



Bild: prastiwi/
123rf.com

Früher sagte man dazu „Plagiat“, heute heißt es „KI-Training“

Neue rechtliche Herausforderungen – der Gesetzgeber ist gefordert

Stephan Holländer

» Entwickler und Produzenten von KI-Technologien sind bei der Beschaffung von Daten und maschinenlesbaren Texten zu Trainingszwecken von KI-Anwendungen wenig zimperlich und beschaffen sich, was im World Wide Web zu finden ist, ohne sich um allfällige bestehende Urheberrechte und Datenschutzbestimmungen zu kümmern. Bibliothekarinnen und Bibliothekare brauchen Kenntnisse, wie sie diese KI-Technologien rechtskonform einsetzen können und ihre Nutzenden über den Gebrauch von KI-basierten Anwendungen beraten können. Die Bibliotheksleitungen sind gefordert, strategische Konzepte zu entwickeln wie diese Technologien gesetzeskonform eingesetzt werden und dass ihre digitalen Bestände, vor allem auch der Open-Access-Bereich, auf den Servern vor dem Abgreifen durch AI-Crawler technisch und rechtlich geschützt werden. Dabei sind rechtlich noch nicht alle Fragen geklärt.

Erste Gerichtsfälle zum KI-Training sind in den USA anhängig

Bekannt ist die anhängige Klage um Urheberrechtsverletzung der New York Times gegen OpenAI wegen der Verwendung von deren Texten zum Training von ChatGPT¹.

Die New York Times hat OpenAI und Microsoft wegen der unerlaubten Verwendung von Times-Artikeln zum Trainieren der großen Sprachmodelle ihrer Generative Pre-Trained Transformer (GPT) verklagt. Der Fall könnte erhebliche Auswirkungen auf das Verhältnis zwischen generativer KI und dem Urheberrecht haben, insbesondere im Hinblick auf die faire Nutzung, und könnte letztlich bestimmen, ob und wie KI-Modelle erstellt werden dürfen.

Der Sachverhalt, der der aktuellen Klage der New York Times zugrunde liegt, dreht sich um die Verwendung

¹ <https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html>

urheberrechtlich geschützter Werke bei der Entwicklung der generativen KI-Tools ChatGPT von OpenAI und Bing Chat (oder „Copilot“) von Microsoft, die beide auf dem GPT-Modell von OpenAI aufbauen. Bei diesen Tools handelt es sich um große Sprachmodelle (Large Language Models, LLMs), die durch „Training“ mit umfangreichen Textkorpora erstellt werden. Die Modelle nehmen Informationen aus diesen Datensätzen auf und „lernen“ die syntaktischen Muster von Wörtern in einem bestimmten Kontext. Bei einer Abfrage kann das LLM dann die wahrscheinlichste Wortkombination vorhersagen und eine Antwort in natürlicher Sprache auf die Eingabeaufforderung der Benutzenden erzeugen. Die neuesten Modelle von GPT werden mit Billionen von Wörtern trainiert – ein Datensatz, der so groß ist, dass er einem Microsoft-Word-Dokument von über 3,7 Milliarden Seiten entspricht.

Die New York Times gelangte bereits einmal vor Gericht, aber nicht als Klägerin, sondern als Beklagte. Vor weniger als drei Jahrzehnten kämpfte der Verlag im Fall New York Times gegen Tasini et al. gegen eine Gruppe seiner eigenen freiberuflichen Autoren.² Dabei verfolgte er eine argumentative Strategie, die der heute verwendeten widerspricht.

Der Hauptvorwurf der New York Times lautet, dass OpenAI durch die nicht lizenzierte und nicht genehmigte Nutzung und Reproduktion von Werken der Times während des Trainings ihrer Modelle gegen Urheberrechte verstoße. Die Urheberrechtsverletzung wird gemäß der Zeitung jedoch in zweierlei Hinsicht noch verschärft. Erstens „merken“ sich LLMs manchmal Teile der in den Trainingsdaten enthaltenen Werke. In diesem Fall können die Modelle gelegentlich nahezu wortgetreue Reproduktionen der Werke erzeugen. Damit zusammenhängend erzeugen LLMs zweitens „synthetische“ Suchergebnisse, die „wenn die LLM“ dazu aufgefordert werden, „wesentlich aussagekräftigere Inhalte aus [einem] Originalartikel wiedergeben können als das, was traditionell bei einer Online-Suche angezeigt würde“. Dies ermögliche den Lesern effektiv, die Paywall der Times zu umgehen.

Diese Probleme stellen nach Ansicht der New York Times eine erhebliche Bedrohung für den Qualitätsjournalismus dar. Wenn die Leser mit Hilfe von GPT-Modellen einfach und kostenlos Zusammenfassungen oder nahezu wortgetreue Reproduktionen von Artikeln der Times erstellen können, würde es nicht mehr notwendig sein, den Zugang über den Verlag der New York Times zu erwerben, was die Möglichkeiten des Verlags, seinen Journalismus weiterhin zu finanzieren, einschränke. Wenn dies geschehe, „würden die Kosten für die Gesellschaft enorm sein“, so die New York Times.

