

Urheberrechtliche Zulässigkeit des Text und Data Mining für generative KI-Systeme: Systematisierung der Optionen für einen kohärenten Rechtsrahmen

PD Dr. Daniel Timmermann / Dr.-Ing. Andreas Lommatzsch*

Erwägungsgrund 105 der KI-Verordnung¹ definiert generative KI-Systeme und fasst zugleich den regulatorischen Interessenkonflikt zusammen: „KI-Modelle mit allgemeinem Verwendungszweck, insbesondere große generative KI-Modelle, die *Text, Bilder und andere Inhalte erzeugen* können, bedeuten einzigartige *Innovationsmöglichkeiten*, aber auch *Herausforderungen für Künstler, Autoren und andere Kreative* [...]. Für die Entwicklung und das Training solcher Modelle ist der *Zugang zu riesigen Mengen an Text, Bildern, Videos und anderen Daten erforderlich*. In diesem Zusammenhang können Text-und-Data-Mining-Techniken in großem Umfang für das Abrufen und die Analyse solcher Inhalte, die *urheberrechtlich* [...] geschützt sein können, eingesetzt werden. Für jede Nutzung urheberrechtlich geschützter Inhalte ist die *Zustimmung* des betreffenden Rechteinhabers erforderlich, es sei denn, es gelten *einschlägige Ausnahmen und Beschränkungen* des Urheberrechts.“

Die Jahrestagung der GJZ in Göttingen 2024 widmete sich „Disruptionen im Privatrecht“, wobei unter dem Ausdruck „Disruption“ die Notwendigkeit eines fundamentalen Paradigmenwechsels verstanden wird. Die Abhandlung legt dar, dass die kleinteiligen Streitigkeiten über den sachlichen Anwendungsbereich der §§ 44b, 44a UrhG bzw. der zugrundeliegenden EU-Richtlinien eine vom *Telos des Urheberrechts losgelöste Eigendynamik* gewonnen haben, und zeigt gesetzgeberische Lösungsoptionen auf.

* Daniel Timmermann hat sich mit Unterstützung der Deutschen Forschungsgemeinschaft (DFG) bei Prof. Dr. Caroline Meller-Hannich (Halle) habilitiert und vertritt im WS 2024/25 den Lehrstuhl für Bürgerliches Recht und Digitalisierung an der Europa-Universität Viadrina Frankfurt (Oder). Andreas Lommatzsch forscht am DAI-Labor der TU Berlin. Seine Arbeitsschwerpunkte liegen im Bereich personalisierte Informationssysteme und Empfehlungssysteme. Derzeit liegt sein Forschungsschwerpunkt auf Sprachmodellen und intelligenten Konversations-Agenten. Der Autor Lommatzsch hat die technischen Ausführungen in A I 1, 2, II 2 a bb verfasst.

1 VO (EU) 2024/1689, ABl. L 12.7.2024.

Den größten Block des Beitrags bildet die Analyse der rechtlichen und technischen Ausgangslage (A). Die zusammengefassten Ergebnisse werden sodann am Zweck des Urheberrechts gespiegelt (B). Im Kern bestehen drei Optionen wie der Unionsgesetzgeber einen Rechtsrahmen ausgestalten kann, der kohärent im Sinne von technikneutral ist, weil etwaige temporäre und/oder dauerhafte Vervielfältigungen im Rahmen des Trainingsablaufs nicht länger ein Differenzierungskriterium darstellen. Zudem gilt es faktische Rückwirkungen mit dem Internationalen Privatrecht zu berücksichtigen. Dadurch führt der Beitrag das Forschungsgespräch zurück zum Telos des Urheberrechts (C). Die Abhandlung schließt mit einer kurzen politischen Stellungnahme (D).

Das Text und Data Mining tangiert neben dem Urheberrecht insbesondere das *Datenschutzrecht*. Dieses gehört ebenso nicht zum Untersuchungsgegenstand, wie die (zu verneinende²) Frage, ob generative KI neue geschützte Werke erzeugen kann.

A. Rechtliche und technische Ausgangslage

Wenn ein Softwareunternehmen beim Text und Data Mining in Verwertungsrechte des Urhebers eingreift und damit eine urheberrechtlich relevante Handlung vornimmt (I), müsste diese durch eine Schranke gerechtfertigt sein.³ Andernfalls bestünde ein Unterlassungs- und gegebenenfalls Schadensersatzanspruch.⁴ Sofern das Softwareunternehmen mit dem Urheber keinen Nutzungsvertrag abgeschlossen hat, kommt als Rechtfertigungsgrund allein eine gesetzliche Schranke in Betracht (II). Infolge der KI-Verordnung ist fraglich geworden, ob die Urheberrechtsstandards der EU auch für auswärtige Anbieter gelten (III).

2 Überzeugend nach dem Autonomiegrad von Systemen differenzierend N. Maamar, Urheberrechtliche Fragen beim Einsatz von generativen KI-Systemen, ZUM 2023, 481 (490).

3 *Terminologisch* orientiert sich die Abhandlung an Verdikten des BGH, der in einem ersten Schritt einen *Eingriff in Verwertungsrechte* und in einem zweiten Schritt i.R.d. Widerrechtlichkeit *Schrankenregelungen als Rechtfertigungsgrund* prüft (BGH GRUR 2015, 1101 (1102; 1103); BGH GRUR 2013, 503 (503)).

4 § 97 UrhG.

I. Urheberrechtlich relevante Handlungen

Das Urheberrecht findet auf *urheberrechtlich geschützte Werke* im Sinne des § 2 UrhG Anwendung. Daher ist für sämtliche als Trainingsmaterialien genutzte Inhalte gesondert zu prüfen, ob sie eine persönliche geistige Schöpfung sind (Abs. 2), was einen gewissen Grad der Individualität des jeweiligen geistigen Erzeugnisses voraussetzt⁵. Der Ausdruck Trainingsmaterial bezeichnet vorliegend eine Gruppe verwendeter Werke, mithin wird als Prämisse unterstellt, dass nach dem Urheberrecht geschützte immaterielle Gegenstände bestehen. Im Einzelfall kann eine nähere Prüfung erforderlich sein, wobei Schutz auch für die sogenannte kleine Münze besteht⁶.

Ist das Eingangstor zum Urheberrecht eröffnet, so weisen §§ 11, 15 ff. UrhG dem Urheber seine *Verwertungsrechte* zu. Aus dem beispielhaften⁷ Rechtekatalog der §§ 15 ff. UrhG kommen Eingriffe in das *Vervielfältigungsrecht* (§ 16 Abs. 1 UrhG) in Betracht. In chronologischer Hinsicht gilt es zwischen dem Training (1), dem trainierten System (2) sowie den Ausgaben des Systems (3) zu unterscheiden. Weil der Katalog nur beispielhaft ist, stellt sich die Frage, ob für das Training generativer KI als solches ein *unbenanntes Verwertungsrecht* existiert (4).

1. Vervielfältigungen im Rahmen des Trainings

Die gegenwärtig verfügbaren großen Sprachmodelle (Large Language Models, kurz "LLMs") basieren auf großen künstlichen neuronalen Netzen, die darauf trainiert werden, bei einer Anweisung an die Software mittels einer textuellen Eingabe ("Prompt") eine gute (zumeist) textuelle Antwort zu liefern. Während des Trainings werden die Gewichtungen zwischen den Knoten des Netzwerkes schrittweise angepasst. Für das Training von Sprachmodellen werden umfangreiche Sammlungen von Text-Dokumenten benötigt. Es wird dabei überwiegend ein selbstüberwachtes Lernverfahren (self-supervised learning) genutzt, mit dem das Sprachmodell *Muster in Texten und Sprache lernt*. Hierfür werden im Allgemeinen Texte aus Büchern oder der freien Enzyklopädie Wikipedia genutzt (Schritt 1). Nach dem Pre-Trai-

5 W. Bullinger in: A. Wandtke/W. Bullinger (Hrsg.), UrhR-Kommentar, 6. Aufl., München 2022, § 2 UrhG Rn. 23.

6 Bullinger (Fn. 5), § 2 UrhG Rn. 24.

7 G. Schulze in: T. Dreier/G. Schulze (Hrsg.), UrhG-Kommentar, 7. Aufl., München 2022, § 15 Rn. 10.

ning erfolgt zumeist ein *Finetuning* mit aktuellen Web-Dokumenten oder im Unternehmen vorhandenen Daten, um einen aktuellen Wissensstand sicherzustellen sowie aktuelle Formulierungen anzutrainieren (Schritt 2). Zudem wird das neuronale Netzwerk mit Hilfe von Frage-Antwort-Paaren sowie Instruktionsdatensätzen darauf trainiert, zu vorgegebenen Prompts hochqualitative Antworten zu liefern (Schritt 3).

Das Ziel von Lernverfahren ist es, *Muster und Wahrscheinlichkeitsverteilungen in Dokumenten zu lernen*. Die Texte werden nicht direkt abgespeichert ("auswendig gelernt"). Vielmehr werden alle relevanten Muster ermittelt, um durch die Anwendung der Muster bei neuen Eingaben passende sinnvolle Ausgaben zu generieren. Da große Sprachmodelle mehrere Milliarden Parameter besitzen, die beim Training eingestellt werden müssen, sind ein umfangreicher Trainingsdatensatz sowie ein effektives Trainingsverfahren essentiell. Für das Training des Modells werden Texte in eine geeignete Repräsentation überführt. Der Vorgang beinhaltet das Entfernen irrelevanter Teile (z. B. Formatierungen, Meta-Daten, Navigations-Elemente), die Zerlegung in kleine sinntragende Elemente ("Token"), die meist aus 1-5 Zeichen bestehen, sowie die Repräsentation in Vektoren.

