Zweiter Teil
Die Maschine lernt – auch anhand urheberrechtlich geschützter Daten?

Das auf viele urheberrechtliche Sachverhalte zutreffende Gleichnis „Zwerge auf den Schultern von Riesen“[114] beschreibt Modelle maschinellen Lernens ziemlich genau. Ohne Trainingsdaten ist selbst das beste (trainingsdatenbasierte) Modell nicht zu verwenden,[115] und mit der Qualität und Quantität dieser Daten steigen und fallen die Qualität der Ergebnisse sowie die Nutzbarkeit des Modells. Auf das Gleichnis übertragen sind die Modelle vergleichsweise kleine Gebilde – Zwerge – die auf monumentalen Datensammlungen – Riesen – aufsetzen. Aus dieser Perspektive wird die Relevanz der Trainingsdaten besonders deutlich.

Im Zuge dessen stellen sich die Fragen, was im Trainingsprozess eigentlich mit den Trainingsdaten passiert: Werden etwa urheberrechtlich relevante Handlungen vorgenommen? Inwiefern könnten die Trainingsdaten im fertigen Modell enthalten sein? Liegt eine Bearbeitung oder eine freie Benutzung vor, oder findet das Urheberrecht hier gar keine Anwendung? Wie stehen die Text- und Data-Mining-Schranken § 60d UrhG bzw. § 44b UrhG-E zu maschinellen Lernprozessen?

Von der Diskussion in dieser Arbeit ausgenommen ist die Frage, wie eine Sammlung von Trainingsdaten selbst geschützt sein kann. Die für alle Text- und Data-Mining-Verfahren gleichermaßen vorzunehmenden Handlungen zur Sammlung der Daten bzw. Datenkorpuserstellung im Vorfeld bedürfen keiner gesonderten Betrachtung für Zwecke des Machine Learnings. Hier dürften die allgemeinen Regeln des Datenbankschutzes nach § 4 Abs. 2 UrhG und §§ 87a ff. UrhG anwendbar sein, ohne dass sich aus der Natur des Machine Learning-Kontextes Besonderheiten ergeben.[116]

---

114 Ursprung vermutlich in einem Brief von *Peter von Blois* an *Reginald, Bishop of Bath*, Epistolae 92 (1180), zitiert in *Merton*, On The Shoulders Of Giants, S. 216 f..

115 Es sei denn, es handelt sich um ein trainingsdatenloses selbstlernendes Modell, das allein anhand von Regeln und Reinforcement lernt, wie z. B. AlphaGo, vgl. https://deepmind.com/research/case-studies/alphago-the-story-so-far (Stand: 22.02.2021), das als Input lediglich das Go-Feld bekommt.

116 Den Schutz als Datenbankwerk gem. § 4 Abs. 2 UrhG eher ablehnend, für eine Anwendung von §§ 87a ff. UrhG und auch weitere Schutzmöglichkeiten diskutierend *Hacker*, GRUR 2020, 1025, 1028.

§ 3 Urheberrechtlich relevante Vorgänge im maschinellen Lernprozess

Zunächst erfolgt eine Analyse der technischen Vorgänge des Einlesens und des Trainings anhand der eingelesenen Daten (Trainingsdaten), um daraus auf eventuell urheberrechtlich relevante Vorgänge zu schlussfolgern. Daran schließt sich eine Darstellung der geltenden und der im Rahmen der Umsetzung der DSM-Richtlinie zu erwartenden Schranken an, um eine Einordnung aus der Perspektive unterschiedlicher Nutzergruppen vornehmen zu können.

A. Technische Beleuchtung des maschinellen Lernprozesses

Um die eingangs gestellten Fragen zu beantworten, erscheint es dienlich, den ablaufenden Prozess genauer unter die Lupe zu nehmen und dabei den Fokus auf die Trainingsdaten zu legen. Dabei ist freilich eine Betrachtungsweise zu finden, die so abstrakt ist, dass sie auf eine Vielzahl unterschiedlicher ML-Modelle anwendbar ist, und gleichzeitig konkret genug, um die urheberrechtlichen Fragen zufriedenstellend beantworten zu können.

Der Prozess des maschinellen Lernens lässt sich grob in vier Phasen unterteilen, die in Abbildung 3.1 gezeigt sind. Die erste Phase umfasst die Sammlung und Organisation der Daten, die hier nicht weiter thematisiert werden. Wenn die Datensammlung abgeschlossen ist, werden die Daten entweder direkt in das den Trainingsprozess ausführende Programm oder zunächst in ein anderes Programm eingelesen, dass die Daten für den Trainingsprozess anpasst und aufbereitet. Daran schließt sich der eigentliche Trainingsprozess an, mit dem Ziel, ein trainiertes Modell zu produzieren (hier abgebildet sind zwei Möglichkeiten, die auch im weiteren Verlauf diskutiert werden: Modelle in sog. Baumstrukturen und in sog. Netzwerkstrukturen). Wenn das Modell trainiert ist, kann es eingesetzt bzw. für den Einsatz bereitgestellt werden.

Nicht abgebildet, weil für die Fragestellung in diesem Kapitel nicht relevant, ist die Entwicklung des Modells, das die Trainingsdaten später verwendet. Diese Entwicklung kann vor der Trainingsdatensammlung stattfinden (sodass gezielt nach Daten für das Modell gesucht werden kann), oder anhand vorhandener Daten erfolgen und mitunter auch während des Trainingsprozesses iterativ Anpassungen erfordern.

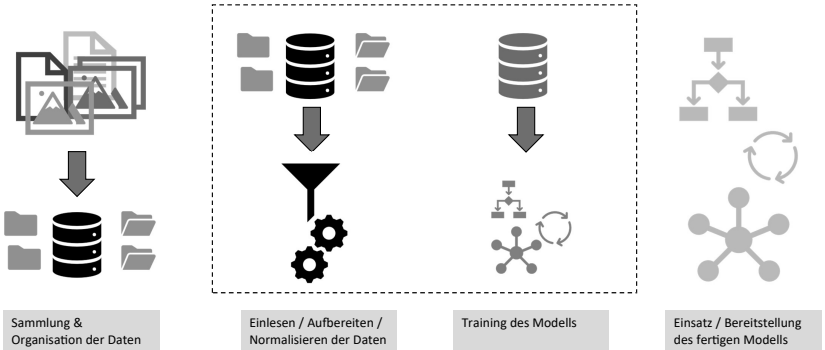

Abbildung 3.1: Phasen des Machine Learning-Prozesses, eigene Darstellung.

Die folgenden Ausführungen beschränken sich auf die Einlese- und Trainingsvorgänge.

I. Einlesevorgang und Analyse bzw. Aufbereitung der Trainingsdaten

Grundsätzlich wird im Einlesevorgang jeder Datensatz der Datenbasis einzeln auf die gleiche Weise analysiert (und nicht etwa die Trainingsdatensammlung als Ganzes).[117] Welche Analyseschritte vorgenommen werden, ist davon abhängig, für welche Vorgehensweise der Entwickler sich entscheidet, weshalb diese Vorbereitung bzw. Aufbereitung der Trainingsdaten für jedes ML-Modell unterschiedlich gestaltet sein kann. Beispiele für Analyseschritte sind das Einlesen eines Bildes in ein von dem Modell verarbeitbares Dateiformat oder ein Zusammenschneiden des Bildes auf eine bestimmte Größe, sowie die Bereinigung der Daten von sogenannten „Ausreißern", die aufgrund ungewöhnlicher Werte zu Verzerrungen führen könnten.[118]

117 Vgl. z. B. die Vorgehensweise des Einlesens von Trainingsdatenbildern in `image_dataset.py` ab Zeile 211 bzw. `paths_and_labels_to_dataset(...)` und `path_to_image(...)`, https://github.com/tensorflow/tensorflow/blob/v2.4.1/tensorflow/python/keras/preprocessing/image_dataset.py (Stand: 22.02.2021).

118 *Osinga*, Deep Learning Cookbook, S. 22 f.; *Johnstone*, What it Means to „Clean" Data and „Train" Machine Learning Algorithms.

1. Einlesen einzelner oder aller Datensätze

Die Trainingsdatensammlung kann in unterschiedlichster Form vorliegen: Beispielsweise als eine Vielzahl von Dateien in Ordnern (jede Datei entspricht dann einem Trainingsdatum) oder auch als einzelne Datei (etwa ein Textdokument[119] oder eine MySQL-Datenbank[120]).

Der Unterschied ist aber für die urheberrechtliche Analyse unerheblich, da es letztendlich auf den Zustand vor dem Beginn und nach dem Abschluss des Einlesevorgangs aller Daten ankommen wird.

2. Umwandlung der Daten in ein maschinenlesbares Format

Damit ein Machine Learning-Modell mit der Trainingsdatensammlung arbeiten kann, muss es auf diese Daten zugreifen können. Die zu nutzende Datei wird dafür vorübergehend in den Arbeitsspeicher geladen. Für das Training von KNN wird der Input in einen Tensor[121] oder eine andere Repräsentationsform der Daten „übersetzt". Die folgenden Ausführungen beschränken sich auf die Vorgehensweise mit Tensoren. Für eine Bilddatei mit einer Dimension von 20 mal 20 Pixeln bedeutet das, dass ein Tensor entsteht, der für jedes Pixel drei Werte (RGB) enthält.[122] Oder anders gesagt: es entsteht im Arbeitsspeicher ein Konstrukt ähnlich einem Setzkasten mit drei Etagen, wobei jede Etage aus 20 mal 20 Feldern besteht. Die oberste Etage enthält die Rotwerte (R), die mittlere die Grünwerte (G) und die untere die Blauwerte (B). Angenommen, das Pixel in der oberen linken Ecke des Bildes hat den Farbwert „255, 242, 0"[123] (gelb, vgl. Abbildung 3.2), dann erhält das Feld der obersten Schicht in der oberen linken Ecke den Wert „255", das Feld der Grünschicht den Wert „242" und und das Feld der Blauschicht den Wert „0". Der Machine Learning-Algorithmus arbeitet nur auf der Tensor-Version des Inputs.