Pikant dabei ist, dass die New York Times im Fall des oben erwähnten Urteils New York Times gegen Tasini et al. aus dem Jahr 1997 eine gegenteilige Rechtsauffassung in ihrer Klageantwort vertreten hat. Damals befand sich die New York Times auf der Beklagtenseite einer Urheberrechtsklage, als eine Gruppe freiberuflicher Autoren den Verlag wegen der Aufnahme ihrer Artikel in neue digitale Archive verklagte. In Zusammenarbeit mit einem Host und Datenbankanbieter, jedoch ohne die Zustimmung der freien Mitarbeiter, lud die New York Times alle in ihren Zeitschriften veröffentlichten Artikel in drei Datenbanken hoch. Es ging dabei um die Frage, ob diese Datenbanken eine „Überarbeitung“ der ursprünglichen Sammelwerke darstellen würden, in denen die Artikel der freien Mitarbeiter zuerst veröffentlicht worden waren. Der Oberste Gerichtshof kam damals zum Schluss, dass die Datenbanken keine „Überarbeitung“ der Sammelwerke darstellen können, da die Artikel in den Datenbanken jeweils einzeln dargestellt werden und damit nicht die rechtlichen Anforderungen von Sammelwerken in ihrer Gesamtheit erfüllen. Daher wurde festgestellt, dass die Times die Urheberrechte der Freiberufler verletzt hatte.

War also mit den bisher im Einsatz befindlichen Suchtechnologien die automatische Erstellung von nahezu wortgetreuen Reproduktionen oder von Rechnern erstellte Zusammenfassungen von Artikeln der New York Times quasi unmöglich, so machte der Einsatz von KI dies nun technisch möglich. Damit waren mit Hilfe der bisherigen Technologie entstandene Artikel und Zusammenfassungen von wenig großer wirtschaftlicher Bedeutung und auch ohne urheberrechtlichen Schutz. Dies änderte sich schlagartig mit dem Aufkommen der KI-Technologie und dem großen Erfolg bei der Einführung von ChatGPT, da solche Reproduktionen und Zusammenfassungen nun in großem Umfang möglich wurden.

Die Generierung von maschinenlesbaren Texten durch Digitalisierung und Zusammenfassungen von bereits existierenden maschinenlesbaren Texten wurde so zu einer gewinnbringenden wichtigen Ressource beim Training und der Anwendung von KI-Technologien. Das haben die Medienkonzerne in den USA zuerst zu spüren bekommen.

Einige wenige Nachrichtenverlage sind dem Beispiel der New York Times gefolgt und haben OpenAI und andere KI-Unternehmen wegen der unerlaubten Nutzung ihrer veröffentlichten Texte verklagt.

Viele andere haben inzwischen jedoch Verträge mit den KI-Unternehmen unterzeichnet, die in der Regel die Verwendung der Verlagstexte als Bezugspunkte

2 <https://www.law.cornell.edu/supct/pdf/00-201P.ZO>

für Nutzeranfragen in Tools wie ChatGPT (mit derzeit versprochener Rückverweisung auf die Webseiten der Verlage) und den KI-Unternehmen die Nutzung der Technologie zur Entwicklung ihrer eigenen Produkte ermöglicht.³

Um etwaigen Initiativen der amerikanischen Bundesstaaten zuvorzukommen, stimmte das US-Repräsentantenhaus am 22. Mai 2025 der Verabschiedung eines 10-jährigen Moratoriums für die Durchsetzung von staatlichen Gesetzen zur Künstlichen Intelligenz (KI) zu. Sobald dieses KI-Moratorium in Kraft tritt, wird es die Möglichkeiten der Bundesstaaten zur Regulierung von KI-Aktivitäten in allen Branchen einschränken. Die Bestimmung ist jedoch nur ein Teil eines umfassenderen Textes, des One Big Beautiful Bill Act (OBBBA), der vom Repräsentantenhaus verabschiedet wurde. Die Bestimmung ist jedoch ein wichtiger Schritt in Richtung des Ziels der Regierung, Hindernisse für KI-Innovationen zu beseitigen und die Führungsrolle der USA im Bereich KI zu stärken.

Das KI-Moratorium in den USA

Die KI-Industrie und die republikanische Partei seien bei diesem Thema unterschiedlicher Meinung, berichtete die Financial Times (FT)⁴. „Es ist eine Machtübernahme durch Tech-Bro-Oligarchen, die versuchen, noch mehr Reichtum und Macht zu konzentrieren“, sagte Max Tegmark, ein MIT-Professor und Präsident des Future of Life Institute, einer gemeinnützigen Organisation, die sich für eine Regulierung der KI einsetzt.

Letzten Monat hat eine Gruppe von 140 Organisationen einen Brief an die Führung des Repräsentantenhauses geschickt und sie gebeten, das 10-jährige Verbot abzulehnen. „Dieses Moratorium würde bedeuten, dass selbst dann, wenn ein Unternehmen absichtlich einen Algorithmus entwickelt, der einen vorhersehbaren Schaden verursacht – unabhängig davon, wie vorsätzlich oder ungeheuerlich das Fehlverhalten ist oder wie verheerend die Folgen sind – das Unternehmen, das diese schlechte Technologie entwickelt, dem Gesetzgeber und der Öffentlichkeit gegenüber nicht rechenschaftspflichtig wäre“, hieß es in dem Brief.

Der FT-Artikel stellte fest, dass das Moratorium die Republikaner gespalten hat, deren Parteivorsitzender den Erlass seines Vorgängers zur KI rückgängig gemacht hat. Die Zeitung zitiert dazu die Abgeordnete Marjorie Taylor Greene, eine Republikanerin aus Georgia, auf X mit dem Post: „Wir haben keine Ahnung, wozu KI in

den nächsten 10 Jahren fähig sein wird, und ihr freien Lauf zu lassen und den Staaten die Hände zu binden, ist potenziell gefährlich. Das muss im Senat gestrichen werden.“

Zu Anfang des Monats Mai 2025 hatte die Trump-Regierung zudem die Direktorin der Kongressbibliothek, Carla Hayden, und die Urheberrechtsbeauftragte Shira Perlmutter, die das Register für Urheberrechte als Teil der Kongressbibliothek leitete, fristlos entlassen. Beide hatten sich für eine Öffnung der Kongressbibliothek, beziehungsweise sehr entschieden für die konsequente Respektierung geltenden Urheberrechts⁵ und des „Fair Use“-Prinzips eingesetzt.

