerfG*, BeckRS 1973, 104803 (Rn. 75); *Weichert*, ZD 2020, 18 (19); *Roßnagel*, ZD 2019, 157 (158); *T. Radtke*, ZGE 17 (2025), 1 (39 f.) m.w.N.

293 *LfDIBW*, Diskussionspapier: Rechtsgrundlagen im Datenschutz beim Einsatz von Künstlicher Intelligenz, S. 35.

nahme der angewandten Forschung legt nahe, dass auch Forschungsprojekte dem Anwendungsbereich unterfallen können, die auf einen praktischen Einsatz gerichtet sind.

In systematischer Hinsicht ist bei einer Gesamtbetrachtung mit den übrigen Öffnungsklauseln des Art. 9 Abs. 2 DSGVO aber jedenfalls ein Bezug zu einem konkreten Forschungsprojekt erforderlich, das nicht bloß Durchgangsstadium zu einem von vornherein feststehenden Produktiveinsatz ist. Denn eine zu weite Auslegung würde dazu führen, dass ansonsten letztlich die Entwicklung jeder Software und die Untersuchung ihrer Effektivität in der Praxis bereits als wissenschaftliche Forschung zu qualifizieren wäre, womit es bereits nicht mehr der übrigen Ausnahmen in Art. 9 Abs. 2 DSGVO bedürfte.

Vor diesem Hintergrund ist im GSJ-Projekt zwar grundsätzlich denkbar, einzelne Verarbeitungen, die schwerpunktmäßig der Forschung dienen, auf die Rechtsgrundlage des Art. 9 Abs. 2 lit. j DSGVO zu stützen (z.B. Debugging-Maßnahmen zu optimiertem Lernverhalten). Keinesfalls aber können sämtliche Verarbeitungen im Trainingsprozess und im späteren Einsatz hierauf gestützt werden.

c. Personenbezogene Daten über strafrechtliche Verurteilungen und Straftaten (Art. 10 DSGVO)

Sofern vereinzelt eine Verarbeitung personenbezogener Daten über strafrechtliche Verurteilungen und Straftaten[294] erfolgt, z.B. im Zusammenhang mit der Beiziehung von strafrechtlichen Akten im Zivilverfahren, ist Art. 10 S. 1 DSGVO zu beachten. Nach dieser Vorschrift darf die Verarbeitung nur unter behördlicher Aufsicht, aber ohne das Erfordernis einer über Art. 6 Abs. 1 DSGVO hinausgehenden Rechtsgrundlage,[295] oder gestützt auf eine entsprechende Rechtsvorschrift erfolgen. Die Einzelheiten des Anwendungsbereichs von Art. 10 DSGVO mit Blick auf Zeugen, schuldloses Handeln und mutmaßliche Straftaten sind umstritten,[296] wirken sich vorliegend mit Blick auf die Beschränkung auf zivilrechtliche Verfahren aber ohnehin

294 Ordnungswidrigkeiten sind nach vorzugswürdiger, aber bestrittener Ansicht aufgrund des Wortlauts nicht mitgemeint, weil die DSGVO gerade zwischen Straftaten und bspw. Bußgeldern nach Art. 83 DSGVO unterscheidet. Zum Meinungsstand *Nolde*, in: Taeger/Gabel, Art. 10 DSGVO Rn. 9 m.w.N.

295 *Petri*, in: Simitis/Hornung/Spiecker gen. Döhmann, Art. 10 DSGVO Rn. 1.

296 Überblick etwa bei *Nolde*, in: Taeger/Gabel, Art. 10 DSGVO Rn. 9 ff.

kaum aus. Besonders für etwaige Strafverfahren gilt es vorgelagert, die Voraussetzungen nach Art. 10 DSGVO zu prüfen und zu wahren.

Die behördliche Aufsicht schließt zwar nicht per se die Weitergabe an Dritte aus, verlangt aber eine in diesen Fällen fortbestehende Aufsicht und effektive Steuerung durch die Behörde.[297] Eine Sachverhaltskonstellation, die einer nicht-behördlichen Stelle eine eigenständige Gestaltung der Verarbeitung ermöglicht und eine gemeinsame Verantwortlichkeit begründet, wird diesen Anforderungen nicht gerecht. Die hier zwischen den Ministerien und den ausführenden Stellen anzunehmende Auftragsverarbeitung erfüllt grundsätzlich das Merkmal der behördlichen Aufsicht und damit die Anforderungen aus Art. 10 S. 1 Var. 1 DSGVO.

Eine spezifische Regelung i.S.d. zweiten Tatbestandsalternative des Art. 10 S. 1 DSGVO ist für das untersuchungsgegenständliche GSJ-Projekt nicht ersichtlich.

d. Verhältnis zu Anforderungen an die Zweckänderung (Art. 5 Abs. 1 lit. b, Art. 6 Abs. 4 DSGVO)

Die Frage nach dem Verhältnis von Anforderungen an die Zweckänderung nach Art. 6 Abs. 4 DSGVO und allgemeinen Rechtsgrundlagen[298] wirkt sich auch für Art. 9 DSGVO aus. Für die Kategorien personenbezogener Daten als maßgebliches Abwägungskriterium in Art. 6 Abs. 4 lit. c DSGVO könnte auch in diesen Fällen keine separate Rechtsgrundlage aus Art. 9 Abs. 2 DSGVO für eine Zweckänderung bei der Weiterverarbeitung durch den ursprünglichen Verantwortlichen erforderlich sein.

Im Ergebnis bedarf diese Frage vorliegend jedoch keiner Entscheidung. Denn selbst wenn man eine separate Rechtsgrundlage aus Art. 9 Abs. 2 DSGVO für nicht erforderlich hält, sind ähnliche Erwägungen mit hohem Gewicht in die Abwägung nach Art. 6 Abs. 4 DSGVO einzubeziehen. Damit sind nach beiden Lösungswege strenge Anforderungen an die Verarbeitung besonderer Kategorien personenbezogener Daten zu stellen.

297 *Frenzel*, in: Paal/Pauly, Art. 10 DSGVO Rn. 6; *Schiff*, in: Ehmann/Selmayr, Art. 10 DSGVO Rn. 7.

298 S. oben unter D.V.4.e.

e. Schlussfolgerungen für das untersuchungsgegenständliche Projekt

Da nach den vorherigen Ausführungen kein Rechtfertigungstatbestand aus Art. 9 Abs. 2 DSGVO vorliegen dürfte, auf den die Verarbeitungen des GSJ-Projekts belastbar in allen Fällen gestützt werden könnten, empfiehlt sich eine weitergehende Filterung und gegebenenfalls händische Anonymisierung mit Blick auf besondere Kategorien personenbezogener Daten.

Wenn und soweit durch das Erlanger Tool bereits ein Anonymisierungsstandard erreicht wurde, der den allgemeinen Anforderungen an eine Urteilsveröffentlichung entspricht, kann grundsätzlich auch die Weiterverarbeitung zu Zwecken des KI-Trainings als eingriffsarme Anonymisierungstechnik i.w.S. erfolgen.

Zunächst empfiehlt sich eine Clusterung nach:

- auszuschließenden Verfahrenskategorien (z.B. Asylrecht, Arzthaftungssachen, Strafrecht, Familienrecht) und
- zu berücksichtigenden Verfahrenskategorien (z.B. Mietrecht, Verkehrsrecht).

Beispielhaft können die folgenden Merkmale als Ausgangspunkte für eine weitergehende Filterung und gegebenenfalls händische Anonymisierung herangezogen werden:

- Geltendmachung von deliktsrechtlichen Schadensersatzansprüchen (insbesondere § 823 Abs. 2 BGB i.V.m. StGB);
- Geltendmachung von Schmerzensgeld nach § 253 Abs. 1, 2 BGB;
- Geltendmachung von Unterlassungsansprüchen mit Blick auf (politische) Aussagen, wobei gegebenenfalls eine offensichtliche Öffentlichmachung nach Art. 9 Abs. 2 lit. e DSGVO in Betracht kommen kann;
- Arztberichte innerhalb der Aktenauszüge;
- Beigezogene Strafakten als Teil der zivilrechtlichen Aktenauszüge;
- Vorkommen von Begriffen wie „Gewerkschaft“ sowie geläufigen Gewerkschafts- und Parteibezeichnungen innerhalb der Aktenauszüge.

6. Weitere Pflichten aus der DSGVO

Die weiteren Pflichten aus der DSGVO unterliegen keinen projektbezogenen Besonderheiten, sind aber von Relevanz mit Blick auf den Einsatz eines KI-Systems und werden daher nachfolgend im Überblick erläutert, soweit diese DSGVO-Pflichten gerade Fragen mit Blick auf das Training

des Sprachmodells und dessen Einsatz als KI-System in der Justiz aufwerfen.

a. Informationspflichten (Art. 13, 14 DSGVO)

Die Informationspflicht aus Art. 14 Abs. 1-4 DSGVO findet im Ergebnis aufgrund der in Art. 14 Abs. 5 lit. b Hs. 1 Var. 2 DSGVO niedergelegten Ausnahme des unverhältnismäßigen Aufwands der Informationserteilung keine Anwendung. Es sind allerdings geeignete Maßnahmen zum Schutz betroffener Personen zu treffen, wobei diese geeigneten Maßnahmen die Bereitstellung von Informationen für die Öffentlichkeit einschließen.[299]

Die Informationspflicht, auch im Hinblick auf eine Zweckänderung (siehe Art. 14 Abs. 4 DSGVO), richtet sich mangels Direkterhebung durch das Ministerium als erste relevante Stelle im Rahmen dieser Untersuchung nach Art. 14 DSGVO. Konkret wird regelmäßig eine Direkterhebung durch das jeweilige Gericht oder die Parteien und deren Bevollmächtigten erfolgt sein.

Ausnahmen von der Informationspflicht sind insbesondere vorgesehen, wenn die Erteilung der Informationen einen unverhältnismäßigen Aufwand erfordern würde (Art. 14 Abs. 5 lit. b Hs. 1 Var. 2 DSGVO) – so vor allem auch bei wissenschaftlichen Forschungszwecken. Die Artikel-29-Datenschutzgruppe verlangt insoweit einen erhöhten Begründungsaufwand für die Unverhältnismäßigkeit bei anderen als wissenschaftlichen Forschungszwecken.[300]

Die hier untersuchungsgegenständliche KI-Entwicklung dient zwar nicht primär einem Forschungsprojekt, weist aber eine Nähe zur Forschung i.S.d. in Art. 14 Abs. 5 lit. b Hs. 1 Var. 2 DSGVO verankerten Regelbeispiels auf. Diese Forschungsnähe zeigt sich nicht nur an dem Einsatz innovativer Technologien, sondern vor allem an dem Zugriff auf große Mengen an Daten, um hieraus losgelöst von den einzelnen betroffenen Personen neue Informationen bzw. Inhalte zu entwickeln. Weiterhin sind die hohe Anonymisierungsrate, die nur sehr vereinzelt mögliche Datenextraktion und das Vorhandensein von zusätzlichen Schutzmaßnahmen zu berücksichtigen. Nach alledem wird das untersuchungsgegenständliche GSJ-Projekt im Zuge

299 Diese Informationsbereitstellung kann sich gegebenenfalls auf die urheberrechtliche Bewertung auswirken, indem vermehrt mit der Erklärung von Vorbehalten i.S.d. § 44b Abs. 3 UrhG zu rechnen ist, s. E.II.2.c.cc(3).

300 *Artikel-29-Datenschutzgruppe*, WP 260 rev.01, Rn. 61

einer Gesamtbetrachtung ebenfalls auf diese Ausnahme gestützt werden können.

Die Ausnahme der ernsthaften Beeinträchtigung einer Verwirklichung der Ziele der Verarbeitung (Art. 14 Abs. 5 lit. b Hs. 2 DSGVO) zielt vor allem auf Verarbeitungen, die gerade eine Geheimhaltung erfordern (z.B. interne Ermittlungen), und wird daher im vorliegenden Fall grundsätzlich nicht zur Anwendung gelangen.

Der Verantwortliche muss in jedem Fall weitere Schutzmaßnahmen prüfen. Hierzu zählt insbesondere die regelmäßig erforderliche[301] öffentliche Bereitstellung dieser Informationen, z.B. über eine Internetseite. Die weiteren Maßnahmen können etwa eine Datenschutzfolgenabschätzung sowie technische und organisatorische Maßnahmen umfassen; insoweit kann auf die nachfolgenden Ausführungen unter c. verwiesen werden.

b. (Weitere) Betroffenenrechte und Datenschutzgrundsätze

Darüber hinaus sind die weiteren Anforderungen aus den Art. 15 ff. DSGVO zu beachten, die beispielsweise auf Berichtigungs- und Löschungsersuchen abstellen, zugleich aber auch eigeninitiativ über die Datenschutzgrundsätze i.V.m. Art. 24 ff. DSGVO den Verantwortlichen auferlegt werden können. Diese Anforderungen an die Ausgestaltung stellt den Umgang mit KI-Systemen, wie er im GSJ-Projekt vorgesehen ist, vor Herausforderungen.[302]

Für diese Herausforderungen können in der Praxis zwar belastbare Lösungen gefunden werden, die aber wegen der charakteristischen Besonderheiten von KI-Systemen das Risiko eines Verstoßes gegen Art. 5, 15 ff. DSGVO vielfach nicht werden restlos ausräumen können. Im GSJ-Projekt dürfte es aufgrund der grundsätzlich nur justizinternen Weiterverarbeitung der KI-Ausgabe[303] allerdings ohnehin allenfalls selten zur Geltendmachung von Betroffenenersuchen kommen.

Im Fall der Geltendmachung ist der Verantwortliche nach Art. 11 Abs. 1, 2 DSGVO gegebenenfalls nicht zur eigenständigen Verknüpfung der Trai-

301 *Paal/Hennemann*, in: Paal/Pauly, Art. 14 DSGVO Rn. 41; a.A. *Schmidt-Wudy*, in: BeckOK Datenschutzrecht, Art. 14 DSGVO Rn. 99.

302 *Berz/Engel/Hacker*, ZUM 2023, 586 (589); *Werry*, MMR 2023, 911 (914).

303 Denkbar bleibt in den Use-Cases allerdings eine Übernahme generierter KI-Ausgaben in eine Gerichtsentscheidung, die dann der Öffentlichkeit zugänglich gemacht werden.

ningsdaten im Einzelfall zur Identifizierung der betroffenen Personen verpflichtet, muss aber zusätzliche, von der betroffenen Person bereitgestellte Informationen zur Identifizierung und Bearbeitung eines Ersuchens berücksichtigen.[304]

aa. Auskunft und Datenkopie (Art. 15 DSGVO)

Nach Art. 15 Abs. 1, 3 DSGVO hat die betroffene Person ein Recht auf Auskunft und auf eine Kopie der verarbeiteten personenbezogenen Daten. Im Hinblick auf den Einsatz des KI-Systems dürfte vor allem die nach Art. 15 Abs. 1 S. 1 Hs. 1 DSGVO vorzunehmende Bestätigung über die Verarbeitung personenbezogener Daten die Verantwortlichen vor Herausforderungen stellen. Denn der Verantwortliche hat regelmäßig gerade keine Kenntnis davon, ob Daten einer einzelnen Person verarbeitet werden oder nicht. Erst wenn dem Verantwortlichen ein Datenleck als Ergebnis einer KI-Ausgabe bekannt wird, kann der Verantwortliche die betreffende Datenverarbeitung bestätigen, gegebenenfalls anhand des Trainingskorpus lokalisieren und die weiteren Informationen nach Art. 15 Abs. 1 DSGVO bereitstellen.

Mithin besteht das Risiko, einem Betroffenenersuchen nach Art. 15 DSGVO mangels Kenntnis des Verantwortlichen über die konkreten Datenverarbeitungen nicht hinreichend entsprechen zu können. Für diesen Fall kann unter bestimmten Voraussetzungen ein Unmöglichkeitseinwand ins Feld geführt werden, der sich in der Rechtsprechung des EuGH zumindest für die Auskunft über einzelne Informationskategorien abzeichnet.[305] Ob der EuGH einen solchen Einwand in diesem Umfang anerkennt, bleibt allerdings abzuwarten. Der EuGH könnte insoweit auch bloß eine aufgrund der Technologie selbst herbeigeführte Unmöglichkeit erkennen.[306]

Das vorgenannte Risiko eines Verstoßes gegen Art. 15 DSGVO wird aber durch folgende Überlegung zumindest begrenzt: Das Sprachmodell weist regelmäßig für sich genommen keinen Personenbezug auf,[307] sondern es kommt erst durch eine Zusammenführung verschiedener Informationen in der KI-Ausgabe zu einer Verarbeitung i.S.d. DSGVO. Diese Zusammenführung im Rahmen der Ausgabe wird grundsätzlich durch verschiedene Maßnahmen verhindert. Sofern es zu einer solchen Zusammenführung

304 Hierzu etwa *Ashkar*, ZD 2023, 523 (528).
305 *EuGH*, NJW 2023, 973 (Rn. 48) – Österreichische Post.
306 *Mertens/D. Meyer*, K&R 2023, 563 (568).
307 S. unter D.IV.1.

kommt, ist durch technische und organisatorische Maßnahmen sicherzustellen, dass die Verantwortlichen über bekanntgewordene Datenlecks und damit auch über das Vorliegen einer Verarbeitung informiert werden, z.B. durch die Möglichkeit für die das KI-System einsetzende Richterschaft, die KI-Ausgabe zur Überprüfung zu melden.

bb. Berichtigung und Mitteilungspflicht (Art. 16 und 19 DSGVO)

Nach Art. 16 DSGVO hat die betroffene Person das Recht, die Berichtigung unrichtiger bzw. Vervollständigung unvollständiger, sie betreffender personenbezogener Daten zu verlangen. Die Erfüllung dieses Rechts kann aus den vorgenannten Gründen in Ansehung von KI-Systemen (ebenfalls) nicht nur unerhebliche Schwierigkeiten für Verantwortliche aufwerfen.

Allerdings ist zu bedenken, dass die Richtigkeit oder Vollständigkeit der personenbezogenen Daten aus den Trainingsdaten selten mit Erfolg wird bestritten werden können.[308] Denn die vollständige Wiedergabe einer Gerichtsentscheidung in den Trainingsdaten als Dokumentation der Entscheidung des Gerichts ist gerade nicht unrichtig. Denn die Richtigkeit verlangt eine Übereinstimmung der Informationen[309] mit der Realität[310] unter Berücksichtigung des Informationskontexts:[311] Die Information muss als Tatsache geäußert werden,[312] z.B. als durch objektive oder durch ein Gericht ordnungsgemäß festgestellte Tatsache,[313] oder im Sinne eines graduellen Ansatzes zumindest einen Tatsachengehalt in Form von – gegebenenfalls auch wertenden – Prognosen enthalten.[314]

308 Anders z.B. beim Scraping *Mertens/D. Meyer*, K&R 2023, 563 (569).

309 Also nicht bloß den Prozess der mathematisch richtigen Erzeugung, *Engeler/Rolfes*, ZD 2024, 423 (428).

310 *Herbst*, in: Kühling/Buchner, Art. 5 DSGVO Rn. 60; *Voigt*, in: Taeger/Gabel, Art. 5 DSGVO Rn. 30.

311 *Stevens*, CR 2020, 73 (Rn. 8) m.w.N.

312 *Herbst*, in: Kühling/Buchner, Art. 5 DSGVO Rn. 60.

313 Vgl. auch *Roßnagel*, in: Simitis/Hornung/Spiecker gen. Döhmann, Art. 5 DSGVO Rn. 140.

314 *Hoeren*, ZD 2016, 459 (462) unter Berücksichtigung der unterschiedlichen Sprachfassungen; *Hallinan/Zuiderveen Borgesius*, International Data Privacy Law 10 (2020), 1; *Stevens*, CR 2020, 73 (Rn. 9); *Schantz*, in: BeckOK Datenschutzrecht, Art. 5 DSGVO Rn. 27; im Hinblick auf die Datengrundlagen für automatisierte Entscheidungen *Artikel-29-Datenschutzgruppe*, WP251 rev.01, S. 12 f.; diff. *Pesch/Böhme*, MMR 2023, 917 (923); die Vorschrift hingegen auf reine Tatsachenbehauptungen beschränkend *Herbst*, in: Kühling/Buchner, Art. 5 DSGVO Rn. 60; *Werry*,

Erst die neue Zusammensetzung durch die Generierung der KI-Ausgabe kann zu einer Unrichtigkeit führen (z.B. in Gestalt einer „Halluzination"),[315] sofern man annimmt, das KI-System stelle insoweit überhaupt eine Behauptung mit Tatsachengehalt auf.[316] Ein solcher Tatsachengehalt wird insbesondere in Abhängigkeit davon anzunehmen sein, wie zuverlässig das KI-System eingeschätzt[317] und wofür es eingesetzt wird. Im untersuchungsgegenständlichen GSJ-Projekt dürften gelegentliche personenbezogene[318] Ausgaben aus den Trainingsdaten sowie den Eingabedaten aus der neuen Akte in den Use-Cases einer Sachverhaltszusammenstellung einer neuen Akte als Tatsache wahrgenommen werden. Die Unrichtigkeit der Angaben kann aber gegebenenfalls durch diese entsprechende Temperatur-Einstellung eingedämmt werden.[319]

Unter dieser Prämisse müsste die betreffende KI-Ausgabe, soweit sie denn gespeichert wurde,[320] durch den Verantwortlichen bearbeitet oder gelöscht und müssten etwaige Empfänger nach Art. 19 DSGVO hierüber in Kenntnis gesetzt werden. Sofern eine Berichtigung im Gesamtkontext der KI-Ausgabe nicht zielführend erscheint, wird es mit Blick auf die gerade unbeabsichtigte und aus Sicht des Verantwortlichen zu vermeidende Verarbeitung sowie der regelmäßig verwirklichten Befriedigung der Interessen der betroffenen Personen durch die Löschung als zulässig anzusehen sein, erst recht und nur die Löschung oder gegebenenfalls Einschränkung der Verarbeitung (Art. 17 und 18 DSGVO) vorzunehmen.[321]

cc. Löschung und Mitteilungspflicht (Art. 17 und 19 DSGVO), auch im Zusammenhang mit einem Widerspruch (Art. 21 DSGVO)

Die Löschung ist von Gesetzes wegen (Art. 5 Abs. 1, Art. 25 Abs. 1 DSGVO) oder auf Ersuchen einer betroffenen Person hin auch und gerade dann

---

MMR 2023, 911 (914); wohl auch *Roßnagel*, in: Simitis/Hornung/Spiecker gen. Döhmann, Art. 5 DSGVO Rn. 140.

315 Hierzu schon oben unter D.IV.1.b.dd(2) und D.IV.2.

316 Hierzu auch *EDPB*, Report of the work undertaken by the ChatGPT Taskforce, Rn. 30 f.; dies i.E. annehmend *Pesch/Böhme*, MMR 2023, 917 (922).

317 *Pesch/Böhme*, MMR 2023, 917 (922).

318 Also gerade nicht Angaben bezogen auf eine fiktive Person, s. *LfDIBW*, Diskussionspapier: Rechtsgrundlagen im Datenschutz beim Einsatz von Künstlicher Intelligenz, S. 9.

319 S. bspw. https://platform.openai.com/docs/api-reference/chat/create.

320 Vgl. *Paal*, in: Paal/Pauly, Art. 16 DSGVO Rn. 15.

321 Vgl. auch *EDPB*, Report of the work undertaken by the ChatGPT Taskforce, Rn. 34.

vorzunehmen, wenn die betroffene Person wirksam einen Widerspruch eingelegt hat (Art. 21 Abs. 1 DSGVO) und keine vorrangigen berechtigten Gründe des Verantwortlichen vorliegen (Art. 17 Abs. 1 lit. c DSGVO). Derartige Fallkonstellationen dürften bei der justizinternen Verarbeitung kaum vorkommen.

Falls in einem KI-basiert vorformulierten Sachverhaltsabschnitt einer Gerichtsentscheidung (siehe den zweiten Use-Case) personenbezogene Daten aus den Trainingsdaten enthalten sind und über die Gerichtsentscheidung der Öffentlichkeit zugänglich gemacht werden, mag die Löschung nach Art. 17 DSGVO allerdings relevant werden. Die nach Art. 6 Abs. 1 UAbs. 1 lit. e, Abs. 4 DSGVO (oder Art. 6 Abs. 1 UAbs. 1 lit. f DSGVO) vorgenommene Interessenabwägung[322] ist dann gegebenenfalls für den konkreten Einzelfall zu überprüfen.

Der Verantwortliche[323] muss nach dem Eingang eines berechtigten Löschersuchens sicherstellen, dass das Datenleck den jeweiligen Gerichtsentscheidungen aus dem Trainingsdatensatz zugeordnet wird und die Urteile in dem Training künftiger KI-Modelle ausgeschlossen oder vollständig anonymisiert werden. Das gilt in Ansehung der erzielten Ausgabe als – gegebenenfalls erneute reproduzierbare – Verarbeitung personenbezogener Daten unabhängig davon, ob das Sprachmodell selbst als eine Speicherung personenbezogener Daten anzusehen ist.

Nach Art. 19 DSGVO sind erforderlichenfalls etwaige Empfänger über die Löschung zu informieren. In diesem Zusammenhang sollte der Verantwortliche als technische und organisatorische Maßnahme zur Sicherstellung der Rechtmäßigkeit der Verarbeitung (Art. 5 Abs. 1 lit. a, Art. 24 f. DSGVO) gegenüber den Nutzern des KI-Systems eine (dienst-)vertragliche Meldepflicht etablieren oder zumindest auf die Meldemöglichkeit prominent hinweisen.

Die Löschung muss grundsätzlich binnen einer Frist von einem Monat nach Art. 12 Abs. 3 S. 1 DSGVO erfolgen; die Löschungsfrist verlängert sich ausnahmsweise „unter Berücksichtigung der Komplexität und der Anzahl von Anträgen“ um zwei weitere Monate. Diese Ausnahme ist restriktiv anzuwenden (vgl. ErwGr. 59 DSGVO). Zwar kann die Schwierigkeit, die konkreten Informationen auszulesen, im Hinblick auf einzelne Anträge eine

322 S. oben unter D.V.4.c.

323 In der Regel die jeweiligen Ministerien, s. oben unter D.V.3.b.dd.

Fristverlängerung rechtfertigen,[324] nicht aber pauschal der Einsatz eines besonderen (KI-)Systems eine Fristverlängerung zum Regelfall erheben.

Der Verantwortliche kann sich insoweit jedoch gegebenenfalls anderweitig behelfen: Unter Beachtung des oben aufgestellten Maßstabs[325] kann die primäre Anonymisierung und sekundäre eingeschränkte justizinterne Verarbeitung grundsätzlich auf Art. 6 Abs. 1 UAbs. 1 lit. e , f DSGVO oder Art. 6 Abs. 4 DSGVO gestützt werden. Das gilt im Einzelfall auch für besondere Kategorien personenbezogener Daten nach Art. 9 Abs. 2 lit. f DSGVO. Einzelne Löschungen können auf Ausgabeebene realisiert werden: System-Prompts und Ausgabenfilter, die die konkrete Verarbeitung im Rahmen der Ausgabe verhindern, können faktisch einer Löschung gleichstehen oder gar eine Anonymisierung herbeiführen, falls keine wahrscheinlichen Mittel zur Reproduktion der personenbezogenen Daten in der Ausgabe in Betracht kommen.

Den Befund der Zulässigkeit dieses Vorgehens stützt überdies § 35 Abs. 1 BDSG, der eine Ausnahme von dem Recht auf Löschung definiert. Für Daten i.S.d. Art. 9, 10 DSGVO, die trotz der ergriffenen Maßnahmen offengelegt werden, kann im Einzelfall abhängig von der Komplexität eine Fristverlängerung nach Art. 12 Abs. 3 S. 2 DSGVO gerechtfertigt sein.

dd. Einschränkung der Verarbeitung und Mitteilungspflicht (Art. 18 und 19 DSGVO)

Die Einschränkung der Verarbeitung (Art. 18 DSGVO) als „Einfrieren" zur Klärung oder zu Dokumentationszwecken und zur Mitteilung nach Art. 19 DSGVO lässt sich entsprechend umsetzen, z.B. wenn die Richtigkeit einer Ausgabe bestritten wird oder die betroffene Person Widerspruch eingelegt hat und (noch) nicht feststeht, ob die berechtigten Gründe des Verantwortlichen gegenüber denen der betroffenen Person überwiegen. Wie bereits unter D.V.5.b aufgezeigt, dürften diese Fälle im GSJ-Projekt praktisch allenfalls (sehr) selten vorkommen.

324 *EDPB*, Leitlinien 01/2022 zu den Rechten der betroffenen Person – Auskunftsrecht, Rn. 163.

325 S. oben unter D.V.4.

c. Datenschutzfolgenabschätzung und Pflicht zur Implementierung technischer und organisatorischer Maßnahmen

Als eine besondere technische und organisatorische Maßnahme ist die Datenschutzfolgenabschätzung (Art. 35 Abs. 1 DSGVO) hervorzuheben, die insbesondere die Anonymisierungsmaßnahmen (einschließlich der Ausgabefilter) und das verbleibende Restrisiko nachvollzieht. Ein solches Erfordernis ergibt sich zwar nicht zwingend aus Art. 35 Abs. 3 DSGVO, kann aber gleichwohl mit Blick auf die Verwendung neuer Technologien i.S.d. Art. 35 Abs. 1 DSGVO[326] sowie die Quantität der möglicherweise verarbeiteten personenbezogenen Daten einschließlich der zu vermeidenden Verarbeitung besonderer Kategorien personenbezogener Daten i.S.d. Art. 9, 10 DSGVO angezeigt sein (vgl. Art. 35 Abs. 3 lit. b DSGVO).[327]

Darüber hinaus sind zur Umsetzung der Datenschutzgrundsätze, so auch und gerade der Datenminimierung, Speicherbegrenzung, Integrität und Vertraulichkeit (Art. 5 Abs. 1 lit. c, e, f DSGVO) technische und organisatorische Maßnahmen zu ergreifen (Art. 24, 25, 32 DSGVO). Das gilt auch mit Blick auf die landesrechtlichen Vorgaben (z.B. § 3 DSG NRW). Diese Verpflichtung umfasst die regelmäßige Überprüfung, ob das eingesetzte Anonymisierungstool, die Auswahl der ein- und ausgeschlossenen Urteilskategorien, die Ausgabefilter, der System Prompt sowie die eingesetzte Software und Übermittlungswege dem Stand der Technik und IT-Sicherheitsanforderungen entsprechen.[328] Diese Überprüfung sollte proaktiv in regelmäßigen Abständen und zudem als Reaktion auf etwaige Datenschutzverletzungen i.S.d. Art. 33, 34 DSGVO erfolgen. Grundlage für diese Überprüfung ist eine sorgfältige Dokumentation einschließlich der Informationen über die verwendeten Trainingsdaten[329] und erfolgten Anonymisierungsmaßnahmen.

Zudem empfiehlt es sich, den Nutzern des KI-Systems eine einfache Möglichkeit zur Verfügung zu stellen, um KI-Ausgaben gegebenenfalls zur Überprüfung an die Verantwortlichen zu melden. Eine generelle Markie-

326 *Ashkar*, ZD 2023, 523 (529).

327 Vgl. *Artikel-29-Datenschutzgruppe*, WP248 rev.01, S. 10 ff.; *Dewitte*, in: AI Meets the GDPR, S. 155.

328 Zur Handlungspflicht bei Änderungen des Stands der Technik *Hornung/B. Wagner*, CR 2019, 565 (Rn. 25 ff.).

329 *BayLDA*, Datenschutzkonforme Künstliche Intelligenz, S. 5; *EDPB*, Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, Rn. 56-58.

rung der KI-Ausgaben mit einem Wasserzeichen[330] kann darüber hinaus eine Nachvollziehbarkeit des Ursprungs der KI-Ausgabe bei unbeabsichtigter Weiterverarbeitung sicherstellen.

