Liebe Mitglieder des Netzwerkes der ehemaligen SSAB, liebe Interessierte

In beschleunigtem Tempo werden immer noch potentere KI-Tools entwickelt. Es gibt eigentliche Durchbrüche in Richtung einer AGI (Artificial General Intelligence), von KI-Systemen, die einmal gesetzte Ziele autonom verfolgen. Wir beleuchten im vorliegenden und im nächsten E-Mail einige beunruhigende Aspekte, die ein wichtiges Diskussionsthema in der Bildung sein sollten.

Warnungen vor einer AGI – unmittelbar vor der KI-Sicherheitskonferenz in Seoul
Kurz vor der Sicherheitskonferenz in Seoul vom 21./22.5.2024 warnten 25 KI-Koryphäen, unter ihnen die Turing-Award Preisträger von 2018, Yoshua Bengio und Geoffrey Hinton, in der renommierten Fachzeitschrift Science (ganzer Artikel bezahlpflichtig) vor einer unkontrollierten Entwicklung einer AGI. Die KI-Unternehmen investierten enorme Summen in immer leistungsfähigere Tools und viel zu wenig in die nötige Forschung zu Sicherheitsfragen. Es drohe die Gefahr, dass autonome KI-Systeme entwickelt werden, die die von uns vorgegebenen roten Linien überschreiten. Dies könnte versehentlich wegen fehlerhafter Programmierung erfolgen oder auch, weil sich die KI-Systeme selbstlernend über menschliche Vorgaben, auch Abschaltmechanismen, hinwegsetzen. Die Folgen wären irreversibel. Die Autoren sprechen davon, dass sie uns Menschen marginalisieren oder gar die Menschheit auslöschen könnten. Das Autorenteam verlangt proaktive, international vereinbarte Sicherheitsregelungen, die automatisch beim Erreichen bestimmter Etappen in der künftigen KI-Entwicklung in Kraft treten.

Denkpause? – Viele kritische Stimmen
Die Warnung, eine übermenschliche KI könnte zur Auslöschung der Menschheit führen, ist nicht neu. Bereits im März 2023 wurde in einem offenen Brief (hier) mit damals über 26'000 Unterschriften (heute sind es über 33'000) eine sechsmonatige Denkpause beim Training neuer KI-Modelle gefordert. Als Reaktion gab es viele kritische Stimmen, national (Prof. Dr. Andreas Krause und Dr. Alexander Ilic vom ETH AI Center, hier) und international (DAIR Institute, eine gemeinnützige US-Organisation, hier). Hier einige Kritikpunkte. Wäre ein Moratorium überhaupt kontrollier- und durchsetzbar? Wollen einige heimlich weiterforschen, um den Rückstand aufzuholen? Soll mit einer apokalyptischen Story vom bereits aktuellen Gefahrenpotenzial abgelenkt werden? Zielgerichteter sei es, von den Tech-Giganten Transparenz zu fordern und sie zur Rechenschaft zu verpflichten.

Richtlinien für KI-Sicherheitstests
Seit der Publikation des offenen Briefes ging der KI-Wettlauf beschleunigt weiter. Unter den Erstunterzeichnern – neben prominenten KI-Forschungskapazitäten – fand sich auch Elon Musk, der inzwischen den grössten KI-Supercomputer (hier) für sein am 9.3.2023 gegründetes Startup xAI bauen will. Immerhin scheinen sich die Tech-Giganten der Sicherheitsrisiken bei der Entwicklung immer leistungsfähigerer KI-Tools bewusst zu sein. Am AI-Summit vom 21./22.5.2024 in Seoul haben 16 KI-Unternehmen ein Dokument unterzeichnet, wonach Richtlinien für KI-Sicherheitstests mit Risikoschwellen erarbeitet werden sollen, hier.

Eine autonom agierende KI, die sich gegen die Menschen richtet?
Ist es wirklich denkbar, dass sich eine AGI, wie sie die Tech-Giganten im Wettlauf entwickeln wollen, nicht mehr kontrollieren liesse und in ihrer destruktiven Kraft nicht mehr zu stoppen wäre? Auch Jürgen Schmidhuber, Direktor am IDSIA (KI-Forschungsinstitut im Tessin), als «Vater fortgeschrittener KI» bezeichnet (in der New York Times, hier), ist überzeugt, dass mit der AGI Akteure entstehen, die uns Menschen hoch überlegen sind. Er sieht jedoch keine existenzielle Gefahr für die Menschheit, denn es gebe kein Motiv, uns nur deshalb auszurotten, weil wir dümmer sind, hier. Wirklich gefahrlos? Hierzu gibt es einige beunruhigende Informationen im nächsten E-Mail morgen Freitag.

Unbestritten – das Gefahrenpotenzial in den Händen destruktiver Menschen
In dieser Frage besteht Einigkeit. Der italienische Dichter Italo Svevo thematisierte das Gefahrenpotenzial technischer Erfindungen bereits in seinem 1923 publizierten Roman «La coscienza di Zeno». Einer stiehlt und zündet den Explosionsstoff eines genialen Erfinders und sprengt die Welt in die Luft («…e la terra ritornata alla forma di nebulosa errerà nei cieli priva di parassiti e di malattie»).

Mit freundlichen Grüssen
Für das Netzwerk der ehemaligen SSAB: Hanna Muralt Müller

Neues Datenschutzrecht: Falls Sie diese E-Mail nicht mehr erhalten möchten, melden Sie sich bitte bei mir!