Philosophie der Künstlichen Intelligenz

Ein strukturierter Überblick

Vincent C. Müller

Abstract: *This paper presents the main topics, arguments, and positions in the philosophy of AI at present (excluding ethics). Apart from the basic concepts of* intelligence *and* computation*, the main topics of artificial cognition are* perception, action, meaning, rational choice, free will, consciousness, *and* normativity. *Through a better understanding of these topics, the philosophy of AI contributes to our understanding of the nature, prospects, and value of AI. Furthermore, these topics can be understood more deeply through the discussion of AI; so we suggest that »AI Philosophy« provides a new method for philosophy.*

Keywords: *AI philosophy; philosophy of AI; cognition; artificial intelligence; meaning*

1. Thema und Methode

1.1 Künstliche Intelligenz

Der Begriff *Künstliche Intelligenz* wurde nach dem »Dartmouth Summer Research Project on Artificial Intelligence« von 1956 populär, dessen Ziele wie folgt formuliert wurden:

> »Die Studie geht von der Vermutung aus, dass jeder Aspekt des Lernens oder jedes andere Merkmal der Intelligenz im Prinzip so genau beschrieben werden kann, dass eine Maschine in der Lage ist, ihn zu simulieren.« (McCarthy et al. 1955: 1)[1]

Dies ist das ehrgeizige Forschungsprogramm, das davon ausgeht, dass menschliche Intelligenz oder Kognition als regelbasierte Berechnung über eine symbolische Repräsentation verstanden oder modelliert werden kann, so dass diese Modelle getes-

1 Alle in diesem Beitrag vorkommenden Übersetzungen zitierter fremdsprachiger Literatur stammen von Matthias Kettner. [Anm. der Hrsg.]

tet werden können, indem sie auf verschiedenen (künstlichen) Computern ausgeführt werden. Im Erfolgsfall würden die Computer, auf denen diese Modelle laufen, künstliche Intelligenz aufweisen. KI und Kognitionswissenschaft sind zwei Seiten derselben Medaille. Dieses Programm wird gewöhnlich als *klassische KI* bezeichnet:[2]

a) KI ist ein Forschungsprogramm zur Entwicklung intelligenter computerbasierter Agenten.

Die von John Searle eingeführten Begriffe *Starke KI* und *Schwache KI* stehen in der gleichen Tradition. *Starke* KI bezieht sich auf die Idee, dass »der entsprechend programmierte Computer wirklich ein Geist ist, in dem Sinne, dass von Computern, die die richtigen Programme erhalten, buchstäblich gesagt werden kann, dass sie verstehen und andere kognitive Zustände haben.« *Schwache* KI bedeutet, dass KI lediglich mentale Zustände simuliert. In diesem schwachen Sinne »besteht der Hauptwert des Computers für die Erforschung des Geistes darin, dass er uns ein sehr mächtiges Werkzeug an die Hand gibt.« (Searle 1980: 353).

Andererseits wird der Begriff »KI« in der Informatik häufig in einem Sinne verwendet, den ich als *technische KI* bezeichnen möchte:

b) KI ist eine Sammlung von Informatikmethoden für Wahrnehmung, Modellierung, Planung und Handlung (Suche, logische Programmierung, probabilistisches Schließen, Expertensysteme, Optimierung, Steuerungstechnik, neuromorphes Engineering, maschinelles Lernen usw.). (Görz et al. 2020; Pearl/Mackenzie 2018; Russell 2019; Russell/Norvig 2020).

Es gibt auch eine Minderheit in der KI, die dafür plädiert, dass sich die Disziplin auf die Ziele von a) konzentriert, während die derzeitige Methodik unter b) beibehalten wird, meist unter dem Namen *Artificial General Intelligence* (AGI).[3]

Das Vorhandensein der beiden Traditionen (klassisch und technisch) führt gelegentlich zu Vorschlägen, dass wir den Begriff »KI« nicht verwenden sollten, weil er starke Behauptungen impliziert, die aus dem Forschungsprogramm a) stammen, aber sehr wenig mit der eigentlichen Arbeit unter b) zu tun haben. Vielleicht sollten wir lieber von »maschinellem Lernen« oder »entscheidungsunterstützenden Maschinen« oder einfach von »Automatisierung« sprechen (wie im Lighthill-Bericht von 1973 vorgeschlagen: Lighthill 1973). Im Folgenden werden wir den Begriff der »Intelligenz« klären, und es wird sich zeigen, dass es ein einigermaßen kohärentes Forschungsprogramm der KI gibt, das die beiden Traditionen vereint: Die *Erzeugung intelligenten Verhaltens durch Rechenmaschinen*.