Nur kurz erwähnt sei an dieser Stelle[331] der Abschluss von etwaigen notwendigen datenschutzrechtlichen Vereinbarungen (insbesondere einer Auftragsverarbeitungsvereinbarung) sowie Nutzungsvereinbarungen bzw. Nutzungshinweisen.

d. Grundsatz der Speicherbegrenzung

Das Trainingskorpus kann für weitere Trainings des Sprachmodells gespeichert werden, da die Verarbeitung erforderlich ist für das KI-Training.[332] Eine restriktive Auslegung der Erforderlichkeit i.S.d. Art. 5 Abs. 1 lit. c, e DSGVO ist nicht geboten, wenn der Verantwortliche durchgehend die vorgenannten Anforderungen an die technischen und organisatorischen Maßnahmen erfüllt.

Mit Blick auf Besonderheiten einzelner Sachverhalte bedarf es jedoch gegebenenfalls einer Prüfung einzelner Datencluster, wenn sich die Sensibilität personenbezogener Daten durch Zeitablauf erhöht und sich die Abwägung nach Art. 6 Abs. 1 UAbs. 1 lit. e DSGVO im Laufe der Zeit unter Berücksichtigung des Zwecks tendenziell zugunsten der betroffenen Person verschiebt. Von Bedeutung sind insoweit beispielsweise Informationen über eine finanzielle Notlage,[333] wie etwa im Zusammenhang mit der Privatinsolvenz, Entscheidungen der Sozialgerichtsbarkeit oder vollstreckungsgerichtliche Entscheidungen.

Für verschiedene weitere Urteilscluster mag zwar die Intensität eines Datenlecks mit der Zeit zunehmen; dieser Effekt dürfte aber zugleich durch eine höhere Anonymisierungswahrscheinlichkeit mit Zeitablauf aufgehoben werden. Sofern beispielsweise einer Person aufgrund einer Ortsbeschreibung im Urteil bestimmte Informationen zugeordnet werden konnten, dürfte diese Zuordnung mit Zeitablauf (in Verbindung etwa mit Um-

330 *EDPB*, Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, Rn. 107.

331 S. zu weiteren möglichen Maßnahmen auch *EDPB*, Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, Rn. 102: Wartefrist und Opt-Out-Möglichkeit.

332 *Paal*, ZfDR 2024, 129 (141).

333 Vgl. *EuGH*, NJW 2024, 418 – Schufa; NJW 2014, 2257 – Google Spain.

zügen und Veränderungen der örtlichen Umstände) schwerer fallen und gegebenenfalls sogar ausscheiden.

7. Pflichten unter der JI-RL

Die JI-RL sieht besondere datenschutzrechtliche Regelungen vor für die „Verarbeitung personenbezogener Daten durch die zuständigen Behörden zum Zwecke der Verhütung, Ermittlung, Aufdeckung oder Verfolgung von Straftaten oder der Strafvollstreckung, einschließlich des Schutzes vor und der Abwehr von Gefahren für die öffentliche Sicherheit", die in den §§ 45 ff. BDSG und landesrechtlichen Regelungen[334] umgesetzt wurden. Hierunter fallen insbesondere Strafsachen und die zugehörigen Ermittlungsakten, die im GSJ-Projekt jeweils ausgenommen sind.

Vor diesem Hintergrund ist keine nähere Betrachtung der betreffenden Regelungen für das untersuchungsgegenständliche Projekt angezeigt.

8. Ergänzende Anforderungen mit Blick auf Eingabedaten

Mit Blick auf personenbezogene Daten in den Eingabeaufforderungen des entwickelten KI-Systems sind ebenfalls die oben dargestellten datenschutzrechtlichen Anforderungen zu beachten. Als Hilfsmittel im Rahmen der Bearbeitung eines konkreten Falls in Betracht kommt grundsätzlich auch die nicht-anonymisierte Eingabe im Zusammenhang mit den öffentlichen Aufgaben der Rechtspflege i.S.d. Art. 6 Abs. 1 UAbs. 1 lit. e DSGVO i.V.m. landesrechtlichen Generalklauseln. Allerdings sind die zuvor aufgestellten Maßstäbe entsprechend anzuwenden, wenn die Eingabedaten unmittelbar und ohne weitere Anonymisierung zum Zwecken des weiteren Trainings des KI-Systems verwendet werden. Eine solche Verwendung wirft nicht nur Bedenken mit Blick auf eine Rechtsgrundlage nach Art. 6 DSGVO auf, sondern löst zusätzlich gegebenenfalls ergänzende Anforderungen aus der gemeinsamen Verantwortlichkeit aus.[335]

334 S. zum Verhältnis insb. § 1 Abs. 1 S. 1 Nr. 2 BDSG.
335 Hierzu ausf. *Hüger/T. Radtke*, KIR 2025, 154.

VI. Besonderheiten bei der Einbeziehung von Aktenauszügen

Für die Verwertung vollständiger Aktenauszüge und deren Bearbeitung durch das Erlanger Tool sind die Besonderheiten der Informationsanordnung zu berücksichtigen.

1. Inhalte von gerichtlichen Akten

Zunächst gilt es, sich hierzu der typischen Akteninhalte zu vergewissern, wobei diese typischen Akteninhalte hier in Ansehung der Vielfältigkeit nicht abschließend, sondern nur exemplarisch beschrieben werden können:

- Anwaltliche Schriftsätze (z.B. Klage und Klageerwiderung);
- Anlagen zu anwaltlichen Schriftsätzen (z.B. Wiedergabe von Zeitungsartikeln, Internetseiten, Fotografien, eidesstattliche Versicherungen und Zeugenaussagen);
- Urkunden und amtliche Auskünfte;
- Sitzungsprotokolle;
- Hinweis-, Beweis- und sonstige Beschlüsse;
- Terminsladungen und zusammenhängende gerichtliche Verfügungen;
- Beigezogene Akten (z.B. Ermittlungsakten der Staatsanwaltschaft);
- Vermerke;
- Gutachten (z.B. Unfallgutachten oder psychologische Gutachten);
- Empfangsbekenntnisse;
- Gerichtliche (End-)Entscheidung, z.B. in Form eines Urteils;
- Handschriftliche Anmerkungen auf Aktenbestandteilen, Post-Its o.ä., die gegebenenfalls die digitale Bearbeitung erschweren.

Diese Aktenbestandteile können in unterschiedlicher Quantität und Qualität personenbezogene Daten enthalten. Das betrifft insbesondere Kennungen, wie etwa Namen, Anschriften, E-Mail-Adressen, Telefonnummern und Identifikationsnummern (z.B. der Parteien, Bevollmächtigter und Kanzleiangestellter, Zeugen, Gutachter sowie Angehörige der Staatsanwaltschaft, der übrigen Rechtspflege oder sonstiger Behörden). Hinzu kommen auch hier[336] – in einer gegenüber einer Gerichtsentscheidung unstrukturierten Form – Informationen über Aktenzeichen, Ort- und Zeitangaben sowie

336 S. oben unter D.I.3.

sonstige Sachverhaltsumstände, die unter Zuhilfenahme weiterer Quellen die Identifizierung natürlicher Personen ermöglichen.

Aus dem Aufbau und Inhalt der Aktenauszüge ergeben sich also vor allem technische Herausforderungen betreffend eine qualitativ gleichbleibende Anonymisierung.

2. Rechtliche Leitplanken für die (Teil-)Anonymisierung der Aktenauszüge

Für diese technischen Herausforderungen gilt es, rechtliche Leitplanken zu herauszuarbeiten, die einerseits quantitativen Anforderungen im Hinblick auf den Recall-Wert bzw. Anonymisierungsanteil genügen, andererseits aber auch qualitative Anforderungen an die zu anonymisierenden Datenkategorien beinhalten.

a. Kein Veröffentlichungsstandard

Im Ausgangspunkt ist festzuhalten, dass die Veröffentlichung von Aktenauszügen gerade keine öffentliche, verfassungsrechtlich gebotene Aufgabe ist. Anders als bei Gerichtsentscheidungen[337] existiert also kein Anonymisierungsgrad, ab dem ohnehin eine Veröffentlichung und damit regelmäßig auch eine Weiterverarbeitung zulässig wäre.

b. Quantitativ: Recall

Die für die Anonymisierung von Gerichtsentscheidungen angenommenen Recall-Werte können zwar grundsätzlich den datenschutzrechtlichen Anforderungen auf der ersten Stufe der Anonymisierung und der zweiten Stufe der Rechtfertigung nach Art. 6 Abs. 1 UAbs. 1 lit. e, f, Abs. 4 DSGVO genügen, sofern man das KI-Training sowie technische und organisatorische Maßnahmen berücksichtigt. Es dürfte aber der Recall-Wert i.H.v. 94-96 % für Kennungen und i.H.v. 68-96 % für sonstige identifizierende Merkmale signifikant unterschritten werden. Dieser reduzierte Anonymisierungsgrad kann sich sowohl auf die Frage auswirken, ob die Daten mangels hinreichender Identifizierungswahrscheinlichkeit anonym sind, als auch auf die

337 S. hierzu ausf. unter D.II.5.b.

Rechtfertigung nach Art. 6 Abs. 1 UAbs. 1 lit. e DSGVO bei vereinzelt vorkommenden personenbezogenen Daten.

Von besonderer Relevanz bleibt auch bei Aktenauszügen ein hoher Recall-Wert mit Blick auf Kennungen. Für die übrigen identifizierenden Merkmale dürfte bei der justizinternen Verarbeitung keine hinreichende Wahrscheinlichkeit bestehen, dass mittels weiterer Quellen ein Personenbezug hergestellt wird. Für diese sonstigen identifizierenden Merkmale dürfte daher auch ein Recall-Wert von unter 50 % nicht schaden, solange die qualitativen Anforderungen gewahrt bleiben.

Enthaltene Kennungen führen hingegen unmittelbar zu einem Personenbezug. Je häufiger es zu Verarbeitungen von derartigen Kennungen kommt, desto schwerer fällt die Annahme eines belastbaren Zusammenhangs der Verarbeitung der verfahrensfremden Trainingsdaten in einem neuen Verfahren als Wahrnehmung einer öffentlichen Aufgabe i.S.d. Art. 6 Abs. 1 UAbs. 1 lit. e, Abs. 4 DSGVO.

Geringere Abweichungen des Recall-Werts für Kennungen nach unten müssen sich zunächst nicht unmittelbar auf die Frage der Anonymisierung und der Rechtfertigung auswirken. Aufgrund des Trainings des Sprachmodells und der Abhängigkeit der Ausgabe von der konkreten Eingabeaufforderung ist es selbst bei einem Recall-Wert von beispielsweise nur 90 % denkbar, dass es in weniger als 0,1 % der Fälle des konkreten KI-System-Einsatzes zu einer Ausgabe dieser Kennungen kommt. In den konkreten Use-Cases 1-3, in denen regelmäßig ein umfangreicher Eingabekontext und eine auf diese abstellende Eingabeaufforderung übergeben werden, dürfte es – vorbehaltlich einer technisch-empirischen Untersuchung – höchst unwahrscheinlich sein, dass in ein Relationsgutachten oder eine Sachverhaltszusammenfassung ein vollständiger Name oder eine Anschrift „aus" den Trainingsdaten einfließt.

Eine signifikante Unterschreitung dieses Recall-Werts von 90 % mit Blick auf Kennungen führt aber zu einer erhöhten Wahrscheinlichkeit von nicht mehr nach Art. 6 DSGVO zulässigen Verarbeitungen personenbezogener Daten. Ein solcher Mindest-Recall-Wert dürfte sich auch unter Berücksichtigung weiterer technischer und organisatorischer Maßnahmen nach dem Stand der Technik erzielen lassen. Sofern Anonymisierungsdefizite aus einer schlecht(er)en digitalen Lesbarkeit (z.B. im Rahmen des Einsatzes von Optical Character Recognition (OCR)-Software) herrühren, können vereinzelt Kennungen schon gar nicht erst erfasst werden – und hierdurch eine höhere Anonymisierung auslösen.

Soweit zugleich Kennungen einfließen, die aufgrund von Formatänderungen nicht erkannt werden (z.B. falsch zugeordnete Zeichen, zusätzlich ergänzte Leerzeichen innerhalb eines Namens, einzelne wegen handschriftlicher Anmerkungen nicht lesbare Adressbestandteile), empfiehlt sich die Implementierung eines gesonderten Mechanismus zur Qualitätssicherung. Dokumentbestandteile, deren Digitalisierung Herausforderungen mit sich bringt, könnten separat identifiziert und gegebenenfalls aussortiert oder nachbearbeitet werden.

Ferner kann der ursprüngliche Recall-Wert stets faktisch durch effektive Ausgabefilter erhöht werden, die insbesondere typische Kennungsformate (z.B. Namen, E-Mail-Adressen, Telefonnummern und Identifikationsnummern) anhand ihres Formats durch sog. reguläre Ausdrücke erkennen und mit deren Vorkommen im Eingabekontext abgleichen. Nur wenn sich diese Angaben auch im Eingabekontext wiederfinden, sprich dem konkreten Verfahren und nicht den Trainingsdaten zuzuordnen sind, wird die Ausgabe insoweit zugelassen.

c. Qualitativ: Clusterung

Mit Blick auf die besonderen Kategorien personenbezogener Daten nach Art. 9 DSGVO erlangen die rechtlichen Leitplanken in qualitativer Hinsicht besondere Bedeutung.

Im Ausgangsfall stellt in Ansehung der justizinternen Verwendung des KI-Systems vor allem eine kombinierte Ausgabe von Kennungen (siehe zuvor zu den quantitativen Anforderungen) und besonderen Kategorien personenbezogener Daten die Verantwortlichen vor erhebliche Herausforderungen. Während die Kennung regelmäßig den Trainingsdaten entstammen dürfte, könnten unter Umständen auch halluzinierte Gesundheitsdaten oder ähnlich sensible Informationen einen besonderen Rechtfertigungsbedarf nach Art. 9 DSGVO auslösen.[338] Diese Kombination kann bereits ausreichend adressiert werden, indem das Vorkommen von Kennungen in den Ausgabedaten reduziert wird.

Darüber hinaus sind an dieser Stelle besondere Kategorien personenbezogener Daten aus den Trainingsdaten in den Blick zu nehmen, deren Vorkommen in einer KI-Ausgabe zusammen mit Kennungen aus den Trainingsdaten strikt zu verhindern ist. Hierzu bedarf es einer entsprechenden Clusterung der entsprechenden Bestandteile der Aktenauszüge.

338 S. zu Halluzinationen z.B. schon oben unter D.V.6.b.bb.

Der Verantwortliche ist insoweit als Ausfluss der Pflicht zur Implementierung von technischen und organisatorischen Maßnahmen nach den Art. 24 ff. DSGVO gefordert, sich einen Überblick über die in den Aktenauszügen typischerweise enthaltenen Dokumentenkategorien und deren Sensibilität mit Blick auf Art. 9 DSGVO zu verschaffen.

Auf dieser Grundlage sind einzelne Dokumentenkategorien entweder vollständig auszunehmen oder deren Anonymisierung unter Berücksichtigung des gesamten Akteninhalts händisch sorgfältig zu prüfen. Das kann z.B. – unter der Prämisse eines nicht hinreichend hohen Recall-Werts – die folgenden Dokumentenkategorien betreffen, die typischerweise eine Vielzahl an besonderen Kategorien personenbezogener Daten enthalten:

- Beigezogene Akten der Staatsanwaltschaft im Zusammenhang mit Verfahren, die typischerweise über die Straftat i.S.d. Art. 10 DSGVO hinaus sensible Informationen enthalten (z.B. Straftaten gegen die sexuelle Selbstbestimmung, schwere Körperverletzungsdelikte);
- Psychologische Gutachten;
- Ärztliche Gutachten.

Weitere Dokumentenkategorien sind gegebenenfalls nur durch einen speziell angepassten Anonymisierungsalgorithmus oder nach händischer Nachbearbeitung in das Trainingskorpus aufzunehmen. Das kann z.B. die folgenden Dokumentenkategorien betreffen:

- Beigezogene Akten der Staatsanwaltschaft im Übrigen;
- Behandlungsberichte und Überblick über Behandlungsmaßnahmen, verschriebene Medikamente u.Ä.

VII. Exkurs: Eigene Forschungszwecke der ausführenden Stellen

Im Rahmen einer Zusammenarbeit mit Forschungsorganisationen, wie etwa den ausführenden Stellen, kann sich ein Bedürfnis nach der Verfolgung eigener Forschungszwecke durch die Forschungsorganisationen ergeben. Zur Untersuchung dieser Konstellation soll daher nachfolgend hypothetisch angenommen werden, die ausführenden Stellen würden (Grundlagen-)Forschung mit sämtlichen den ausführenden Stellen von den Ministerien übermittelten Daten betreiben (nachfolgend als hypothetisches Szenario bezeichnet). In diesem Fall ergeben sich verschiedene Abweichungen zu der vorangehend vorgenommenen datenschutzrechtlichen Beurteilung. Diese Abweichungen betreffen sowohl die datenschutzrechtliche Verant-

wortlichkeit (unter 1.) und das Erfordernis einer Rechtsgrundlage für die Verarbeitungsvorgänge (unter 2.) als auch weitere datenschutzrechtliche Anforderungen (unter 3.).

In dem hypothetischen Szenario bestehen auch und gerade mit Blick auf die Übermittlung personenbezogener Daten aus den Gerichtsentscheidungen und Aktenauszügen an die ausführenden Stellen Zweifel im Hinblick auf das Vorliegen einer belastbaren Rechtsgrundlage aus Art. 6 Abs. 1 UAbs. 1 DSGVO. Etwas anderes dürfte gelten, wenn und soweit vor der Übermittlung eine Anonymisierung vorgenommen wird und der Anonymisierungsgrad den unter D.II.5.b herausgearbeiteten Anforderungen an eine Urteilsveröffentlichung entspricht, sprich über den Einsatz und die Wirkmöglichkeiten des Erlanger Tools in der derzeitigen Form hinausgeht.

1. (Gemeinsame) Verantwortlichkeit der ausführenden Stellen

Die Festlegung eigener (Forschungs-)Zwecke durch die ausführenden Stellen geht regelmäßig mit weiteren eigenen Festlegungen zu wesentlichen Mitteln der Verarbeitung (z.B. zur Speicherdauer und Empfängern) einher und begründet grundsätzlich nach Art. 4 Nr. 7 DSGVO abweichend vom Sachverhalt des GSJ-Projekts eine Verantwortlichkeit der ausführenden Stellen.[339]

Da die sich ergänzenden[340] Verarbeitungsvorgänge der ausführenden Stellen und der Ministerien gemeinsame Datensätze zum Gegenstand haben,[341] auf einer für betroffene Personen in diesem Umfang nicht konkret vorhersehbaren Verarbeitung beruhen[342] und eine umfängliche Zusammenarbeit zwischen den Ministerien sowie den ausführenden Stellen stattfindet,[343] dürften die Ministerien und die ausführenden Stellen als gemeinsam verantwortlich für die Übermittlung an die ausführenden Stellen und die Weiterverarbeitung für das GSJ-Projekt anzusehen sein. Soweit und sobald die ausführenden Stellen personenbezogene Daten über das Projekt hinaus (nur) für eigene (Grundlagen-)Forschung verarbeiten, geht

339 Ausf. zu den Anforderungen an eine (gemeinsame) Verantwortlichkeit unter D.V.3.a.
340 Vgl. *EuGH*, ZD 2024, 209 (Rn. 43); *EDPB*, Leitlinien 07/2020 zu den Begriffen „Verantwortlicher" und „Auftragsverarbeiter" in der DSGVO, Rn. 55.
341 *T. Radtke*, Gemeinsame Verantwortlichkeit unter der DSGVO, S. 193 ff.
342 Vgl. *EuGH*, ZD 2019, 455 (Rn. 75) – Fashion ID.
343 Vgl. *EuGH*, ZD 2024, 209 (Rn. 43 f.); ZD 2024, 328 (Rn. 59 ff.) – IAB Europe.

die gemeinsame Verantwortlichkeit gegebenenfalls in eine eigenständige Verantwortlichkeit der jeweiligen ausführenden Stellen über.[344]

Vorliegen und Umfang der gemeinsamen Verantwortlichkeit können im GSJ-Projekt beschränkt werden, indem zunächst die ausführenden Stellen streng weisungsgebunden als Auftragsverarbeiter die (Teil-)Anonymisierung durchführen. Hierdurch bestünde dann erst für die anschließende Verarbeitung zu Projekt- und Forschungszwecken mit den zumindest teilweise anonymisierten Daten eine gemeinsame Verantwortlichkeit, wobei die Verarbeitung in Gestalt der Bereitstellung (teil-)anonymisierter Daten an die ausführenden Stellen nach Art. 6 Abs. 1 DSGVO tendenziell eher gerechtfertigt werden kann.

2. Rechtsgrundlage nach Art. 6, 9 DSGVO für die Verarbeitung

Für die Beurteilung der Rechtmäßigkeit nach Art. 6, 9 DSGVO ergeben sich in dem hypothetischen Szenario weitere Abweichungen in der rechtlichen Beurteilung gegenüber dem Sachverhalt des GSJ-Projekts.

Wie im Sachverhalt des GSJ-Projekts kommen wegen Art. 6 Abs. 1 UAbs. 2 DSGVO berechtigte Interessen nach Art. 6 Abs. 1 UAbs. 1 lit. f DSGVO nicht als Rechtsgrundlage in Betracht.[345] Ein Rückgriff auf Art. 6 Abs. 4 DSGVO als (mögliche) eigenständige Rechtsgrundlage ist aufgrund der Neuerhebung durch die ausführenden Stellen als Verantwortliche ebenfalls ausgeschlossen.[346]

Im Fall einer gemeinsamen Verantwortlichkeit müssen sich die gemeinsam Verantwortlichen jeweils auf eine Rechtsgrundlage berufen können.[347] Eine solche Differenzierung nach den jeweiligen Verarbeitungszwecken ist zudem in Art. 89 Abs. 4 DSGVO angelegt. Insoweit sind die Anforderungen an eine mitgliedstaatliche Rechtsvorschrift i.S.d. Art. 6 Abs. 1 UAbs. 1 lit. e DSGVO aus Art. 6 Abs. 3 DSGVO erneut in Erinnerung zu rufen. Danach bedarf es einer Rechtsvorschrift mit festgelegtem Datenverarbeitungsbezug,[348] die „klar und präzise und [deren] Anwendung […] für die Rechtsunterworfenen […] vorhersehbar" ist (ErwGr. 41 S. 2 DSGVO).

344 Vgl. *EuGH*, ZD 2019, 455 (Rn. 70) – Fashion ID.
345 S. oben unter D.V.4.a.
346 Hierzu oben unter D.V.4.e.
347 *EuGH*, ZD 2019, 455 (Rn. 96 f.) – Fashion ID.
348 *Albers/Veit*, in: BeckOK Datenschutzrecht, Art. 6 DSGVO Rn. 57.

Die Anforderungen an die Konkretisierung der Aufgaben im öffentlichen Interesse i.S.d. Art. 6 Abs. 1 UAbs. 1 lit. e DSGVO in der Rechtsvorschrift stehen in einer Wechselwirkung zur unionsautonom auszulegenden[349] Erforderlichkeit einer Verarbeitung für die Wahrnehmung ebendieser Aufgabe. Beispielsweise ist die Verarbeitung personenbezogener Daten zu Zwecken der Forschung und Lehre grundsätzlich zulässig unter Rekurs auf allgemeine Aufgabenzuweisungen der Hochschulen i.V.m. Art. 6 Abs. 1 UAbs. 1 lit. e DSGVO.[350] Je gravierender sich eine Verarbeitung für die betroffenen Personen mit Blick auf die Verarbeitungszwecke, die Sensibilität der verarbeiteten Daten und den Verarbeitungsumfang darstellt, desto präziser muss die mitgliedstaatliche Rechtsvorschrift ausgestaltet sein.[351]

In diesem Zusammenhang ist auch und gerade unter Beachtung der Grundsätze der Datenminimierung (Art. 5 Abs. 1 lit. c DSGVO) und Speicherbegrenzung (Art. 5 Abs. 1 lit. e DSGVO) mit Blick auf Forschungszwecke zu prüfen, ob und in welchem Umfang es eines Personenbezugs bedarf oder ob alternativ mit pseudonymisierten oder anonymisierten Daten der verfolgte Zweck gleichermaßen erreicht werden kann (vgl. auch Art. 89 Abs. 1 DSGVO).[352]

Das aus Art. 89 Abs. 1 DSGVO ableitbare Anonymisierungserfordernis gleicht gerade das an anderer Stelle in der DSGVO vorgesehene Forschungsprivileg aus (z.B. in Art. 5 Abs. 1 lit. b DSGVO, ErwGr. 33 DSGVO).

a. Übermittlung an die ausführenden Stellen

Gegenüber dem Sachverhalt des GSJ-Projekts ergeben sich insbesondere aus dem Vorgang der Übermittlung personenbezogener Daten an die ausführenden Stellen besondere Herausforderungen: Aufgrund der gemeinsamen Verantwortlichkeit bedarf es sowohl einer Rechtsgrundlage der Ministerien für die Offenlegung durch Übermittlung als auch einer Rechtsgrundlage der ausführenden Stellen für die Entgegennahme in Gestalt der Erhe-

349 *Buchner/Petri*, in: Kühling/Buchner, Art. 6 DSGVO Rn. 118.

350 *Golla*, in: Specht-Riemenschneider/Mantz, § 23, Rn. 83.

351 *Schulz*, in: Gola/Heckmann, Art. 6 DSGVO Rn. 57 mit Verweis auf die ständige Rechtsprechung des BVerfG; *Heberlein*, in: Ehmann/Selmayr, Art. 6 DSGVO Rn. 63; *U. Spies*, ZD 2022, 75 (77); vgl. auch *Taeger*, in: Taeger/Gabel, Art. 6 DSGVO Rn. 96; *Roßnagel*, ZD 2020, 296 (298); s. auch zu § 3 BDSG BT-Drs. 18/11325, 81; im Überblick zu großzügigeren Stimmen *Reimer*, in: Sydow/Marsch, Art. 6 DSGVO Rn. 66 m.w.N.

352 *Golla*, in: Specht-Riemenschneider/Mantz, § 23, Rn. 48.

bung dieser Daten durch die ausführenden Stellen. Während die Erhebung durch die ausführenden Stellen im Wesentlichen den gleichen Anforderungen wie die weitere Verarbeitung unterliegt (s. nachfolgend unter b.), ist an dieser Stelle vor allem die Rechtfertigung der Übermittlung aus Sicht der Ministerien gesondert zu prüfen.

aa. Art. 6 Abs. 1 UAbs. 1 lit. e, Abs. 2, 3 DSGVO i.V.m. Art. 5 Abs. 1 S. 1 Nr. 2 BayDSG bzw. § 8 Abs. 2 Nr. 2 DSG NRW

In Betracht kommt zunächst ein Rekurs auf die Bestimmungen der Art. 6 Abs. 1 UAbs. 1 lit. c, e, Abs. 2, 3 DSGVO i.V.m. Art. 5 Abs. 1 S. 1 Nr. 2 BayDSG bzw. § 8 Abs. 2 Nr. 2 DSG NRW, die eine Übermittlung an Dritte aufgrund einer Interessenabwägung zulassen.

Allerdings hat das BVerwG mit Blick auf die nach Art. 6 Abs. 1 UAbs. 2 DSGVO ausgeschlossene Interessenabwägung die Vereinbarkeit des Art. 5 Abs. 1 S. 1 Nr. 2 BayDSG mit den Vorgaben der DSGVO in Abrede gestellt.[353] Diese zutreffende Argumentation des BVerwG dürfte auf den im Kern inhaltsgleichen § 8 Abs. 2 Nr. 2 DSG NRW übertragbar sein.

Vor diesem Hintergrund dürften diese landesrechtlichen Vorschriften nicht als eine belastbare Rechtsgrundlage herangezogen werden können für die Übermittlung an (gemeinsam) Verantwortliche, die eigene Zwecke verfolgen.

bb. Art. 6 Abs. 1 UAbs. 1 lit. e, Abs. 3 DSGVO i.V.m. Art. 4 Abs. 1 BayDSG bzw. § 3 Abs. 1 DSG NRW

Zu denken ist ferner an Art. 6 Abs. 1 UAbs. 1 lit. e, Abs. 3 DSGVO i.V.m. Art. 4 Abs. 1 BayDSG bzw. § 3 Abs. 1 DSG NRW, die als Generalklauseln auf eine dem Verantwortlichen übertragene Aufgabe abstellen. Als eine solche Aufgabe der Justizministerien im öffentlichen Interesse lässt sich auch die Forschung bezogen auf die Verbesserung von Arbeitsabläufen und die Entwicklung von Software für den Einsatz in der Justiz einordnen.

Wenn und soweit allerdings den ausführenden Stellen im Hinblick auf die Zwecke und den Umfang der Verarbeitung ein Gestaltungsspielraum zukommt (z.B. für die grundlegende KI-Forschung ohne spezifischen Er-

353 *BVerwG*, BeckRS 2018, 35172 (Rn. 25 f.); vgl. auch zu § 25 BDSG *Frenzel*, in: Paal/Pauly, § 25 BDSG Rn. 2.

kenntnisgewinn für die Justiz), wird die hierauf gerichtete Übermittlung personenbezogener Daten nicht erforderlich sein für die Erfüllung öffentlicher Aufgaben der *Ministerien*.

Dieser Befund wird in Ansehung der jeweils und kumulativ erforderlichen Rechtsgrundlage(n) der gemeinsam Verantwortlichen unabhängig davon gelten, ob es sich bei der Forschung um eine öffentliche Aufgabe der *ausführenden Stellen* handelt.

Vor diesem Hintergrund erklären sich die Bemühungen und Überlegungen betreffend die Schaffung eines Forschungsdatengesetzes, welches gerade auf einen erleichterten Forschungszugang zu Daten der öffentlichen Hand und entsprechende Rechtsgrundlagen zielt.[354]

Sofern die Ministerien ausschließlich mit dem Erlanger Tool vorbearbeitete und somit teilanonymisierte Datensätze an die ausführenden Stellen übermitteln, sinken im Rahmen der Erforderlichkeit die Anforderungen an den Funktionszusammenhang zwischen Verarbeitung und den Aufgaben im öffentlichen Interesse. Dieser Befund betrifft insbesondere Konstellationen, in denen keine sensiblen personenbezogenen Daten verarbeitet werden (z.B. im Miet- und Verkehrsrecht unter Ausschluss von Fällen, die im Rahmen der Schadensbemessung allfällige Gesundheitsangaben enthalten).

Selbst in diesen Konstellationen kann es aber potenziell zum Vorhandensein von Zehntausenden betroffenen Personen kommen, obwohl aus der Nähe zu Forschungszwecken gerade strenge Anonymisierungsanforderungen aus Art. 89 DSGVO i.V.m. landesrechtlichen Vorschriften[355] erwachsen. Vor diesem Hintergrund bestehen nicht unerhebliche Zweifel daran, ob die Verarbeitung des Restbestands – und gegebenenfalls auch überhaupt die Verarbeitung – personenbezogener Daten zur (Grundlagen-)Forschung als erforderlich anzusehen ist für die Aufgabenerfüllung der (Justiz-)Ministerien.

cc. Art. 6 Abs. 1 UAbs. 1 lit. e, Abs. 3 DSGVO i.V.m. Hochschulgesetzen der Länder

Die Hochschulgesetze der Länder Bayern und Nordrhein-Westfalen weisen den Justizministerien zudem keine Aufgabe im öffentlichen Interesse zu, die

354 *BMBF*, Eckpunkte BMBF Forschungsdatengesetz; hierzu etwa *Specht-Riemenschneider/Wehde*, ZGI 2022, 3 (6); s. auch den aus der Mitte der FDP-Fraktion eingebrachten Gesetzesentwurf, der sich allerdings grundsätzlich auf registerführende Behörden und statistische Daten beschränkt BT-Drs. 20/14262.

