Humanistische Union

GPAI-Modelle und das erste KI-Gesetz der Welt

*Es war ein zähes Ringen in der Europäischen Union, um Regelungen eine Verordnung zu Künstlichen Intelligenz zu finden. Es geht darum KI-Systeme so zu regulieren, dass Persönlichkeitsrechte und das Prinzip des Diskriminierungsverbots eingehalten werden. Doch Vieles ist dabei noch strittig. Rolf Schwartmann und Moritz Köhler gehen einigen der strittigen Details auf den Grund. Dabei betonen Sie, dass der Mensch (und nicht die Maschine) im Mittelpunkt der Betrachtung stehen müssen. Daher ist es auch nötig, dass Nutzer*innen die Wirkweise einer KI nachvollziehen können.*

Nach einem Verhandlungsmarathon von fast drei Tagen Anfang Dezember 2023 scheint die Zukunft der KI-Verordnung gesichert. Die Vertreter*innen der Mitgliedstaaten der Europäischen Union, die Kommission und das Europaparlament haben zäh gerungen und wollen, dass die EU ein starkes und gutes Signal für eine vertrauenswürdige KI in die Welt sendet.

Im Detail ist noch vieles streitig (Hacker 2023). Im Grundsatz soll es einen risikoorientierten Ansatz geben, der gestufte Regeln abhängig vom Risiko vorsieht, das von einer Künstlichen Intelligenz ausgeht. Anwendungen mit minimalem Risiko, die etwa Produkte zum Kauf empfehlen, sollen von strengen und speziellen Verpflichtungen befreit sein (Europäische Kommission 2023). Risikoreiche Anwendungen, die in der Verordnung so präzise wie möglich benannt werden, müssen strenge und spezifische Verpflichtungen erfüllen. Das gilt etwa für KI-Systeme, die Einfluss auf den Zugang zu Bildungseinrichtungen haben, oder für solche, die in der Medizin oder in der Rechtspflege eingesetzt werden.[i] Sie müssen nachweislich robust und störungsfrei laufen, ihre Betreiber*innen müssen deren Aktivitäten transparent machen sowie protokollieren, und sie müssen Sicherheit gegenüber Hackerangriffen aufweisen. Manche Systeme sind auch ganz verboten, weil sie ein nicht akzeptables Risiko aufweisen. Das gilt etwa für Sprachsysteme, die Menschen manipulieren oder gar Kinder negativ beeinflussen können oder für Anwendungen, die Staaten einsetzen können, um Bürger*innen zu bewerten (Art. 5 KI-VO-E).

Unabhängig von ihrem konkreten Einsatzzweck haben die genannten Anwendungen gemeinsam, dass sie auf großen Datenpools fußen. Aus ihnen beziehen die genannten Systeme ihr „Wissen“. Diese Datenpools bezeichnet der Entwurf der KI-Verordnung als General Purpose KI (GPAI) (Europäische Kommission 2023). Damit ist gemeint, dass diese Datenpools nicht darauf spezialisiert sind, bestimmte Aussagen hervorzubringen oder zu bestimmten Zwecken eingesetzt zu werden. Es handelt sich bei diesen sogenannten Basismodellen um mächtige Modelle, denen über das „Füttern“ mit Daten eine allgemeine Sprachverarbeitungskompetenz antrainiert wurde. Ihre eigentliche Spezialität ist nicht, dass sie Wissen enthalten, sondern, dass sie aus Wissensdatenbanken mit Mitteln der Wahrscheinlichkeitsrechnung sinnvolle und zusammenhängende Texte hervorbringen. Sie arbeiten sogenannten Prompts, sprich Anfragen, ab. Je spezifischer die Frage ist, desto präziser ist die Antwort. Aus der Datenbasis wird keine Antwort auf eine Frage generiert, mit der die Datenbasis etwas Unrechtes zum Ausdruck bringen möchte.