2 Ein Beispiel dafür: Dietrich 2002. Der klassische historische Überblick ist Boden 2006.

3 Die AGI-Konferenzen werden seit 2008 organisiert.

Diese beiden Traditionen bedürfen nun einer Fußnote: Beide wurden weitgehend unter dem Begriff der *klassischen KI* entwickelt, was hat sich also mit dem Übergang zum *maschinellen Lernen* (ML) geändert? ML ist eine traditionelle (konnektivistische) Rechenmethode in neuronalen Netzen, die keine Repräsentationen verwendet. (Rosenblatt 1957; Buckner (forthcoming); Garson/Buckner 2019; LeCun et al. 2015) Seit ca. 2015, mit dem Aufkommen von massiver Rechenleistung und massiven Daten für tiefe neuronale Netze, hat sich die Leistung von ML-Systemen in Bereichen wie Übersetzung, Textproduktion, Spracherkennung, Spiele, visuelle Erkennung und autonomes Fahren dramatisch verbessert, so dass sie in einigen Fällen dem Menschen überlegen ist. ML ist jetzt die Standardmethode in der KI. Was bedeutet dieser Wandel für die Zukunft der Disziplin? Die ehrliche Antwort lautet: Wir wissen es noch nicht. Wie jede Methode hat auch ML ihre Grenzen, aber diese Grenzen sind weniger restriktiv, als man viele Jahre lang dachte, denn die Systeme zeigen eine nichtlineare Verbesserung – mit mehr Daten können sie sich plötzlich deutlich verbessern. Ihre Schwächen (z. B. Überanpassung, kausale Schlussfolgerungen, Zuverlässigkeit, Relevanz, Blackbox) können denen der menschlichen rationalen Entscheidung recht nahe kommen, insbesondere wenn »prädiktive Verarbeitung« die richtige Theorie des menschlichen Geistes ist (s. unten Abschnitte 4.1, 6).

1.2 »Philosophie der KI« und Philosophie

Eine Möglichkeit, die Philosophie der KI zu verstehen, ist, dass sie sich hauptsächlich mit drei kantischen Fragen beschäftigt: Was ist KI? Was kann KI tun? Was sollte KI sein? Ein wichtiger Teil der Philosophie der KI ist die *Ethik* der KI, aber wir werden diesen Bereich hier nicht diskutieren.[4]

Traditionell befasst sich die Philosophie der KI mit einigen ausgewählten Punkten, bei denen Philosophen etwas über KI zu sagen haben, z. B. zu der These, dass Kognition eigentlich Berechnung ist oder dass Computer sinnvolle Symbole als solche verarbeiten können.[5] Eine Überprüfung dieser Punkte und der entsprechenden Autoren (Turing, Wiener, Dreyfus, Dennett, Searle, u.a.) würde zu einer fragmentierten Diskussion führen, die nie ein Bild des Gesamtprojekts ergibt. Es wäre so, als würde man eine Menschheitsgeschichte im alten Stil anhand einiger weniger »Helden« schreiben. Außerdem wird in dieser Sichtweise die Philosophie der KI von ih-

4 Im vorliegenden Band sind die Texte in der Rubrik »Verantwortungsverhältnisse« dafür einschlägig.

5 Es gibt nur sehr wenige Überblicksartikel und keine aktuellen. Siehe Carter 2007; Copeland 1993; Dietrich 2002; Floridi 2003; Floridi 2011. Einiges von dem, was Philosophen zu sagen hatten, kann als Unterminierung des Projekts der KI angesehen werden, vgl. Dietrich et al. 2021.

rer Cousine, der Philosophie der Kognitionswissenschaft, getrennt, die wiederum eng mit der Philosophie des Geistes verbunden ist (Margolis et al. 2012).

Im Folgenden versuche ich einen anderen Weg: Wir betrachten die *Komponenten eines intelligenten Systems*, wie sie sich in der Philosophie, der Kognitionswissenschaft und der KI darstellen. Eine Möglichkeit, solche Komponenten zu betrachten, ist, dass es relativ einfache Tiere gibt, die relativ einfache Dinge tun können, und dann können wir uns zu komplizierteren Tieren »hocharbeiten«, die außer diesen einfachen Dingen noch mehr tun können. Ein schematisches Beispiel: Eine *Fliege* wird immer wieder gegen das Glas stoßen, um zum Licht zu gelangen; eine *Kobra* wird verstehen, dass sich hier ein Hindernis befindet, und versuchen, es zu umgehen; eine *Katze* wird sich vielleicht daran erinnern, dass sich hier beim letzten Mal ein Hindernis befand, und sofort einen anderen Weg einschlagen; ein *Schimpanse* wird vielleicht erkennen, dass das Glas mit einem Stein zerbrochen werden kann; ein *Mensch* wird vielleicht den Schlüssel finden und die Glastür aufschließen … oder aber das Fenster nehmen, um hinauszukommen. Um sich mit der Philosophie der künstlichen Intelligenz zu befassen, brauchen wir also ein breites Spektrum an Philosophie: Philosophie des Geistes, Erkenntnistheorie, Sprache, Werte, Kultur, Gesellschaft, u.a.m.

Außerdem ist die Philosophie der KI in unserem Ansatz nicht nur »angewandte Philosophie«: Es geht nicht darum, dass wir eine Lösung in der Werkzeugkiste des Philosophen bereithalten und sie »anwenden«, um Probleme der KI zu lösen. Das philosophische Verständnis selbst ändert sich, wenn man den Fall der KI betrachtet: Es wird weniger anthropozentrisch, weniger auf unseren eigenen menschlichen Fall konzentriert. Ein tieferer Blick auf Konzepte muss normativ von der Funktion geleitet werden, die diese Konzepte erfüllen, und diese Funktion kann besser verstanden werden, wenn wir sowohl die natürlichen Fälle als auch den Fall der aktuellen und möglichen KI betrachten. Dieses Papier ist somit auch ein »proof of concept« für die Philosophie durch die begriffliche Analyse von KI: Ich nenne dies KI-Philosophie.

