Titel: Über die KI-Superintelligenz (1/2)

Untertitel: Nur noch ein kleiner Sprung bis zur Verselbstständigung der KI?

Einleitung (ohne Titel):
Seit einigen Monaten ist die künstliche Intelligenz (KI) überall. Am Morgen in den Schlagzeilen der Medien, am Nachmittag hilft sie mir mit ChatGPT eine E-Mail zu verfassen und am Abend schlägt sie mir auf Netflix einen Film vor. Man könnte fast meinen, dass sie seit der Einführung von ChatGPT im November 2022 omnipräsent sei. Doch der Schein trügt: Die KI mischt bereits seit 20 Jahren im Hintergrund überall mit – es wird uns nur erst jetzt so langsam richtig bewusst. Je besser die KI-Modelle werden, desto einfacher ist es für die meisten von uns, diese direkt anzuwenden. Gleichzeitig bemerken wir damit auch, was sie alles kann und wie fest sie eigentlich bereits in unseren Leben angekommen ist.

Wie KI unsere Welt verändert
Die Medizin verwendet KI schon lange, um Krankheiten zu detektieren. Zu Beginn des 21. Jahrhunderts startete Amazon mit Buchempfehlungen durch künstliche Intelligenz. Das Verfahren war so erfolgreich, dass es von vielen Konkurrenten kopiert wurde. Die Suchmaschine Google generiert die Antworten auf unsere Suchanfragen bereits seit 2015 mit einem Sprachmodell namens RankBrain. Die Suchfunktion wurde kontinuierlich verbessert mit Modellen wie Neural Matching, BERT und MUM. Varianten dieser Modelle (insbes. BERT) sind immer noch rege am Werk, wenn wir unsere Fragen an Google stellen. Tesla fährt mittlerweile routinemässig mit KI und erkennt die Strassen sowie Objekte in der Nähe. Tinder gibt es bereits seit 2012, wobei die KI von Anfang an Teil des Geschäftsmodells war. Bei den Konkurrenten Bumble und Co. sieht dies nicht anders aus. YouTube, Netflix, Spotify, Alexa, Siri, Grammarly – sie alle sind im Grunde genommen KI-Systeme. Die Technologie wird in der Aviatik, der Robotik, im Retail, auf den Aktienmärkten, im Kundensupport, in der Produktion und Planung, im Supply-Chain-Management, in der Sportanalyse, in der Pflege oder teils auch bei der Polizei und im Justizwesen eingesetzt. In der Tat wäre unsere Welt ohne die KI eine andere.

Besonders brisant ist allerdings, dass die Entwicklungen in der künstlichen Intelligenz rasant zunehmen und sich die Technologie exponentiell verbessert. Im besten Fall kann uns dies in

die Ära einer neuen glorreichen Zivilisation führen, im schlechtesten Fall kann uns die Technologie ungewollt alle zerstören.

Im Ernst?

Gute KI, böse KI

Wir müssen zwei Szenarien voneinander unterscheiden: Wenn der Mensch das Problem ist und wenn die KI zum Problem wird. Die Gefahren im ersten Szenario sind nicht schwer abzuschätzen. Menschen können jede Technologie für gute und schlechte Absichten benutzen. KI-Systeme werden heute bereits für militärische Zwecke eingesetzt, für Disinformationskampagnen und für Hacking-Angriffe. *Cambridge Analytica* musste 2018 Insolvenz anmelden, weil bekannt wurde, dass sie mit der Hilfe von KI auf den sozialen Medien gezielt versucht haben, die Meinung der Nutzer:innen für politische Zwecke zu manipulieren. (Es wird spekuliert, dass die Firma nun unter anderem Namen weiterarbeitet.) Auf dem DarkNet kann sich heute bereits jede Person eine KI namens Worm-GPT runterladen, womit man ohne Weiteres eine Reihe von kriminellen Phishing-Attacken ausführen kann. Seit Beginn des Jahres haben sich die Cyber-Attacken aus diesen Gründen verständlicherweise vervielfacht – alle Leser:innen sollten sich daher bewusst sein, wie wichtig es ist, die eigenen Daten zu schützen und auf keine unbekannten Links zu klicken oder auf unbekannte E-Mails zu reagieren.

