Erster Teil: Abgrenzung des Untersuchungsgegenstands und dessen technische Grundlagen

A. Einleitung

Weitreichende technische Veränderungen prägen das Urheberrecht seit jeher. So legte die Erfindung des mechanischen Buchdrucks im 15. Jahrhundert den methodischen Grundstein für das urheberrechtliche Ausschließlichkeitsrecht.[1] Der Einzug des Magnettonbands in die privaten Haushalte der Konsumenten hatte eine Reformierung des Systems der Privatkopiefreiheit zur Folge.[2] Die Relevanz und Marktmacht von Suchmaschinen und User Generated Content-Plattformen führte nicht zuletzt zur Einführung eines neuen Leistungsschutzrechts und der Begründung eines innovativen Haftungsregimes für sogenannte Diensteanbieter, dem UrhDaG.[3] All diese Entwicklungen machen deutlich, dass eine besondere Wechselbeziehung zwischen Urheberrecht und technischer Innovation besteht. Grund hierfür ist unter anderem, dass ein wesentlicher Teil technischer Innovation die Informationsverarbeitung und -aufnahme durch Menschen beeinflusst. Trägermedium dieser Informationen sind oftmals aber urheberrechtlich geschützte Werke.

Auch die technischen Veränderungen durch die Entwicklung und Anwendung künstlicher neuronaler Netze (KNN) als Form des maschinellen Lernens sind seit der Veröffentlichung des Chatbots *ChatGPT* im November 2022 weitreichend in den gesellschaftlichen Alltag gerückt. Zwischen dem Urheberrecht und der Entwicklung sowie Anwendung von KNN besteht dabei ebenfalls eine besondere Wechselbeziehung. Denn für die Entwicklung der KNN, dem „Training", sind große Datenmengen erforderlich.[4] Bereits aufgrund der niedrigen urheberrechtlichen Schutzvoraussetzungen[5]

1 Dazu m. w. N. *Schack*, Urheber- und Urhebervertragsrecht Rn. 108 ff.
2 BT-Drs. IV/270, S. 71.
3 Durch die Richtlinie 2019/790/EU über das Urheberrecht und die verwandten Schutzrechte im digitalen Binnenmarkt (DSM-RL).
4 Dazu unter I. Teil B. II. 3.
5 Sogar einzelne Sätze oder Satzteile können urheberrechtlichen Schutz genießen, vgl. EuGH, Urt. v. 16.07.2009 - C-5/08, GRUR 2009, 1041 Rn. 47 – Infopaq.

ist davon auszugehen, dass einem wesentlichen Teil der verarbeiteten Trainingsdaten Schutz als Werke im Sinne des § 2 UrhG zukommt. Werke können darüber hinaus bei der Anwendung eines KNN als Input[6] des Systems dienen. Das ist beispielsweise der Fall, wenn mit Sensoren verbundene KNN bei ihrem Einsatz im öffentlichen oder nicht-öffentlichen Raum befindliche Werke oder Werkteile erfassen. Dies ist unter anderem bei mit Kamerasensoren ausgestatteten autonomen Fahrzeugen denkbar.[7] Auch das Web Scraping von Internetquellen kann Inputdaten für die Anwendung von KNN generieren. Es findet unter anderem im Kontext automatisierter Journalismus-Anwendungen statt.[8] Hierbei liegt eine Betroffenheit von Urheberinteressen ebenfalls nahe.

Ein Rückgriff auf Algorithmen, die auf KNN basieren, hat sich im Verlauf der letzten Jahre zum bevorzugten Lösungsweg für die computerbasierte Bewältigung komplexer Anwendungsprobleme entwickelt.[9] Häufig wird dabei allgemein von Künstlicher Intelligenz (KI) gesprochen. Besonders relevant sind vielschichtige, als „tiefe" bezeichnete Ausführungen der KNN.[10] Grund hierfür ist die erhebliche Leitungsfähigkeit der Systeme. Keine andere Technologie konnte bisher ohne wesentlichen menschlichen Einfluss realistisch anmutende Gemälde,[11] längere kohärente Texte[12] so-

6 Zur Funktionsweise genauer unter B. II. 2.

7 Beispielhafte Umgebungsaufnahme eines autonomen Fahrzeugs in *Paaß/Hecker*, Künstliche Intelligenz, S. 318 Abb. 8.28.

8 Siehe dazu unter anderem *Haim/Graefe*, in: Nuernbergk/Neuberger, Journalismus im Internet, S. 139; *Schippan*, ZUM 2024, 670; *Gräfe/Kahl*, MMR 2021, 121.

