Synthetische KI-Lerndaten – Voraussetzungen für einen Personenbezug

Daniel Maslewski

Einleitung

Der zentrale Grundgedanke des europäischen Datenschutzrechts ist es, natürliche Personen vor einer beeinträchtigenden Datenverarbeitung zu schützen.[1] Diesen Schutz von natürlichen Personen gilt es auch bei der Verwendung und Entwicklung von KI-Modellen zu gewährleisten, da in diesem Bereich regelmäßig größere Datenmengen verarbeitet werden.[2] Die Qualität der KI-Systeme hängt dabei entscheidend von der zugrundeliegenden Datenqualität und Datenquantität ab, mithin auch von der Verwendung personenbezogener Daten.[3] Bei einer schlechten Datenqualität, welche unter anderem auch aus einer schlechten Datenquantität resultieren kann, können KI-Systeme unter Umständen fehleranfällig werden oder gar diskriminierende Muster annehmen.[4] Dies führt zu einem gewissen Spannungsverhältnis zwischen dem Datenschutz der Betroffenen auf der einen Seite und dem Interesse an einer effektiven Datennutzung für KI-Systeme auf der anderen Seite.[5] Die Schwierigkeit in der Praxis besteht vor allem darin, dieses Konfliktpotenzial zwischen „Daten-Hunger" und Datenschutz bestmöglich aufzulösen,[6] um so einen gerechten Ausgleich zwischen den Interessen herzustellen zu können. Eine Möglichkeit ist es, den Anwendungsbereich der DS-GVO durch die Verwendung von Daten ohne Personenbezug bzw. anonymisierten Daten auszuschließen, um so dieses Konfliktpotenzial grundsätzlich zu vermeiden. Nach Art. 2 Abs. 1 DS-GVO findet die DS-GVO nämlich nur dann Anwendung, wenn es um die Verarbeitung personenbezogener Daten geht. Eine Lösung hierfür

1 Vgl. Hornung/Spiecker gen. Döhmann,, in Simitis/Hornung/Spiecker Art. 1, Rn. 3; Datenschutzkonferenz (2018), S. 1.
2 Vgl. Valkanova 2020, S. 336, Rn. 1 f.
3 Vgl. ebd., S. 336, Rn. 2; Raji, DuD 2021, S. 304.
4 Vgl. Brink/Bäßler/Groß-Karrais 2021, S. 501, Rn. 28, ebd. S. 304.
5 Vgl. Paal 2020, S. 427 Rn. 1 f.; ebd. S. 304.
6 Vgl. ebd., S. 427, Rn. 1a f.

kann es beispielsweise sein, wenn für das Training von KI-Systemen sog. synthetische Daten verwendet werden.[7] Bei der Nutzung solcher Daten ist dann insbesondere entscheidend, unter welchen Voraussetzungen von einem fehlenden Personenbezug auch mit Blick auf Verwendung von innovativen Technologien ausgegangen werden kann und ob ein Personenbezug unter gewissen Voraussetzungen dennoch herstellbar ist.

Im Mittelpunkt des Beitrags steht daher die Frage, unter welchen Voraussetzungen synthetische KI-Lerndaten einen Personenbezug aufweisen. Zum Verständnis soll zunächst der Begriff der sog. „synthetischen Daten" näher erläutert und deren Herstellungsprozess kurz aufgezeigt werden (dazu Abschnitt „Synthetische KI-Lerndaten"). Hierauf aufbauend wird sich sodann mit dem Personenbezug bei synthetischen KI-Lerndaten beschäftigt (dazu Abschnitt „Voraussetzungen eines Personenbezugs synthetischer KI-Lerndaten"), wobei hier zunächst die allgemeinen Maßstäbe der DS-GVO zur Bestimmung des Personenbezugs herangezogen werden, um anschließend auf die synthetischen KI-Lerndaten eingehen zu können.

Synthetische KI-Lerndaten

Beim maschinellen Lernen, einem Teilgebiet der KI,[8] wird in der Regel eine große Menge an (personenbezogenen) Daten benötigt.[9] Hierfür werden sog. Trainingsdaten verwendet.[10] Darunter sind Daten zu verstehen, anhand derer die KI-Modelle trainiert werden.[11] Damit für den KI-Lernprozess ausreichend Daten zur Verfügung stehen, welche zur Entwicklung und Validierung von KI-Modellen benötigt werden, und gleichzeitig die datenschutzrechtlichen Interessen von natürlichen Personen hinreichend gewahrt werden, können sog. „synthetische Daten" verwendet werden.[12] Hierunter sind anonymisierte Datensätze zu verstehen, die möglichst nah an die Originaldaten herankommen, ohne dass dabei grundsätzlich eine Re-Identifizierung möglich ist, mithin kein Personenbezug besteht.[13] Bei synthetischen Daten handelt es sich dementsprechend um eine Art von Er-

