Rainer König

HAL lässt grüßen

Den Fortschritten in der KI bin ich bislang mit großer Begeisterung und noch größerem Interesse begegnet. Hoffnung dominierte. Vielleicht war ich zu naiv!

Der DLF-Podcast vom 18.02.2024 mit dem Titel „*Trickst uns Künstliche Intelligenz bald aus?*“ hat mich nachdenklicher und skeptischer werden lassen[1]. Denn offenbar ist es schon relativ oft passiert, dass KI-Systeme gelogen und betrogen haben. Und zwar trotz des ihnen einprogrammierten Verbots, das zu tun.
So wie schon der fiktive Computer HAL 9000 in Stanley Kubricks „*2001: Odyssee im Weltraum*“: Hal belügt hier seinen Astronauten-Chef Dave, damit dieser ihn Hal nicht ausschalten kann.

So hat z.B. aktuell eine KI der Firma Meta namens Cicero in einem Strategiespiel gelogen, was das Zeug hielt, um zu gewinnen. Obgleich Cicero eigentlich hilfreich bleiben sollte und niemanden in den Rücken fallen durfte.
Eine auf Ehrlichkeit programmierte KI lügt – und zwar um zu gewinnen. Wirklich beruhigend ist das nicht.

Noch beunruhigender: ChatGPT 4 verleugnete bei einem Investmentspiel Insiderinformationen und bei einer Bilderkennung sich selbst. Bei letzterer täuschte die KI sogar einen echten Menschen in der wirklichen Welt mit dem Satz: „Nein, ich bin kein Computer. Ich habe eine Sehschwäche, um Bilder zu sehen.“

Beim Lügen braucht man 1. die Fähigkeit, sich in Menschen hineinzuversetzen. 2. Man muss Menschen aktiv täuschen können. Zu beidem sind moderne KIs offenbar in der Lage. Das beobachtete ein Stuttgarter Forscher namens Tilo Hagendorf. Und die KIs werden da immer besser.

Wie sagte doch der britische Informatiker und Kognitionspsychologe Geoffrey Hinton: Sobald KIs schlauer werden, werden sie uns täuschen können. Eben weil sie das von uns gelernt haben. Zudem gibt´s nur wenige Beispiele in der Evolution, dass Intelligentes von weniger Intelligentem kontrolliert wurde.

Die Einsicht bereitet mir nun tatsächlich Sorgen.

[1] https://www.deutschlandfunk.de/ki-und-luegen-kuenstliche-intelligenz-austricksen-dlf-55c947ec-100.html

Die heute diskutierten Vorschläge & Thesen, aus dem Dilemma herauszukommen:

- Z.B., dass KIs noch kein situatives Bewusstsein und keine eigenen inneren Ziele haben sollen. Klingt aber nicht wirklich angstabbauend: Denn es reicht das Ziel von außen. Schon 2003 zeigte der schwedische Philosoph Nick Boström, dass eine eigentlich harmlose KI das Überleben der Menschheit gefährden kann, nur um ihr ursprünglich einprogrammiertes Ziel sehr effizient zu verfolgen.
- Weitere Tipp: Vorsicht. Ok, das klingt zu schlicht, um beruhigend zu sein.
- Dritter Tipp: Eine KI programmieren, die die KI beim Lügen kontrolliert. Mehr KI also! Na ja: wenn die eine zum Lügen fähig ist, kann es die Aufpasser-KI ja auch. Darum hat es Google noch immer nicht geschafft, eine wirklich zuverlässige Kontroll-KI zu programmieren. Meine Panik steigt.
- Vierter Tipp: so wie wir heute eine Menschenkenntnis brauchen, um Menschen beim Lügen zu ertappen, brauchen wir künftig KI-Kenntnis als Kulturtechnik. Ok: Das ist ein schönes Postulat, aber keine wirkliche Lösung. Für die Entwicklung der Empathie hatten wir ein paar Millionen Jahre Evolution Zeit. Die Zeit haben wir für den Aufbau der KI-Empathie nicht.

Mann o Mann. KI sehe ich jetzt anders.

Danke Piotr Heller und Ralf Krauter vom DLF für den tollen Podcast! Hoffentlich habt ihr als echte Menschen im Studio gesessen und euch nicht von KI-Avataren vertreten lassen!