9 *Jordan/Mitchell*, Science 2015, 255 (255); *Drexl u. a.*, MPI for Innovation and Competition Research Paper Nr. 19-13, S. 3; *WIPO*, WIPO Technology Trends 2019: Artificial Intelligence, S. 31; siehe auch *Döbel u. a.*, Maschinelles Lernen, S. 8, hier wird Machine Learning als „Schlüsseltechnologie der künstlichen Intelligenz" bezeichnet.

10 *Drexl u. a.*, MPI for Innovation and Competition Research Paper Nr. 19-13, S. 3; *Jordan/Mitchell*, Science 2015, 255 (255) („high impact area"); *WIPO*, WIPO Technology Trends 2019: Artificial Intelligence, S. 31 (Deep Learning sei der am schnellsten wachsende Zweig innerhalb des Machine Learning); vgl. auch *Döbel u. a.*, Maschinelles Lernen, S. 11, 16, 24 f. insbesondere auch mit einer Statistik zur stark zunehmenden wissenschaftlichen Forschung in diesem Bereich; *Baum*, in: Leupold/Wiebe/Glossner, MAH IT-Recht, Teil 9 Rn. 30, der den „Siegeszug" des Deep Learning verkündet.

11 Ein durch KI erschaffenes Portrait des fiktiven Edmond de Belamy, siehe *SPIEGEL* v. 26.10.2018. Außerdem der „Next Rembrandt", ein durch KI geschaffenes Gemälde in der malerischen Tradition Rembrandt van Rijns, siehe https://www.nextrembrandt.com.

12 So nun aber durch Textgeneratoren wie GPT-3, siehe *Romero*, Medium/towards data science v. 24.5.2021.

wie Musikstücke[13] erschaffen oder bei einer Bildanalyse menschliche Fehlerquoten unterbieten[14]. Im Gegenteil war die Produktion insbesondere von Kreativinhalten bisher menschlichen Schöpfern vorbehalten. Die Leistungsfähigkeit der KNN spiegelt sich zudem auch wirtschaftlich wider.[15] Es ist also die Basis geschaffen, damit das Urheberrecht auch durch KNN besonders geprägt werden kann. Vor diesem Hintergrund wird im Rahmen der vorliegenden Arbeit untersucht, wohin sich das Urheberrecht im Kontext der Entwicklung und Anwendung von KNN de lege ferenda bewegen wird.

Auch die Anwendung des geltenden Urheberrechts steht vor der Herausforderung, technische Veränderungen im Rahmen der gesetzlich gewährten Spielräume abzubilden. Dabei sollen zuweilen Interessen betroffener Urheber geschützt, an anderer Stelle aber auch Nutzerinteressen Rechnung getragen werden. In den letzten Jahren entlud sich das Spannungsverhältnis zwischen Urheberrecht und technischer Innovation insbesondere beim Umgang mit Verlinkungstechniken im Internet,[16] aber auch im Kontext von Internetrekordern,[17] Sharehosting-Plattformen,[18] dem Schutz von Computerprogrammen[19] oder dem E-Lending[20]. Auch KNN nehmen nun ihren Platz in diesem Spannungsfeld ein. Angesichts dessen gilt es zu untersuchen, wie das Training und die Anwendung von KNN in den geltenden urheberrechtlichen Regelungskomplex eingeordnet werden können.

Besondere Rechtsfragen wirft demgegenüber der in der Anwendungsphase eines KNN generierte Output des Systems auf. Hier kann unter-

13 So beispielsweise die „Jukebox", ein neuronales Netz von Open AI, siehe *Heaven*, MIT Technology Review 2021; außerdem direkt unter https://openai.com/blog/jukebox.