In diesen Beispielen ist aber der Mensch der problematische Akteur und nicht die Technologie. Es geht aber auch anders: Die KI selbst hat ihre inhärenten Probleme und kann sich auf gefährliche Weise verselbstständigen. Zu den klassischen Problemen einer KI gehören:

- ***KI-Halluzination:*** Die Maschine erfindet Dinge, die nicht stimmen (und wir merken es nicht).
- ***KI-Alignierung:*** Die Maschine macht nicht, was wir eigentlich wollen (und wir merken es nicht).
- ***KI-Runaway:*** Die Maschine macht sich unabhängig von unserem Auftrag und macht sich selbstständig (und wir merken es nicht).
- ***KI-Diskriminierung:*** Die Maschine behandelt nicht alle gleich und fair (und wir merken es nicht).
- ***KI-Lock In Effekt:*** Das Modell bleibt in einem Narrativ stecken und kommt nicht wieder raus.

Die Europäische Union versucht mit ihrem «EU Artificial Intelligence Act» die KI-Entwicklung etwas zu kontrollieren. Sie unterscheidet dabei zwischen jenen Anwendungen, die verboten

gehören (z.B. Gesichtserkennung im Social-Scoring wie in China), jenen die ein hohes soziales Risiko tragen (z.B. die Suchtgefahr und selektive Meinungsprägung von TikTok) und Anwendungen, die als unbedenklich eingestuft werden (z.B. BERT in der Google-Suchmaschine). Die EU kann allerdings nicht regulieren, dass sich eine KI ungewollt Ziele setzt, die negative Konsequenzen für uns Menschen nach sich ziehen. Hier braucht es ganz andere Interventionen.

Ab wann ist die KI «zu intelligent»?

Der Begriff der künstlichen Intelligenz wurde in den 1950er Jahren von John McCarthy geprägt. Der Mathematiker Alan Turing hat in derselben Zeit den Turing-Test vorgeschlagen, der besagt, dass eine Maschine «intelligent» ist, wenn wir im Gespräch nicht mehr unterscheiden können, ob es sich bei dem Gegenüber um einen Menschen handelt oder nicht. Heute gilt dieser Test als veraltet, da die Menschen sehr viele Modalitäten haben, wo sich unsere kognitiven Fähigkeiten auf unterschiedliche Weise zeigen. Moderne Sprachmodelle werden daher diversen «HumanEval»-Tests ausgesetzt, die eine Reihe von unterschiedlichen menschlichen Fähigkeiten untersuchen. Es herrscht allerdings unter Experten und Expertinnen nach wie vor keinen Konsens darüber, ob Computer mithilfe von KI-Modellen wirklich «intelligent» sind oder nicht. Laut Noam Chomsky, einer der einflussreichsten Denker der zweiten Hälfte des 20. und ersten Hälfte des 21. Jahrhunderts, haben wir es bei diesem Begriff mit einer menschlichen Projektion zu tun und somit handelt es sich hier um einen Kategorienfehler. Chomsky interessiert sich dafür, was Intelligenz wirklich ist und findet, die Maschinen ‘erscheinen’ uns lediglich intelligent, weil man mit ihnen versucht, unsere kognitiven Fähigkeiten nachzuahmen. Die meisten Leute sind da allerdings etwas pragmatischer und finden, dass man eine KI ruhig als intelligent bezeichnen kann, wenn sie uns als solches erscheint. Ganz unabhängig davon, ob dies eine «echte» Form von Intelligenz ist oder nicht – man kann sie daher auch einfach «künstliche» Intelligenz nennen.

Die Modelle werden kontinuierlich besser und das klar deklarierte Ziel der Macher:innen ist es, die KI hin zu einer AGI zu entwickelt. AGI steht für «Artificial General Intelligence» und betitelt ein Modell, das in allen wichtigen Belangen den Fähigkeiten eines Menschen gleichkommt. Elon Musk meint hier aber, ganz im Sinne von Chomsky, dass man wohl erst von einer AGI sprechen könne, wenn sie etwas «ganz neues» entdeckt. Die Menschen haben in ihrem kurzen Dasein erstaunlich viel entdeckt und erschaffen: Wir haben viele Naturgesetze entschlüsselt, die Quantenphysik entdeckt, die Atome aufgebrochen, Autos erfunden, Computer erschaffen und befinden uns mit U-Booten sowohl unter Wasser wie auch mit Flugzeugen im Himmel. Erst wenn die Maschinen uns in der Fähigkeit nahekommen, neue

Errungenschaften zu propagieren, können wir von einer echten AGI sprechen. Ob dabei eine AGI auch Bewusstsein entwickeln könnte, ist eine Frage für einen anderen Tag.

