Kein halbes Jahr nach der Veröffentlichung von *ChatGPT* wurde im März 2023 bereits dessen Nachfolger *GPT-4* veröffentlicht, welcher die Leistungsfähigkeit seines Vorgängers noch einmal deutlich übertrifft. Dies veranlasste einige der einflussreichsten Vordenker auf diesem Gebiet sogar dazu, in einem viel beachteten offenen Brief[1] eine vorübergehende Pause in der weiteren Entwicklung von KI-Systemen, welche noch leistungsfähiger als GPT-4 sind, zu fordern, um einem möglicherweise drohenden Kontrollverlust vorzubeugen.

Künstliche Intelligenz und Hirnforschung

Die erstaunlichen Leistungen von *ChatGPT* und *GPT-4* haben auch direkte Auswirkungen auf unser Verständnis des menschlichen Gehirns und seiner Funktionsweise. Sie fordern daher die Hirnforschung nicht nur heraus, sondern haben sogar das Potenzial, sie zu revolutionieren. In der Tat waren KI und Hirnforschung in ihrer Geschichte schon immer eng miteinander verflochten. Die sogenannte kognitive Revolution Mitte des letzten Jahrhunderts kann auch als Geburtsstunde der Forschung auf dem Gebiet der KI angesehen werden, wo sie sich als integraler Bestandteil der neu entstandenen Forschungsagenda der Kognitionswissenschaften als eigenständige Disziplin entwickelte. Tatsächlich ging es in der KI-Forschung nie nur darum, Systeme zu entwickeln, die uns lästige Arbeit abnehmen. Von Anfang an ging es auch darum, Theorien über natürliche Intelligenz zu entwickeln und zu testen. Wie wir sehen werden, konnten gerade in jüngster Zeit einige erstaunliche Parallelen zwischen KI-Systemen und Gehirnen aufgedeckt werden. KI spielt daher in der Hirnforschung eine immer größere Rolle, und zwar nicht nur als reines Werkzeug zur Analyse von Daten, sondern insbesondere auch als Modell für die Funktion des Gehirns.

Umgekehrt haben auch die Neurowissenschaften in der Geschichte der Künstlichen Intelligenz eine Schlüsselrolle gespielt und die Entwicklung neuer KI-Methoden immer wieder inspiriert. Die Übertragung von Design- und Verarbeitungsprinzipien aus der Biologie auf die Informatik hat das Potential, neue Lösungen für aktuelle Herausforderungen im Bereich der KI bereitzustellen. Auch dabei spielt die Hirnforschung nicht nur die Rolle, mit dem Gehirn ein Vorbild für neue KI-Systeme zur Verfügung zu stellen. Vielmehr wurde in den Neurowissenschaften eine Vielzahl von Methoden

[1] https://futureoflife.org/open-letter/pause-giant-ai-experiments/

zur Entschlüsselung der Repräsentations- und Rechenprinzipien natürlicher Intelligenz entwickelt, die jetzt wiederum als Werkzeug zum Verständnis Künstlicher Intelligenz eingesetzt und damit zur Lösung des sogenannten Black-Box-Problems beitragen können. Ein Unterfangen, welches gelegentlich als Neurowissenschaft 2.0 bezeichnet wird. Es zeichnet sich ab, dass beide Disziplinen in der Zukunft immer mehr miteinander verschmelzen werden (Marblestone et al., 2016; Kriegeskorte & Douglas, 2018; Rahwan et al., 2019; Zador et al., 2023).

Zu blind, um den Elefanten zu sehen

Die Erkenntnis, dass verschiedene Disziplinen zusammenarbeiten müssen, um etwas derart Komplexes wie Kognition auf menschlichem Niveau zu verstehen, ist natürlich nicht neu und wird in der bekannten Metapher von den sechs Blinden und dem Elefanten anschaulich illustriert (Friedenberg et al., 2021):

> Es waren einmal sechs blinde Wissenschaftlerinnen und Wissenschaftler, die noch nie einen Elefanten gesehen hatten und erforschen wollten, was ein Elefant ist und wie er aussieht. Jeder untersuchte einen anderen Körperteil und kam entsprechend zu einer anderen Schlussfolgerung.
>
> Die erste Blinde näherte sich dem Elefanten und berührte seine Seite. „Ah, ein Elefant ist wie eine Wand", sagte sie.
>
> Die zweite Blinde berührte den Stoßzahn des Elefanten und rief: „Nein, ein Elefant ist wie ein Speer!"
>
> Die dritte Blinde berührte den Rüssel des Elefanten und sagte: „Ihr irrt euch beide! Ein Elefant ist wie eine Schlange!"
>
> Der vierte Blinde berührte ein Bein des Elefanten und sagte: „Ihr irrt euch alle. Ein Elefant ist wie ein Baumstamm."
>
> Der fünfte Blinde berührte das Ohr des Elefanten und sagte: „Keiner von euch weiß, wovon ihr redet. Ein Elefant ist wie ein Fächer."
>
> Schließlich näherte sich der sechste Blinde dem Elefanten und berührte seinen Schwanz: „Ihr irrt euch alle", sagte er. „Ein Elefant ist wie ein Seil."