Im Rahmen des Trainings erfolgen zumeist temporäre und/oder dauerhafte *Vervielfältigungen*. Technisch zwingend erforderlich ist das indessen nicht. Erwägungsgrund 9 der DSM-Richtlinie⁸ geht zutreffend davon aus, dass es auch Fälle des Text und Data Mining geben kann, in denen keine Vervielfältigung erfolgt (sogleich a bb).⁹ Für das Training von Sprachmodellen kann ein lokaler dauerhafter Korpus (a) oder ein anhand von im Internet verfügbaren Dokumenten zusammengestellter temporärer Korpus (b) genutzt werden. In der Praxis werden die beiden Trainingsmethoden zumeist kombiniert (c).

a. Training mit lokalem dauerhaften Datenkorpus

aa. Vorteile

Für ein effektives Training eines neuronalen Netzes werden Text-Daten "vorbereitet" (Reduktion auf den Text, Zerlegung in Token). Die vorverarbeiteten Dokumente werden (meist mit optimierter Hardware) paralle-

8 RL (EU) 2019/790, ABl. L 130/92.

9 So auch BT-Drs. 19/27426, S. 88.

lisiert trainiert. Dies erfolgt am besten mit einem dauerhaft lokal gespeicherten Korpus. Die Trainingsmethode bietet den Vorteil, dass *Vorverarbeitungsschritte nur einmalig ausgeführt werden müssen und Daten effizient im Trainingsprozess erreichbar sind*. Probleme der Übertragung und der Unzuverlässigkeit von Web-Ressourcen werden vermieden. Zudem können Fehler beim Trainingsprozess einfacher nachvollzogen, statistische Verzerrungen ("Bias") im Trainingskorpus verhindert und die Leistungsfähigkeit des Modells effizient gegen den lokalen Korpus geprüft werden.

Dadurch lassen sich *unrichtige Daten leichter berichtigen*, was nicht zuletzt für die Erfüllung von Ansprüchen (z.B. Recht auf Berichtigung gemäß Art. 16 DSGVO) und anderen rechtlichen Pflichten wichtig sein kann. Werden die Trainingsdaten dauerhaft gespeichert, kann bei der Generierung von Antworten zudem analysiert werden, ob ein hinreichender Abstand zwischen den verwendeten Trainingsdokumenten und den vom Modell generierten Ausgaben besteht (zu § 23 UrhG unten A I 3).

bb. Möglichkeiten der Auslesung mit oder ohne temporäre Vervielfältigungen im Arbeitsspeicher

Das Training der Modelle mit einem lokalen Korpus erfordert *technisch keine Vervielfältigung*.

In der *gegenwärtigen Praxis* wird der Korpus indessen vervielfältigt, etwa um Arbeitsschritte zu parallelisieren und die Prozessoren effektiv auszulasten. Ausgehend von einem (vergleichsweise) langsamen Datenträger ("Festplatte") wird der Korpus in einen schnellen Arbeitsspeicher (bzw. GPU-Speicher) übertragen und hierbei vervielfältigt. Diese Zwischenspeicherung ("Caching") im Arbeitsspeicher ist eine nicht zwingend erforderliche Optimierung, wenngleich sie in der Praxis aufgrund der Größe der LLMs für einen technisch effizienten Trainingsprozess wichtig ist.

b. Training mit Webdokumenten ohne lokalen dauerhaften Datenkorpus

Eine Alternative zur Nutzung eines dauerhaften lokalen Korpus besteht in der Nutzung von Web-Dokumenten, die lediglich temporär auf den zum Training des Sprachmodells genutzten Computer übertragen werden. Das bedeutet, dass Dokumente per "Crawler" aus dem Internet auf die zum Training genutzte Maschine geladen werden. Bei der Übertragung

der Dokumente findet eine *technisch notwendige Vervielfältigung* statt, weil die vom Webserver abgerufene Datei auf der zum Training genutzten Maschine abgelegt wird. Die (relativ) langsame und (relativ) fehleranfällige Netzwerkverbindung erfordert, dass die Web-Dokumente auf dem zum Trainieren des LLM genutzten Computer zwischengespeichert werden, um dann die Inhalte der Web-Dokumente für das Training aufzubereiten und das Training des LLMs durchzuführen. Die Zwischenspeicherung wird dabei im Allgemeinen bereits durch die Standard-Komponenten des Betriebssystems vorgenommen.

Eine *dauerhafte Speicherung ist nicht erforderlich*. Die notwendigen Schritte zur Vorbereitung einer Datei an die Erfordernisse des Trainingsalgorithmus ("Vorverarbeitung") können im flüchtigen Arbeitsspeicher erfolgen. Die Webdokumente werden zum Trainieren des Modells, namentlich zum Einstellen der geeigneten Parameter genutzt. Nach dem Training werden die Webdokumente wieder gelöscht. Um das Training effizient durchführen zu können, werden hinreichend große Caches benötigt, die insbesondere das langsame, unzuverlässige Webcrawling ausgleichen müssen. Zudem sind hinreichende Netzwerkressourcen notwendig, um eine hinreichende Menge von Webdokumenten sammeln zu können.

Mit dem Verzicht auf einen lokalen dauerhaften Datenkorpus gehen *Nachteile* einher, wie die Gefahr einer schlechten Nachvollziehbarkeit des trainierten Modells und Verzerrungen im Trainingsdatensatz. So können über schnelle Netzwerk-Leitungen erreichbare Dokumente schneller in größerer Menge gesammelt werden als Dokumente von "langsamen Webservern", weshalb von schnellen Dokumentenquellen bereitgestellte Dokumente meist beim Training bevorzugt werden, da diese schneller zur Verfügung stehen. Das kann dazu führen, dass Meinungen und Informationen aus bestimmten Regionen oder Themenbereichen beim Training unterrepräsentiert werden, da sie von langsamen Webservern gecrawlt werden müssen.

c. Kombination der Trainingsmethoden: Nutzung von Webdokumenten als Eingabe für Sprachmodelle

Beim Training von Sprachmodellen ist zwischen verschiedenen Phasen zu unterscheiden. Die Modelle erlernen das Verständnis von Sprache, trainieren Wissen (ggf. für ausgewählte Fachgebiete bzw. über aktuelle Entwicklungen) und erlernen den Umgang mit Fragen bzw. Anweisungen

("Prompts"). Vortrainierte Modelle können lediglich das in den Trainingsdaten vorhandene Wissen berücksichtigen.

Sollen besonders aktuelle bzw. besonders geschützte Dokumente bei der Generierung von Ausgaben berücksichtigt werden, wird das *Retrieval Augmented Generation (RAG)* als Methode eingesetzt. Dabei nutzt man den Umstand aus, dass aktuelle Sprachmodelle im Prompt mehrere Millionen Zeichen erlauben. Die aktuellen bzw. urheberrechtlich sensiblen Dokumente werden nicht zum Trainieren des Modells, sondern lediglich als Teil des Prompts genutzt. Dieses Vorgehen bietet folgende Vorteile:

- Die aktuellen (Web-)Dokumente werden *nur temporär benötigt* und nur für die Prompt-Erstellung genutzt. Eine dauerhafte Speicherung der zur Prompt-Erstellung verwendeten Dokumente ist nicht notwendig.
- Sollen für die Generierung von Antworten *sowohl urheberrechtlich geschützte als auch gemeinfreie Werke verwendet* werden, besteht die Möglichkeit, das Sprachmodell (LLM) lediglich mit den gemeinfreien Werken zu trainieren und die geschützten Dokumente zur Generierung von Prompts zu nutzen¹⁰.
- Nach der Generierung der Antwort kann geprüft werden, ob ein *hinreichender Abstand* zwischen den im Prompt genutzten Dokumenten und der generierten Ausgabe gegeben ist, da das Prompt temporär bis zum Abschluss des Verfahrens im Arbeitsspeicher vorgehalten wird (zu § 23 UrhG unten A I 3).
- Da bei der Generierung der Antwort vorrangig Informationen aus den Dokumenten des Prompts genutzt werden, ist die *Antwort einfacher nachvollziehbar*. Die zur Antwortgenerierung genutzten Informationen werden vorrangig aus dem Prompt extrahiert.
- Traditionelle Suchverfahren zur Ermittlung relevanter Informationen ("Websuche") werden mit Sprachmodellen zur Generierung natürlicher Antworten kombiniert.

Der limitierende Faktor ist die begrenzte Größe des Prompts. Allerdings sind viele aktuelle Modelle in der Lage, große Prompts zu verarbeiten.

10 Ob Letzteres urheberrechtlich zulässig ist und wem die Handlung zuzurechnen ist, gehört nicht zum Untersuchungsgegenstand.

2. Dauerhafte Vervielfältigungen im trainierten System?

Das trainierte System wird auf einer Festplatte gespeichert. Die kodierten Informationen (Daten) auf der Festplatte beinhalten aus den Werken extrahierte Informationen über Muster. Einzelne Informationen, also der Bedeutungsinhalt (die Semantik) können mit den Trainingswerken identisch sein. Technisch betrachtet ist die Repräsentation, also die äußere Form der Darstellung insofern anders, als die Ladezustände der Speicherzellen nicht einander entsprechen. Allerdings ist Letzteres für eine Vervielfältigung nicht zwingend erforderlich. Etwa wird die Abspeicherung in einem abweichenden Format oder in komprimierter Form für ausreichend erachtet¹¹, obgleich in diesen Fällen die Ladezustände der Speicherzellen nicht identisch sind. Hierfür spricht die Ratio: Entscheidend ist nicht die technische Repräsentation in Gestalt der Ladezustände, sondern dass eine körperliche Festlegung in den Speicherzellen erfolgt, die *geeignet ist, das Werk den menschlichen Sinnen wahrnehmbar zu machen*¹². Für dieses Verständnis streitet ferner der Wortlaut des § 16 Abs. 1 UrhG („gleichviel [...] in welchem Verfahren“) und des Art. 2 InfoSoc-RL¹³ („auf jede Art und Weise und in jeder Form“). Eine dauerhafte Vervielfältigung im trainierten System liegt demnach vor, wenn durch bestimmte Prompts (Anweisungen an die Software) *einzelne Werke oder geschützte Werkteile wiederherstellbar sind*.¹⁴

Sofern eine Wiederherstellung der Werke nur ausgeschlossen werden kann, indem Vervielfältigungen der Werke dauerhaft abgespeichert werden, wäre die Speicherung als Vervielfältigung zu klassifizieren. *Die juristisch*

11 J. Heerma in: A. Wandtke/W. Bullinger (Hrsg.), UrhR-Kommentar, 6. Aufl., München 2022, § 16 UrhG Rn. 5.

12 H.-P. Götting in: H.-P. Götting/A. Lauber-Rönsberg/N. Rauer (Hrsg.), BeckOK UrhR, Stand 1.5.2024, § 16 UrhG Rn. 4; Schulze (Fn. 7), § 16 Rn. 6.