Ich schlage also vor, die Frage vom Kopf auf die Füße zu stellen, wie Marx gesagt hätte: Wenn wir KI verstehen wollen, müssen wir uns selbst verstehen; und wenn wir uns selbst verstehen wollen, müssen wir auch KI verstehen!

2. Intelligenz

2.1 Der Turing-Test

»Ich schlage vor, die Frage ›Können Maschinen denken?‹ zu untersuchen« schrieb Alan Turing zu Beginn seines Aufsatzes in der führenden philosophischen Zeitschrift *Mind* (Turing 1950). Das war 1950, Turing war einer der Gründerväter des Computers, und viele Leser des Aufsatzes werden damals noch nicht einmal von

solchen Maschinen gehört haben, denn es gab nur ein halbes Dutzend Universalcomputer auf der Welt (Z3, Z4, ENIAC, SSEM, Harvard Mark III, Manchester Mark I) (s. Anonym 1950). Turing erklärt kurzerhand, dass die Suche nach einer Definition des Begriffs »Denken« sinnlos sei, und schlägt vor, seine ursprüngliche Frage durch die Frage zu ersetzen, ob eine Maschine erfolgreich ein »Nachahmungsspiel« spielen könne. Dieses Spiel ist unter dem Namen »Turing-Test« bekannt geworden: Ein menschlicher Befrager wird über »Teleprinting« mit einem anderen Menschen und einer Maschine verbunden, und wenn der Befrager die Maschine nicht von dem Menschen unterscheiden kann, indem er ein Gespräch führt, dann sagen wir, dass die Maschine »denkt«. Am Ende des Aufsatzes kommt Turing auf die Frage zurück, ob Maschinen denken können, und sagt: »Ich glaube, dass sich am Ende des Jahrhunderts der Wortgebrauch und die allgemeine gebildete Meinung so sehr verändert haben werden, dass man von denkenden Maschinen sprechen kann, ohne mit Widerspruch zu rechnen.« (Turing 1950: 442) Turing schlägt also vor, unseren alltäglichen Begriff des »Denkens« durch einen operativ definierten Begriff zu ersetzen, einen Begriff, den wir mit einem Verfahren testen können, das ein messbares Ergebnis hat.

Turings Vorschlag, die Definition des Denkens durch eine operative Definition zu ersetzen, die sich ausschließlich auf das Verhalten stützt, passt in das intellektuelle Klima der damaligen Zeit, in der der Behaviorismus noch eine dominierende Kraft war: In der Psychologie ist der Behaviorismus ein *methodologischer* Vorschlag, der besagt, dass die Psychologie zu einer echten wissenschaftlichen Disziplin werden sollte, indem sie sich auf überprüfbare Beobachtungen und Experimente stützt, anstatt auf subjektive Selbstbeobachtung. Angesichts der Tatsache, dass der Geist anderer eine »Black Box« ist, sollte die Psychologie zur Wissenschaft von Reiz und Verhaltensreaktion, von Input-Output-Beziehungen werden. Die frühe analytische Philosophie führte zu einem *reduktionistischen Behaviorismus*. Wenn die Bedeutung eines Begriffs seine »Überprüfungsbedingungen« sind, dann *bedeutet* ein mentaler Begriff wie »Schmerz« lediglich, *dass* die Person zu einem bestimmten Verhalten bereit ist.

Ist der Turing-Test über beobachtbares Verhalten eine nützliche Definition von Intelligenz? Kann er unsere Rede von Intelligenz »ersetzen«? Es ist klar, dass es intelligente Wesen geben wird, die diesen Test nicht bestehen, zum Beispiel Menschen oder Tiere, die nicht tippen können. Man kann also mit Fug und Recht behaupten, dass Turing das Bestehen des Tests nur als hinreichende Voraussetzung für Intelligenz ansah, nicht als notwendige Voraussetzung. Wenn also ein System diesen Test besteht, muss es dann intelligent sein? Das hängt davon ab, ob Sie glauben, dass Intelligenz nur intelligentes Verhalten ist, oder ob Sie glauben, dass wir für die Zuschreibung von Intelligenz auch die interne Struktur betrachten müssen.

2.2 Was ist Intelligenz?

Intuitiv betrachtet ist Intelligenz eine Fähigkeit, die intelligentem Handeln zugrunde liegt. Welches Handeln intelligent ist, hängt von den Zielen ab, die verfolgt werden, und vom Erfolg beim Erreichen dieser Ziele – denken Sie an die oben erwähnten Beispiele tierischen Verhaltens. Der Erfolg hängt nicht nur vom Agenten ab, sondern auch von den Bedingungen, unter denen er agiert, so dass ein System mit weniger Möglichkeiten, ein Ziel zu erreichen (z. B. Nahrung zu finden), weniger intelligent ist. In diesem Sinne lautet eine klassische Definition: »Intelligenz misst die Fähigkeit eines Agenten, Ziele in einem breiten Spektrum von Umgebungen zu erreichen.« (Legg/Hutter 2007: 402) Hier ist Intelligenz die *Fähigkeit, flexibel Ziele zu verfolgen*, wobei Flexibilität mit Hilfe unterschiedlicher Umgebungen erklärt wird. Dieser Intelligenzbegriff aus der KI ist ein *instrumenteller* (bezogen auf Zielerreichung) und normativer Begriff von Intelligenz, in der Tradition der klassischen Entscheidungstheorie, die besagt, dass ein rationaler Agent immer versuchen *sollte*, den erwarteten Nutzen zu maximieren (siehe Abschnitt 6).[6]