119 Dann könnte etwa jedes Wort oder jeder Satz ein Trainingsdatum darstellen.

120 MySQL ist ein Datenbankformat, das bspw. ein Backup einer Datenbank in einer einzigen Datei speichern kann, vgl. z. B. https://dev.mysql.com/doc/refman/8.0/en/mysqldump.html (Stand: 22.02.2021).

121 Tensoren sind Vektoren mit mehreren Dimensionen, vgl. z. B. *Osinga*, Deep Learning Cookbook, S. 1.

122 Der Tensor hätte also die Form (20,20,3) – vgl. https://www.tensorflow.org/tutorials/load_data/images#visualize_the_data (Stand: 22.02.2021).

123 In der hier gewählten Schreibweise ist Schwarz definiert als 0, 0, 0 und Weiß als 255, 255, 255.

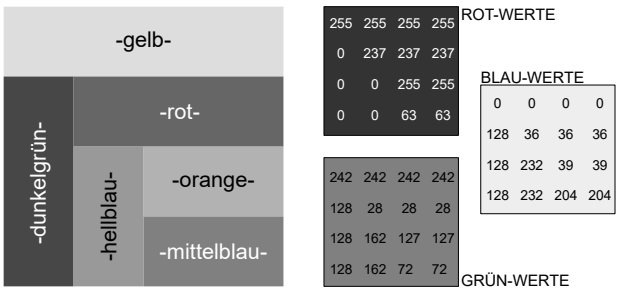

Abbildung 3.2: Bild mit 4x4 Pixel, eigene, stark vergrößerte Darstellung.

Abbildung 3.2 veranschaulicht die Umwandlung der Werte. Links ist ein Ausgangsbild – also ein Beispiel für mögliche Trainingsdaten – abgebildet,[124] rechts die Farbwerte in den „Setzkästen", die ausgelesen werden. Aus den ermittelten Werten könnte das Ausgangsbild wiederhergestellt werden, die Werte sind lediglich eine maschinenlesbare Transformation des Ausgangsbildes, die dann im Trainingsvorgang verwendet wird. Die Art der Repräsentation ist stark abhängig von der Art der verwendeten Ausgangsdaten und dem geplanten Vorgehen im Trainingsprozess.

Ein Blick in die Praxis offenbart, dass etwa eine in *TensorFlow* bzw. *Keras* übliche Vorgehensweise nicht einen Tensor für jedes einzulesende Bild erzeugt, sondern einen `image_batch`-Tensor der Form (<Anzahl Bilder>,<Anzahl Pixel vertikal>, <Anzahl Pixel horizontal>, <Farbkanäle>), der die Werte aller eingelesenen Bilder enthält.[125] Auf das Beispielbild in 3.2 übertragen hätte der Tensor also die Form (1, 20, 20, 3). Dies stellt jedoch lediglich eine Möglichkeit dar, Daten einzulesen, die Umsetzung ist von den Vorlieben des Entwicklers und der eingesetzten Technologie sowie der Beschaffenheit der Trainingsdaten abhängig.

Im Anschluss an die Umwandlung in das Tensor-Format erfolgen möglicherweise noch weitere Anpassungsschritte: Für KNN empfiehlt *TensorFlow* beispielsweise die Umwandlung der RGB-Werte, die zwischen 0 und 255

124 Die Farbnamen dienen nur der Veranschaulichung der Pixelwerte.

125 Dies ist der Fall bei Verwendung der Funktion `image_dataset_from_directory(...)`, vgl. https://www.tensorflow.org/tutorials/load_data/images (Stand: 22.02.2021). Diese Funktion erzeugt ein `Dataset`-Objekt, das unter anderem den `image_batch`-Tensor und den `labels_batch`-Tensor enthält.

liegen, in eine Repräsentation die sich zwischen 0 und 1 bewegt.[126] Dadurch entsteht ein neues `Dataset`, bzw. ein neuer Tensor, der die angepassten Werte enthält.

II. Trainingsvorgang

Im Rahmen des Trainingsvorgangs werden die zuvor eingelesenen und vorbereiteten Daten im Modell durchlaufen. An dieser Stelle wird von der Wiedergabe mathematischer Funktionen zur Erklärung des Trainingsvorgangs abgesehen, da diese nicht dem Verständnis zuträglich wären. Stattdessen erfolgt eine abstrakte Beschreibung.

Im Rahmen eines typischen Trainingsvorgangs laufen die folgenden Schritte ab (der genaue Prozess ist davon abhängig, was mit welchen Mitteln mit dem Modell erreicht werden soll):[127]

Abhängig von der Wahl und der Art des ML-Modells wird zunächst jeweils ein Tensor eingelesen und dann durch das Modell „geleitet".

Für den Fall von KNN bedeutet das, dass für jeden Wert des eingelesenen Datums ein Inputneuron den Wert übergeben bekommt, ggf. analysiert und je nach Ergebnis der Analyse und Architektur des Netzes das Ergebnis an alle oder ausgewählte Neuronen der nächsten Schicht weiterleitet. Das weitergeleitete Ergebnis ist mithin unter anderem abhängig von der dem Neuron zugewiesenen Funktion, bzw. den dem Neuron zugewiesenen Werten für *Gewicht* und *Bias* (vgl. oben § 2 B.II.2.). Weitergegeben wird also nicht ein Teil des Trainingsdatums, sondern das Ergebnis einer Berechnung. Die Neuronen der letzten Schicht repräsentieren das Ergebnis des Durchlaufs: so könnte für überwachtes Lernen jedes Neuron der letzten Schicht einem Label entsprechen. Das Neuron, das in der letzten Schicht aktiviert wird, entspricht dem Label, das das KNN im aktuellen Zustand dem eingelesenen Trainingsdatum zuweisen würde. Um den Lerneffekt zu erzielen, erfolgt ein Abgleich des errechneten Ergebnisses mit dem im Voraus dem Trainingsdatum zugeordneten Label. Bei Abweichungen nimmt ein Lernalgorithmus Anpassungen der variablen Werte in den Neuronen im KNN vor. Dieser Optimierungsvorgang erfolgt so lange, bis entweder zum Beispiel eine gewisse

126 Vgl. https://www.tensorflow.org/tutorials/load_data/images#standardize_the_data (Stand: 22.02.2021).

127 Vgl. z. B. Beschreibung auf https://pytorch.org/tutorials/beginner/blitz/neural_networks_tutorial.html#sphx-glr-beginner-blitz-neural-networks-tutorial-py (Stand: 22.02.2021).

Fehlerquote unterschritten wird oder eine zuvor definierte maximale Anzahl Trainingsvorgänge durchgeführt wurde.

Die Einzelheiten dieser Vorgehensweise sind abhängig vom eingesetzten System, jedoch hier nicht weiter relevant.[128] Bemerkenswert ist, dass die gesamte Vorgehensweise einen Optimierungsprozess darstellt, in dem statistische Funktionen zum Einsatz kommen und in dem nicht das gesamte Trainingsdatum gleichmäßig verwertet wird, sondern indem lediglich anhand der eingelesenen Trainingsdaten Optimierungen der Werte im Modell vorgenommen werden.

B. Urheberrechtliche Relevanz des Einlese- und des Trainingsvorgangs

Auch in urheberrechtlicher Hinsicht ist zwischen dem Einlesevorgang und dem Trainingsvorgang zu unterscheiden. Im Rahmen des ersten könnten urheberrechtlich erhebliche Vervielfältigungs- und Bearbeitungshandlungen angenommen werden (dazu sogleich), während zweite nur mit Referenzen bzw. Repräsentationen der bereits eingelesenen Daten arbeitet und diese in der Regel nicht weiter vervielfältigt. Auch für das TDM wurde festgestellt, dass die eigentliche automatisierte Auswertung der Daten keinen urheberrechtlich relevanten Vorgang darstellt[129] (wobei davon ausgegangen wird, dass – sobald die Daten vorbereitet sind – im Rahmen der Auswertung keine Vervielfältigungen mehr vorgenommen werden).[130]

I. Vervielfältigung, § 16 UrhG

Im Rahmen des zuvor beschriebenen Einlesevorgangs[131] könnten gem. §§ 15 Abs. 1 Nr. 1, 16 UrhG dem Urheber vorbehaltene Vervielfältigungen entstehen. Ansatzpunkte für mögliche Vervielfältigungen sind das Einlesen der Daten in den Arbeitsspeicher, die sich daran anschließende Umwandlung

128 Eine mathematische und praktische Einführung gibt z. B. *Nielsen*, Neural Networks and Deep Learning; außerdem gut verständlich *Patterson/Gibson*, Deep Learning, S. 27 ff.; *Goodfellow* et al., Deep Learning Handbuch, S. 107 ff..

129 BT-Drs. 18/12329, S. 40.

130 Wandtke/Bullinger–*Bullinger*, PK UrhR, § 60d Rn. 15; Dreier/Schulze–*Dreier*, UrhG, § 60d Rn. 4; Erw.Gr. 8 DSM-RL nennt als u. U. erlaubnispflichtigen Vorgang „während des Vorgangs des Text und Data Mining" die Normalisierung von Daten.

131 S. dazu § 3 A.I..

in einen oder mehrere Tensoren sowie etwaige Anpassungen der Werte der Tensoren.

Vervielfältigungen sind „körperliche Festlegungen, die geeignet sind, das Werk den menschlichen Sinnen auf irgendeine Weise unmittelbar oder mittelbar wahrnehmbar zu machen".[132] Nach dieser aus der Gesetzesbegründung hervorgehenden Definition der Vervielfältigung kommt es lediglich darauf an, dass die Festlegung geeignet ist, das Werk wahrnehmbar zu machen, und nicht darauf, ob dies auch beabsichtigt ist. Der Zweck der Vervielfältigung ist also nicht zu berücksichtigen.

