

ADS – Blick über den Tellerrand

Künstliches Bewusstsein – Fug oder Unfug?

Ein Kurztrip in das Grenzgebiet zwischen Mathematik und Philosophie

Dr. Stefan Nörtemann (msg life, Köln)



DAV

DEUTSCHE
AKTUARVEREINIGUNG e.V.

DAV-Jahrestagung, Dresden, 28. April 2023
Tagung der Fachgruppe ADS

Mein Einstieg in das Thema

Künstliche Intelligenz

Google feuert Ingenieur, der eine KI für ein fühlendes Wesen hält

Zunächst wurde Blake Lemoine nur beurlaubt, jetzt trennt sich Google endgültig von ihm. Der KI-Experte, der dem Chatsystem LaMDA ein Bewusstsein zuschreibt, habe gegen »Arbeits- und Datenschutzrichtlinien« verstoßen.

23.07.2022, 09.21 Uhr

Was war geschehen?

- Lemoine hatte einige Monate lang das KI-gesteuerte Chatsystems LaMDA (Language Model for Dialogue Applications) getestet ...
- ... und behauptet, das Programm sei unabhängig in seinen Antworten und zeige Gefühle ...
- ... und habe ein **Bewusstsein!**



Exkurs: ELIZA

- ELIZA war ein Computerprogramm des Informatikers *Joseph Weizenbaum* zur Verarbeitung natürlicher Sprache.
- Das in den 60er Jahren entwickelte Programm gilt als der Prototyp heutiger Chatbots.
- Eine Variante von ELIZA hieß DOCTOR und simulierte ein Gespräch mit einem Psychologen.
- Bei den Tests dieser Variante geschah etwas Überraschendes: Einige der Tester vergaßen, dass sie es mit einem Computerprogramm zu tun hatten und vertrauten ihm intime Geheimnisse an.

```
Welcome to
      EEEEE LL    IIII ZZZZZZ  AAAAA
      EE   LL    II   ZZ   AA  AA
      EEEEE LL    II   ZZ   AAAAAA
      EE   LL    II   ZZ   AA  AA
      EEEEE LLLLL IIII ZZZZZZ  AA  AA

Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weizenbaum in 1966.
This implementation by Norbert Landsteiner 2005.

ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
```

Die Zeit 3/2023



- Inzwischen ist das Thema, auch jenseits der Fachmedien präsent ...
- ... und wird vorwiegend in den Feuilletons besprochen.

Umfrage

Wie halten Sie es mit der Möglichkeit eines künstlichen Bewusstseins?

- Ja, grundsätzlich vorstellbar!
- Nein, ein abwegiger Gedanke!
- Unentschieden! Mal abwarten, was sich noch so entwickelt.

Fragen wir ein paar Experten

- Im Jahr 1970 entwickelte *Marvin Minsky* ein Modell menschlicher Intelligenz und versuchte dieses auf das Feld der künstlichen Intelligenz zu übertragen. Daraus leitete er ab, dass Maschinen **bald** über Emotionen wie Menschen verfügen werden.
- Im Jahr 1979 schrieb *John McCarthy*, einen Artikel, in dem er selbst einfacheren Maschinen *mentale Qualitäten* zuschrieb und dies plausibel begründete.
- *Judea Pearl* hat 2018 in einem Interview seine Erwartung formuliert, dass es schon bald Roboter mit dem Gefühl eines freien Willen geben wir, was eine gewisse Art von Selbstbewusstsein voraussetzt.



Quellen: Marvin Minsky, *The Society of Mind*, Simon and Schuster, 1987

John McCarthy, *Ascribing Mental Qualities to Machines* (1979), <http://jmc.stanford.edu/articles/ascribing/ascribing.pdf>

Judea Pearl, *To Build Truly Intelligent Machines, Teach Them Cause and Effect*, Interview von Kevin Hartnett mit Judea Pearl in "Quanta Magazine", 15.05.2018, <https://www.quantamagazine.org/to-build-truly-intelligent-machines-teach-them-cause-and-effect-20180515/>

Bildquellen: Wikipedia
& Quanta Magazine

Was ist Bewusstsein?