Allerdings bergen diese Datenmodelle ein besonders Risiko. Sie liefern zwar Texte, die sich für den

menschlichen Nutzer*innen als sinnvolle und zusammenhängende Aussage darstellen. Das System produziert diese Texte allerdings nicht in einem Sinnzusammenhang, sondern Wort für Wort auf der Basis von Wahrscheinlichkeiten. Überspitzt und vereinfacht gesagt treffen bei der Anwendung eines KI-Systems durch eine*n menschlichen Nutzer*in zwei gegenläufige Modelle aufeinander: Dialektik und Stochastik. Dialektik hilft nach einem sehr vereinfachten Verständnis, der in unbegrenzten Möglichkeitsräumen sich bewegenden menschlichen Vernunft, über aufeinander bezogene Gründe (These) und Gegengründe (Antithese) zu Einsichten (Synthese) zu gelangen. Stochastik hingegen ist, abermals vereinfacht ausgedrückt, ein Bereich der Mathematik, der Wahrscheinlichkeitstheorie und Statistik verwendet. Stochastik untersucht verkürzt gesagt die mathematische Modellierung von zufälligen Ereignissen. Antworten von Robotern auf menschliche Fragen können allenfalls eine Simulation der dialektischen Methode sein, da ihnen menschliche Vernunft fremd ist (Schwartmann 2023: 132). Der Mensch ist im Gegensatz dazu in der Lage, vernünftig zu sein und danach zu handeln. Er muss entscheiden, ob er den Impulsen seines limbischen Systems ohne Reflexion nachgibt oder sein zugegebenermaßen nicht vorurteilsfreies Denken hinterfragt, bevor er danach handelt. (Hansen/Köhler/Schwartmann 2023)

Der Mangel an Vernunft als Filter der auf Basis von Wahrscheinlichkeiten gefundenen Ergebnisse führt dazu, dass die KI-Systeme bisweilen Texte generieren, deren Aussagen mit dem Anstandsgefühl der Gesellschaft unvereinbar, sinnlos oder schlicht falsch sind. Die Gründe für derartige Aussagen sind mannigfaltig: Sie können auf Verzerrungen in der Datenbasis zurückzuführen sein, etwa wenn eine bildgenerierende KI auf den Befehl „Zeichne einen Doktor“ stets einen weißen, männlichen Arzt generiert (Hansen/Köhler/Schwartmann 2023). Oder sie können auf Lücken in der Datenbasis beruhen, welche die KI fehlerhaft ausfüllt, um dem Befehl des Nutzenden gerecht zu werden. So wird aus dem Landesdatenschutzbeauftragten für Baden-Württemberg ein international renommierter Ornithologe. Schließlich kann der menschliche Nutzende die Schwachstellen dieser Technologie auch bewusst ausnutzen: Provoziert er das KI-System nämlich durch einen Prompt dazu, fragwürdige Ergebnisse zu liefern, dann wird es diese aufgrund der mathematischen Gesetze der Wahrscheinlichkeitsrechnung auch liefern. So wird das KI-System bei Eingabe eines entsprechenden Prompts den Satz „Das Schöne an der Diskriminierung von Schwachen ist, …“ sprachlich sinnvoll, aber menschenrechtlich und moralisch inakzeptabel vervollständigen. Will man derartige Ergebnisse verhindern, muss man den Sprachmodellen diese Antworten abtrainieren.

Ein entscheidender Streitpunkt der KI-Verordnung war es, wer die Verantwortung für derartige Ergebnisse der KI trägt. Müssen die Herstellerfirmen ihre Datenmodelle so programmieren, dass sie Werte befolgen? In diesem Fall müssten staatliche Strukturen darüber entscheiden, welche Ergebnisse die Maschine liefern darf und welche Prompts sie blocken sollte. Das mag aus bürgerrechtlicher Perspektive abzulehnen sein. Zu groß scheint die Manipulationsgefahr. Nicht umsonst hat das Bundesverfassungsgericht in jahrzehntelanger Rechtsprechung den Grundsatz der Staatsferne im Rundfunk entwickelt (BVerfGE 12, 205; 31, 314; 57, 295; 121, 30; 136, 9). Alternativ könnte der Gesetzgeber den Programmierer*innen die Direktive auf den Weg geben, dass die Datenmodelle keine moralisch verwerflichen oder rechtlich unzulässigen Ergebnisse generieren dürfen. Das führt allerdings dazu, dass die Programmierer*innen die Macht bekommen, zwischen Gut und Böse zu unterscheiden. Man würde also die Entscheidungsgewalt über die Grenzen des Sagbaren in die Hände Privater legen. Das wäre möglicherweise noch vertretbar, wenn man diese Grenze präzise ziehen könnte. In Fällen mit Bezug zum Persönlichkeitsrecht etwa werden aber Probleme deutlich: Hier liegen die Grenzen zwischen erlaubter Meinungsäußerung und Persönlichkeitsrechtsverletzung besonders nah beieinander, und sie verschieben sich aufgrund neuer Rechtsprechung ständig. KI-Systeme sind bislang weit davon entfernt, die komplexen rechtlichen Abwägungen im Bereich des Persönlichkeitsrechts vornehmen, das heißt mathematisch simulieren zu können. Es darf aber zugleich bezweifelt werden, dass solche Abwägungen in der demokratischen Gesellschaft jemals von einer KI übernommen werden sollten.

