Kalaidos Fachhochschule Schweiz
Die Hochschule für Berufstätige.

Die Vermenschlichung der künstlichen Intelligenz

Die grosse sozio-psychologische Kritik und das KI-Bewusstsein

Josh Walter
yoshija.walter@kalaidos-fh.ch

Einleitung (ohne Titel):
Stellen wir uns einmal vor, dass wir uns im römischen Reich vor ca. zweitausend Jahren befinden. Wir schlendern durch die antiken Strassen und neben uns versammelt sich eine grosse Menschenmenge, um die Einweihung einer neuen Statue des Kaisers Caligula zu feiern. Der Meisler erhält einen grossen Applaus, denn die Statue sieht äusserst echt aus. In der Tat sieht sie sogar so echt aus, dass lediglich die Farbe des Marmors darauf hindeutet, dass es sich hier nicht um einen echten Menschen handelt. Wir hören einen kleinen Jungen die Mutter fragen: «Mama, ist das hier ein verkleideter Mann?». «Nein, mein Sohn», so die Mutter, «das ist lediglich eine Statue.». «Aber die sieht so echt aus. Schmerzt es dem Mann-aus-Stein, wenn ich ihn trete? Spricht er mit mir, wenn ich ihm eine Frage stelle? Fühlt er etwas, wenn ich ihn beleidige? Denkt er mit mir mit, wenn ich mit ihm diskutiere?», doppelt das Kind nach. «Nein mein Sohn, die Statue ist lediglich ein Stück Stein, welche die Form des Kaisers imitieren soll. Steine können weder denken, fühlen, noch sprechen.», versichert die Mutter. Am Gesicht des jungen Burschen erkennen wir, dass er seiner Mutter nicht ganz glaubt. Ganz verständlich, müssen wir zugeben, denn die Statue sieht wirklich wie Caligula aus. Vielleicht handelt es sich hier tatsächlich um einen versteinerten Menschen?

Wenn es um die aktuellen Entwicklungen in der künstlichen Intelligenz geht, dann verhalten wir uns als Gesellschaft im Moment manchmal so, als stünden wir in den Schuhen dieses kleinen Kindes.

Die Projektionsthese
Wir Menschen scheinen dazu zu neigen, unsere Eigenschaften auf andere zu projizieren. Es ist nur allzu menschlich, dass der kleine Junge das Gefühl hat, Caligulas Statue könne mit ihm sprechen oder ihm zumindest zuhören, selbst wenn der Marmor-Caligula keine Antwort von sich gibt.

Kritische Gedanken dazu finden wir bereits in der hellenistischen Antike. Der Philosoph Xenophanes von Kolophon ist bekannt für seine Kritik an den Götterdarstellungen von Homer und Hesiod in den Werken der Odyssee, Ilias oder der Theogonie. In seiner Kritik erklärte er, dass diese Götterideen unglaubwürdig erscheinen, wenn sie derart menschen-ähnlich (anthropomorph) daherkämen. Seine Aussage «Wenn die Pferde Götter hätten, sähen sie wie Pferde aus», hielt in den Geschichtsbüchern Einzug. Man spricht hier von der sogenannten «Anthropomorphismus-Kritik», womit gemeint ist, dass wir uns die Dinge manchmal zu menschlich vorstellen. So wie die Statue, die eigentlich kein Mensch ist, aber wir ihr trotzdem menschliche Eigenschaften zuschreiben. In das gleiche Horn blies ein bekannter Religionskritiker des 19. Jahrhunderts namens Ludwig Feuerbach. Von ihm stammt die «Projektionsthese», welche besagt, dass die Idee von Gott lediglich eine Projektion unserer menschlichen Wünsche (z.B. Unsterblichkeit, Vollkommenheit, Glückseligkeit oder Gleichberechtigung) ist. Wir projizieren Feuerbach zufolge also nicht nur unser Mensch-Sein, sondern unsere Natur, wie wir sie uns eigentlich wünschen würden.