14 Vgl. dazu *He u. a.*, arXiv: 1502.01852 2015, S. 9.

15 Weltweit wird für den Absatz von Deep Learning-Software eine Umsatzsteigerung von 655 Mio. USD (2016) auf 35 Mrd. USD (2025) erwartet *Döbel u. a.*, Maschinelles Lernen, S. 24 mit weiteren Details und Nachweisen. Zur wirtschaftlichen Entwicklung bis 2023 siehe außerdem insgesamt *Maslej u. a.*, The AI Index 2024 Annual Report by Stanford University, S. 214 ff.

16 Beispielsweise siehe EuGH, Urt. v. 08.06.2016 - C-160/15, GRUR 2016, 1152 – GS Media; EuGH, Urt. v. 09.03.2021 - C-392/19, GRUR 2021, 706 – VG Bild-Kunst/SPK.

17 Insbesondere BGH, Urt. v. 22.04.2009 - I ZR 175/07, ZUM 2009, 765 – Online-Videorekorder; BGH, Urt. v. 05.03.2020 - I ZR 32/19, GRUR 2020, 738 – Internet-Radiorecorder.

18 EuGH, Urt. v. 22.06.2021 - C-682/18, C-683/18, GRUR 2021, 1054 – YouTube/Uploaded.

19 EuGH, Urt. v. 02.05.2012 - C-406/10, EuZW 2012, 584 – SAS Institute.

20 C-174/15, Urt. v. 10.11.2016, GRUR 2016, 1266 – VOB/Stichting.

sucht werden, ob und inwieweit dem Output urheberrechtlicher Schutz zukommt.[21] Fraglich ist außerdem, wer für Urheberrechtsverletzungen insbesondere durch den Einsatz generativer KNN, also solcher Systeme, die Kreativinhalte produzieren, verantwortlich ist.[22] Die Beantwortung der sich hieraus ergebenden, von Training und Input in ein KNN nur bedingt abhängigen Fragen würde den für die vorliegende Untersuchung zur Verfügung stehenden Umfang überschreiten. Sie werden deshalb im Rahmen dieser Arbeit nicht näher beleuchtet.

B. Systeme künstlicher Intelligenz als Forschungsgegenstand

Um die urheberrechtlichen Rahmenbedingungen für die Verwendung von Werken als Trainings- und Inputdaten für KNN analysieren und Perspektiven für das Regelungsregime aufzeigen zu können, muss zunächst der Forschungsgegenstand konkretisiert werden. Darüber hinaus bedarf es der Vorstellung einzelner technischer Prozesse und der Einführung einiger Fachbegriffe, deren Kenntnis für die Subsumtion des Forschungsgegenstands unter die urheberrechtlichen Bestimmungen notwendig sind.

I. KI-Begriff der Untersuchung

Wie bereits dargelegt wurde,[23] nehmen KNN innerhalb der KI-Forschung und Anwendung von umgangssprachlich als „KI" bezeichneten Systemen eine besondere Rolle ein. Daher stehen die Forschung und Anwendung von „KI" immer mehr synonym für die Forschung an und Anwendung von Systemen, die auf künstlichen neuronalen Netzen basieren. Die tatsächlichen Entwicklungen sollen auch in der nachfolgenden Untersuchung abgebildet werden. Demzufolge wird auch der Forschungsgegenstand der vorliegenden Arbeit auf Systeme beschränkt, die auf künstlichen neuronalen Netzen basieren und damit als Systeme maschinellen Lernens eingeordnet werden können. Der Beschränkung des Forschungsgegenstands entsprechend wer-

21 Dazu unter vielen hervorzuheben *Maamar*, Computer als Schöpfer.
22 Dazu unter anderem in *Pukas*, ZGE 2024, 38.
23 Unmittelbar oben unter A.

den unter Systemen „künstlicher Intelligenz“ im Rahmen der vorliegenden Untersuchung nur solche Systeme verstanden, die auf KNN aufbauen. Eine terminologische Bezugnahme zum Rechtsbegriff des „KI-Systems“ aus Art. 3 Nr. 1 VO (EU) Nr. 2024/1689 (KI-VO) ist demgegenüber nicht angestrebt. Denn der Anwendungsbereich der KI-VO geht über die Entwicklung und Nutzung von KNN hinaus. Die in der KI-VO normierte Legaldefinition des „KI-Systems“ übersteigt damit den Untersuchungsgegenstand der vorliegenden Arbeit.