Damit hat die KI-Industrie des Silicon Valley dank Lobbying eines der grundlegenden Prinzipien des amerikanischen Urheberrechts, das Fair-Use-Prinzip, das so im europäischen Recht nicht existiert, außer Kraft gesetzt. „Fair Use“ ist ein rechtlicher Begriff im US-amerikanischen Urheberrecht, der es erlaubt, urheberrechtlich geschütztes Material unter bestimmten Umständen ohne die Zustimmung der Urheber zu verwenden. Das Ziel des Fair Use besteht darin, eine ausgewogene Balance zwischen den Interessen der Urheber und Urheberinnen und den Interessen der Öffentlichkeit zu schaffen. Damit soll ermöglicht werden, dass geschütztes Material in begrenztem Umfang für bestimmte Zwecke genutzt werden kann, ohne dass es zu einer Verletzung des Urheberrechts kommt. Die Einstellung im Silicon Valley war schon immer: „Don’t ask for permission, ask for forgiveness later.“ Es geht weniger um die Vorstellung von der Rechtmäßigkeit des eigenen Tuns als um die Überzeugung, dass im Interesse der Innovation als einer „guten Sache“ eben auch kurzfristige Disruptionen mit den begleitenden Rechtsverstößen möglich sein sollten.

Die Entscheidungen der amerikanischen Regierung und des Kongresses haben Wirkung über die USA hinaus. Bisher gelang es in anderen Ländern nicht, entsprechende Gesetzesänderungen zur Regelung der Verwendung urheberrechtlich geschützter, maschinenlesbarer Texte zu Trainingszwecken bei KI-Anwendungen auf parlamentarischem Weg zu regeln. Dies trifft unter anderem für Großbritannien⁶ und Südkorea⁷ zu.

In einem weiteren Schritt hat die amerikanische Regierung nun beschlossen, mit einem neuen Strategiepplan den Wettlauf um den technologischen Fortschritt mit China zu gewinnen. Ein Punkt des Plans sieht wei-

³ <https://pressgazette.co.uk/platforms/news-publisher-ai-deals-lawsuits-openai-google/>

⁴ Siehe Financial Times vom 18. Juni 2025 <https://www.ft.com/content/52ae52f1-531e-462f-898f-e9f86b3b1869>

⁵ Siehe <https://americanlibrariesmagazine.org/2025/06/18/five-months-into-the-trump-presidency/>

⁶ <https://dig.watch/updates/ai-copyright-clash-stalls-uk-data-bill>

⁷ <https://www.koreatimes.co.kr/lifestyle/people-events/20250613/govt-to-issue-ai-copyright-guidelines-amid-growing-legal-disputes>

tere Deregulierungen im Bereich des Urheberrechts im Umgang mit urheberrechtlich geschützten Daten beim Training von großen KI-Sprachmodellen und weitere Erleichterungen vor⁸. Der amerikanische Präsident steht hinter diesbezüglichen Forderungen der KI-Branche nach einer Deregulierung im Bereich des Urheberrechts. Die diesbezüglichen gerichtlichen Prozesse laufen in den USA noch.

Verstößt das KI-Training mit Texten aus dem Internet gegen das Urheberrecht?

Ob das Training von urheberrechtlich geschützten digitalen Texten aus dem World Wide Web ein rechtlich relevanter Vorgang ist, der gegen bestehende Rechtsvorschriften der Urheberrechtsgesetze verstößt, ist gerichtlich weder im angelsächsischen Recht und auch im kontinentaleuropäischen Recht noch nicht letztinstanzlich entschieden worden. Erste Klagen laufen bereits, so hat in Deutschland die GEMA gegen OpenAI geklagt⁹. Die GEMA wirft OpenAI vor, urheberrechtlich geschützte Songtexte ohne Genehmigung verwendet zu haben. Zwei Nutzungsarten stehen in der Kritik:

Zum einen soll ChatGPT beim Training des KI-Modells auf geschützte Liedtexte zurückgegriffen haben. Das bedeutet, dass im Prozess des „Lernens“ komplette Songtexte kopiert und analysiert wurden, um das Sprachmodell zu verbessern.

Zum anderen kann ChatGPT auf Anfrage hin originalgetreue Songtexte ausgeben, gemäß dem Vorwurf der GEMA. Das heißt, formuliert ein Nutzer einen einfachen Prompt, wie etwa „Gib mir den Liedtext von ‚Ein bißchen Frieden‘“, liefert der Chatbot den vollständigen Text dieses urheberrechtlich geschützten Songs. Lizenzen für diese Vorgänge hat OpenAI bei der GEMA nicht erworben, und Vergütungen an die Urheber wurden nicht gezahlt.

Die Gerichte werden zu beurteilen haben, ob hier das sogenannte Text und Data Mining (TDM) gemäß § 44b UrhG und § 60d UrhG greift. Diese Ausnahmebestimmung erlaubt es beispielsweise, große Mengen an Texten aus dem Internet zu kopieren und auszuwerten, wenn dies zu Analyse Zwecken geschieht und keine Veröffentlichung der vollständigen, geschützten Werke erfolgt. Die diesbezügliche gerichtliche Beurteilung des Falles dazu steht noch aus, denn hinzukommt, dass juristisch noch nicht ab-

schließend geklärt ist, ob KI-Training tatsächlich von der TDM-Ausnahme gedeckt ist. Die Rechtslage ist neu und komplex.