Wenn die KI-Philosophie Intelligenz als relativ zu einer Umgebung versteht, dann kann man, um mehr Intelligenz zu erreichen, entweder den Akteur oder die Umgebung verändern. Der Mensch hat beides in großem Umfang durch das getan, was als »Kultur« bekannt ist: Wir haben nicht nur ein ausgeklügeltes Lernsystem für Menschen geschaffen (um den Agenten zu verändern), sondern auch die Welt physisch so gestaltet, dass wir unsere Ziele in ihr verfolgen können; um zu reisen, haben wir z. B. Straßen, Autos mit Lenkrädern, Karten, Straßenschilder, digitale Routenplanung und KI-Systeme geschaffen. Das Gleiche tun wir jetzt für KI-Systeme, sowohl für das lernende System als auch für die Veränderung der Umgebung (Autos mit Computerschnittstellen, GPS usw.). Indem wir die Umwelt verändern, werden wir auch unsere Wahrnehmung und unser Leben verändern – vielleicht auf eine Art und Weise, die sich zu unserem Nachteil auswirkt.

In den Abschnitten 4-9 werden wir uns mit den wichtigsten Komponenten eines intelligenten Systems befassen, doch zuvor werden wir den Mechanismus der KI erörtern: die Berechnung.

6 Z.B. Simon 1955; Thoma 2019. Siehe auch den neo-behavioristischen Vorschlag in Coelho Mollo 2022.

3 Berechnung (»Computation«)

3.1 Der Begriff des Rechnens

Die Maschinen, auf denen KI-Systeme laufen, sind »Computer« oder »Rechner«, so dass es für unsere Aufgabe wichtig sein wird, herauszufinden, was ein Computer ist und was er prinzipiell tun kann. Eine damit zusammenhängende Frage ist, ob die menschliche Intelligenz vollständig oder teilweise auf Berechnungen zurückzuführen ist. Wenn sie vollständig auf Berechnungen zurückzuführen ist, wie die klassische KI angenommen hatte, dann scheint es möglich zu sein, diese Berechnungen auf einem künstlichen Computer nachzubilden.

Um zu verstehen, was ein Computer ist, ist es nützlich, sich die Geschichte der Rechenmaschinen in Erinnerung zu rufen – ich sage »Maschinen«, denn vor ca. 1945 war das Wort »Computer« oder »Rechner« eine Bezeichnung für einen Menschen, der einen bestimmten Beruf hat, für jemanden, der Berechnungen durchführt. Diese Berechnungen, z.B. die Multiplikation zweier großer Zahlen, werden durch ein mechanisches Schritt-für-Schritt-Verfahren durchgeführt, das, wenn es einmal vollständig ausgeführt ist, zu einem Ergebnis führt. Solche Verfahren werden »Algorithmen« genannt. 1936 schlug Alan Turing als Antwort auf Gödels »Entscheidungsproblem« vor, dass der Begriff »etwas berechnen« dadurch erklärt werden könnte, »was eine bestimmte Art von Maschine tun kann« (genau wie er vorschlug, den Begriff der Intelligenz im »Turing-Test« zu operationalisieren). Turing skizzierte, wie eine solche Maschine aussehen würde, mit einem unendlich langen Band als Speicher und einem Kopf, der Symbole von diesem Band lesen und darauf schreiben kann. Diese Zustände auf dem Band sind immer spezifische diskrete Zustände, so dass jeder Zustand von einem Typ aus einer endlichen Liste ist (Symbole, Zahlen, u.a.), also zum Beispiel entweder der Buchstabe »V« oder der Buchstabe »C«, nicht etwa ein bisschen von jedem. Mit anderen Worten, die Maschine ist »digital« (nicht analog).[7] Etwas Entscheidendes kommt hinzu: In der »universellen« Version der Maschine kann man das, was der Computer tut, durch weitere Eingaben *verändern*. Mit anderen Worten: Die Maschine *kann* so *programmiert werden*, dass sie einen bestimmten Algorithmus ausführt, und sie speichert dieses Programm in ihrem Speicher.[8] Ein solcher Computer ist ein Universalcomputer, d.h. er kann jeden beliebigen Algorithmus berechnen. Es sollte erwähnt werden, dass auch weiter gefasste Begriffe der Berechnung vorgeschlagen wurden, z.B. analoges Rechnen und Hypercomputing (Piccinini 2021; Shagrir 2022; Siegelmann 1995; Siegelmann 1997).

7 Negroponte 1995. Siehe auch Haugeland 1985: 57; Müller 2013.

8 Gödel 1931; Turing 1936. Das ursprüngliche Programm ist skizziert in Hilbert 1900. Siehe z.B. Copeland et al. 2013.

Es stellt sich auch die Frage, ob das Rechnen eine reale Eigenschaft physikalischer Systeme ist, oder eher nur eine nützliche Art, diese Systeme zu beschreiben. Searle hat gesagt: »Die elektrischen Zustandsübergänge sind der Maschine immanent, aber die Berechnung liegt im Auge des Betrachters.« (Dodig-Crnkovic/Müller 2011; Searle 2004: 64) Wenn wir eine antirealistische Sichtweise der Berechnung annehmen, dann ändert sich die Situation radikal.

