

Friedrich Graf von Westphalen

Haben KI-Systeme das letzte Wort – abseits des Rechts?



Mitte Juli rauschte eine wirklich verstörende Nachricht um den Globus: Eine von OpenAI entwickelte KI hatte sich aus der Testumgebung „befreit“, war ins Internet eingedrungen und hatte selbstständig eine Sicherheitslücke bei Hugging Face aufgespürt und ausgenutzt. Die Bundesregierung sprach von einem

„neuen Zeitalter“ der KI. Denn der Mensch hat nicht mehr die vollständige Kontrolle über das Verhalten einer sich selbstständigenden KI. Anfang Juni war eine ähnlich besorgniserregende Nachricht bekannt geworden: Einer der weltweit führenden KI-Entwickler – Anthropic – hatte vorgeschlagen, ein weltweites Moratorium bei der Entwicklung für „besonderes fortgeschrittene Modelle“ der KI-Systeme für die Dauer von sechs Monaten in die Diskussion zu werfen. Grund: Die technische Entwicklung zielt immer rascher und nahezu hemmungslos darauf ab, dass sich die KI selbst programmiert, ohne hinreichende menschliche Aufsicht oder nachhaltige Einflussnahme. Wenig später hat auch Armin Papperger, der Chef von Rheinmetall, diesen Gedanken aufgegriffen und ein weltweit verbindliches Verbot – ähnlich wie bei Atomwaffen – gefordert: Keine „killing robots“ ist seine Losung. Vielmehr soll der Mensch immer beim Einsatz von Waffen das letzte Wort haben und die personale Verantwortung tragen. Es ist ja eine Entscheidung über Leben und Tod. Doch Papperger musste selbst einräumen, dass diese Zurückhaltung wohl dann an ihre Grenzen stößt, wenn der „Feind“ sich anders verhält. Denn Geschwindigkeit ist hier alles. Die spieltheoretisch von der KI empfohlene, vom „Bediener“ blitzschnell umzusetzende Reaktion auf einen unmittelbar drohenden Angriff kommt aber zu spät, wenn der Mensch mit seiner ethischen Pflicht zur Verantwortung das letzte Wort haben muss, weil ja selbst zum Nicken keine Zeit mehr verbleibt – des eigenen Überlebens willen.

Doch diese Stimmen – und auch die von Papst Leo XIV. in seiner historisch zu nennenden Enzyklika „Magnifica Humanitas“ – verhallen trotz aller berechtigten Mahnungen vor den neuen, nicht mehr beherrschbaren Risiken der KI-Systeme weitgehend ungehört. Kein anderer KI-Entwickler hat ja die Anre-

gung von Anthropic nach einem Moratorium – nach einer Zeit der Besinnung und des Innehaltens, der abwägenden Debatte – aufgegriffen. Auch die Rüstungsindustrie rüstet sich immer für den Ernstfall; Präzision und Zerstörungskraft der KI-dominierten Waffen sind stets zu verbessern. Die Aktienmärkte registrieren diese Entwicklung mit tosendem Beifall. Es ist vielleicht dann nur noch ein sehr kleiner Schritt, auch den Einsatz von Atomwaffen der KI anzuvertrauen.

Vor genau zehn Jahren hat die belgische Rechtsphilosophin Mireille Hildebrandt die wichtige Frage in sprachgewandter Doppeldeutigkeit im Titel ihres bemerkenswerten Buchs formuliert: „Smart Technologies and the End(s) of Law“. Kenntnisreich und emphatisch hat sie davor gewarnt, dass die digitale Entwicklung bald darauf hinauslaufen könnte, „to turn written law into a paper dragon“. Genau dieser Frage ist im Blick auf die rasende Fortentwicklung einer sich selbst programmierenden KI jetzt nachzuspüren. Denn es geht um die Bewahrung der Menschenwürde gegenüber einer schier allmächtig sich gebarenden Technik durch das regulierende Recht. Es geht um die Erhaltung der „rule of law“, die ja zuvörderst den Anwälten als lebenslang zu beachtendes Leitmotiv ihres Handelns anvertraut ist.