7 Vgl. Raji, DuD 2021, S. 305; Schild, in BeckOK Datenschutzrecht, Art. 4, Rn. 27b.
8 Vgl. Kaulartz 2020, S. 462 Rn. 1.
9 Vgl. Niemann/Kevekordes 2020a, S. 18; Meyer, ZRP 2018, S. 233.
10 Vgl. Kaulartz 2020, S. 34 Rn. 11.
11 Vgl. ebd.
12 Vgl. Raji, DuD 2021, S. 305; Schild, in BeckOK Datenschutzrecht, Art. 4 Rn. 27b.
13 Vgl. Paal 2020, S. 439, Rn. 28; Schild, in BeckOK Datenschutzrecht, Art. 4, Rn. 27b.

satzdaten.[14] Zur Herstellung von synthetischen Daten gibt es eine Vielzahl möglicher Verfahren.[15] Beispielsweise können hierfür Generative Adversial Networks (GAN) verwendet werden, bei denen die Daten durch zwei konkurrierende neuronale Netzwerke – einem Diskriminator und einem Generator – erzeugt werden.[16] Da die beiden konkurrierenden Netzwerke in der Regel den gleichen Datensatz verwenden, wird grundsätzlich nur ein verhältnismäßig kleiner Datenpool an personenbezogenen Daten für die Generierung der neuen synthetischen Daten benötigt.[17] Ein wesentlicher Vorteil bei der Verwendung synthetischer Daten ist es insoweit, dass diese wegen ihres grundsätzlich fehlenden Personenbezugs in den Trainingsphasen beliebig verwendet oder produziert werden können und sie aufgrund ihrer Nähe zu den Originaldatensätzen für KI-Lernprozesse besonders geeignet sind.[18]

Voraussetzungen eines Personenbezugs synthetischer KI-Lerndaten

Es kann allerdings sein, dass synthetische KI-Lerndaten unter gewissen Umständen dennoch einen Personenbezug aufweisen.[19] Unter welchen Voraussetzungen ein Personenbezug vorliegt, die natürlichen Personen also identifiziert oder identifizierbar sind, steht in unmittelbarem Zusammenhang mit der Frage, ob es sich bei den Lerndaten um anonyme Informationen handelt.[20] Nach dem EwGr. 26 S. 5 gelten die Grundsätze des Datenschutzes nicht für anonyme Informationen, konkret also für Informationen, die sich nicht auf eine identifizierte oder identifizierbare natürliche Person beziehen, oder solche personenbezogenen Daten, die in einer Weise anonymisiert worden sind, dass die betroffene Person nicht oder nicht mehr identifiziert werden kann. Demzufolge hängt der rechtliche Status von synthetischen Daten entscheidend von dem Vorliegen eines Personenbezugs ab.[21] Die DS-GVO enthält allerdings keine speziellen Vorschriften

14 Vgl. Meents 2020, S. 457, Rn. 47.
15 Vgl. Drechler/Jentzsch 2018, S. 11.
16 Vgl. hierzu ausführlich Meents 2020, S. 457, Rn. 47; Raji, DuD 2021, S. 305.
17 Vgl. Meents 2020, S. 457, Rn. 47.
18 Vgl. Datenethikkommission 2019, S. 132; Raji, DuD 2021, S. 305 f.
19 Vgl. Paal 2020, S. 439, Rn. 28
20 Vgl. ebd. S. 439, Rn. 28; Raji, DuD 2021, S. 306; Schild, in BeckOK Datenschutzrecht, Art. 4, Rn. 27c.
21 Vgl. Drechsler/Jentzsch 2018, S. 19.

für KI-Sachverhalte, weshalb sich die Bestimmung des Personenbezugs nach den allgemeinen Vorschriften der Verordnung richtet, die im Wege der Auslegung anzuwenden sind.[22] Es kommt somit letztlich auf die DS-GVO an, welche konkreten Anforderungen an einen Personenbezug zu stellen sind, wobei hierfür insbesondere die Begriffsbestimmung der personenbezogenen Daten aus Art. 4 Nr. 1 DS-GVO heranzuziehen ist.

Anforderungen der DS-GVO an einen Personenbezug

Als personenbezogene Daten im Sinne des Art. 4 Nr. 1 DS-GVO sind alle Informationen zu verstehen, die sich auf eine identifizierte oder identifizierbare natürliche Person beziehen (sog. betroffene Person). Voraussetzung ist somit neben dem Vorliegen von Informationen über eine natürliche Person insbesondere, dass es sich um eine konkret bestimmbare Person handelt.