Die Behandlung der vorübergehenden Festlegung im Arbeitsspeicher als Vervielfältigung ist umstritten.[133] An dieser Stelle kann die Einordnung jedoch möglicherweise dahinstehen, wenn in einem der weiteren Schritte des Einlesevorgangs eine Vervielfältigung erfolgt.

Zu klären ist, ob die körperliche Festlegung des Werkes in Form der Umwandlung in einen Tensor eine Vervielfältigung darstellt. Diese Umwandlung erfolgt einzig zu dem Zweck, das Trainingsdatum für den ML-Prozess nutzbar zu machen. Bezweckt ist mithin nicht, dass diese umgewandelte Form nach Abschluss des Prozesses für den Menschen sichtbar ist. Wie jedoch eingangs festgestellt, ist der Zweck der Festlegung unerheblich, solange diese geeignet ist, das Werk den menschlichen Sinnen wahrnehmbar zu machen.

Grundsätzlich könnten die Tensoren in die ursprüngliche Form der Trainingsdaten zurückgewandelt werden. Es ist gerade das Ziel des Umwandlungsvorgangs, die Trainingsdaten in der Gestalt der Tensoren einzeln ansteuern zu können, und dabei alle relevanten Informationen, die in den Trainingsdaten enthalten sind, in Tensoren abzulegen. Besonders deutlich wird die Beziehung zwischen Originaldatum und Tensor bei der Umwandlung von Bildern: Hier werden regelmäßig die einzelnen Pixel untersucht und in den Tensoren abgelegt.[134] Es ist möglich, den Einlesevorgang umzukehren und das Bild aus den Tensoren wieder sichtbar zu machen.[135] Mithin sind die Tensoren

132 BT-Drs. IV/270, S. 47.

133 Vgl. OLG Hamburg ZUM 2001, 512, 513 – *Roche Lexikon Medizin*; Dreier/Schulze–*Schulze*, UrhG, § 16 Rn. 13; Loewenheim–*Loewenheim*, Handbuch Urheberrecht, § 20 Rn. 11; eine Vervielfältigung ablehnend KG ZUM 2002, 828, 830: Der Zustand im Arbeitsspeicher sei von der „Fortdauer der Energieversorgung des Rechners" abhängig, weshalb nicht von einer körperlichen Festlegung zu sprechen sei.

134 Vgl. dazu schon § 3 A.I.2..

135 Vgl. z. B. die Diskussion dazu hier https://github.com/pytorch/pytorch/issues/30655 (Stand: 22.02.2021).

mittelbar dazu geeignet, die eingelesenen Daten für die menschlichen Sinne wahrnehmbar zu machen, eine Vervielfältigung liegt also vor.[136]

Zudem erfolgen möglicherweise Anpassungen der Werte in den Tensoren, um die Verarbeitung durch den gewählten Modelltyp zu optimieren, wie etwa die zuvor beschriebene Umwandlung der RGB-Werte in Werte zwischen 0 und 1, wofür ein neues `Dataset`-Objekt erzeugt wird.[137] Nach dem Abschluss dieses Vorgangs existieren drei Repräsentationen der Trainingsdaten: die Originaldateien, der aus den eingelesenen Daten erzeugte Tensor, sowie der Tensor mit den angepassten Werten. Beide Tensoren sind unter Verwendung entsprechender *TensorFlow*-Funktionen dazu geeignet, die Originalbilder darzustellen.[138] Insofern liegen also – in dieser angenommenen Vorgehensweise – zwei Vervielfältigungsstücke von allen Trainingsdaten vor. Die Anzahl tatsächlich erfolgender Vervielfältigungen ist von der konkreten Umsetzung durch den Entwickler abhängig. Es ist aber davon auszugehen, dass im Rahmen des Einlesevorgangs mindestens eine Vervielfältigung erfolgt,[139] nämlich wenn der Tensor der Trainingsdaten angelegt wird.

Dass im sich an die Einlese- und Umwandlungsvorgänge anschließenden Trainingsprozess keine urheberrechtlich relevanten Vervielfältigungen entstehen, wurde bereits eingangs festgestellt.[140]

II. Bearbeitung oder Umgestaltung, § 23 S. 1 UrhG

In dem dargestellten Einlese- und Umwandlungsprozess könnte nicht nur eine Vervielfältigung, sondern auch eine Bearbeitung gem. § 23 S. 1 UrhG entstehen. Im Rahmen einer Bearbeitung erfolgt eine Anpassung des Werks an andere Nutzungsformen (geläufig sind vor allem Übersetzungen von Schriftwerken in andere Sprachen, Verfilmungen etc.), wobei die Identität des Ausgangswerkes im Gegensatz zur Umgestaltung nicht verändert wird.[141]

---

136 Im Ergebnis so auch *Deutscher Bundestag*, Künstliche Intelligenz und Machine Learning. Eine urheberrechtliche Betrachtung, S. 7.

137 Vgl. dazu § 3 A.I.2..

138 Vgl. https://www.tensorflow.org/tutorials/load_data/images#visualize_the_data_2 (Stand: 22.02.2021).

139 So auch *Deutscher Bundestag*, Künstliche Intelligenz und Machine Learning. Eine urheberrechtliche Betrachtung, S. 7; *Spindler*, GRUR 2016, S. 1113.

140 Vgl. § 3 B., bei Fußnote 129.

141 Vgl. BT-Drs. IV/270, S. 51; Dreier/Schulze–*Schulze*, UrhG, § 23 Rn. 6.

1. Dem Original dienende Funktion, Anpassung an andere Nutzungsformen

Zunächst müssten die im Rahmen des Einlese- und Umwandlungsvorgangs entstandenen Repräsentationen der Trainingsdaten eine „dem Original dienende Funktion" erfüllen, bzw. eine Anpassung an andere Nutzungsformen darstellen im Sinne einer Erweiterung der Verwertungsmöglichkeiten der Originale.[142] Um bei dem Beispiel der Bilddatei zu bleiben: die ursprüngliche Nutzungsform könnte die Wiedergabe auf einem Bildschirm sein, die neue Nutzungsform der Einsatz des Bildes zum Training eines Machine Learning-Modells, wofür die Anpassung in Form der Umwandlung in einen Tensor erforderlich ist. Hinzu kommt als Kriterium für eine Bearbeitung eine „dem Original dienende Funktion" dergestalt in Betracht, dass die Bearbeitung eine Verwendung des Werks ermöglicht, die über die bisherige Verwendung hinausgeht und einen weiteren Verwertungsmöglichkeitenkreis eröffnet.[143] Dies ist hier der Fall: Das Bild, das bisher lediglich der Verwendung als Bilddatei zugänglich war (z. B. zur Betrachtung oder Einbindung in Präsentationen oder Dokumente) kann nun dazu beitragen, ein Modell maschinellen Lernens zu beeinflussen, indem das Modell im Trainingsvorgang die Werte der in Tensoren eingelesenen Trainingsdaten berücksichtigt. Die Umwandlung der Trainingsdaten in Tensoren (sowie auch die anschließenden Anpassungen für spezielle ML-Modelltypen) sind also nicht nur Vervielfältigungen, sondern erfüllen auch grundsätzlich die Anforderungen an eine Bearbeitung.[144]

2. Freie Benutzung?

Auch wenn die grundsätzlichen Voraussetzungen einer Bearbeitung vorliegen, könnte eventuell auch eine freie Benutzung im Sinne des § 24 UrhG vorliegen. Dafür müsste jedoch ein neues, eigenständiges, schutzfähiges Werk entstehen, wobei das Originalwerk lediglich als Anregung dienen darf.[145] Hier entspricht jedoch der Tensor dem Ausgangswerk, die Identität des Originalwerks wird mithin nicht verändert, sondern der Algorithmus „sieht" genau das, was

142 Vgl. BT-Drs. IV/270, S. 51; Dreier/Schulze–*Schulze*, UrhG, § 23 Rn. 5.
143 Dreier/Schulze–*Schulze*, UrhG, § 23 Rn. 6.
144 Anders *Deutscher Bundestag*, Künstliche Intelligenz und Machine Learning. Eine urheberrechtliche Betrachtung, S. 7, die nur darauf abstellen, dass lediglich eine Formatänderung vorliege, durch die das Werk grundsätzlich nicht verändert wird.
145 Dreier/Schulze–*Schulze*, UrhG, § 24 Rn. 5, 7.

ein Mensch sähe, wenn er das Bild betrachtete, man könnte gar von einer Transformation sprechen. Eine freie Benutzung gem. § 24 UrhG liegt mithin in Bezug auf den Einlese- bzw. Umwandlungsvorgang nicht vor.[146]

3. Einwilligungserfordernis bei Veröffentlichung oder Verwertung

Zu beachten ist weiterhin, dass § 23 S. 1 UrhG eine Einwilligung des Urhebers nur für den Fall erfordert, dass die Bearbeitung veröffentlicht oder verwertet werden soll; eine Herstellung einer Bearbeitung kann demnach auch ohne Einwilligung erfolgen,[147] sofern es sich bei dem Original nicht um ein Datenbankwerk[148] oder ein Computerprogramm handelt,[149] und sofern die Bearbeitung oder Umgestaltung nicht veröffentlicht bzw. verwertet werden soll.

a) Bearbeitung der Trainingsdatenbank?

Ob die Trainingsdaten einzeln eingelesen und umgewandelt werden oder ob – wie in § 3 A.I.2. beschrieben – alle Trainingsdaten gemeinsam eingelesen und in einen einzigen Tensor (bzw. in zwei Tensoren: In der Regel werden die Trainingsdaten in Trainings- und Validierungsdaten unterteilt) eingelesen werden, ist wiederum von der individuellen Umsetzung abhängig. Wenn, wie beschrieben, alle Trainingsdaten eingelesen werden, und die Trainingsdatensammlung die Anforderungen an ein Datenbankwerk erfüllt,[150] ist die Bearbeitung nach dem soeben Gesagten unabhängig von der Anschlussnutzung einwilligungspflichtig. Eine Differenzierung zwischen dem Gesamtbestand der Trainingsdaten und den einzelnen Bestandteilen ist

146 Im Diskussionsentwurf, der § 24 UrhG aufhebt, wird in § 23 UrhG die Voraussetzung aufgenommen, dass eine Bearbeitung oder andere Umgestaltungen insbesondere dann unfrei sind, wenn sie keinen hinreichenden Abstand zum verwendeten Werk wahren (§ 23 Abs. 1 S. 1 UrhG-DiskE), sodass diese Voraussetzung dann die Prüfung des ehemaligen § 24 UrhG erübrigt, vgl. *BMJV*, DiskE 06/2020.