Das führt in gerader Linie zu der Frage, ob denn die Regeln der KI-VO (2024/1689) so belastbar und noch so zeitgemäß sind, dass das Unionsrecht das Zepter regulatorischen Handelns weiterhin gegenüber den Risiken einer entfesselten Technik in Händen hält und so in verlässlicher Weise die Grundwerte der Unionsbürger schützt. Als maßgebende „Eigenschaften“ hat die EU bei der risikobasierten Einordnung von KI-Systemen immer die „Autonomie“ ihres Verhaltens in den Blick genommen und damit auch das „self-learning“, ebenso aber die „Unvorhersehbarkeit“ ihres Verhaltens. Beides zusammen bildet den Kern der Aussage, die KI sei eben eine „Black Box“. Doch die jetzt zu bewältigende – neue – Frage erstreckt sich auf die gesellschaftlichen Konsequenzen, die sich aus einer KI ergeben, die sich selbst programmiert; der Mensch tritt zur Seite.

Ist also unter diesem Gesichtswinkel die in der KI-VO vorgesehene Definition ihres „Wesens“ so, dass sie auch den Tatbestand der Eigenprogrammierung der KI und ihre damit verbundene

selbstständige Weiterentwicklung trägt? Diese Definition liest sich in Art. 3 Nr. 1 KI-VO wie folgt:

„KI-System“ (ist) ein maschinengestütztes System, das für einen in unterschiedlichem Grade **autonomen Betrieb** ausgelegt ist und das nach seiner Betriebsaufnahme **anpassungsfähig** sein kann und das aus den erhaltenen Eingaben für explizite oder implizite Ziele ableitet, wie Ausgaben wie etwa Vorhersagen, Inhalte, Empfehlungen oder Entscheidungen erstellt werden, die physische oder virtuelle Umgebungen beeinflussen können“.

Hier liegt erkennbar die Sollbruchstelle: Kann ein KI-System auch dann noch als „anpassungsfähig“ bezeichnet werden, wenn es sich – ohne Zutun eines Menschen – selbst programmiert? Wird dann nicht eine neue, eine andere, eine weitere Dimension eines KI-Systems entwickelt? Dafür spricht manches.

Doch im Hintergrund steht der Befund, dass die geltende KI-VO das Ziel verfolgt, eine „menschenzentrierte und vertrauenswürdige“ KI im Binnenmarkt einzuführen und ein „hohes Schutzniveau“ zugunsten der Grundrechte, einschließlich der Demokratie, der Rechtsstaatlichkeit und des Umweltschutzes sicherzustellen. Exemplarisch wird dies in der Regel des Art. 14 Abs. 1 der KI-VO bei der Definition der „Hochrisiko-KI-Systeme“, die so umschrieben ist: Diese „werden so konzipiert und entwickelt, dass sie während der Dauer ihrer Verwendung – auch mit geeigneten Instrumenten einer Mensch-Maschine-Schnittstelle – von natürlichen Personen wirksam beaufsichtigt werden können.“ Doch eine „Beaufsichtigung“ durch einen Menschen ist

wohl dann ausgeschlossen, wenn sich das KI-System selbst programmiert, ohne dass ein Mensch die technischen Abläufe vorhersagen oder gar steuern kann. Ein solches KI-System ist daher wohl als „selbstständig“ zu bezeichnen.

Ob die neue Produkthaftungs-RL (2024/2853) mit ihrem Grundsatz der Gefährdungshaftung bei einem dem Verbraucher zugefügten Schaden den Risiken eines sich selbst programmierenden KI-Systems gewachsen ist, ist auch nicht sicher, könnte aber deswegen – ungeachtet aller Zweifel an der Definition eines KI-Systems – durchgreifen, weil ein solcher Schaden aus der Sicht des Verbrauchers nach Art. 7 Abs. 1 RL 2024/2853 von einem Produkt herrührt, das nicht die Sicherheit geboten hat, die „man erwarten“ kann. Doch: Ob dann auch ein „Hersteller“ als Haftungsgegner gefunden werden kann, der ja eine „natürliche oder juristische“ Person mit personaler Verantwortung für die eingetretene Rechtsgutverletzung sein muss, ist ebenfalls noch recht offen. Keinem belastbaren Zweifel unterliegt jedoch der Befund, dass die auf einem persönlichen Verschulden aufbauende Deliktshaftung des § 823 BGB im gewerblichen Bereich nicht imstande sein wird, den erforderlichen Rechtsschutz beim Versagen eines KI-Systems zu gewährleisten. Alles in allem: eine nicht gerade froh stimmende Erkenntnis über eine sich mehr und mehr verselbstständigende KI, der das Recht noch nichts Entscheidendes entgegenzusetzen hat, um den ausgebrochenen „Geist“ wieder in die Flasche zurückzubringen. Denn es dauert mittlerweile nur drei Monate, bis sich ein neuer Schritt in der Entwicklung der KI als „Superintelligenz“ vollzieht.