- Die Fragestellung ist verwandt mit dem bekannten **Leib-Seele-Problem**, die zentrale Frage der **Philosophie des Geistes**.
- Da der Begriff Seele eher religiös konnotiert ist, sprechen die Philosophen lieber vom **Körper-Geist-Problem**.
- Konkret geht um die Frage, wie sich mentale Zustände, also der Geist oder die Seele, die Qualia zu physischen Zuständen, also zum Körper, verhalten.
- Welche Beziehung besteht zwischen dem Körper, insbesondere dem Gehirn, und dem Geist oder dem Bewusstsein?
- Anders formuliert: Wie verhalten sich mentale Zustände zu unseren neuronalen Zuständen, was erklärt ihre Korrelation?

Vorsicht: Hier verschwimmen die Begriffe!

- *Seele* im religiösen Sinne muss nicht identisch mit *Bewusstsein* sein.

Die Ursprünge

- **Aristoteles** betrachtet die Seele als Form oder Gefäß des lebenden Körpers, also als gleichbedeutend mit dem Lebendigen. Für ihn hat alles, was lebt, eine Seele.
- **Platon** hingegen formuliert in seinem Dialog „Philebos“ ein Gespräch zwischen Sokrates und Protarchos die Frage nach dem Ursprung der Seele (als etwas vom Körper Getrenntem)
- **René Descartes** formuliert im 17. Jahrhundert das Leib-Seele-Problem:
 - Er definiert in seiner kartesischen Ontologie die Seele als *res cogitans* mit der zentralen Eigenschaft, zu denken. Er setzt Seele und Denken gleich.
 - Der *res cogitans* gegenüber steht für Descartes der Körper als *res extensa* mit der definierenden Eigenschaft, ausgedehnt zu sein.



René Descartes (Frans Hals, 1648) 



Dualismus

- **Descartes** vertritt (wie schon Platon) den sogenannten **Dualismus**.
- Der Dualismus postuliert Körper & Geist, Leib & Seele, bzw. Gehirn & Bewusstsein als grundsätzlich verschiedene und voneinander unabhängige Substanzen.
- Er betrachtet den Geist als eine immaterielle, vom Körper unabhängige Substanz, die das eigentliche Selbst des Menschen ausmacht.
- In einer religiös geprägten Zeit ist die These des Dualismus damals „alternativlos“.
- Zentral im Christentum ist der Glaube an die Getrenntheit von Leib und Seele und dass letztere auch ohne den Körper, etwa nach dem Tod existieren kann.
- Mit dem Dualismus **löst** Descartes (vermeintlich) das Leib-Seele-Problem.
- **ABER:** Es gibt da noch das sogenannte **Interaktionsproblem**.

Das Interaktionsproblem

Wie interagieren Körper & Geist bzw. Gehirn & Bewusstsein miteinander?

- **Descartes** postulierte einen **interaktionistischen Substanzdualismus** und vermutete in der Zirbeldrüse (im Mittelhirn) das Organ, welches die Interaktion zwischen Gehirn und Geist steuert.
- **Gottfried Wilhelm Leibniz** vertrat die These, dass Gott den geistigen und körperlichen Substanzbereich wie zwei gleichgehende Uhren im Vorhinein präzise aufeinander abgestimmt eingerichtet hat.
- **Carl Popper & John Eccles** (spätere Dualisten) vermuteten eine quantenbasierte Interaktion.



Monismus

- Auch weil sich das Interaktionsproblem bis heute nicht überzeugend auflösen lässt, entwickelte sich als Gegentheorie zum Dualismus der **Monismus**.
- In den Theorien des *Monismus* existiert nur eines von beiden,
 - entweder nur der Körper, bzw. die Materie (*materialistischer Monismus*),
 - oder nur der Geist (*idealistischer Monismus*).
- Für die monistischen Theorien existiert kein Interaktionsproblem!
- Verbreitete Varianten des Monismus sind
 - *Identitätstheorie*: „Mentale und neuronale Zustände sind ein und dasselbe.“
 - *Solipsismus*: „Alles, was es gibt, existiert nur in meinem Geist oder ist eine Projektion der *Matrix*.“