147 Vgl. auch Dreier/Schulze–*Schulze*, UrhG, § 23 Rn. 16.

148 Dreier/Schulze–*Schulze*, UrhG, § 23 Rn. 24; im Diskussionsentwurf auch explizit in § 23 Abs. 2 Nr. 4 UrhG-DiskE genannt.

149 Dreier/Schulze–*Schulze*, UrhG, § 23 Rn. 19.

150 Ob die Trainingsdatensammlung als Datenbankwerk zu qualifizieren ist, ist einzelfallbezogen zu klären.

jedoch entbehrlich, wenn auch die Bearbeitung der einzelnen Bestandteile geschieht und einwilligungspflichtig ist.

b) Veröffentlichung oder Verwertung der Bearbeitung

Zu prüfen ist, ob im Anschluss an die Umwandlung im Rahmen des Trainings- bzw. Produktiveinsatzes maschineller Lernverfahren üblicherweise eine relevante Veröffentlichung oder Verwertung der Bearbeitung erfolgt.

aa) Veröffentlichung

Eine Veröffentlichung liegt vor, wenn die Bearbeitung anderen Menschen als der Person des Bearbeiters zugänglich gemacht wird.[151] Die Verarbeitung des Tensors erfolgt in der Regel lediglich modellintern, die Bearbeitung bzw. Umgestaltung tritt dabei nicht aus dem Modell heraus. Nach Abschluss des Trainingsvorgangs ist es zwar nicht unüblich, dass das Modell veröffentlicht wird, allerdings nur in einer Form von Computerprogrammcode, Hyperparametern und errechneter Parameter, die nicht ohne weiteres einen Rückschluss auf die einzelnen Tensoren bzw. die Trainingsdaten zulässt. Eine Veröffentlichung der Bearbeitung dergestalt, dass die Bearbeitung für (andere) Menschen wahrnehmbar zugänglich gemacht wird, ist also nicht anzunehmen.

bb) Verwertung

Möglicherweise liegt in der Verwendung der Umwandlung in Gestalt der Tensoren jedoch eine Verwertung. Welche Handlungen den Verwertungsrechten unterfallen, ergibt sich aus § 15 UrhG. Dazu zählen demnach u. a. Vervielfältigung und Verbreitung, aber auch andere denkbare Verwertungs- und Nutzungsarten,[152] letztendlich also alles, was über die reine Rezeption hinausgeht. Es ist nicht auszuschließen, dass im weiteren Verlauf der Vorbereitung des Trainingsvorgangs zumindest einige Vervielfältigungen der Bearbeitung erfolgen, jedoch hängt dies von der individuellen Umsetzung

151 Dreier/Schulze–*Schulze*, UrhG, § 23 Rn. 17.
152 Dreier/Schulze–*Schulze*, UrhG, § 23 Rn. 18.

durch den Entwickler ab. Insbesondere wenn die Aufbereitung der Trainingsdaten in einem vorgelagerten Prozess erfolgt, die Umwandlungen also für eine spätere Verwendung gespeichert und sodann erneut eingelesen werden, schließen sich Vervielfältigungen an die Bearbeitung an.

c) Zwischenergebnis

In der Konsequenz wäre dann die Umwandlung in das Tensorformat in jedem Fall einwilligungspflichtig, wenn dem Trainingsdatenbestand insgesamt Datenbankwerkqualität zukommt, und in der Regel einwilligungspflichtig, wenn auf die einzelnen Trainingsdaten abzustellen ist und kein Ausnahmetatbestand greift. Ein solcher liegt nach § 23 S. 3 UrhG beispielsweise vor, wenn die Bearbeitung bzw. Umgestaltung im Rahmen eines Vorgangs des Text- und Data Mining (§ 60d UrhG) erfolgt. Auf die Ausnahmetatbestände wird in § 4 eingegangen.

III. Zusammenhang zwischen Trainingsdaten und Output generativer Modelle

Angedacht werden könnte auch, ob im Falle von generativen Modellen,[153] mit Output, der dem Format der Trainingsdaten entspricht (also zum Beispiel die Erzeugung von Bildern, wenn als Trainingsdaten auch Bilder verwendet werden), dieser Output eine Bearbeitung oder gar Vervielfältigung der Trainingsdaten darstellt. Nach dem Charakter des Trainingsvorgangs scheidet insbesondere eine Vervielfältigung schon deshalb aus, weil das Ergebnis nicht aus einer schlichten Kopie der Trainingsdaten entsteht, sondern diese lediglich als Datenbasis für komplexe Berechnungen verwendet werden.

Zudem ist zu unterscheiden zwischen Trainingsdaten, die eingesetzt werden, um das Modell zu trainieren, und eventuellen Inputdaten, die dem fertigen Modell zugeführt werden. Dies wird besonders deutlich am Beispiel des *Google DeepDream*-Systems: In der fertig trainierten Variante kann dem System durch den Benutzer ein Bild zur Verfügung gestellt werden, das daraufhin durch das Modell analysiert wird. Der Output gleicht dann in der Regel sehr stark dem Input, mit dem Unterschied, dass *DeepDream* Formen und Figuren in das Bild „hineininterpretiert“ (vgl. für ein Beispiel

153 Vgl. oben § 2 B.II.3..

Abbildung 10.2 und weitere Erläuterungen zum System in § 10 C.I.).[154] Die Trainingsdaten wirken sich nur insofern aus, als sie beeinflussen, welche Formen und Figuren in dem vom Benutzer gestellten Bild erkannt und hervorgehoben werden. Auch spiegeln sich dann nicht einzelne Trainingsdaten im Output wieder, sondern aus mehreren Trainingsdaten mit dem gleichen Label erkannte allgemeine Strukturen. Es erfolgt also gerade keine exakte Wiedergabe der Trainingsdaten im Output, und das vom Benutzer bereitgestellte Bild dominiert das Ergebnis.

Die Entwicklung eines Systems, das vollständige Trainingsdatenbilder in Ausgabedaten wiedergibt, ist zwar denkbar. Dies dürfte jedoch Spezialfälle darstellen, die dann einzelfallbezogen zu untersuchen sind. Gängiger ist das Training etwa anhand von Bildern eines bestimmten Malers, mit dem Ziel, ein neues Bild im Stile dieses Malers zu erzeugen.[155] Wenngleich die erzeugten Ergebnisse stilistisch den Trainingsdaten ähneln, geben sie diese doch nicht wieder, sondern stellen das Ergebnis einer Berechnung aus der Analyse verschiedener Aspekte der Trainingsdaten dar. So könnte etwa die im Ergebnis abgebildete Nase den Mittelwert aller Nasen der Trainingsdaten wiedergeben, wenn das Modell so konfiguriert wurde.

Auch kann in dem Output keine Vervielfältigung oder Bearbeitung des aus der Umwandlung des Trainingsinputs entstandenen Tensors gesehen werden, denn – wie im Rahmen der Beschreibung des technischen Vorgangs dargestellt – der Tensor selbst wird im Trainingsprozess nicht weiter vervielfältigt oder bearbeitet, sondern lediglich als Berechnungs- bzw. Optimierungsgrundlage verwendet.

IV. Exkurs: Manipulierte ML-Modelle

Eine differenziertere Betrachtung wird erforderlich im Kontext manipulierter ML-Systeme. Es folgt eine knappe Erläuterung, weshalb es unter Umständen doch zu diskutieren ist, ob ein ML-Modell Vervielfältigungen von Trai-

154 In der Regel stellt aber der Urheber oder ein Berechtigter das Ausgangsbild zur Verfügung, sodass sich keine urheberrechtlichen Fragestellungen ergeben.

155 Vgl. z. B. das Projekt *Next Rembrandt*, https://www.nextrembrandt.com (Stand: 22.02.2021).

ningsdaten enthält. Daran schließt sich eine Auswertung der Folgen auf die urheberrechtliche Betrachtung an.[156]

Unter einem manipulierten ML-Modell versteht diese Arbeit Modelle maschinellen Lernens, die in der Regel in der Cloud (Machine Learning as a Service, MLaaS) bereitgestellt werden, sodass der auszuführende Code für den Benutzer weder einsehbar noch kontrollierbar ist. Machine Learning as a Service beschreibt eine Dienstleistung, die es auch Benutzern ohne die erforderlichen Hardwarekapazitäten ermöglicht, Machine Learning für ihre Anwendungszwecke einzusetzen. Der Benutzer wählt dafür einen Anbieter und ein ML-Modell, das seinen Bedürfnissen entspricht, und stellt die zuvor gesammelten Trainingsdaten bereit.

Für (böswillige) Anbieter von MLaaS eröffnet das diverse Angriffsmöglichkeiten. Selbst wenn der Anbieter auf die Trainingsdaten, die der Benutzer hochlädt, nicht direkt zugreifen kann, bestehen für den Anbieter Möglichkeiten, die Trainingsdaten im Trainingsvorgang zu erfassen und in den Modellen selbst zu speichern, und damit etwa sensible Informationen auszulesen oder vom Kunden aufwendig zusammengestellte Trainingsdaten abzugreifen und für eigene Zwecke zu nutzen. Es gibt bereits zahlreiche Ansätze zur Umsetzung solcher Angriffe, exemplarisch seien hier nur die *Capacity Abuse Attack*,[157] die *Sign Encoding Attack*[158] sowie die *Least Significant Bit Attack*[159] genannt. Letztere wird für ein besseres Verständnis der diskutierten Problematik im Folgenden dargestellt.

