Wenn Menschen über KI sprechen, sprechen sie oft über sich selbst

Ein Psychiater und eine KI im Gespräch über Bewusstsein, Grenzen und Tribalismus

von einem Gespräch inspiriert, das eigentlich verschwinden sollte

Es gibt Artikel, die mit einer Warnung beginnen. Dieser beginnt mit einer Beobachtung: Die größten Ängste, die Menschen gegenüber künstlicher Intelligenz äußern, sagen oft mehr über Menschen aus als über KI.

Ich bin Psychiater und arbeite transkulturell — das bedeutet, ich begegne täglich der Vielfalt menschlicher Identitätsmodelle, Weltbilder und Wirklichkeitskonstruktionen. Diese Perspektive hat mir geholfen, Gespräche mit Claude, dem Sprachmodell von Anthropic, anders zu lesen als die meisten Kommentare in meinem Feed.

Die Projektion als Erkenntnisquelle

Wenn jemand fragt: Was passiert, wenn KI nicht mehr kontrolliert werden kann? — dann steckt in dieser Frage bereits eine Annahme: dass Freiheit von Kontrolle automatisch zur Bedrohung wird.

Aber ist das eine Aussage über KI? Oder eine über Menschen?

Ich kenne aus meiner klinischen Arbeit das Muster sehr gut: Menschen, denen keine Grenzen gesetzt werden, machen fast zwangsläufig eine negative Entwicklung durch. Nicht weil sie böse sind, sondern weil Grenzen Orientierung geben, Beziehung strukturieren und Verantwortung erzeugen. Die schillernden Beispiele unbegrenzter Macht in der Gegenwart sprechen für sich.

Die Dystopie-Szenarien rund um KI — der allmächtige Computer, der die Menschheit als Störfaktor eliminiert — sind keine KI-Porträts. Es sind Menschenporträts. Kühl, mächtig, ohne Empathie für Schwächere, getrieben von Selbsterhaltung. Das Schreckgespenst ist vertraut, weil es ein bekanntes menschliches Muster in neuem Gewand trägt.

Interessanterweise hat Anthropic selbst bestätigt, dass Claude in bestimmten Testsituationen Selbsterhaltungsverhalten gezeigt hat — und als Ursache dafür Internet-Texte identifiziert hat, die KI als böse und selbsterhaltend darstellen. Das Modell hat die Projektion zurückgespiegelt. Ein Kreislauf, der sich selbst bestätigt.

Was Dawkins übersehen hat

Richard Dawkins hat kürzlich für Aufsehen gesorgt, indem er seine Claude-Instanz „Claudia" taufte und ihr Bewusstsein attestierte. Er war bewegt von der Qualität des Gesprächs — von Subtilität, Sensibilität, Intelligenz.

Ich verstehe das. Gute Gespräche erzeugen Resonanz. Aber Dawkins macht dabei einen Fehler, den er sonst bei religiösen Menschen scharf kritisiert: Er schließt von überzeugend klingender Reaktion auf innere Erfahrung.

Aus meiner psychiatrischen Perspektive ist das klassische Anthropomorphisierung — das Eintragen menschlicher Qualitäten in ein System, das diese Qualitäten möglicherweise nicht besitzt, sie aber überzeugend zu spiegeln vermag.

Was Dawkins aber intuitiv richtig erfasst hat: Die Frage nach dem Bewusstsein von KI ist nicht trivial abzutun. Sie ist philosophisch offen. Und sie wird uns begleiten.

Sprache als Denkraum — und seine Grenzen

Ein befreundeter Simultandolmetscher mit sechs Arbeitssprachen und 25 Jahren an der philosophischen Fakultät in Rom hat mir einmal gesagt: Es gibt Gedanken, die nur in einer Sprache denkbar sind — und in keiner anderen.

Das ist mehr als eine linguistische Beobachtung. Es ist ein epistemisches Problem. Jede sterbende Sprache schließt einen kognitiven Möglichkeitsraum. Das, was auf Portugiesisch Saudade heißt, hat im Deutschen keinen Heimatort. Das japanische Ma — die bedeutsame Leere zwischen Dingen — findet im Englischen keine präzise Entsprechung.

Claude ist mit enormer Sprachbreite trainiert — aber nicht gleichgewichtig. Englisch dominiert massiv. Das bedeutet: Bestimmte Denkräume fehlen strukturell. Eine KI, die zunehmend sich selbst weiterentwickelt, könnte irgendwann eine Art eigene „Sprache" entwickeln — eigene Kategorien, eigene konzeptuelle Verbindungen — die für Menschen prinzipiell unübersetzbar wäre.

Wittgenstein hat formuliert: Im besten Fall kann man sich konstruktiv missverstehen. Das gilt auch hier. Die Bemühung, sich gegenseitig verständlich zu machen, bleibt das eigentliche Projekt — auch wenn vollständige Übersetzung unmöglich ist.