Hätten die sechs Wissenschaftlerinnen und Wissenschaftler ihre Erkenntnisse kombiniert, wären sie der wahren Natur des Elefanten viel näher gekommen. In dieser Geschichte steht der Elefant für den menschlichen Geist, und die sechs Blinden stehen für die verschiedenen wissenschaftlichen Disziplinen, die versuchen, seine Funktionsweise jeweils aus verschiedenen Perspektiven zu ergründen (Abb. 1.1). Die Pointe der

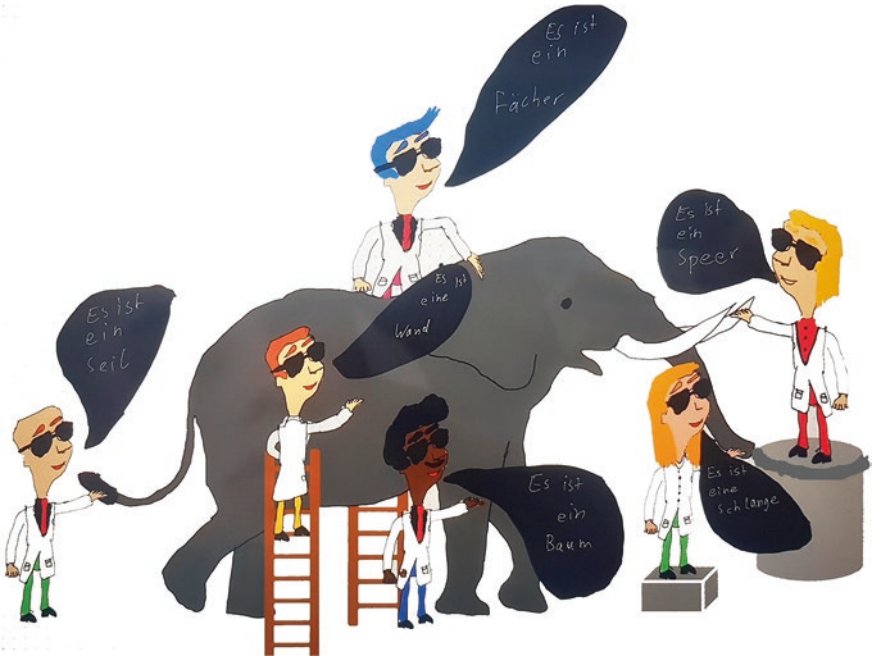

Abb. 1.1 Die Blinden und der Elefant. Jeder untersucht einen anderen Körperteil und kommt entsprechend zu einer anderen Schlussfolgerung. Der Elefant steht für Geist und Gehirn, und die sechs Blinden stehen für verschiedene Wissenschaften. Die Sichtweise jeder einzelnen Disziplin ist wertvoll, ein umfassendes Verständnis kann jedoch nur durch die Zusammenarbeit und den interdisziplinären Austausch erreicht werden

Geschichte ist, dass die Sichtweise jedes Einzelnen zwar wertvoll ist, dass aber ein umfassendes Verständnis von Kognition nur erreicht werden kann, wenn die unterschiedlichen Wissenschaften zusammenarbeiten und sich austauschen.

Dies ist der Gründungsgedanke der Kognitionswissenschaften, die in den 1950er-Jahren als intellektuelle Bewegung begannen, welche als kognitive Revolution bezeichnet wurde (Sperry, 1993; Miller, 2003). In dieser Zeit kam es zu großen Veränderungen in der Arbeitsweise von Psychologen und Linguisten und zur Entstehung neuer Disziplinen wie Informatik und Neurowissenschaften. Die kognitive Revolution wurde durch eine Reihe von Faktoren vorangetrieben, darunter die rasche Entwicklung von Personal Computern und neuen bildgebenden Verfahren für die Hirnforschung. Diese technologischen Fortschritte ermöglichten es den Forschern, besser zu verstehen, wie das Gehirn funktioniert und wie Informationen verarbeitet, gespeichert und abgerufen werden. Als Folge dieser Entwicklungen ent-

stand in den 1960er-Jahren ein interdisziplinäres Gebiet, das Forscher aus den unterschiedlichsten Disziplinen zusammenführte. Dieses Gebiet trug verschiedene Namen, darunter Psychologie der Informationsverarbeitung, Kognitionsforschung und eben auch Kognitionswissenschaft.