355 Hierzu sogleich unter D.VII.2.a.cc.

die Datenübermittlung zur Forschungszwecken i.V.m. Art. 4 Abs. 1 BayDSG bzw. § 3 Abs. 1 DSG NRW erfordern würde. Etwaige Aufgabenzuweisungen (z.B. § 2 Abs. 2, § 3 Abs. 1, 3 HG NRW bzw. Art. 2 Abs. 1 S. 2, Abs. 2 S. 2, 3, Art. 3 Abs. 1 S. 1, Art. 6 Abs. 1 S. 1 BayHIG) adressieren grundsätzlich nur die jeweiligen Hochschulen.

Soweit bspw. Art. 6 Abs. 1 S. 1 BayHIG darauf abstellt, dass die Hochschulen „mit dem Bund, den Ländern und anderen juristischen Personen des öffentlichen Rechts" zur Erfüllung ihrer öffentlichen Aufgaben zusammenwirken, dürfte sich aus dieser Vorschrift zwar ein allgemeines Bekenntnis zu einer Zusammenarbeit entsprechend den vorstehenden Erwägungen zu Art. 4 Abs. 1 BayDSG bzw. § 3 Abs. 1 DSG NRW, nicht aber eine hinreichend konkrete Aufgabenzuweisung an das bayerische Justizministerium ergeben, nicht-anonymisierte oder nur teilweise anonymisierte Daten zur freien Forschung in dem untersuchungsgegenständlichen Umfang zur Verfügung zu stellen.

dd. Art. 6 Abs. 1 UAbs. 1 lit. e, Abs. 3 DSGVO i.V.m. § 17 Abs. 1 DSG NRW bzw. Art. 4 Abs. 1 BayDSG i.V.m. BayHIG oder Erst-Recht-Schluss aus Art. 8 Abs. 1 S. 1 Nr. 5 BayDSG i.V.m. Art. 6 Abs. 2 Nr. 3 lit. c BayDSG

Der Rekurs auf Art. 6 Abs. 1 UAbs. 1 lit. e, Abs. 3 DSGVO i.V.m. § 17 Abs. 1 DSG NRW ermöglicht ebenfalls keine belastbar rechtssichere Übermittlung personenbezogener Daten zu wissenschaftlichen Forschungszwecken in dem hypothetischen Szenario.

Zunächst stellt § 17 Abs. 1 DSG NRW keine Anforderungen an die öffentliche Stelle, sodass die Rechtsgrundlage im Hinblick auf eine Verarbeitung für Forschungszwecke im Ausgangspunkt auch den verantwortlichen Ministerien offensteht. Die Rechtsgrundlage kann im Hinblick auf den dort festgelegten Forschungszweck somit grundsätzlich den Anforderungen der DSGVO aus Art. 6 Abs. 3 DSGVO genügen.[356]

§ 17 DSG NRW statuiert allerdings weitere Anforderungen, darunter eine Abwägung mit den schutzwürdigen Belangen der betroffenen Personen (§ 17 Abs. 1 DSG NRW), die Einhaltung der allgemeinen Garantien aus § 15 DSG NRW (i.e. technische und organisatorische Maßnahmen einschließlich der nachträglichen Überprüfbarkeit der Verarbeitung und der Sensibilisierung der an Verarbeitungsvorgängen Beteiligten) sowie im Einklang

356 Vgl. zu dem Verhältnis der Alternativen in Art. 6 Abs. 3 S. 2 DSGVO *Heberlein*, in: Ehmann/Selmayr, Art. 6 DSGVO Rn. 63.

mit Art. 89 Abs. 1 DSGVO in § 17 Abs. 3 S. 1 DSG NRW (ähnlich: Art. 25 Abs. 2 S. 1 BayDSG) eine Anonymisierungspflicht, „sobald dies nach dem Forschungs- oder Statistikzweck möglich ist".

Entsprechend den zu den anderen Rechtsvorschriften herausgearbeiteten Maßgaben stehen die schutzwürdigen Belange betroffener Personen insbesondere einer Verarbeitung personenbezogener Daten entgegen, wenn die personenbezogenen Daten von vornherein nicht für die Forschungszwecke benötigt werden, zugleich jedoch massenhaft übermittelt werden. Zu dem gleichen Ergebnis führt die in § 17 Abs. 3 S. 1 DSG NRW vorgesehene Anonymisierungspflicht, die nicht auf die technische Möglichkeit einer automatisierten Anonymisierung abstellt[357] (z.B. durch das Erlanger Tool), sondern vielmehr nur darauf, ob durch eine Anonymisierung der Forschungszweck beeinträchtigt wird. Eine solche Beeinträchtigung des Forschungszwecks dürfte nach dem untersuchungsgegenständlichen Sachverhalt vorliegend nicht anzunehmen sein.

Im Ergebnis erwachsen in diesem hypothetischen Szenario vor allem aus den an Forschungszwecke gestellten strengeren Maßnahmenanforderungen in Verbindung mit der Übermittlung durch die Ministerien erhebliche datenschutzrechtliche Bedenken mit Blick auf die beabsichtigte Zusammenarbeit im Rahmen des GSJ-Projekts.

Die Bewertung nach bayerischem Landesrecht ist grundsätzlich gleichlaufend (s. auch Art. 25 Abs. 2 BayDSG zum Anonymisierungserfordernis). Für die Verarbeitung personenbezogener Daten in Gerichtsentscheidungen der bayerischen Justiz fehlt es aber an einer Entsprechung zu § 17 Abs. 1 DSG NRW, der über Art. 9 Abs. 2 DSGVO hinaus funktional zugleich als Rechtsvorschrift i.S.d. Art. 6 Abs. 1 UAbs. 1 lit. e DSGVO wirkt. Denn die bayerischen Regelungen privilegieren Forschungszwecke nur mit Blick auf die Zweckbindung (Art. 6 Abs. 2 Nr. 3 lit. c BayDSG) und die Verarbeitung besonderer Kategorien personenbezogener Daten i.S.d. Art. 9 DSGVO (Art. 8 Abs. 1 S. 1 Nr. 5 i.V.m. Art. 6 Abs. 2 Nr. 3 lit. c BayDSG). Für nichtbesondere Kategorien personenbezogener Daten kommt daher nur ein Rückgriff auf Art. 6 Abs. 1 UAbs. 1 lit. e DSGVO, Art. 4 Abs. 1 BayDSG i.V.m. BayHIG in Betracht oder eine mit erheblichen Unsicherheiten behaftete

357 Vgl. zu § 27 BDSG *Koch*, in: BeckOK Datenschutzrecht, § 27 BDSG Rn. 39 ff.; *Hense*, in: Sydow/Marsch, § 27 BDSG Rn. 20: „Ob und inwieweit Anonymisierung unter den Vorzeichen fortschreitender Datenverarbeitungstechniken und damit verbundener Verknüpfungsmöglichkeiten mittels Algorithmen überhaupt eine Option bleibt, sei dahingestellt."; wohl auch *Petri*, in: Simitis/Hornung/Spiecker gen. Döhmann, § 27 BDSG Rn. 14 f.; mit einem etwas weiteren Verständnis *Buchner/Tinnefeld*, in: Kühling/Buchner, § 27 BDSG Rn. 24.

dogmatische Herleitung in Form eines Erst-Recht-Schlusses der Zulässigkeit aus Art. 8 Abs. 1 S. 1 Nr. 5 i.V.m. Art. 6 Abs. 2 Nr. 3 lit. c BayDSG auch für nicht-besondere Kategorien personenbezogener Daten.

ee. Besonderheiten mit Blick auf besondere Kategorien personenbezogener Daten nach Art. 9 DSGVO

Für die Übermittlung besonderer Kategorien personenbezogener Daten in den Gerichtsentscheidungen kommen als Rechtsgrundlage zwar grundsätzlich Art. 9 Abs. 2 lit. j DSGVO i.V.m. § 17 Abs. 1 DSG NRW bzw. Art. 8 Abs. 1 S. 1 Nr. 5 i.V.m. Art. 6 Abs. 2 Nr. 3 lit. c BayDSG in Betracht.

Nach den vorherigen Ausführungen bestehen insoweit aber erst recht erhebliche Bedenken, eine umfangreiche Verarbeitung sensibler, nicht vollständig anonymisierter Daten auf eine Forschungsgeneralklausel zu stützen. Das in Art. 89 Abs. 1 DSGVO angelegte und in den landesrechtlichen Vorschriften ausgeprägte Anonymisierungserfordernis wird sich in Ansehung der der sensiblen Daten in besonderem Maße auswirken.

b. Weiterverarbeitung durch die ausführenden Stellen

Soweit eine (gerechtfertigte Übermittlung) erfolgt ist, sind die Anforderungen an die Weiterverarbeitung grundsätzlich gleichlaufend mit den Anforderungen aus dem Sachverhalt des GSJ-Projekts (s. oben unter D.V), da die zur Realisierung des GSJ-Projekts notwendigen Verarbeitungsaktivitäten weiterhin zu den auch im Sachverhalt des GSJ-Projekts maßgeblichen Zwecken erfolgen.

Die ergänzende Verarbeitung zu Forschungszwecken über das GSJ-Projekt hinaus kann auf die landesrechtlichen Forschungsgeneralklauseln bzw. Aufgabenzuweisungen in den Hochschulgesetzen i.V.m. Art. 6 Abs. 1 UAbs. 1 lit. e, Art. 9 Abs. 2 lit. j DSGVO gestützt werden (z.B. Art. 6 Abs. 1 UAbs. 1 lit. e DSGVO, Art. 4 Abs. 1 BayDSG i.V.m. Art. 2 Abs. 1 S. 2, Abs. 2 S. 2 und 3, Art. 3 Abs. 1 S. 1, Art. 6 Abs. 1 S. 1 BayHIG bzw. §§ 3, 17 DSG NRW i.V.m. § 2 Abs. 2, § 3 Abs. 1, 3 HG NRW), wenn bereits die Übermittlung dieser Daten zu Forschungszwecken zulässig war und ein ausreichender Anonymisierungsgrad erreicht wurde. Denn die ausführenden Stellen üben in diesem Fall ihre Kerntätigkeit in der Forschung aus; somit steht gerade nicht (mehr) die Frage im Raum, ob die Verfolgung von Forschungszwecken eine Aufgabe der (*Justiz-)Ministerien* ist.

3. Datenschutzrechtliche Anforderungen im Übrigen

Im Fall einer gemeinsamen Verantwortlichkeit zwischen den Ministerien und den ausführenden Stellen sind die weitergehenden Pflichten zu einer Vereinbarung (Art. 26 Abs. 1 S. 2, 3, Abs. 2 S. 1 DSGVO) sowie zur Bereitstellung des Wesentlichen der Vereinbarung (Art. 26 Abs. 2 S. 2 DSGVO) zu beachten.

Das gesteigerte Risiko[358] für betroffene Personen ist im Rahmen der Datenschutzfolgenabschätzung nach Art. 35 DSGVO zu reflektieren. Diese Verpflichtung gilt umso mehr in Ansehung der Forschungsgeneralklauseln, denen gerade keine konkrete Folgenabschätzung zugrunde liegt, die den vorliegenden Fall mit seinen Besonderheiten ausreichend abdeckt.[359]

Für die zu Forschungszwecken erfolgenden Verarbeitungen sind die in Ausprägung des Art. 89 DSGVO ergänzenden Anforderungen aus dem Landesrecht zu beachten. So statuieren beispielsweise § 17 Abs. 4 DSG NRW bzw. Art. 25 Abs. 3 BayDSG zusätzliche Anforderungen an die Veröffentlichung von Forschungsdaten, wobei sich diese zusätzlichen Anforderungen gegebenenfalls auf die Veröffentlichung des Sprachmodells auswirken können.

Zudem schränkt § 17 Abs. 5 DSG NRW bzw. Art. 25 Abs. 4 BayDSG (vgl. Art. 89 Abs. 2 DSGVO) zugunsten der ausführenden Stellen Betroffenenrechte mit Blick auf die Verarbeitung zu Forschungszwecken ein.

4. Zwischenergebnis

In dem hier zugrunde gelegten hypothetischen Szenario erwachsen besondere Herausforderungen vor allem aus dem Erfordernis einer Rechtsgrundlage für die Übermittlung nicht-anonymisierter Gerichtsentscheidungen durch die Ministerien an die ausführenden Stellen. An der Möglichkeit zu einem rechtlich belastbaren Rückgriff auf die Aufgabenzuweisungen und Forschungsgeneralklauseln in den landesrechtlichen Vorschriften bestehen insoweit erhebliche Zweifel.

Dieser Befund gilt umso mehr, wenn und soweit die Gerichtsentscheidungen ohne Vorbearbeitung durch ein Anonymisierungstool und unter Einschluss im Einzelfall sensibler Gerichtsentscheidungen übermittelt werden (z.B. Gesundheitsdaten im Rahmen von Schmerzensgeld im Verkehrs-

358 *T. Radtke*, Gemeinsame Verantwortlichkeit unter der DSGVO, S. 411 ff.
359 Vgl. *Reimer*, in: Sydow/Marsch, Art. 6 DSGVO Rn. 64.

recht oder sogar über das Miet- und Verkehrsrecht hinaus im Falle von Arzthaftungsentscheidungen).

Selbst wenn eine (Teil-)Anonymisierung mit dem Erlanger Tool erfolgt, dürften die strengeren Anonymisierungsanforderungen an eine Verarbeitung zu Forschungszwecken aus Art. 89 DSGVO mit dem derzeit durch das Erlanger Tool erreichten Anonymisierungsstandard kaum eingehalten werden können.

VIII. Ableitung von datenschutzrechtlichen Handlungsempfehlungen

Aus der vorangegangenen Untersuchung lassen sich konkrete Handlungsempfehlungen für das untersuchungsgegenständliche GSJ-Projekt ableiten, die an dieser Stelle zusammengeführt werden.

1. Datenschutzrechtliche Ausgangslage in drei Stufen: Anonymisierung – Rechtfertigung nach Art. 6 DSGVO – Rechtfertigung nach Art. 9, 10 DSGVO

Mit Blick auf die Anwendbarkeit der DSGVO und die Einhaltung grundlegender datenschutzrechtlicher Anforderungen lässt sich zwischen drei, nachfolgend näher zu behandelnden Stufen unterscheiden. Demgegenüber sind andere Rechtsakte, insbesondere die JI-RL, mit Blick auf den gewählten Zuschnitt des Projekts (i.e. Ausschluss von Strafsachen) von der Vertiefung auszunehmen.

Auf erster Stufe steht die Frage des Vorliegens personenbezogener Daten i.S.d. Art. 4 Nr. 1 DSGVO und mithin einer (erfolgreichen) Anonymisierung. Ein Personenbezug setzt entweder eine eindeutige Kennung voraus oder sonstige Informationen, die nach allgemeinem Ermessen wahrscheinlich zur Identifizierung einer natürlichen Person genutzt werden. Für diese Wahrscheinlichkeit kommt es bei KI-Systemen maßgeblich auf die Ausgabe an, da auf andere Weise den justizinternen Anwendern im vorliegenden Fall kein Zugriff auf die Daten, mithin auch keine Identifizierung möglich ist. Eine (Teil-)Anonymisierung durch das Erlanger Tool kann vor diesem Hintergrund in vielen Fällen zu einer nahezu vollständigen Anonymisierung führen, indem die Informationen zunächst lediglich die Grundlage für ein wahrscheinlichkeitsbasiertes Training des KI-Systems sind und erst im Rahmen der Ausgabe eine risikobehaftete Verarbeitung erfolgen kann.

Soweit keine hinreichenden Maßnahmen zum Ausschluss des Personenbezugs ergriffen wurden, können auf zweiter Stufe ein verbleibender Personenbezug im Modell im Rahmen einer Bearbeitung (z.B. betreffend die Verarbeitung von Kennungen) sowie die vereinzelte Ausgabe von Kennungen gegebenenfalls auf Art. 6 Abs. 1 UAbs. 1 lit. e DSGVO i.V.m. landesrechtlichen Generalklauseln Art. 4 Abs. 1 BayDSG, § 3 Abs. 1 DSG NRW oder nach einer weiteren in Betracht kommenden Auffassung auf Art. 6 Abs. 1 UAbs. 1 lit. f DSGVO oder auf Art. 6 Abs. 4 DSGVO gestützt werden, da insoweit ein enger Zusammenhang mit einer öffentlichen Aufgabe besteht und eine risikoarme justizinterne Verarbeitung erfolgt.

Auf dritter Stufe ist die Verarbeitung besonderer Kategorien personenbezogener Daten i.S.d. Art. 9, 10 DSGVO angesprochen, die nur im Einzelfall auf Rechtfertigungstatbestände, wie etwa Art. 9 Abs. 2 lit. e, j DSGVO, gestützt werden oder die Anforderungen des Art. 10 S. 1 DSGVO erfüllen kann. Daher ist im Hinblick auf das Vorliegen dieser Daten besonders sorgfältig zu clustern und zu filtern.

Bei Beachtung der jeweils auf dieser Stufe zu stellenden Anforderungen ist das untersuchungsgegenständliche GSJ-Projekt grundsätzlich datenschutzrechtskonform realisierbar.

Auf eine echte Privilegierungswirkung bereits veröffentlichter Gerichtsentscheidungen können sich Verantwortliche hierbei allerdings nicht berufen. Gleichwohl wird die Weiterverarbeitung einer zulässigerweise, mit entsprechend hohem Anonymisierungsgrad veröffentlichten Gerichtsentscheidung aufgrund eines engeren Funktionszusammenhangs mit öffentlichen Aufgaben und gegebenenfalls nicht signifikant erhöhten Risiken für betroffene Personen grundsätzlich auf Art. 6 Abs. 1, 4 DSGVO, Art. 9 Abs. 2 lit. f DSGVO gestützt werden können.

2. Anforderungen an die Zusammenstellung des Trainingskorpus aus Gerichtsentscheidungen und Aktenauszügen (insbesondere Clusterung)

Soweit personenbezogene Daten vorliegen, ergibt sich zunächst vor allem mit Blick auf Art. 9, 10 DSGVO ein Bedarf zur Clusterung. Im Rahmen einer solchen Clusterung nach Verfahrensarten können insbesondere Miet- und Verkehrssachen grundsätzlich einzuschließen, Arzthaftungssachen, Familien- und Strafsachen demgegenüber grundsätzlich auszuschließen sein.

Aus den Aktenauszüge sind unzureichend bzw. mangelhaft digitalisierte Dokumentbestandteile zu filtern und gegebenenfalls händisch nachzubearbeiten.

Außerdem sind für die Aktenauszüge entsprechende Dokumentenkategorien festzulegen mit:

- Hohem Risiko (regelmäßig vollständiger Ausschluss oder Nachbearbeitung), z.B.:
  - o Beigezogene Akten der Staatsanwaltschaft im Zusammenhang mit Verfahren, die typischerweise über die Straftat i.S.d. Art. 10 DSGVO hinaus sensible Informationen enthalten (z.B. Straftaten gegen die sexuelle Selbstbestimmung und schwere Körperverletzungsdelikte);
  - o Psychologische Gutachten;
  - o Ärztliche Gutachten.
- Mittlerem Risiko (angepasstes anonymisiertes Anonymisierungsverfahren), z.B.:
  - o Beigezogene Akten der Staatsanwaltschaft im Übrigen;
  - o Behandlungsberichte und Überblick über Behandlungsmaßnahmen, verschriebene Medikamente u.Ä.

Darüber hinaus ist nach Merkmalen innerhalb der Gerichtsentscheidungen und Aktenauszüge zu filtern. Das betrifft insbesondere:

- Geltendmachung von deliktsrechtlichen Schadensersatzansprüchen (insbesondere § 823 Abs. 2 BGB i.V.m. StGB);
- Geltendmachung von Schmerzensgeld nach § 253 Abs. 1, 2 BGB;
- Geltendmachung von Unterlassungsansprüchen mit Blick auf (politische) Aussagen, wobei gegebenenfalls eine offensichtliche Öffentlichmachung nach Art. 9 Abs. 2 lit. e DSGVO in Betracht kommen kann;
- Vorkommen von Begriffen wie „Gewerkschaft" sowie geläufigen Gewerkschafts- und Parteibezeichnungen innerhalb der Aktenauszüge.

Soweit einzelne Textabschnitte sich immer wieder in ähnlicher Form in verschiedenen Gerichtsentscheidungen wiederfinden, ist wegen der verstärkten Repräsentation in den Trainings- und womöglich auch den Ausgabedaten in besonderem Maße darauf zu achten, dass Kennungen und sonstige identifizierende Merkmale identifiziert wurden. Das betrifft auch und gerade Konstellationen, in denen Textbausteine aus einer Gerichtsentscheidung in ähnlicher Form in andere Entscheidungen übernommen wurden.

3. Anforderungen an das Anonymisierungstool (insbesondere Recall-Wert)

Das Anonymisierungstool muss insbesondere für Kennungen einen hohen Recall-Wert erreichen. Diese Anforderung gilt gleichermaßen für Gerichtsentscheidungen und Aktenauszüge. Zwar ist selbst bei einem Recall-Wert von unter 90 % aufgrund des zwischengelagerten Trainings, der eng eingegrenzten Eingabeaufforderung im Rahmen der Use-Cases 1-3 sowie der Ausgabe eine datenschutzkonforme Umsetzung mit hohem Anonymisierungsanteil auf erster Stufe und gerechtfertigten systeminternen Verarbeitungen auf zweiter Stufe denkbar. Zugleich steigt aber mit einem niedrig(er)en Recall-Wert für Kennungen jedenfalls die Wahrscheinlichkeit von Datenschutzverstößen, z.B. mit Blick auf vermehrte und gegebenenfalls nicht mehr zu rechtfertigende Verarbeitungen in der Ausgabe oder – auf dritter Stufe – der Kombination von Kennungen mit besonderen Kategorien personenbezogener Daten.

Für sonstige identifizierende Merkmale kann unter Umständen auch noch ein Recall-Wert von deutlich unter 50 % genügen. Denn insoweit ergibt sich regelmäßig schon auf erster Stufe eine Anonymisierung dadurch, dass die Heranziehung weiterer Quellen zur Identifizierung natürlicher Personen durch justizinterne Nutzer im Fall einer unbeabsichtigten Ausgabe aus den Trainingsdaten unwahrscheinlich ist. Dieser Befund gilt vorbehaltlich der übrigen Maßnahmen, insbesondere der Clusterung mit Blick auf besondere Kategorien personenbezogener Daten.

Das in Rede stehende Anonymisierungstool, übrige Software und Prozesse sind entsprechend dem Stand der Technik auch im Hinblick auf IT-Sicherheitsanforderungen fortlaufend zu evaluieren und gegebenenfalls fortzuentwickeln.

4. Anforderungen an das KI-System und dessen Einsatz (insbesondere System Prompt und Ausgabefilter)

Das KI-System ist nur für den justizinternen Einsatz und auf ausgewählte Use Cases ausgelegt. Diese Beschränkung sollte zugleich durch technische und organisatorische Maßnahmen sichergestellt werden, z.B. durch einen möglichst sicheren Login-Mechanismus nach dem Stand der Technik sowie, nicht notwendigerweise, durch technische Einschränkungen der Eingabeaufforderung.

In den genannten Use-Cases kann es sich beispielsweise zur weiteren, nicht aber zwingend erforderlichen Risikoreduktion anbieten, dem Nutzer

die Eingabeaufforderung verbindlich vorzugeben (z.B. die Auswahl jeweils eines Prompts für die Use-Cases 1-3) und nur die Auswahl des Eingabekontexts zu ermöglichen (z.B. die Auswahl der jeweiligen E-Akte oder eines PDF-Dokuments). Eine etwaige Ausgabe personenbezogener Trainingsdaten sollte zudem auf einfachem Wege durch die Nutzer gemeldet werden können.

Ferner ist das KI-System über den initialen sog. System Prompt in geeigneter Weise anzuhalten, vor allem mit den neuen Eingaben zu arbeiten sowie die Wiedergabe von Kennungen oder sonstigen personenbezogenen Daten aus den Trainingsdaten zu vermeiden.

Besonderes Gewicht kommt vor allem effektiven Ausgabefiltern zu, die schon auf erster Stufe eine weitgehende Anonymisierung herbeiführen können. Die Ausgaben sollten auf typische Kennungsformate (z.B. Namen, E-Mail-Adressen, Telefonnummern und Identifikationsnummern) durch sog. reguläre Ausdrücke durchsucht und diese mit dem Vorkommen im Eingabekontext abgeglichen werden. Nur wenn sich diese Angaben in ähnlicher Form auch im Eingabekontext wiederfinden (z.B. eine dort verwendete Telefonnummer oder ein dort verwendeter Nachname), sprich dem konkreten Verfahren und nicht den Trainingsdaten zuzuordnen sind, sollte die Ausgabe zugelassen werden.

Eine weitergehende, aber ebenfalls nicht zwingend erforderliche Risikoreduktion kann erfolgen, indem über einfache reguläre Ausdrücke hinaus die Ausgabe anhand des Erlanger Tools klassifiziert wird und auf dieser Basis mit dem Eingabekontext abgeglichen wird.

Die KI-Ausgabe kann zudem durch ein digitales Wasserzeichen gekennzeichnet werden. Außerdem empfiehlt sich generell eine niedrige Temperatur-Einstellung, um das Auftreten datenschutzrelevanter Halluzinationen zu reduzieren. Eingaben dürfen nicht ohne weitere Bearbeitung zur Verbesserung und zum weiteren Training des bestehenden Sprachmodells verwendet werden.

5. Anforderungen an die Veröffentlichung des KI-Modells oder KI-Systems

Die öffentliche Zurverfügungstellung des KI-Systems, sprich beispielsweise eines öffentlichen abrufbaren Justiz-Chatbots, berührt vor allem die erste Stufe, i.e. die Anonymisierung. Durch den unbeschränkten, auch justizexternen Adressatenkreis steigt die Wahrscheinlichkeit einer Identifizierung anhand von KI-Ausgaben, z.B. durch Extrahierungsangriffe. Dieser Effekt

kann durch stark beschränkte Eingabemöglichkeiten (z.B. nur zum eigentlichen Eingabekontext) und effektive Ausgabenfilter eingedämmt werden, sodass in Abhängigkeit von der konkreten Umsetzung die öffentliche Zurverfügungstellung des KI-Systems datenschutzrechtlich vertretbar ausgestaltet werden können dürfte.

Bei der Veröffentlichung des trainierten Sprachmodells (d.h. insbesondere der Gewichte), gegebenenfalls einschließlich der Trainingsdaten, werden sämtliche dieser Sicherungsmechanismen jedoch grundsätzlich ausscheiden. Insoweit bestehen datenschutzrechtliche Bedenken gegen die Veröffentlichung, die zumindest auf dritter Stufe mit Blick auf Art. 9, 10 DSGVO durch eine effektive Clusterung zu einem hohen Grad ausgeräumt werden können.

Soweit allerdings ein Anonymisierungsgrad erreicht wird, der dem Standard für eine rechtskonforme Urteilsveröffentlichung i.R.d. Art. 6 Abs. 1 UAbs. 1 lit. c, e, f, Abs. 4 DSGVO entspricht (d.h. Recall-Werte von Kennungen i.H.v. annähernd 99 %), kann auch die Veröffentlichung des hierauf aufbauenden Sprachmodells unter Berücksichtigung der hierdurch nicht stärker tangierten Interessen betroffener Personen auf die Rechtsgrundlage aus Art. 6 Abs. 1 UAbs. 1 lit. c, e, f, Abs. 4 DSGVO gestützt werden. Ein solches Vorgehen ist allerdings nur für ein Sprachmodell eröffnet, das lediglich auf Grundlage der Gerichtsentscheidungen, nicht aber auch auf den Aktenauszügen trainiert ist.

6. Beachtung weiterer datenschutzrechtlicher Pflichten

Darüber hinaus sind weitere datenschutzrechtliche Pflichten zu beachten. Diese Pflichten betreffen vielgestaltige technische und organisatorische Maßnahmen, so etwa

- nach der hiesigen Auffassung eine Auftragsverarbeitungsvereinbarung zwischen den Ministerien und den ausführenden Stellen (Art. 28 DSGVO) sowie eine Vereinbarung über die gemeinsame Verantwortlichkeit zwischen den Ministerien (Art. 26 DSGVO);
- die Durchführung einer Datenschutzfolgenabschätzung (Art. 35 DSGVO);
- die Dokumentation der Zusammenstellung der Trainingsdaten bei gleichzeitiger Festlegung und Beachtung von Löschfristen;

- die fortlaufende Überprüfung des Standards des Anonymisierungstools sowie weiterer technischer und organisatorischer Maßnahmen anhand des Stands der Technik (Art. 24, 25, 32 DSGVO).

Im Hinblick auf die Betroffenenrechte ergeben sich weniger aus dem konkreten Projekt als vielmehr aus dem Einsatz eines KI-Systems verschiedene Herausforderungen.

Für die Informationspflicht kann die Ausnahme des Art. 14 Abs. 5 lit. b Hs. 1 Var. 2 DSGVO zur Anwendung gelangen, erfordert allerdings weitere Maßnahmen zum Schutz der betroffenen Personen (z.B. die öffentliche Information über das Projekt und die damit einhergehende Verarbeitung personenbezogener Daten). Auskunftsansprüche nach Art. 15 DSGVO dürften im GSJ-Projekt aufgrund der justizinternen Verarbeitung praktisch kaum eine Bedeutung erlangen und werden sich zudem auch regelmäßig nicht auf das GSJ-Sprachmodell, sondern nur auf eine etwaige Ausgabe richten, denn das Sprachmodell beinhaltet regelmäßig keine personenbezogenen Daten.

Die Richtigkeit der Trainingsdaten, z.B. im Zusammenhang mit einem Berichtigungsbegehren nach Art. 16 DSGVO, wird in Ansehung der einzunehmenden Perspektive im Projekt selten in Frage stehen. Denn es wird sich regelmäßig um richtige Daten als Niederlegung des gerichtlichen Verfahrensablaufs und der rechtlichen Würdigung des Gerichts handeln. Das Risiko des Auftretens etwaiger Halluzinationen kann durch technische und organisatorische Maßnahmen (z.B. die Temperatur-Einstellung des KI-Systems) reduziert werden.

Eine Löschung, z.B. im Zusammenhang mit einem entsprechenden Begehren nach Art. 17, 21 DSGVO, dürfte sich vor allem auf die Ergebnisse der Use-Cases als Sachverhaltsabschnitt einer neuen Gerichtsentscheidung richten und ist insoweit einer Einzelfallprüfung zu unterwerfen.

E. Urheberrechtliche Anforderungen an die Verwendung von Aktenauszügen zu Trainingszwecken

Die Verwendung von Aktenauszügen zum Training eines Sprachmodells wird im Folgenden auf der Grundlage und am Maßstab des UrhG zu prüfen sein.

I. Urheberrechtlicher Schutz und Zuordnung

Zunächst ist die Frage aufzuwerfen, ob und inwieweit Aktenauszüge urheberrechtlich geschützt sind. Als Schöpfer (§ 7 UrhG) kommt stets nur eine natürliche Person in Betracht, wobei Nutzungsrechte auch juristischen Personen (wie etwa dem Arbeitgeber) eingeräumt werden können.

Schwerpunktmäßig sind Schriftsätze in den Blick zu nehmen (unter 1.), wobei etwaige Besonderheiten für Aktenauszüge im Übrigen anschließend adressiert werden (unter 2.).