1. *Informationen über natürliche Personen*

Die Begriffsbestimmung des personenbezogenen Datums in Art. 4 Nr. 1 DS-GVO ist sehr weit zu verstehen, sodass grundsätzlich sämtliche Informationen umfasst sind, die eine natürliche Person betreffen.[23] Insofern gibt es auch bei der automatisierten Verarbeitung kein belangloses Datum.[24] Bereits aus dem Wortlaut der Norm ergibt sich, dass es sich um Informationen über eine natürliche Personen handeln muss, mit der Folge, dass Informationen über juristische Personen in der Regel nicht erfasst werden.[25] Unter Umständen kann jedoch auch bei juristischen Personen ein Personenbezug bestehen. Dies kann insbesondere dann der Fall sein, wenn aufgrund ihrer tatsächlichen Ausgestaltung (z. B. als Ein-Mann-GmbH) Rückschlüsse auf die hinter der Gesellschaft stehende natürliche Person möglich sind.[26]

22 Vgl. Niemann/Kevekordes 2020b, S. 184; Paal 2020, S. 427, Rn. 1f.
23 Vgl. Arning/Rothkegel, in Taeger/Gabel, Art. 4, Rn. 5; Klar/Kühling, in Kühling/Buchner, Art. 4, Nr. 1, Rn. 8.
24 Vgl. BVerfG, 15.2.1983, 1 BvR 209/83 – NJW 1984, S. 422.
25 Vgl. Arning/Rothkegel, in Taeger/Gabel, Art. 4, Rn. 16.
26 Vgl. ebd., Art. 4, Rn. 17.

2. *Bestimmbarkeit der betroffenen Person*

Eine wesentliche Voraussetzung für einen Personenbezug ist, dass die natürliche Person bestimmbar und damit identifiziert oder identifizierbar ist.

a) *Identifizierte Person*
Als identifiziert gilt eine natürliche Person, wenn sich ihre Identität unmittelbar aus der Information ergibt.[27] Die Information muss sich also einer Person objektiv eindeutig zuordnen lassen,[28] sodass sich diese von einer Personengruppe abgrenzbar hervorhebt.[29] Ein Beispiel hierfür ist ein aussagekräftiger Name einer Person, der eine Wiedererkennung der Person ermöglicht und sich somit von anderen Namen klar abgrenzt.[30] Die Frage nach der Identifizierbarkeit einer Person wird dabei allerdings stets von den jeweiligen Umständen abhängen und muss daher im konkreten Einzelfalls beurteilt werden.[31]

b) *Identifizierbare Person*
Wesentlich schwieriger wird sich hingegen regelmäßig die Beantwortung der Frage gestalten, wann eine Person identifizierbar ist. Von einer Identifizierbarkeit ist grundsätzlich auszugehen, wenn eine Information über eine natürliche Person für sich alleine genommen zur Identifizierung noch nicht ausreicht, sondern Rückschlüsse auf sie erst durch weitere zusätzliche Informationen möglich sind.[32] In Art. 4 Nr. 1 Hs. 2 DS-GVO werden als solche zusätzlichen Informationen beispielhaft die Zuordnung zu einer Kennung oder zu weiteren besonderen Merkmalen genannt.[33] Bei der Prüfung, ob eine natürliche Person identifizierbar ist, sind gemäß des EwGr. 26 S. 3 sämtliche 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. Nach dem EwGr. 26

27 Vgl. EuGH, 19.10.2016, C-582/14 – ZD 2017, S. 25; Arning/Rothkegel, in Taeger/Gabel, Art. 4, Rn. 24; Borges, in BeckOK IT-Recht, Art. 4, Rn. 10; Karg, in Simitris/Hornung/Spiecker, Art. 4, Nr. 1, Rn. 54; Klar/Kühling, in Kühling/Buchner, Art. 4, Nr. 1, Rn. 18.
28 Vgl. Karg, in Simitris/Hornung/Spiecker, Art. 4, Nr. 1, Rn. 54.
29 Vgl. Arning/Rothkegel, in Taeger/Gabel, Art. 4, Rn. 24.
30 Vgl. ebd.
31 Vgl. ebd. Art. 4, Rn. 25; Karg, in Smitris/Hornung/Spiecker, Art. 4 Nr.1, 54.
32 Vgl. Klar/Kühling, in Kühling/Buchner, Art. 4, Nr. 1, Rn. 19.
33 Vgl. Arning/Rothkegel, in Taeger/Gabel, Art. 4, Rn. 30; Borges, in BeckOK IT-Recht, Art. 4, Rn. 12.

S. 4 sind bei der Feststellung, dass die Mittel nach „allgemeinem Ermessen wahrscheinlich" genutzt werden, insbesondere auch objektive Faktoren zu berücksichtigen, wie etwa die Kosten der Identifizierung sowie der dafür erforderliche Zeitaufwand. Daneben sind die zum Zeitpunkt der Verarbeitung verfügbare Technologie und technologische Entwicklungen zu berücksichtigen. Der Personenbezug ist damit mittels einer Risikoanalyse hinsichtlich der Identifizierungswahrscheinlichkeit festzustellen.[34] Auf wessen Wissen und Mittel es bei Beurteilung der Identifizierbarkeit ankommt und welche konkreten Anforderungen an die erforderlichen Zusatzinformationen zu stellen sind, ist jedoch umstritten.[35] Hierzu haben sich verschiedene Ansätze herausgebildet.

