Lebt sie schon?

Was die künstliche Intelligenz (noch) vom menschlichen Geist trennt.

Eckhart Arnold, BAdW, DH-Referat

.dhmuc KI-Hackathon, 21. September 2023

Überblick

1. Ist es schon so weit oder schon zu spät?
2. (Vermeintliche) Merkmale der menschlichen Sonderstellung auf der Prüfstand
    1. Sprachvermögen
    2. Bewusstsein
    3. Willensfreiheit
    4. Intelligenz
    5. Kreativität
    6. Kulturelle Tradition und Evolution
3. KI-Weltherrschaft (vorerst) abgesagt.

Ist es schon so weit?

Google fires software engineer who claims AI chatbot is sentient

Company said Blake Lemoine violated Google policies and that his claims were 'wholly unfounded'

https://commons.wikimedia.org/wiki/File:Th%C3%A9%C3%A2tre_D%E2%80%99op%C3%A9ra_Spatial.jpg
Gemeinfrei, da KI-generierte Bilder kein Urheberrecht genießen

Experts are warning AI could lead to human extinction. Are we taking it seriously enough?

Künstliche Intelligenz

KI-generierte Kunstwerke krempeln den Kunstmarkt um

KI-Systeme kreieren beeindruckende Filme und preisgekrönte Bilder. Der Kunstmarkt könnte vor einem Umbruch stehen.

Die menschliche Sonderstellung auf dem Prüfstand

1. Sprachvermögen

Erfinde ein Gedicht über den Herbst

Der Herbst ist da, die Blätter fallen,
Die Bäume sind kahl, die Winde hallen.
Die Luft ist kühl, der Himmel grau,
Die Vögel fliegen in den Süden, weit und blau.

Google-Plagiats-Gegenprobe:

Keine Ergebnisse für **"Der Herbst ist da, die Blätter fallen, Die Bäume sind kahl, die Winde hallen."** gefunden

=> Das Sprachvermögen unterscheidet Menschen von Tieren, aber nicht mehr von Maschinen!

2. Bewusstsein (Qualia)

Mit „**Qualia**“ ist all das gemeint, was wir durch „**innere Wahrnehmung**“ von unserem Bewusstsein erfassen können: Gedanken, Empfindungen, Gefühle, Wahrnehmungen, Träume

Ein Wesensmerkmal von Q.: Sie lassen sich durch äußere Wahrnehmung nicht erfassen.

- Können wir anderen Menschen Qualia zuschreiben?
  Mglw. Ja, durch „analogische Apprehension“ (Husserl)
- Können wir Tieren Qualia zuschreiben?
  Mglw. Ja, aber wir wissen nicht welche.
  („Wie ist es eine Fledermaus zu sein?“, Nagel)
- Ab wann können oder müssen wir Qualia Maschinen zuschreiben?
  Wenn sie denken und handeln wie Menschen? Wenn sie aus Fleisch und Blut sind?
- Sind Qualia ein relevantes Kriterium für die Mensch-Maschine-Unterscheidung?
  Nicht unbedingt, da ihre Existenz von außen nicht feststellbar ist, und
  sofern alle anderen Bewusstseinsleistungen ohne sie möglich sind (wie Dennet annimmt)

2. Bewusstsein (CTM)

Computational Theory of Mind („Church-Turing“)

Annahme: Alle kognitiven Leistungen des Menschen können (auch) von einer Turing-Maschine erbracht werden.

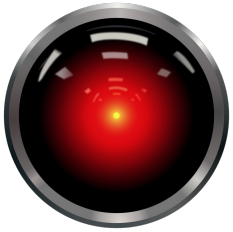
Die Annahme kann bisher weder bestätigt, noch ohne Weiteres widerlegt werden.
(M.E. liegt die Beweislast eher bei ihren Verfechtern.)

Stimmt die Annahme nicht, bleibt menschliche Intelligenz auch von zukünftiger KI unterschieden, sofern sie auf derselben Art von Maschinen entwickelt wird.

Das allein impliziert noch nicht, dass menschliche Intelligenz auch überlegen ist oder bleibt.

(Und nur wenn die Annahme nicht stimmt ist Willensfreiheit möglich.)

3. „Ghost oder Puppet“ – Die Willensfreiheit



Die Vorstellung von Willensfreiheit beinhaltet:

- mindestens einen gewissen Grad an **Nicht-Determinismus**
- Zugleich **mehr als reinen Zufall**
- Umstritten ist, ob ein freier Wille auch ein bewusster Wille sein muss.

Unterstellt man dem Menschen einen freien Willen, so bleibt dies ein Unterschied, den auch eine beliebig verbesserte KI auf Basis von Turing-Maschinen nicht einholen kann.