Ist das KI-Training rechtlich keine Kopie sondern eine Transformation?

Hingegen ist zurzeit ein Prozess vor einem Gericht in San Francisco anhängig, der die weitere Entwicklung der KI-Technologie nachhaltig beeinflussen wird. Eine Gruppe von Autorinnen und Autoren hatte gegen Meta geklagt, die ihr generatives KI-Modell Llama mit 70 Millionen Parametern und mittels einer großen Menge unter Verletzung der Urheberrechte ihrer Inhaber heruntergeladener E-Books aus der Schattenbibliothek Library Genesis trainiert hat¹⁰. Die Kläger machen nun geltend, dass ihnen dadurch ein materieller Schaden entstanden sei, den Meta zu entschädigen habe. Meta hingegen macht geltend, dass beim Herunterladen und Training von Llama keine Kopien im juristischen Sinne eines endgültigen Downloads gemacht wurden, sondern dass nur eine vorübergehende Umwandlung der Bücher zu Trainingszwecken unternommen wurde, was über das Fair-Use-Prinzip im amerikanischen Urheberrecht zugelassen sei, denn die Urheberrechtsinhaber hätten keinen Anspruch auf „Schutz vor Wettbewerb auf dem Markt der Ideen“. Der Richter entgegnete dem Anwalt von Meta: „Aber wenn ich Dinge vom Markt der Ideen stehle, um meine eigenen Ideen zu entwickeln, ist das doch eine Urheberrechtsverletzung, oder?“ An den Anwalt der Kläger gewandt, argumentierte der Richter weiter: „Ich denke, dass die faire Nutzung wegfällt, es sei denn, ein Kläger kann zeigen, dass der Markt für sein eigentliches urheberrechtlich geschütztes Werk dramatisch beeinträchtigt wird.“¹¹. Dies sind für beide Seiten knifflige Fragen, die für eine endgültige Beurteilung der Rechtssache an einem weiteren Termin zu belegen sind, wie der Zwischenentscheid festhält¹². Der Zwischenentscheid lässt anklingen, dass es sich um einen sehr ungewöhnlichen Fall handelt, da das Kopieren, obwohl es laut Meta zu einem hochgradig transformativen Zweck geschieht, mit hoher Wahrscheinlichkeit zu einer Überflutung der Märkte mit ähnlichen KI-generierten Werken führt, was den Markt urheberrechtlich geschützter Werke beeinträchtigen wird. KI-Unternehmen argumentieren, dass ihre Systeme urheberrechtlich geschütztes Material auf faire Weise nutzen, um zu lernen, ähnliche

8 <https://www.zeit.de/digital/2025-07/donald-trump-ki-regulierung-usa>

9 <https://www.heise.de/news/GEMA-verklagt-OpenAI-Auf-Klau-gebaute-Songtexte-10032255.html>

10 <https://www.reuters.com/technology/artificial-intelligence/meta-knew-it-used-pirated-books-train-ai-authors-say-2025-01-09/> und <https://www.reuters.com/legal/litigation/meta-says-copying-books-was-fair-use-authors-ai-lawsuit-2025-03-25/>

11 <https://www.reuters.com/legal/litigation/judge-meta-case-weighs-key-question-ai-copyright-lawsuits-2025-05-01/>

12 <https://cdn.arstechnica.net/wp-content/uploads/2025/03/Kadrey-v-Meta-Motion-for-Summary-Judgment-3-10-25.pdf>

Inhalte zu erstellen. Sie behaupten zudem, dass es die wachsende KI-Industrie behindern könnte, wenn sie gezwungen wären, Urheberrechtshabern Gebühren zu bezahlen.

Ein US-Bundesgericht in San Francisco hat nun in einem richtungsweisenden Zwischenurteil entschieden, dass die Nutzung urheberrechtlich geschützter Bücher zum Training eines KI-Modells unter bestimmten Bedingungen rechtmäßig ist. Zugleich stellte das Gericht aber auch fest, dass die Speicherung und Nutzung illegal beschaffter Bücher eine Urheberrechtsverletzung darstellen. Das Technologieunternehmen Anthropic, unterstützt von Amazon und Alphabet, stand wegen der Nutzung von Büchern zur Schulung seines Sprachmodells Claude vor Gericht. Die Klage wurde von mehreren Autoren eingereicht, darunter Andrea Bartz, Charles Graeber und Kirk Wallace Johnson. Sie warfen Anthropic vor, ihre Werke ohne Genehmigung und ohne Entlohnung verwendet zu haben. Das Gericht in San Francisco hielt drei Punkte in seinem Urteil fest¹³:

1. Das Gericht hat entschieden, dass das Training von KI-Modellen mit urheberrechtlich geschützten Büchern als Fair-Use gelten kann, sofern die Werke legal beschafft wurden.
2. Für Bücher aus Quellen wie Books3, LibGen und PiLiMi gab das Gericht keinen Fair-Use-Schutz: Die dauerhafte Speicherung und Nutzung raubkopierter Werke auch für Trainingszwecke stellt laut Urteil einen klaren Urheberrechtsverstoß dar.
3. Das Verfahren gegen Anthropic wird fortgesetzt, da das Gericht über die Haftung für die Nutzung der Raubkopien und mögliche Schadensersatzforderungen wegen willentlicher Verletzung noch nicht entschieden hat.

Das Urteil stellt einen wichtigen Präzedenzfall für die Bewertung von KI-Trainingspraktiken im Lichte des Urheberrechts dar. Während die Nutzung urheberrechtlich geschützter Inhalte unter bestimmten Bedingungen als legal gilt, mahnt das Gericht zugleich deutlich zur Sorgfalt bei der Beschaffung der Daten. Der Ausgang des Prozesses im Dezember könnte wegweisend für die gesamte Branche sein.