– *Objektiver Ansatz*
  Nach dem objektiven Ansatz sind zur Beurteilung der Identifizierbarkeit sämtliche Informationen und Mittel zu berücksichtigen, die irgendeine Person oder Stelle hat.[36] Erfasst werden sollen zudem auch rechtswidrige Zugriffe auf die Datensätze der verantwortlichen Stelle.[37] Das bedeutet, dass die Personenbezogenheit dadurch grundsätzlich für jedes Datum rein objektiv feststellbar ist.[38] Für diesen Ansatz spreche insbesondere der Umstand, dass die Verordnung unter Heranziehung des EwGr. 26 von einem weiten und damit absoluten Verständnis ausgehe.[39] Insoweit wird sich auf dessen Wortlaut gestützt, wonach „[…] alle Mittel berücksichtigt werden, die von dem Verantwortlichen oder einer anderen Person nach allgemeinem Ermessen wahrscheinlich genutzt werden […]"[40].
– *Subjektiver Ansatz*
  Nach dem subjektiven Ansatz hingegen kommt es bei der Beurteilung der Identifizierbarkeit nur auf die Sphäre und die Mittel der verantwortlichen Stelle an und dementsprechend darauf, ob diese die natürliche Person unter einem vertretbaren Aufwand identifizieren

34 Vgl. Klar/Kühling, in Kühling/Buchner, Art. 4, Nr. 1, Rn. 22.
35 Vgl. Arning/Rothkegel, in Taeger/Gabel, Art. 4, Rn. 33; Bergt, ZD 2015, S. 365; Borges, in BeckOK IT-Recht, Art. 4, Rn. 14.
36 Vgl. Breyer, ZD 2014, S. 404 f.; Buchner, DuD 2016, S. 156.; Düsseldorfer Kreis 2014, S. 12.
37 Vgl. Arning/Rothkegel, in Taeger/Gabel, Art. 4, Rn. 33; Auer-Reinsdorff/Conrad, in Auer-Reinsdorff/Conrad, § 34, Rn. 95; Brink/Eckhardt, ZD 2015, S. 206.
38 Vgl. Brink/Eckhardt, ZD 2015, S. 206.
39 Vgl. Buchner, DuD 2016, S. 156.
40 Vgl. ebd.

kann.[41] Dies ergebe sich bereits aus dem EwGr. 30, welcher die Möglichkeit aufzeige, dass IP-Adressen „in Kombination mit einer eindeutigen Kennung" eine natürliche Person identifizieren könne und damit im Ergebnis als personenbezogene Daten zu klassifizieren seien, ohne dass hiervon trotz Zuordnungsmöglichkeit des Accessproviders stets ausgegangen werde.[42] Darüber hinaus spreche für ein relatives Verständnis, dass der EwGr. 26 S. 3 eine nach „allgemeinen Ermessen wahrscheinliche" Nutzung verlange, welche nur dann anzunehmen sei, wenn eine dritte Person beispielsweise infolge einer Übermittlung der Daten durch die verantwortliche Stelle in Kontakt mit den Informationen komme.[43] Es bedürfe somit eines konkreten Bezuges zum Verantwortlichen.[44]

– *Breyer-Urteil des EuGH*
Auch der EuGH hat sich bereits in einem Vorabentscheidungsverfahren zur Datenschutz-Richtlinie[45] mit den Anforderungen an einen Personenbezug auseinandergesetzt. Darin hatte der Gerichtshof zu entscheiden, ob und inwieweit eine dynamische IP-Adresse für einen Anbieter von Online-Mediendiensten ein personenbezogenes Datum darstellt.[46] Nach Auffassung des Gerichtshofs liegt für den Anbieter ein personenbezogenes Datum vor, wenn er über rechtliche Mittel verfügt, mithilfe deren er auf Zusatzinformationen des Internetzugangsanbieters zugreifen kann, um so die betreffende Person identifizieren zu können.[47] Mit seiner Breyer-Entscheidung hat sich der Gerichtshof insoweit an einem subjektiven Ansatz orientiert, wobei er in seiner Urteilsbegründung gleichzeitig klargestellt hat, dass der Wortlaut des EwGr. 26 ein Indiz dafür darstelle, dass es nicht erfor-

41 Vgl. OLG Hamburg, MMR 2011, S. 282; Redeker, in IT-Recht, Rn. 1012; Schmitz, in Spindler/Schmitz, § 13, Rn. 13; Schulz, in Roßnagel, § 11 TMG, Rn. 23; Schulz, in Gola/Heckmann, § 46, Rn. 15. Ähnlich Klar/Kühling, in Kühling/Buchner, Art. 4, Nr. 1, Rn. 26.

42 Vgl. Schulz, in Gola/Heckmann, § 46, Rn. 15.

43 Vgl. Klar/Kühling, in Kühling/Buchner, Art. 4, Nr. 1, Rn. 26.

44 Vgl. ebd.

45 Vgl. Richtlinie 95/46/EG des Europäischen Parlaments und Rates vom 24. Oktober 1995 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten zum freien Datenverkehr. Am 25. Mai 2018 mit Wirksamkeit der DS-GVO außer Kraft getreten.