1. Schriftsätze

Mit Blick auf anwaltliche Schriftsätze sind neben dem Schutz als persönliche geistige Schöpfung weitere (Schutz-)Rechte und Einschränkungen aus dem UrhG zu erörtern.

a. Schutz als Werk nach § 2 UrhG

Als Werke[360] geschützt sind nach der nicht-abschließenden[361] Aufzählung des § 2 Abs. 1 UrhG insbesondere „Sprachwerke, wie Schriftwerke"

360 Noch nicht vollständig unionsrechtlich harmonisiert nach *Schmid*, KIR 2025, 69 (76); weitgehende Harmonisierung durch die Infopaq-Entscheidung annehmend *Wandtke*, MMR 2017, 367 (369); s. auch *EuGH*, GRUR 2019, 73 (Rn. 35 ff.) – Levola Hengelo BV/Smilde Foods BV.

361 Statt aller *Raue*, in: Dreier/Schulze, § 2 UrhG Rn. 101 f.

(Nr. 1),[362] sofern es sich um eine persönliche geistige Schöpfung handelt (§ 2 Abs. 2 UrhG). Über das Merkmal der persönlichen geistigen Schöpfung schützt das Urheberrecht das Vorkommen von einzelnen Zeichen (Syntax) als eine konkrete Form, sofern diese Syntax nach ihrer Bedeutung (Semantik) den (kreativen) geistigen Gehalt einer natürlichen Person (i.e. des Urhebers) zum Ausdruck bringt (so insbesondere „Form und Art der Sammlung, Einteilung und Anordnung des dargebotenen Stoffes"[363] sowie „schöpferische Gedankenformung und -führung des dargebotenen Inhalts"[364]).[365] Ein solcher geistiger Gehalt ist einem (Sprach-)Werk unter Berücksichtigung eines Freihaltebedürfnisses[366] nicht zu entnehmen, wenn und soweit die Form des Werks „aus der Natur der Sache oder nach den Gesetzen der Zweckmäßigkeit vorgegeben ist".[367] Zudem folgt auch aus einem „innovativen Charakter" des Inhalts eines Schriftsatzes noch keine persönliche geistige Schöpfung.[368]

Grundsätzlich können anwaltliche Schriftsätze als Sprachwerk urheberrechtlichen Schutz genießen.[369] Der BGH stellte zwar zunächst aufgrund des begrenzten Spielraums für schöpferische Kreativität an wissenschaftliche Sprachwerke und insbesondere auch Anwaltsschriftsätze erhöhte Schutzanforderungen[370] über die sonst angenommene sog. kleine Münze[371] hinaus (so z.B. über die chronologische Schilderung eines Sachverhalts, auf der Hand liegende rechtliche Anforderungen und Anträge hinaus).

An den Maßgaben aus dieser Rechtsprechung dürfte unter Berücksichtigung der unionsrechtlichen Vorgaben[372] aber nicht mehr festgehalten wer-

362 Ein Datenbankwerk i.S.d. § 4 Abs. 2 UrhG wird mangels systematischer und methodischer Anordnung von Literaturquellen in vereinzelten Fußnoten in den Anwaltsschriftsätzen nicht in Betracht kommen.
363 *OLG Düsseldorf*, MDR 2023, 1401 (1401); *BGH*, MMR 2011, 182 (Rn. 36).
364 *BGH*, GRUR 1986, 739 (740).
365 *Zech*, Information als Schutzgegenstand, S. 37 f., 246 ff., 355 ff.; *Hofmann*, WRP 2024, 11 (14); *Hofmann*, ZUM 2024, 166 (167).
366 *Bullinger*, in: Wandtke/Bullinger, § 2 UrhG Rn. 25.
367 *BVerwG*, GRUR 2020, 189 (Rn. 19); *BGH*, GRUR 2002, 958 (960).
368 *BVerwG*, GRUR 2020, 189 (Rn. 19).
369 Schon *BGH*, GRUR 1986, 739 (740); *Rauer/Bibi*, in: Götting/Lauber-Rönsberg/Rauer, § 2 UrhG Rn. 195; *OVG Hamburg*, GRUR-RS 2021, 34061; *OLG Düsseldorf*, MDR 2023, 1401.
370 *BGH*, GRUR 1986, 739.
371 Statt aller *BGH*, GRUR 2014, 175 (Rn. 18).
372 S. etwa *EuGH*, GRUR 2019, 73 – Levola Hengelo BV/Smilde Foods BV.

den können.[373] Es sind daher für Sprachwerke, wie es der BGH schon für Werke der angewandten Kunst entschieden hat,[374] grundsätzlich einheitliche Schutzanforderungen zu stellen, wobei gegebenenfalls in Abhängigkeit von dem jeweiligen Werktyp und fehlendem Raum für Kreativität ein geringerer Schutzumfang anzunehmen sein kann.[375]

Unter Anwendung dieses Maßstabs verneinte jüngst das OLG Düsseldorf den Schutz für eine Klageschrift, die lediglich „eine chronologische Darstellung des prozessrelevanten Sachverhalts mit einer Bezugnahme auf beigefügte Anlagen [enthält] und […] im Folgenden kurz die angekündigten Klageanträge auf Zahlung eines angemessenen Schmerzensgeldes, auf Begleichung außergerichtlicher Rechtsanwaltskosten gegenüber der Rechtsschutzversicherung sowie auf Feststellung einer weitergehenden Schadensersatzpflicht [erläutert]".[376] Für die weiteren verfahrensgegenständlichen Schriftsätze wurde ein Schutz ebenfalls abgelehnt, da diese Schriftsätze „soweit sie inhaltlich zum Prozessstoff Stellung nehmen, auf das Vorbringen der Gegenseite" erwidern.[377] Hervorzuheben ist, dass im konkreten Fall der verfassende Rechtsanwalt allerdings nicht substantiiert zu den Schriftsätzen als Ausdruck einer schöpferischen Tätigkeit vorgetragen hatte.

Das OVG Hamburg hingegen nahm im Zuge einer recht großzügigen Auslegung[378] des § 2 Abs. 2 UrhG einen Schutz für einen Schriftsatz an und stützte sich hierzu unter anderem auf den Umfang (acht Seiten), eine zwar nicht besonders kreative, aber zumindest auch nicht zwingende Gliederung (z.B. „b. Werbung/und weiter hilfsweise auch für die Packung i) ‚Organic' ii) ‚aus ökologischem Anbau'") und die nicht vorgegebene Wortwahl.[379]

Aus alledem ergibt sich, dass anwaltliche Schriftsätze zwar urheberrechtlich geschützt sein können; ein solcher Schutz ist aber jedenfalls nicht zwingend und gegebenenfalls in seinem Umfang begrenzt. Die Frage des Vorliegens eines solchen Schutzes und von dessen Reichweite lässt sich

373 Vgl. etwa *BVerwG*, GRUR 2020, 189 (Rn. 22); s. auch *OVG Hamburg*, GRUR-RS 2021, 34061; *OLG Düsseldorf*, MDR 2023, 1401 (1401).

374 *BGH*, GRUR 2014, 175; GRUR-RS 2025, 2426 (Rn. 26): für alle Werkarten wird „eine nicht zu geringe Schöpfungshöhe" vorausgesetzt.

375 Ausf. *Rauer/Bibi*, in: Götting/Lauber-Rönsberg/Rauer, § 2 UrhG Rn. 80 m.w.N.; *Loewenheim/Leistner*, in: Schricker/Loewenheim, § 2 UrhG Rn. 60; *Nordemann-Schiffel/Nordemann*, in: Peifer/Kubis/Stieper/Raue, Alles Käse oder vielleicht doch mehr?, S. 241; s. zur Schutzfähigkeit einer Datenschutz-Erklärung *OLG München*, MMR 2023, 974 (Rn. 19).

376 *OLG Düsseldorf*, MDR 2023, 1401 (1402).

377 *OLG Düsseldorf*, MDR 2023, 1401 (1402).

378 *Raue*, in: Dreier/Schulze, § 2 UrhG Rn. 146: „sehr großzügig".

379 *OVG Hamburg*, GRUR-RS 2021, 34061 (Rn. 55 ff.).

nicht abstrakt anhand äußerer Merkmale (z.B. über den Umfang eines Schriftsatzes) beantworten, sondern bedarf vielmehr jeweils einer Einzelfallbetrachtung. Für die automatisierte Verwertung von Schriftsätzen sind daher urheberrechtliche Anforderungen im Ausgangspunkt grundsätzlich zu beachten.

b. Kein Ausschluss nach § 5 Abs. 1 UrhG

Nach § 5 Abs. 1 UrhG genießen „Gesetze, Verordnungen, amtliche Erlasse und Bekanntmachungen sowie Entscheidungen und amtlich verfaßte Leitsätze zu Entscheidungen" keinen urheberrechtlichen Schutz. Die Regelung trägt dem Allgemeininteresse an dem Zugang zu amtlichen Dokumenten Rechnung.[380] Hierdurch werden sämtliche gerichtliche Entscheidungen unabhängig von ihrer Art in Bezug genommen, das umfasst insbesondere Urteile i.S.d. §§ 300 ff. ZPO, Beschlüsse und Verfügungen im Gesamten einschließlich der Entscheidungsgründe[381] (z.B. auch Hinweisbeschlüsse und wohl auch Terminsladungen). Soweit Anwaltsschriftsätze zum Bestandteil der Gerichtsentscheidung werden (z.B. durch eine wörtliche Übernahme von Textpassagen oder die ausdrückliche Inbezugnahme des Anwaltsschriftsatzes), unterfallen auch sie § 5 Abs. 1 UrhG.[382]

Im Übrigen stellt der Akteninhalt einschließlich der anwaltlichen Schriftsätze allerdings keine gerichtliche Entscheidung dar und ist daher nicht nach § 5 Abs. 1 UrhG vom Urheberrechtsschutz ausgenommen.[383] Zur Begründung verweist die Rechtsprechung insoweit auf die Ausnahme für öffentliche Reden in § 48 Abs. 1 Nr. 2 UrhG, die grundsätzlich den Schutz von vor Gericht vorgetragenen Plädoyers und damit auch erst recht von anwaltlichen Schriftsätzen voraussetzt.[384]

Im Regelfall sind anwaltliche Schriftsätze daher schutzfähig und an den Vorgaben des UrhG zu messen.

---

380 *Dreier*, in: Dreier/Schulze, § 5 UrhG Rn. 3.

381 *Arnold*, Amtliche Werke im Urheberrecht, S. 94; *Marquardt*, in: Wandtke/Bullinger, § 5 UrhG Rn. 12; zust. *Ahlberg/Lauber-Rönsberg*, in: Götting/Lauber-Rönsberg/Rauer, § 5 UrhG Rn. 16; *Katzenberger/Metzger*, in: Schricker/Loewenheim, § 5 UrhG Rn. 46, wenngleich mit weitergehender Diff. in Rn. 37.

382 *LG Köln*, GRUR-RR 2011, 5 (6).

383 *BGH*, GRUR 1986, 739 (740); s. auch die Gesetzeshistorie *Katzenberger*, GRUR 1972, 686 (690 f.).

384 *BGH*, GRUR 1986, 739 (740).

c. Verwandte Schutzrechte

Ein etwaiges Leistungsschutzrecht aufgrund der *Zusammenstellung* der Akte als Datenbank i.S.d. §§ 87a ff. UrhG wäre der Justiz als Datenbankherstellerin[385] und nicht dem jeweiligen Rechtsanwalt zuzuordnen. Aus den §§ 87a ff. UrhG ergeben sich also keine im untersuchungsgegenständlichen Fall erhöhten Anforderungen an die Verwertung von anwaltlichen Schriftsätzen.

d. Zwischenergebnis

Anwaltliche Schriftsätze können urheberrechtlichen Werkschutz genießen, wenngleich der Schutzumfang regelmäßig auf einzelne Abschnitte beschränkt sein und verschiedene Schriftsätze ausschließen dürfte.

2. Aktenauszüge im Übrigen

Von den oben unter D.VI.1 dargestellten Akteninhalten sind zahlreiche Inhaltstypen bereits den gerichtlichen Entscheidungen i.S.d. § 5 Abs. 1 UrhG zuzuordnen[386] oder verwertungsrechtlich dem jeweiligen (Justiz-)Arbeitgeber zugeordnet (vgl. § 43 UrhG).[387] Insbesondere Anlagen zu den Schriftsätzen sowie eingeholte Gutachten stechen insoweit allerdings heraus und bedürfen daher einer weitergehenden Erörterung.

385 S. § 87a Abs. 2 UrhG und ErwGr. 41 Datenbank-RL. Hersteller ist in der Regel der Arbeitgeber, s. *Paul*, in: Hilber/Borges, § 87a UrhG Rn. 23; im Überblick zum Begriff des Datenbankherstellers *Vogel*, in: Schricker/Loewenheim, § 87a UrhG Rn. 71 f. m.w.N.

386 S. zuvor unter E.I.1.b.

387 Der Arbeitgeber bzw. Dienstherr erwirbt hiernach i.V.m. dem jeweiligen Arbeitsverhältnis und unter Berücksichtigung der Zweckübertragungslehre Nutzungsrechte an den in Ausübung der Tätigkeit hergestellten Werken. s. hierzu etwa *Wandtke*, in: Wandtke/Bullinger, § 43 UrhG Rn. 49; das betrifft auch Richter, *Nordemann/Obergfell*, in: Loewenheim, § 69 Sonderfragen bei Arbeits- und Dienstverhältnissen, Rn. 10; *Klass*, GRUR 2019, 1103 (1106), wobei die Vorschrift für Hochschullehrer wegen deren eigenverantwortlicher Tätigkeit für nicht anwendbar gehalten wird. Angesichts der auf die gerichtliche Entscheidung i.S.d. § 5 UrhG gerichteten Tätigkeit und insoweit vorbereitenden Tätigkeiten dürften aber auch Richter dem § 43 UrhG unterfallen; s. jüngst zur Unzulässigkeit eines über § 43 UrhG hinausgehenden gesetzlichen Übergangs der Rechte *EuGH*, GRUR-RS 2025, 3103 – Orchestre national de Belgique.

Die Anlagen zu den Schriftsätzen können grundsätzlich urheberrechtlich geschützt sein als Sprachwerke bzw. als Lichtbildwerke (§ 2 Abs. 1 Nr. 1, 5 UrhG) oder leistungsschutzrechtlich geschützt sein mit (noch) geringeren Anforderungen als Lichtbilder (§ 72 UrhG).[388] Eine pauschale Bewertung ist wegen der Vielgestaltigkeit der Form dieser Inhalte nicht möglich. Bei einer automatisierten Verarbeitung dieser Inhalte wird daher zugrunde zu legen sein, dass jedenfalls Teile der betreffenden Anlagen urheber- bzw. leistungsschutzrechtlich geschützt sein dürften.

Für Gutachten als Form einer wissenschaftlichen Ausarbeitung gelten die oben dargestellten Grundsätze zum urheberrechtlichen Schutz von Anwaltsschriftsätzen entsprechend. Ein Gutachten basierend auf (zwingenden) mathematischen Berechnungen zu der Geschwindigkeit(sspanne) eines verunfallten Kraftfahrzeugs ohne nennenswerten Textanteil dürfte regelmäßig keinen urheberrechtlichen Schutz genießen.[389] Demgegenüber liegt für ein umfangreiches psychologisches oder sonst textbasiertes Gutachten die Annahme eines Urheberrechtsschutzes grundsätzlich näher. Vor diesem Hintergrund wird also regelmäßig von einem urheberrechtlichen Schutz auszugehen sein. Im Übrigen können potenziell nicht schutzfähige Gutachten mit höherem Anteil an mathematischen Berechnungen und geringerem Textanteil (z.B. Gutachten zur Wertberechnung) gegebenenfalls anhand der Fachdisziplin identifiziert und geclustert werden.

II. Nutzung innerhalb der Justiz

Zunächst ist eine Verwendung der Aktenauszüge für das Training eines Sprachmodells zu untersuchen, das ausschließlich justizintern eingesetzt wird. Mit Blick auf die geschützten Werke sind zunächst die betroffenen Verwertungsrechte des jeweiligen (Urheber-)Rechts am Werk zu thematisieren (unter 1.), bevor in einem zweiten Schritt untersucht wird, ob sich diese Verwertungs- bzw. Nutzungshandlungen belastbar auf Nutzungsrechtseinräumungen oder urheberrechtliche Schrankenbestimmungen stützen lassen (unter 2.), wobei die Einbindung weiterer Stellen zu berücksichtigen ist (unter 3.). Hieraus werden sodann Bedingungen für das Training und den Einsatz des Sprachmodells abgeleitet (unter 4.).

388 S. etwa *LG Hamburg*, ZUM-RD 2010, 80 (82).

389 Schutz für ein Wertermittlungsgutachten bejahend *LG Hamburg*, ZUM-RD 2010, 80 (82).

1. Betroffene Verwertungsrechte

Für die untersuchungsgegenständlichen Szenarien und Use-Cases ist mit Blick auf urheberrechtliche Verwertungsrechte[390] zwischen der Zusammenstellung der Trainingsdaten, dem eigentlichen Training i.w.S. und der Generierung der Ausgabe zu differenzieren. Das Training selbst als „Umwandlung" von Sprachwerken in eine mathematische, wahrscheinlichkeitsbasierte Repräsentation ist urheberrechtlich nicht relevant,[391] ebenso wenig der reine Werkgenuss;[392] urheberrechtlich von Bedeutung sind vielmehr etwaige Vervielfältigungen i.S.d. § 16 UrhG vor, während und nach dem (KI-)Trainingsprozess.

a. Zusammenstellung des Trainingskorpus und Vervielfältigungen im Trainingsprozess

Die urheberrechtliche Vervielfältigung betrifft sämtliche „körperlichen Festlegungen, die geeignet sind," wiedererkennbare, prägende Bestandteile des Werks den „menschlichen Sinnen" wahrnehmbar zu machen.[393] Entscheidend sind somit die Wahrnehmbarkeit und – vor allem auch im Hinblick auf veränderte Werkbestandteile – die Wiedererkennbarkeit[394] des Werks oder von prägenden Werkteilen, die ihrerseits die Voraussetzungen des § 2 UrhG erfüllen.[395] Nach dem deutschsprachlichen Begriffsverständnis soll auch die Eignung zur mittelbaren Wahrnehmung genügen,[396] worauf nachfolgend zurückzukommen sein wird (unter bb.). Demgegenüber

390 Zu dem Handlungsbezug der Verwertungsrechte s. *Sesing-Wagenpfeil*, ZGE 16 (2024), 212 (221) m.w.N.

391 *Schack*, NJW 2024, 113 (Rn. 8); *Bomhard*, InTeR 2023, 174 (175); vgl. BT-Drs. 18/12329, 40.

392 Statt aller *Sucker*, Der digitale Werkgenuss im Urheberrecht, S. 50 ff., 59, 87 ff.; *Beurskens*, RDi 2025, 1 (Rn. 7); *Schack*, NJW 2024, 113 (Rn. 9).

393 BT-Drs. IV/270, 47; *BGH*, GRUR 2017, 793 (Rn. 41); GRUR 2009, 942 (Rn. 25).

394 *EuGH*, GRUR 2019, 929 (Rn. 36 ff.) – Pelham ua [Metall auf Metall III]; *Grisse*, RuZ 2020, 143 (147).

395 *Raue*, in: Dreier/Schulze, § 16 UrhG Rn. 14; zur str. Abgrenzung zu § 23 UrhG *Raue*, in: Dreier/Schulze, § 23 UrhG Rn. 16.

396 BT-Drs. IV/270, 47; *BGH*, GRUR 2017, 793 (Rn. 41); GRUR 2009, 942 (Rn. 25).

sind der Zweck, das Format[397] oder die beabsichtigte Speicherdauer[398] ohne Belang für die Frage nach dem Vorliegen einer Vervielfältigung.

aa. Vervielfältigungen im Einzelnen

Falls Aktenauszüge in Vorbereitung des Trainingsprozesses zunächst digitalisiert werden, ist hierin bereits eine urheberrechtliche Vervielfältigung zu sehen.[399] Das Zusammenführen der digitalen Aktenauszüge an einem Ort[400] in Vorbereitung des Trainings des Sprachmodells geht ebenfalls mit Vervielfältigungen einher.[401]

Im GSJ-Projekt lassen sich Verarbeitungen gegebenenfalls durch die Weiterverarbeitung an der Dokumentquelle und abgeleitete Textformate[402] vermeiden.[403] Die Weiterverarbeitung an der Dokumentquelle meint, dass möglichst viele Verarbeitungs- und Trainingsschritte lokal ohne erneute Zwischenspeicherung in einer Datei vorgenommen werden. Nach Abschluss des Verarbeitungsprozesses werden die benötigten Daten (i.e. nicht das ursprüngliche Dokument) sodann als abgeleitetes Textformat an die KI-Trainingsserver übermittelt. Dieses abgeleitete Textformat könnte mangels Wahrnehmbarkeit und Wiedererkennbarkeit nicht mehr als Abbildung eines urheberrechtlich geschützten Werkes anzusehen sein (hierzu sogleich unter bb.),[404] sodass die Daten grundsätzlich ohne urheberrechtli-

397 Etwa *BGH*, GRUR 2010, 628 (Rn. 17) m.w.N.

398 Das zeigt insb. die in § 44a UrhG niedergelegte Schrankenbestimmung, die temporäre Vervielfältigungen adressiert und mithin die Anwendbarkeit des UrhG auf diese temporären Vervielfältigungen unterstellt, s. *Sucker*, Der digitale Werkgenuss im Urheberrecht, S. 75 ff.; *Käde*, Kreative Maschinen und Urheberrecht, S. 69.

399 Vgl. etwa *OLG Hamburg*, GRUR-RR 2008, 378 (380 f.); *BGH*, GRUR 1999, 325 (327).

400 Z.B. das Anlegen eines digitalen Trainingskorpus oder das Hochladen einer Datei auf einen KI-Trainingsserver. S. aber zur Rechtfertigung nach § 44a UrhG nachfolgend unter E.II.2.d.

401 Vgl. auch *La Durantaye*, ZUM 2023, 645 (647).

402 Hierzu etwa *Raue/Schöf*, RuZ 2020, 118; *Grisse*, RuZ 2020, 143 m.w.N.; *Organisciak/Downie*, in: Golub/Y.-H. Liu, Research access to in-copyright texts in the humanities; *Schöch et al.*, Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen; *Schöch et al.*, RuZ 2020, 160; *Brockmeyer*, Text und Data Mining, S. 200 ff.

403 Im Überblick zu der Frage, ob sich Vervielfältigungen vermeiden lassen, s. *Jager*, Artificial Creativity?, S. 329 m.w.N.

404 *Raue/Schöf*, RuZ 2020, 118 (119): „durch eine gezielte Informationsreduktion so zu transformieren, dass die urheberrechtlich geschützte Textstruktur irreversibel verloren geht"; s. zur gegebenenfalls möglichen Rekonstruierbarkeit von abgeleiteten

che Implikationen verwertet werden könnten.[405] Da Informationsverluste denkbar sind,[406] bedarf die Frage der Realisierbarkeit im vorliegend zu untersuchenden Sachverhalt einer technischen Prüfung, die allerdings nicht Gegenstand dieses Rechtsgutachtens ist.

bb. Vervielfältigung bei möglicher Rekonstruierbarkeit aus abgeleiteten Textformaten oder anhand eines Sprachmodells

Die Auswirkungen einer Rekonstruierbarkeit auf die Annahme einer Vervielfältigung (insbesondere über eine Eignung zur mittelbaren Wahrnehmbarkeit) betreffen nicht nur abgeleitete Textformate, sondern auch und gerade Sprachmodelle. Analog zur datenschutzrechtlichen Einordnung[407] könnten auch Sprachmodelle in Abhängigkeit von den möglichen Ausgaben als Vervielfältigung anzusehen sein. Sprachmodelle ähneln den abgeleiteten Textformaten, indem ihre Gewichte eine – allerdings wahrscheinlichkeitsbasierte – mathematische Repräsentation von Wort-Vektoren darstellen.

Es ist nicht abschließend geklärt, ob und unter welchen Voraussetzungen eine mögliche Rekonstruierbarkeit des Werks anhand der abgeleiteten Textformate[408] oder eines Sprachmodells[409] über eine mittelbare Wahrnehm-

Textformaten *Kugler et al.*, in: InvBERT: Reconstructing Text from Contextualized Word Embeddings by inverting the BERT pipeline.

405 Das steht unter dem Vorbehalt, dass diese Daten nicht wieder zusammengeführt werden (können) und das ursprüngliche Werke erkennbar abbilden.

406 Zu möglichen Formaten *Grisse*, RuZ 2020, 143 (152 ff.): Term-Dokument-Matrizen, Lemmatisierung, gestörte Sequenzinformationen, N-Gramm-Tabellen und Wort-Vektoren; zu den Wort-Vektoren auf Basis von Word Embedding im Natural Language Processing *Schöch et al.*, RuZ 2020, 160 (171 ff.).

407 S. oben unter D.I.1.b.bb.

408 Offen mit Tendenz zur Berücksichtigung der Rekonstruierbarkeit *Grisse*, RuZ 2020, 143 (150 f.); *Raue/Schöf*, RuZ 2020, 118 (124): nicht „mit verhältnismäßigem Aufwand rekonstruierbar".

409 Ein KI-Modell stellt kein Vervielfältigungsstück dar, nach *La Durantaye*, AfP 2024, 9 (Rn. 19); *Leistner*, GRUR 2024, 1665 (1668 f.): wegen wahrscheinlichkeitsbasiertem Aufbau nicht vergleichbar mit einer Kopie; ähnl. *Konertz/Schönhof*, WRP 2024, 534 (Rn. 8); *Antoine*, GRUR 2025, 118 (120); implizit *Maamar*, ZUM 2023, 481 (486), wonach ein einmal im Ausland trainiertes Sprachmodell ohne weitere urheberrechtliche Implikationen in der EU angeboten werden kann; ähnl. *Peukert*, GRUR Int. 2024, 497 (506); Tendenz gegen eine Vervielfältigung bei *Baumann*, NJW 2023, 3673 (Rn. 10 ff.); „im Regelfall" abzulehnen nach *Schmid*, KIR 2025, 69 (72); abhängig von der Konzeption und intendierten Einsatzes des Modells *Käde*,

barkeit zur Annahme einer Vervielfältigung führt. Diese Unsicherheit ist auch auf die fehlende (EuGH-)Rechtsprechung zu der konkreten Fragestellung zurückzuführen; es ist unklar, ob und inwieweit die Eignung zur mittelbaren menschlichen Wahrnehmung für die Annahme einer Vervielfältigung genügt. Im Ausgangspunkt adressiert Art. 2 InfoSoc-RL jede „unmittelbare oder mittelbare, vorübergehende oder dauerhafte Vervielfältigung auf jede Art und Weise und in jeder Form" und wurde durch den EuGH bereits weit ausgelegt mit Blick auf Werkteile.[410]

Aus der bisher insoweit ergangenen Rechtsprechung des EuGH lässt sich allerdings jedenfalls eine Tendenz ableiten, auch subjektive Aspekte in die Prüfung einer Vervielfältigung einzustellen[411] (z.B. das fehlende Interesse einer Person, ein Werk tatsächlich zu rekonstruieren) und die objektiv ermittelte Rekonstruierbarkeit nicht ausreichen zu lassen. Insoweit wird zutreffend insbesondere auch darauf hingewiesen,[412] dass der EuGH – allerdings auf Ebene des Schutzumfangs statt mit Blick auf die Verwertung – über den Werkbegriff eine „Ausdrucksform des urheberrechtlichen Schutzobjekts [erfordert], die es mit hinreichender Genauigkeit und Objektivität identifizierbar werden lässt".[413] Eine solche Ausdrucksform ist bei stets wahrscheinlichkeitsbasierten Sprachmodellen nicht gegeben; dieser Befund gilt auch bei den hier untersuchungsgegenständlichen Use-Cases und dem Einsatz von Ausgabefiltern.

Ergänzend legt ein teleologischer Vergleich des jeweils europarechtlich determinierten sowie präformierten Urheberrechts und Datenschutzrechts

ZUM 2024, 174 (182); an einer Vervielfältigung zweifelnd *Hofmann*, WRP 2024, 11 (Rn. 12); a.A. *Dornis/Stober*, Urheberrecht und Training generativer KI-Modelle, S. 54 f.; *Jager*, Artificial Creativity?, S. 333; *Sesing-Wagenpfeil*, ZGE 16 (2024), 212 (236 ff.): soweit sich Ausgaben nachweisbar auf Trainingsdaten zurückführen lassen; *Beurskens*, RDi 2025, 1 (Rn. 9); *Pesch/Böhme*, GRUR 2023, 997 (1004 f.), nach denen auch die „Schwierigkeit der Rekonstruktion" nicht schaden soll; womöglich auch *Welser*, GRUR-Prax 2023, 516 (Rn. 8), der darauf abstellt, dass Modelle Trainingsdaten „speichern" können.

410 *EuGH*, GRUR 2009, 1041 (Rn. 38 ff.) – Infopaq; *Spindler*, in: Peifer/Kubis/Stieper/Raue, Künstliche Intelligenz und Urheberrecht aus europäischer Perspektive, S. 349.

411 Zum subjektiven Vervielfältigungsrecht *Raue*, ZGE 9 (2017), 514; *Grisse*, RuZ 2020, 143 (151) mit einem Überblick über die EuGH-Rechtsprechung in Fn. 51; objektiv hingegen nach *Dornis/Stober*, Urheberrecht und Training generativer KI-Modelle, S. 51.

412 *Leistner*, GRUR 2024, 1665 (1669); *EuGH*, GRUR 2019, 73 (Rn. 40) – Levola Hengelo BV/Smilde Foods BV.

413 *EuGH*, GRUR 2019, 73 (Rn. 40) – Levola Hengelo BV/Smilde Foods BV.

die zurückhaltende Annahme einer Vervielfältigung durch das Sprachmodell (oder abgeleitete Textformate) nahe.

Für das Datenschutzrecht ist die Verknüpfung von Datum und natürlicher Person grundlegend (i.e. Art. 4 Nr. 1 DSGVO, Art. 8 GRCh, Art. 2 Abs. 1 i.V.m. Art. 1 Abs. 1 GG). Gerade diese Verknüpfung ermöglicht Rückschlüsse auf die natürliche Person und Auswirkungen auf deren Handeln sowie ihr Privat- und Familienleben (Art. 7 GRCh, Art. 8 EMRK), die ein entscheidender Grund für die grundrechtliche Gewährleistung in Art. 8 GRCh sind (siehe auch Art. 2 Abs. 1 i.V.m. Art. 1 Abs. 1 GG).[414] Der EuGH legt die DSGVO vor diesem Hintergrund weit aus.[415] Bereits eine nur mittelbare Verknüpfung im Wege der Identifizierbarkeit mit verfügbaren Mitteln kann hiernach die betroffene Person tangieren.

Dem Urheberrecht ist eine derart enge Verknüpfung zwischen Werk und Urheber nicht fremd. So stellt sich das Urheberpersönlichkeitsrecht nach §§ 12 ff. UrhG dar als Ausdruck der ideellen Beziehung des Urhebers zu seinem Werk aus dem allgemeinen Persönlichkeitsrecht.[416] Das Werk ist eine individuelle schöpferische Leistung des Urhebers und damit Ausdruck seiner Persönlichkeit. Zugleich und von vorrangiger Bedeutung für das untersuchungsgegenständliche Projekt sieht das Urheberrecht Verwertungsrechte in §§ 15 ff. UrhG vor, die nicht primär persönlichkeitsrechtlich unterlegt sind, sondern vielmehr Vermögensinteressen ausgehend von dem geschaffenen Werk schützen (Art. 14 GG). Ein Werkformat, aus dem sich das ursprüngliche Werk rekonstruieren lässt, berührt vor diesem Hintergrund noch nicht die Verwertungsinteressen des Urhebers. Die bloße Wahrscheinlichkeit einer möglichen Beeinträchtigung der Verwertungsinteressen genügt insoweit gerade nicht; im Fall einer *tatsächlichen* Rekonstruktion liegt eine Vervielfältigung nach § 16 UrhG vor und ermöglicht dem Urheber die Kontrolle der Verwertung. Die (mittelbare) Wahrnehmbarkeit als Voraussetzung einer Vervielfältigung ist hierbei vor dem Hintergrund zu sehen, dass die menschliche Wahrnehmung eines Werks (der „Werkgenuss") nicht

414 Beispielhaft *EuGH*, MMR 2023, 105 (Rn. 57 ff.) – Recht auf Vergessenwerden; *Ehmann/Selmayr*, in: Ehmann/Selmayr, Einl. Rn. 30 f.; *Buchner*, in: Kühling/Buchner, Art. 1 DSGVO Rn. 10: „Recht auf Privatleben ist Grundvoraussetzung für die Entwicklung einer freien und selbstbestimmten Persönlichkeit".