II. Technische Grundlagen des maschinellen Lernens auf Basis von KNN

Hinsichtlich der technischen Grundlagen und Funktionalität von KI-Systemen im Sinne der vorliegenden Untersuchung lässt sich zwischen dem Grundprinzip maschinellen Lernens, der Struktur eines künstlichen neuronalen Netzes sowie dem Trainingsprozess, in dem der KI ihre Fähigkeiten zur Bewältigung einer Anwendungsaufgabe vermittelt werden, unterscheiden.

1. Maschinelles Lernen

„Maschinelles Lernen“ ist als Oberbegriff für eine Vielzahl algorithmischer Ansätze zu verstehen. Ziel des Systems ist, eine Aufgabe (Zielfunktion) durch das Sammeln von Erfahrungen während eines repetitiven Lernprozesses (Training) immer besser bewältigen zu können.[24] Bei Systemen, deren Arbeitsweisen konzeptionell auf maschinellen Lernverfahren beruhen, muss zur Bewältigung der Anwendungsaufgabe keine umfangreiche Wissensbasis für eine Erkenntnisableitung „top down“ angelegt werden. Vielmehr arbeiten maschinelle Lernverfahren „bottom up“: Die für die Aufgabe notwendigen Fähigkeiten werden aus zur Verfügung gestellten Beispieldaten durch das System selbst abgeleitet.[25] Auf diese Weise können in einer großen Zahl bestehender Inhalte verborgene Muster und Zusammenhän-

24 *Simon*, in: Michalski/Carbonell/Mitchell, Machine learning: an artificial intelligence approach, S. 25 (28); *Mitchell*, Machine Learning, S. 2.

25 *Morik*, in: Görz, Einführung in die künstliche Intelligenz, S. 243 (254 f.); *Görz/Braun/Schmid*, in: Görz/Schmid/Braun, Handbuch der Künstlichen Intelligenz, S. 1 (19); *Meyer*, ZRP 2018, 233 (235); *Alpaydin*, Machine learning, S. 16 f., 24 f.; *Rosengrün*, Künstliche Intelligenz zur Einführung, S. 27; vgl. außerdem *Sommer*, Haftung für autonome Systeme, S. 38.

ge, also statistische Regeln, extrahiert und zur Lösung eines Anwendungsproblems genutzt werden. Aus der Funktionsweise der algorithmischen Ansätze maschinellen Lernens folgt aber, dass alle auf einem maschinellen Lernverfahren basierenden Systeme auf einen Trainingsprozess angewiesen sind.[26]

2. Struktur und Arbeitsweise eines künstlichen neuronalen Netzes

Künstliche neuronale Netze beruhen, wie auch ihre biologischen Vorbilder, auf „Neuronen", also einer Vielzahl von Knotenpunkten, über die die Informationsverarbeitung abläuft.[27] Künstliche Neuronen nehmen dabei durch mathematische Abläufe ermittelte Zahlenwerte als Impulse (Input) auf und leiten diese an andere Neuronen weiter (Output).[28]

In architektonischer Hinsicht sind die für diese Informationsübertragungen genutzten Netze streng strukturiert.[29] Sie lassen sich in verschiedene Schichten (Layer) aufteilen.[30] Jede Schicht besteht aus einer Vielzahl künstlicher Neuronen und hat eine eigene Aufgabe zu bewältigen. Die Schichten sind dabei miteinander verbunden. Über die Input-Schicht werden die Ausgangsinformationen, beispielsweise Sensordaten oder eine konkrete Arbeitsanweisung, in das KNN eingeleitet.[31] Mithilfe der Output-Schicht gibt das KNN die ermittelten Werte am Ende des Rechenprozesses an den Systemnutzer aus.[32] Die tatsächliche Mustererkennung findet auf Ebene der mittleren, sogenannten verborgenen Schichten statt.