Dieses Teilurteil und die noch zu klärenden Rechtsfragen müssten Bibliotheken und Verlage in Europa eigentlich wachrütteln. Es stellt sich einerseits die Frage, ob Biblio-

theken dies als Text- und Data Mining (TDM) gelten lassen müssen und ob andererseits bei gleichen Vorgängen dies doch eine Urheberrechtsverletzung ist. Ein Rechtsprofessor und ein IT-Professor zeigten in ihrer interdisziplinären Untersuchung auf, dass sich generatives KI-Training fundamental vom Text- und Data Mining unterscheidet¹⁴. Der Europaabgeordnete Axel Voss hat als Berichterstatter des AIDA-Sonderausschusses des Europaparlaments einen Bericht zu den wirtschaftlichen Auswirkungen zu dieser Frage vorgelegt¹⁵. Sein Fazit: der Gesetzgeber muss handeln.

Das Verfassen von Büchern dank KI-Technologie

Die KI-Technologie dient auch dazu Bücher zu generieren, die nicht von Menschen geschrieben oder nur teilweise von Menschen verfasst wurden.

So hat die KI-Technologie dazu geführt, dass Plattformen wie Amazon, aber auch Öffentliche Bibliotheken, mit KI-generierten Reiseführern und Kinderbüchern, die Fehlinformationen aufweisen, konfrontiert¹⁶ sind. Dies kann einerseits als Chance für Öffentliche Bibliotheken gesehen werden, da diese Bibliotheken mit einem gut kuratierten Medienbestand aus verlässlicher Herkunft ein Alleinstellungsmerkmal aufweisen werden, andererseits versuchen Amateure dank Chatbots und digitalem Eigenverlag, mit KI-generierten Romanen oder Ratgebern schnell Geld zu verdienen¹⁷. Einige dieser KI-Bücher imitieren sogar den Stil etablierter Autoren und erzielen respektable Verkaufserfolge. Amazon führte nach monatelangen Diskussionen mit dem amerikanischen Schriftstellerverband „Authors Guild“, eine neue Regel ein, die Kindle Direct Publisher'n vorschreibt, den Einsatz von KI in ihren Werken offenzulegen¹⁸. Die Authors Guild unterscheidet zwischen KI-generierten und KI-unterstützten Werken. Generell kann KI aber auch ein nützliches Werkzeug für Autoren sein. Aber KI-generierte Texte können auch gefährlich sein, wie beispielsweise ungeprüfte KI-generierte Pilzratgeber zeigen, die fehlerhafte und damit gefährliche Tipps enthalten. Davor sind auch die wissenschaftlichen Bibliotheken nicht gefeit, wie eine wissenschaftliche Buchveröffentlichung dieses Jahr gezeigt hat. Ende März brachte der Verlag „Springer Nature“ ein Buch mit dem Titel „Advanced Nanovaccines for Cancer Immunotherapy“ he-

13 <https://storage.courtlistener.com/recap/gov.uscourts.cand.434709/gov.uscourts.cand.434709.231.0.pdf>

14 Tim W. Dornis und Sebastian Stober. Urheberrecht und Training generativer KI-Modelle. Technologische und juristische Grundlagen. Nomos Verlag, 2024 ...

15 <https://www.axel-voss-europa.de/kuenstliche-intelligenz/>

16 NDR, Der Nachmittag Sendung <https://www.ndr.de/kultur/buch/KI-Kinderbuecher-erobern-Amazon-Fast-Food-Literatur,kinderbuecher320.html>

17 <https://www.straitstimes.com/world/chatgpt-launches-boom-in-ai-written-e-books-on-amazon>

18 <https://authorsguild.org/news/ai-driving-new-surge-of-sham-books-on-amazon/>

raus¹⁹. Autor ist ein gewisser Nanasheb Thorat, Associate Professor für Medizinische Physik an der irischen University of Limerick. Das Buch sollte einen detaillierten Einblick in die Anwendung von Impfstoffen bei der Behandlung von Krebs, die durch Fortschritte in der Nanotechnologie verbessert wird, bieten. Auf Seite 25 stand im Kapitel „1.6.3. Why Cancer Vaccines Are Better than Chemotherapy“ der Satz: „It is important that as AI modell, I can provied a general perspective, but you should consult with medical professionals for personalized advice“. Dieser Standardsatz stammt aus dem Dialog mit einem KI-Chatbot, den der Autor vergessen hat zu löschen. Ist das ein Einzelfall oder die bekannte Spitze des Eisbergs? Dies scheint öfter vorzukommen, und auch renommierte Verlage und Zeitschriften trotz des Einsatzes von Peer Review und Lektorat zu betreffen²⁰. Springer Nature hat das Buch in der Zwischenzeit zurückgezogen.

Bibliotheken als Trainingcamps für das KI-Training?

Aus Befürchtung, dass nicht genügend Daten für das Training von KI-Anwendungen zur Verfügung stehen könnten, haben sich die großen Techkonzerne auch an Bibliotheken großer Institutionen in den USA gewandt. So hat die Harvard Universität bereits im letzten Dezember die „Institutional Data Initiative“ ins Leben gerufen. Beginnend mit der Bibliothek der Rechtsfakultät wurde die ganze Fallrechtssammlung für das Training zugänglich gemacht. Die in Harvard ansässige Institutional Data Initiative arbeitet mit Bibliotheken und Museen auf der ganzen Welt daran, ihre historischen Sammlungen so für die KI fit zu machen, dass sie auch den Gemeinschaften zugutekommen, denen sie dienen.