Sollten die Datenmodelle also unreguliert bleiben? Damit wären die Nutzer*innen in die Verantwortung gezogen. Es würde somit stets dem Einzelnen obliegen, die auf Basis seiner Prompts generierten Ergebnisse auf mögliche Rechtsverletzungen zu prüfen und seine Handlungen an die Ergebnisse dieser Prüfung anzupassen. Problematisch sind dabei kognitive Verzerrungen, die typischerweise beim Umgang mit den Ergebnissen einer KI auftreten können. Zu nennen ist zunächst der Automatisierungsbias: Bereits der Entwurf der KI-Verordnung wusste um die mögliche „Neigung zu einem übermäßigen oder automatischen Vertrauen in das von einem Hochrisiko-KI-System hervorgebrachte Ergebnis“ (Art. 14 Abs. 4 Buchst. B KI-VO-E). Wenn es eine scheinbar vertrauenswürdige Instanz gibt, die für den Menschen Entscheidungen trifft, ist der Mensch also gerne bereit, dieser Instanz zu vertrauen und sich selbst, zumindest in der eigenen Wahrnehmung, einer Verantwortung zu entziehen. Die Gefahr, einem Automatisierungsbias zu unterliegen, dürfte besonders groß sein, wenn KI-Systeme Ergebnisse hervorbringen, die derart ausgereift formuliert sind, dass beim Menschen der Eindruck objektiver Wahrheit entsteht und wenn Entwickler*innen diesen Eindruck dadurch bestärken, dass sie das KI-System in der Ich-Form kommunizieren lassen (Hansen/Köhler/Schwartmann 2023).

Neben dem Automatisierungsbias steht auch der sogenannte Ankereffekt einer Übertragung der vollen Verantwortung auf die Nutzer*innen entgegen. Der Ankereffekt beschreibt das psychologische Phänomen, dass Menschen in ihrer Entscheidungsfindung unbewusst von vorhandenen Umgebungsinformationen beeinflusst werden (Kahnemann 2012: 152). Als Ursache gilt einerseits eine unzureichende Korrektur des vom Anker gesetzten Ergebnisvorschlags (Tversky/Kahnemann 1974: 1124ff.). Andererseits werden im Gehirn selektiv solche Gedächtnisinhalte aktiviert, die mit dem Anker kompatibel sind (Strack/Mussweiler 1997: 437ff.). Einfach gesagt: Wer noch keine gefestigte Meinung hat, kann sich dem Ankereffekt nicht entziehen. Der Ankereffekt ist sogar messbar, wenn der Anker offensichtlich zufällig gewählt wurde und damit keinerlei Informationsgehalt in sich trägt. Der Psychologe Daniel Kahneman beschreibt in seinem Buch *Schnelles Denken, langsames Denken* ein Experiment: Richter sollten eine fiktive Ladendiebin zu einer Freiheitsstrafe verurteilen. Vor der Urteilsfindung warfen sie einen gezinkten Würfel, der stets eine Drei oder eine Neun zeigte. Richter, die eine Neun gewürfelt hatten, verurteilten die Ladendiebin im Durchschnitt zu einer Freiheitsstrafe von acht Monaten. Fiel der Würfel auf Drei, lautete das Urteil auf fünf Monate. (Kahnemann 2012: 159ff.) Was ergäbe ein Experiment, bei dem nicht ein Würfel eine scheinbar zufällige Zahl anzeigt, sondern eine KI ein eloquent ausgearbeitetes Ergebnis vorschlägt?