415 Z.B. mit Blick auf personenbezogene Daten, vgl. *EuGH*, NJW 2018, 767 (Rn. 33) – Nowak m.w.N.; mit Blick auf den Begriff des Verantwortlichen ZD 2024, 328 (Rn. 54 f.) – IAB Europe.

416 *Specht-Riemenschneider*, in: Dreier/Schulze, Vor § 12 UrhG Rn. 1, 5 ff.; *Peukert*, in: Schricker/Loewenheim, Vor § 12 UrhG Rn. 29 ff.

geschützt ist[417] und daher bereits der unmittelbar einer Wahrnehmung vorgelagerte Akt erfasst sein soll. Die Konzentration des Urheberrechts auf Verwertungsformen (§ 15 UrhG) sperrt eine Erstreckung des Urheberrechts auf das Vorfeld einer Verwertung, ebenso wie auf den Nachgang in Form des Werkgenusses.

Das Datenschutzrecht beschränkt sich nicht auf einzelne Verwertungsformen, sondern umfasst nach Art. 4 Nr. 2 DSGVO sämtliche (teil-)automatisierten Vorgänge im Zusammenhang mit personenbezogenen Daten. Anders als das Urheberrecht verlangt die DSGVO umfangreiche technische und organisatorische Maßnahmen (Art. 24 ff. DSGVO) zur sicheren Aufbewahrung und Verarbeitung personenbezogener Daten. Diese datenschutzrechtlichen Anforderungen drohten konterkariert zu werden, wenn Daten mit Identifizierungsmöglichkeit der betroffenen Person ohne jede Sicherheitsmaßnahme gespeichert werden könnten und die DSGVO-Vorgaben erst ab der tatsächlichen Identifizierung eingreifen würden.

Hinzu kommt, dass der EuGH in seiner einschlägigen Rechtsprechung im Urheberrecht verschiedene Wertungen in die Auslegung der Verwertungsrechte einfließen lässt.[418] Die Berücksichtigung der vorgenannten Wertungen bei der Auslegung der urheberrechtlichen Vervielfältigung steht also nicht im Widerspruch zur bisherigen EuGH-Judikatur. Es liegt daher nahe, dass sowohl abgeleitete Textformate als auch trainierte Sprachmodelle grundsätzlich keine Vervielfältigung von urheberrechtlichen Werken darstellen werden. Unberührt hiervon bleibt eine tatsächlich erfolgende Vervielfältigung aufgrund des Sprachmodells in der Ausgabe des KI-Systems.[419]

Da diese Frage durch den EuGH zum jetzigen Zeitpunkt (noch) ungeklärt ist, besteht ein (Rest-)Risiko, dass die Rechtsprechung geringere Anforderungen an die Annahme einer Vervielfältigung stellt. Um diese Unsicherheit zu adressieren, sollten die bereits mit Blick auf die DSGVO unter D.VIII.4 dargestellten Mechanismen auch zur Verhinderung einer Rekonstruierbarkeit urheberrechtlich geschützter Werke zum Einsatz kommen. Diese Empfehlung betrifft insbesondere die Implementierung eines geeigneten System Prompts und der Ausgabefilter.

417 *Beurskens*, RDi 2025, 1 (Rn. 7); zum Hintergrund auch *Schack*, NJW 2024, 113 (Rn. 9).
418 Ausf. *Grisse*, RuZ 2020, 143 (151) m.w.N. in Fn. 51.
419 S. nachfolgend unter E.II.1.b.

b. Einsatz und Ausgabe des KI-Systems

Bei dem Einsatz des KI-Systems sind Vervielfältigungen unter Anwendung der vorstehenden Maßgaben systemintern regelmäßig abzulehnen.[420] Denn die mathematischen Operationen basieren nicht auf wahrnehmbaren oder rekonstruierbaren Werkbestandteilen.

Allerdings kann die Ausgabe eines KI-Systems durchaus Bestandteile aus Trainingsdaten enthalten,[421] die durch die Ausgabe vervielfältigt werden (§ 16 Abs. 1 UrhG). Sofern die Ausgabe im Einzelfall über eine Internetseite öffentlich zugänglich gemacht wird, greift die speziellere Vorschrift des § 19a UrhG ein.[422] Eine solche Öffentlichkeit[423] dürfte im untersuchungsgegenständlichen Szenario der Sachverhaltsaufbereitung in den Use-Cases 1-3 jedoch kaum in Betracht kommen, sodass es regelmäßig bei einer möglichen Vervielfältigung bleiben wird.

Die Wahrscheinlichkeit einer solchen Vervielfältigung kann durch verschiedene Mechanismen reduziert werden, wie etwa durch den System Prompt und durch die Implementierung von Ausgabefiltern. Ferner können entsprechende (Dienst-)Anweisungen zum Umgang mit dem KI-System Einfluss auf die Eingabeaufforderungen nehmen und hierüber die Wahrscheinlichkeit einer Vervielfältigung erheblich reduzieren. Sofern es im Einzelfall dennoch zu Vervielfältigungen kommt (in der Regel im Zusammenhang mit missbräuchlichen Eingabeaufforderungen), bestehen Zweifel an der Passivlegitimation der Projektverantwortlichen.[424] Im Übrigen kann für vereinzelte Vervielfältigungen auch die Schranke für unwesentliches Beiwerk nach § 57 UrhG in den Blick genommen werden.[425]

420 Anders etwa *Schack*, NJW 2024, 113 (Rn. 4).

421 Etwa *Karamolegkou et al.*, Copyright Violations and Large Language Models. S. auch schon oben unter D.IV.1.b.cc.

422 *Dornis/Stober*, Urheberrecht und Training generativer KI-Modelle, S. 55; vgl. auch *Raue*, CR 2017, 656 (660).

423 *Dreier*, in: Dreier/Schulze, § 60d UrhG Rn. 10.

424 Hierzu auch unter E.III.

425 Dies diskutierend im Hinblick auf das Textkorpus *Raue/Schöf*, RuZ 2020, 118 (123 f.); krit. *Brockmeyer*, Text und Data Mining, S. 203 f.

c. Bearbeitungen des Werkes im Rahmen des Trainingsprozesses und des Einsatzes des KI-Systems

Wenn und soweit das Werk im Laufe des Trainingsprozesses und bei dem Einsatz des KI-Systems verändert wird, ist darüber hinaus § 23 UrhG[426] mit gesonderten Anforderungen an Bearbeitungen und Umgestaltungen eines Werks zu beachten.[427] Technisch bedingte Änderungen bei der automatisierten Analyse von Werken (dem sog. Text und Data Mining)[428] sind allerdings nach § 23 Abs. 3 UrhG ausgenommen. Das betrifft nach vorzugswürdiger Auffassung beispielsweise die modellinterne Verarbeitung der Werke und zufällige Bearbeitungen von rekonstruierbaren Werkbestandteilen, repräsentiert als Zahlen im Modell.

d. Zwischenergebnis

Aus urheberrechtlicher Sicht sind insbesondere Vervielfältigungen der in den Aktenauszügen enthaltenen Werke relevant. Sofern diese Vervielfältigungen bereits in digitaler Form vorliegen, lässt sich eine Anwendbarkeit des Urheberrechts über abgeleitete Textformate verhindern. Eine solche Verhinderung steht ihrerseits unter dem Vorbehalt der technischen Machbarkeit. Nach vorzugswürdiger, wenngleich teilweise bestrittener Ansicht liegt auch dann regelmäßig keine Vervielfältigung vor, wenn sich anhand des gespeicherten Sprachmodells allfällige Werkteile mit nicht nur unerheblichem Aufwand rekonstruieren lassen. Insoweit ist mit Blick auf die Vervielfältigungen ein zugunsten der urheberrechtlichen Zulässigkeit großzügigerer Maßstab als unter der DSGVO mit Blick auf die Identifizierung natürlicher Personen anzusetzen.

In jedem Fall empfiehlt es sich, durch (Dienst-)Anweisungen zum Umgang mit dem KI-System, durch Ausgabefilter und den System Prompt Vervielfältigungen im Rahmen der Ausgabe des KI-Systems möglichst weitgehend zu verhindern. Sofern es dennoch zu entsprechenden Vervielfältigungen kommt, sind etwaige Nutzungsrechtseinräumungen und die urheberrechtlichen Schrankenbestimmungen zu prüfen.

---

426 Unabhängig davon, ob dem Bearbeitungsrecht noch eine eigenständige Bedeutung zukommt, s. *Schack*, ZGE 15 (2023), 263 (271 f.); vgl. auch *BGH*, GRUR 2010, 628 (Rn. 17); gegen eine eigenständige Bedeutung etwa *Raue*, in: Dreier/Schulze, § 23 UrhG Rn. 16; a.A. *Götting*, in: Götting/Lauber-Rönsberg/Rauer, § 16 UrhG Rn. 10.

427 *Schack*, NJW 2024, 113 (Rn. 4); *Jager*, Artificial Creativity?, S. 333 ff.

428 S. nachfolgend unter E.II.2.c.

Die Veränderung der Werke durch das Sprachmodell ist wegen § 23 Abs. 3 UrhG und der vorrangigen Vervielfältigung nach § 16 UrhG von untergeordneter Bedeutung.

2. Zulässigkeit aufgrund von Nutzungsrechtseinräumungen oder Schrankenbestimmungen

Allfällige Verwertungshandlungen (wie etwa die Vervielfältigung) sind gegebenenfalls aufgrund einer Nutzungsrechtseinräumung oder einschlägigen Schrankenbestimmungen unter Beachtung des Dreistufentests i.S.d. Art. 5 Abs. 5 InfoSoc-RL[429] zulässig.

a. Nutzungsrechtseinräumungen

Während für innerhalb der Justiz erarbeitete Inhalte regelmäßig bereits ein Nutzungsrecht eingeräumt worden sein wird (vgl. § 43 UrhG),[430] ist eine (konkludente)[431] Rechtseinräumung nach § 29 Abs. 2, § 31 UrhG für das KI-Training nicht durch die Einreichung des Schriftsatzes oder eines angehängten Gutachtens bei Gericht anzunehmen. Selbst wenn bereits die Einreichung eines Schriftsatzes als eine solche Rechtseinräumung anzusehen wäre,[432] wird eine solche Einräumung entsprechend dem Zweck der Einreichung auf die Verwertung mit Blick auf das konkrete gerichtliche Verfahren beschränkt sein (vgl. § 31 Abs. 5 UrhG). Bei gerichtlich beauftragten Sachverständigengutachten (z.B. §§ 411, 411a ZPO) kann eine solche Nut-

429 S. u.a. Art. 5 Abs. 5 InfoSoc-Richtlinie: „Die […] Beschränkungen dürfen nur [1.] in bestimmten Sonderfällen angewandt werden, in denen [2.] die normale Verwertung des Werks oder des sonstigen Schutzgegenstands nicht beeinträchtigt wird und [3.] die berechtigten Interessen des Rechtsinhabers nicht ungebührlich verletzt werden."

430 S. oben unter E.I.2.

431 Nur *Paul*, in: Holznagel/Hoeren/Sieber, Teil 7.4, Rn. 93.

432 Gegebenenfalls ist eher eine Rechtseinräumung zugunsten des Mandanten denkbar, vgl. *OVG Hamburg*, BeckRS 2020, 30248.

zungsrechtseinräumung zwar tendenziell eher angedacht werden,[433] scheint aber im Schrifttum kaum Anklang zu finden.[434]

Grundsätzlich denkbar ist die Beschaffung entsprechender Nutzungsrechte mit Blick auf Werke von Anwälten und Gutachtern, die gemäß dem geänderten Wahrnehmungsvertrag der VG Wort die sog. „KI-Lizenz" eingeräumt haben.[435] In der Praxis dürfte gegenwärtig nur eine geringe Anzahl der Anwälte und Gutachter bei der VG Wort gemeldet sein sowie eine entsprechende Lizenz eingeräumt haben.

b. Rechtspflege (§ 45 UrhG)

Nach § 45 Abs. 1 UrhG dürfen „einzelne Vervielfältigungsstücke von Werken zur Verwendung in Verfahren vor einem Gericht" hergestellt werden und nach Abs. 3 ist „unter den gleichen Voraussetzungen […] auch die Verbreitung, öffentliche Ausstellung und öffentliche Wiedergabe der Werke zulässig".[436] Die Schranke stützt sich auf Art. 5 Abs. 3 lit. e, Abs. 4 InfoSoc-RL, die mitgliedstaatliche Schrankenbestimmungen zur „Sicherstellung des ordnungsgemäßen Ablaufs von […] Gerichtsverfahren" zulässt. Die Regelung soll die Aufgabenerfüllung der Gerichte sicherstellen,[437] indem Vervielfältigungen zu Beweiszwecken und als sonstige Hilfsmittel für eine Entscheidung ermöglicht werden.[438]

Der historische Gesetzgeber ging von einer bloß gerichtsinternen und auf das jeweilige Verfahren beschränkten Verwertung aus und schloss daher unveröffentlichte Werke ein.[439] Dementsprechend verlangt auch die (wohl) herrschende Auffassung zutreffend eine Verwendung in konkreten

433 Für in der Vergangenheit erstattete Gutachten wäre nicht ohne Weiteres auch eine Übertragung der Rechte mit Blick auf Text und Data Mining anzunehmen, sondern § 31a UrhG anzuwenden, vgl. *Hamann*, ZGE 16 (2024), 113 (138 f.).

434 Etwa *Ulrich*, DSB 2011, 308 (315), der keine Nutzungsrechteinräumung annimmt, sondern § 45 UrhG anwendet; s. aber zu Privatgutachten *T. Heinrich*, NZV 2015, 68 (68 f.).

435 S. https://www.vgwort.de/veroeffentlichungen/aenderung-der-wahrnehmungsbedingungen/fragen-/-antworten-zur-aenderung.html.

436 Zum Änderungsverbot und der Quellenangabe nach §§ 62, 63 UrhG *Lüft*, in: Wandtke/Bullinger, § 45 UrhG Rn. 1.

437 *Lüft*, in: Wandtke/Bullinger, § 45 UrhG Rn. 1.

438 BT-Drs. IV/270, 63.

439 BT-Drs. IV/270, 63; *BVerwG*, GRUR 2020, 189 (Rn. 43); *Melichar/Stieper*, in: Schricker/Loewenheim, § 45 UrhG Rn. 13.

Gerichtsverfahren,[440] wobei die Anzahl der Vervielfältigungen auf den Bedarf für das konkrete Verfahren zu beschränken ist.[441] Da der Begriff der Vervielfältigung nicht zwischen digitalen und analogen Niederlegungen differenziert,[442] sind im Ausgangspunkt auch digitale Vervielfältigungen einschließlich der Digitalisierung bei Führung einer digitalen Akte umfasst. Das gilt ebenso für Vervielfältigungen aus übergebenen (neuen) Akten in den Use-Cases 1-3.

Eine Digitalisierung von bestehenden Aktenauszügen nur zur Durchführung des KI-Trainings weist allerdings keinen konkreten Verfahrensbezug auf, weshalb dieser Vorgang nicht auf § 45 UrhG gestützt werden kann.

Unter Berücksichtigung des Schutzzwecks und des angelegten konkreten Verfahrensbezugs kann die Vorschrift zudem auch nicht im Wege einer Analogie auf Vervielfältigungen für das Training eines Sprachmodells als Schritt zur Verbesserung der Arbeitsabläufe in allen Gerichtsverfahren erstreckt werden. Denn die maßgebliche Interessenlage unterscheidet sich maßgeblich durch den weitergehenden Verwertungsumfang im Fall des Trainings eines Sprachmodells, selbst wenn das Sprachmodell zwar nur justizintern, aber eben gerichts- und verfahrensübergreifend zum Einsatz kommt. Zwar könnte der deutsche Gesetzgeber womöglich im Einklang mit dem Dreistufentest und der InfoSoc-RL die Schranke des § 45 UrhG, gegebenenfalls i.V.m. einer Vergütungspflicht, erweitern. Die derzeitige Regelung durch den deutschen Gesetzgeber gibt aber (noch) keine solche Auslegung her.

c. Text und Data Mining (§§ 44b UrhG und 60d UrhG)

Die § 44b, § 60d UrhG lassen gestützt auf[443] Art. 3, 4 DSM-Richtlinie Vervielfältigungen für das sog. Text und Data Mining ohne Einschränkungen

440 *Lüft*, in: Wandtke/Bullinger, § 45 UrhG Rn. 3; *BGH*, GRUR 2021, 711 (713); *LG Düsseldorf*, GRUR-RR 2007, 193 (194); *Dreier*, in: Dreier/Schulze, § 45 UrhG Rn. 6; *Melichar/Stieper*, in: Schricker/Loewenheim, § 45 UrhG Rn. 6; *Paul*, in: Hilber/Borges, § 45 UrhG Rn. 1; *Schulz*, in: Götting/Lauber-Rönsberg/Rauer, § 45 UrhG Rn. 8; vgl. auch *BGH*, GRUR 2010, 623; *Wiebe*, in: Spindler/Schuster, § 45 UrhG Rn. 2; wohl auch *Konertz*, ZUM 2024, 355 (363).

441 *Schulz*, in: Götting/Lauber-Rönsberg/Rauer, § 45 UrhG Rn. 5.

442 Statt aller *Heerma*, in: Wandtke/Bullinger, § 16 UrhG Rn. 4.

443 S. zum Begriff der Ausnahme und Beschränkung in der DSM-Richtlinie und den Konsequenzen für die Mitgliedstaaten *Brockmeyer*, Text und Data Mining, S. 70 f.; schon zuvor *Schack*, GRUR 2021, 904.

des Zwecks mit Widerspruchslösung (§ 44b Abs. 2, 3 UrhG) bzw. für wissenschaftliche Forschungszwecke ohne Widerspruchsmöglichkeit zu (§ 60d UrhG). Text und Data Mining ist nach § 44b Abs. 1 UrhG „die automatisierte Analyse von einzelnen oder mehreren digitalen oder digitalisierten Werken, um daraus Informationen insbesondere über Muster, Trends und Korrelationen zu gewinnen".[444]

aa. KI-Training als Text und Data Mining

Nicht nur die Digitalisierung von analogen Dokumenten,[445] sondern auch das weit verstandene[446] KI-Training, unter Einschluss eines Sprachmodells, ist mit der (wohl) herrschenden Auffassung[447] als Text und Data Mining i.S.d. § 44b UrhG zu qualifizieren.

Die Diskussion um die Anwendung und Reichweite des Text und Data Mining wurde unlängst vor allem durch ein Werk von *Dornis und Stober* neu entfacht.[448] Unbestritten konnte der europäische Gesetzgeber bei Verabschiedung der DSM-RL vom 17. April 2019 noch nicht die Dimension des Scrapings für das Training von KI-Modellen vorhersehen,[449] die in

444 S. auch die materiell übereinstimmende Definition aus Art. 2 Nr. 2 DSM-RL: „Text und Data Mining bezeichnet eine Technik für die automatisierte Analyse von Texten und Daten in digitaler Form, mit deren Hilfe Informationen unter anderem — aber nicht ausschließlich — über Muster, Trends und Korrelationen gewonnen werden können." sowie *T. Radtke*, ZGE 17 (2025), 1 (30 f.).

445 *Raue*, CR 2017, 656 (658); BT-Drs. 18/12329, 41.

446 D.h. die Zusammenstellung des Trainingskorpus, Pre-Training, Finetuning und weitere Schritte auf dem Weg zu einem KI-Modell, das in ein KI-System implementiert wird, sind umfasst.

447 S. hierzu insb. die Nachweise bei *T. Radtke*, ZGE 17 (2025), 1 (36) in Fn. 176; *Hamann*, ZGE 16 (2024), 113 (121) m.w.N.; *Bomhard*, in: Götting/Lauber-Rönsberg/Rauer, § 44b UrhG Rn. 11b ff. m.w.N.; s. auch BT-Drs. 19/27426, 60: „für das maschinelle Lernen als Basis-Technologie für Künstliche Intelligenz von besonderer Bedeutung"; *Schwartmann/Köhler*, in: Schwartmann/Keber/Zenner, 3. Kapitel, Rn. 113; einen Abgleich i.R.d. der Kuratierung von Trainingsdaten als TDM einordnend, die Frage aber im Übrigen offenlassend *LG Hamburg*, GRUR 2024, 1710 (Rn. 40 ff.); a.A. *Dornis/Stober*, Urheberrecht und Training generativer KI-Modelle, S. 94 ff.; außerdem sei die Diskussion noch nicht beendet nach *Geiger*, in: Thouvenin/Peukert/Jaeger/Geiger, When the Robots (Try to) Take Over: Of Artificial Intelligence, Authors, Creativity and Copyright Protection, S. 77

448 *Dornis/Stober*, Urheberrecht und Training generativer KI-Modelle.

449 Zurecht aber auf die absehbare (Weiter-)Entwicklung von generativer KI hinweisend *Leistner*, GRUR 2024, 1665 (1669); zur Diskussion im Vorfeld der Regelung s. auch *Beurskens*, RDi 2025, 1 (Rn. 17).

Chatbots wie ChatGPT zum Einsatz kommen. Maßgeblich aus diesem Umstand leitet *Dornis* ab,[450] dass Text und Data Mining womöglich zwar Teilaspekte eines KI-Trainingsprozesses umfassen können, nicht aber sämtliche Vervielfältigungen im Zusammenhang mit dem KI-Training.[451] Schon zuvor wurden in der Literatur verschiedentlich Zweifel mit Blick auf einen Informationsgewinn unmittelbar durch den Trainingsprozess und[452] eine gebotene enge Auslegung als Ausnahmebestimmung[453] geäußert.

Dieser Auffassung ist allerdings entgegenzuhalten, dass es dem Gesetzgeber grundsätzlich freisteht, ein weites Begriffsverständnis zu unterstellen. Die Definition des Text und Data Mining ist dementsprechend weitgefasst[454] und schon vor einigen Jahren wurde ein enormes Wertschöpfungspotenzial in Text und Data Mining erblickt.[455] Ferner bestätigen Wortlaut, Systematik unter Abgrenzung des Text und Data Mining zu wissenschaftlichen Forschungszwecken mit einem Erkenntnisgewinn und Genese der entsprechenden Vorschriften ein weites Verständnis des Text und Data Mining. Diesem Verständnis steht zudem auch nicht der Dreistufentest i.S.d. Art. 5 Abs. 5 InfoSoc-RL entgegen, da es sich (1.) bei dem Text und Data Mining weiterhin um einen Sonderfall handelt und insbesondere mangels der Entwicklung eines Konkurrenzprodukts auch (2. und 3.) die normale Verwertung des Werks und die sonstigen berechtigten Interessen des Rechtsinhabers nicht ungebührlich verletzt werden.[456] Zutreffend wird insoweit darauf hingewiesen, dass aus ökonomischer Perspektive bloße Kopien zu verhindern sind, nicht aber – wie regelmäßig im Rahmen von KI-Ausgaben von Bedeutung – eine Substitution, die auf (abstrakten) Ideen und Gestaltungen bestehender Werke aufbaut.[457]

Vor allem aber stützen die jüngeren Art. 53 Abs. 1 lit. c, ErwGr. 105 S. 2 ff. KI-VO eine solche Auslegung.[458] Nach Art. 53 Abs. 1 lit. c KI-VO sind Anbieter von KI-Modellen mit allgemeinem Verwendungszweck ver-

450 Aus diesem Grund auch zweifelnd *Baumann*, NJW 2023, 3673 (Rn. 14).
451 *Dornis/Stober*, Urheberrecht und Training generativer KI-Modelle, S. 77, 87 ff.; *Dornis*, KIR 2024, 156 (158 ff.).
452 *Schack*, NJW 2024, 113 (Rn. 8).
453 *Welser*, GRUR 2024, 1406 (1413).
454 Etwa *Kleinkopf*, Text- und Data-Mining, S. 34.
455 *Geiger/Frosio/Bulayenko*, CEIPI Research Paper 2019 08 (19) m.w.N.; vgl. auch *Steinrötter/Schauer*, in: Barudi, § 4 Text und Data Mining, Forschung und Lehre, S. 146.
456 Ausf. *T. Radtke*, ZGE 17 (2025), 1 (34 ff.).
457 *Stieper/Denga*, GRUR 2024, 1473 (1475) m.w.N.
458 Überblick zum Meinungsstand bei *T. Radtke*, ZGE 17 (2025), 1 (36) in Fn. 177; etwa auch *Leistner*, GRUR 2024, 1665 (1669); *Hamann*, ZGE 16 (2024), 113 (122);

pflichtet, „eine Strategie zur Einhaltung des Urheberrechts der Union und damit zusammenhängender Rechte und insbesondere zur Ermittlung und Einhaltung eines gemäß *Artikel 4 Absatz 3 der Richtlinie (EU) 2019/790 geltend gemachten Rechtsvorbehalts*, auch durch modernste Technologien, auf den Weg [zu bringen]" (Hervorhebung d. d. Verf.). Hierdurch wird unterstellt, dass einem Vorbehalt nach Art. 4 Abs. 3 DSM-RL bzw. § 44b Abs. 3 UrhG entscheidende Bedeutung zukommt, also der Anwendungsbereich der Schrankenbestimmung auch und gerade die KI-Trainingsprozesse umfasst. Wäre hingegen das KI-Training als Ganzes nicht von der Schrankenbestimmung umfasst oder lediglich ein Teilaspekt des KI-Trainings als Text und Data Mining einzuordnen, hätte es gerade nicht der Herausstellung des betreffenden Vorbehalts bedurft.

Soweit *Dornis* der vorgenannten Vorschrift keinen Willen des Gesetzgebers entnehmen will, mit der Verabschiedung der KI-VO konkludent den Anwendungsbereich der DSM-RL zu erweitern, [459] folgt dies wiederum (insoweit konsequent) der unzutreffenden Prämisse eines zuvor enger gefassten Begriffs des Text und Data Mining in der DSM-RL. Nach zutreffender Lesart hat der europäische Gesetzgeber durch die KI-VO in Kenntnis der jüngsten KI-Entwicklungen allerdings bloß die Reichweite des Text und Data Mining aus der DSM-RL *bestätigt*.[460]

Dementsprechend lehnt auch das überwiegende Schrifttum die durch *Dornis* vorgebrachten Einwände ab und hält zurecht an einer weiten Auslegung des Begriffs des Text und Data Mining fest.[461] Das KI-Training ist daher grundsätzlich als Text und Data Mining i.S.d. §§ 44b, 60d UrhG einzustufen.

Peukert, GRUR Int. 2024, 497 (504); *Beurskens*, RDi 2025, 1 (Rn. 17); wohl auch *Buchalik/Gehrmann*, CR 2024, 145 (Rn. 59); anders vor Verabschiedung der KI-VO *Jager*, Artificial Creativity?, S. 336, 349 ff.

459 *Dornis/Stober*, Urheberrecht und Training generativer KI-Modelle, S. 95 ff.

460 *T. Radtke*, ZGE 17 (2025), 1 (36 f.).

461 *Stieper/Denga*, GRUR 2024, 1473 (1474 ff.); *Leistner*, GRUR 2024, 1665 (1669 f.); *Käde*, KIR 2024, 162 (168); *Dreier*, in: Dreier/Schulze, § 44b Rn. 5; wohl auch *Antoine*, GRUR 2025, 118 (122) bei Fn. 59; *LG Hamburg*, GRUR 2024, 1710 (Rn. 49); *Pesch*, LTZ 2025, 72 (76): „flawed argument"; nicht Position beziehend, aber zumindest eine große Bedeutung des TDM für das Training generativer KI annehmend *Schmid*, KIR 2025, 69; offengelassen in der KI-VO nach *Welser*, GRUR 2024, 1406 (1415); *Geiger*, in: Thouvenin/Peukert/Jaeger/Geiger, When the Robots (Try to) Take Over: Of Artificial Intelligence, Authors, Creativity and Copyright Protection, S. 77.

bb. KI-Einsatz und Ausgabe als Text und Data Mining

Sofern für den konkreten Einsatz eines KI-Systems durch eine Eingabeaufforderung innerhalb des Rechenprozesses des Sprachmodells allfällige Vervielfältigungen angenommen werden,[462] sind diese als Teil des Analyseprozesses anzusehen. Denn derartige Vervielfältigungen stellen die Grundlage dar für den späteren Informationsgewinn aus trainiertem Sprachmodell und Eingabe.

Vervielfältigungen im Rahmen der durch das KI-System generierten Ausgabe erfolgen allerdings regelmäßig nicht (mehr), „um daraus [mittels automatisierter Analyse] Informationen insbesondere über Muster, Trends und Korrelationen zu gewinnen" (§ 44b Abs. 1 UrhG), sprich Text und Mining vorzunehmen. Vielmehr wurden die in Rede stehenden Informationen bereits analysiert und dürften nunmehr dem (manuellen) Werkgenuss dienen. Dieses Ergebnis bestätigt auch der Dreistufentest auf den Stufen zwei und drei, wonach bei der wörtlichen Übernahme von Schriftsätzen und Gutachten die normale Verwertung des Werks gefährdet und die berechtigten Interessen des Rechtsinhabers verletzt sein könnten.

An wörtlichen, eng an den Trainingsdaten orientierten Ausgaben dürfte im untersuchungsgegenständlichen GSJ-Projekt regelmäßig allerdings ohnehin kein Interesse bestehen. Denn es wird eine spezifische Befassung mit neuen Anliegen und Aktenauszügen gewollt sein (z.B. im Rahmen der Sachverhaltsaufbereitung und -zusammenfassung in den Use-Cases 1-3), nicht dagegen eine Ausgabe eng am Stil und Wortschatz eines Anwalts oder Gutachters. Entsprechende Ausgaben sind nach den Ausführungen zum Datenschutzrecht[463] hiernach durch angemessene technische und organisatorische Maßnahmen zu verhindern bzw. zu minimieren. Sollten ausnahmsweise durch einzelne Anwender gezielt rechtsverletzende Ausgaben durch Eingabeaufforderungen generiert werden (z.B. durch sog. Prompt Injections), liegt zudem mit Blick auf diese Vervielfältigung und eine hieraus resultierende Urheberrechtsverletzung eine Passivlegitimation primär des Anwenders anstelle des GSJ-Projektbetreibers nahe.[464]

462 S. aber oben unter E.II.1.a.bb

463 S. oben unter C.VI.4.

464 *T. Radtke*, ZGE 17 (2025), 1 (12); s. auch *Determann/Paal*, KI-Recht international, S. 139.

cc. Anwendbarkeit des § 44b Abs. 2 S. 1 UrhG

Die Anwendbarkeit des § 44b Abs. 2 S. 1 UrhG ist sowohl aus intertemporaler Perspektive als auch mit Blick auf einen möglicherweise erklärten Vorbehalt i.S.d. § 44b Abs. 3 UrhG zu untersuchen.