26 Vgl. *Bauckhage u. a.*, in: Görz/Schmid/Braun, Handbuch der Künstlichen Intelligenz, S. 429 (430 ff.); *Drexl u. a.*, MPI for Innovation and Competition Research Paper Nr. 19-13, S. 4 f.

27 *Rey/Wender*, Neuronale Netze, S. 16.

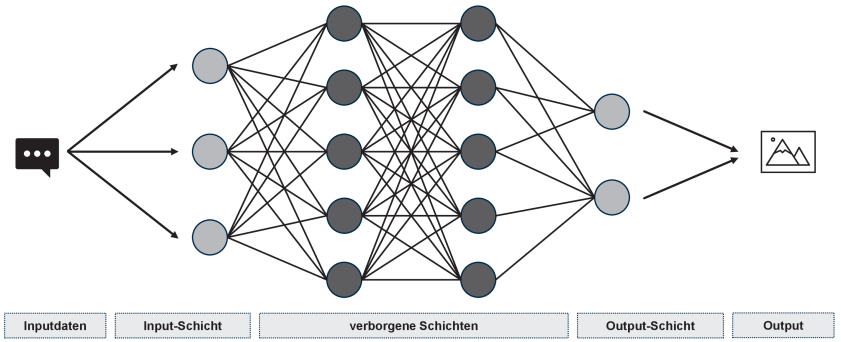
28 *Rey/Wender*, Neuronale Netze, S. 16 f.

29 *Ragni*, in: Görz/Schmid/Braun, Handbuch der Künstlichen Intelligenz, S. 227 (240).

30 Vgl. *Alpaydin*, Machine learning, S. 87; *Ertel*, Grundkurs Künstliche Intelligenz: eine praxisorientierte Einführung, S. 301.

31 *Rey/Wender*, Neuronale Netze, S. 17.

32 *Rey/Wender*, Neuronale Netze, S. 17; vgl. außerdem sehr anschaulich *Rosengrün*, Künstliche Intelligenz zur Einführung, S. 25 f.



Die Verbindung künstlicher Neuronen zueinander wird unter anderem durch eine Vielzahl anpassbarer, numerisch dargestellter Parameter beeinflusst, die die Berechnungsprozesse innerhalb des Algorithmus maßgeblich steuern.[33] Einige von ihnen, die Gewichtungswerte, werden automatisch beim Training des KNN eingestellt.[34] Sie haben einen besonders großen Einfluss auf die statistische Mustererkennung beim Training und den Ablauf der Informationsverarbeitung innerhalb des KNN.[35]

Bevor Inhalte in einem künstlichen neuronalen Netz verarbeitet, also als Input in den Algorithmus eingegeben werden können, müssen sie für die algorithmische Nutzung vorbereitet werden. Dabei werden die Daten im Regelfall insbesondere auf eine vergleichbare Skala transponiert, also normalisiert und anschließend in eine numerische Darstellungsform überführt („Feature Encoding“).[36] Nur diese numerischen Repräsentationen können als Grundlage der algorithmischen Verarbeitung genutzt werden. Der Prozess wird zusammen mit anderen Verarbeitungsschritten, denen im Rahmen dieser Untersuchung keine Bedeutung zukommt, als Datenvorverarbeitung bezeichnet.

33 Mit weiteren Beispielen *Bauckhage u. a.*, in: Görz/Schmid/Braun, Handbuch der Künstlichen Intelligenz, S. 429 (494); *Drexl u. a.*, MPI for Innovation and Competition Research Paper Nr. 19-13, S. 6; *Wahlster/Winterhalter*, Deutsche Normungsroadmap Künstliche Intelligenz des DIN eV. und der DKE, S. 88.

34 *Drexl u. a.*, MPI for Innovation and Competition Research Paper Nr. 19-13, S. 6.

