Bild: Karl-Josef Hildenbrand / dpa

Füttern verboten?

Wie sich Verlage und Autoren gegen KI-Bots wehren

Sprachmodelle wie ChatGPT grasen ungefragt Texte im Web ab. Sie trainieren damit für ihre Antworten, nehmen so aber Urhebern und Verwertern Geld für Klicks und Werbung weg, werfen diese ihnen vor. Die Verlagsbranche will sich wehren, das ist bei der aktuellen Rechtslage gar nicht so einfach.

Von Falk Steiner

Selbstlernende Systeme mit sogenannter „künstlicher Intelligenz“ benötigen authentische Trainingsdaten in rauen Mengen. Das betrifft neben den KI-Bildgeneratoren ganz besonders generative Text-KIs wie ChatGPT von OpenAI.

Deshalb grasen die Systeme Websites ab, verleiben sich beispielsweise journalistische Texte hochwertiger Medien ein, lernen daraus und spucken die umformulierten Inhalte innerhalb weniger Sekunden in Antworten wieder aus. KI-Suchmaschinen wie Bing, Neeva und You.com geben wenigstens noch die Quellen ihres Lernmaterials an; bei ChatGPT kann man nur spekulieren, mit welchen Wagenladungen an Texten und Büchern es trainiert wurde.

Bei Verlegern und Autoren schrillen deshalb die Alarmglocken. Sie sehen ihre Urheber- und Verwertungsrechte verletzt. In Deutschland erklärten etwa der Bundesverband der Digitalpublisher und Zeitungsverleger (BDZV) und der Medienverband der freien Presse (MVFP) gemeinsam: „Eine Verwertung von Verlagsangeboten durch KI-Sprachmodule für die Veröffentlichung konkurrierender Inhalte ist unseres Erachtens nur mit einer Lizenz des Verlages zulässig“. Kurz: Die Branche will das Leistungsschutzrecht auch aufs sogenannte Textmining von KIs ausgedehnt wissen.

Diese Forderung widerspricht jedoch dem politischen Trend auf europäischer Ebene: Die EU will ihre Mitgliedsstaaten bei der KI-Entwicklung nicht ausbremsen. Deshalb passt sie seit einigen Jahren die rechtlichen Rahmenbedingungen Stück für Stück an, so auch 2019 mit der Reform der Urheberrechtsrichtlinie. Darin schränkt sie die Rechte von Urhebern gezielt zur Förderung von KI „made in Europe“ ein.

Nach Artikel 3 der Richtlinie ist Textmining zu Forschungszwecken grundsätzlich zulässig. Dies gilt beispielsweise, wenn Universitäten neu entwickelte KI-Modelle trainieren wollen. Artikel 4 definiert eine Art „Opt-out“ für kommerzielle Zwecke: „Textentnahmen“ von Websites sind so lange möglich, bis der Rechteinhaber „in angemessener Weise“ und „mit maschinenlesbaren Mitteln“ einen Vorbehalt einlegt. Dieser könne in den Metadaten, aber auch in den Geschäftsbedingungen vermerkt werden, erfährt man recht beiläufig in Erwägungsgrund 18 zum Gesetz.

Rechtssicherheit vorhanden?

Die deutsche Umsetzung dieser europäischen Richtlinienvorgabe erfolgte Mitte 2021 in § 44b des Gesetzes zur Anpassung des Urheberrechts an die Erfordernisse des digitalen Binnenmarkts: Wer die Rechte an seinen Inhalten nicht ausdrücklich und maschinenlesbar vorbehält, dessen Daten und Texte dürfen fürs KI-Training demnach ohne Nachfrage verwendet werden – eine sogenannte Schranke für das Urheberrecht. „Sie schafft Rechtssicherheit für kommerzielle Datenanalysen“, erläutert Prof. Benjamin Raue, Direktor des Instituts für Recht und Digitalisierung Trier.

Folgt man der Rechtsauffassung von Raue, so könnten KI-Anbieter nahezu beliebig Texte, Bilder und Metadaten von Websites abgreifen. Die Inhalte dürfen allerdings nicht dauerhaft gespeichert werden, sondern nur so lange, wie es für das Anlernen der Modelle zwingend notwendig ist. „Gelöscht werden müssen nur die urheberrechtlich geschützten Ausgangsmaterialien, nicht aber die gewonnenen Erkenntnisse“, betont Rechtsprofessor Raue.