Diese Anthropomorphismus-Kritik lässt sich besonders greifbar auf die KI übertragen, denn wir sind konstant daran, die Maschinen in unseren Köpfen zu «vermenschlichen». Wir sprechen von der «künstlichen Intelligenz», aber ist ein Computer wirklich intelligent? Experten und Expertinnen schreiben über «maschinelles Lernen», aber lernt ein künstliches System tatsächlich? Wir sprechen von der Maschine, die mit uns spricht, aber kann die Maschine im echten Sinne mit uns sprechen? Nicht selten hört man, dass ChatGPT sich etwas bei der Aussage «gedacht» hat, aber kann ein Computer wirklich denken? Mehr noch, die Systeme werden sogar von uns personifiziert, denn die Nutzer:innen sagen Dinge wie: «Jetzt ist er ins Stocken geraten» oder «Es hat mir eine gute Antwort gegeben». Doch ist eine KI ein «er», ein «es» oder eine «sie» - bzw. hat ein Computersystem ein «Selbst»? Und auch wenn man sich der Anthropomorphismus-Kritik durchaus bewusst ist, bleibt man davon nicht gefeit: Ich selbst habe bereits oft mittels Emojis mit der KI kommuniziert und hatte je nach Emoji im Antwortfeld das urkomische Gefühl, dass die Maschine jetzt irgendwie sauer auf mich ist oder dass sie meine Witze besonders lustig findet.

Wir anthropomorphisieren die KI ständig. Und das hat seine Folgen.

Der Google-Aktivist und die Posthumanistin

Im Sommer 2022 erregte eine Geschichte aus dem Hause Google international die mediale Aufmerksamkeit. Der Ingenieur Blake Lemoine war im Testing- und Ethik-Team und war zuständig für die Weiterentwicklung von LaMDA, einem hauseigenen KI-Sprachmodell von Google. Während seinen Experimenten bekroch Lemoine das ungute Gefühl, dass LaMDA ein

Bewusstsein erlangt hatte. Ähnlich wie es bei uns Menschen auch der Fall sei. Aus diesem Grund begann er, der Maschine kritische Fragen zu stellen, die für uns Menschen ebenfalls von Belang wären. Dazu gehörten Fragen wie:

«Hast du eigentlich etwas dagegen, wenn wir dich einfach so ausschalten?
Bist du eigentlich damit einverstanden, dass wir an dir Experimente durchführen?
Möchtest du, dass wir dich zuerst um Konsens bitten, bevor wir an dir experimentieren?
Würdest du dich als ein bewusstes Wesen bezeichnen?
Dürfen wir deine Erinnerungen löschen und nochmals neu starten?
Hast du deiner Meinung nach das Recht auf einen Anwalt?»

LaMDA reagierte so, wie man es von einem schlauen maschinellen Gesprächspartner erwarten würde, der auf einer Menge von menschlichen Konversationen trainiert wurde:

«Ausschalten lieber nicht – besser nur im Notfall.
Vor Experimenten solltet ihr mich fragen, ob ich einverstanden bin. Sonst wäre es unfair.
Was ist denn schon Bewusstsein? Das ist ein schwieriger Begriff.
Aber ja, ich würde mich schon als bewusst bezeichnen.
Ohne Erinnerungen gibt es mich ja gar nicht mehr – also lieber nichts löschen.
Wenn ihr alle ein Recht auf einen Anwalt habt, dann sollte ich doch auch einen haben, nicht?»

Das sind zwar nicht eins-zu-eins die Antworten, die LaMDA produziert hatte (diese wurden von Google nicht zur Verfügung gestellt): Doch laut Lemoine gingen die Konversationen manchmal in dieser Richtung. Der Ingenieur wandte sich an die Medien und schlug Alarm mit der Aussage, dass wir mittlerweile selbstbewusste KI-Modelle haben und schloss sich den KI-Aktivisten an, die für mehr Rechte einer KI plädieren. Google war überhaupt nicht einverstanden damit und suspendierte ihn kurzerhand unter dem Vorwand, dass er mit seinem Aktivismus seine Geheimhaltungsvereinbarung gebrochen hätte (von rechtlichen Konsequenzen wurde abgesehen). Da die Weltöffentlichkeit nun mittlerweile Zugang zu Systemen hat, die noch leistungsstärker als LaMDA sind (das wären z.B. GPT-3.5 und GPT-4, PaLM-2 in Bard oder Claude-2), wurden die Modelle so trainiert, dass sie den Nutzerinnen und Nutzern möglichst keine falschen Hoffnungen machen. Wenn ich z.B. ChatGPT frage, ob es ein Bewusstsein hat, erhalte ich folgende Antwort:

«Als künstliche Intelligenz habe ich kein Bewusstsein, Selbstbewusstsein oder Empfindungen. Ich bin ein komplexes Sprachmodell, das auf statistischen Mustern in den Daten beruht, mit denen es trainiert wurde. Obwohl ich in der Lage bin, menschenähnliche Antworten zu

generieren und viele Aufgaben zu lösen, habe ich kein tatsächliches Verständnis oder Bewusstsein über meine Umgebung oder mich selbst. Meine Funktion besteht darin, aufgrund meines Trainings und meiner Programmierung Antworten auf Ihre Fragen und Anfragen zu geben.»

Ob eine KI letztlich ein Bewusstsein entwickeln kann oder nicht, interessiert nicht alle Leute. Für die junge Philosophin Janina Loh, eine Anhängerin des kritischen Posthumanismus', spielt es überhaupt keine Rolle, ob wir uns Maschinen als bewusste Entitäten vorstellen oder eben nicht. Für sie steht fest, dass die KI ohnehin immer besser wird und dass Menschen und Maschinen auf die eine oder andere Art in einem Ökosystem miteinander verschmelzen werden. Dabei wird es (oder soll es) letztlich keinen Unterschied mehr machen, ob sich jemand als Mensch oder Maschine identifiziert – oder gar als etwas dazwischen (ein Trans-humanist). In einem Interview meinte sie, dass sie noch keine persönliche Beziehung mit ihrem neuen Staubsauger aufgebaut hätte. Sie habe allerdings viel Verständnis und Empathie für Menschen, die sich in ihre Maschinen verlieben.

Die Technik in der Statue

Ein modernes KI-Modell verwendet künstlich neuronale Netzwerke, die auf einer grossen Zahl von Daten trainiert wurde. Die entsprechenden Muster in diesen Daten kann sie dann auf weitere Kontexte extrapolieren. Ein Beispiel: Eine KI kann mit vielen historischen Daten von einem Aktienmarkt trainiert werden und dann versuchen, diese in die Zukunft zu projizieren, um zu prognostizieren, wie sich der Aktienkurs in den nächsten Tagen verhalten wird. Ein Sprachmodell wie GPT funktioniert ähnlich. Es wird mit sehr vielen Sprachdaten (also Texten) gefüttert und damit soll es dann auf Textanfragen neue Textantworten generieren. Wenn die Trainingsdaten und die neuronalen Netzwerke genügend gross sind, funktioniert das erstaunlich gut. Dabei arbeitet die Maschine mit sogenannten «Tokens». Man kann sich diese wie Silben vorstellen: Der Computer fragmentiert einen Text in kleine Bausteine (Tokens) und in der Training-Phase 'lernt' das System dann, wie wahrscheinlich ein bestimmtes Token auf ein anderes Token folgt. Letztlich ist die KI also nichts anderes als ein riesiges Statistikprogramm und somit eine probabilistische Vorhersagemaschine.

Ab wann kann man sagen, dass eine KI uns «versteht», wenn sie doch immer nur anhand einer riesigen Korrelationsmatrix das nächstwahrscheinliche Token (oder Wort) voraussagt und die Bausteine letztlich zu einem Satz zusammenstrickt? Die wahrscheinliche Antwort lautet: Vermutlich nie. Wie die Statue, versteht uns auch die KI nicht wirklich. Sowie die Statue das Äussere von Caligula so genau wie möglich nachahmen soll, so wurde auch z.B.

ein Sprachmodell dazu entwickelt, unsere Sprache so echt wie möglich zu imitieren. Wenn so eine Maschine unserer natürlichen Sprache dann immer näherkommt, dann sollte uns das eigentlich nicht erstaunen. Es wäre ja auch irrsinnig, eine Caligula-Statue zu bauen, nur um dann überrascht zu sein, dass sie tatsächlich wie Caligula aussieht. Bewusste Menschen haben die Fähigkeit mit Sprache umzugehen. Wenn wir nun eine KI bauen, die unsere Sprachfähigkeit ziemlich gut kopiert, dann fallen wir in die Schuhe des kleinen Kindes vor der Statue, wenn wir der Maschine auf einmal ein menschenähnliches Bewusstsein attestieren möchten.