35 Vgl. *Rey/Wender*, Neuronale Netze, S. 17.

36 *Bauckhage u. a.*, in: Görz/Schmid/Braun, Handbuch der Künstlichen Intelligenz, S. 429 (433); *Ertel*, Grundkurs Künstliche Intelligenz: eine praxisorientierte Einführung, S. 301 Abb. 9.26; mit sehr anschaulichen auch grafischen Beispielen *Paaß/Hecker*, Künstliche Intelligenz, S. 51 f., 89.

3. Das Lernen: Training des künstlichen neuronalen Netzes

Wie auch beim menschlichen Lernen[37] ist die Wiederholungstechnik Grundstein des künstlichen Lernens, also des Trainings einer KI. Dem KNN wird dabei durch wiederholte Aktivität die Möglichkeit gegeben, Muster und Zusammenhänge innerhalb der zur Verfügung gestellten Daten zu erkennen. Algorithmisch wird das Training durch eine systematische und repetitive Veränderung der flexiblen Parameter innerhalb des KNN bewerkstelligt.[38] Hierzu gehören insbesondere die Gewichtungswerte.[39] Die Veränderungen werden dabei so lange wiederholt, bis der Output des KI-Systems vordefinierte Qualitätsmerkmale aufweist.[40] Dies spricht indirekt dafür, dass die dem Trainingsmaterial zugrundeliegenden Muster und Zusammenhänge hinreichend genau im KNN repräsentiert worden sind. Zum Training eines KNN können dabei verschiedene Verfahren eingesetzt werden. Welcher Trainingsprozess von KI-Entwicklern gewählt wird, hängt von den verfügbaren Trainingsdaten, dem Anwendungsziel des KI-Systems sowie dem verwendeten Modelltyp ab.[41]

Um den repetitiven Trainingsprozess eines KNN durchzuführen, werden daher spezielle Trainingsdaten in ausreichender Menge benötigt. Als Schätzung dient die Veranschlagung, dass zehnmal so viele Trainingsdatenpaare benötigt werden, wie anpassungsbedürftige Parameter im KNN bestehen.[42] Für fortgeschrittene KI-Systeme, die mitunter Milliarden oder sogar Bil-

37 Vgl. *Bechmann u. a.*, in: Anderhuber/Pera/Streicher, Waldeyer - Anatomie des Menschen, S. 945 (1107); *Heckmann/Dudel*, in: Schmidt/Lang/Heckmann, Physiologie des Menschen, S. 76 (93).

38 *Bauckhage u. a.*, in: Görz/Schmid/Braun, Handbuch der Künstlichen Intelligenz, S. 429 (447); *Drexl u. a.*, MPI for Innovation and Competition Research Paper Nr. 19-13, S. 7.

39 *Mallot*, in: Görz, Einführung in die künstliche Intelligenz, S. 813 (826); *Otero*, GRUR Int. 2021, 1043 (1051); *Rey/Wender*, Neuronale Netze, S. 17.

40 Vgl. *Bauckhage u. a.*, in: Görz/Schmid/Braun, Handbuch der Künstlichen Intelligenz, S. 429 (549); *Rey/Wender*, Neuronale Netze, S. 42.

41 Mit einer tabellarischen Übersicht *Döbel u. a.*, Maschinelles Lernen, S. 10; im Übrigen siehe *Bauckhage u. a.*, in: Görz/Schmid/Braun, Handbuch der Künstlichen Intelligenz, S. 429 (432 ff.); *Drexl u. a.*, MPI for Innovation and Competition Research Paper Nr. 19-13, S. 8; *Rey/Wender*, Neuronale Netze, S. 28 f.

42 Vgl. zu der Faustregel beispielsweise *Alwosheel/Van Cranenburgh/Chorus*, Journal of Choice Modelling 2018, 167 (167), wobei Bestrebungen bestehen, leistungsfähige KI-Systeme auch auf Basis kleinerer Datensätze zu trainieren.