46 Vgl. EuGH, 19.10.2016, C-582/14 – MMR 2016, S. 842.

47 Vgl. ebd. S. 844.

derlich sei, wenn sich die zur Identifikation erforderlichen Informationen ausschließlich bei einer Stelle befinden.[48]

– *Vermittelnder Ansatz*
Auf Grundlage der Rechtsprechung des EuGH hat sich inzwischen ein weiterer vermittelnder Ansatz herausgebildet, bei welchem zur Bestimmung der Identifizierbarkeit sowohl subjektive als auch objektive Elemente einen entsprechenden Niederschlag finden.[49] Danach muss der Verantwortliche sich auch das Zusatzwissen Dritter zurechnen lassen, wenn er über Mittel verfügt, um dieses Zusatzwissen nutzen zu können und eine solche Nutzung zugleich wahrscheinlich ist.[50] Für diesen Ansatz spreche zum einen der EwGr. 26 S. 3, der hinsichtlich der Identifizierbarkeit einer Person nicht nur auf Mittel des Verantwortlichen abstelle, sondern explizit auch Mittel anderer Personen einbeziehe, sofern eine Verwendung dieser Mitteln nach allgemeinem Ermessen wahrscheinlich sei.[51] Zum anderen lege der EwGr. 30 auch den Schluss nahe, dass Online-Kennungen mit entsprechendem Zusatzwissen eine Identifizierung eben nur erlauben können.[52] Im Übrigen müsse eine teleologische Auslegung erfolgen, da andernfalls eine Anonymisierung, also die Aufhebung des Personenbezugs[53], für den Verantwortlichen faktisch nicht möglich sei.[54] Zur Feststellung, ob eine Person identifizierbar ist, sind bei dem erforderlichen Zusatzwissen Dritter neben den faktischen Mitteln darüber hinaus auch die rechtlichen Mittel einzubeziehen.[55] Bei der Nutzung von rechtlichen Mittel soll dann auf den konkreten Einzelfall unter Beachtung der tatsächlichen Nutzungswahrschein-

48 Vgl. ebd. S. 843.
49 Vgl. Arning/Rothkegel, in Taeger/Gabel, Art. 4, Rn. 35; Gola, in Gola, Art. 4, Rn. 18; Schantz, in Schantz/Wolff, Rn. 279. Ähnlich Brink/Eckhardt 2015, S. 205, 210 f.; Klar/Kühling, in Kühling/Buchner, Art. 4, Nr. 1, Rn. 26.
50 Vgl. Arning/Rothkegel, in Taeger/Gabel, Art. 4, Rn. 35; Brink/Eckhardt, ZD 2015, S. 210 f.
51 Vgl. Arning/Rothkegel, in Taeger/Gabel, Art. 4, Rn. 35; Brink/Eckhardt, ZD 2015, S. 209.
52 Vgl. ebd.
53 Vgl. Schild, in BeckOK Datenschutzrecht, Art. 4, Rn. 15f.
54 Vgl. Arning/Rothkegel, in Taeger/Gabel, Art. 4, Rn. 35.
55 Vgl. EuGH, 19.10.2016, C-582/14 – MMR 2016, S. 843f.; Arning/Rothkegel, in Taeger/Gabel, Art. 4, Rn.37; Borges, in BeckOK IT-Recht, Art. 4, Rn. 18.

lichkeit abgestellt werden.[56] Auch sind hierbei nach dem EwGr. 26 S. 4 die zum Verarbeitungszeitpunkt maßgebliche Technologie sowie technischen Fortschritte einzubeziehen, was unter Umständen die Gefahr einer nachträglichen Re-Identifizierbarkeit nach sich ziehen kann.[57] Auf Grundlage des EwGr. 26 S. 4, welcher bei der Wahrscheinlichkeit einer Nutzung der Mittel alle objektiven Faktoren einbezieht, soll nach überwiegender Auffassung zugleich auch der Einsatz illegaler Mittel zur Identifizierung berücksichtigt werden.[58]

– *Ergebnis*

Mit Blick auf das Breyer-Urteil des EuGH vermag der vermittelnde Ansatz zu überzeugen, da hierbei sowohl Elemente des subjektiven als auch des objektiven Ansatzes berücksichtigt werden. Bei der Feststellung des Personenbezugs sind demzufolge alle Informationen zu berücksichtigen, die mit vernünftigen (rechtlichen) Mitteln erlangt werden können, wobei nach dem EwGr. 26 S. 4 die zum Verarbeitungszeitpunkt maßgebliche Technologie sowie technische Fortschritte einzubeziehen sind.