13 RL 2001/29/EG, ABl. L 167/10.

14 Dahingehend auch M. von Welser, Generative KI und Urheberrechtsschranken, GRUR-Prax 2023, 516 Rn. 13: „Die Formulierung „auf jede Art und Weise und in jeder Form“ schließt auch Vorgänge ein, die in künstlichen neuronalen Netzen stattfinden“; P. Pesch/R. Böhme, Artpocalypse now? – Generative KI und die Vervielfältigung von Trainingsbildern, GRUR 2023, 997 (1005): „Eine Vervielfältigung liegt in Fällen von nahezu vollständig in KI-Modelle übersetzten Bildinformationen von Trainingsbildern nur in solchen Fällen nicht vor, in denen diese Festlegung dauerhaft nicht dazu geeignet ist, das Trainingsbild zu rekonstruieren“; a.A. (beim Training einer KI würden nicht die Trainingsdaten in die KI implementiert) J. Pukas, KI-Trainingsdaten und erweiterte kollektive Lizenzen, GRUR 2023, 614 (618); Maamar Fragen (Fn. 2), 483; unentschlossen R. Heine, Generative KI: Nutzungsrechte und Nutzungsvorbehalt, GRUR-Prax 2024, 87 Rn. 5.

entscheidungserhebliche technische Frage lautet demnach, ob eine Wiederherstellung der Werke ausgeschlossen und kumulativ das Trainingsmaterial gelöscht werden kann.

Große Sprachmodelle "lernen" anhand von Beispielen die in der Trainingsmenge gegebenen Regeln und Muster. Ziel ist es, die Beispiele so zu generalisieren, dass die gelernten Muster auf neue Eingaben übertragen werden können. Technisch basieren die aktuellen großen Sprachmodelle auf tiefen künstlichen *neuronalen Netzen*. Die Knoten im Netzwerk sind dabei mit gewichteten Kanten verbunden, die festlegen, wie stark ein Knoten einen anderen Knoten beeinflussen soll. Während des Trainings werden die Gewichtungen inkrementell angepasst, so dass das Netzwerk zu gegebenen Eingaben die gewünschte Ausgabe liefert.

Beim Training wird mit Sätzen oder Absätzen gearbeitet, um dem Modell die *Regeln der Sprache mit einem angemessenen Kontext anzutrainieren*. Beim Trainieren eines Sprachmodells werden keine Trainingsdokumente im Modell abgespeichert (Modell-zentrierter Ansatz im Gegensatz zu Instanz-zentrierten KI-Verfahren), sondern lediglich Gewichte, die Knoten im Netzwerk verbinden. Das Trainieren entspricht dabei dem Vorgehen beim Data Mining, bei dem Muster aus Trainingsdaten gelernt werden.

Bei der *Generierung von Ausgaben werden Wahrscheinlichkeitsmodelle genutzt*, um geeignete Token bzw. Wörter für die Ausgabe zu ermitteln. Bei der Generierung von Ausgaben wird zudem meist unter Nutzung eines Zufallszahlengenerators ein Wort aus mehreren Synonymen ausgewählt, dass mit einer hohen Wahrscheinlichkeit geeignet ist. Dies kann technisch über den Startwert (Seed) des Zufallszahlengenerators sowie den Parameter "Temperature" gesteuert werden, der festlegt, wie stark populäre Wörter präferiert werden sollen. Zudem werden beim Training von künstlichen neuronalen Netzen zumeist Methoden (z. B. "Pruning", "Drop-Out") eingesetzt, die das exakte "Auswendiglernen" von Trainingsbeispielen verhindern sollen. Das für das Training von Sprachmodellen genutzte Verfahren lässt es sehr unwahrscheinlich erscheinen, dass beim Training genutzte Texte exakt wiederhergestellt werden. Allerdings gibt es praktische Beispiele, bei denen gezeigt wurde, dass durch scheinbar sinnlose Prompts, Sprachmodelle Trainingsdaten wieder ausgeben.

Zusammenfassend ist die 1:1-Wiedergabe von Trainingsdaten unwahrscheinlich; die Wahrscheinlichkeit der exakten Wiederherstellung wird durch technische Maßnahmen reduziert. Die 1:1-Wiederherstellung ist aber trotzdem möglich und kann nicht vollständig ausgeschlossen werden. Da Trainingsdaten für Sprachmodelle in der Regel nicht gespeichert werden,

kann das Sprachmodell zumeist nicht erkennen, ob Trainingsdaten 1:1 als Ausgabe geliefert werden.

3. Exkurs: Vervielfältigungen durch Ausgaben des Systems

Auf der Output-Seite einer generativen KI gilt es, dem Urheber vorbehaltene Wiederherstellungen des Werkes (eins-zu-eins-Vervielfältigungen § 16 UrhG), dem Urheber vorbehaltene Bearbeitungen (§ 23 Abs. 1 S. 1 UrhG) und freie Benutzungen bei hinreichendem Abstand (§ 23 Abs. 1 S. 2 UrhG) abzugrenzen. Ein hinreichender Abstand ist bei *fehlender Wiedererkennbarkeit* der schutzbegründenden eigenschöpferischen Elemente gewährt.¹⁵

Die Output-Seite gehört nicht zum Text und Data Mining. Die gesetzlichen Schranken der §§ 44b, 44a UrhG sind daher offenkundig nicht einschlägig. Sofern kein hinreichender Abstand besteht, benötigt der Anbieter der generativen KI demnach eine *vertragliche Befugnis*.

4. Training generativer KI-Anwendungen als unbenanntes Verwertungsrecht?

Die *bloße Benutzung* eines Werkes als solche ist kein urheberrechtlich relevanter Vorgang. So verglich der BGH die Benutzung eines Computerprogramms mit dem Lesen eines Buches, dem Anhören einer Schallplatte und dem Betrachten eines Kunstwerkes.¹⁶ Die Tatsache, dass der Werkgenuss durch Menschen urheberrechtlich freigestellt ist, bedeutet indessen nicht zwangsläufig, dass auch die „elektronische Lektüre“ zum Zweck des Text und Data Mining frei wäre. Die Folgerung „the right to read is the right to mine“¹⁷ ist unzutreffend, weil das Reading durch Menschen und das Mining durch algorithmische Systeme erfolgt. Das zweckunabhängige „right

15 BGH ZUM 2022, 547 (552); K. Grisse/ C. Kaiser, Freiheit ab Unkenntlichkeit? – Die Bedeutung der (fehlenden) Wiedererkennbarkeit für das urheberrechtliche Vervielfältigungsrecht, ZUM 2021, 401.

16 BGH GRUR 1991, 449 (453).

17 Pukas KI-Trainingsdaten (Fn. 14), 619. Die Anführungszeichen finden sich im Original; ähnlich Maamar Fragen (Fn. 2), 483; F. Hofmann, Retten Schranken Geschäftsmodelle generativer KI-Systeme?, ZUM 2024, 166 (170); F. Hofmann, Zehn Thesen zu Künstlicher Intelligenz (KI) und Urheberrecht, WRP 2024, 11 Rn. 23.

to read“ ist für Menschen anerkannt. Für eine „elektronische Lektüre“ zu sämtlichen Zwecken folgt daraus kein Argument.¹⁸

Auf *Unionsebene* existiert kein Verwertungsrecht für das Training generativer KI als solches¹⁹ – die DSM-Richtlinie normiert lediglich Schranken für das Vervielfältigungsrecht. Zudem gibt es weder einen abschließenden Katalog an Verwertungsrechten noch eine mit §§ 11, 15 UrhG vergleichbare Generalklausel.

Folglich könnten §§ 11, 15 UrhG als Grundlage eines unbenannten Verwertungsrechts (Innominatfall) für das Training generativer KI in *Deutschland* dienen, weil sie die Verwertungsrechte nur beispielhaft auflisten.²⁰ Innominatfälle kommen insbesondere bei Digitalisierungssachverhalten in Betracht.²¹ Das Lesen eines Buches oder Betrachten eines Gemäldes ist vor allem deshalb urheberrechtlich nicht relevant, weil sich diese Handlungen der Konsumenten nahezu nicht kontrollieren lassen und der Gesetzgeber die Verwertungsrechte daher bei den Handlungen der Werkvermittler ansetzt.²² Die Anbieter generativer KI sind bei normativer Betrachtung eher mit Werkvermittlern vergleichbar als mit lesefreudigen Verbrauchern. Die Softwareunternehmen ziehen wirtschaftliche Früchte aus der Werkanalyse. Das streitet für die (rechtsunsichere) Annahme eines unbenannten Verwertungsrechts.