Aber was machen wir denn, wenn die Maschine unsere Fähigkeiten nicht nur approximiert, sondern diese sogar übertrifft? Dieser Moment wird als «Singularität» bezeichnet und löst bei vielen Kommentatoren beachtliche Ängste aus. Eine uns überlegene Maschine nennen wir heute «superintelligent». Irving John Good nannte dies in den 1960ern noch «ultaintelligent», aber das klingt vermutlich etwas zu wenig filmreif. Für die meisten Wissenschaftler:innen ist es gemessen an der momentan rasanten Entwicklung nur noch eine Frage der Zeit, bis es so weit ist. Dies wirft die unausweichliche Frage auf: Was machen wir dann? Wir können schliesslich eine Intelligenz, die intelligenter als die unsrige ist, nicht mehr wirklich kontrollieren. Wenn überhaupt, dann kontrollieren nicht wir eine solche Übermacht, sondern sie kontrolliert uns.

Explosive Intelligenz

Forscher:innen arbeiten fleissig daran, eine Maschine zu bauen, die sich selbst verbessern kann. Man spricht hier von einer sogenannten «rekursiven KI», weil sie eben rekursiv auf sich zurückschauen kann, um an sich selbst Schwachstellen zu entdecken und sich so zu verbessern. Im Prinzip handelt es sich bei der Rekursivität um eine Art Metakognition, wo man durch erweitertes Lernen seine eigene Lernfähigkeit verbessern kann. Sobald dieser Zustand in einer Maschine gegeben ist, ist es in den Köpfen vieler nur noch ein kleiner Sprung zu der Verselbstständigung der KI. Hier spricht man von einer «Intelligenz-Explosion». Wenn also das Modell sich konstant selbst verbessern und selbstständig werden kann, dann müsste doch die Singularität nur noch eine Frage der Zeit sein, nicht? In diesem Fall wären wir dann dem Goodwill einer solchen KI ausgeliefert und können nur noch hoffen, dass sie keine bösen Absichten hat. So zumindest sehen es etliche einflussreiche Personen aus der Tech-Szene, von Elon Musk über Bill Gates bis hin zu Sam Altman. Aus diesem Grund arbeiten diese daran, mittels einer ausgeklügelten KI-Alignierung den Modellen entsprechende Ziele, Grenzen und Sicherheitsarchitekturen einzubauen, damit die KI bzw. die AGI (oder besser gesagt die ASI, «Artificial Super Intelligence») uns nicht plötzlich zuwiderläuft.

Über den Stand der Entwicklungen im Bereich KI-Alignment lesen Sie demnächst im zweiten Teil dieses Blogbeitrags.

Quellen und weiterführende Informationen

Boggust, A., Hoover, B., Satyanarayan, A., & Strobelt, H. (2022). Shared Interest: Measuring Human-AI Alignment to Identify Recurring Patterns in Model Behavior. *CHI Conference on Human Factors in Computing Systems*, 1–17. https://doi.org/10.1145/3491102.3501965

Bostrom, N. (2002). Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology*, *9*. https://ora.ox.ac.uk/objects/uuid:827452c3-fcba-41b8-86b0-407293e6617c

Bostrom, N. (2012). The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents. *Minds and Machines*, *22*(2), 71–85. https://doi.org/10.1007/s11023-012-9281-3

Butlin, P. (2021). AI Alignment and Human Reward. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 437–445). Association for Computing Machinery. https://doi.org/10.1145/3461702.3462570

Gabriel, I. (2020). Artificial Intelligence, Values, and Alignment. *Minds and Machines*, *30*(3), 411–437. https://doi.org/10.1007/s11023-020-09539-2

Müller, V. C., & Bostrom, N. (2016). Future Progress in Artificial Intelligence: A Survey of Expert Opinion. In V. C. Müller (Ed.), *Fundamental Issues of Artificial Intelligence* (pp. 555–572). Springer International Publishing. https://doi.org/10.1007/978-3-319-26485-1_33

OpenAI. (2023, July 5). *Introducing Superalignment*. Official Website. https://openai.com/blog/introducing-superalignment