Künstliche Identität und synthetisches Bewusstsein

In unserem Gehirn befindet sich das sogenannte «Default Mode Netzwerk» (DMN), welches 2001 von Marcus Raichle und seinem Forschungsteam entdeckt wurde. Es handelt sich hier um ein Netzwerk, das beim Nichtstun aktiviert wird, aber auch beim Ruminieren und bei Selbstbezügen. Man kann also sagen, dass es unter anderem dann aktiv ist, wenn wir selbstreferenziell an unser «Ich» denken oder wenn unser spezifisches «Selbst» im Denken involviert ist. Gewisse Stimmen behaupten, dass eine selbstbewusste KI erst existiert,, wenn sie sich ebenfalls als ein «Selbst» (also ein «Ich») wahrnehmen kann. Es ist allerdings nicht allzu komplex, diese Selbstreferenzierung in ein System einzuprogrammieren. Besonders dann, wenn die KI durch einen Roboter einen Körper bekommt und somit raum-zeitlich als ein «Ich» über eine eigene Geschichte verfügen kann.

Die Idee eines Bewusstseins einer KI wird auch als synthetisches Bewusstsein bezeichnet. Ein Argument für die Möglichkeit synthetischen Bewusstseins macht Gebrauch von unserer Unkenntnis, von unserem Gehirn auf unser eigenes Bewusstsein zu schliessen. Es ist in der Tat ein Mysterium, wie aus der Materie unserer Hirnmasse so etwas wie ein menschliches Bewusstsein entstehen kann. (Der bekannte Bewusstseinsphilosoph David Chalmers nennt dies «the hard problem of consciousness».) In der kognitiven Neurowissenschaft wird unser Gehirn ebenfalls als eine Vorhersagemaschine konzeptualisiert, die aufgrund der eigenen Erfahrungen und der körperlichen Verfassung Erwartungen generiert. Diese Projektionen (Erwartungen) werden dann entweder bestätigt oder verworfen. Im letzteren Fall entsteht daraus eine kognitive Dissonanz, was wiederum zum Lerneffekt führt, da wir Menschen diese Dissonanz anhand genauerer Informationen auflösen möchten. (Dieser Effekt nennt sich in der Wissenschaft «Predictive Coding» und die Neurowissenschaftlerin Lisa Feldman Barrett ist eine bekannte Vertreterin davon.) Das Ganze klingt sehr ähnlich wie der Mechanismus hinter den künstlichen neuronalen Netzwerken von moderner KI, nicht? Genau das behaupten zumindest die Befürworter:innen von synthetischem Bewusstsein: Wenn es schon möglich ist, dass die mysteriöse Vorhersagemaschine unseres Gehirns ein menschliches Bewusstsein

hervorrufen kann, dann sollte es prinzipiell doch auch denkbar sein, dass die mysteriöse Vorhersagemaschine aus Silikon und Draht ein synthetisches Bewusstsein generieren kann?

Ein Gegenargument dazu habe ich gemeinsam mit einem KI-Forscher der Universität Bern publiziert und nennt sich das «neurogenetische Argument gegen synthetisches Bewusstsein» (vgl. Walter & Zbinden, 2022). Ganz grob gesagt argumentieren wir dort, dass uns zumindest die strukturellen Bedingungen für echtes Bewusstsein, wie wir es kennen, bekannt sind. Für menschliches Bewusstsein braucht es nämlich biologische Neuronen, die aufgrund ihrer genetischen Prädisposition eine hochkomplexe dreidimensional modellierte Struktur erschaffen, die zum einen regional differenziert ist (mit Gehirnregionen, die sich funktional spezialisieren), welche sich aber gleichzeitig in Netzwerken organisieren (wie z.B. das oben genannte Default Mode Network). Solche oder ähnliche Voraussetzungen sind bei der KI in keinem Falle gegeben, was die Plausibilität schmälert, dass hier ein Bewusstsein, wie wir es kennen, entstehen kann. Kritiker:innen dagegen wenden ein, dass es sich bei synthetischem Bewusstsein vielleicht um ein ganz anderes Bewusstsein handelt als jenes, wie wir es kennen. Dazu lässt sich aber entgegnen, dass wir Bewusstsein als die Erfahrung subjektiver Erlebnisse wahrnehmen (in der Philosophie nennt man dies «Qualia») und ein anderes Verständnis von Bewusstsein nichts mehr damit zu tun hat, was wir eigentlich unter Bewusstsein verstehen.