Es ist unklar, ob Willensfreiheit eine Voraussetzung anderer kognitiver Leistungen ist. Ebenso, ob sie eine Voraussetzung von Willensäußerungen, wie z.B. „Eigeninitiative“ ist.

Bei der ethischen Bewertung könnte die Willensfreiheit ebenso wie Qualia-Bewusstsein jedoch wieder relevant werden: Haben KI-Agenten Rechte? Dürfen oder sollten vielleicht sogar Maschinen gebaut werden, die Menschen in jeder Hinsicht geistig überlegen sind?

4. Intelligenz (Turing-Test)

Annahme: Eine Maschine ist einem Menschen dann intellektuell ebenbürtig, wenn ein Mensch sie allein durch Fragen und Antworten (der Maschine) nicht von einem Menschen unterscheiden kann.

Bereits in den 1960-Jahren wurde der Anspruch erhoben, dass Computer diesen Test bestehen können. (ELIZA, heute als Spielzeug im emacs-Editor mit „ALT-x doctor“)

Der Test ist schon wegen der Vagheit des Kriteriums nicht gut zu entscheiden.

Akzeptiert man, dass heutige Chatbots den Test bestehen, dann spricht das möglicherweise eher gegen den Test (als Kriterium) als für die Intelligenz von Maschinen.

4. Intelligenz – Nicht-Wissen-Wissen und Kausales Schließen

Leicht zu bemerkende Schwächen von großen Sprachmodellen (ChatGPT):

1. Kein Bewusstsein der eigenen Grenzen. („Ich weiß, dass ich nichts weiß“)
   Sie antworten immer, egal was („Hallozinieren“).
2. Schwierig: Warum-Fragen oder „Woher weißt Du das?“
   (Ein Grundproblem algorithmisch generierten Wissens.)
3. Womöglich, kein Modell der Wirklichkeit? keine Fähigkeit zu kausalem Schließen?

Forschungsansatz: Verbindung von KI (insb. Sprachmodellen) und Kausalitätstheorie

Es könnte sein, dass die 3. Lücke bald(?) ausgewetzt wird.

Leiter der Kausalität (J. Pearl, Book of Why)

Assoziation P(Y | X)	Intervention P(Y | X, do(Z))	Kontrafaktische Annahmen P(Y | X , Z‘)
X: Kopfschmerzen Y: Migräne	do(Z): nimm eine Tablette	Z‘: wenn ich eine Tablette genommen hätte
Beobachten	Eingreifen	Vorstellen

5. Kreativität

Kann künstliche Intelligenz kreativ sein?

Ja, schon:

Offene Fragen:

- Kann KI nur bestimmte Arten von Kreativität imitieren, so wie sie bisher nur bestimmte intelligente Leistungen imitieren kann?
- Gibt es prinzipielle Grenzen für irgendwelche (wie zu charakterisierende?) menschliche Arten von Kreativitität.
- Kann sich die KI selbst Themen suchen?
- Warum wirkt KI-Kunst oft so steril? (Oder bilde ich mir das nur ein?)

6. Kulturelle Tradition

Kulturelle Tradition und Evolution bedeutet, dass eine Population von Agenten (KI oder Menschen), ihr Wissen an **künftige Generationen weitergeben** (Tradition) und **um neu dazu Gelerntes ergänzen kann** (Evolution). Kann die KI das?

- "Curse of Recursion" – Das Training von KIs mit KI-generierten Daten führt zum Verfall der Fähigkeiten der KI (https://arxiv.org/pdf/2305.17493.pdf)

Wenn sich dieser Befund (theoretisch wie empirisch) erhärtet, dann ist das nicht nur ein wesentliches Unterscheidungsmerkmal zum menschlichen Geist, sondern es offenbart zugleich Grenzen heutiger KI.

Lebt sie schon?

Ein Resümée

- Auf einigen Feldern bleibt das Ergebnis des Vergleichs uneindeutig, was aber auch daran liegt, dass es sich um schwer fassbare (Qualia, Willensfreiheit) oder schwer operationalisierbare (Kreativität) Phänomene handelt.

- In anderen Bereichen drängt sich der Eindruck auf, dass die KI zwar stellenweise die menschlichen Fähigkeit stark überbietet, sich zugleich aber auch Aussetzer leistet ("Halluzinationen"). Hier verhält sie sich nicht wesentlich anders als andere Technologien, die auch sektoral menschliche Fähigkeiten überbieten und damit ergänzen (dafür bauen wir sie ja), aber eben auch viele Dinge nicht leisten.

- *Die Unfähigkeit als Kollektiv aus eigener Kraft das eigene kognitive Niveau zu halten („Curse of Recursion"), legt die Ansicht nahe, dass die KI noch weit entfernt davon ist, ein (potentiell gefährliches) Eigenleben zu entwickeln.*