(1) Intertemporale Anwendbarkeit

Nach Art. 26 Abs. 1, 2 DSM-RL findet das dort niedergelegte unionsrechtliche Vorbild für § 44b UrhG auch Anwendung auf Werke, die (auch) ab dem 7. Juni 2021 geschützt sind, nicht aber auf „Handlungen und Rechte, die vor dem 7. Juni 2021 abgeschlossen bzw. erworben wurden". Das deutsche Umsetzungsgesetz ist am gleichen Tag in Kraft getreten,[465] greift die Regelung des Art. 26 Abs. 2 DSM-RL zwar nicht auf, ist aber unter Berücksichtigung der unionsrechtlichen Vorschrift auszulegen. Danach sind grundsätzlich alle geschützten Werke und Schutzgegenstände umfasst.[466] Lediglich KI-Modelle, die auf Vervielfältigungen vor dem 7. Juni 2021 basieren, können unter Beachtung von Art. 26 Abs. 2 DSM-RL bei Auslegung des § 44b UrhG nicht auf die Schrankenbestimmung gestützt werden.[467]

Indem § 44b Abs. 2 S. 1 UrhG die rechtmäßige Zugänglichkeit von Werken voraussetzt, könnte die Anwendbarkeit auf schon vor dem 7. Juni 2021 rechtmäßig zugängliche Werke als erworbenes Recht infragestehen. Allerdings kann ein derart weitreichender Anwendungsbereich dem Art. 26 Abs. 2 DSM-RL mit Blick auf den Zusammenhang zu Art. 26 Abs. 1 DSM-RL nicht entnommen werden. Vielmehr bleiben gerade zuvor erworbene Rechte „unberührt" und werden nicht durch Art. 26 Abs. 2 DSM-RL eingeschränkt.

(2) Vorbehalt der Vervielfältigungen zu Zwecken des Text und Data Mining

Rechteinhaber können sich nach § 44b Abs. 3 S. 1 UrhG ihre Rechte an Vervielfältigungen zu Zwecken des Text und Data Mining vorbehalten.

465 Vgl. Art. 5 des Gesetzes zur Anpassung des Urheberrechts an die Erfordernisse des Digitalen Binnenmarkts vom 31. Mai 2021, Bundesgesetzblatt 2021 Teil I Nr. 27, 1204 (4. Juni 2021).

466 Vgl. *Beurskens*, RDi 2025, 1 (Rn. 16).

467 *Schack*, NJW 2024, 113 (Rn. 7); *Welser*, GRUR-Prax 2023, 516 (Rn. 41); *Dregelies*, GRUR 2024, 1484 (1487).

(a) Form des Vorbehalts

Die Anforderung der Maschinenlesbarkeit[468] des Vorbehalts sieht § 44b Abs. 3 S. 2 UrhG nur für über das Internet zugängliche Werke vor.[469] Für die hier untersuchungsgegenständlichen, gerade nicht über das Internet frei zugänglichen Aktenauszüge schließen daher auch andere Formen eines ausdrücklichen[470] Vorbehalts die Nutzung ex nunc[471] aus. Ein solcher Vorbehalt dürfte allerdings in Ansehung der für das Text und Data Mining in der Justiz bisher typischerweise kaum sensibilisierten Adressaten allenfalls (sehr) selten[472] vorkommen. In einem Schriftsatz oder Gutachten wären daher typischerweise Vorbehalte in natürlicher Sprache zu berücksichtigen, die dem Dokument aufgrund einer entsprechenden Formulierung, Positionierung und gegebenenfalls Hervorhebung „ohne allzu große Anstrengungen zu entnehmen sind".[473] Die Voraussetzung der „Ausdrücklichkeit" schließt konkludente Vorbehalte aus sowie verlangt aber zugleich einen konkreten Bezug zu Text und Data Mining.[474] Die Darlegungs- und Beweislast für einen fehlenden Vorbehalt soll nach der gesetzgeberischen Intention der Nutzende tragen.[475]

Konkrete Vorgaben zur Formulierung des Vorbehalts macht der Gesetzgeber nicht, sodass vielgestaltige Formulierungen denkbar sind, wie etwa: „Kein Text oder Data Mining",[476] „Kein TDM" oder „Kein KI-Training". Grenzfälle dürften allgemeine Vorbehalte darstellen, wie etwa „Alle Rechte vorbehalten",[477] „Vervielfältigung vorbehalten" oder die in Gutachten übliche[478] Untersagung der Weitergabe des Gutachtens. Ein bloßer Urheberrechtshinweis (z.B. „© Autor, 2025") ist weder im Hinblick auf das Text und Data Mining noch einen Vorbehalt hinreichend ausdrücklich,

468 Hierzu etwa *LG Hamburg*, GRUR 2024, 1710 (1713 f.); *Hamann*, ZGE 16 (2024), 113 (128 ff.).
469 S. auch ErwGr. 18 UAbs. 2 DSM-RL.
470 S. Art. 4 Abs. 3 DSM-RL: „ausdrücklich in angemessener Weise".
471 *Paul*, in: Hilber/Borges, § 44b UrhG Rn. 5; *Hamann*, ZGE 16 (2024), 113 (124) m.w.N.
472 S. allgemein *Bomhard*, DSRITB 2023, 255 (265).
473 *Bomhard*, in: Götting/Lauber-Rönsberg/Rauer, § 44b UrhG Rn. 25; *Bomhard*, InTeR 2023, 174 (178).
474 *Hamann*, ZGE 16 (2024), 113 (134 f.); vgl. BT-Drs. 19/27426, 89.
475 BT-Drs. 19/27426, 89.
476 *Paul*, in: Hilber/Borges, § 44b UrhG Rn. 5.
477 *Paul*, in: Hilber/Borges, § 44b UrhG Rn. 5; wohl aber nicht nach *Hamann*, ZGE 16 (2024), 113 (134 f.), der einen konkret-individuellen Bezug verlangt.
478 S. *T. Heinrich*, NZV 2015, 68.

sondern kann vielmehr auch als bloße Nennung des Rechtsinhabers bzw. gegebenenfalls auch des Urhebers i.S.d. § 13 UrhG verstanden werden.[479] Die Vervielfältigung für eine automatisierte Analyse von Dokumenten,[480] um einen Vorbehalt zu identifizieren, ist zwar ebenfalls Text und Data Mining, aber in systematischer Hinsicht mit Blick auf § 44b Abs. 3 UrhG und gegebenenfalls auch als vorübergehende Vervielfältigung nach § 44a UrhG[481] zulässig.

(b) Rechtsinhaber

Die Rechte vorbehalten kann sich nur der „jeweilige" Rechtsinhaber i.S.d. Art. 3 Abs. 3 DSM-RL, der nicht mit dem Urheber identisch sein muss. Für die Frage der Rechtsinhaberschaft kann es daher darauf ankommen, wem Nutzungsrechte i.S.d. §§ 31 ff. UrhG eingeräumt wurden.[482] Unter Berücksichtigung der Zweckübertragungslehre sind die dem Vorbehalt zugrunde liegenden Rechte mit Blick auf Text und Data Mining für Rechteeinräumungen in der Vergangenheit nach einer Stimme in der Literatur nicht ohne Weiteres erfasst, sondern sollen als neue Nutzungsart zu behandeln sein (insbesondere § 31a UrhG).[483]

Sofern mit Blick auf Schriftsätze ein Rechteübergang auf den Mandanten angenommen wird,[484] umfasste dieser Rechteübergang jedenfalls bis vor wenigen Jahren nicht die dem Vorbehalt zugrunde liegenden Rechte. Ein etwaiger Vorbehalt (konnte) also regelmäßig durch den jeweiligen Rechtsanwalt oder Gutachter als Rechtsinhaber erklärt werden. Ein ähnliches Ergebnis ergibt sich, wenn man die Übertragung von Nutzungsrechten verlangt und gegebenenfalls die Wertungen aus §§ 177 ff. BGB für eine etwaige (rückwirkende) Genehmigung eines nicht durch den jeweiligen Rechtsinhaber erklärten Vorbehalts heranzieht.[485]

Für Aktenauszüge aus jüngerer Vergangenheit, insbesondere solche nach dem 7. Juni 2021, kann auch ein Übergang der zum Vorbehalt berechtigten

479 Hierzu etwa *Mantz*, in: Dreier/Schulze, § 10 UrhG Rn. 13.
480 S. etwa *LG Hamburg*, GRUR 2024, 1710 (Rn. 67 f.).
481 S. nachfolgend unter E.II.2.d
482 *Leistner*, GRUR 2024, 1665 (1671); *Vesala*, IIC 2023, 351 (357); *Brockmeyer*, Text und Data Mining, S. 79.
483 *Hamann*, ZGE 16 (2024), 113 (138 f.); a.A. *Leistner*, GRUR 2024, 1665 (1672): allgemeine Rechtsgeschäftslehre.
484 S. unter E.II.2.a.
485 Vgl. *Leistner*, GRUR 2024, 1665 (1671 f.).

Nutzungsrechte vorliegen. In der Folge könnte nur der neue Rechtsinhaber (z.B. der Mandant) den Vorbehalt erklären.

Der hieraus resultierenden erhöhten Komplexität kann praktisch etwa dadurch begegnet werden, dass die Wirksamkeit sämtlicher in den Aktenauszügen erklärten Vorbehalte unterstellt wird und entsprechende Dokumente gegebenenfalls von dem Trainingsprozess ausgenommen werden.

(c) Rechtmäßig zugängliche Werke

Die Schrankenbestimmung ermöglicht die Vervielfältigungen von „rechtmäßig zugänglichen Werken". Unter Berücksichtigung von ErwGr. 14 S. 1, 2 DSM-RL[486] ist auf die Rechtmäßigkeit des Zugangs, nicht jedoch darauf abzustellen, „ob das Werk mit oder ohne Zustimmung des Rechteinhabers zugänglich gemacht wurde".[487] Es schadet der Verwirklichung des Tatbestandsmerkmals also nur, wenn Schutzmaßnahmen umgangen wurden.[488] Für die Justiz wird im GSJ-Projekt regelmäßig ein solcher rechtmäßiger Zugang gegeben sein.[489]

Entsprechende Überlegungen für weitere Akteure, denen die Aktenauszüge durch die Justiz übermittelt werden, wobei gegebenenfalls weitere Anforderungen zu beachten sind.[490]

(d) Weitere Anforderungen

Die weiteren Anforderungen an einen wirksamen Vorbehalt betreffen einerseits einen möglichen Ausschluss abstrakt-genereller Vorbehalte aufgrund des Ausdrücklichkeitserfordernisses,[491] andererseits (zulässige) nach-

---

486 In ErwGr. 14 S. 1, 2 DSM-RL wird ein „rechtmäßig[er] Zugang" verlangt.

487 So zutreffend *Raue*, CR 2017, 656 (658); *Specht-Riemenschneider*, OdW 2018, 285 (286); *Döhl*, RuZ 2020, 195 (217); anders hingegen *Schwartmann/Köhler*, in: Schwartmann/Keber/Zenner, 3. Kapitel, Rn. 118.

488 *Dornis/Stober*, Urheberrecht und Training generativer KI-Modelle, S. 98 f.; anschaulich *Talke*, Urheberrecht in Bildung, Wissenschaft und Kulturerbe, S. 30 f.; *K. Wagner*, MMR 2024, 298 (299).

489 Insb. auch mit Blick darauf, dass keine Zugangsbeschränkungen umgangen werden, s. *Paul*, in: Hilber/Borges, § 44b UrhG Rn. 3.

490 S. hierzu unter D.II.3.

491 So etwa *Hamann*, ZGE 16 (2024), 113 (148 f.)

träglich[492] erklärte Vorbehalte. Nur in Ausnahmefällen, so insbesondere wenn ein Rechtsinhaber nachträglich gegenüber den jeweiligen Gerichten[493] einen Vorbehalt erklärt, gegebenenfalls pauschal bezogen auf seine sämtlichen Schriftsätze, werden sich diese Fragen stellen.[494] Die Frage nach einer *ex tunc*- oder *ex nunc*-Wirkung[495] ist ebenfalls zu vernachlässigen, da selbst eine *ex nunc*-Wirkung faktisch zurückwirken kann, wenn das einmal vervielfältigte Werk nach Erklärung des Vorbehalts für ein weiteres Training erneut vervielfältigt wird.

Die Möglichkeit eines vor Inkrafttreten des § 44b UrhG erklärten Vorbehalts[496] dürfte für das untersuchungsgegenständliche Projekt in Bezug auf in den Aktenauszügen enthaltene Schriftsätze und Gutachten gleichfalls kaum von Relevanz sein.

(3) Zwischenergebnis

Die Regelung des § 44b Abs. 2 S. 1 UrhG kann belastbar das Training des Sprachmodells ermöglichen, setzt allerdings die Filterung von Werken voraus, für die ein Nutzungsvorbehalt erklärt wurde. Da tendenziell solche Erklärungen nur äußerst selten vorkommen dürften, besteht ein sehr geringes Restrisiko für nicht identifizierte und sodann übergegangene Vorbehalte. Selbst in diesen Fällen kann zudem gegebenenfalls eine andere Schrankenbestimmung eingreifen, was nachfolgend näher zu prüfen ist.

Die Regelung des § 44b UrhG dürfte in der Praxis zur Folge haben, dass ab Bekanntwerden entsprechender KI-Projekte in der Anwaltschaft pauschal entsprechende Vorbehalte, etwa in Kanzlei-Briefköpfen erklärt werden.[497] Die hieraus folgende Begünstigung des GSJ-Projekts durch „Ge-

492 Nachträglich in diesem Sinne meint, einen Vorbehalt separat nach Einreichung eines Schriftsatzes oder Gutachtens zu erklären, der ex nunc zukünftiges Text und Data Mining verhindern könnte.

493 Eine allgemeine Bekanntmachung über die Website des Rechtsinhabers genügt nicht den Anforderungen an die einfache Auffindbarkeit des ausdrücklichen Vorbehalts.

494 Eine etwaige Genehmigung eines durch eine nichtberechtigte Person erklärten Vorbehalts kann nach § 184 BGB zurückwirken *Leistner*, GRUR 2024, 1665 (1671 f.) unter Verweis auf die LAION-Entscheidung des LG Hamburg.

495 *Brockmeyer*, Text und Data Mining, S. 79 f.

496 S. die Konstellation in *LG Hamburg*, GRUR 2024, 1710.

497 Gegebenenfalls besteht insoweit aber auch kein bzw. ein nur eingeschränkter Markt und kein Interesse an der Erklärung eines Vorbehalts, vgl. *Raue*, ZUM 2020, 172 (173).

heimhaltung“ hat keine unmittelbaren urheberrechtlichen Implikationen (siehe aber zum Datenschutzrecht unter D.V.6.a), denn der Vorbehalt nach § 44b Abs. 3 UrhG hängt nicht von einem konkreten Projekt ab, sondern konnte bereits vorher bezogen auf beliebige Werke durch die Anwaltschaft und Gutachter erklärt werden. Die Frage des Umgangs mit dem jeweiligen Projekt dürfte daher vor allem eine Frage der Öffentlichkeitsarbeit und der öffentlich-rechtlichen Anforderungen an die Auftraggeber sein.

Vorliegend nicht im Detail zu erörtern sind ferner Fragen nach einem allfälligen Umgestaltungsbedarf des besonderen elektronischen Anwaltspostfachs (beA) zur Möglichkeit der einfachen Erklärung eines Vorbehalts.[498]

dd. Anwendbarkeit des § 60d UrhG

§ 60d UrhG erlaubt Text und Data Mining unabhängig von einem erklärten Vorbehalt für wissenschaftliche Forschungszwecke durch Forschungsorganisationen. Unter den Begriff der „Forschungsorganisationen“ fallen insbesondere Hochschulen nach § 60d Abs. 2 S. 2 UrhG.

Allerdings genügt es entgegen einem insoweit kritisch zu betrachtenden Urteil des LG Hamburg nicht, dass der durch Text und Data Mining generierte Datensatz später einmal für wissenschaftliche Forschungszwecke verwendet werden *kann*.[499] Mit Blick auf die Abgrenzung von Text und Data Mining einerseits und wissenschaftlichen Forschungszwecken andererseits bedarf es gerade des Bezugs zu einem konkreten Forschungsprojekt.[500] Für anwendungsbezogene Forschung, wie im GSJ-Projekt, ist darüber hinaus die Frage aufzuwerfen, ob es genügt, wenn die wissenschaftlichen Forschungszwecke von vornherein ein bloßes Durchgangsstadium für die spätere Umsetzung in der Praxis sind. Ein derartiges Verständnis würde dazu führen, dass letztlich jeder kommerzielle Einsatz gerechtfertigt wäre, wenn und soweit sich Elemente aus dem Entwicklungsstadium auch als Forschung einer Forschungsorganisation einordnen ließen. Für die Anwendung der Schrankenbestimmung aus § 44b UrhG bliebe somit kaum noch Raum.

498 In ErwGr. 18 UAbs. 2 S. 5 DSM-RL heißt es nämlich: „Die Rechteinhaber sollten in der Lage sein, Maßnahmen zu treffen, mit denen sie sicherstellen, dass ihre diesbezüglichen Vorbehalte Beachtung finden.“

499 *LG Hamburg*, GRUR 2024, 1710 (Rn. 76 f.); begrüßenswert hingegen nach *Leistner*, GRUR 2024, 1665 (1667).

500 Hierzu ausf. *T. Radtke*, ZGE 17 (2025), 1 (44 ff.).

ErwGr. 11 S. 2 DSM-RL lässt sich ein Hinweis darauf entnehmen, dass die Zusammenarbeit mit der Privatwirtschaft der persönlichen Anwendbarkeit der Vorschrift nicht schaden muss.[501] Forschungsergebnisse aus öffentlich-privaten Partnerschaften werden typischerweise in der Praxis umgesetzt. Außerdem verlangen § 60d UrhG, Art. 3 DSM-RL nach ihrem Wortlaut im Gegensatz zu anderen Schrankenbestimmungen (wie etwa § 44a UrhG, Art. 5 Abs. 1 InfoSoc-RL („alleiniger Zweck")) keine Exklusivität der Verwendung zu wissenschaftlichen Forschungszwecken. Zugleich verfolgte Sekundärzwecke stehen daher der Anwendung des § 60d UrhG grundsätzlich nicht entgegen.

Gleichwohl bestehen mit Blick auf das Konkurrenzverhältnis zum allgemeinen § 44b UrhG gegebenenfalls Bedenken, wenn das Forschungsprojekt von *vornherein* nicht primär auf den wissenschaftlichen Erkenntnisgewinn ausgerichtet ist, sondern auf den Einsatz des trainierten Sprachmodells in der (Justiz-)Praxis zur Sachverhaltsaufbereitungen in konkreten Verfahren (siehe Use-Cases 1-3) und durch Forschung lediglich begleitet werden soll.[502] Denn in dieser Konstellation sind die Vervielfältigungen unmittelbare Ausgangsbasis für das in der Praxis eingesetzte Produkt und nicht ein bloßes Forschungsobjekt, auf dessen Grundlage sodann ein separates Produkt für den operativen Einsatz entwickelt wird. Es ist somit festzuhalten: Nur weil eine Tätigkeit im Rahmen der Produktentwicklung auch wissenschaftliche Erkenntnisse zu liefern vermag, kann die produktbezogene Tätigkeit nicht ohne Weiteres pauschal und einschließlich der hierzu notwendigen Verwertungshandlungen den Schranken für wissenschaftliche Forschungszwecke zugerechnet werden.

Diese Ablehnung wissenschaftlicher Forschungszwecke für das GSJ-Projekt wird zudem bestätigt durch die Inbezugnahme des Dreistufentests (vgl. insoweit unter anderem Art. 5 Abs. 5 InfoSoc-RL): Eine weite Auslegung der wissenschaftlichen Forschungszwecke, wonach begleitende Forschung durch Text und Data Mining genügt bei einem von vornherein gerade auf

501 „Im Einklang mit der derzeitigen Forschungspolitik der Union, die Hochschulen und Forschungsinstitute zur Zusammenarbeit mit der Privatwirtschaft anhält, sollten auch Forschungsorganisationen eine solche Ausnahme nutzen dürfen, sofern ihre Forschungstätigkeit im Rahmen öffentlich-privater Partnerschaften durchgeführt wird".

502 Vgl. auch *T. Radtke*, ZGE 17 (2025), 1 (21, 44 ff.); sowie schon den deutschen Gesetzgeber in BT-Drs. 18/12329, 39: „Forschung, die ein Unternehmen betreibt, um Waren oder Dienstleistungen zu entwickeln [sic!] und diese dann zu vermarkten, dient allerdings kommerziellen Zwecken".

den operativen Einsatz ausgelegten und entsprechend beauftragten Endprodukt (i.e. das KI-Modell), würde sich gerade nicht mehr auf bestimmte Sonderfälle[503] im Sinne der ersten Stufe des Dreistufentests beschränken. Eine solche Auslegung würde vielmehr die umfassende Anfertigung von Vervielfältigungen als Ausgangspunkt für die weitere Verwertung ermöglichen. Die Vervielfältigung i.S.d. § 60d UrhG würde hiernach im Ergebnis zu einem Regelfall. Diese weite Auslegung ist daher auch mit Blick auf die dritte Stufe des Dreistufentests bedenklich, sprich die ungebührliche Verletzung der berechtigten Interessen des Rechtsinhabers unter Berücksichtigung der Vergütungsfreiheit des § 60d UrhG[504] und des Nutzungsumfangs.[505]

Eine andere Bewertung mag zwar mit entsprechender Argumentation noch vertretbar sein, geht aber jedenfalls mit einem erheblichen Risiko von Urheberrechtsverstößen einher. Nach der vorzugswürdigen Ansicht ist eine solche andere Bewertung allerdings gerechtfertigt, wenn und soweit Text und Data Mining auf ein (Teil-)Produkt gerichtet ist, das gerade nicht von vornherein nur dem operativen Einsatz dient (z.B. ein eigenständiges, zunächst trainiertes KI-Modell zur Klärung von Forschungsfragen für die spätere Entwicklung eines KI-Modells für den Produktiveinsatz). Wenn und soweit (privilegierte) wissenschaftliche Forschungszwecke in solchen Konstellationen belastbar angenommen werden können, werden sich zugleich strengere Anonymisierungsanforderungen aus dem Datenschutzrecht ergeben, sprich aus Art. 89 Abs. 1 DSGVO i.V.m. Art. 25 Abs. 2 BayDSG, § 17 Abs. 3 DSG NRW.

d. Vorübergehende Vervielfältigungshandlungen (§ 44a UrhG)

Nach § 44a Nr. 2 UrhG sind „vorübergehende Vervielfältigungshandlungen [zulässig], die flüchtig oder begleitend sind und einen integralen und wesentlichen Teil eines technischen Verfahrens darstellen und deren alleiniger Zweck es ist, eine rechtmäßige Nutzung eines Werkes oder sonstigen Schutzgegenstands zu ermöglichen, und die keine eigenständige wirtschaft-

503 Vgl. zur engen Auslegung der Ausnahmen mit Blick auf den Dreistufentest *EuGH*, GRUR 2017, 610 (Rn. 63) – Stichting Brein/Jack Frederik Wullems; GRUR 2009, 1041 (Rn. 58) – Infopaq.

504 *Ott*, ZUM 2009, 345 (353) in Fn. 77; *Senftleben*, GRUR Int. 2004, 200 (211); vgl. auch *Hofmann*, ZUM 2024, 166 (172).

505 Vgl. *EuGH*, EuZW 2014, 868 (Rn. 56) – TU Darmstadt/Eugen Ulmer KG.

liche Bedeutung haben". Die rechtmäßige Nutzung umfasst sowohl den Werkgenuss als auch eine von den Schrankenbestimmungen gedeckte Nutzung.[506] Die Regelung aus § 44a Nr. 2 UrhG adressiert insbesondere das sog. Caching.[507]

Daher kann die Vorschrift für temporäre Zwischenspeicherungen im Zusammenhang mit dem Training eines Sprachmodells Bedeutung erlangen, z.B. wenn temporäre (digitalisierte) Kopien auf den Trainingsservern angelegt werden, die der Ausgangspunkt für abgeleitete Textformate sind. Diese Kopie muss als integraler Teil eines technischen Verfahrens automatisiert nach dessen Abschluss gelöscht werden,[508] wobei eine entsprechende Programmierung der Software genügt.[509] Für die wirtschaftliche Bedeutung ist unter Berücksichtigung des Schutzzwecks des Urheberrechts darauf abzustellen, ob spezifisch die urheberrechtlich geschützten Bestandteile eine wirtschaftliche Bedeutung haben.[510] Eine entsprechende wirtschaftliche Bedeutung ist also nicht bereits deshalb anzunehmen, weil auf Grundlage der vorübergehenden Vervielfältigung eine Verwertung von nicht urheberrechtlich geschützten Bestandteilen möglich ist (wie etwa in Form eines Sprachmodells, das nicht einzelne Werke vervielfältigt und Urheber imitiert).[511] Im untersuchungsgegenständlichen Projekt können also gestützt auf § 44a UrhG digitale Aktenauszüge auf einem Server abgelegt werden, wenn diese automatisiert als Übergangsstadium zur Übertragung in ein abgeleitetes Textformat oder sonstigen Aufbereitung vor dem Training angelegt und unmittelbar nach der Übertragung automatisiert gelöscht werden.

Grundsätzlich können auch vorübergehende Vervielfältigungen zu Zwecken des Text und Data Mining oder flüchtige Vervielfältigungen im Rah-

506 *Schulz*, in: Götting/Lauber-Rönsberg/Rauer, § 44a UrhG Rn. 13; *Dreier*, in: Dreier/Schulze, § 44a UrhG Rn. 8.

507 ErwGr. 33 S. 3 InfoSoc-RL.

508 *EuGH*, GRUR 2009, 1041 (Rn. 64) – Infopaq.

509 Zu eng daher *LG Hamburg*, GRUR 2024, 1710 (Rn. 35); wie hier auch *Leistner*, GRUR 2024, 1665 (1666).

510 *EuGH*, GRUR 2009, 1041 (Rn. 50) – Infopaq; MMR 2011, 817 (Rn. 175) – Football Association Premier League Ltd. u.a.

511 Vgl. *T. Radtke*, ZGE 17 (2025), 1 (25 f.); ebenfalls ablehnend, da nicht die Werke, sondern deren Datenpunkte Gegenstand des Trainings seien *La Durantaye*, AfP 2024, 9 (Rn. 22); wegen Änderung des Vervielfältigungsstücks *Hofmann*, ZUM 2024, 166 (169); von einer solchen Bedeutung aber regelmäßig im Rahmen des KI-Trainings ausgehend *Wissenschaftliche Dienste des Bundestags*, Künstliche Intelligenz und Machine Learning - Eine urheberrechtliche Betrachtung, S. 12; *Pesch/Böhme*, GRUR 2023, 997 (1006). S. auch ErwGr. 9 DSM-RL.

men eines Trainingsprozesses[512] auf § 44a Nr. 2 UrhG gestützt werden. Der europäische Gesetzgeber erkennt in ErwGr. 9 S. 2 DSM-RL ausdrücklich an, dass es auch „Fälle des Text und Data Mining geben [kann], in denen […] die Vervielfältigungen unter die in Artikel 5 Absatz 1 der Richtlinie 2001/29/EG vorgesehene verbindliche Ausnahme für vorübergehende Vervielfältigungshandlungen fallen, die auch künftig auf Verfahren des Text und Data Mining angewandt werden sollte […]". Allerdings dürfte die Aufbewahrung des Trainingskorpus zur weiteren Verbesserung des Sprachmodells und gegebenenfalls des Trainings einer weiteren Version gewünscht sein, sodass es eines Rekurses auf die speziellen Text und Data Mining-Ausnahmen in §§ 44b, 60d UrhG bedarf.[513]

e. Zitate (§ 51 UrhG)

§ 51 UrhG erlaubt Vervielfältigungen zum Zweck des Zitats. Vereinzelte KI-Ausgaben mit geschützten Werkbestandteilen könnten grundsätzlich der geistigen Auseinandersetzung[514] mit dem jeweiligen Werk dienen und daher unter den § 51 UrhG fallen. Allerdings setzt die Vorschrift, wie sich aus einer Gesamtschau der § 51 S. 2 Nr. 1-3 UrhG ergibt, grundsätzlich die Auseinandersetzung im Rahmen eines Werks eines menschlichen Urhebers voraus und dürfte daher nicht die Aufnahme in einen KI-generierten Text umfassen. Ferner wird in den Use-Cases des GSJ-Projekts keine abgrenzende Auseinandersetzung mit dem Werk, sondern eine bloße Vervielfältigung eines zudem noch nicht veröffentlichten Werks erfolgen.[515] Die Vervielfältigung dürfte insbesondere auch nicht die nach § 63 Abs. 1 S. 1 UrhG erforderliche Quellenangabe enthalten. Die Zitatschranke kommt daher grundsätzlich nicht zur Rechtfertigung etwaiger Vervielfältigungen in Betracht.

f. Wissenschaftliche Forschung (§ 60c UrhG)

Die Vorschrift des § 60c UrhG stellt im untersuchungsgegenständlichen Projekt keine geeignete Schranke dar. § 60c UrhG umfasst beispielsweise

512 *Jager*, Artificial Creativity?, S. 348; vgl. auch *Stieper/Denga*, GRUR 2024, 1473 (1475).
513 S. auch *Chiou*, JIPITEC 2019, 398 (405 ff.).
514 *BGH*, NJW 1968, 1875 (1876 f.).
515 Zu den Anforderungen *Peifer*, ZUM 2020, 342.

„die Unterrichtung über den Stand der Wissenschaft",[516] ist nach § 60h UrhG vergütungspflichtig und erfordert nach § 63 Abs. 1 S. 1 UrhG eine Quellenangabe.

Selbst wenn man im GSJ-Projekt die Verfolgung wissenschaftlicher Forschungszwecke unterstellt,[517] ist eine quantitative Beschränkung vorgesehen von jeweils 15 Prozent (§ 60c Abs. 1 UrhG) bzw. 75 Prozent eines Werkes (§ 60c Abs. 2 UrhG). Grundsätzlich wäre für das GSJ-Projekt der geringere Wert von 15 Prozent aus § 60c Abs. 1 UrhG zugrunde zu legen, da § 60c Abs. 2 UrhG nicht die Weitergabe über die eigene Einrichtung hinaus umfasst;[518] solche Weitergaben über die eigene Einrichtung hinaus dürften im GSJ-Projekt bereits durch die Zusammenarbeit zwischen den ausführenden Stellen und den Ministerien erfolgen. In Ansehung der speziell(er)en Regelungen der §§ 44b, 60d UrhG zu Vervielfältigungen für Zwecke des Text und Data Mining ist zudem zweifelhaft, ob § 60c UrhG in diesen Konstellationen überhaupt Anwendung finden kann.

Die Verwertung von bloß 15 Prozent der Werke bei verbleibender Rechtsunsicherheit mit Blick auf die spezielleren §§ 44b, 60d UrhG und die Verfolgung wissenschaftlicher Forschungszwecke ermöglicht nicht belastbar die Verwertung im Rahmen des GSJ-Projekts.

3. Einbindung weiterer Stellen

Im GSJ-Projekt wirken die Ministerien und die ausführenden Stellen zusammen, was die Frage nach dem berechtigten Personenkreis mit Blick auf die jeweiligen Schranken aufwirft. Im Ausgangspunkt bedarf diejenige Person einer Rechtfertigung, die selbst eine Nutzungshandlung begeht bzw. an dieser mitwirkt und ohne eine einschlägige Schrankenbestimmung als Täter, Teilnehmer, Störer oder Hilfsdienst[519] für eine Urheberrechtsverletzung passiv legitimiert wäre.[520]

---

516 *Grübler*, in: Götting/Lauber-Rönsberg/Rauer, § 60c UrhG Rn. 1.

517 S. aber unter E.II.2.c.dd.

518 *Grübler*, in: Götting/Lauber-Rönsberg/Rauer, § 60c UrhG Rn. 13; *Stieper*, in: Schricker/Loewenheim, § 60c UrhG Rn. 19; BT-Drs. 18/12329, 40: „Die so hergestellten Kopien dürfen in keiner Form weitergegeben werden".

