

(Nur) Zur Zulässigkeit von Vervielfältigungen für die Erstellung und Überprüfung eines Datensatzes

Besprechung zu LG Hamburg, Urteil vom 27.09.2024 - 310 O 227/23 (ZUM 2025, 64)

[Erstveröffentlicht in: ZUM 2025, 71]

I. Sachverhalt

Die Entscheidung des Landgerichts Hamburg in der Sache Kneschke ./ LAION e.V. hat eine ungeheure Resonanz in Presse und Wissenschaft erfahren. Aber handelt es sich, wie teilweise zu lesen war, tatsächlich um eine Grundsatzentscheidung zum Training generativer KI mit urheberrechtlich geschütztem Material? Ein näherer Blick auf den Verfahrensgegenstand belegt das nicht. Denn die streitgegenständliche Vervielfältigung durch den beklagten Verein LAION erfolgte nicht zum Training eines KI-Modells – sie diente allein dem Zweck, zu überprüfen, ob Inhalt und Beschreibung von Bildern übereinstimmten, deren URL und Beschreibung LAION aus einem Datensatz der Organisation Common Crawl entnommen hatte. In diesem Datensatz befand sich auch die URL eines Bildes des Klägers, das er zuvor über die Webseite bigstockphoto.com zugänglich gemacht hatte. Nachdem LAION nicht hinreichend übereinstimmende Bild-Text-Paare durch eine Software erkannt und herausgefiltert hatte, entstand daraus schließlich der Datensatz LAION-5B (mit 5,85 Milliarden Einträgen) – bestehend lediglich aus der URL der Bilder und weiteren Informationen, insbesondere einer Bildbeschreibung –, den LAION dann über das Internet frei zur Verfügung stellte. Wie und wofür dieser Datensatz verwendet wurde, stand hingegen nicht mehr im Einflussbereich von LAION (die Linksammlung LAION-5B wurde unter einer CC-BY 4.0-Lizenz zur Verfügung gestellt und eine Nutzung für wissenschaftliche Zwecke lediglich empfohlen).

Das LG Hamburg stellt dementsprechend fest: »Vorliegend keiner Entscheidung bedarf daher die [...] Frage, ob das Training von Künstlicher Intelligenz in seiner Gesamtheit der Schrankenregelung des § 44b UrhG unterfällt oder nicht«. Und doch widmet sich die Entscheidung in weiten Teilen genau dieser Frage. Problematisch daran ist zunächst, dass die tragenden Erwägungen zur Schranke des § 60d UrhG nicht die gleiche (und: zu wenig) Aufmerksamkeit erhalten. Zudem scheint das Gericht zwischen all den *obiter dicta* seine eingangs vorgenommene klare Trennung zwischen den Handlungen LAIONs und der Verwendung der Linksammlung durch Dritte (zum Zwecke des Trainings von KI-Modellen) selbst aus dem Blick zu verlieren: Die spätere Nutzung des Datensatzes durch »auf dem Gebiet künstlicher neuronaler Netze Forschende« ist ausschlaggebend dafür, dass LAION als Forschungsorganisation eingeordnet wird und sich auf die Privilegierung für wissenschaftliche Forschung berufen kann.

II. Rechtliche Erwägungen

1. § 60d UrhG

Tragend für die Abweisung der Klage ist § 60d UrhG: Die Vervielfältigungen der Bilder durch LAION seien gerechtfertigt, weil sie für Text und Data Mining (TDM) zu Zwecken der wissenschaftlichen Forschung erfolgten.

Ob die Handlungen LAIONs als TDM einzuordnen sind, prüft das Gericht schon im Zusammenhang mit § 44b UrhG. Richtigerweise wird dabei nur auf die Analyse des Bilds auf

Übereinstimmungen zwischen Bild und Bildbeschreibung abgestellt. Diese lässt sich in der Tat recht problemlos als TDM qualifizieren, weil sie der Gewinnung von Informationen über Korrelationen (zwischen Beschreibung und Inhalt) dient. Die daran anschließende intensive Auseinandersetzung mit der Frage, ob eine spätere Nutzung der unter den von LAION kuratierten Links abrufbaren Inhalte zum Zweck des Trainings generativer KI ebenfalls ein TDM darstellt, ist nicht entscheidungserheblich (siehe dazu unten 2.).