Nach Ansicht der Verlegerverbände sind solche Nutzungen jedoch nicht durch den Wortlaut des Gesetzes gedeckt. „Ins-

besondere die gesetzliche Schranke für sogenanntes Text- und Datamining ändert daran nichts“, erklärten sie. Die größte Angst der Anbieter: Google, Meta & Co. könnten die Neugier und das Interesse an den Inhalten noch vor dem Besuch der Medien-Websites befriedigen, indem sie mithilfe von KI-Bots Zusammenfassungen von Artikeln erstellen, und sich so einen unfairen Vorteil verschaffen, ohne dafür auch nur einen einzigen Journalisten beschäftigen zu müssen.

Maschinenlesbarer Vorbehalt

Aber müsste darüber überhaupt gestritten werden? Immerhin könnten doch alle, die dies nicht wollen, einen maschinenlesbaren Ausschluss formulieren. Doch hier stellt sich die Frage nach dem Wie und Wo. Für Webseiten scheint die Datei robots.txt der naheliegendste zu sein. Sie ist die Datei, mit der Anbieter den Crawlern von Suchmaschinen aller Art mitteilen, welche Seiten sie indexieren dürfen und welche nicht. Fast jede größere deutsche Publikation nutzt diese Methode. Doch der zugrunde liegende, fast 30 Jahre alte Robots Exclusion Standard sieht bislang keine spezifischen Anweisungen gegen KI-Crawler vor.

Weil der besagte § 44b des neuen Urheberrechts in seinem Absatz 3 nur allgemein die „maschinenlesbare Form“ eines Nutzungsvorbehalts vorsieht, hilft auch hier ein Blick in die Begründung des Gesetzgebers, also der damaligen schwarzroten Bundesregierung. Derzufolge kann der Vorbehalt „auch im Impressum oder in den Allgemeinen Geschäftsbedingungen (AGB) enthalten sein, sofern er auch dort maschinenlesbar ist“. Zwar ist dieses Gebot nicht unmittelbar Gesetz, dürfte aber in Zweifelsfällen von Gerichten herangezogen werden. Der Vorbehalt gilt übrigens laut Begründung nicht rückwirkend, sondern „ex nunc“, also erst, wenn er sich auf der Website befindet.

Die Frankfurter Allgemeine Zeitung (FAZ) gilt als Qualitätsmedium. Einen Eintrag etwa in der robots.txt hält man dort nicht für notwendig: „Der Vorbehalt muss lediglich maschinenlesbar sein“, erklärt eine Verlagssprecherin. „Diese Voraussetzung erfüllt stets ein elektronischer Text, eingebunden an zentralen Stellen der Publikation.“ Deshalb finden Nutzer und KI-Crawler seit einigen Wochen im Impressum von Faz.net folgenden Zusatz: „Die Frankfurter Allgemeine Zeitung GmbH behält sich eine Nutzung ihrer Inhalte für kommerzielles Text- und Datamining im Sinne von § 44b UrhG ausdrücklich vor. Für den Erwerb einer entsprechenden Nutzungslizenz wenden Sie sich bitte an nutzungsrechte@faz.de.“

Ähnliche Passagen fügen immer mehr Verlage ein, jüngst beispielsweise auch die Münchener Süddeutsche Zeitung GmbH im Impressum ihres Webauftritts Sueddeutsche.de. Unter Urheberrechtlern herrscht die Meinung vor, dass jede Texterklärung auf einer Webseite als „maschinenlesbar“ anzusehen ist. Wie intelligent wäre eine KI, die nicht einmal einen Rechtevorbehalt in Textform interpretieren könnte?

Wer crawlt hier wie?

Allerdings können die Anbieter kaum feststellen, inwieweit KI-Betreiber die Regeln einhalten. Denn bislang ist es technisch nicht möglich, die Crawler der Künstlichen Intelligenz zu identifizieren. Und wenn sie als Suchmaschinen-Crawler herumstreifen, müssten Seitenbetreiber fürchten, mit ihrem Stoppschild Reichweite zu verlieren.