519 *Reber*, in: Loewenheim, § 86 Ansprüche aus Verletzung des Urheber- oder Leistungsschutzrechts, Rn. 52-142.

520 Vgl. *Wolff/Wandtke*, in: Wandtke/Bullinger, § 97 UrhG Rn. 38.

Die urheberrechtlichen Schranken sind grundsätzlich neutral mit Blick auf die berechtigten Personen und ermöglichen demzufolge die Verwendung von Hilfsmitteln sowie die Einbindung Dritter im Auftrag eines Berechtigten,[521] verlangen allerdings beispielsweise in § 45 UrhG einen engen Zusammenhang mit einem Gerichtsverfahren, in §§ 44b, 60d UrhG den rechtmäßigen Zugang, wobei § 60d UrhG zusätzliche Anforderungen an die berechtige Person stellt (i.e. das Erfordernis einer Forschungsorganisation). Dieser rechtmäßige Zugang unterliegt allerdings keinen hohen Anforderungen,[522] sondern ist bereits gewahrt, wenn die Werke den ausführenden Stellen durch die Ministerien zugänglich gemacht werden.

Der Rekurs auf § 45 Abs. 1 UrhG scheidet für das untersuchungsgegenständliche Projekt aus, da sich das Projekt – anders als der Einsatz des entwickelten Tools in den einzelnen Use-Cases – nicht auf ein konkretes gerichtliches Verfahren bezieht.[523] Auf §§ 44a, 44b UrhG wirkt sich die Einbindung der ausführenden Stellen zur Durchführung des Sprachmodell-Trainings für die Ministerien bzw. die Justiz nicht aus. Denn trotz Auslagerung bleibt es bei der gleichen Nutzungshandlung, die auch bei eigenständiger Ausführung durch die ausführenden Stellen gerechtfertigt wäre.

§ 60d UrhG scheidet im GSJ-Projekt aus und bedarf daher keiner weiteren Prüfung mit Blick auf die Beschränkung auf Forschungsorganisationen nach § 60d Abs. 2 UrhG.

4. Ableitung von Einsatzbedingungen

Aus dem Vorherigen lassen sich generelle und wesentliche Einsatzbedingungen für das untersuchungsgegenständliche Projekt ableiten.

Das eigentliche Training eines Sprachmodells ist urheberrechtlich grundsätzlich neutral. Allerdings bedürfen Vervielfältigungen und gegebenenfalls Bearbeitungen vor, während oder nach dem Trainingsprozess eines eingeräumten Nutzungsrechts oder einer einschlägigen Schrankenbestimmung.

Soweit nicht schutzfähige Bestandteile von einzelnen Aktenauszügen wahrnehmbar verarbeitet werden, ist das Projekt urheberrechtlich neutral.

521 *T. Radtke*, ZGE 17 (2025), 1 (24) m.w.N. in Fn. 109; s. insb. am Beispiel der Privatkopie *BGH*, MMR 2020, 538 (Rn. 22); MMR 2024, 1037 (Rn. 15 ff.); *EuGH*, MMR 2018, 80 (Rn. 25) – VCAST Limited/RTI SpA.

522 S. oben unter E.II.2.c.cc(2)(c).

523 S. oben unter E.II.2.b.

Es empfiehlt sich daher auf erster Stufe sowohl (i) die Filterung von wahrscheinlich urheberrechtlich geschützten Werken, die zugleich von geringem Mehrwert für den Trainingsprozess sein dürften (z.B. ausgewählte Gutachtenkategorien) als auch (ii) die Vermeidung von Vervielfältigungen dadurch, dass bereits digitalisierte Aktenauszüge unmittelbar in abgeleitete Textformate übertragen und zum Training des Sprachmodells verwendet werden. Nicht vermeidbare Vervielfältigungen können auf zweiter Stufe für eine automatisierte Zwischenspeicherung zur Weiterverarbeitung auf § 44a Nr. 2 UrhG gestützt werden. Vervielfältigungen, die nicht nur vorübergehend gespeichert werden sollen, können auf dritter Stufe vorzugswürdig auf § 44b Abs. 2 S. 1 UrhG gestützt werden. Die Forschungsausnahme nach § 60d UrhG wird mangels Bezugs zu einem konkreten Forschungsprojekt vorliegend nicht in Anspruch genommen werden können. Für eine Berufung auf § 44b UrhG sind die schutzfähigen Bestandteile der Aktenauszüge auf etwaige Vorbehalte (gegebenenfalls automatisiert) zu prüfen, die sodann zu berücksichtigen sind.

Einzelne Werkausschnitte, die während des Trainingsprozesses oder der Ausführung eines LLM verarbeitet werden, sind nach vorzugswürdiger Auffassung mangels Wahrnehmbarkeit bereits nicht als urheberrechtliche Vervielfältigungen zu behandeln. Sofern man dieser Ansicht nicht folgt, sind die betreffenden Vervielfältigungen ebenfalls als Text und Data Mining über § 44b UrhG zulässig, indem sie die Ausgangsbasis für die weitere Analyse auf dem Weg zu einer Ausgabe im konkreten Fall darstellen. Entsprechende Überlegungen gelten für Bearbeitungen nach § 23 Abs. 3 UrhG.

Eine Ablehnung der Wahrnehmbarkeit setzt insbesondere voraus, dass nicht einzelne Werkbestandteile über die Ausgabe des KI-Systems vervielfältigt werden. Solche Vervielfältigungen sind daher durch geeignete Maßnahmen zu verhindern, die analog zu den datenschutzrechtlich gebotenen technischen und organisatorischen Maßnahmen neben Privacy-Konzepten auf Trainingsebene auch Filtermaßnahmen, einen entsprechenden System Prompt und Benutzungsanweisungen sowie gegebenenfalls vertragliche Benutzungseinschränkungen umfassen. Zudem wirkt sich die organisatorische und institutionelle Einbindung zugunsten der Zulässigkeit aus.[524] Sollten unter Verstoß gegen diese umfangreichen Maßnahmen durch eine Prompt Injection oder vergleichbare Angriffe gezielt Werkbestandteile im Rahmen der Ausgabe extrahiert werden, kommt im Regelfall nur noch der

524 S. schon oben unter D.V.4.c.bb.

jeweilige Anwender als passivlegitimierter Täter dieser Urheberrechtsverletzung in Betracht.

III. Besonderheiten bei der Veröffentlichung des Sprachmodells

Die Veröffentlichung des Sprachmodells kann durch die Eröffnung eines öffentlichen Zugriffs mittels eines KI-Systems, durch die Veröffentlichung auch der Gewichte des Sprachmodells oder durch die Veröffentlichung auch des Trainingskorpus erfolgen. Analog zur datenschutzrechtlichen Betrachtung[525] ergeben sich aufsteigend in der dort beschriebenen Reihenfolge (i.e. KI-System, Sprachmodell und Trainingskorpus) verschiedene urheberrechtliche Implikationen.

Die Bereitstellung eines öffentlich zugänglichen KI-Systems ist urheberrechtlich grundsätzlich vertretbar. Zwar erhöht diese Form der Veröffentlichung das Risiko missbräuchlicher Eingabeaufforderungen durch den nicht beschränkten Nutzerkreis bereits nicht unerheblich; zugleich bleiben aber System Prompts und Ausgabefilter erhalten und es dürfte mangels Kennungen schwerfallen, auf urheberrechtlich geschützte Ausdrucksweisen aufzubauen (z.B. durch die Aufforderung: „Schreibe einen Schriftsatz im Stil von x").

Falls es im (seltenen) Einzelfall zu einer Urheberrechtsverletzung kommt, bleibt eine Passivlegitimation der Stelle denkbar, die das KI-System öffentlich zur Verfügung stellt.[526] Umfangreiche Vorsorgemaßnahmen können allerdings dieser Passivlegitimation in Form einer Täterschaft, Teilnahme oder sonstiger Haftung aufgrund von Verkehrspflichtverletzungen die Grundlage entziehen.

Die Veröffentlichung auch der Gewichte oder gar des Trainingskorpus steigert das Risiko einer Urheberrechtsverletzung erheblich. In diesem Fall entfallen weitere Schutzmechanismen, wie z.B. die Filtermaßnahmen und der System Prompt. Eine solche Veröffentlichung führt daher regelmäßig zu Urheberrechtsverletzungen, für die eine Passivlegitimation der Projektanbieter denkbar ist. Etwas anderes würde nur gelten, wenn sich die berechtigten Interessen an einer Open Source-Veröffentlichung in den urheberrechtlichen Schrankenbestimmungen niederschlagen würden. Zu-

525 S. oben unter D.IV.3.

526 S. zur Haftung einer Plattform bei der Bereitstellung von geschützten Trainingsdaten *Siglmüller/Gassner*, RDi 2023, 124 (Rn. 32).

mindest auf geschriebene Schrankenbestimmungen wird sich die erhöhte Transparenz durch Open Source, soweit ersichtlich, nicht stützen können.

IV. Exkurs: Eigene Forschungszwecke der ausführenden Stellen

Sofern im hypothetischen Szenario[527] die ausführenden Stellen eigene Forschungszwecke verfolgen würden, blieben die grundlegenden urheberrechtlichen Erwägungen mit Blick auf abgeleitete Textformate und § 44b UrhG grundsätzlich unverändert. Unter der Prämisse der Verfolgung eigener Forschungszwecke liegt allerdings eine Anwendbarkeit des Forschungsprivilegs für Text und Data Mining aus § 60d UrhG näher. Dieser Prämisse mit besonderer Relevanz für die Zusammenarbeit von Praxis und Wissenschaft soll nachfolgend ebenfalls mit Blick auf die urheberrechtlichen Implikationen in Form eines Exkurses nachgegangen werden.

1. Vervielfältigungen im Rahmen des GSJ-Projekts

Abhängig von den Einzelheiten der Verwendung der Vervielfältigungen für das Training des Sprachmodells bleibt es bei dem primären Nichtforschungszweck in Form der Entwicklung eines in der Justiz einsetzbaren Sprachmodells und der Nichtanwendbarkeit des § 60d UrhG (s. oben unter E.II.2.c.dd). Die Forschungszwecke könnten aber im Einzelfall durchaus auch überwiegen sowie die Anwendbarkeit des § 60d UrhG für Vervielfältigungen zu Zwecken des Text und Data Mining eröffnen, womit es nicht auf erklärte Vorbehalte nach § 44b UrhG ankommen würde.

Die Anwendbarkeit des § 60d UrhG dürfte insbesondere für Vervielfältigungen anzunehmen sein, die nicht zwingend und unmittelbar über das Text und Data Mining in das Endprodukt eines GSJ-Sprachmodells für die Justiz münden, sondern lediglich Zwischenschritte für die Produktentwicklung, aber auch die (Grundlagen-)Forschung darstellen (z.B. Vervielfältigungen für das Training eines kleinen Testmodells, das selbst nicht in der Justiz operativ zum Einsatz gebracht werden soll).

527 S. zu den datenschutzrechtlichen Implikationen unter D.VII.1.

2. Vervielfältigungen über das GSJ-Projekt hinaus

Vervielfältigungen der ausführenden Stellen zu Zwecken des Text und Data Mining, die von vornherein über die Entwicklung des GSJ-Sprachmodells hinausgehen und der eigenen (Grundlagen-)Forschung der ausführenden Stellen dienen, können grundsätzlich ebenfalls auf § 60d UrhG gestützt werden.

V. Schlussbetrachtung aus urheberrechtlicher Perspektive

Das Urheberrecht erfasst das KI-Training insbesondere über Vervielfältigungen im Vorfeld des KI-Trainings. Diese Vervielfältigungen können im Kontext des GSJ-Projekts gestützt werden auf § 44b UrhG bzw. § 44a UrhG, soweit die Vervielfältigungen temporär als integraler Bestandteil des automatisierten Prozesses angelegt werden. Das GSJ-Sprachmodell selbst enthält (nach allerdings umstrittener Auffassung) grundsätzlich keine Vervielfältigungen.

Vervielfältigungen im Rahmen der Ausgabe sind durch angemessene urheberrechtliche Schutzmaßnahmen zu verhindern, die im Grundsatz den datenschutzrechtlich erforderlichen Schutzmaßnahmen entsprechen. Derartige urheberrechtliche Schutzmaßnahmen verlieren an Wirksamkeit bzw. entfallen sogar, wenn ein KI-System oder das gesamte trainierte Sprachmodell der Öffentlichkeit zugänglich gemacht werden. Eine Veröffentlichung des Sprachmodells ist aus urheberrechtlicher Sicht daher bedenklich.

Daraus lassen sich die folgenden Anforderungen und hierauf bezogenen Handlungsempfehlungen ableiten.

1. Anforderungen an die Zusammenstellung der Trainingsdaten

Durch den Einsatz abgeleiteter Textformate und das KI-Training an der Quelle sind Vervielfältigungen möglichst zu vermeiden.

Wenn und soweit sich Vervielfältigungen nicht vermeiden lassen, sind die folgenden Einsatzbedingungen mit Blick auf die Aktenauszüge zu beachten:

- Sichtung und Clusterung der typischen Aktenbestandteile im Hinblick auf gegebenenfalls auszusortierende Werkkategorien;

- Aussortierung und gegebenenfalls Prüfung von Werken mit besonders hohem urheberrechtlichen Schutzstandard (z.B. umfangreiche, vor allem textbasierte Gutachten);
- Automatisierte Löschung von notwendigen, zwischengespeicherten Kopien (z.B. zur Übermittlung an einen Trainingsserver);
- Prüfung der Aktenauszüge auf Rechtsvorbehalte i.S.d. § 44b Abs. 3 UrhG (z.B. „Kein Text oder Data Mining", „Kein Text und Data Mining", „Kein TDM", „Kein KI-Training", „Kein LLM-Training", gegebenenfalls auch „Alle Rechte vorbehalten" oder „Vervielfältigung vorbehalten") und Aussortierung der entsprechenden Inhalte.

2. Anforderungen an das KI-System und dessen Einsatz

Im Hinblick auf den Einsatz des KI-Systems sind weitere Bedingungen zu beachten, die zugleich auch auf die Frage zurückwirken, ob das KI-Modell Vervielfältigungen enthält. Diese weiteren Bedingungen stimmen teilweise mit den datenschutzrechtlich erforderlichen Maßnahmen überein (siehe unter D.VIII.4: insbesondere Login-Mechanismus, Vorgabe der Eingabeaufforderung).

Darüber hinaus sind die notwendigen Maßnahmen vor allem im Fall einer frei wählbaren Eingabeaufforderung auch auf die Anforderungen des Urheberrechts auszurichten:

- Gegebenenfalls Einrichtung eines System Prompts, wonach der Stil ausgewählter Dokumentenkategorien aus den Aktenauszügen nicht nachgeahmt werden darf (z.B. Verhinderung von Prompts, einen Schriftsatz im Stil von Aktenauszügen aus einem bestimmten Themengebiet des Verkehrsrechts zu schreiben, der maßgeblich von einem bestimmten Rechtsanwalt dominiert wird);
- Gegebenenfalls Implementierung von Mechanismen zur Prüfung der Ausgabe (z.B. ein separates Sprachmodell zur Überprüfung, ob die Ausgabe vor allem einen konkreten Stil imitiert bzw. imitieren soll).

In Ansehung der bereits datenschutzrechtlich gebotenen Entfernung von Kennungen dürfte es ohnehin grundsätzlich nicht möglich sein, den Stil einer einzelnen Person durch eine entsprechende Eingabeaufforderung gezielt zu imitieren.

3. Anforderungen an die Veröffentlichung des Sprachmodells

Von einer Veröffentlichung des Sprachmodells gehen für das Urheberrecht weniger und geringere urheberrechtliche Implikationen aus als es für das Datenschutzrecht der Fall ist.

Die Veröffentlichung in der Form eines öffentlich zugänglichen KI-Systems erhöht geringfügig die Wahrscheinlichkeit von Urheberrechtsverletzungen wegen des unbeschränkten Nutzerkreises. In Abhängigkeit davon, ob die Eingabeaufforderung frei gewählt werden kann, sind gezielte urheberrechtsverletzende Ausgaben gegebenenfalls unwahrscheinlicher. Dieser Befund gilt insbesondere in Ansehung der fehlenden Kennungen, sodass nicht ohne Weiteres der Bezug zu dem gegebenenfalls urheberrechtlich geschützten Stil einer Person hergestellt werden kann.

Die Veröffentlichung der Gewichte oder gar des Trainingsdatensatzes machen eine Urheberrechtsverletzung allerdings erheblich wahrscheinlicher, da insoweit urheberrechtsrelevante Werkbestandteile mit gewisser Wahrscheinlichkeit extrahiert werden können. In diesen Fällen ist auch eine Passivlegitimation und damit eine Haftung der Ministerien denkbar, sodass dem Grunde nach von einer Veröffentlichung des Modells (Open Source) abzuraten ist.

F. Literaturverzeichnis

Adrian, Axel/Dykes, Nathan/Evert, Stephanie/Heinrich, Philipp/Keuchen, Michael/Proisl, Thomas, Manuelle und automatische Anonymisierung von Urteilen, in: Axel Adrian, Michael Kohlhase, Stephanie Evert et al. (Hrsg.), Digitalisierung von Zivilprozess und Rechtsdurchsetzung, 2022, S. 173-198.

Adrian, Axel/Kohlhase, Michael/Evert, Stephanie/Zwickel, Martin (Hrsg.), Digitalisierung von Zivilprozess und Rechtsdurchsetzung, 2022.

Albalak, Alon/Elazar, Yanai/Xie, Sang Michael/Longpre, Shayne/Lambert, Nathan/Wang, Xinyi/Muennighoff, Niklas/Hou, Bairu/Pan, Liangming/Jeong, Haewon/Raffel, Colin/Chang, Shiyu/Hashimoto, Tatsunori/Wang, William Yang, A Survey on Data Selection for Language Models, 26. Februar 2024, http://arxiv.org/pdf/2402.16827.

Albrecht, Jan Philipp, Das neue EU-Datenschutzrecht – von der Richtlinie zur Verordnung, Überblick und Hintergründe zum finalen Text für die Datenschutz-Grundverordnung der EU nach der Einigung im Trilog, CR 2016, 88-98.

Al-Rubaie, Mohammad/Chang, J. Morris, Privacy Preserving Machine Learning: Threats and Solutions, 27. März 2018, http://arxiv.org/pdf/1804.11238.

Antoine, Lucie, Die Pflichten aus Art. 53 I KI-VO – Hilfreich für das Urheberrecht?, GRUR 2025, 118-129.

Arning, Marian/Forgó, Nikolaus/Krügel, Tina, Datenschutzrechtliche Aspekte der Forschung mit genetischen Daten, DuD 2006, 700-705.

Arnold, Claudius, Amtliche Werke im Urheberrecht, Zur Verfassungsmäßigkeit und analogen Anwendung des § 5 UrhG, zugl. Dissertation, Baden-Baden 1994.

Artikel-29-Datenschutzgruppe, WP 216, Stellungnahme 5/2014 zu Anonymisierungstechniken, https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_de.pdf.

Artikel-29-Datenschutzgruppe, WP 136, Stellungnahme 4/2007 zum Begriff „personenbezogene Daten", 20. Juni 2007, https://www.lda.bayern.de/media/wp136_de.pdf.

Artikel-29-Datenschutzgruppe, WP 203, Opinion 03/2023 on purpose limitation, 2. April 2013.

Artikel-29-Datenschutzgruppe, WP248 rev.01, Leitlinien zur Datenschutz-Folgenabschätzung (DSFA) und Beantwortung der Frage, ob eine Verarbeitung im Sinne der Verordnung 2016/679 wahrscheinlich ein hohes Risiko mit sich bringt", 4. April 2017.

Artikel-29-Datenschutzgruppe, WP251 rev.01, Leitlinien zu automatisierten Entscheidungen im Einzelfall einschließlich Profiling für die Zwecke der Verordnung 2016/679, 6. Februar 2018.

Artikel-29-Datenschutzgruppe, WP 260 rev.01, Leitlinien für Transparenz gemäß der Verordnung 2016/679, 11. April 2018.

Ashkar, Daniel, Wesentliche Anforderungen der DS-GVO bei Einführung und Betrieb von KI-Anwendungen, Überblick und Praxishinweise zur Umsetzung, ZD 2023, 523-530.

Assion, Simon, Die Entwicklung des Datenschutzrechts, NJW 2023, 2619-2624.

Barudi, Malek (Hrsg.), Das neue Urheberrecht, UrhG, UrhDaG, VGG, NomosPraxis, Baden-Baden 2021.

Baumann, Generative KI und Urheberrecht – Urheber und Anwender im Spannungsfeld, NJW 2023, 3673-3678.

Baumgartner, Ulrich, Anmerkung zu EuG, Urt. v. 26.04.2023 – T-557/20, ZD 2023, 402-404.

BayLDA, Datenschutzkonforme Künstliche Intelligenz, Checkliste mit Prüfkriterien nach DS-GVO, 24. Januar 2024.

Becher, Jonathan D./Berkhin, Pavel/Freeman, Edmund, Automating exploratory data analysis for efficient data mining, in: Raghu Ramakrishnan, Sal Stolfo, Roberto Bayardo et al. (Hrsg.), Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD00: The Second Annual International Conference on Knowledge Discovery in Data in Boston Massachusetts USA, New York 2000, S. 424-429.

Bergt, Matthias, Die Bestimmbarkeit als Grundproblem des Datenschutzrechts, Überblick über den Theorienstreit und Lösungsvorschlag, ZD 2015, 365-371.

Berz, Amelie/Engel, Andreas/Hacker, Philipp, Generative KI, Datenschutz, Hassrede und Desinformation – Zur Regulierung von KI-Meinungen, ZUM 2023, 586-594.

Beurskens, Michael, Training generativer KI nur auf Lizenzgrundlage?, Eine Analyse des rechtlichen Gutachtens von Dornis/Stober, RDi 2025, 1-7.

BfDI, Positionspapier zur Anonymisierung unter der DSGVO unter besonderer Berücksichtigung der TK-Branche, 29. Juni 2020.

Biallaß, Isabelle Désirée, Die Auswirkungen der KI-VO auf die Justiz, Nutzung von KI durch Gerichte und Staatsanwaltschaften, MMR 2024, 646-651.

BMBF, Eckpunkte BMBF Forschungsdatengesetz, 28. Februar 2024, https://www.bmbf.de/SharedDocs/Downloads/DE/gesetze/forschungsdatengesetz/sonstige/Eckpunktepapier.pdf?__blob=publicationFile&v=3.

Bomhard, David, KI-Training mit fremden Daten, IP-rechtliche Herausforderungen rund um § 44 b UrhG, DSRITB 2023, 255-269.

Bomhard, David, Text und Data Mining auf Grundlage von Webcrawling und Webscraping, InTeR 2023, 174-179.

Brink, Stefan/Eckhardt, Jens, Wann ist ein Datum ein personenbezogenes Datum?, Anwendungsbereich des Datenschutzrechts, ZD 2015, 205-212.

Britz, Thomas/Indenhuck, Moritz/Langerhans, Tom, Die Verarbeitung „zufällig" sensibler Daten, Einschränkende Auslegung von Art. 9 DS-GVO, ZD 2021, 559-564.

Brockmeyer, Henning, Text und Data Mining, zugl. Dissertation, München 2022.

Buchalik, Barbara/Gehrmann, Mareike Christine, Von Nullen und Einsen zu Paragraphen: Der AI Act, ein Rechtscode für Künstliche Intelligenz, Der horizontale und risikobasierte Ansatz für Produktsicherheitsaspekte von KI-Systemen und Allzweck-KI, CR 2024, 145-153.

Burghoff, Ramon, Praxisgerechter Umgang mit der Verfremdung personenbezogener Daten, Empfehlungen für Unternehmen, ZD 2023, 658-664.

Carlini, Nicholas/Ippolito, Daphne/Jagielski, Matthew/Lee, Katherine/Tramer, Florian/Zhang, Chiyuan, Quantifying Memorization Across Neural Language Models, 15. Februar 2022, http://arxiv.org/pdf/2202.07646.

Carlini, Nicholas/Liu, Chang/Erlingsson, Úlfar/Kos, Jernej/Song, Dawn, The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks, 22. Februar 2018, http://arxiv.org/pdf/1802.08232.

Carlini, Nicholas/Tramer, Florian/Wallace, Eric/Jagielski, Matthew/Herbert-Voss, Ariel/Lee, Katherine/Roberts, Adam/Brown, Tom/Song, Dawn/Erlingsson, Ulfar/Oprea, Alina/Raffel, Colin, Extracting Training Data from Large Language Models, 14. Dezember 2020, http://arxiv.org/pdf/2012.07805.

Chiou, Theodoros, Copyright lessons on Machine Learning: what impact on algorithmic art?, JIPITEC 2019, 398-411.

Culik, Nicolai/Döpke, Christian, Zweckbindungsgrundsatz gegen unkontrollierten Einsatz von Big Data-Anwendungen, Analyse möglicher Auswirkungen der DS-GVO, ZD 2017, 226-230.

Dänische Datatilsynet, Offentlige myndigheders brug af kunstig intelligens, Oktober 2023, https://www.datatilsynet.dk/Media/638321084132236143/Offentlige%20myndigheders%20brug%20af%20kunstig%20intelligens%20-%20Inden%20I%20g%C3%A5r%20i%20gang.pdf.

Determann, Lothar/Paal, Boris P., KI-Recht international, Compliance Field Guide, Baden-Baden 2025.

Deuber, Dominic/Keuchen, Michael, Anonymisierung von Gerichtsentscheidungen im Lichte der IT-Sicherheit, Nachweis der Unsicherheit eines geheim gehaltenen Anonymisierungsverfahrens und Alternativen, MMR 2023, 338-344.

Deuber, Dominic/Keuchen, Michael/Christin, Nicolas, Assessing Anonymity Techniques Employed in German Court Decisions: A De-Anonymization Experiment 2023.

Devlin, Jacob/Chang, Ming-Wei/Lee, Kenton/Toutanova, Kristina, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 11. Oktober 2018, http://arxiv.org/pdf/1810.04805.

Dewitte, Pierre, AI Meets the GDPR, in: Nathalie A. Smuha (Hrsg.), The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence, Cambridge 2025, S. 133-157.

Dieker, Amon, Datenschutzrechtliche Zulässigkeit der Trainingsdatensammlung, Wie Scraping und Crawling zur KI-Entwicklung eingesetzt werden können, ZD 2024, 132-137.

Döhl, Frédéric, Digitaler Game Changer für die Gedächtnisinstitutionen?, – Digital Humanities, die anstehende Schranke der Urheber- und Leistungsschutzrechte zugunsten der wissenschaftlichen Nachnutzung von Korpora bei Text und Data Mining und die unadressiert gebliebene Herausforderung des rechtmäßigen Zugangs –, RuZ 2020, 195-218.

Dornis, Tim W., Generatives KI-Training und Text- und Data-Mining, Eine funktionale Unterscheidung, KIR 2024, 156-161.

Dornis, Tim W./Stober, Sebastian, Urheberrecht und Training generativer KI-Modelle, Technologische und juristische Grundlagen, Baden-Baden 2024.

Dregelies, Max, KI-Training unter dem AI Act, GRUR 2024, 1484-1493.

Dreier, Thomas/Raue, Benjamin/Specht-Riemenschneider, Louisa/Mantz, Reto/Schulze, Gernot (Hrsg.), Dreier/Schulze, Urheberrechtsgesetz, Urheberrechts-Diensteanbieter-Gesetz, Verwertungsgesellschaftengesetz, Nebenurheberrecht, Portabilitätsverordnung, Marrakeschverordnung, Kunsturhebergesetz : Kommentar, 8. Aufl., München 2025.

DSK, Erfahrungsbericht der unabhängigen Datenschutzaufsichtsbehörden des Bundes und der Länder zur Anwendung der DS-GVO, November 2019, https://www.datenschutzkonferenz-online.de/media/dskb/20191213_erfahrungsbericht_zur_anwendung_der_ds-gvo.pdf.

DSK, Orientierungshilfe der Aufsichtsbehörden zur Verarbeitung von personenbezogenen Daten für Zwecke der Direktwerbung unter Geltung der Datenschutz-Grundverordnung, Februar 2022, https://www.datenschutzkonferenz-online.de/media/oh/OH-Werbung_Februar%202022_final.pdf.

DSK, Orientierungshilfe Künstliche Intelligenz und Datenschutz, Version 1.0, 6. Mai 2024.

Eckhardt, Jens, Anwendungsbereich des Datenschutzrechts – Geklärt durch den EuGH?, Der europarechtliche Grundsatz des Personenbezugs, CR 2016, 786-790.

EDPB, Leitlinien 02/2019, für die Verarbeitung personenbezogener Daten gemäß Artikel 6 Absatz 1 Buchstabe b DSGVO im Zusammenhang mit der Erbringung von Online-Diensten für betroffene Personen, 8. Oktober 2019.

EDPB, Leitlinien 07/2020 zu den Begriffen „Verantwortlicher" und „Auftragsverarbeiter" in der DSGVO, Version 2.0, 7. Juli 2021.

EDPB, Verbindlicher Beschl. 1/2021, 28. Juli 2021, https://www.edpb.europa.eu/system/files/2022-03/edpb_bindingdecision_202101_ie_sa_whatsapp_redacted_de.pdf.

EDPB, Leitlinien 01/2022 zu den Rechten der betroffenen Person – Auskunftsrecht, Version 2.1, 28. März 2023.

EDPB, Report of the work undertaken by the ChatGPT Taskforce, 23. Mai 2024, https://www.edpb.europa.eu/system/files/2024-05/edpb_20240523_report_chatgpt_taskforce_en.pdf.

EDPB, Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, 17. Dezember 2024.

EDPB, Guidelines 01/2025 on Pseudonymisation, 16. Januar 2025, https://www.edpb.europa.eu/system/files/2025-01/edpb_guidelines_202501_pseudonymisation_en.pdf.

Ehmann, Eugen/Selmayr, Martin (Hrsg.), Datenschutz-Grundverordnung Kommentar, 3. Aufl., München 2024.

Engeler, Malte/Rolfes, Louis, Datenschutzrechtliche Korrekturansprüche bei Erzeugung von Falschinformationen durch LLMs, Potenziale einer alternativen Datenökonomie, ZD 2024, 423-429.

Franzen, Martin/Gallner, Inken/Oetker, Hartmut (Hrsg.), Kommentar zum europäischen Arbeitsrecht, 5. Aufl., München 2024.

Fredrikson, Matt/Jha, Somesh/Ristenpart, Thomas, Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures, in: Indrajit Ray, Ninghui Li, Christopher Kruegel (Hrsg.), Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, CCS'15: The 22nd ACM Conference on Computer and Communications Security in Denver Colorado USA, New York 2015, S. 1322-1333.

Freye, Merle/Schnebbe, Maximilian, Digitale Gerichtsverhandlung, Datenschutzrechtliche Analyse einer Verhandlung nach § 128a ZPO, ZD 2020, 502-506.

Fuchs, Thomas/Wünschelbaum, Markus, Wahrscheinlichkeiten reichen nicht: Warum KI-Modelle keine Datenspeicher sind, RDV 2024, 314-315.

Geiger, Christophe, When the Robots (Try to) Take Over: Of Artificial Intelligence, Authors, Creativity and Copyright Protection, in: Florent Thouvenin, Alexander Peukert, Thomas Jaeger et al. (Hrsg.), Kreation Innovation Märkte - Creation Innovation Markets, Festschrift Reto M. Hilty, Berlin, Heidelberg 2024, S. 67-87.

Geiger, Christophe/Frosio, Giancarlo/Bulayenko, Oleksandr, Text and Data Mining: Articles 3 and 4 of the Directive 2019/790/EU, CEIPI Research Paper 2019, 10.2139/ssrn.3470653.

Gersdorf, Hubertus/Paal, Boris P. (Hrsg.), BeckOK Informations- und Medienrecht, 47. Aufl., München 2025.