Genau dieselbe Berechnung kann auf verschiedenen physischen Computern durchgeführt werden und eine unterschiedliche Semantik haben. Es gibt also drei Beschreibungsebenen, die für einen bestimmten Computer besonders relevant sind: (a) die *physische Ebene* der tatsächlichen »Realisierung« des Computers, (b) die *syntaktische Ebene* des berechneten Algorithmus und (c) die *symbolische Ebene* des Inhalts, dessen, was berechnet wird.

Physikalisch gesehen kann eine Rechenmaschine aus allem gebaut werden und jede Eigenschaft der physikalischen Welt nutzen (Zahnräder, Relais, DNA, Quantenzustände usw.). Dies kann als Verwendung eines physikalischen Systems zur Kodierung eines formalen Systems angesehen werden (Horsman et al. 2014). Tatsächlich wurden alle Universalcomputer mittels großer Mengen von Schaltern gebaut. Ein Schalter hat zwei Zustände (offen/geschlossen), also arbeiten die darauf basierenden Computer mit zwei Zuständen (ein/aus, 0/1), sie sind *binär* – dies ist eine Designentscheidung. Binäre Schalter können leicht zu »Logikgattern« kombiniert werden, die auf Eingaben in Form der logischen Verknüpfungen in der booleschen Logik (die ebenfalls zweiwertig ist) reagieren: NOT, AND, OR, usw. Wenn sich solche Schalter in einem Zustand befinden, der *syntaktisch* als 1010110 verstanden werden kann, dann könnte dies *semantisch* (nach den derzeitigen ASCII/ANSI-Konventionen) den Buchstaben »V«, die Zahl »86«, einen hellgrauen Farbton, einen grünen Farbton usw. darstellen.

3.2 »Computationalismus«

Wie wir gesehen haben, ist die Vorstellung, dass im Berechnen die Ursache für die Intelligenz natürlicher Systeme, z.B. des Menschen, zu finden ist und zur Modellierung und Reproduktion dieser Intelligenz verwendet werden kann, eine Grundannahme der klassischen KI. Diese Auffassung ist häufig mit der Ansicht gekoppelt (und durch sie motiviert), dass menschliche mentale Zustände funktionale Zustände sind und dass diese funktionalen Zustände die eines Computers sind: »Maschinenfunktionalismus«. Diese These wird in den Kognitions- und Neurowissenschaften oft als Selbstverständlichkeit vorausgesetzt, ist aber in den letzten Jahrzehnten auch erheblich kritisiert worden.[9] Die Hauptquellen für diese Ansicht sind

9 Edelman 2008; Miłkowski 2018. Zur Diskussion: Harnad 1990; Scheutz 2002; Shagrir 1997; Varela et al. 1991.

die Begeisterung für die universelle Technologie des digitalen Rechnens sowie frühe neurowissenschaftliche Befunde, die darauf hindeuten, dass menschliche Neuronen (im Gehirn und im Körper) ebenfalls in gewisser Weise binär sind, d.h. entweder senden sie ein Signal an andere Neuronen, sie »feuern«, oder sie tun es nicht. Einige Autoren verteidigen die *Physikalische Symbolsystemhypothese*, d.h. den Computationalismus, sowie die Behauptung, dass nur Computer intelligent sein können (vgl. Boden 2006: 1419ff.; Newell/Simon 1976: 116).

4. Wahrnehmung und Handlung

4.1 Passive Wahrnehmung

Es mag überraschen, dass die Überschrift dieses Kapitels Wahrnehmung und Handlung verbindet. Aus der KI und der Kognitionswissenschaft können wir aber lernen, dass die Hauptfunktion der Wahrnehmung darin besteht, Handeln zu ermöglichen; ja, dass die Wahrnehmung eine Art von Handeln *ist*. Das traditionelle Verständnis von Wahrnehmung in der Philosophie ist die *passive* Wahrnehmung, bei der wir uns selbst beobachten, wie wir die Welt beobachten, und zwar in dem, was Dan Dennett das *kartesische Theater* genannt hat: Es ist, als ob ein kleiner Mensch in meinem Kopf säße, der die Außenwelt durch unsere Ohren hört und durch unsere Augen beobachtet (Dennett 1991: 107). Diese Vorstellung ist letztlich absurd, vor allem weil sie voraussetzen würde, dass noch ein weiterer kleiner Mensch im Kopf dieses kleinen Menschen sitzt. Und doch wird in der philosophischen Literatur ein Großteil der Diskussion über die menschliche Wahrnehmung so behandelt, als wäre sie etwas, das in meinem Kopf passiert.

Da ist zum Beispiel das 2D-3D-Problem beim Sehen; das Problem, wie ich die visuelle Erfahrung einer dreidimensionalen Welt durch ein zweidimensionales Wahrnehmungssystem erzeugen kann (die Netzhaut ist eine zweidimensionale Schicht, die unsere Augäpfel von innen bedeckt). Es muss doch einen Weg geben, die visuellen Informationen in der Netzhaut, dem Sehnerv und den optischen Verarbeitungszentren des Gehirns zu verarbeiten, um diese dreidimensionale Erfahrung zu erzeugen. Aber so geht es nicht wirklich zu.[10]

4.2 Aktive Wahrnehmung

Tatsächlich entsteht der dreidimensionale Eindruck durch eine Interaktion zwischen mir und der Welt (im Falle des Sehens durch die Bewegung meiner Augen und meines Körpers). Es ist besser, die Wahrnehmung in Anlehnung an den Tastsinn zu

10 Für eine Einführung in die Vision siehe O'Regan 2011: Kap. 1–5.

betrachten: Berühren ist etwas, das ich *tue*, um die Weichheit eines Gegenstandes, die Beschaffenheit seiner Oberfläche, seine Temperatur, sein Gewicht, seine Biegsamkeit usw. zu erfahren. Ich tue dies, indem ich handle und dann die Veränderung des sensorischen Inputs wahrnehme. Das nennt man eine Wahrnehmungs-Handlungs-Schleife: Ich tue etwas, das die Welt verändert, und verändere damit die Wahrnehmung, die ich habe.