Wesentlich schwieriger zu beantworten ist dagegen, ob auch die übrigen Voraussetzungen des § 60d UrhG erfüllt sind. Die Schranke greift nämlich nur für »Zwecke der wissenschaftlichen Forschung«. Absatz 2 S. 1 präzisiert insofern: »Zu Vervielfältigungen berechtigt sind Forschungsorganisationen«. Dies sind nach S. 2 »Hochschulen, Forschungsinstitute oder sonstige Einrichtungen, die wissenschaftliche Forschung betreiben«. Zusätzliche Voraussetzung ist, dass sie (1.) entweder nicht kommerzielle Zwecke betreiben, (2.) sämtliche Gewinne in die wissenschaftliche Forschung reinvestieren oder (3.) im Rahmen eines staatlich anerkannten Auftrags im öffentlichen Interesse tätig sind. Im Gegensatz zu § 60d UrhG aF ist eine nicht-kommerzielle Tätigkeit also nicht mehr zwingend. Entscheidendes Gewicht kommt hingegen dem Charakter der Tätigkeit zu: Die Einrichtung muss wissenschaftliche Forschung *betreiben* (Wandtke/Bullinger/Bullinger, Urheberrecht, 6. Aufl. 2022, UrhG § 60d Rn. 17). Die Vorschrift wurde in Umsetzung der DSM-Richtlinie umformuliert. Diese definiert Forschungsorganisationen u.A. als »Einrichtung, deren vorrangiges Ziel die wissenschaftliche Forschung oder die Lehrtätigkeit [...] ist«, Art. 2 Nr. 1 DSM-RL. Unter wissenschaftlicher Forschung wird das »methodisch geleitete Generieren neuen Wissens« verstanden (Caliess/Ruffert/Ruffert, EUV/AEUV, 6. Aufl. 2022, GRCh Art. 13 Rn. 6). Hieran orientiert sich auch das Landgericht Hamburg und stellt zudem richtigerweise fest, dass das tatsächliche Erreichen eines späteren Erkenntnisgewinns nicht erforderlich ist und auch Tätigkeiten privilegiert sind, die nicht unmittelbar mit der Gewinnung neuen Wissens verbunden sind, sondern diese lediglich vorbereiten. Das entspricht dem Schutzgehalt von Art. 13 GRCh, der für die Auslegung des europäischen Forschungsbegriffs heranzuziehen ist und nach allgemeiner Ansicht auch vorbereitende und unterstützende Tätigkeiten umfasst (Jarass/Jarass, EU-Grundrechte-Charta, 4. Aufl. 2021, GRCh Art. 13 Rn. 8). Vor diesem Hintergrund stellt das Gericht fest: »Zwar mag die Erstellung des Datensatzes als solche noch nicht mit einem Erkenntnisgewinn verbunden sein; sie ist aber grundlegender Arbeitsschritt mit dem Ziel, den Datensatz zum Zwecke späteren Erkenntnisgewinns einzusetzen« (ZUM 2025, XXX). Daraus wird die Schlussfolgerung gezogen, dass LAION wissenschaftliche Forschung betreibe und daher nach § 60d UrhG privilegiert sei (ZUM 2025, XXX).

Diese Subsumtion vermag nicht zu überzeugen. Sie vermengt zwei unterschiedliche Aspekte, nämlich den Begriff der Forschungseinrichtung (Abs. 2 S. 2) und jenen der (einzelnen) Vervielfältigungshandlung zu wissenschaftlichen Zwecken (Abs. 1). Während nicht jede Handlung in der wissenschaftlichen Forschung unmittelbar einem eigenen Erkenntnisgewinn dienen muss, sollte eine Forschungsorganisation insgesamt sehr wohl das Ziel des Erkenntnisgewinns verfolgen. Nicht genügen kann es, wenn eine Einrichtung sich ausschließlich auf vorbereitende Arbeitsschritte beschränkt und ein Erkenntnisgewinn dann allenfalls bei Dritten erfolgt. Sonst wären etwa auch Bleistifthersteller (die ihre Gewinne reinvestieren) Forschungsorganisationen. Auf den Fall angewendet bedeutet das: Dass die von LAION erstellte Linksammlung von Dritten (auch) zu Forschungszwecken, insbesondere zum Training neuronaler Netze, genutzt wird, ist für die Qualifikation des Vereins als Forschungseinrichtung unter Abs. 2 S. 2 nicht ausschlaggebend. Stattdessen wäre ein genauerer Blick auf LAIONs (Forschungs-)Tätigkeiten zu richten gewesen. Sind zB die Research Paper, die Mitarbeiter von LAION veröffentlicht haben, Zeugnis einer