Der objektive und der subjektive Ansatz lassen aufgrund ihrer extremen Positionen für sich gesehen wesentliche Aspekte ungeachtet. So würde eine strikte Anwendung des subjektiven Ansatzes dem Wortlaut des EwGr. 26 S. 3 zuwiderlaufen, da dieser ausdrücklich bei den zu berücksichtigen Mitteln „von dem Verantwortlichen oder einer anderen Person" spricht. Der Verordnungsgeber hat damit explizit neben dem Verantwortlichen weitere Personen in seiner Formulierung aufgenommen, sodass eine andere Auffassung nur schwer zu überzeugen vermag. Aber auch der objektive Ansatz verkennt, dass es nach dem EwGr. 26 S. 3 gerade der Wahrscheinlichkeit einer Nutzung der Mittel bedarf, was zumindest bei einem Abstellen auf irgendeine beliebige Person nur schwer zu begründen sein wird. Für einen vermittelnden Ansatz spricht hingegen insbesondere der Sinn und Zweck der DS-GVO, wonach der Schutz von natürlichen Personen bei der Verarbeitung personenbezogener Daten im Vorder-

56 Vgl. Arning/Rothkegel, in Taeger/Gabel, Art. 4, Rn.3; Borges, in BeckOK IT-Recht, Art. 4, Rn. 19.

57 Vgl. Borges, in BeckOK IT-Recht, Art. 4, Rn. 21.

58 Vgl. Borges, in BeckOK IT-Recht, Art. 4, Rn. 20; Ernst, in Paal/Pauly, Art. 4, Rn. 13; Klabunde, in Ehmann/Selmayr, Art. 4, Rn. 17; Klar/Kühling, in Kühling/Buchner, Art. 4, Nr. 1, Rn. 29. Andere Ansicht: Karg, in Simitis/Hornung/Spiecker, Art. 4, Rn. 64.

> grund steht, vgl. Art. 1 Abs. 1 DS-GVO. Für einen effektiven Schutz Betroffener vor einer beeinträchtigenden Datenverarbeitung wird es in der Regel nicht ausreichen, wenn ausschließlich auf das Wissen und die Mittel des Verantwortlichen abgestellt wird. Es bedarf vielmehr zusätzlich auch einer Berücksichtigung der Umstände außerhalb der Sphäre des Verantwortlichen, da vor allem durch die stetig voranschreitende Digitalisierung sowie die Entwicklung neuer Technologien die Möglichkeiten für einen Rückgriff auf das Zusatzwissen Dritter steigen, was zugleich die Identifizierungswahrscheinlichkeit erhöht. Dies ist bei der Auslegung entsprechend zu berücksichtigen. Gleichzeitig darf ein zu weites Verständnis des Personenbezugs jedoch nicht dazu führen, dass Informationen generell als personenbezogen zu werten sind, was einen Ausschluss des Anwendungsbereichs faktisch unmöglich machen würde.
> Überzeugend ist insoweit ein Mittelweg, mit welchem versucht wird, die Schutzinteressen der Betroffenen und die Interessen der Verantwortlichen an der Nutzung der Daten unter Berücksichtigung der datenschutzrechtlichen Vorgaben bestmöglich in Ausgleich zu bringen.

Herstellung eines Personenbezugs durch Re-Identifizierung

Das dargelegte Verständnis zur Identifizierbarkeit ist auch bei der Beantwortung der Frage, unter welchen Voraussetzungen synthetische KI-Lerndaten einen Personenbezug aufweisen können, zugrunde zu legen. Bei der Risikoanalyse hinsichtlich der Identifizierungswahrscheinlichkeit sind somit alle Informationen zu berücksichtigen, die mit vernünftigen (rechtlichen) Mitteln erlangt werden können, wobei nach EwGr. 26 S. 4 auch die zum Zeitpunkt der Verarbeitung verfügbare Technologie und technologische Entwicklungen eine wesentliche Rolle spielen.

Die Gefahr einer Re-Identifikation bei synthetischen KI-Lerndaten kann daher vor allem aufgrund der wachsenden Informationsmenge und der stetigen Weiterentwicklung neuer leistungsfähiger KI-Modelle nicht generell ausgeschlossen werden.[59] Obwohl synthetische KI-Lerndaten in der Regel zunächst keinen Personenbezug aufweisen, kann wegen ihrer Nähe zu den Originaldatensätzen unter Umständen die Gefahr bestehen, dass

59 Vgl. Hornung/Wagner, CR 2019, S. 568; Niemann/Kevekordes, CR 2020a, S. 20.

eine Re-Identifizierung der Datensubjekte, welche zur Herstellung der synthetischen Daten verwendet wurden, möglich ist und eine Gewährleistung der Anonymität infrage stellt.[60] Entscheidend für die Wahrscheinlichkeit einer Re-Identifizierung ist daneben auch die Wahl des Verfahrens zur Gewinnung synthetischer Daten.[61] So kann bei teilweise synthetisierten Daten das Risiko einer Re-Identifizierung größer sein als bei einer vollständigen Synthetisierung der Datensätze, da bei einer teilweisen Synthetisierung ein Teil der Daten gerade nicht verändert wird und das Individuum bereits im originalen Datenpool vorhanden war.[62] Auch eine Zusammenführung zweier Datensätze kann das Risiko einer Re-Identifikation steigern.[63] Es wird dementsprechend bei der Beurteilung eines Personenbezugs für Verantwortliche vor allem darauf ankommen, inwieweit eine Re-Identifikation synthetischer Daten unter Hinzuziehung der derzeit zur Verfügung stehenden innovativen Mittel möglich ist, was letztlich im Einzelfall festgestellt und regelmäßig an den aktuellen Stand der Technik angepasst werden muss.