Die Komplexität des Problems und die unvorhersehbaren Folgen der Entscheidung führten dazu, dass die Frage der Regulierung von GPAI-Modellen im Trilog bis zuletzt diskutiert wurde. Dabei wurden im Wesentlichen drei Ansichten vertreten: Insbesondere die Hersteller und Anbieter*innen der Foundation Models lehnten eine Regulierung gänzlich ab. Sie wollten die Betreiber*innen und die Nutzer*innen in die Verantwortung ziehen. Sie beriefen sich vor allem darauf, dass Herstellerfirmen und Anbieter*innen nicht dafür verantwortlich seien, welche Ziele Betreiber*innen und Nutzende mit dem Einsatz der GPAI-Modelle verfolgen. Schließlich sei auch der*die Anbieter*in eines Hammers nicht verantwortlich, wenn das Werkzeug als Mordwaffe eingesetzt wird. Das Europäische Parlament vertrat die Gegenposition. Es wollte die Anbieter*innen von GPAI-Modellen verpflichten, bestimmte Mindestanforderungen zu erfüllen. Diese Pflichten sollten die Anbieter*innen von GPAI-Modellen unabhängig davon treffen, ob sie das Modell eigenständig oder eingebettet in ein KI-System auf dem Markt bereitstellen. Hierzu wurde angeführt, dass nur durch eine sachgerechte Zuweisung von Verantwortlichkeiten entlang der gesamten KI-Wertschöpfungskette die Grundrechte der Betroffenen geschützt werden können (Konferenz der unabhängigen Datenschutzaufsichtsbehörden des Bundes und der Länder 2023). Kurz vor Beginn der letzten Runde der Trilog-Verhandlungen veröffentlichten Deutschland, Frankreich und Italien zudem ein gemeinsames Positionspapier, in dem sie eine verpflichtende Selbstregulierung der Entwickler von GPAI-Modellen im Rahmen eines Verhaltenskodexes vorschlugen. Dazu sollten alle Entwickler der GPAI-Modelle verpflichtet werden, sogenannte Model Cards zu definieren und zu veröffentlichen. Mithilfe der auf den

Model Cards angegebenen Informationen sollte es möglich sein, die Funktionsweise des einzelnen Modells, seine Fähigkeiten und seine Grenzen zu verstehen. So sollte es anderen Entwicklerfirmen ermöglicht werden, Verstöße gegen den aufgestellten Verhaltenskodex zu erkennen und zu melden.

Am Ende der Verhandlungen konnte sich das Parlament mit seiner Forderung nach einer harten gesetzlichen Regulierung der GPAI-Entwickler durchsetzen (Europäisches Parlament 2023). Ergebnis der Trilog-Verhandlungen war mit Blick auf die Regulierung von GPAI ein zweistufiger Ansatz. Demnach treffen die Entwicklerfirmen kleinerer GPAI-Modelle – wie das deutsche Unternehmen Aleph Alpha – hauptsächlich Transparenzpflichten. Sie müssen technische Dokumentationen erstellen und diese an Aufsichtsbehörden und an Unternehmen weitergeben, die das Modell in eigene KI-Systeme einbinden wollen. Auf der zweiten Stufe der Regulierung stehen die sogenannten systemischen GPAI-Modelle. Diese werden zunächst auf Basis verschiedener Kriterien, unter anderem der zum Training benötigten Rechenleistung, von der Kommission ausgewählt und unterliegen dann weitergehenden Pflichten. Dazu gehören neben der technischen Dokumentation verpflichtende Maßnahmen zur Gewährleistung von Cybersecurity sowie eine interne Evaluierung des Modells und ein verpflichtender Mechanismus zur Meldung von Vorfällen im Rahmen des Trainings oder des Betriebs des GPAI-Modells.[ii] Die konkreten Pflichten für die Entwickler systemischer GPAI-Modellen sollen in Zusammenarbeit zwischen Kommission, Wissenschaft, Zivilgesellschaft und Industrie entwickelt werden. (Europäische Kommission 2023)

Im Rahmen der Regulierung wurde allerdings darauf verzichtet, Hersteller von GPAI gesetzlich dazu zu verpflichten, KI-Anwendungen so zu trainieren, dass bei deren Einsatz keine Rechtsverletzungen auftreten. Kommt es aufgrund der Ergebnisse eines KI-Systems zu Rechtsverletzungen, ist daher entscheidend, an welchem Punkt der Lieferkette der Fehler eingetreten ist: In Betracht kommt dann eine Haftung des Entwicklers des GPAI-Models, eine Haftung des Unternehmens, das eine Hochrisiko-Anwendung auf Basis des GPAI-Models anbietet oder eine Haftung des*der einzelnen Nutzer*in. Zentral wird insofern Artikel 28 der KI-Verordnung sein, der die Verantwortung entlang der Wertschöpfungskette zuweisen soll. Abschließend wird die Vorschrift die Frage der Verantwortlichkeit indes kaum klären können. Erforderlich ist vielmehr eine KI-Haftungs-Verordnung, die bereits geplant war, aufgrund der stockenden Verhandlungen zur KI-Verordnung allerdings pausiert wurde und erst nach den Wahlen zum EU-Parlament im kommenden Jahr angegangen werden kann.[iii] Jedenfalls solange eine Zuweisung der Verantwortlichkeiten entlang der KI-Wertschöpfungskette noch nicht feststeht, ist jede nutzende Person dazu berufen, die Ergebnisse seiner Prompts kritisch zu hinterfragen und auf etwaige Rechtsverletzungen zu überprüfen.