Forschungstätigkeit? Oder ist in der eigenen Entwicklung von Datensätzen, KI-Modellen und Tools wissenschaftliche Forschung zu erblicken? Dass die von LAION verfolgten Projekte genügen, um den Verein letztlich als Forschungseinrichtung auszuweisen, ist sehr gut möglich. Das Gericht hätte hierauf aber tatsächlich eingehen müssen (entgegen ZUM 2025, XXX).

2. § 44b UrhG

Nachdem das Gericht knapp die Anwendung von § 44a UrhG ablehnt, verwendet es den Großteil der Ausführungen auf § 44b UrhG. Dabei grenzt das Gericht die hier streitgegenständliche Vervielfältigung bei der Erstellung eines Datensatzes von einem anschließenden KI-Training mit diesem Datensatz ab. Die Frage, welche Fernziele oder -wirkungen einzubeziehen sind, taucht wiederholt in der Entscheidung auf. Die Entscheidung, ein potenzielles späteres KI-Training an dieser Stelle außer Betracht zu lassen, überzeugt. Anderenfalls würde es für die Beurteilung, ob eine Rechtsverletzung vorliegt, keinen Unterschied machen, ob der Datensatz unmittelbar nach der Analyse gelöscht würde, obwohl es dann niemals zu einem Training käme. Die Einbeziehung des KI-Trainings ist zum Schutz der Interessen von Urheberinnen auch nicht erforderlich: Vervielfältigungen von Nutzern des Datensatzes, die die Bilder zum Training gebrauchen möchten, sind in der Regel ebenfalls an § 44b UrhG zu messen, wo etwaige Unterschiede (und auch ein Nutzungsvorbehalt, dazu sogleich unter b)) Berücksichtigung finden können. Gleiches gilt für den Drei-Stufen-Test.

a) Ausführungen zur Zulässigkeit des KI-Trainings

Das Gericht hätte es bei dieser Abgrenzung belassen können. Dennoch positioniert es sich im Folgenden auch *obiter* zur Zulässigkeit des KI-Trainings selbst. Das Gericht sieht die Anwendbarkeit der TDM-Schranke nicht als Auslegungsfrage, sondern erwägt eine teleologische Reduktion der vom Wortlaut grundsätzlich einschlägigen Schranke, was die Argumentationslast umkehrt. Dabei setzt es sich insbesondere mit den Ansichten von *Dornis/Stober* (Urheberrecht und Training generativer KI-Modelle, 2024, S. 96 ff. [Gutachten im Auftrag der Initiative Urheberrecht]) und *Schack* (NJW 2024, 113, 114) zur Unterscheidung von Inhalt und Ausdruck auseinander.

Zu folgen ist dem Gericht darin, dass die aus der Informationswissenschaft stammende Differenzierung nach Informationsdimensionen, die zwischen struktureller, syntaktischer und semantischer Information unterscheidet (vgl. etwa *Zech*, Information als Schutzgegenstand, 2012, S. 45 et passim), nur begrenzt geeignet ist, digitalisierte Werke zu strukturieren (so schon *Kuschel*, Der Erwerb digitaler Werkexemplare, 2019, S. 124 ff.). Syntaktische Information ist die »durch eine Menge von Zeichen und deren Beziehung zueinander bestimmte Information« (*Zech*, Information als Schutzgegenstand, 2012, S. 37); die Semantik wiederum betrifft die Bedeutungsebene eines Textes (*Grimm*, Digitale Kommunikation, 2005, S. 8). Es trifft zwar zu, dass Werke sich in syntaktischer Information ausdrücken, der Werkschutz ist einerseits aber keinesfalls auf diese syntaktische Ebene begrenzt. Man denke etwa an literarische Figuren: Ihr Schutz erschöpft sich nicht in der der konkreten sprachlichen Beschreibung, sondern erstreckt sich auf die durch »unverwechselbare Kombinationen äußerer Merkmale, Charaktereigenschaften, Fähigkeiten und typischen Verhaltensweisen« geprägte »Persönlichkeit« (vgl. BGH ZUM 2014, 234 Rn. 26 – Pippi-Langstrumpf-Kostüm). Auf der anderen Seite ist nicht überzeugend, dass die TDM-Schranke auf die Auswertung der Inhaltsebene von Werken beschränkt sein sollte. So wird TDM etwa in der Literaturwissenschaft eingesetzt, um rhythmische Muster in Gedichten zu erkennen (Rhythmicalizer. A digital tool to identify free verse prosody, perma.cc/DN92-JMYN) oder Hinweise auf eine voranschreitende Alzheimer-Erkrankung von Agatha *Christie* zu entdecken