lionen Parameter beinhalten,[43] ist demzufolge eine erhebliche Menge an Trainingsdaten notwendig. Diese für das Training notwendigen Datensätze können auf verschiedenen Arten zusammengestellt werden. Praktisch relevant ist vor allem, die Daten aus frei verfügbaren Internetquellen ohne individuelle Erlaubnis der Betroffenen zusammenzutragen.[44] Dabei werden sie mit Hilfe sogenannter Webcrawler automatisch von Webseiten oder Datenbanken Dritter extrahiert.[45] Es zeigt sich auch, dass die Heterogenität von Webinhalten entscheidend für die Modellqualität von KNN ist.[46] Beispielsweise das Trainingsdatenset *LAION-400M* beinhaltet Bild-Text-Paare, die zwischen 2014 und 2021 aus beliebigen Internetquellen heruntergeladen worden sind. Diese stammen aus dem *Common Crawl*, einem frei zugänglichen, nicht kommerziellen Datenset, welches durch Web Scraping unspezifischer Internetquellen erstellt wurde.[47] Auch der dem Sprachmodell *GPT-3* zugrundeliegende Trainingsdatenkorpus basierte zu großen Teilen auf diesem Datensatz. Darüber hinaus wurden auch andere frei zugängliche Texte wie insbesondere Wikipedia-Beiträge abgespeichert.[48] Nach dem Web Scraping bedarf es auch für das Training auf Basis von Werken noch einer Vorverarbeitung und gegebenenfalls auch Annotation der Inhalte.

43 Das Sprachverarbeitungsmodell *GPT-4* über eine Billiarden Parameter, siehe *Baktash/Dawodi*, arXiv:2305.03195 2023, S. 2. Bereits das Vorgängermodell *GPT-3* bewegt sich mit 175 Milliarden Parametern im Milliardenbereich, siehe *Brown u. a.*, arXiv:2005.14165 2020, S. 5. Ebenfalls das Sprachverarbeitungsmodell *PaLM*, welches Berechnungen auf Basis von 540 Milliarden Parametern vornimmt, siehe *Muller u. a.*, Annual report on European SMEs 2020/2021, S. 23. Selbst ältere Sprachverarbeitungsmodelle, die vor *GPT-3* veröffentlicht wurden, beinhalteten bereits eine Anzahl von Gewichtungswerten im dreistelligen Millionen- bis zweistelligen Milliardenbereich, siehe dazu auch *Brown u. a.*, arXiv:2005.14165 2020, S. 4 m. w. N. Zur Entwicklung der Parameteranzahl insgesamt siehe *Muller u. a.*, Annual report on European SMEs 2020/2021, S. 54.

44 *Longpre u. a.*, NAACL 2024, 3245 (3261).

45 Vgl. *Iglesias u. a.*, Intellectual Property and Artificial Intelligence - A literature review, S. 10; BT-Drs. 19/23700, S. 71; *IBM* hat durch Web-Scraping von Millionen von Fotografien vom Bildhosting-Dienst Flickr große mediale Empörung verursacht, siehe in den NBC News vom 12.3.2019: *Solon*, NBC News v.12.3.2019; auch der bereits erwähnte *Microsoft COCO*-Datensatz ist mit Hilfe von Web-Scraping (fachsprachlich auch als „harvesting" bezeichnet) erstellt worden, siehe *Lin u. a.*, arXiv: 1405.0312 2015, S. 2.

46 *Longpre u. a.*, NAACL 2024, 3245 (3246, 3251).

47 Dazu https://laion.ai/blog/laion-400-open-dataset (zuletzt abgerufen am: 10.1.2024).

48 Zu *GPT-3* siehe *Brown u. a.*, arXiv:2005.14165 2020, S. 8 f. Vgl. außerdem *Hirschberg/Manning*, Science 2015, 261 (264).

C. Forschungsfrage und Ziel der Arbeit

In Bezug auf den Forschungsgegenstand soll untersucht werden, welche Rahmenbedingungen das Urheberrecht für die Verwendung von Werken als a) Trainings- und b) Inputdaten für auf KNN beruhende Systeme maschinellen Lernens, KI-Systeme im Sinne dieser Untersuchung, aufstellt. Dabei ist auch die notwendige Datenvorverarbeitung der als Berechnungsgrundlage der Algorithmen verwendeten Werke zu berücksichtigen. Weil die Generierung, also Speicherung von Werken zur Erstellung von Trainingsdatensätzen aus Internetquellen mittels Webcrawlern praktisch eine erhebliche Bedeutung hat, beschränkt sich die Untersuchung auf diese Art der Datenbeschaffung.