Es können allerdings auch Maßnahmen ergriffen werden, um einer Identifizierungswahrscheinlichkeit zu entgegnen. Ein präventiv technischer Ansatz kann es beispielsweise sein, Datensätze vor neuen Zusatzinformationen abzuschotten, um dadurch so das Risiko der Identifizierbarkeit zu vermindern.[64] Aber auch eine regelmäßige Analyse von Angriffsszenarien auf Datensätze sowie eine Weiterentwicklungen im Bereich von De-Anonymisierungstechniken zählen zu den technischen Ansätzen.[65] Daneben werden zudem präventive rechtliche Maßnahmen diskutiert, wie etwa eine Selbstverpflichtung Verantwortlicher anonymisierte Daten nicht zu re-identifizieren.[66] Ein weiterer Vorschlag ist es, durch den Europäischen Datenschutzausschuss oder durch die Schaffung von Codes of Conduct Konkretisierungen der DS-GVO in diesem Bereich vorzunehmen.[67]

60 Vgl. Datenethikkommission 2019, S. 132; Paal 2020, S. 439, Rn. 28; Raji, DuD 2021, S. 306.
61 Vgl. Drechsler/Jentzsch 2018, S. 10.
62 Vgl. ebd.
63 Vgl. ebd., S. 19f.; Hornung/Wagner, CR 2019, S. 568.
64 Vgl. ebd. S. 571
65 Vgl. Hornung/Wagner, CR 2019, S. 571.
66 Vgl. Hornung/Wagner, CR 2019, S. 573.
67 Vgl. ebd. S. 573.

Fazit

Mit voranschreitendem technologischen Fortschritt wird auch bei der Entwicklung und Verbesserung von KI-Modellen voraussichtlich der Bedarf an (personenbezogenen) Daten zunehmen. Zur Steigerung des Schutzes Betroffener vor einer beeinträchtigenden Datenverarbeitung können allerdings synthetische KI-Lerndaten eine sinnvolle Alternative darstellen. Die Voraussetzungen, unter denen synthetische KI-Lerndaten einen Personenbezug aufweisen können, richten sich in Ermangelung anderer gesetzlicher Regelungen nach den Grundsätzen und Vorschriften der DS-GVO. Es kommt also vor allem darauf an, wie der Begriff des personenbezogenen Datums ausgelegt und verstanden wird. Unter Zugrundlegung des vermittelnden Ansatzes wird die Identifizierbarkeit von natürlichen Personen und damit die Herstellung eines Personenbezugs derzeit grundsätzlich weit verstanden. Dieses Verständnis gilt es auch bei der Prüfung zu Grunde zu legen, ob synthetische KI-Lerndaten einen Personenbezug aufweisen. Wesentlich ist daher, dass bei der vorzunehmenden Risikoanalyse hinsichtlich der Identifizierungswahrscheinlichkeit alle Informationen berücksichtigt werden, die mit vernünftigen (rechtlichen) Mitteln erlangt werden können. Hierbei müssen als Mittel insbesondere auch die technologischen Entwicklungen (z. B. im Bereich der Künstlichen Intelligenz) berücksichtigt werden, was dazu führt, dass die Wahrscheinlichkeit einer Re-Identifizierung nicht generell ausgeschlossen werden kann und die Identifizierungsgefahr mit zunehmendem technologischen Fortschritt in der Regel steigen kann. Es bedarf somit im konkreten Einzelfall einer Wahrscheinlichkeitsprüfung, ob und inwieweit die Möglichkeit einer Re-Identifizierung tatsächlich besteht und ob dadurch synthetische KI-Lerndaten einen Personenbezug aufweisen.

Literatur

Auer-Reinsdorff, Astrid/Conrad, Isabell (Hrsg.) (2019): Handbuch IT- und Datenschutzrecht. München: C. H. Beck.

Brink, Stefan/Wolff, Heinrich Amadeus (Hrsg.) (2022): Beck'scher Online-Kommentar Datenschutzrecht. München: C. H. Beck.

Brink, Stefan/Eckhardt, Jens (2015): Wann ist ein Datum ein personenbezogenes Datum? – Anwendungsbereich des Datenschutzrechts. In: Zeitschrift für Datenschutz 1/2015, S. 205-212.

Borges, Georg/Hilber, Marc (Hrsg.) (2022): Beck'scher Online-Kommentar IT-Recht. München: C. H. Beck.

Bergt, Matthias (2015): Die Bestimmbarkeit als Grundproblem des Datenschutzrechts – Überblick über den Theorienstreit und Lösungsvorschlag. In: Zeitschrift für Datenschutz Ausgabe Nr.8/2015, S. 365-371.

Breyer, Patrick (2014): Personenbezug von IP-Adressen – Internetzugang und Datenschutz. In: Zeitschrift für Datenschutz 8/2014, S. 400-405.

Buchner, Benedikt (2016): Grundsätze und Rechtmäßigkeit der Datenverarbeitung unter der DS-GVO. In: Datenschutz und Datensicherheit – DuD, Nr. 40, S. 155-161.