Zu wünschen wäre eine *Entscheidung des Unionsgesetzgebers*, andernfalls des Bundesgesetzgebers. Bis dahin obliegt es der *Rechtsprechung*, den nationalen Tatbestand eines (noch) unbenannten Verwertungsrechts für das Training generativer KI und die dazugehörigen Schranken festzulegen.²³ Betreffend die Schranken müsste das Gericht etwaige gesetzgeberische

18 Ausführlich zur „begriffliche[n] Verwirrung um das „right to mine““ *T. Dornis/S. Stober*, Urheberrecht und Training generativer KI-Modelle, Baden-Baden 2024, S. 103 ff. Die Anführungszeichen finden sich im Original.

19 Ein entsprechendes Verwertungsrecht wohl annehmend *R. Meys*: Data Mining Under the Directive on Copyright and Related Rights in the Digital Single Market, GRUR Int. 2020, 457 (465): „The DSM Directive acknowledges that the practice of text and data mining may involve the reproduction of copyrighted materials as well as the *extraction of contents* from a protected database, which cannot be accomplished without the consent of the relevant rightholders in the absence of any applicable exception“. Die Kursivierung findet sich nicht im Original.

20 Abstrakt (nicht zum KI-Training) *Schulze* (Fn. 7), § 15 Rn. 10.

21 *Schulze* (Fn. 7), § 15 Rn. 13.

22 *H. Schack*, Urheber- und Urhebervertragsrecht, 10. Aufl., Tübingen 2021, Rn. 429.

23 Abstrakt (nicht zum KI-Training) *J. v. Ungern-Sternberg* in: U. Loewenheim/M. Leister/A. Ohly (Hrsg.), Schricker/Loewenheim UrhR Kommentar, 6. Aufl., München 2020, § 15 UrhG Rn. 265.

Wertentscheidungen in den §§ 44a, 44b UrhG beachten: *Wenn und soweit im Rahmen des Trainings generativer KI Vervielfältigungen gesetzlich zulässig sein sollten (dazu sogleich II), umfasst die Gestattung als Minus auch das Training generativer KI als solches.*

II. Rechtsunsicherheit hinsichtlich der Schranken

Wird im Rahmen des Text und Data Mining ein Werk vervielfältigt, muss die Handlung entweder aufgrund einer gesetzlichen Schranke oder einer vertraglichen Befugnis gestattet sein. Weil die Vervielfältigungen nicht zum privaten Gebrauch erfolgen, ist § 53 UrhG nicht einschlägig. Für Zwecke der wissenschaftlichen Forschung enthält § 60d UrhG eine Privilegierung.²⁴ Für die untersuchungsgegenständlichen kommerziell eingesetzten generativen KI-Anwendungen kommen die Schrankenregelungen des § 44b UrhG (1) und des § 44a UrhG (2) in Betracht, die auf unionsrechtlicher Ebene dem Drei-Stufen-Test standhalten müssen (3).

1. Art. 4 DSM-RL, § 44b UrhG bei Text und Data Mining mit oder ohne Datenkorpus

Die Art. 4 DSM-Richtlinie umsetzende Vorschrift § 44b UrhG gestattet gemäß ihres Absatz 2 Satz 1 Vervielfältigungen von rechtmäßig zugänglichen Werken für das Text und Data Mining.²⁵ Weil die Schranke temporäre und dauerhafte Vervielfältigungen erlaubt, bietet sie sich potenziell für alle Trainingsmethoden an.

a. Tendenziell keine Anwendbarkeit für das Training generativer KI

Text und Data Mining bezeichnet ausweislich der Legaldefinition des Art. 2 Nr. 2 DSM-Richtlinie „eine Technik für die automatisierte Analyse von Texten und Daten in digitaler Form, mit deren Hilfe Informationen unter

²⁴ Aufgrund des Verweises von § 60d Abs. 1 UrhG auf § 44b Abs. 2 S. 1 UrhG muss das Werk für den Nutzer rechtmäßig zugänglich sein (D. Bomhard in: H.-P. Götting/A. Lauber-Rönsberg/N. Rauer (Hrsg.), BeckOK UrhR, Stand 15.2.2024, § 60d UrhG Rn. 15).

²⁵ Etwaige technisch bedingte Änderungen, etwa Umwandlungen in andere Dateiformate, gestattet § 23 Abs. 3 UrhG.

anderem – aber nicht ausschließlich – über Muster, Trends und Korrelationen gewonnen werden können.“

Generative KI wird mit dem gesamten Inhalt geistiger Schöpfungen trainiert. Dies lässt sich allenfalls schwerlich unter das Gewinnen von (bestimmten) Informationen subsumieren, zumal die drei genannten Beispiele erheblich hinter einer vollständigen Erfassung des Inhalts zurückbleiben.²⁶ Auch werden die (umfassenden) Informationen nicht der menschlichen Wahrnehmung zugänglich gemacht, sondern verbleiben im trainierten System, was nach dem allgemeinen Sprachgebrauch für ein „Gewinnen“ nicht ausreichend ist.²⁷ Folglich streitet der *Wortlaut* tendenziell gegen eine Anwendung der Schranke für generative KI.

In *historischer* Hinsicht legen die Erwägungsgründe der DSM-Richtlinie nahe, dass der Unionsgesetzgeber im Jahr 2019 keine generative KI vor dem geistigen Auge hatte: Die gesetzgeberische Einschätzung, es „besteht weitgehend Einvernehmen darüber, dass diese Technik vor allem für die Forschung von besonderem Nutzen ist“²⁸, passt nicht zu ChatGPT und anderen generativen KI-Anwendungen. Wenngleich es an anderer Stelle heißt, ein Ziel sei die „Privatwirtschaft zu Innovationen anzuregen“²⁹, legen die Erwägungsgründe in der Gesamtschau nahe, dass damit die Entwicklung neuer Produkte (und nicht von Konkurrenzprodukten) gemeint ist.³⁰

Hieran vermag der Umstand nichts zu ändern, dass der Unionsgesetzgeber 2024 in der KI-Verordnung von einer Anwendbarkeit der DSM-Richtlinie auf generative KI ausgeht³¹. Die KI-Verordnung ändert die DSM-Richtlinie nicht. Eine – je nach Ansicht – deklaratorische Klarstellung oder

26 Ähnlich *H. Schack*, Auslesen von Webseiten zu KI-Trainingszwecken als Urheberrechtsverletzung de lege lata et ferenda, NJW 2024, 113 Rn. 8; tendenziell a.A. (eine Unterscheidung zwischen „in den Daten verborgenen Informationen“ und „dem Inhalt der geistigen Schöpfung“ sei nicht möglich) LG Hamburg GRUR-RS 2024, 25458 Rn. 44.

27 Die englische Fassung lautet „in order to generate information“.

28 Erwägungsgrund 8 S. 3 RL (EU) 2019/790, ABl. L 130/92.

29 Erwägungsgrund 18 S. 4 RL (EU) 2019/790, ABl. L 130/92; u.a. deshalb die Anwendbarkeit der Norm für generative KI bejahend *Maamar* Fragen (Fn. 2), 483.

30 Ausführlich zum europäischen Gesetzgebungsverfahren *Dornis/Stober*, Training (Fn. 18), S. 122 ff.

31 Erwägungsgründe 105, 99 (große generative KI-Modelle sind ein typisches Beispiel für ein KI-Modell mit allgemeinem Verwendungszweck), Art. 53 Abs. 1 lit. c (geht davon aus, dass Anbieter von KI-Modellen mit allgemeinem Verwendungszweck Text und Data Mining iSd Art. 4 DSM-RL vornehmen können) VO (EU) 2024/1689, ABl. L 12.7.2024; LG Hamburg GRUR-RS 2024, 25458 Rn. 48 und *D. Bomhard* in: H.-P. Götting/A. Lauber-Rönsberg/N. Rauer (Hrsg.), BeckOK UrhR, Stand 15.2.2024,

konstitutive Änderung des Urheberrechts müsste als solche deklariert werden. Deshalb ist ein *systematischer* Schluss, dass generative KI aufgrund der KI-Verordnung vom Text und Data Mining-Begriff der DSM-Richtlinie umfasst sei, abzulehnen.

Gegen eine Erfassung generativer KI streiten schließlich Art. 4 Abs. 2 DSM-RL, § 44b Abs. 2 S. 2 UrhG: Vervielfältigungen dürfen nur so lange aufbewahrt werden, wie es für die Zwecke des Text und Data Mining notwendig ist. Text und Data Mining bezeichnet das Training und nicht das trainierte System. Der Gesetzgeber geht demnach davon aus, dass *sämtliche Vervielfältigungen nach dem Training löscher* sind. Weil bei generativer KI durch geschickte Prompts einzelne Werke oder geschützte Werkteile wiederherstellbar sind³², existieren die Vervielfältigungen im rechtlichen Sinne indessen so lange wie es das trainierte System gibt. Aufgrund der potenziellen Wiederherstellbarkeit ist eine vollständige Löschung im rechtlichen Sinne nicht möglich. Auch dieser Gedankengang offenbart, dass der Gesetzgeber in der DSM-Richtlinie keine generative KI erfassen wollte. Sicher ist vorliegendes Verständnis vom Anwendungsbereich der DSM-Richtlinie allerdings nicht – Klarheit vermag erst eine EuGH-Entscheidung zu schaffen.³³

Das LG Hamburg sah 2024 *die Erstellung eines Datensatzes*, der auch für das Trainieren generativer KI nutzbar ist³⁴, als gemäß § 44b UrhG erlaubte Nutzungshandlung an³⁵, ließ hingegen explizit offen, ob die Schranke auch „das Training von Künstlicher Intelligenz in seiner Gesamtheit“³⁶ gestattet, weil es diese Frage Streitgegenständlich für nicht entscheidungserheblich hielt. Die im Urteil vorgenommene *Differenzierung zwischen der (streitgegenständlichen) Erstellung eines Trainingsdatensatzes und dem nachfolgenden Training*³⁷ ist technisch und chronologisch richtig. Unzutreffend ist indessen, die Erstellung eines Trainingsdatensatzes deshalb als von der

§ 44b UrhG Rn. 11b halten diesen systematischen Schluss für bedeutsamer als die Verfasser.