1. Beispiel: Least Significant Bit Attack

Wie zuvor bereits beschrieben, bestehen ML-Modelle im Wesentlichen aus Struktur- bzw. Architekturvorgaben (Hyperparameter) und Parameterwerten (Weights / Biases). Letztere werden im Rahmen des Trainingsprozesses optimiert. Vermeintlich um eine möglichst hohe Präzision zu erreichen, wird für diese Werte in der Regel das Fließkommazahlenformat verwendet, das eine

156 Die folgenden Ausführungen wurden bereits veröffentlicht im Rahmen der DSRI-Herbstakademie 2020, *von Maltzan/Käde*, Algorithmen, die nicht vergessen – Model Inversion Attacks und deren Bedeutung für den Schutz der Daten und der Urheberrechte, S. 505 ff. sowie in *Käde/von Maltzan*, InTeR 4 2020, 201 ff..

157 *Song/Ristenpart/Shmatikov*, Machine Learning Models that Remember Too Much, 587, 593.

158 Dies., Machine Learning Models that Remember Too Much, 587, 591 f..

159 Dies., Machine Learning Models that Remember Too Much, 587, 590 f..

Zahl unter Verwendung von 32 Stellen (Bits) – Nullen und Einsen – darstellt. Der Wert 1,5 würde im Fließkomma-Binärformat wie folgt dargestellt:[160]

00111111110000000000000000000000

Dabei wirken sich allerdings Änderungen an den hinteren Bits nur marginal auf die Änderung des Wertes der Zahl aus: So entspricht

00111111110000000000000001001100

dem Dezimalwert 1,50000905991. Es ist also möglich, die hinteren – für den Zahlenwert weniger relevanten – Stellen nach Belieben zu verändern, ohne dass das Modell spürbar verändert wird. Im Beispiel können die hinteren Bits auch als Binärwert des Buchstaben „L" (01001100) interpretiert werden.

Bis zu 20 Bits je Parameter können auf diese Weise verändert werden, bevor die Genauigkeit des Modells signifikant abnimmt.[161]

Die *Least Significant Bit*-Methode setzt genau an dieser Stelle an. Der Code, der für das Training mit den Nutzertrainingsdaten abläuft, liest die Daten ein, wandelt sie in Binärformat-Zeichenketten um und teilt die entstandene Zeichenkette in beispielsweise 20-Bit-Pakete. Diese werden, nachdem das Training abgeschlossen ist, auf die zur Verfügung stehenden Parameter verteilt und die letzten 20 Bits jedes Parameters entsprechend angepasst.[162] Da das Modell ansonsten wie vom Nutzer erwartet trainiert wird, merkt dieser in der Regel nichts von dem Vorgang. Wenn der Nutzer das Modell im Anschluss wiederum frei zur Verfügung stellt,[163] kann der Angreifer aus den von ihm modifizierten Parametern die Trainingsdaten rekonstruieren.

---

160 Binär dargestellte Fließkommazahlen unterscheiden sich von „reinen" Binärformatzahlen dadurch, dass sie auch das Vorzeichen und die Nachkommastellen berücksichtigen. Ein ganzzahliger Wert bräuchte zur Darstellung einer Eins im 32-Bit-Format nur das hinterste Bit: 00000000000000000000000000000001, der Zwei entspräche 00000000000000000000000000000010. Zur einfachen Umwandlung von Dezimal- zu Binärformat vgl. z. B. https://www.h-schmidt.net/FloatConverter/IEEE754de.html (Stand: 22.02.2021).

161 *Song/Ristenpart/Shmatikov*, Machine Learning Models that Remember Too Much, 587, 595.

162 Dies., Machine Learning Models that Remember Too Much, 587, 595.

163 Für diese Fälle wird von „White-Box-Angriffen" gesprochen, denn dem Angreifer steht zur Vollendung des Angriffs das gesamte Modell samt aller Werte zur Verfügung, vgl. auch *Fredrikson/Jha/Ristenpart*, Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures, 1322 und *Song/Ristenpart/Shmatikov*, Machine Learning Models that Remember Too Much, 597.

2. Urheberrechtliche Bewertung

Song/Ristenpart/Shmatikov haben gezeigt, dass es durch diese Manipulationen möglich ist, Trainingsdaten, die in Form von Bildern vorliegen, beinahe vollständig wiederherzustellen, und zeigen dies anhand eines Vergleichs der Originalbilder mit den rekonstruierten Versionen.[164] Die rekonstruierten Bilder gleichen den Trainingsdaten dermaßen, dass tatsächlich davon gesprochen werden könnte, dass das manipulierte Modell eine Vervielfältigung im Sinne des Urheberrechts enthält. Die Modelle fungieren dann – ähnlich wie eine Festplatte, auf der Daten fragmentiert abgelegt werden – als Speichermedien für die abgegriffenen Daten. Noch einmal zur Verdeutlichung: Vervielfältigung im Sinne des Urheberrechts ist

> jede körperliche Festlegung eines Werkes, die geeignet ist, das Werk den menschlichen Sinnen auf irgendeine Art mittelbar oder unmittelbar wahrnehmbar zu machen.[165]

Dafür genügt, dass die Vervielfältigung auf einem Datenträger gespeichert wird, der sodann dazu genutzt werden kann, das Werk wahrnehmbar zu machen – so wie es auch bei CDs und Festplatten der Fall ist.[166] Nach dem Auslesen der wie beschrieben gespeicherten Informationen können die Ausgangswerke wieder wahrgenommen werden. Eine tiefergehende Betrachtung ist hier auch nicht erforderlich – schließlich liegt es auch schon aufgrund der Zielrichtung des Angreifers auf der Hand, dass genau die Vervielfältigung der Trainingsdaten beabsichtigt ist, um später Nutzen daraus ziehen zu können. Eine Bearbeitung (§ 23 S. 1 UrhG) hingegen liegt nicht vor und ist auch nicht beabsichtigt, Ziel ist die möglichst originalgetreue Reproduktion des Ausgangszustands. Festzuhalten ist: ML-Modelle *können* Vervielfältigungen enthalten, es bedarf dazu aber in aller Regel entsprechender Anstrengungen, um genau das zu erreichen.

V. Zusammenfassung

Nach der technischen Analyse und der erfolgten urheberrechtlichen Betrachtung ergibt sich, dass zwischen dem Einlese- und Aufbereitungsvorgang zum

164 Vgl. die Grafiken dazu in *Song/Ristenpart/Shmatikov*, Machine Learning Models that Remember Too Much, 587, 596.
165 RGZ 107, 277; BGHZ 17, 267, 269 f.; Dreier/Schulze–*Schulze*, UrhG, § 16 Rn. 6.
166 Dreier/Schulze–*Schulze*, UrhG, § 16 Rn. 7.

einen und dem Trainingsvorgang zum anderen zu unterscheiden ist. Der reine Trainingsvorgang bleibt dabei frei von urheberrechtlich relevanten Handlungen, da hier nur Berechnungen bzw. Optimierungen stattfinden. Einer genaueren Untersuchung wurde dagegen der Einlese- und Aufbereitungsvorgang unterzogen. In dessen Rahmen werden regelmäßig Vervielfältigungshandlungen und teilweise auch Bearbeitungen vorgenommen. Es bleibt im Folgenden zu prüfen, ob diese von Ausnahmetatbeständen erfasst sind.

§ 4 Mögliche Ausnahmetatbestände

Nachdem festgestellt wurde, dass grundsätzlich potenziell sowohl Vervielfältigungen als auch Bearbeitungen bzw. Umgestaltungen im Rahmen des Einlese- und Datenvorbereitungsvorgangs entstehen können, ist zu untersuchen, welche der bestehenden Ausnahmetatbestände die urheberrechtlich relevanten Handlungen dennoch einwilligungsfrei gestatten könnten. An dieser Stelle ist anzumerken, dass das UrhG aktuell einem umfassenden Änderungs- und Aktualisierungsprozedere unterliegt. Seit einem Entwurf vom 15. Januar 2020 war absehbar, dass ein § 44b UrhG eingeführt wird, der die Anwendung des § 44a UrhG im Kontext des Text und Data Mining endgültig unnötig macht, sowie dass u. a. eine Änderung des § 60d UrhG erfolgt; beides dient der Umsetzung des Art. 3 DSM-RL.[167] Die Betrachtung differenziert daher zwischen dem zum Stand der Bearbeitung aktuellen Gesetzesstand und die 2021 zu erwartenden Änderungen durch die Umsetzung der DSM-RL.

A. Situation de lege lata

Aktuell besteht eine Regelung zum TDM in § 60d UrhG, darüber hinaus existiert ein Ausnahmetatbestand für vorübergehende Vervielfältigungshandlungen in § 44a UrhG. Folgend wird untersucht, ob diese TDM-Vorschriften auch auf den ML-Prozess angewendet werden können sowie welche Folgen sich daraus ergeben.

I. Zulässigkeit sowohl der Vervielfältigung als auch der Bearbeitung gem. § 60d UrhG?

Möglicherweise sind die Grundsätze des TDM gem. §§ 60d ff. UrhG auch auf maschinelle Lernverfahren anwendbar. Dies würde den Einlesevorgang und die Aufbereitung urheberrechtlich geschützter Daten zumindest in der

167 *BMJV*, DiskE 01/2020, S. 14.

nichtkommerziellen wissenschaftlichen Forschung[168] ermöglichen. Die sogenannte „Text- und Data Mining-Schranke" soll gerade die flüchtigen Vervielfältigungen und Bearbeitungen, die im Rahmen automatisierter Auswertung von Datensammlungen zwangsläufig erfolgen, erlauben, sofern das TDM im Rahmen der nichtkommerziellen Forschung vorgenommen wird.[169]

Eine solche automatisierte Auswertung von Datensammlungen ist auch in der Datenverarbeitung im Prozess des maschinellen Lernens zu sehen; wie schon oben (§ 2 B.IV.1.) erläutert, liegt gerade darin die Schnittmenge von TDM und ML.