Datenethikkommission der Bundesregierung (2019): Gutachten der Datenethikkommssion. Online: https://www.bundesregierung.de/breg-de/service/publikationen/gutachten-der-datenethikkommission-langfassung-1685238 (letzter Zugriff: 30.1.2023).

Datenschutzkonferenz (2018): Kurzpapier Nr. 18: Risiko für die Rechte und Freiheiten natürlicher Personen. Online: https://www.datenschutzkonferenz-online.de/media/kp/dsk_kpnr_18.pdf (letzter Zugriff: 05.06.2023).

Drechsler, Jörg/Jentzsch, Nicola (2018): Synthetische Daten, Innovationspotenzial und gesellschaftliche Herausforderungen, Stiftung Neue Verantwortung 2018. Online: https://www.stiftung-nv.de/sites/default/files/synthetische_daten.pdf (letzter Zugriff: 31.01.2023).

Düsseldorfer Kreis (2014): Orientierungshilfe – Cloud Computing. Online: https://www.datenschutzkonferenz-online.de/media/oh/20141009_oh_cloud_computing.pdf (letzter Zugriff: 05.06.2023).

Ehmann, Eugen/Selmayr, Martin (Hrsg.) (2018): DS-GVO – Datenschutz-Grundverordnung Kommentar. München: C. H. Beck.

Gola, Peter/Heckmann, Dirk (Hrsg.) (2019): Bundesdatenschutzgesetz Kommentar. München: C. H. Beck.

Hornung, Gerrit/Wagner, Bernd (2019): Der schleichende Personenbezug – Die Zwickmühle der Re-Identifizierbarkeit in Zeiten von Big Data und Ubiquitous Computing. In: Computer und Recht, Nr. 9/2019, S. 565-574.

Kaulartz, Markus (2020): Personenbezug von KI-Modellen. In: Kaulartz, Markus/Braegelmann, Tom (Hrsg.): Rechtshandbuch Artificial Intelligence und Machine Learning. München: C. H. Beck, S. 464-477.

Kaulartz, Markus (2020): Trainieren von Machine-Learning-Modellen. In: Kaulartz, Markus/Braegelmann, Tom (Hrsg.): Rechtshandbuch Artificial Intelligence und Machine Learning. München: C. H. Beck, S.32-36.

Kühling, Jürgen/Buchner, Benedikt (Hrsg.) (2020): Datenschutz-Grundverordnung. München: C. H. Beck.

Meents, Jan Geert (2020): Datenschutz durch KI. In: Kaulartz, Markus/Braegelmann, Tom (Hrsg.): Rechtshandbuch Artificial Intelligence und Machine Learning. München: C. H. Beck, S. 445-461.

Meyer, Stephan (2018): Künstliche Intelligenz und die Rolle des Rechts für Innovation. In: Zeitschrift für Rechtspolitik, S. 233-238.

Niemann, Fabian/Kevekordes, Fabian (2020a): Machine Learning und Datenschutz (Teil 1). In: Computer und Recht, Band 36, H. 1/2020, S. 17-25.

Niemann, Fabian/Kevekordes, Fabian (2020b): Machine Learning und Datenschutz (Teil 2). In: Computer und Recht, Band 36, H. 3, S. 179-184.

Paal, Boris (2020): Spannungsverhältnis von KI und Datenschutzrecht. In: Kaulartz, Markus/Braegelmann, Tom (Hrsg.): Rechtshandbuch Artificial Intelligence und Machine Learning. München: C. H. Beck, S. 427-444.

Raji, Behrang (2021): Rechtliche Bewertung synthetischer Daten für KI-Systeme. In: Datenschutz und Datensicherheit – DuD, Nr. 45, S. 303-309.

Roßnagel, Alexander (Hrsg.) (2013): Beck'scher Kommentar zum Recht der Telemediendienste. München: C. H. Beck.

Redeker, Helmut (2020): IT-Recht. München: C. H. Beck.

Schantz, Peter/Wolff, Heinrich Amadeus (2017): Das neue Datenschutzrecht, Datenschutz-Grundverordnung und Bundesdatenschutzgesetz in der Praxis. München: C. H. Beck.

Simitris, Spiros/Hornung, Gerrit/Spiecker gen. Döhmann, Indra (Hrsg.) (2021): Datenschutzrecht, DSGVO mit BDSG. München: C. H. Beck.

Spindler, Gerald/Schmitz, Peter/Liesching, Marc (2018): Telemediengesetz mit Netzwerkdurchsetzungsgesetz Kommentar. München: C. H. Beck.

Taeger, Jürgen/Gabel, Detlev (Hrsg.) (2022): Kommentar DSGVO – BDSG – TTDSG, Frankfurt am Main: Deutscher Fachverlag GmbH.

Valkanova, Monika (2020): Trainieren von KI-Modellen. In: Kaulartz, Markus/Braegelmann, Tom (Hrsg.): Rechtshandbuch Artificial Intelligence und Machine Learning. München: C. H. Beck, S. 336-351.