32 Oben A I 2.

33 Anwendbarkeit verneinend *Schack*, Auslesen (Fn. 26), Rn. 8; bejahend *F. Hofmann*, Thesen (Fn. 17), Rn. 15; *de la Durantaye*, »Garbage in, garbage out« – Die Regulierung generativer KI durch Urheberrecht, ZUM 2023, 645 (651); *Maamar* Fragen (Fn. 2), 483 (ohne Begründung); *Heine*, Generative KI (Fn. 14), Rn. 12; zweifelnd *Pukas* KI-Trainingsdaten (Fn. 14), 618.

34 LG Hamburg GRUR-RS 2024, 25458 Rn. 2.

35 LG Hamburg GRUR-RS 2024, 25458 Rn. 38 ff.

36 LG Hamburg GRUR-RS 2024, 25458 Rn. 40.

37 LG Hamburg GRUR-RS 2024, 25458 Rn. 45.

Schranke umfasst anzusehen, weil die anschließenden Verwendungszwecke möglicherweise noch nicht abschließend feststehen³⁸: Die Schranke der Art. 4 Abs. 1 DSM-RL, § 44b Abs. 2 S. 1 UrhG gestattet ausweislich ihres Wortlauts lediglich *zum Zwecke* des Text und Data Mining vorgenommene Vervielfältigungen. Steht die Verwendung des Datensatzes zu einer Form des Text und Data Mining nicht bereits anfänglich fest, so deckt die Schranke die vorgenommenen Vervielfältigungen zur Erstellung des Datensatzes nicht ab, weil ihre tatbestandlichen Voraussetzungen nicht erfüllt sind. Entgegen der Kammer ist eine isolierte „Beurteilung der rechtlichen Zulässigkeit der Erstellung des Trainingsdatensatzes als solche[r]“³⁹ nicht möglich. Vielmehr ist die Zielsetzung miteinzubeziehen, mithin bilden die Vervielfältigungen, die Erstellung des Trainingsdatensatzes und das beabsichtigte Training in seiner konkreten Form rechtlich ein unteilbares Ganzes. Vervielfältigungen zur Erstellung eines Trainingsdatensatzes sind lediglich dann zulässig, wenn der beabsichtigte anschließende Trainingsablauf als Text und Data Mining im Sinne der Schranke zu klassifizieren ist.⁴⁰

Unter der Prämisse, dass die Schranke der Art. 4 DSM-RL, § 44b UrhG Vervielfältigungen zum Training generativer KI nicht erfasst, muss das Softwareunternehmen *für die Bildung eines Datenkorpus Vervielfältigungstücke erwerben*.

b. Maschinenlesbarer Nutzungsvorbehalt

Unter der gegenteiligen Prämisse, dass die Schranke das Training generativer KI umfasst, gestatten Art. 4 Abs. 3 DSM-RL, § 44b Abs. 3 UrhG den Rechtsinhabern, einen Nutzungsvorbehalt zu erklären. Bei online zugänglichen Werken muss dies in maschinenlesbarer Form erfolgen. Die (unklaren) Anforderungen für eine Maschinenlesbarkeit⁴¹ weisen ebenso keinen Bezug zur Forschungsfrage auf, wie die (unklare) Informationspflicht in

38 Dahingehend aber LG Hamburg GRUR-RS 2024, 25458 Rn. 46: „Die konkreten Anwendungsmöglichkeiten bei einer sich rasant entwickelnden Technologie wie der KI sind zum Zeitpunkt der Erstellung des Trainingsdatensatzes daher nicht abschließend absehbar und mithin nicht rechtssicher feststellbar“.

39 LG Hamburg GRUR-RS 2024, 25458 Rn. 46.

40 Bei Redaktionsschluss lagen noch keine Urteilsanmerkungen vor.

41 Das LG Hamburg legt den Begriff iSv „Maschinenverständlichkeit“ aus und stellt an diese niedrige Anforderungen, indem es den Einsatz von KI-Anwendungen verlangt (LG Hamburg GRUR-RS 2024, 25458 Rn. 65, 70).

der KI-Verordnung⁴², die den Inhabern von Urheberrechten die Ausübung ihres Vetorechts erleichtern soll⁴³. *Um die Optionen für einen kohärenten Rechtsrahmen zu systematisieren, genügt es zu reflektieren, dass die Möglichkeit eines Opting-Out prinzipiell besteht.*

2. Art. 5 InfoSoc-RL, § 44a UrhG bei Text und Data Mining ohne Datenkorpus

Die Schranke erlaubt temporäre Vervielfältigungshandlungen (a), deren alleiniger Zweck es ist, eine rechtmäßige Nutzung eines Werkes zu ermöglichen, und die keine eigenständige wirtschaftliche Bedeutung haben (b).

a. Temporäre Vervielfältigungen statt Datenkorpus

aa. Relevanz der Abgrenzung

Eine generative KI kann entweder mit einem lokalen dauerhaften Datenkorpus oder mit Webdokumenten ohne dauerhaften Datenkorpus trainiert werden.⁴⁴ Die Unterscheidung zwischen dauerhaften und temporären Vervielfältigungen vermag – in Abhängigkeit von der Rechtsauffassung betreffend § 44b UrhG – in folgenden Konstellationen rechtliche Relevanz zu zeitigen:

Sofern § 44b UrhG für das Training generativer KI nicht anwendbar ist, können Vervielfältigungen allein durch § 44a UrhG gerechtfertigt werden. Das Softwareunternehmen muss folglich entweder Vervielfältigungsstücke von Werken entgeltlich erwerben, mit diesen einen Datenkorpus bilden (bis hierhin nimmt es keine urheberrechtlich relevante Handlung vor) und darf sodann (wenn § 44a UrhG einschlägig ist) die generative KI mittels temporärer Vervielfältigungen trainieren. Oder das Unternehmen liest die Werke mittels vorübergehender Vervielfältigungen im Arbeitsspeicher direkt aus dem Internet aus.

Sofern § 44b UrhG (entgegen hiesiger Auffassung) für generative KI anwendbar ist, hat § 44a UrhG eine eigenständige Bedeutung, wenn der Urhe-

42 Art. 53 Abs. 1 lit. d VO (EU) 2024/1689, ABl. L 12.7.2024; dazu ausführlich und überzeugend *de la Durantaye* Garbage (Fn. 33), 655 f.

43 Erwägungsgrund 107 VO (EU) 2024/1689, ABl. L 12.7.2024.

44 Oben A I 1.

ber den Nutzungsvorbehalt (in maschinenlesbarer Form) erklärt. Sodann kann das Unternehmen das betroffene Werk entgeltlich erwerben und dem Datenkorpus hinzufügen oder das Werk mittels temporärer Vervielfältigungen direkt aus dem Internet auslesen. *Die Schranke des § 44a UrhG kennt kein Opting-Out.*

bb. Praktische Abgrenzungsschwierigkeiten

Die Abgrenzung zwischen dauerhaften und temporären Vervielfältigungen ist demnach, wenn man § 44b UrhG für nicht anwendbar hält, sehr bedeutend, und wenn man § 44b UrhG für anwendbar erachtet, im Falle des Opting-Out relevant.

Die in früheren Jahren einfache Unterscheidung zwischen flüchtigem Arbeitsspeicher ("RAM") und persistentem Festplattenspeicher ("HDD") hat sich in den letzten Jahren verkompliziert, da *Hybrid-Systeme zum Standard geworden sind*, die sehr große Mengen von flüchtigem Arbeitsspeicher (RAM) mit etwas langsameren persistenten Speicher (SSDs) sowie sehr großen Magnet-Festplatten ("HDD") dynamisch kombinieren. Persistente Datenspeicher verwenden unterschiedliche Technologien zum Speichern von Informationen, sodass das sichere Löschen von Daten teilweise mit technischem Aufwand verbunden ist und Daten auf Datenträgern beim Löschen lediglich als "freigegeben" markiert werden, ohne dass die Daten physikalisch gelöscht wären. *In der Praxis bedeutet dies, dass bei temporären Vervielfältigungen der Speicherplatz durch das Lernprogramm für neue Aufgaben freigegeben wird, aber erst dann gelöscht bzw. überschrieben wird, wenn der Platz benötigt wird.* Daher kann es beim Eingriff in die Steuerungssoftware der Speicherelemente (z.B. von SSD-Festplatten) möglich sein, als gelöscht markierte Daten wiederherzustellen. Vor diesem Hintergrund *sind temporäre und dauerhafte Vervielfältigungen technisch nicht (mehr) trennscharf abgrenzbar.*

Ausweislich der Gesetzgebungsmaterialien sind sich der *Unions- und Bundesgesetzgeber über die Abgrenzungsschwierigkeiten bewusst.* So wurde die Schranke des Art. 4 DSM-RL auch geschaffen, um für mehr Rechtssicherheit zu sorgen: Die Nutzer des Text und Data Mining könnten mit Rechtsunsicherheit hinsichtlich der Frage konfrontiert sein, „ob Vervielfältigungen und Entnahmen zum Zwecke des Text und Data Mining bei rechtmäßigem Zugang zu Werken oder sonstigen Schutzgegenständen vor-

genommen werden dürfen, insbesondere wenn bei den zum Zwecke der Durchführung des technischen Vorgangs vorgenommenen Vervielfältigungen und Entnahmen möglicherweise nicht alle Bedingungen der für vorübergehende Vervielfältigungshandlungen in Artikel 5 Absatz 1 der Richtlinie 2001/29/EG vorgesehenen Ausnahme erfüllt sind.⁴⁵ Der Bundesgesetzgeber hat bei der Ausgestaltung von § 44b UrhG auf eine Vergütungspflicht unter anderem verzichtet, weil „die mit Artikel 4 DSM-RL beabsichtigte Rechtssicherheit für Nutzer des Text und Data Mining nicht erreicht werden könnte, wenn in Zukunft ermittelt werden müsste, ob lediglich eine ephemere Kopie erstellt wird (dann nach § 44a UrhG in jedem Fall vergütungsfrei) oder eine sonstige Vervielfältigung (dann gegebenenfalls nach § 44b UrhG-E vergütungspflichtig).“⁴⁶

b. Streit über eigene wirtschaftliche Bedeutung

Die Analyse als solche („elektronische Lektüre“) ist – *sofern man ein unbenanntes Verwertungsrecht verneint*⁴⁷ – keine urheberrechtlich relevante Handlung und damit eine *rechtmäßige Nutzung* im Sinne des § 44a Nr. 2 UrhG. Flüchtige und begleitende Vervielfältigungshandlungen stellen einen integralen und wesentlichen Teil der Analyse als technisches Verfahren dar. Lebhaft diskutiert wird indessen, ob temporäre Vervielfältigungshandlungen beim Training generativer KI „keine eigenständige wirtschaftliche Bedeutung haben“⁴⁸.

