Liebe Mitglieder des Netzwerkes der ehemaligen SSAB, liebe Interessierte

Ist es möglich, dass autonome KI-Systeme unprogrammiert, selbstlernend unerwünschte Effekte hervorbringen und Vorgaben – rote Linien – nicht einhalten (Alignment-Problem)? Es gibt dazu einige beunruhigende Informationen. Besonders alarmierend ist jedoch, dass es gewisse Dilemmata gibt, die die wichtigsten Akteure in ihrem Handlungsspielraum einschränken. Es ist höchste Zeit, darüber – nicht nur in der Bildung – zu diskutieren.

Gefährliche KI-Systeme – obwohl mit ethischen Vorgaben trainiert?
Forschende am MIT (Massachusetts Institute of Technology) haben das Game Cicero von Meta untersucht und festgestellt, dass die KI, obwohl trainiert, ehrlich zu sein und niemals Menschen absichtlich zu hintergehen, sich selbst Tricks und Täuschungen beibrachte, um das im Algorithmus festgelegte Ziel zu erreichen, nämlich zu gewinnen. Der Courthouse News Service, ein amerikanischer Nachrichtendienst, berichtete hierüber Anfang Mai 2024, hier.

Wehe, wenn sie selbständig agierten…
Wenig vertrauenserweckend sind die Resultate einer wissenschaftlichen Studie der Stanford University zu den Vorschlägen verschiedener KI-Chatbots in simulierten Kriegssituationen, hier. Alle Modelle – untersucht wurden u.a. GPT-3.5, GPT-4 von OpenAI, Claude 2 von Anthropic und Llama 2 von Meta – zeigten schwer voraussehbare Eskalationsmuster. Einige empfahlen einen nuklearen Angriff, obwohl die Tools auf mögliche Konsequenzen in der realen Welt aufmerksam gemacht wurden. Dieses Phänomen ist in der Alignment-Forschung als «power-seeking» zur Zielerreichung bekannt. Was, wenn diese Tools nicht zur Strategieschöpfung von uns Menschen dienten, sondern als autonome KI-Systeme agierten?

Das Dilemma von Tech-Giganten – die sich gegenseitig misstrauen?
Es ist erfreulich, dass 16 KI-Giganten am AI-Summit vom 21./22.5.2024 in Seoul beschlossen haben, Richtlinien für KI-Sicherheitstests mit Risikoschwellen zu erarbeiten, hier. Der AI-Summit schliesst an jenen in London an, der zur Bletchley-Erklärung führte. Wir warten auf Wirkungen – noch vor dem nächsten Summit, der in Frankreich stattfinden soll. Könnte es sein, dass die Tech-Giganten zwar wissen, dass es Sicherheitsrichtlinien braucht, aber keiner die eigene Innovationskraft stärker als andere bremsen will? Und wer kontrolliert die Einhaltung der Regeln? Wer sich zuerst bewegt, sich zu strikt an die Richtlinien hält, hat im Innovationswettlauf schon verloren. Auf die Selbstkontrolle der Tech-Giganten ist wohl kaum Verlass. Bremsend könnte eine Haftpflicht oder ein Bussensystem wirken.

Das Dilemma staatlicher Organisationen – die geopolitische Dimension
Könnten staatliche Organisationen den international agierenden Tech-Giganten Paroli bieten und wollen sie das? Der KI-Wettlauf der Tech-Giganten hat auch eine geopolitische Dimension. Seit längerem setzen die USA mit den Tech-Sanktionen gegen China alles daran, ihren Vorsprung z.B. in der für die KI-Entwicklung wichtigen Chip-Produktion zu sichern. Peking reagierte mit rekordhohen Investitionen. Eine zu frühe staatliche Regulierung könnte die eigene KI-Entwicklung gegenüber der Konkurrenz benachteiligen. Das Innovationspotenzial von KI-Systemen ist gewaltig und für Volkswirtschaften zukunftsbestimmend. Die USA wollen dieses nicht abbremsen. Welche Rolle könnte die EU einnehmen, die in der KI-Entwicklung nachhinkt, aber als grosser Absatzmarkt Druckpotenzial hat? Der geopolitische Wettkampf zwischen den USA und China ist auch ein Wettrennen zwischen Demokratien und Autokratien. Deshalb wird auch die EU nicht vorpreschen.

Ein Wettrennen mit der Zeit
Sinnvoll wäre eine wissenschaftlich-technische Behörde innerhalb der UNO mit autonomem Status, die – ähnlich der Internationalen Atomenergie-Organisation – zuständig wäre, die Entwicklung immer potenterer KI-Systeme zu überwachen und befugt wäre, proaktiv beim Erreichen bestimmter Etappen international vereinbarte Sicherheitsregelungen durchzusetzen. Denn bereits mit der Entwicklung einer AGI, noch bevor diese auf dem Markt lanciert wird, könnten gefährliche irreversible Prozesse in Gang kommen. Der Wettlauf der Tech-Giganten beschleunigt die Entwicklung gefährlicher KI-Systeme. Die politischen Prozesse, insbesondere in Demokratien und bei der Erarbeitung internationaler Vereinbarungen, brauchen Zeit.

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