„Wir versuchen, einen Teil des Potentials aus dem aktuellen KI-Moment zurück in diese Institutionen zu verlagern.“, sagt Aristana Scourtas, die die Forschung im Library Innovation Lab der Harvard Law School leitet. „Bibliothekare waren schon immer die Verwalter von Daten und Informationen.“²¹ Die neue KI-Trainingskollektion von Harvard umfasst schätzungsweise 242 Milliarden Token – eine Menge, die für den Menschen kaum vorstellbar ist, aber dennoch nur ein Tropfen auf den heißen Stein ist, mit dem die fortschrittlichsten KI-Systeme gefüttert werden. Die Facebook-Muttergesellschaft Meta hat beispielsweise erklärt, dass die neueste Version ihres großen

KI-Sprachmodells mit mehr als 30 Billionen Token aus Texten, Bildern und Videos trainiert wurde.

Dieser Initiative schließen sich nicht alle Bibliotheken an. So hat OpenAI, das auch gegen eine Reihe von Urheberrechtsklagen kämpft, dieses Jahr 50 Millionen Dollar an eine Gruppe von Forschungseinrichtungen gespendet, darunter die 400 Jahre alte Bodleian Library der Universität Oxford, die seltene Texte digitalisiert und KI einsetzt, um sie zu transkribieren.

Die Kennzeichnungspflicht für die mit KI-Technologie entstandenen Texte und Publikationen sollte in die Gesetzgebung aufgenommen werden und damit einen Beitrag zur Transparenz vor dem Gebrauch durch die Nutzerinnen und Nutzer von Bibliotheken geleistet werden. Die Bibliotheken stehen aber auch in rechtlicher Hinsicht vor Herausforderungen und ungeklärten Rechtsfragen. Die stürmische Entwicklung der generativen KI stellt die Bibliotheken vor einige offene Fragen. Im Urheberrecht, dem Datenschutz oder dem Haftungsrecht müssen noch Antworten gefunden werden, um einen rechtssicheren und gegenüber Nutzerinnen und Nutzern der Bibliothek verantwortungsvollen Umgang mit der neuen Technologie gewährleisten zu können.

Während das Urheberrecht die Antworten der KI in der Regel nicht schützt, da ihr Urheber kein Mensch ist, sind beim Datenschutz mit der Datenschutzgrundverordnung (DSGVO) und bei der Haftung für die KI-Produkte rechtlich hohe Anforderungen zu beachten.

In der EU-Richtlinie zum Urheberrecht und in den entsprechenden nationalen Gesetzen finden sich keine spezifischen Regeln, die die durch KI-Technologie entstandenen Werke regelt. Gesichert ist nur, dass die ausschließlich mit KI entstandenen Publikationen keinen Urheberrechtsschutz genießen. So können KI-generierte Texte und Bilder gemeinfrei genutzt werden, ohne dass Urheberrechte übertragen und entschädigt werden müssten. Und damit dürfen die so entstanden Werke auch ohne Einschränkung kopiert und verwendet werden. Handelt es sich um eine hybride Publikation, die von einem Autor geschriebene Abschnitte enthält, die durch weitere mit KI-Werkzeugen erzeugte Abschnitte ergänzt wurden, so sind nur die von einem Menschen als Urheber geschaffenen Teile urheberrechtlich geschützt. Bei Texten mit „gemischten“ Bestandteilen wird die Einschätzung schwierig, welche Teile urheberrechtlich geschützt sind und welche nicht. Unproblematisch ist der Einsatz einer KI-Software zur Korrektur von Rechtschreib- oder Grammatikfehlern,

19 <https://www.faz.net/aktuell/wissen/forschung-politik/fake-fachbuch-von-springer-nature-wie-der-einsatz-von-ki-die-wissenschaft-bedroht-110391850.html>

20 Abbas, Muhammad, Uses and Misuses of ChatGPT by Academic Community: An Overview and Guidelines (March 28, 2023). Available at SSRN: <https://ssrn.com/abstract=4402510> or <http://dx.doi.org/10.2139/ssrn.4402510>

21 Libraries open their stacks up as training data for artificial intelligence platforms | AP News

sofern der Text durch einen Menschen verfasst wurde. Hier bleibt der urheberrechtliche Schutz bestehen. Es gibt aber auch Grauzonen, bei denen sich Juristen nicht einig sind, ob das Vorgehen rechtlich zulässig ist oder nicht. Das Durchforsten und Lesen von Daten und Texten mit KI-Technologien, die frei verfügbar im Netz vorhanden sind, sind nach der Auffassung einiger Juristen durch die urheberrechtlichen Schrankenregelungen zum Text und Data Mining (§§ 60d und 44b UrhG) zulässig, sofern eine Text- und Data-Mining-Erlaubnis (TDM-Erlaubnis) vorliegt²². Dagegen einzuwenden ist, dass der EU-Gesetzgeber die Anwendbarkeit der TDM-Schranken als Grundlage für das Training von großen Sprachmodellen in der Urheberrechtsrichtlinie (DSM-RL) nicht vorhergesehen hat – die DSM-Richtlinie wurde 2016 vorgeschlagen und 2019 verabschiedet, als die Frage von LLM-Training noch wenig öffentliche Aufmerksamkeit erregte. In jedem Fall wurde die Definition von TDM in der Richtlinie absichtlich weit gefasst, um so zukunftsfähig zu sein.