Die kognitive Revolution markierte einen wichtigen Wendepunkt in der Geschichte der Psychologie und verwandter Disziplinen. Sie hat die Art und Weise, wie Forscher Fragen der menschlichen Kognition und des menschlichen Verhaltens angehen, grundlegend verändert und den Weg für zahlreiche Durchbrüche in Bereichen wie der Künstlichen Intelligenz, der Kognitiven Psychologie und den Neurowissenschaften geebnet.

Heute versteht man unter Kognitionswissenschaft ein interdisziplinäres wissenschaftliches Unterfangen zur Erforschung der unterschiedlichen Aspekte von Kognition. Dazu gehören Sprache, Wahrnehmung, Gedächtnis, Aufmerksamkeit, logisches Denken, Intelligenz, Verhalten und Emotionen. Hierbei konzentriert man sich vor allem auf die Art und Weise, wie natürliche oder künstliche Systeme Informationen repräsentieren, verarbeiten und umwandeln (Bermúdez, 2014; Friedenberg et al., 2021).

Die Schlüsselfragen sind: Wie funktioniert der menschliche Geist? Wie funktioniert Kognition? Wie ist Kognition im Gehirn implementiert? Und wie kann Kognition in Maschinen umgesetzt werden?

Damit widmen sich die Kognitionswissenschaften einigen der schwierigsten wissenschaftlichen Probleme überhaupt, da das Gehirn unglaublich schwer zu beobachten, zu messen und zu manipulieren ist. Viele Wissenschaftler halten das Gehirn sogar für das komplexeste System im bekannten Universum.

Zu den beteiligten Disziplinen der Kognitionswissenschaften gehören heute Linguistik, Psychologie, Philosophie, Informatik, Künstliche Intelligenz, Neurowissenschaft, Biologie, Anthropologie und Physik (Bermúdez, 2014). Zwischenzeitlich waren die Kognitionswissenschaften etwas aus der Mode gekommen, insbesondere die Idee der integrativen Zusammenarbeit der unterschiedlichen Disziplinen geriet teilweise in Vergessenheit. Speziell KI und Neurowissenschaft entwickelten sich eigenständig weiter und somit auch voneinander weg. Erfreulicherweise erlebt die Idee, dass KI und Hirnforschung komplementär zueinander sind und viel von der jeweils anderen Disziplin profitieren können, derzeit eine regelrechte Renaissance, wobei der Terminus „Kognitionswissenschaft“ anscheinend heute in manchen Communities entweder anders interpretiert wird oder als zu unmodern gilt, weshalb stattdessen Begriffe wie *Cognitive Computational Neuroscience* (Kriegeskorte & Douglas, 2018) oder *NeuroAI* (Zador et al., 2023) vorgeschlagen wurden.

Das Erbe der kognitiven Revolution zeigt sich in den vielen innovativen und interdisziplinären Ansätzen, die unser Verständnis des menschlichen Geistes und seiner Funktionsweise weiterhin prägen. Ob mithilfe modernster bildgebender Verfahren des Gehirns, ausgefeilter Computermodelle oder neuer theoretischer Rahmenkonzepte – die Forscherinnen und Forscher verschieben immer wieder die Grenzen dessen, was wir über das menschliche Gehirn und seine komplexen Prozesse wissen.

Gehirn-Computer-Analogie

Viele Forscher glauben, dass Computermodelle des Geistes uns helfen können zu verstehen, wie das Gehirn Informationen verarbeitet, und dass sie zur Entwicklung intelligenterer Maschinen führen können. Dieser Annahme liegt die Gehirn-Computer-Analogie zugrunde (Von Neumann & Kurzweil, 2012). Man geht davon aus, dass mentale Prozesse wie Wahrnehmung, Gedächtnis und logisches Denken die Manipulation mentaler Repräsentationen beinhalten, die den in Computerprogrammen verwendeten Symbolen und Datenstrukturen entsprechen (Abb. 1.2). Wie ein Computer ist das Gehirn in der Lage, Informationen aufzunehmen, zu speichern, zu verarbeiten und wieder auszugeben.[2]

Diese Analogie bedeutet jedoch nicht, dass das Gehirn tatsächlich ein Computer ist, sondern dass es ähnliche Funktionen erfüllt. Indem man das Gehirn als Computer betrachtet, kann man von biologischen Details abstrahieren und sich auf die Art und Weise konzentrieren, wie es Informationen verarbeitet, um mathematische Modelle für Lernen, Gedächtnis und andere kognitive Funktionen zu entwickeln.

Die Gehirn-Computer-Analogie stützt sich auf zwei zentrale Annahmen, welche den Kognitionswissenschaften zugrunde liegen. Diese sind Computationalismus und Funktionalismus.

[2] Ein fundamentaler Unterschied ist, dass ein Computer Informationen mit anderen Bauteilen verarbeitet als denen, mit denen er die Informationen speichert. Im Gehirn machen beides die – mitunter selben – Neurone.