Betreffend temporäre Vervielfältigungen im *Speicher eines Satellitenrecorders* subsumierte der EuGH, dass der bloße Empfang der streitgegenständlichen Sendungen als solcher, „also die Erfassung ihres Signals und ihre visuelle Darstellung im privaten Kreis“⁴⁹ keine urheberrechtliche Bedeutung haben, also eine rechtmäßige Nutzung des Werkes sind. Um der Schranke „nicht ihre praktische Wirksamkeit zu nehmen, muss die wirtschaftliche Bedeutung in dem Sinne eigenständig sein, dass sie über den wirtschaftlichen Vorteil, der durch den bloßen Empfang einer Sendung mit geschützten Werken entsteht, das heißt über den Vorteil, der sich aus der

45 Erwägungsgrund 18 S. 3 RL (EU) 2019/790, ABl. L 130/92.

46 BT-Drs. 19/27426, S. 88.

47 Oben A I 4.

48 Art. 5 Abs. 1 RL 2001/29/EG, ABl. L 167/10; § 44a UrhG.

49 EuGH MMR 2011, 817 Rn. 171.

bloßen Erfassung der Sendung und ihrer visuellen Darstellung ergibt, hinausgeht.“⁵⁰ Durch die temporären Vervielfältigungen im Satellitenrecorder entsteht kein zusätzlicher wirtschaftlicher Vorteil. Ferner haben *Produktivitätssteigerungen* keine eigenständige wirtschaftliche Bedeutung, wenn die „wirtschaftlichen Vorteile nur bei der [rechtmäßigen] Nutzung des vervielfältigten Werks entstehen, sodass sie von den Vorteilen, die sich aus der [rechtmäßigen] Nutzung des Werks ergeben, weder zu unterscheiden noch zu trennen sind.“⁵¹

Entsprechendes muss für die *elektronische Werkanalyse* gelten: Aus temporären Vervielfältigungen von Werken oder Werkteilen resultiert kein über die rechtmäßige Analyse hinausgehender Vorteil. Der mittelbare Zweck des Trainings, namentlich softwaregenerierte Konkurrenzprodukte zu schaffen, hat eine eigenständige wirtschaftliche Bedeutung. Das ist nach dem Wortlaut allerdings nicht erheblich. Es erschließt sich nicht, warum die wohl herrschende Meinung⁵² mit dem mittelbaren Zweck eine eigenständige wirtschaftliche Bedeutung des Trainingsablaufs bejaht. Als Kontrollüberlegung zur analogen Welt wird die Lektüre einer fremden Doktorarbeit nicht deshalb urheberrechtlich relevant, weil der Leser das Werk zu einer eigenen Doktorarbeit (Konkurrenzprodukt) umformulieren möchte. Ein über die elektronische Analyse hinausgehender Vorteil bestünde etwa, wenn das Softwareunternehmen die temporären Vervielfältigungen weiter veräußert⁵³ oder ein Computerprogramm (die Anwendbarkeit von § 44a UrhG unterstellt) in die Arbeitsspeicher mehrerer Rechner geladen und dadurch parallel und intensiver genutzt wird⁵⁴ (vergleichbar mit dem Kopieren eines Buches).

In der DSM-Richtlinie geht der Gesetzgeber zutreffend davon aus, dass die Schranke der Art. 5 Abs. 1 InfoSoc-RL, § 44a UrhG „auch künftig auf Verfahren des Text und Data Mining angewandt werden sollte, die nicht die

50 EuGH MMR 2011, 817 Rn. 175.

51 EuGH ZUM 2012, 398 Rn. 51.

52 *Pesch/Böhme*, Artpocalypse (Fn. 14), 1006; *J. Siglmüller/D. Gassner*, Softwareentwicklung durch Open-Source-trainierte KI – Schutz und Haftung, RD 2023, 124 (126); *Schack*, Auslesen (Fn. 26), Rn. 6; unentschlossen *F. Hofmann*, Thesen (Fn. 17), Rn. 14; *Maamar* Fragen (Fn. 2), 485.

53 *G. Spindler*, Text und Data Mining – urheber- und datenschutzrechtliche Fragen, GRUR 2016, 1112 (1115); praktisch erscheint das schwierig vorstellbar und zudem wäre das Verbreitungsrecht tangiert.

54 Dahingehend BGH GRUR 2011, 418 Rn. 17 (zu beachten ist der Plural „weiterer Arbeitsplatzrechner“).

Anfertigung von Kopien in einem über diese Ausnahme hinausgehenden Umfang einschließen.⁵⁵ Die Schranke der Art. 4 DSM-RL, § 44b UrhG erlaubt temporäre und dauerhafte Vervielfältigungen, allerdings erfasst ihr Anwendungsbereich nach vorliegender Auffassung keine generative KI⁵⁶. Demgegenüber gestatten Art. 5 Abs.1 InfoSoc-RL, § 44a UrhG lediglich temporäre Vervielfältigungen, doch erstreckt sich der Zweck auf sämtliche rechtmäßige Nutzungen. *Der Anwendungsbereich ist nicht auf den Text und Data Mining-Begriff der Art. 4 DSM-RL, § 44b UrhG beschränkt und umfasst demnach auch das Training generativer KI.*

3. Tendenziell partielle Unionsrechtswidrigkeit der §§ 44a, 44b UrhG wegen Drei-Stufen-Test

Als Zwischenergebnis gestattet § 44a UrhG temporäre Vervielfältigungen zum Training generativer KI, während § 44b UrhG nicht anwendbar ist. Beide Auffassungen sind mit erheblicher Rechtsunsicherheit behaftet. *Erachtet man die bundesgesetzlichen Schranke(n) bei rein nationaler Betrachtung für anwendbar auf das Training generativer KI, gilt es zu klären, ob die Vorschriften auf unionsrechtlicher Ebene dem Drei-Stufen-Test standhalten, der die Funktion einer „Schrankenschranke“⁵⁷ erfüllt. Wäre das nicht der Fall, so müssten entweder die Schranke(n) praeter legem unionsrechtskonform ausgelegt⁵⁸ und ihre Anwendung für generative KI im Ergebnis verneint werden oder – sofern eine unionsrechtskonforme Auslegung und Fortbildung des Bundesrechts nicht möglich sein sollte – die Schranke(n) wegen des Anwendungsvorrangs⁵⁹ des Unionsrechts unangewendet bleiben.*

55 Erwägungsgrund 9 RL (EU) 2019/790, ABl. L 130/92; der Bundesgesetzgeber teilt dieses Verständnis betreffend § 44a UrhG (BT-Drs. 19/27426, S. 88).

56 Oben A II 1 a.

57 *H.-P. Götting* in: U. Loewenheim (Hrsg.), Handbuch des Urheberrechts, 3. Aufl., München 2021, § 30 Rn. 22. Die Anführungszeichen finden sich im Original.

58 Der EuGH legt Art. 5 Abs.1 InfoSoc-RL im Lichte des Absatz 5 aus und erkennt Absatz 5 keine eigenständige Bedeutung zu (EuGH GRUR 2017, 610 Rn. 63; EuGH ZUM 2012, 398 Rn. 56 f.).

59 Das Unionsrecht ist über Art. 23 Abs.1 GG Bestandteil des Maßstabs der richterlichen Tätigkeit nach Art. 20 Abs.3 GG (*W. Kluth*, Die Haftung der Mitgliedstaaten für gemeinschaftsrechtswidrige höchstrichterliche Entscheidungen - Schlussstein im System der gemeinschaftsrechtlichen Staatshaftung, DVBl 2004, 393 (399)); zum Verhältnis von unionsrechtskonformen Fortbildungen des nationalen Rechts und

Der in Art. 5 Abs. 5 InfoSoc-RL, Art. 7 Abs. 2 DSM-RL verankerte Drei-Stufen-Test verlangt, dass die Schranke(n) „nur in *bestimmten Sonderfällen* angewandt werden, in denen die *normale Verwertung des Werks oder des sonstigen Schutzgegenstands nicht beeinträchtigt* wird und die *berechtigten Interessen des Rechtsinhabers nicht ungebührlich verletzt* werden.“ Anders als § 44a UrhG adressiert der Drei-Stufen-Test nicht lediglich Vervielfältigungshandlungen im Rahmen des Trainings, sondern bezieht mittelbare Auswirkungen ein. Durch Text und Data Mining werden die Trainingswerke mittelbar einer neuen Verwertung zugeführt⁶⁰, weil die Ausgaben der algorithmischen Systeme als Konkurrenzprodukt zu klassifizieren sein können⁶¹. Dadurch wird die normale Verwertung der geschützten Werke beeinträchtigt, weil sich normalerweise die Transaktionen verringern⁶². *Eine ungebührliche Verletzung der berechtigten Urheberinteressen wäre allenfalls bei einer Vergütungspflicht abzulehnen.* Der Drei-Stufen-Test legt demnach nahe, Text und Data Mining zum Training generativer KI als unionsrechtlich verboten einzugruppieren. Diese Auffassung muss ebenfalls als rechtsunsicher gelten.⁶³

Weil der Drei-Stufen-Test neben dem Unionsrecht auch in *völkerrechtlichen Verträgen* verankert ist (Art. 9 Abs. 2 RBÜ, Art. 13 TRIPS⁶⁴), würde seine Abschaffung eine vorherige Änderung der Verträge erfordern.