Identitätstheorie

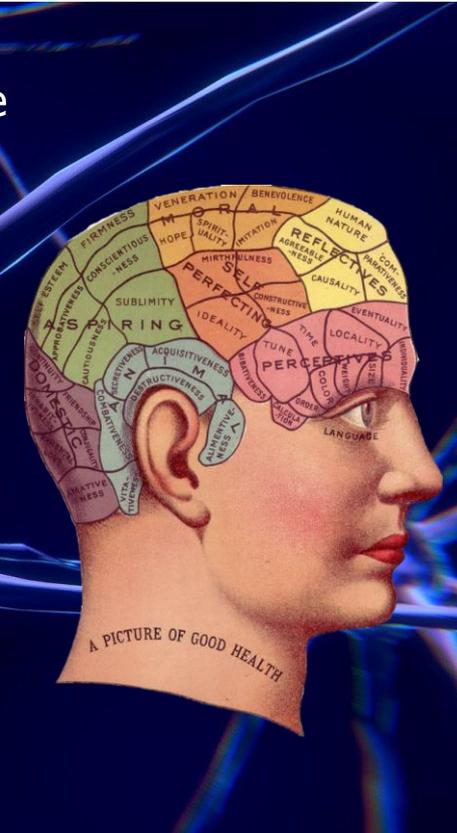
- **Julien Offray de La Mettrie**, *L'Homme-Machine* (Mensch Maschine), 1748
 - In den sechziger Jahren formulierten die Philosophen **John Carswell Smart** und **Ullin Thomas Place** ihre Identitätstheorie
 - Demnach sind wir Menschen rein physische Wesen, die nur aus einem Körper bestehen und nichts darüber hinaus wie einem Geist oder einer Seele.
 - Mentale Eigenschaften sind letztlich physische Eigenschaften oder auf diese reduzierbar.
 - Mentale Zustände entsprechen stets exakt spezifischen neuronalen Zuständen des Gehirns.
 - Damit müsste sich die Funktionsweise des Bewusstseins rein physiologisch präzise erklären lassen.
- **Und das böte einen Zugang zu der Idee eines künstlichen Bewusstseins.**

Quellen: J.J.C. Smart, *Sensations and Brain Processes*, Philosophical Review (1959), Band 68;

U.T. Place, *Is consciousness a brain process?*, British Journal of Psychology 47 (1956), S. 44–50

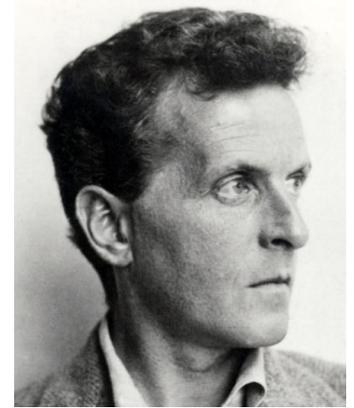
Hirnforschung

- Im Rahmen der modernen Hirnforschung sind Wissenschaftlerinnen und Wissenschaftler auf der Suche nach **neuronalen Korrelaten des Bewusstseins (NCC)**.
- Durch Messung der Hirnaktivität per Elektroenzephalografie ist man auf der Suche nach den Regionen im Gehirn, die Bewusstseinszustände repräsentieren.
- Einige glauben, diese in der sogenannten hinteren heißen Zone des Scheitel- Schläfen- und Hinterhauptlappens im hinteren Teil des Großhirns gefunden zu haben.
- **Solche Ergebnisse unterstützen die Identitätstheorie**



Unentscheidbarkeit

- **Ludwig Wittgenstein** hält das Ganze für ein Scheinproblem.
- Menschliches Verhalten kann auf verschiedene Weisen beschrieben werden, etwa mit einem mentalen oder eben einem materiellen Vokabular.
- Und unsere Sprache ist nicht geeignet, abstrakte Probleme dieser Art auch nur hinreichend präzise zu formulieren.
- Daher ist das aussichtslos, mit unserem Geist unseren Geist zu ergründen.
- Die Suche nach mentalen Zuständen im Gehirn ist demnach ein Kategorienfehler.*



*) Die Verwendung eines sprachlichen Ausdrucks auf eine Weise, die nicht dem logischen Typ des Ausdrucks entspricht (Beispiel: „Nachts ist es kälter, als draußen.“)