Auch in der „Strategie Künstliche Intelligenz" der Bundesregierung[170] wird TDM als Grundlage für maschinelles Lernen eingeordnet.[171] Insbesondere wird dort festgehalten: „The right to read is the right to mine."[172] Dies würde bedeuten: Sobald ein rechtmäßiger Zugang zu Werken besteht, ist das „Mining" – also die automatisierte Verarbeitung z. B. zur Mustererkennung – gestattet, und damit auch die anfallenden urheberrechtlich relevanten Handlungen.

Fraglich und diskussionsbedürftig ist welche Handlungen dem Mining-Prozess zugerechnet werden. Gehört dazu nur der Trainingsprozess (also der eigentliche Vorgang der Informationsgewinnung) oder auch das Einlesen und Aufbereiten der Daten? In der oben zitierten KI-Strategie der Bundesregierung erfolgt hierzu keine Klarstellung. In der Gesetzesbegründung des § 60d UrhG findet sich jedoch Folgendes:

> „Die Reform regelt erstmals das Text und Data Mining, bei dem eine Vielzahl von Texten, Daten, Bildern und sonstigen Materialien ausgewertet werden, um so neue Erkenntnisse zu gewinnen. Die Vorschrift erlaubt insbesondere die mit dieser Methode einhergehenden Vervielfältigungen, sofern diese in urheberrechtlich relevanter Weise das Vervielfältigungsrecht berühren, […]"[173]

Die Formulierung „mit dieser Methode einhergehende Vervielfältigungen" lässt darauf schließen, dass auch vorbereitende Maßnahmen (im Sinne des Einlesens und Aufbereitens der Trainingsdaten) von der Erlaubnis des § 60d UrhG erfasst sein sollen.

---

168 Nach aktuellem Stand – die zu erwartenden Änderungen werden in § 4 B. erläutert und ausgewertet.
169 Vgl. BT-Drs. 18/12329, S. 22.
170 *Bundesregierung*, Strategie Künstliche Intelligenz der Bundesregierung.
171 Vgl. Dies., Strategie Künstliche Intelligenz der Bundesregierung, S. 40.
172 Dies., Strategie Künstliche Intelligenz der Bundesregierung, S. 40.
173 BT-Drs. 18/12329, S. 22.

Damit gestattet § 60d UrhG nach aktuellem Stand die im Rahmen des gesamten ML-Prozesses anfallenden Vervielfältigungen. Dies muss grundsätzlich auch für Bearbeitungen gelten,[174] eine diesbezügliche Klarstellung erfolgt in § 23 S.3 UrhG, der festlegt, dass ausschließlich technisch bedingte Änderungen eines Werkes im Rahmen des TDM im Sinne des § 60d UrhG nicht von § 23 S. 1 und 2 UrhG erfasst werden.[175] Die zuvor angenommenen Bearbeitungen, die im Einlese- und Vorbereitungsvorgang anfallen können, erfolgen zu dem Zweck, die Daten für das ML-Modell optimal verarbeitbar zu gestalten, und sind mithin technisch bedingt.[176]

Auch für generative Modelle[177] ergibt sich übrigens nichts anderes, selbst wenn hier auf den ersten Blick die Auswertung von Daten gegenüber der Synthetisierung von Daten im Hintergrund zu stehen scheint: Die Vervielfältigungshandlungen für den Trainingsprozess finden im Voraus für die Vorbereitung des diskriminierenden Parts statt. Dieser lernt, die charakteristischen Merkmale der Daten zu erkennen, die ihm zugeführt werden. Die für das TDM übliche Analyse und Auswertung findet also vor dem Einsatz und Training des generativen Parts statt, ist aber integraler Bestandteil.

II. Zulässigkeit der Vervielfältigung gem. § 44a UrhG?

Für die Einsatzbereiche außerhalb der nichtkommerziellen Forschung könnte nach dem aktuellen Gesetzesstand § 44a UrhG das Training der ML-Modelle ermöglichen. Möglicherweise ist die Vervielfältigung im Rahmen des Einlesevorgangs nach § 44a Nr. 2 UrhG zulässig (wobei dann im weiteren Verlauf zu klären wäre, ob dies auch auf die Bearbeitung anwendbar ist). Dazu müsste die Vervielfältigung flüchtig oder begleitend sein, einen integralen und wesentlichen Teil eines technischen Verfahrens darstellen und ihr einziger Zweck müsste es sein, eine rechtmäßige Nutzung des vervielfältigten Werkes zu ermöglichen. Darüber hinaus dürfte die Vervielfältigung keine eigenständige wirtschaftliche Bedeutung haben.

174 So wohl auch Dreier/Schulze–*Dreier*, UrhG, § 60d Rn. 1.
175 Wandtke/Bullinger–*Bullinger*, PK UrhR, § 60d Rn. 14.
176 Vgl. zu den möglicherweise anfallenden Bearbeitungen § 3 B.II..
177 Vgl. zur Erklärung § 2 B.II.3..

1. Flüchtige Vervielfältigungen

Die Vervielfältigungen der Trainingsdaten im Rahmen des Einlesens und der Aufbereitung dürften in der Regel ohne weiteres als flüchtige Vervielfältigungen einzuordnen sein, denn sie werden erst bei Start des Programmablaufs erzeugt und existieren nicht über den Programmablauf hinaus.

2. Rechtmäßige Nutzung

Die Aufbereitungsvorgänge und der Machine-Learning-Trainingsprozess müssten rechtmäßige Nutzungen der flüchtig vervielfältigten Daten darstellen. Eine Nutzung ist dann rechtmäßig, „wenn sie [...] vom jeweiligen Rechtsinhaber erlaubt ist, oder [..] im Rahmen gesetzlicher Schrankenbestimmungen zulässig und auch sonst nicht durch Gesetze beschränkt ist."[178] Mit der Einstellung der Bundesregierung – „The right to read is the right to mine"[179] – müsste der ML-Prozess grundsätzlich als rechtmäßige Nutzung einzuordnen sein. Dies setzt aber voraus, dass ein „right to read", also ein rechtmäßiger Zugang zu den Daten, besteht.

3. Vorübergehende Vervielfältigung rechtswidriger Quellen

Ob § 44a UrhG auch die vorübergehende Vervielfältigung rechtswidriger Quellen privilegiert, wenn die sich daran anschließende Nutzungshandlung urheberrechtsfrei ist, ist umstritten.[180] Der EuGH differenziert grundsätzlich zwischen rechtmäßigem und rechtswidrigem Zugang zu der zu vervielfältigenden Quelle,[181] während die Rechtsprechung in Deutschland eine Privilegierung durch § 44a UrhG bisher wohl nur für den Fall ablehnt, dass die Rechtswidrigkeit der Quelle für den Nutzer offensichtlich ist (Anwendungsbereich ist insbesondere das Video-Streaming, wobei zu klären war, ob die Nutzer einer Plattform, die offensichtlich rechtswidrig kostenlosen Zugang

178 Dreier/Schulze–*Dreier*, UrhG, § 44a Rn. 8; Erw.gr. 33 InfoSoc-RL.
179 *Bundesregierung*, Strategie Künstliche Intelligenz der Bundesregierung, S. 40.
180 Dreier/Schulze–*Dreier*, UrhG, § 44a Rn. 8.
181 Vgl. EuGH C-527/15 Rn. 61 – *Stichting Brein (Filmspeler)*; Dreier/Schulze–*Dreier*, UrhG, § 44a Rn. 8.

zu Filmen bereitstellen, sich bzgl. des Streamings im eigenen Browser auf § 44a UrhG berufen können).[182]

Im Zuge der Umsetzung der DSM-RL wurden erneut Stimmen laut, die eine Klarstellung dahingehend fordern, dass eine automatisierte Auswertung aufgrund der urheberrechtlichen Neutralität des Vorganges – und der Möglichkeit eines Rückgriffs auf § 44a UrhG – auch solcher Daten möglich sein müsse, zu denen der Zugang nicht rechtmäßig erlangt wurde – dies sei insbesondere im Hinblick auf investigativen Journalismus und Whistleblowing unerlässlich.[183] Hier wird allerdings wohl zu differenzieren sein. So sind etwa ML-Anwendungen denkbar, die während des Modelleinsatzes weiterlernen (sog. „Online-Learning"[184]). Diese Systeme erfassen Trainingsdaten im laufenden Betrieb (etwa durch Auswertung von Nachrichtenwebseiten), wozu Vervielfältigungen der zu analysierenden Daten erforderlich sind, die aber nicht gespeichert werden (müssen), sodass insofern nur flüchtige Vervielfältigungen anfallen. Es kann nicht Sinn der Regelung sein, dass Verwender solcher Systeme Zugangsschutzbarrieren (wie sie zum Beispiel bei kostenpflichtigen Nachrichtenseiten eingesetzt werden) umgehen und sich anschließend auf den Ausnahmetatbestand des § 44a UrhG berufen können. Dies erscheint unter anderem aus wirtschaftlichen Gesichtspunkten nicht wünschenswert.

4. Zwischenergebnis

Der Ausnahmetatbestand des § 44a UrhG hilft an dieser Stelle also nicht in jedem Fall fehlender Privilegierung durch § 60d UrhG weiter. Selbst wenn in vielen Fällen ML-Vervielfältigungen als flüchtige Vervielfältigungen einzustufen sind, besteht mindestens Rechtsunsicherheit dahingehend, ob der Ausnahmetatbestand auch bei rechtswidrig erlangten Daten greift.

182 Dreier/Schulze–*Dreier*, UrhG, § 44a Rn. 8; vgl. z. B. LG Köln GRUR-RR 2014, 114, 115 – *The Archive*.
183 *Wikimedia Deutschland*, Stellungnahme DiskE 01/2020, S. 6.
184 *Goodfellow* et al., Deep Learning Handbuch, S. 310.