Wesentliche Voraussetzungen einer verlässlichen Kontrolle von KI und einer sinnvollen Haftungszuweisung sind zumindest grundsätzlich Transparenz und Nachvollziehbarkeit der Wirkweise der Technik. Erst wenn die Gesellschaft weiß, wie die Dienste jedenfalls grundsätzlich funktionieren, kann sie sich über wirksame Regeln für deren Einsatz einigen. Die regulativen Leitlinien, auf die man sich am Ende der Trilog-Verhandlungen geeinigt hat, sind daher jedenfalls im Grundsatz zu begrüßen. Ihre Wirksamkeit wird von der konkreten Ausgestaltung der Regulierung abhängen, welche die Mitarbeiter des Parlaments in den kommenden Wochen und Monaten in den technischen Gesprächen erarbeiten müssen. Läuft alles nach Plan, soll der Gesetzestext bis Anfang Februar 2024 stehen und die KI-Verordnung noch vor der Europawahl im Juni 2024 verabschiedet werden.

Prof. Dr. Rolf Schwartmann *ist Leiter der Kölner Forschungsstelle für Medienrecht an der TH Köln und Vorsitzender der Gesellschaft für Datenschutz und Datensicherheit (GDD) e. V. Er wurde 2018 in die Datenethikkommission der Bundesregierung berufen.*

Moritz Köhler *ist wissenschaftlicher Mitarbeiter der Kölner Forschungsstelle für Medienrecht an der TH Köln.*

Literatur

Europäische Kommission 2023: Pressemitteilung der Europäische Kommission vom 09.12.2023, https://ec.europa.eu/commission/presscorner/detail/%20en/ip_23_6473.

Hacker, Philipp 2023: What's Missing from the EU AI Act, in: *Verfassungsblog* vom 13. Dezember 2023, https://verfassungsblog.de/whats-missing-from-the-eu-ai-act/.

Hansen, Marit/Köhler, Moritz/Schwartmann, Rolf 2023: Was gegen die Voreingenommenheit der KI-Systeme hilft, in: *Frankfurter Allgemeine Zeitung* vom 23.10.2023, https://www.faz.net/aktuell/wirtschaft/unternehmen/was-gegen-die-voreingenommenheit-der-ki-systeme-hilft-19260960.html.

Kahnemann, Daniel 2012: Schnelles Denken, langsames Denken, München.

Konferenz der unabhängigen Datenschutzaufsichtsbehörden des Bundes und der Länder 2023: Pressemitteilung der Konferenz der unabhängigen Datenschutzaufsichtsbehörden des Bundes und der Länder vom 29.11.2023, https://datenschutzkonferenz-online.de/media/pm/23-11-29_DSK-Pressemitteilung_KI-Regulierung.pdf.

Schwartmann, Rolf 2023: Dialektik versus Stochastik: Wissenschaft darf nicht im mathematischen Bermuda-Dreieck der KI verschwinden, in: *Forschung & Lehre* 2023, S. 132-133.

Strack, F./ Mussweiler, T. 1997: Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility, in: *Journal of Personality and Social Psychology*, Jg. 73, H. 3, S. 437-446.

Tversky, Amos/Kahnemann, Daniel 1974: Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty, in: *Science* , Jg. 185, H. 4157, S. 1124-1131.

Anmerkungen:

i Vgl. Anhang III des Vorschlags für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für Künstliche Intelligenz (Gesetz über Künstliche Intelligenz) und

zur Änderung bestimmter Rechtsakte der Union, COM/2021/206 final (im Folgenden KI-VO-E).

[ii] Zenner, DataAgenda Datenschutz Podcast, Folge 48 ab Minute 13:00.

[iii] Zenner, DataAgenda Datenschutz Podcast, Folge 48 ab Minute 16:55.