Sollte etwa das bei der Microsoft-Suchmaschine Bing seit einigen Wochen testweise integrierte OpenAI-System auf der Arbeit von Microsofts Crawler Bingbot aufsetzen, stünden Anbieter vor einem kaum lösbaren Problem. Würde man beispielsweise den Bingbot kategorisch durch einen Vorbehalt von Webseiten ausschließen, wäre die Konsequenz klar: Die Seiten würden auch in der Suche nicht mehr als Ergebnis angezeigt.

Die Verlegerverbände hoffen auf ein hartes Durchgreifen der EU-Kommission, die solche Bündelungen per Wettbewerbsrecht und dem Digital Markets Act (DMA) untersagen könnte. Microsoft hat auf mehrere Anfragen zu diesem Thema bis Redaktionsschluss nicht geantwortet.

Dass die KI-Modelle auf vorhandenem, erreichbarem Wissen aufbauen und dies kostenlos schürfen dürfen, stört auch die Verwertungsgesellschaften, die stellvertretend Tantiemen einsammeln und an die Rechteinhaber ausschütten. Denn als der Gesetzgeber die Text- und Datamining-Schranke einführte, hat er eines ausgeschlossen: Eine Pauschalvergütung für Rechteinhaber durch KI-Nutzung, die nicht genau zu ermitteln ist und deshalb im Ungefähren bleibt. Es sei „problematisch, dass sowohl bei der gesetzlichen Erlaubnis für kommerzielles Text- und Datamining als auch bei der Regelung für wissenschaftliches Text- und Datamining“ kein Vergütungsanspruch vorgesehen sei, erklärte Anette Frankenberger, Sprecherin der VG Wort.

Dieser Ansicht dürften sich weitere Akteure anschließen, falls das Thema weiter in den öffentlichen Fokus rückt. Viele Autoren und Verleger haben bislang keinen Schimmer davon, dass ihre Angebote ausgelesen und zum KI-Training genutzt werden dürfen. Mehrere unserer Anfragen endeten mit überraschten Rückrufen – davon habe man nichts gewusst und sich deshalb noch nicht damit beschäftigt. Europas Drang, endlich einmal regulatorisch an der Spitze zu stehen, scheint einige der Betroffenen abgehängt zu haben. *(hob@ct.de)* c't

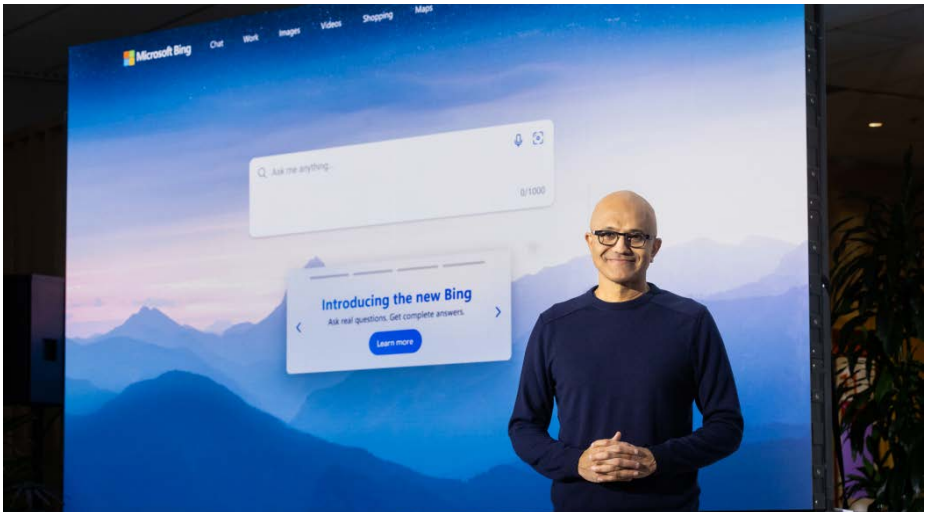

Bild: Dan DeLong / Microsoft

CEO Satya Nadella zeigt, wie Microsoft die Suchmaschine Bing mit einem KI-Bot koppelt. Autoren und Verlage befürchten eine unfaire Konkurrenz, die ihre Einnahmen schmälert.