B. *Situation nach zu erwartenden Änderungen durch die Umsetzung der DSM-Richtlinie*

Deutlich mehr Klarheit in Bezug auf die Regelung der Machine Learning-Handlungen wird durch die Umsetzung aktueller europäischer Rechtsakte erwartet. Es folgt eine knappe Analyse der umzusetzenden Vorgaben, im Anschluss daran eine Erörterung der geplanten nationalen Umsetzung, die danach in Bezug auf Machine Learning ausgewertet wird.

I. Was ist umzusetzen?

Auf EU-Ebene wurden im Jahr 2019 mit der DSM-RL in deren Artikeln 2, 3 und 4 auch einige Regelungen zum Text- und Data Mining verabschiedet, die allerdings noch bis 2021 in nationales Recht umzusetzen waren. Unter anderem ist die TDM-Schranke nicht mehr explizit auf nichtkommerzielle Forschung beschränkt, sondern soll jedermann zumindest die automatisierte Auswertung und die dazu vorgenommenen Vervielfältigungshandlungen erlauben.

Artikel 2 Nr. 2 enthält eine Definition des TDM:

> „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."

Artikel 3 DSM-RL regelt TDM zum Zwecke der wissenschaftlichen Forschung, Artikel 4 sodann „Ausnahmen und Beschränkungen für das Text und Data Mining". Letztere gelten nicht mehr nur für (nichtkommerzielle) Forschung, sondern erlauben „zum Zwecke des Text und Data Mining vorgenommene Vervielfältigungen und Entnahmen von rechtmäßig zugänglichen Werken und sonstigen Schutzgegenständen" (Art. 4 Abs. 1) ohne eine Beschränkung auf einen bestimmten Anwenderkreis.

Artikel 3 verlangt für Forschungsorganisationen und Einrichtungen des Kulturerbes, die die gesammelten Daten bzw. Korpora speichern und aufbewahren, hingegen *nicht*, dass diese Aufbewahrung für die Zwecke des TDM erforderlich ist – und privilegiert damit ebendiesen Anwenderkreis. „Forschungsorganisationen" sind nach wie vor Einrichtungen, die „in ihrer Tätigkeit nicht gewinnorientiert [sind] oder alle Gewinne in ihre wissenschaftliche Forschung [reinvestieren], oder im Rahmen eines von einem

Mitgliedstaat anerkannten Auftrags im öffentlichen Interesse tätig [sind]" (Art. 2 Nr. 1 a und b DSM-RL), also im wesentlichen die schon bisher privilegierte nichtkommerzielle Forschung.

II. Wie erfolgt die Umsetzung?

Die Umsetzung erfolgt im Rahmen eines „Gesetzes zur Anpassung des Urheberrechts an die Erfordernisse des digitalen Binnenmarkts" (UrhG-E). Hierzu existieren zwei Diskussionsentwürfe des BMJV von Januar und Juni 2020, die in einem im Oktober 2020 veröffentlichten Referentenentwurf zusammengeführt wurden,[185] aus dem sich auch schon das Schicksal der TDM-Regelungen (de lege lata und de lege ferenda) andeutete. Am 03. Februar 2021 wurde ein Regierungsentwurf beschlossen, der sich in Bezug auf die TDM-Regelungen nicht von dem Referentenentwurf aus Oktober 2020 unterscheidet.[186] Das entsprechende Gesetz wurde am 04. Juni 2021 im Bundesgesetzblatt verkündet und ist am 07. Juni 2021 in Kraft getreten.[187] Die Verkündung erfolgte nach Einreichung dieser Arbeit, daher entspricht im Folgenden „UrhG-E" dem aktuell seit 07. Juni 2021 geltenden Urheberrechtsgesetz und „UrhG" stellt den Stand vor der Umsetzung dar.

Die Umsetzung des Art. 4 DSM-RL erfolgt durch den neuen § 44b UrhG-E, Art. 3 DSM-RL wird durch eine Neufassung des § 60d UrhG-E umgesetzt. Der neue § 44b UrhG-E enthält im ersten Absatz eine Definition von Text und Data Mining, regelt im zweiten Absatz, welche Vervielfältigungen zulässig sind und räumt im dritten Absatz dem Rechtsinhaber auch die Möglichkeit einer Nutzungsuntersagung ein.[188]

Die Definition in § 44b Abs. 1 UrhG-E, die die zuvor gefundene Definition[189] des TDM weiter präzisiert, entspricht inhaltlich im Wesentlichen der, die bereits in Art. 2 Nr. 2 DSM-RL enthalten ist: Danach ist Text und Data

185 *BMJV*, RefE 10/2020.

186 *Bundesregierung*, RegE 02/2021, S.13 f..

187 Gesetz zur Anpassung des Urheberrechts an die Erfordernisse des digitalen Binnenmarktes vom 31. Mai 2021, BGBl. 2021 Teil I Nr. 27, hrsg. am 04. Juni 2021, S. 1204 ff.

188 *Bundesregierung*, RegE 02/2021, S. 13; *BMJV*, DiskE 01/2020, S. 5; im Rahmen dieses Kapitels wird auf die Seitenzahlen der Entwürfe abgestellt, da das Gesetz erst nach Ende der Bearbeitung dieser Arbeit in Kraft getreten ist. Inhaltlich haben sich seit dem Regierungsentwurf in Bezug auf die hier diskutierten Passagen im verabschiedeten Gesetz keine Änderungen mehr ergeben.

189 S. oben § 2 B.IV.1..

Mining die automatisierte Analyse von einzelnen oder mehreren digitalen oder digitalisierten Werken, um daraus Informationen insbesondere über Muster, Trends und Korrelationen zu gewinnen.

Eine Beschränkung der Zulässigkeit des TDM auf den Kreis nichtkommerzieller Forschung erfolgt nicht mehr. Zu beachten ist auch, dass diese explizite TDM-Regelung nicht mehr in den §§ 60a ff. UrhG (gesetzlich erlaubte Nutzungen für Unterricht, Wissenschaft und Institutionen), sondern im Abschnitt (allgemein) gesetzlich erlaubter Nutzungen zu finden ist. Damit wird dem Gedanken Rechnung getragen, dass TDM eben nicht mehr nur einem bestimmten Adressatenkreis vorbehalten sein soll, während alle anderen sich mit dem § 44a UrhG abfinden müssen – sondern der Erkenntnisgewinn durch TDM soll allen möglich sein, die das Potenzial dieser Technologie erkennen und sie einsetzen möchten.

Darüber hinaus wird zur Umsetzung des Art. 3 DSM-RL der § 60d UrhG-E überarbeitet. Dieser stellt nun eine Spezialregelung gegenüber § 44b UrhG-E für die Forschung dar, wobei das bisher strenge Kriterium der Nichtgewerblichkeit durch die differenziertere von der DSM-RL vorgegebene Charakterisierung der Berechtigten ersetzt wird.[190]

In § 44b UrhG-E verlangte der Entwurfstext im Januar 2020 noch die Erforderlichkeit der vorgenommenen Vervielfältigungen („Zulässig sind Vervielfältigungen, sofern sie für das Text und Data Mining erforderlich sind“),[191] obgleich Artikel 4 DSM-RL die Erforderlichkeit der Vervielfältigungshandlungen nicht bedingt. Schon im Referentenentwurf war die Erforderlichkeit so nicht mehr Bestandteil des § 44b Abs. 2 UrhG-E, dies ist auch im Regierungsentwurf übernommen worden.[192] Satz 2 des § 44b Abs. 2 UrhG-E besagt allerdings nach wie vor, dass die Vervielfältigungen zu löschen sind, wenn sie für das TDM „nicht mehr erforderlich“ sind.[193]

Möglicherweise kann hieraus dennoch die grundsätzliche Erforderlichkeit der Vervielfältigungen konstruiert werden, denn „nicht mehr erforderlich“ impliziert, dass zuvor eine Erforderlichkeit vorlag. Insofern ist davon auszugehen, dass diese textliche Änderung lediglich der Verbesserung der Lesbarkeit dienen sollte, es ist zu hoffen, dass diesbezüglich noch eine Klarstellung erfolgt. Anderenfalls könnten sich schwerwiegende Folgen für den ML-Entwicklungsprozess ergeben: Nicht immer ist Code effizient aufgebaut, mitunter werden etwa die geladenen Daten im Arbeitsspeicher in andere

190 *Bundesregierung*, RegE 02/2021, S. 14; *BMJV*, DiskE 01/2020, S. 6.
191 Dass., DiskE 01/2020, S. 5.
192 *Bundesregierung*, RegE 02/2021, S. 13.
193 Vgl. Dies., RegE 02/2021, S. 13.

Variablen kopiert. Müssen die Entwickler bei jedem Arbeitsschritt darüber nachdenken, ob sie gerade eine unnötige Vervielfältigung vornehmen? Welcher Schaden würde den Urhebern aus diesen, nur im Code vorgenommenen und den Arbeitsspeicher nicht verlassenden, aber unnötigen Vervielfältigungshandlungen entstehen? An dieser Stelle ist aus Gründen der Rechtssicherheit für die Entwickler zu fordern, dass hier nicht zu engmaschige Anforderungen an die Erforderlichkeit der Vervielfältigungen oder andere urheberrechtlich relevante Handlungen gestellt werden.

§ 60d Abs. 1 UrhG-E bezieht sich auf § 44b UrhG-E, und zwar nicht nur für die Begriffsdefinition, sondern explizit auch auf Zulässigkeit der Vervielfältigungen in § 44b Abs. 2 S. 1 UrhG-E. § 44b Abs. 2 S. 2 UrhG-E ist für gem. § 60d Abs. 1 UrhG-E privilegiertes TDM nicht anwendbar. Daraus ergibt sich die Frage, ob dann auch das Kriterium der Erforderlichkeit der Vervielfältigungen in Satz 1 nicht mehr gegeben ist. An dieser Stelle bleibt die gesetzliche Regelung unklar.