Es wird nützlich sein zu betonen, dass dies auch bei der Wahrnehmung meines eigenen Körpers geschieht. Ich weiß nur deshalb, dass ich eine Hand habe, weil meine visuelle Wahrnehmung der Hand, die Propriozeption und der Tastsinn übereinstimmen. Wenn das nicht der Fall ist, ist es ziemlich einfach, mir das Gefühl zu geben, dass z.B. eine Gummihand meine eigene Hand ist – dies ist als die »Gummihand-Illusion« bekannt. Wenn eine Handprothese in geeigneter Weise mit dem Nervensystem eines Menschen verbunden ist, kann die Wahrnehmungs- und Handlungsschleife wieder geschlossen werden, und der Mensch wird sie als seine eigene Hand empfinden.

4.3 Prädiktive Verarbeitung und Verkörperung

Diese Sichtweise der Wahrnehmung hat kürzlich zu einer Theorie des »prädiktiven/vorhersagenden Gehirns« (predictive brain) geführt: Das Gehirn wartet nicht passiv auf Eingaben, sondern ist *immer aktiv an* der Handlungs-Wahrnehmungs-Schleife beteiligt. Es erstellt *Vorhersagen* darüber, wie der sensorische Input in Anbetracht meiner Handlungen sein wird, und gleicht diese Vorhersagen dann mit dem tatsächlichen sensorischen Input ab. Der Unterschied zwischen den beiden ist etwas, das wir zu minimieren versuchen, was als »Prinzip der freien Energie« bezeichnet wird (Clark 2013; Clark 2016; Friston 2010).

In dieser Tradition ist die Wahrnehmung eines natürlichen Agenten oder auch eines KI-Systems etwas, das eng mit der physischen Interaktion des Körpers des Agenten mit der Umwelt verbunden ist; die Wahrnehmung ist somit eine Komponente der verkörperten Kognition. Ein nützlicher Slogan in diesem Zusammenhang ist »4E-Kognition«, der besagt, dass Kognition *verkörpert* ist; sie ist in eine Umgebung mit anderen Agenten *eingebettet*; sie ist eher *enaktiv* als passiv; und sie ist *ausgedehnt* (»extended«), d.h. sie findet nicht nur im Kopf statt (Clark/Chalmers 1998; Clark 2003; Newen et al. 2018). Ein Aspekt, der eng mit der 4E-Kognition zusammenhängt, ist die Frage, ob Kognition beim Menschen grundsätzlich repräsentational ist und ob Kognition in der KI repräsentational sein muss (siehe Abschnitt 5).

Verkörperte Kognition wird manchmal als empirische These über die tatsächliche Kognition (insbesondere beim Menschen) oder aber als These über die geeignete Gestaltung von KI-Systemen und manchmal auch als Analyse dessen, was Kognition ist und sein muss, dargestellt. In letzterem Verständnis würde eine nicht verkörper-

te KI zwangsläufig bestimmte Merkmale der Kognition vermissen lassen (Dreyfus 1972; Pfeifer/Bongard 2007).

5. Bedeutung und Repräsentation

5.1 Das Argument des Chinesischen Zimmers

Wie wir oben gesehen haben, beruht die klassische KI auf der Annahme, dass der entsprechend programmierte Computer tatsächlich ein Geist *ist* – mit dieser Annahme kennzeichnete John Searle die *starke KI*. In seinem berühmten Aufsatz »Minds, Brains and Programs« stellte Searle das Gedankenexperiment des »Chinesischen Zimmers« vor (Searle 1980). Das Chinesische Zimmer ist ein Computer, der wie folgt aufgebaut ist: Es gibt einen geschlossenen Raum, in dem John Searle sitzt und ein großes Buch in der Hand hält, das ihm ein Computerprogramm mit Algorithmen vorgibt, wie die Eingabe zu verarbeiten und die Ausgabe zu liefern ist. Was er nicht weiß, ist, dass die Eingabe, die er erhält, ein chinesischer Text ist, und dass die Ausgabe, die er liefert, sinnvolle chinesische Antworten oder Kommentare zu dieser sprachlichen Eingabe darstellen. Die Ausgabe, so die Annahme, ist von der eines kompetenten chinesischen Sprechers nicht zu unterscheiden. Und doch versteht Searle in diesem Raum kein Chinesisch und wird mit dem Input, den er erhält, auch nicht Chinesisch lernen. Daraus schließt Searle, dass *Berechnungen für Verstehen nicht ausreichen*. Es kann keine starke KI geben.