Bei einer Open-Access-Publikation ist die Vergabe einer Creative-Commons-Lizenz Standard und die Nachnutzung der Inhalte durch die entsprechende Lizenz geregelt. Die Regelungen in der Datenschutzgrundverordnung sind wesentlich strenger und greifen bereits bei der Verwendung persönlicher Daten wie E-Mail-Adressen oder IP-Adressen. Auch bei der Formulierung von Prompts sind die Regelungen der DSGVO anwendbar, wenn persönliche Angaben von Personen benutzt werden. Dies gilt auch beim Hochladen von Bildern, auf denen Personen abgebildet sind. Die Einschätzung bei der Formulierung von Prompts ist oft heikel, der Hamburger Beauftragte für Datenschutz und Informationsfreiheit hat eine gute Checkliste für auf großen Sprachmodellen basierende Chatbots erstellt²³. Als Faustregel sollte gelten, dass man bei der Formulierung von Prompts auf personenbezogene Angaben am besten verzichtet.

Wie bereits oben ausgeführt, neigen Chatbots mit KI-Technologie zum Halluzinieren. Rechtlich gesehen stellt sich die Frage, wer dafür haftbar gemacht werden kann. Wurde beispielsweise ein KI-Chatbot mit Fehlinformationen trainiert, so stellt sich die Frage, ob der Produzent der Anwendung oder die Institution, die den Chatbot verwendet, für falsche Angaben haftet, die der Chatbot ausgibt. In Kanada wurde ein diesbezüglicher Fall entschieden²⁴. Ein Passagier hatte gegen Air Canada geklagt, da ihr Chatbot dem Passagier beim Buchen eines Flugtickets auch nach dem Flug einen

nachträglichen Rabatt auf den Preis eines Vollpreistickets versprach, was offensichtlich eine Fehlinformation war, da der Rabatt nur vor dem Flug beim Buchen erhältlich ist. Die Fluggesellschaft erklärte, der Chatbot sei eine „separate juristische Person, die für ihre eigenen Handlungen verantwortlich ist“. Air Canada argumentierte, dass der Passagier auf den vom Chatbot bereitgestellten Link hätte klicken sollen, worauf er die richtige Richtlinie gesehen hätte.

Das British Columbia Civil Resolution Tribunal wies dieses Argument zurück und entschied, dass Air Canada dem klagenden Passagier Schadenersatz und Gerichtsgebühren zahlen müsse. „Air Canada sollte sich darüber im Klaren sein, dass sie für alle Informationen auf ihrer Website verantwortlich ist.“, heißt es in der schriftlichen Urteilsbegründung. „Es macht keinen Unterschied, ob die Informationen von einer statischen Webseite oder einem Chatbot stammen.“ Mit Verweis auf dieses Urteil wurden weitere ähnliche Fälle in Hongkong und in den USA im gleichen Sinne entschieden.

Die KI-Verordnung der EU und die Regulierung der Schweiz

Die Europäische Union hat mit der KI-Verordnung (AI-Act) einen ersten Schritt zur Regulierung von KI-Systemen gemacht. Mit der neuen EU-Verordnung über Künstliche Intelligenz (KI-VO), die am 1. August 2024 in Kraft getreten ist, sind Anbieter und Betreiber von KI-Systemen zu neuen Transparenzmaßnahmen verpflichtet. Insbesondere der Artikel 50 der KI-VO regelt, dass KI-generierte Inhalte klar als solche gekennzeichnet werden müssen. Der Artikel der Verordnung schreibt vor:

- dass KI-Interaktionen mit Menschen erkennbar sein müssen.
- KI-generierte Texte, Bilder, Videos und Audios als solche markiert werden müssen.
- Deepfakes ausdrücklich als manipulierte Inhalte gekennzeichnet werden müssen.

Hier soll aber noch kurz auf die rechtliche Regelung der Schweiz eingegangen werden. Da die Schweiz Nichtmitglied der EU ist, kann sie ihre Gesetzgebung unabhängig von der EU-Gesetzgebung bestimmen. Der Bundesrat, die Schweizer Regierung, hat im Gegensatz zur EU beschlossen, dem Parlament kein spezifisches KI-Gesetz zu unterbreiten, sondern, wo nötig gesetzliche Regelungen in bereits bestehenden Gesetzen anzupassen oder neu einzuführen, das mehrheitlich auch von Vertretern der Rechtswissenschaft gestützt wird²⁵. Das Urheberrechts-

22 So das Landgericht Hamburg (Urt. v. 27.09.2024, Az. 310 O 227/23) und Rack, F. (2024). Rechtsfragen zur generativen KI. *ABI Technik*, 44(1), Link: <https://www.degruyterbrill.com/document/doi/10.1515/abitech-2024-0005/html>

23 Siehe <https://datenschutz-hamburg.de/news/checkliste-zum-einsatz-llm-basierter-chatbots>

24 <https://www.canlii.org/en/bc/bccrt/doc/2024/2024bccrt149/2024bccrt149.html>

25 So etwa Florent Thouvenin, Stephanie Volz, *Ein Rechtsrahmen für den Einsatz von Künstlicher Intelligenz (KI) in der Schweiz*, in: Zeitschrift des

gesetz der Schweiz hat einige Ähnlichkeiten mit der EU-Richtlinie zum Urheberrecht, doch hat der Gesetzgeber die TDM-Regelung und insbesondere das Opt-out-Recht nicht in Schweizer Recht übernommen. Erst kürzlich hat die Regierung dem Parlament für eine Novellierung des Urheberrechtsgesetzes ein Leistungsschutzrecht für Inhalte der Schweizer Medienkonzerne und Verlage vorgeschlagen. Damit sollen die Techkonzerne eine Abgabe zahlen, wenn sie Ausschnitte aus Verlagseinhalten für Nachrichten auf ihren Plattformen verwenden wollen, ganz nach dem Prinzip „Fair Use ist Fairpay“.