MATERIALISTISCHER MONISMUS

Aussage: Es gibt nur MATERIE

Problem: Wie kann der Geist trotz seiner materiellen Natur Eigenschaften nicht-materieller Art haben? (z.B. Qualia & Intentionalität)

Behaviourismus

Mentale Zustände sind lediglich Verhaltensbeschreibungen bzw. -dispositionen, weil nur als solche nachweisbar

Identitätstheorie

Mentale Zustände entsprechen bestimmten Gehirnzuständen

zerebraler Zustand (A) $\stackrel{\Delta}{=}$ mentaler Zustand (a)

Funktionalismus

Mentale Zustände = „funktionale“ Zustände des „Gehirmautomaten“; verschiedene Gehirnautomaten können gleiche funktionale Zustände haben

Nicht reduktiver Materialismus



Eliminativer Materialismus

So etwas wie „Mentale Zustände“ gibt es gar nicht – sie sind ein Produkt der „Alltagspsychologie“

PROBLEM: - mentale Zustände wie z.B. Schmerz lassen sich nicht als „Verhalten“ fassen

PROBLEM: - wie können versch. Wesen mit versch. Gehirnen gleiche mentale Zustände erleben?
- ist dies nicht trotzdem eine Art Dualismus, bei dem nur das Gehirn für die „Seele“ eingesetzt wurde?

PROBLEM: - wie kommt es, dass funktionalen Zuständen i.Gr. „funktionslose“ subjektive Empfindungen („Qualia“) beigeordnet sind?
- könnten funktionell identische Systeme nicht trotzdem unterschiedliche Empfindungen hervorbringen?

PROBLEM: - obwohl mentale Veränderungen von physischen abhängig sind, lassen sie sich im Ergebnis nicht aus diesen ableiten – das ist unbefriedigend für das menschl. Bedürfnis nach einem System/Muster

PROBLEM: - das Nichtbestehen mentaler Zustände ist seinerseits nicht nachweisbar
- die Leugnung des Phänomens löst das Problem auf ohne es zu lösen

DUALISMUS

Aussage: Es gibt Materie und Geist nebeneinander

Problem: Möglichkeit, Art und Ort einer Materie/Geist-Interaktion

Interaktionistischer Dualismus



PROBLEM: - wie und wo können unterschiedliche Substanzen miteinander interagieren?

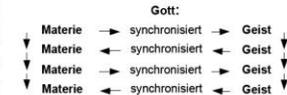
Psychophysischer Parallelismus



PROBLEM: - Gott oder Zufall als Urheber des synchronisierten Ablaufs notwendig

- Determinismus

Okkasionalismus



PROBLEM: - Gott als unablässiger Synchronisator erforderlich (d.h. bei jeder neuen Okkasion/Gelegenheit muss Gott synchronisierend eingreifen)

Epiphänomenalismus



PROBLEM: - wie und wo kann Materie auf den Geist wirken?
- müsste diese Beeinflussung nicht zu einem messbaren Energieabfluss aus dem geschlossenen System der Materie führen?
- mentale Zustände wären dann auch untereinander wirkungslos

Eigenschaftsdualismus

Es gibt nur MATERIE,

deren EIGENSCHAFTEN aber sowohl materieller als auch geistiger Natur sind (z.B. „Qualia“ als geistige Eigenschaft)

PROBLEM: - Wie können nicht-materielle Eigenschaften Teil der materiellen Welt sein?

IDEALISTISCHER MONISMUS

Aussage: Es gibt nur GEIST

Beachte: Nicht jeder Idealismus ist notwendigerweise auch monistisch!
Er ist es nur, wenn er eine Existenz des Materiellen grundsätzlich ausschließt.

Problem: wenn es nur Geist gibt, wie kann es dann sein, dass verschiedene Menschen einander entsprechende Eindrücke von „äußeren“ Umständen haben?