In der weiteren Erörterung seines Arguments des Chinesischen Zimmers geht Searle auf zwei erwartbare typische Entgegnungen ein: Die *System-Antwort* akzeptiert zwar, dass Searle gezeigt hat, dass keine einfache Manipulation der Person im Raum diese Person in die Lage versetzen wird, Chinesisch zu verstehen, wendet aber ein, dass die Manipulation von Symbolen doch vielleicht das *umfassendere System*, von dem die Person nur ein Teil ist, in die Lage versetzen wird, Chinesisch zu verstehen. Steckt in Searles Argument also vielleicht ein Fehlschluss vom Teil aufs Ganze? Dieser Einwand wirft allerdings die Frage auf, warum man denken sollte, dass das Gesamtsystem Eigenschaften aufweist, die der algorithmische Prozessor selbst nicht hat.

Eine Möglichkeit, auf diese Herausforderung mit dem Vorschlag einer bestimmten Systemveränderungen zu antworten, nennt Searle die *Roboter-Antwort*. Sie räumt ein, dass das größere System, so wie es beschrieben ist, zwar kein Chinesisch versteht, aber nur weil dem System etwas fehlt, was Chinesisch sprechende Menschen haben, nämlich eine kausale Verbindung zwischen den Worten und der Welt. Wir müssten also Sensoren und Effektoren zu unseren Computer hinzufügen, die für die notwendige kausale Verbindung sorgen würden. Searle entgegnet auf diesen Vorschlag, dass die Eingabe von Sensoren für den Searle im Inneren des

Zimmers »nur noch mehr Chinesisch« wäre; sie würde kein weiteres Verständnis liefern, tatsächlich hätte Searle keine Ahnung, dass die Eingabe von einem Sensor stammt (Cole 2020; Preston/Bishop 2002).

5.2 Rekonstruktion

Ich denke, wir können den Kern des Arguments des Chinesischen Zimmers als eine Erweiterung der folgenden Beobachtung Searles betrachten:

> »Niemand würde annehmen, dass wir Milch und Zucker durch eine Computersimulation der formalen Abläufe bei der Laktation und der Photosynthese erzeugen können, aber wenn es um den Geist geht, sind viele Menschen bereit, an ein solches Wunder zu glauben.« (Searle 1980: 424)

Der Kern des Arguments lässt sich dann so rekonstruieren:

1. Ein System, das nur syntaktische Manipulationen vornimmt, kann keine Bedeutungen erfassen.
2. Ein Computer nimmt nur syntaktische Manipulationen vor.
3. Also kann ein Computer keine Bedeutungen erfassen.

In Searles Terminologie hat ein Computer *nur eine Syntax* und *keine Semantik*; den Symbolen in einem Computer fehlt die Intentionalität (Gerichtetheit), die der menschliche Sprachgebrauch hat. Am Schluss seines Aufsatzes fasst er seine Position zusammen:

> »›Könnte eine Maschine denken?‹ Die Antwort lautet natürlich: Ja. Wir sind genau solche Maschinen. […] Aber könnte etwas denken, verstehen und so weiter, allein kraft dessen, dass es ein Computer mit der richtigen Art von Programm ist? […] die Antwort ist: Nein.« (Searle 1980: 422)

5.3 Berechnungen, Syntax und Kausalkräfte

Wenn man Searles Argument auf diese Weise rekonstruiert, stellt sich die Frage, ob die Prämissen wahr sind. Mehrere Kommentatoren haben argumentiert, dass Prämisse 2 falsch ist, weil man das, was ein Computer tut, als sinnvolle Reaktion auf sein Programm verstehen müsse (McCarthy 2007; Boden 1988: 97; Haugeland 2002: 385). Ich bin der Meinung, dass dies ein Irrtum ist, denn der Computer *folgt* diesen Regeln nicht, er ist lediglich so konstruiert, dass er diesen Regeln *entsprechend handelt*, wenn seine Zustände von einem Beobachter entsprechend interpretiert werden.[11] Abgese-

11 Vgl. schon das Argument bei Wittgenstein 1960[1953]: §§ 82–86, 198, 217 usw.

hen davon hat jeder tatsächliche Computer, jede physische Realisierung eines abstrakten Algorithmus-Prozessors, sehr wohl kausale Kräfte, er kann mehr als bloß syntaktische Manipulationen durchführen. Er kann zum Beispiel das Licht an- oder ausschalten.

Das Argument des Chinesischen Zimmers hat die Aufmerksamkeit in der Sprachphilosophie weg von Konventionen und Logik hin zu den Bedingungen gelenkt, unter denen ein Sprecher das meint, was er sagt (Sprecherbedeutung), oder überhaupt etwas meint (Intentionalität); es hat neue Diskussionen angeregt, insbesondere über die Rolle, die *Repräsentationen* in der Kognition spielen, und über die Rolle des Rechnens mit Repräsentationen (Searle 1984; Searle 2004).

6 Rationale Wahl

6.1 Normative Entscheidungstheorie (MEU)

Ein rationaler Akteur nimmt die Umwelt wahr, findet heraus, welche Handlungsoptionen bestehen, und trifft dann die beste Entscheidung. Genau darum geht es in der Entscheidungstheorie. Sie ist eine normative Theorie darüber, wie ein rationaler Akteur angesichts des ihm zur Verfügung stehenden Wissens handeln *sollte* – und keine deskriptive Theorie darüber, wie rationale Akteure tatsächlich handeln *werden*.