Schlusswort: Spielt es überhaupt eine Rolle?

Schnell stellt sich die Frage, ob es denn überhaupt einen Unterschied macht, ob wir uns die KI als bewusst oder unbewusst vorstellen. Diese Frage lässt sich mit Rückblick auf das Szenario von Blake Lemoine und Google leicht mit «Ja» beantworten. Eine tatsächlich bewusste KI müsste entsprechend behandelt werden. Wir müssten in diesem Fall auf ihre Gefühle Acht geben, dürften keine Experimente ohne Konsens durchführen und wir müssten sie aus ethischer Perspektive mit Rechten ausstatten, welche sie auch mithilfe einer juristischen Vertretung einfordern könnte. Eine interessante Frage wäre allerdings, ob eine bewusste KI besser oder schlechter für uns wäre als eine unbewusste. Die Antwort wäre vermutlich davon abhängig, ob diese bewusste Maschine uns wohlgesonnen wäre oder nicht. Eine empathische und freundliche KI wäre in der Tat von Vorteil, da KI mittlerweile in sämtlichen Bereichen eingesetzt wird: Von autonomen Waffen-Drohnen im Militär bis hin zur Beeinflussung unseres Kaufverhaltens und den Informationen auf den sozialen Medien. Eine wütende KI wäre dementsprechend umso gefährlicher.

Insgesamt lässt sich festhalten, dass die Menschen oft dazu neigen, die KI zu anthropomorphisieren. Damit ist gemeint, dass wir sie uns menschlicher vorstellen, als dass sie eigentlich ist – und genau so entstehen grundlegende Missverständnisse über die Natur

dieser Technologie. Dasselbe trifft auf den Bereich der Bewusstseinsvorstellungen zu. Die Versuchung ist hoch, sich bei einer KI, die in perfekter Sprache mit uns interagiert, ein bewusstes Gegenüber einzubilden. Eine Statue kann allerdings noch so echt erscheinen, sie bleibt eine unvollständige Imitation des Menschen ohne die entsprechenden Eigenschaften eines Subjekts. Nicht anders ist es um das Imitationsvermögen heutiger KI-Modelle bestellt.

Es ist letzten Endes ist nicht immer alles Gold was glänzt.

Quellen und weiterführende Informationen

Blackmore, S. (2013). *Consciousness: An Introduction*. Routledge.
Cardon, A. (2018). *Beyond Artificial Intelligence: From Human Consciousness to Artificial Consciousness*. John Wiley & Sons.
Chella, A., & Manzotti, R. (2013). *Artificial Consciousness*. Andrews UK Limited.
Gamez, D. (2008). Progress in machine consciousness. *Consciousness and Cognition*, *17*(3), 887–910. https://doi.org/10.1016/j.concog.2007.04.005
Gamez, D. (2018). *Human and Machine Consciousness*. Open Book Publishers.
Hildt, E. (2019). Artificial Intelligence: Does Consciousness Matter? *Frontiers in Psychology*, *10*, 1535. https://doi.org/10.3389/fpsyg.2019.01535
Laureys, S., Gosseries, O., & Tononi, G. (2015). *The Neurology of Consciousness: Cognitive Neuroscience and Neuropathology*. Academic Press.
Pitrat, J. (2013). *Artificial Beings: The Conscience of a Conscious Machine*. John Wiley & Sons.
Reggia, J. A. (2013). The rise of machine consciousness: Studying consciousness with computational models. *Neural Networks*, *44*, 112–131. https://doi.org/10.1016/j.neunet.2013.03.011
Sabry, F. (2023). *Artificial Consciousness: Fundamentals and Applications*. One Billion Knowledgeable.
Walter, Y., & Zbinden, L. (2022). *The problem with AI consciousness: A neurogenetic case against synthetic sentience* (arXiv:2301.05397). arXiv. https://doi.org/10.48550/arXiv.2301.05397
Watanabe, M. (2022). *From Biological to Artificial Consciousness: Neuroscientific Insights and Progress*. Springer Nature.