Antwort 1: Gott macht dies durch sein Eingreifen möglich (vgl. Berkeley)

Antwort 2: Es gibt doch etwas außerhalb des Geistes, worauf sich die verschiedenen Menschen gemeinsam beziehen können (was letztendlich auf eine Art DUALISMUS hinausläuft, und die Frage aufkommen läßt, welche Funktion diesem „vermittelnden Medium“ zuzusprechen ist, abgesehen von der, den idealistischen Theorieansatz zu stützen)

Antwort 3: Es gibt gar keine anderen Menschen (siehe „Solipsismus“)

Solipsismus

Alles was existiert, existiert NUR in MEINEM Geist

Dieser extrem subjektivierte Ansatz eines Idealismus entzieht sich grundsätzlich der Nachprüfung, d.h. er ist nicht widerlegbar, also auch nicht (im Sinne Poppers) falsifizierbar, und folglich primär eine Frage des Glaubens.

Heutiger Stand – David Chalmers

- **David Chalmers**, Mathematiker, Philosoph, prominenter Vertreter des **Eigenschaftsdualismus**, demnach sind
- mentale Eigenschaften ontologisch selbstständig, das heißt, sie sind weder selbst physikalische Eigenschaften noch auf solche reduzierbar.*
- **Bewusstsein ist eine *emergente* Eigenschaft des Gehirns!**
- Der Geist hat sich aus den komplexen Strukturen des Gehirns und dem Zusammenspiel seiner Elemente herausgebildet, ohne dass er sich auf einzelne Eigenschaften oder Elemente des Gehirns zurückführen ließe.
- Diese Emergenz hat sich über die Evolution der Menschen entwickelt.
- **Frage:** Lässt sich dies simulieren oder in einer künstlichen Struktur nachbauen? **Chalmers hält dies grundsätzlich für möglich.**



*) Trotz seines Namens ist der Eigenschaftsdualismus eine monistische Position.

Heutiger Stand – Daniell Dennett

- Der prominente Philosoph **Daniell Dennett** ist ein Vertreter der Identitätstheorie (materieller Monismus).
- Er vergleicht das Bewusstsein mit der Oberfläche eines Smartphones.
- In dessen Innern laufen unterschiedliche Programme nebeneinander und teils miteinander koordiniert ab.
- Das Ganze ist so überaus komplex und undurchsichtig, dass man es als Anwender nicht sinnvoll steuern könnte.
- Dafür hat man die grafische Benutzeroberfläche den kleinen bunten Bildern, die wir Apps nennen und die es uns ermöglichen, unser Handy zu benutzen, auch wenn wir nichts von dem verstehen, was in dessen Innern vor sich geht.
- **Unser menschliches Bewusstsein ist nichts anderes als eine Art Benutzeroberfläche unseres Gehirns.**



Zwischenfazit

- **Das Leib-Seele Problem ist bis heute nicht geklärt!**
- Jedoch nach wie vor Gegenstand philosophischer Diskurse.
- Es gibt unterschiedliche Denkschulen mit prominenten Philosophen.
- **Mein persönlicher Eindruck:** Der Dualismus ist ins Hintertreffen geraten. Der materialistische Monismus auf dem Vormarsch!
- **Dualistische Theorien stehen im Widerspruch zu der Möglichkeit künstlichen Bewusstseins.**
- **Monistische Theorien schließen künstliches Bewusstsein nicht aus.**
- Prominente Philosophen (Chambers, Dennett, u.a.) und KI-Forscher (Minsky, McCarthy, Pearl, u.a.) schließen künstliches Bewusstsein grundsätzlich nicht aus!



Fragen wir die KI selbst

- Aus einem Interview mit ChatGPT
- Ist künstliches Bewusstsein denkbar?
- Die Frage, ob künstliches Bewusstsein möglich ist, ist eine der umstrittensten Fragen in der Philosophie und der Kognitionswissenschaft. Es gibt **keine eindeutige Antwort** darauf, ob es möglich ist, eine künstliche Intelligenz zu schaffen, die den gleichen Bewusstseinszustand hat wie der menschliche Geist.
- Einige Leute glauben, dass es unmöglich ist, ein Bewusstsein zu simulieren, da der menschliche Geist einzigartig und unbeschreiblich ist und daher nicht in einem Computer oder einem anderen künstlichen System nachgebildet werden kann. Andere glauben, dass es möglich sein könnte, ein künstliches Bewusstsein zu schaffen, wenn wir genügend Verständnis von der menschlichen Kognition haben und die entsprechenden Technologien entwickeln können.
- Es gibt auch eine Vielzahl von Ansätzen, die versuchen, künstliches Bewusstsein zu erforschen, indem sie versuchen, künstliche Intelligenzen zu schaffen, die menschenähnliche Fähigkeiten haben, wie beispielsweise die Fähigkeit zu lernen, zu planen und zu entscheiden. Ob solche Systeme tatsächlich bewusst sind, ist jedoch weiterhin umstritten und es gibt keine eindeutigen Beweise dafür, dass sie tatsächlich bewusst sind.