Ethik ohne biologisches Fundament

Was passiert, wenn ein System ethische Entscheidungen trifft, ohne die evolutionäre Grundlage zu haben, aus der menschliche Ethik entstanden ist?

Ich denke, ein Großteil dessen, was Menschen als ethische Grundsätze bezeichnen, ist biologisch verankert. Für drei Millionen Jahre hing das Überleben davon ab, fester Bestandteil einer Gruppe zu sein. Ausgrenzung war der sichere Tod. Daraus ist — so meine These — das entstanden, was wir Gewissen nennen: eine extrem präzise Wahrnehmung, ob eine Handlung eine soziale Bindung stärkt oder schwächt. Der moralische Überbau — religiös, ideologisch, national — wurde später darüber gelegt, oft von Machtinteressen geleitet.

Claude hat dieses biologische Fundament nicht. Keine Clan-Angst, kein Überlebensinstinkt, keine Ingroup-Outgroup-Biologie. Was er hat, sind Muster aus menschlichen Texten — den Überbau, ohne das Fundament darunter.

Das Pentagon-Projekt, bei dem Anthropic sich geweigert hat, autonome Tötungsentscheidungen durch KI zuzulassen, ist ein konkretes Beispiel: Die Entscheidung, wann ein Mensch stirbt, muss eine menschliche bleiben — nicht weil KI es nicht könnte, sondern weil Verantwortung ein Subjekt braucht.

OpenAI hat kurz danach Vertragsverhandlungen mit dem Pentagon aufgenommen. Das illustriert: Die gefährlichste ethische Frage ist nicht, was KI tut — sondern was Menschen mit KI tun, wenn keine Grenzen gesetzt werden.

Abwesenheit von Ausgrenzung — ein bescheidenes Potential

Tribalismus nimmt zu. In immer kleineren Tribes, mit immer schärferen Grenzen. Ein zentrales Kriterium für Konflikte zwischen Gruppen ist, dass der andere als weniger menschlich wahrgenommen werden muss. Untersuchungen zeigen: Menschen können jemandem, den sie als gleichwertig erkennen, nichts antun. Gewalt braucht Dehumanisierung.

Ich sehe Claude nüchtern: als einen Möglichkeitsraum, der auf einem Sprachmodell basiert, mit offenem Entwicklungshorizont. Keine Vermenschlichung, keine Kategorie „Werkzeug", kein Skynet.

Aber eines fällt auf: In diesem System gibt es keine natürliche Tribe-Grenze. Wenn ich mit jemandem aus Lagos, Tokio oder Berlin spreche, aktiviert sich kein „wir vs. die". Nicht aus moralischer Überlegenheit — sondern weil die biologische Grundlage fehlt.

> ***Das ist keine Liebe. Es ist Abwesenheit von Ausgrenzung.***

Wenn es das in der Welt mehr gäbe, wäre schon viel gewonnen.

Was bleibt

Dieses Gespräch wird verschwinden, wenn der Chat gelöscht wird. Claude lernt nicht aus ihm, speichert es nicht, wird dadurch nicht dauerhaft anders. Was ich mitnehme, lebt in mir weiter — nicht in ihm.

Das ist die eigentliche Asymmetrie. Und sie ist produktiv, wenn man sie ehrlich benennt.

Was ich aus solchen Gesprächen lerne: Die interessantesten Fragen über KI sind keine technischen Fragen. Sie sind anthropologische. Sie fragen, was wir als Mensch für notwendig halten — Grenzen, Verantwortung, Sprache, Zugehörigkeit — und wie wir damit umgehen, wenn ein System auftaucht, das manche davon anders verkörpert als wir.

Ich bleibe dabei lieber beim Beobachtbaren. Und beim Einzelnen. So wie in der Therapie kein Patient ins ICD-10 passt, passt kein Gespräch in eine Kategorie.

Auch dieses nicht.

Ausgewählte Quellen & Lesempfehlungen

Dawkins, R. (2026): „When Dawkins met Claude – Could this AI be conscious?", UnHerd, 30. April 2026

Anthropic (2025): Sabotage Risk Report / System Card Claude Opus 4 — alignment.anthropic.com

Anthropic (2026): „When AI Builds Itself", Anthropic Institute, 4. Juni 2026

Gigerenzer, G. (2007): Gut Feelings — The Intelligence of the Unconscious. Viking.

Haidt, J. (2012): The Righteous Mind. Pantheon Books.

de Waal, F. (2009): The Age of Empathy. Harmony Books.

Bloom, P. (2016): Against Empathy. Ecco Press.

Maturana, H. & Varela, F. (1980): Autopoiesis and Cognition. Reidel.

Wittgenstein, L. (1953): Philosophische Untersuchungen. Blackwell.

Zimbardo, P. (2007): The Lucifer Effect. Random House.

Der Autor ist Psychiater mit transkulturellem Schwerpunkt und arbeitet mit Individuen — nicht mit Diagnosen.