(*Le/Lancashire et. al.*, Longitudinal detection of dementia through lexical and syntactic changes in writing: a case study of three British novelists, 2011, perma.cc/HZ58-TMC4). Ausgewertet wird dabei gerade die Ausdrucksebene (zB Wortwahl, Wiederholungen, Ausdrucksspektrum) und nicht die enthaltenen Informationen.

Ganz wesentliche Bedeutung misst das Gericht schließlich Art. 53 Abs. 1 lit. c KI-VO zu, der den Anbietern von KI-Modellen in der Union auferlegt, eine Strategie zur Einhaltung des Urheberrechts der Union zu implementieren und insbesondere einen »gemäß Artikel 4 Absatz 3 der Richtlinie (EU) 2019/790 geltend gemachten Rechtsvorbehalt[]« zu berücksichtigen. Der Vorschrift und den dazugehörigen Erwägungsgründen 105 und 106 entnimmt nicht nur das LG Hamburg, dass der Unionsgesetzgeber das Training generativer KI-Modelle als Fall des TDM ansieht. Allerdings sollten zwei Punkte dabei nicht unberücksichtigt bleiben: Zum einen stellt die KI-VO in Erwägungsgrund 108 klar, dass die Verordnung »nicht die Durchsetzung der Urheberrechtsvorschriften des Unionsrechts [berührt]« und die Einhaltung der Vorgaben der KI-VO nicht etwaigen urheberrechtlichen Ansprüchen vorgreift. Zum anderen betrifft Art. 53 Abs. 1 lit. c KI-VO auch noch andere KI-Modelle als generative KI, sodass der Hinweis auf den TDM-Nutzungsvorbehalt auch auf diese abzielen könnte.

b) Nutzungsvorbehalt

Die Möglichkeit des Nutzungsvorbehalts in § 44b Abs. 3 UrhG kann als Anwendungsfall einer prozeduralen Restrukturierung des Ausschließlichkeitsrechts verstanden werden (vgl. *Rostam*, Handlungslasten des Urhebers, iE). Nutzer werden von Transaktionskosten entlastet und müssen nur prüfen, ob ein Vorbehalt vorliegt – Rechtsinhaber tragen die Handlungslast, für das Verbotsrecht den Vorbehalt zu erklären. Auszugleichen ist daher das Interesse der Nutzer an einer möglichst leichten Erkennbarkeit und das der Rechtsinhaber an einer möglichst einfachen Erklärung. Der hierzu vom Landgericht Hamburg gebildete Maßstab kann durchaus überzeugen: Einerseits ist der Vorbehalt so zu erklären, dass eine automatisierte Auswertung stattfinden kann, die das TDM erst ermöglicht (vgl. zu ähnlich gelagerten Nutzungsszenarien BGH ZUM 2010, 580 Rn. 37 – Vorschaubilder I; EuGH ZUM 2014, 289 Rn. 26 – Svensson; EuGH ZUM 2021, 434 Rn. 38, 45 f. – VG Bild-Kunst). Andererseits ist Nutzern zumutbar, die zum Werknutzungszeitpunkt bestehenden Möglichkeiten, einen Vorbehalt zu erkennen, auch einzusetzen.