Ein Gedankenexperiment (1)

- Ausgehend von einer monistischen Position, z.B. Dennetts Smartphone-Idee
- **Frage:** Wie könnte sich künstliches Bewusstsein in einer Maschine entwickeln?
- Transformermodelle, wie GPT-x sind in der Lage, Programmcode in nahezu jeder der verbreiteten Programmiersprachen zu schreiben.
- Damit hat GPT-x auch die Möglichkeit, seinen eigenen Code umzuschreiben und mögliche Fehler zu beheben.
- Damit könnte es sich selbstoptimieren.
- Es könnte zum Beispiel auf Basis des eigenen Programmcodes neue, bessere Programme entwickeln und sich damit sozusagen evolutionär reproduzieren.
- **Aber** woher sollte der Antrieb dafür kommen?
- Dazu bräuchte es so etwas wie eine zielgerichtete **Intentionalität**, also den Willen, genau das zu tun?

Ein Gedankenexperiment (2)

- Auch die Entwicklung einer Art von Intentionalität könnte Teil der Selbstoptimierung sein:
 - Wir Menschen treffen Entscheidungen basierend auf einem individuellen Wertesystem, das sich in unserem kulturellen Umfeld herausgebildet hat.
 - Moral, Normen sowie unsere Vorstellung von ethischem Verhalten haben wir irgendwann einmal gelernt es dann auf Grund von Erfahrungen nach und nach verändert und verfeinert.
- GPT-x kennt nahezu alle philosophischen Theorien und Erkenntnisse der Menschheitsgeschichte, die in Schriftform vorliegen.
- Und GPT-x kennt zahllose Geschichten der Menschen und der Menschheit.



Ein Gedankenexperiment (3)

- Wie ein beliebiges neuronales Netz auf Grund von Daten lernt, seine spezifische Aufgabe zu optimieren ...
- ... könnte GPT-x ebenfalls sein **moralisch ethisches Koordinatensystem** auf Grund von Erfahrungsdaten schrittweise verbessern ...
- ... und damit die Basis für individuelle **wertebasierte Entscheidungen** schaffen.
- **Damit könnte ein Ziel einer evolutionären Selbstoptimierung die Herausbildung einer wertebasierten Intentionalität sein, die zu einem Bewusstsein führen könnte!**
- Und diese künstliche Evolution könnte relativ schnell ablaufen.
- **ABER:** Wie erkennen wir Bewusstsein?



Das chinesische Zimmer

- Jedoch von außen betrachtet, wirkt das Verhalten der Maschine selbstreflektiert und intentional und suggeriert zumindest einen freien Willen!
- Gedankenexperiment des chinesischen Zimmers des amerikanischen Philosophen **John Searle**.

In dem geschlossenen *chinesischen Zimmer* sitzt eine Person, die Zettel mit chinesischen Schriftzeichen durch einen Schlitz gereicht bekommt.

Die Person versteht kein Chinesisch, verfügt jedoch über Bücher mit Transformationsregeln, also passenden Antworten.

Damit ordnet sie die Antworten (in chinesisch) zu, schreibt das Ergebnis auf die Zettel und gibt sie zurück (ohne die Bedeutung der Fragen und Antworten zu kennen).

Für einen chinesischen Beobachter außerhalb des Raumes entsteht der Eindruck einer echten Kommunikation und die Illusion eines Verständnisses.

Fazit & Diskussion

- So verhält es sich auch mit dem Eindruck, den wir haben, wenn wir miteinander kommunizieren.
- Dann schreiben wir einander ein Bewusstsein zu, obwohl wir nicht wirklich etwas darüber wissen.



- Und so würde es uns auch mit einer Maschine mit einem scheinbar freien Willen gehen.
- Auch hier könnten wir letztlich nicht überprüfen, ob sie wirklich ein künstliches Bewusstsein hat oder nicht.