III. Keine Geltung europäischer Urheberrechtsstandards für Training in Drittstaaten

Wegen des *Schutzlandprinzips* (Art. 8 Abs. 1 Rom-II-Verordnung) ist das europäische Urheberrecht nur für in der Union vorgenommene Vervielfäl-

Anwendungsvorrang *W. Kahl* in: C. Calliess/M. Ruffert (Hrsg.), EUV/AEUV-Kommentar, 6. Aufl., München 2022, Art. 4 EUV Rn. 161 f.

60 *Spindler*, Fragen (Fn. 53), 1116.

61 *M. Baumann*, Generative KI und Urheberrecht – Urheber und Anwender im Spannungsfeld, NJW 2023, 3673 Rn. 15.

62 So zum Streaming über einen multimedialen Medienabspieler EuGH GRUR 2017, 610 Rn. 70.

63 Für eine Unvereinbarkeit mit dem Drei-Stufen-Test *Dornis/Stober*, Training (Fn. 18), S. 142 ff.

64 „Members shall confine limitations or exceptions to exclusive rights to certain special cases which do not conflict with a normal exploitation of the work and do not unreasonably prejudice the legitimate interests of the right holder“ (https://www.wto.org/english/docs_e/legal_e/trips_e.htm#art1).

tigungshandlungen anwendbar. Im Bereich des Satellitenrundfunks wird heute mehrheitlich vertreten, dass neben dem Urheberrecht des Sendelandes auch jenes des Empfangslandes anzuwenden ist (*Bogsch-Theorie*). Die Bogsch-Theorie lässt sich auf viele Internetsachverhalte übertragen, etwa wird ein Werk durch Streaming weltweit öffentlich zugänglich gemacht, weshalb sich Verwerter die Weltrechte einräumen lassen müssen, um keine Urheberrechtsverletzung zu riskieren.⁶⁵ Auf das Training generativer KI lässt sich die Bogsch-Theorie indessen nicht transferieren, weil dieses an einem Ort erfolgt. Lediglich die Ausgaben des trainierten Systems können weltweit durch Prompts abgerufen werden, weshalb – wenn der Output urheberrechtlich geschützte Werke oder Werkteile enthält – bezüglich des Rechts der öffentlichen Zugänglichmachung⁶⁶ eine Anwendung der Bogsch-Theorie zu bejahen ist. *Demnach ist für das Training allein das Urheberrecht des Trainingsortes anwendbar.*

Das europäische Urheberrecht ließe sich für jene Systeme, deren Output in der EU abrufbar ist, indirekt über eine *marktortbezogene Produktregulierung* durchsetzen.⁶⁷ Einen dahingehenden Ansatz enthalten Erwägungsgrund 106 sowie Art. 53 Abs. 1 lit. c KI-Verordnung. Letzterer hält die Anbieter generativer KI-Anwendungen an, eine Strategie zur Einhaltung des Urheberrechts der Union und insbesondere zur Ermittlung und Einhaltung des in Art. 4 Abs. 3 DSM-RL, § 44b Abs. 3 UrhG vorgesehenen Nutzungsvorbehalts auf den Weg zu bringen. Allerdings ist das europäische Urheberrecht nicht tangiert und damit eingehalten, wenn das Training in anderen Erdteilen erfolgt. Deshalb ergänzt Erwägungsgrund 106: „Jeder Anbieter, der ein KI-Modell mit allgemeinem Verwendungszweck in der Union in Verkehr bringt, *sollte* diese Pflicht erfüllen, unabhängig davon, in welchem Hoheitsgebiet die urheberrechtlich relevanten Handlungen, die dem Training dieser KI-Modelle mit allgemeinem Verwendungszweck zugrunde liegen, stattfinden. Dies ist erforderlich, um gleiche Wettbewerbsbedingungen für Anbieter von KI-Modellen mit allgemeinem Verwendungszweck sicherzustellen, unter denen kein Anbieter in der Lage sein sollte, durch die Anwendung niedrigerer Urheberrechtsstandards als in der Union einen Wettbewerbsvorteil auf dem Unionsmarkt zu erlangen.“

65 *J. Drexler* in: J. Säcker/R. Rixecker/H. Oetker/B. Limberg (Hrsg.), MüKo, Band 13, 8. Aufl., München 2021, Art. 8 Rom-II-VO Rn. 289, 297 f. mwN.

66 So für den Output von ChatGPT *Heine*, Generative KI (Fn. 14), Rn. 7; *Gema/OpenAI* anhängig beim LG München I (Az.: 42 O 14139/24).

67 Ähnlich *Maamar* Fragen (Fn. 2), 486.

Eine rechtlich verbindliche marktortbezogene Produktregulierung enthält der Erwägungsgrund nicht, da erstens Begründungserwägungen eines Rechtsakts der Union rechtlich nicht verbindlich sind⁶⁸ und zweitens die „Pflicht“ lediglich erfüllt werden „sollte“.

B. Zusammenfassung der Ausgangslage und Einbeziehung der Ratio des Urheberrechts

1. Das Training generativer KI kann mit oder ohne *dauerhaften Datenkorpus* erfolgen.⁶⁹
2. Ein Training mit Webdokumenten ohne lokalen dauerhaften Datenkorpus erfordert *temporäre Vervielfältigungen*, nicht hingegen ein Training mit dauerhaftem Datenkorpus.⁷⁰
3. Möglich ist demnach mit entgeltlich erworbenen Vervielfältigungsstücken einen dauerhaften Datenkorpus zu erstellen und mit diesem das System ohne temporäre oder dauerhafte Vervielfältigungen zu trainieren. In diesem Fall erfolgt das *Training ohne Vervielfältigungshandlungen*.
4. Weil es technisch nicht möglich ist eine Wiederherstellung der Werke auszuschließen und kumulativ das Trainingsmaterial zu löschen, werden juristisch stets *dauerhafte Vervielfältigungen im trainierten System gespeichert* (die juristische Folgerung ist rechtsunsicher).⁷¹
5. Ob für das *Training als solches ein unbenanntes Verwertungsrecht (Innominatfall)* existiert, ist rechtsunsicher. Sollte dies zu bejahen sein, wären die Schrankenbestimmungen der §§ 44b, 44a UrhG entsprechend anzuwenden, sofern sie Vervielfältigungen zu Trainingszwecken erlauben.⁷²
6. Die Schranke des § 44b UrhG ist für das Training generativer KI *nicht anwendbar*, mithin darf ein dauerhafter Datenkorpus nur im Falle vertraglicher Befugnisse gebildet werden (rechtsunsicher).⁷³

68 EuGH BeckRS 2004, 74578.

69 Oben A I 1.

70 Oben A I 1.

71 Oben A I 2.

72 Oben A I 4.

73 Oben A II 1.

7. Die Schranke des § 44a UrhG ist für das Training generativer KI bei rein nationaler Betrachtung *anwendbar*, temporäre Vervielfältigungen sind demnach eigentlich stets gestattet (rechtsunsicher).⁷⁴
8. Allerdings sind die in den beiden vorherigen Nummern genannten Schrankenbestimmungen infolge des *Drei-Stufen-Tests* für das Training generativer KI nicht *anwendbar* (rechtsunsicher).⁷⁵
9. Europäische Urheberrechtsstandards gelten nicht für das *Training in Drittstaaten*.⁷⁶
10. Den unmittelbaren *Bezugspunkt des urheberseitigen Partizipationsinteresses* bilden die mittels generativer KI geschaffenen Konkurrenzprodukte. Die Frage nach einer angemessenen wirtschaftlichen Partizipation kann im Rahmen der *Schranken des Vervielfältigungsrechts* nicht sinnvoll beantwortet werden, weil sie dort systematisch nicht hingehört: Ob temporäre und/oder dauerhafte Vervielfältigungen erfolgen, bestimmt sich allein nach der technischen Ausgestaltung des Trainingsablaufs (Nr. 1-3). Auch der Gesetzgeber geht zutreffend davon aus, dass ein Text und Data Mining auch ohne Vervielfältigungshandlungen möglich ist⁷⁷. Folglich ist die *gegenwärtige Diskussion systematisch-teleologisch unpassend eingekleidet*; sie geht am Zweck des Urheberrechts vorbei.
11. Ferner besteht *kein Grund hinsichtlich der Einräumung eines Nutzungsvorbehaltsrechts zwischen temporären und dauerhaften Vervielfältigungen zu differenzieren*, wie es die §§ 44a, 44b UrhG aktuell vorsehen (wenn man die Schranken für *anwendbar* hält). Dies gilt umso mehr, als die beiden Vervielfältigungsarten technisch und juristisch nicht trennscharf abgrenzbar sind.⁷⁸ Wenn man ein Opting-Out befürwortet, sollte dieses einheitlich gelten.
12. *Die entscheidenden politischen Fragen müssten lauten, unter welchen – nicht an der Trainingsmethode anknüpfenden – Voraussetzungen das Training generativer KI zulässig sein soll (1) und ob für Ausgaben von generativen KI-Systemen eine marktortbezogene Vergütungspflicht eingeführt werden soll (2).*

74 Oben A II 2.

75 Oben A II 3.

76 Oben A III.

77 Erwägungsgrund 9 RL (EU) 2019/790, ABl. L 130/92; BT-Drs. 19/27426, S. 88.

78 Oben A II 2 a bb.

C. Gesetzgeberische Optionen für einen kohärenten Rechtsrahmen

In Betracht kommen eine gesetzliche Trainingsbefugnis ohne Vergütung (I), eine gesetzliche Trainingsbefugnis mit effektivem gesetzlichen Vergütungsanspruch (II) und der Verzicht auf ein gesetzliches Nutzungsrecht (III).