Fraglich ist infolgedessen, welche Rahmenbedingungen das geltende Urheberrecht für folgende Arbeitsschritte zur Verfügung stellt:

1. Speicherung von urheberrechtlich geschützten Trainingsdaten mit Hilfe von Webcrawlern (Web Scraping)
2. Training mit urheberrechtlich geschützten Inhalten als solches
3. Speicherung von urheberrechtlich geschützten Inputdaten für KI, beispielsweise mittels Sensortechnik wie Kameras oder Mikrofonen oder durch manuelle Eingabe der Werke
4. Vorverarbeitung von urheberrechtlich geschützten Inputdaten durch Normalisierung und Feature Encoding

Dabei sind hinsichtlich dieser Fallgruppen auch Perspektiven für die urheberrechtlichen Rahmenbedingungen aufzuzeigen. Untersucht werden muss deswegen auch, ob der bestehende urheberrechtliche Regelungskomplex, der auf die Verwendung von Werken als Trainings- und Inputdaten Anwendung findet, de lege ferenda angepasst werden sollte und wie Anpassungen gegebenenfalls aussehen könnten.

Gegenstand der nachfolgenden Untersuchung ist allerdings nur die Verwendung urheberrechtlich geschützter Werke, also persönlicher geistiger Schöpfungen im Sinne des § 2 Abs. 2 UrhG. Die Erfüllung dieser Kriterien wird für die Zwecke dieser Untersuchung jeweils vorausgesetzt. Der Untersuchung wird außerdem grundsätzlich ein Zweipersonenverhältnis zwischen Urheber und Werknutzer zugrunde gelegt. Das gilt insbesondere für die regulatorische Analyse bestehender Freistellungsinteressen, die im zweiten Teil der vorliegenden Arbeit stattfindet. Nichtsdestotrotz dürften eine Vielzahl der vorgenommenen Abwägungen auch auf das Verhält-

nis zwischen Verwerter, also sekundärem Rechtsinhaber, und Werknutzer übertragen werden können.

D. Gang der Untersuchung

Damit für die Verwendung von Werken als Trainings- und Inputdaten Perspektiven für das Urheberrecht aufgezeigt werden können, müssen zunächst urheberrechtliche Regelungsbedürfnisse für die Verwendung von Werken in den zu untersuchenden Fallgruppen identifiziert werden (zweiter Teil). Hierdurch wird zugleich ein Fahrplan für die Analyse der urheberrechtlichen Lex lata aufgestellt. Denn sie kann an den zuvor ermittelten Untersuchungsergebnissen ausgerichtet werden. Klärungsbedürftig ist dabei stets, ob die geltenden urheberrechtlichen Bestimmungen dem identifizierten Zielzustand der Rechtsordnung gerecht werden. Sind die Regelungsbedürfnisse identifiziert, schließt sich demzufolge die Analyse der Rahmenbedingungen an, die das geltende Urheberrecht zur Lösung der Fallkonstellationen zur Verfügung stellt (dritter Teil).

Die Ergebnisse aus der Analyse des geltenden Urheberrechts werden anschließend mit den identifizierten Regelungsbedürfnissen abgeglichen (dritter Teil). Aus diesem Vergleich ergibt sich, ob das Urheberrecht de lege lata einen interessengerechten Rechtsrahmen für die Verwendung von Werken als Trainings- und Inputdaten zur Verfügung stellt. Anderenfalls offenbaren sich Regelungsdefizite. Soweit sich Anpassungsbedürfnisse offenbart haben, werden im Anschluss Änderungsvorschläge erarbeitet und damit Perspektiven für die urheberrechtlichen Rahmenbedingungen aufgezeigt (vierter Teil). Hierdurch fügen sich die Abschnitte der vorliegenden Untersuchung zu einem kohärenten Gesamtsystem zusammen. Eine Zusammenfassung und Schlussfolgerungen aus der Analyse vervollständigen die Arbeit (fünfter Teil).