Erst bei der Subsumtion ergeben sich Probleme, indem das Gericht den hier in natürlicher Sprache verfassten Vorbehalt für maschinenlesbar hält, weil der Beklagte über geeignete Technologie zur Erfassung verfügt habe. Schon rein tatsächlich ist das aber zweifelhaft. Die Kammer nennt als einzigen Anhaltspunkt, dass der Beklagte im Rahmen der Analyse eine inhaltliche Erfassung von Bildbeschreibungen durchzuführen vermochte – was aber nur bereits vorhandene Daten betraf und daher kaum mit der Suche und Interpretation vollkommen unbekannter Vorbehaltserklärungen gleichgesetzt werden kann. Der Schluss überzeugt aber auch inhaltlich nicht, weil durch die Fokussierung auf die technischen Möglichkeiten des *Nutzers* die des *Rechtsinhabers* vollkommen unberücksichtigt bleiben. Die Schranke verlangt richtlinienkonform in erster Linie einen angemessenen, maschinenlesbaren Vorbehalt (vgl. Art. 4 Abs. 3 DSM-RL), damit Nutzer kein Rechtsverletzungsrisiko tragen (vgl. Erwgr. 10 f. DSM-RL; dazu Kuschel/Asmussen/Golla/*Rostam*, Intelligente Systeme – Intelligentes Recht, S. 107, 114 ff.). Ein Gebot, den Aufwand der Rechtsinhaber bei der Erklärung maschinenlesbarer Vorbehalte zu minimieren, besteht demgegenüber nicht. Wenn Rechtsinhaber einen Vorbehalt erklären können, der für sie nur geringen (Mehr-)Aufwand verursacht, gleichzeitig für Nutzer aber eine verhältnismäßig große Entlastung bedeutet (weil er etwa besonders einfach auffindbar oder ressourcenschonend ist), ist nicht einzusehen, jedenfalls aber begründungsbedürftig, warum davon

nicht auch Gebrauch gemacht werden sollte (vgl. in diese Richtung schon BGH ZUM 2010, 580 Rn. 37 – Vorschäbilder I; EuGH ZUM 2021, 434 Rn. 46 – VG Bild-Kunst).

Es ist deshalb nicht ausreichend, nur danach zu fragen, ob Nutzer über irgendeine *geeignete* Technologie verfügen, um einen Vorbehalt zu erkennen – es muss auch die Bandbreite zumutbarer Möglichkeiten der Rechtsinhaber in den Ausgleich einbezogen werden. Vieles deutet darauf hin, dass 2021 gegenüber einer einfachsprachlichen Erklärung mit Blick auf diesen Maßstab vorzugswürdige Mittel zur Verfügung standen (vgl. zu robots.txt etwa Data Provenance Initiative, Consent in Crisis: The Rapid Decline of the AI Data Commons, perma.cc/YT4A-8E2X, S. 5, wonach von 14.000 URLs des Datensatzes Common Crawl, in dem auch das streitgegenständliche Bild aufgeführt war, im Jahr 2021 bereits ein Großteil robots.txt-Einschränkungen unterschiedlicher Spezifität einsetzte).

III. Ausblick

Auch wenn die Bedeutung der erstinstanzlichen Entscheidung nicht überschätzt werden sollte, wird sie die weitere Diskussion vorerst (eine Berufung ist anhängig) beschäftigen. Wünschenswert ist, dass dabei der beschränkte Streitgegenstand des Verfahrens nicht aus dem Blick gerät und die unterschiedlichen technischen Vorgänge und ihre Folgen voneinander getrennt werden. Bei der Anwendung des § 60d UrhG sollte klar benannt werden, was LAION zu einer Forschungseinrichtung macht – allein durch die nachgelagerten Handlungen Dritter kann der Verein diesen Status nicht erlangen. Sollte sich daneben die geäußerte Auffassung zur Maschinenlesbarkeit des einfachsprachlichen Vorbehalts durchsetzen, bedeutet das für KI-trainierende Akteure einen wohl erheblichen Mehraufwand und damit – je nach territorialer Reichweite von Art. 53 Abs. 1 lit. c KI-VO – einen Standortnachteil der EU. Insofern kann die Entscheidung aber nur als erster, unverbindlicher Beitrag der Rechtsprechung zur Debatte verstanden werden.

Die Verfasserin Prof. Dr. Linda *Kuschel*, LL.M. (Harvard) ist Juniorprofessorin für Bürgerliches Recht, Immaterialgüterrecht sowie Recht und Digitalisierung an der Bucerius Law School, Hamburg. Der Verfasser Darius *Rostam* ist wissenschaftlicher Mitarbeiter an der Juniorprofessur für Bürgerliches Recht, Immaterialgüterrecht sowie Recht und Digitalisierung an der Bucerius Law School, Hamburg.