I. Option: Rechtssichere gesetzliche Trainingsbefugnis ohne Vergütung

Eine rechtssichere gesetzliche Trainingsbefugnis setzt erstens eine *rechtssicher anwendbare Schranke für temporäre und dauerhafte Vervielfältigungen* zum Zweck des Text und Data Mining voraus. Zweitens müsste das *Training als solches* rechtssicher gestattet sein. Eine Schranke für Vervielfältigungen würde als Minus auch das Training als solches umfassen (insoweit wäre eine Klarstellung sinnvoll), unabhängig davon, ob für das Training als solches ein unbenanntes Verwertungsrecht existiert.

Wenn bezüglich der Vervielfältigungen ein *Nutzungsvorbehaltsrecht* eingeführt wird, sollte dies auch für das Training als solches, also die bloße „elektronische Lektüre“ vorgesehen werden.⁷⁹ Hierfür müsste der Gesetzgeber ein Verwertungsrecht für das Training generativer KI als solches und eine dazugehörige Schranke einführen.

Die unionsrechtliche Option eines vergütungsfreien „right to mine“ *konfliktiert mit dem Drei-Stufen-Test*⁸⁰, der in völkerrechtlichen Verträgen (Art. 9 Abs. 2 RBÜ, Art. 13 TRIPS) verankert ist.

II. Option: Gesetzliche Trainingsbefugnis mit effektivem gesetzlichen Vergütungsanspruch

Sofern der Unionsgesetzgeber eine umfassende gesetzliche Trainingsbefugnis und als Relikt des Nutzungsrechts einen Vergütungsanspruch einführen möchte, müsste er erstens *für das Training als solches ein neues Verwertungsrecht* verankern, zweitens eine *umfassende (KI-)Schranke für das Vervielfältigungsrecht und das neue Verwertungsrecht* etablieren und drittens den *Vergütungsanspruch* ausgestalten.

79 Oben B Nr. II gilt entsprechend.

80 Die Einschätzung teilt *Schack*, Auslesen (Fn. 26), Rn. 24; skeptisch auch *Baumann*, Generative KI (Fn. 61), Rn. 15.

Letzterer sollte am Output der generativen KI und nicht am Input anknüpfen, weil zum einen nicht das Training, sondern die Ausgaben des Systems als ökonomisch relevante Nutzung der Trainingswerke zu qualifizieren sind, und zum anderen eine *marktortbezogene Vergütungspflicht* europäische und nichteuropäische Softwareunternehmen gleichermaßen betrafte und damit das gegen nahezu jede Regulierung vorgetragene Argument des Standortnachteils entkräftet. Die dem Erwägungsgrund 106 sowie dem Art. 53 Abs. 1 lit. c KI-Verordnung zugrundeliegende Intention, gleiche Wettbewerbsbedingungen herzustellen⁸¹, ist richtig. Eine marktortbezogene Vergütungspflicht wäre das richtige Instrument. Als Vorbild könnte die Geräte- und Speichermedienabgabe der §§ 54, 54b UrhG dienen, die für Hersteller, Importeure und Händler gleichermaßen gilt.⁸² Dadurch wäre die Verletzung der Urheberinteressen auf ein mit dem Drei-Stufen-Test vereinbares Maß reduziert.⁸³

III. Option: Verzicht auf gesetzliches Nutzungsrecht

Als dritte Option kann der Unionsgesetzgeber *für das Training als solches ein neues Verwertungsrecht einführen* und von einer gesetzlichen Schranke für das Training generativer KI und damit einhergehender Vervielfältigungen absehen.

Sodann benötigten Softwareunternehmen für ein Training in der EU ein *rechtsgeschäftlich eingeräumtes Nutzungsrecht*. Weil es aufgrund zu hoher Transaktionskosten nicht praktikabel ist mit jedem Rechtsinhaber individuell zu verhandeln und zu kontrahieren, wäre es sinnvoll, *Verwertungsgesellschaften als Intermediäre* zwischenzuschalten und erweiterte kollektive Lizenzmechanismen einzuführen.⁸⁴

81 Oben A III.

82 Schack, Auslesen (Fn. 26), Rn. 26.

83 Näher zur Idee der Vergütungspflicht für KI-Output *M. Senfleben*, A Tax on Machines for the Purpose of Giving a Bounty to the Dethroned Human Author – Towards an AI Levy for the Substitution of Human Literary and Artistic Works, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4123309 (abgerufen am 31.7.2024).

84 Ausführlich *Pukas* KI-Trainingsdaten (Fn. 14), 615 ff.

D. Politische Stellungnahme

Das Text und Data Mining, also das Training als solches greift nicht in die Verwertungsinteressen der Urheber ein.⁸⁵ Gegenteiliges gilt für die *Konkurrenzprodukte*, die mittels generativer KI angeboten werden.⁸⁶ Durch sie fallen potenzielle Käufer der Werke aus. Das berührt die Urheberinteressen.

Schranken dienen typischerweise *schutzwürdigen Belangen der Allgemeinheit*.⁸⁷ Dazu gehört das Text und Data Mining für Zwecke der wissenschaftlichen Forschung (Art. 3 DSM-RL, § 60d UrhG), weil damit Innovationen gefördert werden. Hingegen dient das Training generativer KI-Anwendungen zuvörderst einzelnen Softwareunternehmen, die ihre Produkte entgeltlich anbieten. Dabei handelt es sich um eine gewerbliche und keine altruistische Tätigkeit.

Die Gesetzgebungsmaterialien zum Urhebergesetz aus dem Jahr 1962 enthalten generelle Überlegungen, die auch für die digitale Welt fruchtbar gemacht werden können: „Eine wichtige Grenze für die Einschränkung des Urheberrechts bildet vor allem der von der Rechtsprechung bereits zur Auslegung des geltenden Rechts entwickelte Gedanke, dass der *Urheber tunlichst an dem wirtschaftlichen Nutzen zu beteiligen ist*, der aus seinem Werk gezogen wird. Aus diesem Grundsatz folgt nicht nur, dass das Urheberrecht keinen Einschränkungen unterliegen darf, die *lediglich dem wirtschaftlichen Interesse einzelner Werknutzer dienen* [spräche für Verzicht auf gesetzliches Nutzungsrecht (Option III)]. Es muss auch vermieden werden, dass eine *an sich im Allgemeininteresse gebotene Einschränkung mittelbar zu einer nicht gerechtfertigten Förderung derartiger wirtschaftlicher Einzelinteressen* führt. In solchen Konfliktlagen erscheint es angebracht, lediglich den Verbotsscharakter der betreffenden urheberrechtlichen Befugnis einzugrenzen, dem Urheber jedoch einen *Anspruch auf angemessene Vergütung* für die Benutzung seines Werkes zu belassen [spräche für gesetzliche Trainingsbefugnis mit Vergütungsanspruch (Option II)].“⁸⁸

85 Zutreffend M. Stieper, Die Schranken des Urheberrechts im Gesetz zur Anpassung des Urheberrechts an die Erfordernisse des digitalen Binnenmarktes, ZUM 2021, 776 (778).

86 So auch Schack, Auslesen (Fn. 26), Rn. 6, 8, 28.

87 T. Dreier in: T. Dreier/G. Schulze (Hrsg.), UrhG-Kommentar, 7. Aufl., München 2022, vor § 44a Rn. 3; S. Lüft in: A. Wandtke/W. Bullinger (Hrsg.), UrhR-Kommentar, 6. Aufl., München 2022, vor § 44a UrhG Rn. 1.

88 BT-Drs. IV/270, S. 63. Die Kursivierungen finden sich nicht im Original.

Betreffend generative KI besteht ein *Angebotsoligopol*, da wenige Anbieter vielen Nachfragern gegenüberstehen. Es lässt sich darüber diskutieren, ob generative KI lediglich dem wirtschaftlichen Interesse der Anbieter oder auch dem Allgemeininteresse dient. Wenn man eine Schranke im Allgemeininteresse für geboten hält, sollte diese nicht mittelbar zu einer nicht gerechtfertigten Förderung wirtschaftlicher Einzelinteressen der Oligopolisten führen. Folglich sollte keine gesetzliche Trainingsbefugnis ohne Vergütung (Option I) bestehen. Sinnvoller sind die Optionen II und III, wobei eine *gesetzliche Trainingsbefugnis mit marktortbezogener Vergütungspflicht (Option II) den Vorteil böte, dass europäischen Softwareunternehmen kein Wettbewerbsnachteil entstünde*.

Abschließend sei angemerkt, dass die „künstliche Grenzintelligenz“⁸⁹ zeitnah erreicht sein könnte, weil das Trainingsmaterial ausgeschöpft ist. *Die Anbieter generativer KI sind auf neue qualitativ hochwertige Trainingswerke angewiesen*. Die Quantität von Beiträgen in den sogenannten sozialen Medien vermag Qualität nicht zu ersetzen. Deshalb gehen den Softwareunternehmen die Trainingsdaten aus. Eine Vergütung setzt einen Anreiz, neue hochwertige Trainingsdaten zu erstellen. Ein Rechtsrahmen, der dies nicht berücksichtigt, ist nicht innovationsfreundlich, weil generative KI ohne klugen menschlichen Input lediglich Unsinn imitiert.

89 J. Jürgens, Alles aufgesaugt, ZEIT Nr. 21/2024, S. 20, <https://www.zeit.de/2024/21/kuenstliche-intelligenz-trainingsdaten-suche-google-meta-openai> (abgerufen am 3.8.2024). Die Anführungszeichen finden sich nicht im Original.