Gola, Peter/Heckmann, Dirk (Hrsg.), Datenschutz-Grundverordnung Bundesdatenschutzgesetz, 3. Aufl., München 2022.

Golland, Alexander, KI und KI-Verordnung aus datenschutzrechtlicher Sicht, EuZW 2024, 846-854.

Golub, Koraljka/Liu, Ying-Hsang (Hrsg.), Information and knowledge organisation in digital humanities, Global perspectives, Digital research in the arts and humanities, New York 2021.

Götting, Horst-Peter/Lauber-Rönsberg, Anne/Rauer, Nils (Hrsg.), BeckOK Urheberrecht, 45. Aufl., München 2025.

Graf, Jürgen (Hrsg.), BeckOK GVG, 26. Aufl., München 2025.

Grisse, Karina, Nutzbarmachung urheberrechtlich geschützter Textbestände für die Forschung durch Dritte, Rechtliche Bedingungen und Möglichkeiten, RuZ 2020, 143-159.

Hacker, Philipp, A legal framework for AI training data—from first principles to the Artificial Intelligence Act, Law, Innovation and Technology 13 (2021), 257-301.

Halim, Valentino/Marosi, Johannes, Anmerkung zu EuGH, Urt. v. 07.03.2024 – C-604/22, ZD 2024, 333-334.

Hallinan, Dara/Zuiderveen Borgesius, Frederik, Opinions can be incorrect (in our opinion)! On data protection law's accuracy principle, International Data Privacy Law 10 (2020), 1-10.

Hamann, Hanjo, Nutzungsvorbehalte für KI-Training in der Rechtsgeschäftslehre der Maschinenkommunikation, ZGE 16 (2024), 113-168.

Hanloser, Stefan, Anmerkung zu EuGH, Urt. v. 09.11.2023 – C-319/22, ZD 2024, 175-176.

Hansen, Marit/Walczak, Benjamin, Die KI zaubert nicht, Warum ein Personenbezug in LLMs erhalten bleibt, KIR 2024, 82-86.

Heinrich, Thomas, Schadensgutachten und Urheberschutz, NZV 2015, 68-70.

Hilber, Marc/Borges, Georg (Hrsg.), BeckOK IT-Recht, 17. Aufl., München 2025.

Hirte, Heribert, Mitteilung und Publikation von Gerichtsentscheidungen, Zum Spannungsverhältnis von Persönlichkeitsschutz und Interessen der Öffentlichkeit, NJW 1988, 1698-1705.

HmbBfDI, Diskussionspapier: Large Language Models und personenbezogene Daten, 2024, https://datenschutz-hamburg.de/fileadmin/user_upload/HmbBfDI/Datenschutz/Informationen/240715_Diskussionspapier_HmbBfDI_KI_Modelle.pdf.

Hoeren, Thomas, Big Data und Datenqualität – ein Blick auf die DS-GVO, Annäherungen an Qualitätsstandards und deren Harmonisierung, ZD 2016, 459-463.

Hoffman-Riem, Wolfgang, OVG Bremen, 25.10.1988 – 1 BA 32/88. Zum Gebot der Gleichbehandlung bei der Versorgung von Fachzeitschriften mit veröffentlichungswürdigen Entscheidungen, JZ 1989, 637-638.

Hofmann, Franz, Retten Schranken Geschäftsmodelle generativer KI-Systeme?, ZUM 2024, 166-174.

Hofmann, Franz, Zehn Thesen zu Künstlicher Intelligenz (KI) und Urheberrecht, WRP 2024, 11-18.

Holznagel, Bernd/Hoeren, Thomas/Sieber, Ulrich (Hrsg.), Handbuch Multimedia-Recht, Rechtsfragen des elektronischen Geschäftsverkehrs, 2014. Aufl., München 2014.

Hornung, Gerrit/Wagner, Bernd, Die Zwickmühle der Re-Identifizierbarkeit in Zeiten von Big Data und Ubiquitous Computing, CR 2019, 565-574.

Hornung, Gerrit/Wagner, Bernd, Anonymisierung als datenschutzrelevante Verarbeitung?, Rechtliche Anforderungen und Grenzen für die Anonymisierung personenbezogener Daten, ZD 2020, 223-228.

Hubert, Tom, Die Identifizierbarkeit i.S.d. Art. 4 Nr. 1 DSGVO insbesondere durch Informationen aus dem Internet, (Endlich) alles klar durch die Rechtsprechungslinie des EuGH zum Personenbezug?, CR 2025, 77-85.

Hüger, Jakob, Künstliche Intelligenz und Diskriminierung, zugl. Dissertation, Baden-Baden 2023.

Hüger, Jakob, Die Rechtmäßigkeit von Datenverarbeitungen im Lebenszyklus von KI-Systemen, ZfDR 2024, 263-291.

Hüger, Jakob/Radtke, Tristan, Das Zusammenspiel der Akteure und Verantwortlichkeit unter der KI-Verordnung und der DS-GVO, Synergieeffekte bei der Pflichtenerfüllung mit Blick auf KI-Systeme, KIR 2025, 154-161.

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

International Working Group on Data Protection in Technology, Working Paper on Large Language Models (LLMs), 6. Dezember 2024.

Jager, Lucas Anton, Artificial Creativity?, Zur urheberrechtlichen Verortung von Systemen ›künstlicher Intelligenz‹, deren Erzeugnissen und deren Einsatzmöglichkeiten unter besonderer Berücksichtigung künstlicher neuronaler Netzwerke, zugl. Dissertation, Berlin 2023.

Jansen, Richard/Radtke, Tristan, Interne Prüfungen und besondere Kategorien personenbezogener Daten, PinG 2024, 61-64.

Käde, Lisa, Kreative Maschinen und Urheberrecht, zugl. Dissertation, Baden-Baden 2021.

Käde, Lisa, Do You Remember? – Enthalten KI-Modelle Vervielfältigungen von Trainingsdaten, lassen sich diese gezielt rekonstruieren und welche Implikationen hat das für das Urheberrecht?, ZUM 2024, 174-183.

Käde, Lisa, Training generativer KI-Modelle ist (auch) Text- und Data-Mining, Anwendbarkeit der TDM-Schranke des § 44b UrhG, KIR 2024, 162-169.

Kandpal, Nikhil/Wallace, Eric/Raffel, Colin, Deduplicating Training Data Mitigates Privacy Risks in Language Models, 14. Februar 2022, http://arxiv.org/pdf/2202.06539.

Karamolegkou, Antonia/Li, Jiaang/Zhou, Li/Søgaard, Anders, Copyright Violations and Large Language Models, 20. Oktober 2023, http://arxiv.org/pdf/2310.13771.

Katzenberger, Paul, Die Frage des urheberrechtlichen Schutzes amtlicher Werke, GRUR 1972, 686-695.

Kaulartz, Markus/Braegelmann, Tom (Hrsg.), Rechtshandbuch Artificial Intelligence und Machine Learning, München 2020.

Keppeler, Lutz, „Objektive Theorie" des Personenbezugs und „berechtigtes Interesse" als Untergang der Rechtssicherheit?, Eine Analyse der Schlussanträge des Generalanwalts in der Rechtssache C-582/14 (Speicherung dynamischer IP-Adressen), CR 2016, 360-367.

Keppeler, Lutz/Poncza, Manuel/Wölke, Annika, Lockert der EuGH durch sein FIN-Urt. den strengen „Personenbezug"?, Eine kritische Analyse der EuGH-Rechtsprechung anlässlich EuGH, Urt. v. 9.11.2023 – C-319/22, CR 2024, 18-22.

Klass, Nadine, Das Urheberrecht in Arbeits- und Dienstverhältnissen, Eine Analyse der deutschen Rechtslage im Vergleich mit Systemen angloamerikanischer Prägung, GRUR 2019, 1103-1114.

Kleinkopf, Felicitas Lea, Text- und Data-Mining, Die Anforderungen digitaler Forschungsmethoden an ein innovations- und wissenschaftsfreundliches Urheberrecht, zugl. Dissertation, Baden-Baden 2022.

Kockler, Franz Josef, Publikation von Gerichtsentscheidungen und Anonymisierung, JurPC 1996, 46-54.

Kohn, Matthias/Schleper, Janine, Die (zufällige) Erhebung sensibler Daten, Ein Lösungsvorschlag auf Grundlage der ergangenen EuGH-Urteile, ZD 2023, 723-728.

Konertz, Roman, Urheberrechtliche Fragen der Plagiatskontrolle an Hochschulen, ZUM 2024, 355-364.

Konertz, Roman/Schönhof, Raoul, Rechtsfolgen der Urheberrechtsverletzung bei generativer Künstlicher Intelligenz, Über die Möglichkeit des „Vergessens" in Neuronalen Netzen, WRP 2024, 534-541.

Kroschwald, Steffen, Verschlüsseltes Cloud Computing, Auswirkung der Kryptografie auf den Personenbezug in der Cloud, ZD 2014, 75-80.

Krügel, Tina, Das personenbezogene Datum nach der DS-GVO, Mehr Klarheit und Rechtssicherheit?, ZD 2017, 455-460.

Kugler, Kai/Münker, Simon/Höhmann, Johannes/Rettinger, Achim, InvBERT: Reconstructing Text from Contextualized Word Embeddings by inverting the BERT pipeline, in: Conference Reader: 2nd Annual Conference of Computational Literary Studies, 2023.

Kühling, Jürgen/Buchner, Benedikt (Hrsg.), Datenschutz-Grundverordnung / BDSG Kommentar, 4. Aufl., München 2024.

Kühling, Jürgen/Klar, Manuel, Anmerkung zu EuGH, Urt. v. 19.10.2016 – C-582/14, ZD 2017, 27-29.

La Durantaye, Katharina de, »Garbage in, garbage out« – Die Regulierung generativer KI durch Urheberrecht, ZUM 2023, 645-660.

La Durantaye, Katharina de, Nutzung urheberrechtlich geschützter Inhalte zum Training generativer künstlicher Intelligenz – ein Lagebericht, AfP 2024, 9-22.

Labusga, Jan-Hendrik/Petit, Marc, Die Veröffentlichung gerichtlicher Geschäftsverteilungspläne im Internet, NJW 2022, 300-304.

Leistner, Matthias, TDM und KI-Training in der Europäischen Union, Erste Fingerzeige des LG Hamburg im „LAION"-Urteil, GRUR 2024, 1665-1675.

LfDIBW, Diskussionspapier: Rechtsgrundlagen im Datenschutz beim Einsatz von Künstlicher Intelligenz, Aktualisierte Version 2.0, 17. Oktober 2024, https://www.baden-wuerttemberg.datenschutz.de/rechtsgrundlagen-datenschutz-ki/.

Loewenheim, Ulrich (Hrsg.), Handbuch des Urheberrechts, 3. Aufl., München 2021.

Loewenheim, Ulrich/Leistner, Matthias/Ohly, Ansgar (Hrsg.), Schricker/Loewenheim, Urheberrecht, UrhG, KUG, VGG Kommentar, 6. Aufl., München 2020.

Ludyga, Hannes, Die Veröffentlichung und Anonymisierung von Gerichtsentscheidungen, ZUM 2021, 887-893.

Maamar, Niklas, Urheberrechtliche Fragen beim Einsatz von generativen KI-Systemen, ZUM 2023, 481-491.

Maltzan, Stephanie von/Käde, Lisa, Algorithmen, die nicht vergessen – Model Inversion Attacks und deren Bedeutung für den Schutz der Daten und der Urheberrechte, DSRITB 2020, 505-525.

Mertens, Timon/Meyer, Dominik, Die datenschutzrechtliche Verantwortlichkeit von Data-Scrapern, K&R 2023, 563-570.

Meyer, Lennart, Aufgedrängte Daten und das DS-GVO-Pflichtenprogramm im Kontext generativer KI-Systeme, Verarbeitung, Verantwortlichkeit, Rechtsgrundlagen und Informationspflichten, RDi 2025, 125-135.

Moos, Flemming, Personenbezug von Large Language Models, Eine datenschutzrechtliche Grundsatzfrage bei der Nutzung generativer KI-Modelle, CR 2024, 442-450.

Moos, Flemming/Rothkegel, Tobias, Anm. zu EuGH, Urt. v. 19.10.2016 – C-582/14, MMR 2016, 845-847.

Mueller, Felix B./Görge, Rebekka/Bernzen, Anna K./Pirk, Janna C./Poretschkin, Maximilian, LLMs and Memorization: On Quality and Specificity of Copyright Compliance, 28. Mai 2024, http://arxiv.org/pdf/2405.18492.

Mühlhoff, Rainer/Ruschemeier, Hannah, KI-Regulierung durch Zweckbindung für Modelle, ZfDR 2024, 337-364.

Nabulsi, Selma, Verarbeitungsanforderungen bei Mischdatensätzen, Sensibel + Nichtsensibel = Sensibel?, ZD 2025, 3-7.

Nasr, Milad/Carlini, Nicholas/Hayase, Jonathan/Jagielski, Matthew/Cooper, A. Feder/Ippolito, Daphne/Choquette-Choo, Christopher A./Wallace, Eric/Tramèr, Florian/Lee, Katherine, Scalable Extraction of Training Data from (Production) Language Models, 28. November 2023, http://arxiv.org/pdf/2311.17035.

Naveed, Humza/Khan, Asad Ullah/Qiu, Shi/Saqib, Muhammad/Anwar, Saeed/Usman, Muhammad/Akhtar, Naveed/Barnes, Nick/Mian, Ajmal, A Comprehensive Overview of Large Language Models, 12. Juli 2023, http://arxiv.org/pdf/2307.06435.

Nöhre, Ingo, Anonymisierung und Neutralisierung von veröffentlichungswürdigen Gerichtsentscheidungen, Empfehlungen für den praktischen Umgang mit personenbezogenen Daten bei der Dokumentationsarbeit der Gerichtsverwaltungen, MDR 2019, 136-141.

Nolte, Henrik/Finck, Michèle/Meding, Kristof, Machine Learners Should Acknowledge the Legal Implications of Large Language Models as Personal Data, 3. März 2025, http://arxiv.org/pdf/2503.01630.

Nordemann-Schiffel, Anke/Nordemann, Axel, Alles Käse oder vielleicht doch mehr?, Gedanken zum Werkbegriff des EuGH, in: Karl-Nikolaus Peifer, Sebastian Kubis, Malte Stieper et al. (Hrsg.), Ius Vivum: Kunst – Internationales – Persönlichkeit, FS Schack, 2022, S. 237-244.

noyb, Complaint against OpenAI, 20. März 2025, https://noyb.eu/sites/default/files/2025-03/OpenAI_complaint_redacted.pdf.

Organisciak, Peter/Downie, J. Stephen, Research access to in-copyright texts in the humanities, in: Koraljka Golub, Ying-Hsang Liu (Hrsg.), Information and knowledge organisation in digital humanities, Global perspectives, Digital research in the arts and humanities, New York 2021, S. 157-177.

Ory, Stephan/Weth, Stephan (Hrsg.), juris PraxisKommentar Elektronischer Rechtsverkehr (jurisPK-ERV), 2. Aufl., 2022.

Ott, Stephan, Bildersuchmaschinen und Urheberrecht, Sind Thumbnails unerlässlich, sozial nützlich, aber rechtswidrig?, ZUM 2009, 345-354.

Paal, Boris P., KI-Training mit öffentlich frei zugänglichen Daten im Lichte der DS-GVO-Vorgaben, ZfDR 2024, 129-157.

Paal, Boris P./Pauly, Daniel A. (Hrsg.), Datenschutzgrund-Verordnung Bundesdatenschutzgesetz, 3. Aufl., München 2021.

Pauly, Daniel A./Nabulsi, Selma, Die Rechtmäßigkeit der Datenverarbeitung in der Due Diligence, Datenschutzrechtliche Fragen im Vorfeld des Unternehmenskaufs, ZD 2023, 519-523.

Peifer, Karl-Nikolaus, Kein Maulkorb für die Presse durch einseitige Veröffentlichungsverbote in Anwaltsschriftsätzen, Anmerkung zu BGH, Urt. v. 26.11.2019 – VI ZR 12/19, ZUM 2020, 342-345.

Peifer, Karl-Nikolaus/Kubis, Sebastian/Stieper, Malte/Raue, Benjamin (Hrsg.), Ius Vivum: Kunst – Internationales – Persönlichkeit, FS Schack, 2022.

Pesch, Paulina Jo, The case of LAION: The first public (German) court decision on text and data mining (TDM) in the context of machine learning, LTZ 2025, 72-77.

Pesch, Paulina Jo/Böhme, Rainer, Artpocalypse now? – Generative KI und die Vervielfältigung von Trainingsbildern, GRUR 2023, 997-1007.

Pesch, Paulina Jo/Böhme, Rainer, Verarbeitung personenbezogener Daten und Datenrichtigkeit bei großen Sprachmodellen, ChatGPT & Co. unter der DS-GVO, MMR 2023, 917-923.

Peukert, Alexander, Copyright in the Artificial Intelligence Act – A Primer, GRUR Int. 2024, 497-509.

Purtova, Nadezhda, The law of everything. Broad concept of personal data and future of EU data protection law, Law, Innovation and Technology 10 (2018), 40-81.

Radtke, Tristan, Gemeinsame Verantwortlichkeit unter der DSGVO, zugl. Dissertation, Baden-Baden 2021.

Radtke, Tristan, Das Urheberrecht als (KI-)Innovationsbremse in der Rechtswissenschaft?, Large Language Models in der rechtswissenschaftlichen Forschung, ZGE 17 (2025), 1-52.

Ramakrishnan, Raghu/Stolfo, Sal/Bayardo, Roberto/Parsa, Ismail (Hrsg.), Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD00: The Second Annual International Conference on Knowledge Discovery in Data in Boston Massachusetts USA, New York 2000.

Raue, Benjamin, Das subjektive Vervielfältigungsrecht – eine Lösung für den digitalen Werkgenuss?, ZGE 9 (2017), 514-538.

Raue, Benjamin, Die neue Urheberrechtsschranke des § 60d UrhG, CR 2017, 656-662.

Raue, Benjamin, Die geplanten Text und Data Mining-Schranken (§§ 44b und 60d UrhG-E), ZUM 2020, 172-175.

Raue, Benjamin/Schöf, Christof, Zugang zu großen Textkorpora des 20. und 21. Jahrhunderts mit Hilfe abgeleiteter Textformate, Versöhnung von Urheberrecht und textbasierter Forschung, RuZ 2020, 118-127.

Ray, Indrajit/Li, Ninghui/Kruegel, Christopher (Hrsg.), Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, CCS'15: The 22nd ACM Conference on Computer and Communications Security in Denver Colorado USA, New York 2015.

Reichert, Florian/Radtke, Kristina/Eske, Hermann, KI-Verordnung: Rechtsgrundlagen für die Bereitstellung und Nutzung von KI, Betrachtung des vom LfDI BW initiierten Diskussionspapiers, ZD 2024, 483-490.

Richter, Heiko, Anm. zu EuGH, Urt. v. 19.10.2016 – C-582/14, EuZW 2016, 912-914.

Richter, Philipp, Datenschutz zwecklos?, Das Prinzip der Zweckbindung im Ratsentwurf der DSGVO, DuD 2015, 735-740.

Roßnagel, Alexander, Datenschutz in der Forschung, Die neuen Datenschutzregelungen in der Forschungspraxis von Hochschulen, ZD 2019, 157-164.

Roßnagel, Alexander, Datenschutz im E-Learning, Die neuen Datenschutzregelungen im Lehrbetrieb von Hochschulen, ZD 2020, 296-302.

Roßnagel, Alexander, Anonymisierung personenbezogener Daten und Nutzung anonymer Daten, DuD 2024, 513-520.

Roßnagel, Alexander/Geminn, Christian, Vertrauen in Anonymisierung, Regulierung der Anonymisierung zur Förderung Künstlicher Intelligenz, ZD 2021, 487-490.

Sassenberg, Thomas/Faber, Tobias (Hrsg.), Rechtshandbuch Industrie 4.0 und Internet of Things, Praxisfragen und Perspektiven der digitalen Zukunft, 2. Aufl., München 2020.

Schack, Haimo, Schutzgegenstand, „Ausnahmen oder Beschränkungen" des Urheberrechts, GRUR 2021, 904-909.

Schack, Haimo, Vervielfältigung und/oder Bearbeitung oder freie Benutzung?, ZGE 15 (2023), 263-276.

Schack, Haimo, Auslesen von Webseiten zu KI-Trainingszwecken als Urheberrechtsverletzung de lege lata et ferenda, NJW 2024, 113-117.

Schäfer, Lena, Datenschutz-Compliance im KI-Training, Datenschutzkonformes KI-Training in der Praxis, ZD 2025, 12-17.

Schantz, Peter, Die Datenschutz-Grundverordnung, Beginn einer neuen Zeitrechnung im Datenschutzrecht, NJW 2016, 1841-1847.

Schmid, Gregor, EU-Urheberrecht – weltweit?, Zur Frage der extraterritorialen Wirkung der urheberrechtlichen Regelungen der KI-Verordnung, KIR 2025, 69-79.

Schöch, Christof/Döhl, Frédéric/Rettinger, Achim/Gius, Evelyn/Trilcke, Peer/Leinen, Peter/Jannidis, Fotis/Hinzmann, Maria/Röpke, Jörg, Abgeleitete Textformate: Prinzip und Beispiele, RuZ 2020, 160-175.

Schöch, Christof/Döhl, Frédéric/Rettinger, Achim/Gius, Evelyn/Trilcke, Peer/Leinen, Peter/Jannidis, Fotis/Hinzmann, Maria/Röpke, Jörg, Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen, 2020, 10.17175/2020_006.

Schöch, Christof/Gius, Evelyn/Trilcke, Peer/Ripoll, Élodie, Conference Reader: 2nd Annual Conference of Computational Literary Studies, 2023, 10.5281/ZENODO.8093598.

Schwartmann, Rolf/Jaspers, Andreas/Lepperhoff, Niels/Weiß, Steffen/Meier, Michael, Praxisleitfaden zum Anonymisieren personenbezogener Daten, Anforderungen, Einsatzklassen und Vorgehensmodell, 2022, https://stiftungdatenschutz.org/fileadmin/Redaktion/Dokumente/Anonymisierung_personenbezogener_Daten/SDS_Studie_Praxisleitfaden-Anonymisieren-Web_01.pdf.

Schwartmann, Rolf/Keber, Tobias/Zenner, Kai (Hrsg.), KI-VO Leitfaden, Heidelberg 2024.

Schwartmann, Rolf/Köhler, Moritz, Warum es auf Wahrscheinlichkeiten ankommt: Darum sind personenbezogene Daten auch in LLM geschützt, RDV 2024, 316-317.

Seemann, Michael, Künstliche Intelligenz, Large Language Models, ChatGPT und die Arbeitswelt der Zukunft, Working Paper Forschungsförderung Nr. 304, Hans Böckler Stiftung, September 2023.

Seidel, Hendrik, Der Personenbezug von Daten ist (weiterhin) relativ zu bestimmen – das EuG erinnert an „Breyer", DSB 2023, 212-214.

Senftleben, Martin, Grundprobleme des urheberrechtlichen Dreistufentests, GRUR Int. 2004, 200-211.

Sesing-Wagenpfeil, Andreas, Trainierte KI-Modelle als Vervielfältigungsstücke im Sinne des Urheberrechts, ZGE 16 (2024), 212-268.

Shalev-Shwartz, Shai/Ben-David, Shai, Understanding machine learning, From theory to algorithms, New York 2014.

Shokri, Reza/Stronati, Marco/Song, Congzheng/Shmatikov, Vitaly, Membership Inference Attacks against Machine Learning Models, 19. Oktober 2016, http://arxiv.org/pdf/1610.05820.

Siglmüller, Jonas/Gassner, Daniel, Softwareentwicklung durch Open-Source-trainierte KI – Schutz und Haftung, RDi 2023, 124-132.

Simitis, Spiros/Hornung, Gerrit/Spiecker gen. Döhmann, Indra (Hrsg.), Datenschutzrecht DS-GVO/BDSG, 2. Aufl., Baden-Baden 2025.

Smuha, Nathalie A. (Hrsg.), The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence, Cambridge 2025.

Song, Congzheng/Ristenpart, Thomas/Shmatikov, Vitaly, Machine Learning Models that Remember Too Much, in: Bhavani Thuraisingham, David Evans, Tal Malkin et al. (Hrsg.), Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS '17: 2017 ACM SIGSAC Conference on Computer and Communications Security in Dallas Texas USA, New York 2017, S. 587-601.

Specht-Riemenschneider, Louisa, Die neue Schrankenregelung für Text und Data Mining und ihre Bedeutung für die Wissenschaft, OdW 2018, 285-290.

Specht-Riemenschneider, Louisa/Mantz, Reto (Hrsg.), Handbuch Europäisches und deutsches Datenschutzrecht, Bereichsspezifischer Datenschutz in Privatwirtschaft und öffentlichem Sektor, München 2019.

Specht-Riemenschneider, Louisa/Wehde, Alexander, Forschungsdatenzugang, Rahmenbedingungen, Prinzipien und Leitlinien für einen privilegierten Zugang zu Daten für Forschung und Wissenschaft, ZGI 2022, 3-11.

Spies, Axel, Wann passt die DS-GVO auf KI?, MMR 2024, 289-290.

Spies, Ulrich, Zweckfestlegung der Datenverarbeitung durch den Verantwortlichen, Besteht eine datenschutzrechtliche Zweckbestimmung des Verantwortlichen trotz Zweckbenennung im Gesetz?, ZD 2022, 75-81.

Spindler, Gerald, Künstliche Intelligenz und Urheberrecht aus europäischer Perspektive, in: Karl-Nikolaus Peifer, Sebastian Kubis, Malte Stieper et al. (Hrsg.), Ius Vivum: Kunst – Internationales – Persönlichkeit, FS Schack, 2022, S. 340-352.

Spindler, Gerald/Schuster, Fabian (Hrsg.), Recht der elektronischen Medien, 4. Aufl., München 2019.

Steinrötter, Björn/Markert, Jette, Datenbezogene Vorgaben der KI-Verordnung, RDi 2024, 400-405.

Stevens, Jeremy, Datenqualität bei algorithmischen Entscheidungen, Überlegungen aus Anlass des Gutachtens der Datenethikkommission, CR 2020, 73-79.

Stieper, Malte/Denga, Michael, Die Reichweite des EU-Urheberrechts nach der KI-VO, GRUR 2024, 1473-1483.

Stürmer, Verena, Löschen durch Anonymisieren?, Mögliche Erfüllung der Löschpflicht nach Art. 17 DS-GVO, ZD 2020, 626-631.

Sucker, Reinhard, Der digitale Werkgenuss im Urheberrecht, zugl. Dissertation, Tübingen 2014.

Sydow, Gernot/Marsch, Nikolaus (Hrsg.), DS-GVO | BDSG, Datenschutz-Grundverordnung, Bundesdatenschutzgesetz Handkommentar, 3. Aufl., Baden-Baden, Zürich, Wien 2022.

Taeger, Jürgen/Gabel, Detlev (Hrsg.), DSGVO - BDSG - TDDDG, 4. Aufl., Frankfurt 2022.

Talke, Armin, Urheberrecht in Bildung, Wissenschaft und Kulturerbe, Die Erlaubnisse nach §§ 60a bis 60h UrhG, Passau 2022.

Thouvenin, Florent/Peukert, Alexander/Jaeger, Thomas/Geiger, Christophe (Hrsg.), Kreation Innovation Märkte - Creation Innovation Markets, Festschrift Reto M. Hilty, Berlin, Heidelberg 2024.

Thuraisingham, Bhavani/Evans, David/Malkin, Tal/Xu, Dongyan (Hrsg.), Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS '17: 2017 ACM SIGSAC Conference on Computer and Communications Security in Dallas Texas USA, New York 2017.

Thüsing, Gregor/Rombey, Sebastian, Anonymisierung an sich ist keine rechtfertigungsbedürftige Datenverarbeitung, Eine Auslegung von Art. 4 Nr. 2 DS-GVO nach den Methoden des EuGH, ZD 2021, 548-553.

Ulrich, Jürgen, Urheberrecht für Sachverständige, Teil I: Eigenes Urheberrecht, DSB 2011, 308-316.

Vaswani, Ashish/Shazeer, Noam/Parmar, Niki/Uszkoreit, Jakob/Jones, Llion/Gomez, Aidan N./Kaiser, Lukasz/Polosukhin, Illia, Attention Is All You Need, 12. Juni 2017, http://arxiv.org/pdf/1706.03762.

Veale, Michael/Binns, Reuben/Edwards, Lilian, Algorithms that remember: model inversion attacks and data protection law, Phil. Trans. R. Soc. A 376, 20180083, 10.1098/rsta.2018.0083.

Vesala, Juha, Developing Artificial Intelligence-Based Content Creation: Are EU Copyright and Antitrust Law Fit for Purpose?, IIC 2023, 351-380.

Volodina, Elena/Dobnik, Simon/Tiedemann, Therese Lindström/Vu, Xuan-Son, Grandma Karl is 27 years old, Research agenda for pseudonymization of research data, https://arxiv.org/pdf/2308.16109.

Wachter, Sandra/Mittelstadt, Brent/Russell, Chris, Do large language models have a legal duty to tell the truth?, R Soc Open Sci 11 (2024), 240197.

Wagner, Kristina, Generative KI: Eine „Blackbox" urheberrechtlicher Haftungsrisiken?, Balanceakt zwischen Innovationsförderung und effektivem Rechtsschutz für Werke Dritter, MMR 2024, 298-304.

Walker, Reinhard, Die richterliche Veröffentlichungspraxis in der Kritik, JurPC 1998, 34.

Wandtke, Artur-Axel, Urheberrecht in der Reform oder wohin steuert das Urheberrecht?, Widersprüche in den Reformen des Urheberrechts, MMR 2017, 367-373.

Wandtke, Artur-Axel/Bullinger, Winfried (Hrsg.), Praxiskommentar Urheberrecht, 6. Aufl., München 2022.

Weichert, Thilo, Die Forschungsprivilegierung in der DS-GVO, Gesetzlicher Änderungsbedarf bei der Verarbeitung personenbezogener Daten für Forschungszwecke, ZD 2020, 18-24.

Welser, Marcus von, Generative KI und Urheberrechtsschranken, GRUR-Prax 2023, 516-520.

Welser, Marcus von, Generative KI und Open Source Software, Darf Open Source Software für das Training von generativer KI verwendet werden?, GRUR 2024, 1406-1416.

Werry, Susanne, Generative KI-Modelle im Visier der Datenschutzbehörden, Technische Entwicklungen im Zusammenhang mit KI-Modellen begegnen einer Vielzahl datenschutzrechtlicher Herausforderungen, MMR 2023, 911-917.

Winter, Christian/Battis, Verena/Halvani, Oren, Herausforderungen für die Anonymisierung von Daten, Technische Defizite, konzeptuelle Lücken und rechtliche Fragen bei der Anonymisierung von Daten, ZD 2019, 489-493.

Wissenschaftliche Dienste des Bundestags, Künstliche Intelligenz und Machine Learning - Eine urheberrechtliche Betrachtung, WD 10 – 3000 – 67/18, 23. Oktober 2018.

Wolff, Heinrich Amadeus/Brink, Stefan/Ungern-Sternberg, Antje von (Hrsg.), BeckOK Datenschutzrecht, München 2024.

Wrobel, Mona/Pentzien, Simon, Personenbezug und LLMs: Datenschutzrechtliche Bewertung und Tipps für die Praxis, DSB 2024, 200-204.

Zech, Herbert, Information als Schutzgegenstand, Tübingen 2012.

Ziegenhorn, Gero, Anm. zu EuGH, Urt. v. 19.10.2016 - C-582/14, NVwZ 2017, 213-218.

Ziegenhorn, Gero/Schulz-Große, Stefanie, Der Grundsatz der Zweckbindung, Anforderungen an die Erhebung und weitere Verarbeitung personenbezogener Daten, ZD 2023, 581-587.