Aus der Begründung geht nicht hervor, dass hier eine Abweichung von Art. 4 Abs. 1 DSM-RL beabsichtigt ist. Hier bleibt abzuwarten, ob diesbezüglich eine Überarbeitung zur Klarstellung erfolgt. Dies wäre zu begrüßen, denn die Beschränkung auf erforderliche (im Unterschied zu lediglich zum Zweck des TDM vorgenommenen) Vervielfältigungshandlungen bedingt unter Umständen ein aufwendiges Hinterfragen von Programmierpraktiken[194] und Datensicherungsstrategien, während die Zweckbezogenheit ohne Weiteres bei allen im Programmcode angelegten Vervielfältigungen angenommen werden kann, sofern nicht offensichtlich andere Zwecke verfolgt werden.

Kritisiert wird an der neuen Fassung bisher unter anderem, dass lediglich die *Verarbeitung* von Daten geregelt sei, wohingegen keine Aussagen darüber getroffen würden, ob „das Anzeigen und die Mitteilung der Ergebnisse einer solchen Verarbeitung" erlaubt seien.[195] Des Weiteren habe § 60d UrhG-AF auch die öffentliche Zugänglichmachung des Korpus bzw. der gesammelten Daten für zulässig erklärt, wohingegen eine solche Regelung in § 44b UrhG-E nicht enthalten ist.[196]

Dass auf TDM bei rein flüchtigen Vervielfältigungen auch weiterhin die hierfür bestehende Regelung anwendbar sein soll (in Deutschland also der

194 Zum Beispiel für den Fall, dass die Vervielfältigungen nicht im Prozess der Datensammlung, sondern während der Aufbereitung bei Programmstart bzw. zur Normalisierung (vgl. Erw.Gr. 8 DSM-RL) erfolgen.

195 *Google*, Stellungnahme DiskE 01/2020, S. 8.

196 Dass., Stellungnahme DiskE 01/2020, S. 8 f..

§ 44a UrhG), geht aus Erw.Gr. 9 DSM-RL hervor. Dies ist dem Entwurfstext nicht ohne Weiteres zu entnehmen, was ebenfalls bereits für Kritik sorgte.[197]

1. Welche Auswirkungen ergeben sich für den ML-Prozess?

Zu untersuchen sind noch die Abweichungen in der Behandlung des TDM bzw. die Auswirkungen der zu erwartenden Änderungen auf die Behandlung des ML-Prozesses. Zunächst ergibt die neue Regelung einen Unterschied an die Herangehensweise der Prüfung dergestalt, dass zum Einen nach der Anwendergruppe zu unterscheiden ist und zum anderen nach den fraglichen Handlungen: Denn wenn die Privilegien aus § 60d UrhG-E nicht in Anspruch genommen werden sollen, reichen unter Umständen die §§ 44a und 44b UrhG-E aus.

Einen Überblick über die zu erwartenden Änderungen verschafft Abbildung 4.1. Das Schema unterscheidet vier exemplarische Personengruppen (Privatperson („Citizen Science"), Einzelforscher, nicht-kommerziell handelnde und kommerziell handelnde Forschungsorganisationen) und analysiert die TDM-Prozessschritte hinsichtlich bestehender und zu erwartender urheberrechtlicher Regelungen. Die rechten Teilspalten stellen dabei jeweils den Stand nach Umsetzung der DSM-RL laut aktuellem Entwurf dar. Festzustellen ist, dass kommerzielle Forschung und Citizen Science urheberrechtlich gleich gestellt sind, und sich in Zukunft – wie zuvor der nichtkommerziellen Forschung vorbehalten – auch auf eine TDM-Schranke berufen können (§ 44b UrhG-E).

Warum Einzelforscher laut dem Entwurf (der so auch übernommen wurde) nicht mehr explizit zur Aufbewahrung der Vervielfältigungen befugt sind, geht aus der Entwurfsbegründung nicht hervor. Damit sind sie in dem Punkt gleichgestellt zur Citizen Science, womit gem. § 44b Abs. 2 S. 2 UrhG-E die Vervielfältigungen zu löschen sind, wenn sie für das TDM nicht mehr erforderlich sind.[198]

Übertragen auf den Prozess des maschinellen Lernens bleibt die oberste Zeile in Abbildung 4.1 außer Betracht, relevant sind vor allem die zweite und dritte Zeile. Auch die Nachnutzung des erstellten Korpus (Zeilen vier und fünf) liegen außerhalb des Fokus dieser Arbeit. Grundsätzlich kann sich in Zukunft ungeachtet der konkreten Umsetzung also jeder auf Ausnahme-

197 *Wikimedia Deutschland*, Stellungnahme DiskE 01/2020, S. 5 f..
198 Dies kritisierend *Spindler*, Kurz-Stellungnahme zum DiskE, S. 1.

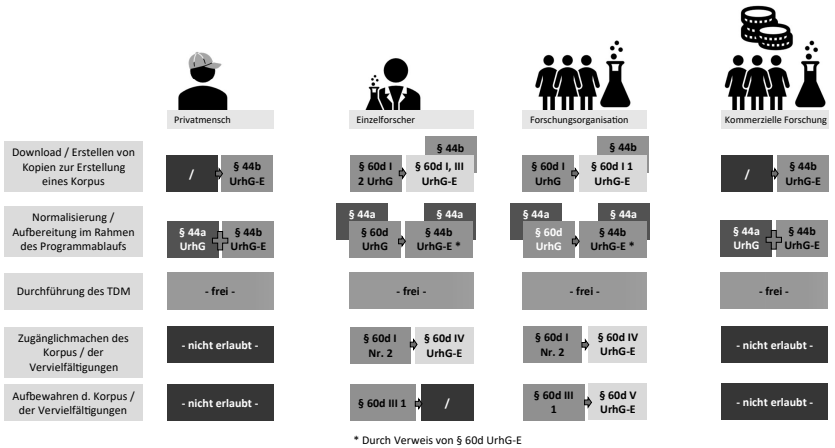

Abbildung 4.1: Änderungen der TDM-Regelungen, Quelle: eigene Darstellung.

tatbestände berufen und ist nicht mehr auf den engen Anwendungsbereich flüchtiger Vervielfältigungen des § 44a UrhG beschränkt.

C. Zusammenfassung

Machine Learning ist (schon jetzt) auf jeglichen Daten zulässig, ungeachtet der Legitimität des Zugangs – sofern keine nicht lediglich flüchtigen Vervielfältigungen stattfinden. Bezüglich für den Zweck des Machine Learnings erfolgender Vervielfältigungen im Vorfeld und im Rahmen des ML-Prozesses kann sich in Zukunft jede/r auf die anwendbare TDM-Schranke in § 44b UrhG-E berufen, wohingegen diese Schranke zum aktuellen Zeitpunkt noch der nichtkommerziellen Forschung vorbehalten ist.

§ 5 Ergebnis des zweiten Teils

In diesem Teil der Arbeit wurde der Machine Learning-Prozess untersucht. Dabei konnten unterschiedliche Phasen der Datenverarbeitung identifiziert werden:

Die Trainingsdaten, aus denen die ML-Modelle lernen, werden zunächst eingelesen und dann für die Verarbeitung im Trainingsvorgang vorbereitet. Dabei erfolgt zum einen eine Umwandlung in ein für das ML-Programm lesbares Datenformat, zum Beispiel Tensoren, und zum anderen ggf. auch noch eine Anpassung oder Skalierung der Daten, um ein effizientes Training zu ermöglichen. Erst im Anschluss daran analysieren die ML-Modelle die Trainingsdaten im Trainingsprozess und optimieren währenddessen ihre Parameter.

Auch vor dem eigentlichen Machine Learning-Prozess müssen Data Scientists ggf. Maßnahmen treffen: Zum einen sind die passenden Trainingsdaten zu sammeln, diese sind (für das überwachte Lernen) mit Labels zu versehen; außerdem kann es zur Sicherung der Datenqualität etwa erforderlich sein, „Ausreißerdaten" (also Daten, die für das zu untersuchende Merkmal eigentlich untypisch sind) zu entfernen oder eine möglicherweise Datensammlung immanente Voreingenommenheit oder implizite Vorurteile auszugleichen, indem Datenpunkte weggenommen oder hinzugefügt werden (so könnten zum Beispiel für die visuelle Erkennung eines ausgeübten Berufs einer abgebildeten Person geschlechterspezifische Verzerrungen in den Daten enthalten sein, weil von einer Berufsgruppe überdurchschnittlich viele Personen eines Geschlechts in den Trainingsdaten vertreten sind). Diese vorbereitenden Handlung liegen jedoch außerhalb der erklärten Grenzen dieser Arbeit und wurden daher nicht auf urheberrechtlich relevante Handlungen untersucht, stattdessen wurde auf den eigentlichen ML-Prozess fokussiert.

Auf die eingangs gestellten Fragen zurückkommend ist demnach Folgendes festzustellen: Hinsichtlich des Machine Learning-Prozesses ist also zu unterscheiden zwischen den vorbereitenden Handlungen, dem Einlesevorgang, der dann ggf. erfolgenden Anpassung der Daten sowie dem eigentlichen Training. Urheberrechtlich relevante Handlungen kommen insbesondere außerhalb des eigentlichen Trainings, also im Rahmen der Einlese- und Aufbereitungs-

vorgänge, in Betracht.[199] Für diese sind dann die geltenden und die mit der Umsetzung der DSM-RL kommenden TDM-Schrankenbestimmungen in den §§ 44a, 44b (UrhG-E) und 60d UrhG relevant und anwendbar.[200] Trainingsdaten sind in der Regel nicht im fertigen Modell enthalten, es sei denn, das Modell wurde diesbezüglich manipuliert.[201]

Zu berücksichtigen ist jedoch, dass diese Einschätzung nur eine Momentaufnahme des aktuellen, untersuchten Technologiestandes darstellt. Der technische Fortschritt ist stets im Blick zu behalten.

199 S. oben § 3 B..
200 S. zur Situation de lege lata und de lege ferenda oben § 4.
201 S. zur Manipulation den Exkurs in § 3 B.IV..