Das Urheberrechtsgesetz ist technologieneutral, so dass die Weitergabe von digitalen Privatkopien gestattet ist, da auch die Weitergabe analoger Privatkopien zulässig ist, wie das Schweizer Bundesgericht in Lausanne in einem Urteil festgehalten hat²⁶. Die Regierung hofft so, die Schweiz als Standort für das Training von KI-Modellen attraktiv zu machen. Daran ändern auch die Vorgaben des EU AI Act für Allzweck-KI-Modelle nichts.

Der EU-Gesetzgeber beabsichtigte eine Regelung zu erlassen, die Nicht-EU-Anbieter von solchen Modellen zwingen sollte, sich beim Trainieren ihrer Modelle namentlich an die TDM-Regelung der EU zu halten. Beim Verfassen der einschlägigen Bestimmung (Art. 53 Abs. 1 Bst. c KI-Verordnung) wurde jedoch übersehen, dass das EU-Urheberrecht bei einem Training in der Schweiz (oder in den USA) nicht zur Anwendung kommen kann. Also kann die Nicht-Beachtung der TDM-Regelung diese EU-Regelung auch gar nicht verletzen, da vor Ort die nationalen gesetzlichen Regelungen des jeweiligen Landes gelten.

Die Datenschutzgesetzgebung der Schweiz ist einerseits durch ein entsprechendes Bundesgesetz geregelt, andererseits haben die 26 Kantone eigene Datenschutzgesetze erlassen, die auch die öffentlichen Bibliotheken betreffen, sofern sie von der öffentlichen Hand finanziert werden. Bei der jüngsten Novellierung hat der Kanton Zürich in sein Datenschutzgesetz auch Bestimmungen zur KI eingefügt²⁷. Dieses Beispiel dürfte auch in anderen Kantonen Schule machen. Mit einer ausdrücklichen Bestimmung in der Benutzungsordnung der Bibliothek oder in den AGB eines Unternehmens kann darauf hingewiesen werden, dass einige Leistungen durch einen KI-Chatbot erbracht werden und dass dessen Ergebnisse nicht durch menschliche Mitarbeiter kontrolliert werden. So kann die Haftung etwa zugunsten einer Bibliothek eingeschränkt werden.

Wie die bisherigen Ausführungen gezeigt haben, stehen die Bibliotheken in ihrer Gesamtheit beim Einsatz der KI-Technologien nicht an der ersten Vorderfront. Das hat verschiedene Gründe. Einerseits sind viele Bibliotheken von ihren Trägern dazu angehalten, Einsparungen vorzunehmen, und andererseits ist das notwendige Know-how unter Bibliothekarinnen und Bibliothekaren noch nicht in genügendem Maße vorhanden. Auch scheinen auf strategischer Ebene in vielen Bibliotheken abgesehen vom schlichten Wunsch, diese Technologien in die Bibliotheksarbeit zu integrieren, noch keine detaillierten Planungen vorzuliegen. Auch rechtlich gesehen sind noch viele Fragen zum rechtssicheren Einsatz der Technologien offen. Der Gesetzgeber ist gefordert, Lücken im Urheber- und Datenschutzrecht zu schließen. Die KI-Verordnung erfordert angesichts der stürmischen Weiterentwicklung der KI-Technologien wohl bald eine Novellierung. Bedenklich stimmt, dass immer mehr Menschen diesen KI-Technologien vertrauen, die in den Händen von wenigen Techkonzernen liegen, die sich gleichzeitig aber auch einen unbeschränkten Zugang zu Mails, Dateien, Bildern und SMS ihrer Nutzerinnen und Nutzer beim Gebrauch der KI-Funktionen von Facebook und Instagram verschaffen wollen, wie das Meta im Moment vorexerziert²⁸. Gegen diese Monopolisierung gibt es ein wirksames Mittel: die Offenlegung der Source Codes und die Entwicklung von Open-Source-Anwendungen. Ein Vorkommnis aus jüngster Vergangenheit beim Ankläger am Internationalen Strafgerichtshof in Den Haag²⁹ hat gezeigt, dass die Abhängigkeit von der Software eines einzelnen Softwareunternehmens riskant sein kann, dem Ankläger wurde auf Anordnung aus der Politik der Zugang zu seinem Mailsystem eines bestimmten Herstellers gesperrt. Bibliotheken sollten sich in ihrer Planung bewusst sein, welche Arbeitsprozesse man an die KI abgibt oder ob man sie „nur“ in einer Assistentenfunktion nutzen will und in wie große Abhängigkeit von Softwareproduzenten man sich begeben will. Es macht einen großen Unterschied, ob man das Verfassen eines Fachbuches an eine KI abgibt oder ob man sich nur Rechtschreibfehler korrigieren lässt oder gar einen Formulierungsvorschlag durch eine KI-Software geben lässt. ■



Stephan Holländer

Basel

stephan@stephan-hollaender.ch

Bernischen Juristenvereins, No. 160, 2024

26 Bundesgerichtsurteil BGE 140 III 616 Entscheid in Sachen Bibliothek ETH Zürich gegen Wissenschaftsverlage Elsevier, Thieme und Springer, Link: http://relevancy.bger.ch/php/clir/http/index.php?highlight_doid=atf%3A%2F%2F140-III-616%3Ade&lang=de&type=show_document

27 https://www.zh.ch/de/politik-staat/gesetze-beschluesse/gesetzessammlung/zhlex-os/erlass-170_4-62-121.html

28 <https://www.tagesschau.de/wirtschaft/verbraucher/meta-widerspruch-datennutzung-100.html>

29 <https://www.wiwo.de/technologie/digitale-welt/sanktionen-gegen-internationalen-gerichtshof-microsoft-steckt-in-der-trump-falle/100129647.html>