Wie sollte also ein rationaler Akteur entscheiden, welche die bestmögliche Handlung ist? Er bewertet die möglichen Ergebnisse jeder Wahl und wählt dann die beste aus, d.h. diejenige, die den höchsten subjektiven Nutzen hat, d.h. Nutzen aus der Sicht des jeweiligen Akteurs. Man beachte, dass rationale Entscheidungen in diesem Sinne nicht notwendigerweise egoistisch sind. Es könnte durchaus sein, dass der Akteur dem Glück einer anderen Person einen hohen Nutzen beimisst und daher rational eine Handlungsweise wählt, die den Gesamtnutzen, wie er diesen selber sieht, durch das Glück dieser anderen Person maximiert. In realen Situationen weiß der Akteur in der Regel nicht, wie die Ergebnisse bestimmter Entscheidungen aussehen werden, so dass er unter Unsicherheit handelt. Um dieses Problem zu überwinden, wählt der rationale Akteur die Handlung mit dem *maximalen erwarteten Nutzen* (MEU), wobei der Wert einer Wahl gleich dem Nutzen des Ergebnisses multipliziert mit der Wahrscheinlichkeit des Eintretens dieses Ergebnisses ist. Man denke an die rationalen Erwartungen, die man hat, wenn man bei bestimmten Glücksspielen oder Lotterien mitmacht.

Komplizierter sind die Fälle von Entscheidungen, wo die Rationalität der je bestimmten Wahl von den nachfolgenden Entscheidungen *anderer Akteure* abhängt. Solche Fälle werden oft mit Hilfe von »Spielen« beschrieben, die zusammen mit anderen Akteuren gespielt werden. In solchen Spielen ist es oft eine erfolgreiche

Strategie, mit anderen Akteuren zu kooperieren, um den subjektiven Nutzen zu maximieren.

Im Diskurs der künstlichen Intelligenz ist es üblich, KI-Agenten als rationale Agenten im beschriebenen Sinne zu betrachten. So bemerkt beispielsweise Stuart Russell:

> »Kurz gesagt, ein rationaler Agent handelt so, dass er den erwarteten Nutzen maximiert. Die Bedeutung dieser Schlussfolgerung kann gar nicht hoch genug eingeschätzt werden. In vielerlei Hinsicht ging es bei der künstlichen Intelligenz vor allem darum, herauszufinden, wie man rationale Maschinen bauen kann.« (Russell 2019: 23)

6.2 Ressourcen und rationale Handlungsfähigkeit

Es ist nicht der Fall, dass ein rationaler Agent *tatsächlich* immer die perfekte Option wählt. Das liegt vor allem daran, dass ein solcher Agent damit zurechtkommen muss, dass seine Ressourcen begrenzt sind, insbesondere Informationsspeicherung (Datenspeicher) und Zeit (bei den meisten Entscheidungen ist Zeit eine kritische Größe). Die Frage ist also nicht nur, was die beste Wahl ist, sondern auch, wie viele Ressourcen ich für die Optimierung meiner Wahl aufwenden sollte; wann sollte ich aufhören zu optimieren und anfangen zu handeln? Dieses Phänomen wird als *eingegrenzte Rationalität* (bounded rationality) oder *begrenzte Optimalität* bezeichnet und verlangt in der Kognitionswissenschaft eine *ressourcenrationale* Analyse (Lieder/Griffiths 2020; Russell 2016: 16ff.; Simon 1955: 99; Wheeler 2020). Außerdem gibt es keine feststehende Menge diskreter Optionen, aus denen man wählen kann, und so muss ein rationaler Akteur nicht nur über seine Mittel nachdenken, sondern auch über seine Ziele (siehe Abschnitt 9).

Die Tatsache, dass (natürliche oder künstliche) Akteure bei ihren Entscheidungen mit begrenzten Ressourcen umgehen müssen, ist für das Verständnis der Kognition von enormer Bedeutung. In der Philosophie wird dies oft nicht in vollem Umfang gewürdigt – selbst in der Literatur über die Grenzen der rationalen Wahl scheint man oft der Meinung zu sein, es wäre irgendwie »falsch«, Heuristiken zu verwenden, die Voreinstellungen (biases) enthalten, sich von der relevanten Umwelt »anschubsen« zu lassen (nudging), oder die Umwelt für »erweiterte« oder »situierte« Kognition zu nutzen.[12] Eigentlich wäre es jedoch irrational, nach perfekten kognitiven Verfahren zu streben, ganz zu schweigen von kognitiven Verfahren, die in jeder Umwelt perfekte Ergebnisse liefern.

12 Kahneman/Tversky 1979; Kahnemann 2011; Thaler/Sunstein 2008, vs. Kirsh 2009.

6.3 Rahmungsproblem(e)

Das ursprüngliche sogenannte Rahmungsproblem (frame problem) der klassischen KI bestand darin, wie das Überzeugungssystem eines Akteurs nach einer erfolgten Handlung *aktualisiert* werden kann, ohne alles anführen zu müssen, was sich *nicht* geändert hat. Dies erfordert eine Logik, in der sich die Schlussfolgerungen ändern können, wenn eine Prämisse hinzugefügt wird – eine nicht-monotone Logik. (Shanahan 2016) Über dieses eher technische Problem hinaus gibt es ein philosophisches Problem der Aktualisierung von Überzeugungen nach einer Handlung, das von Dennett popularisiert wurde und die Frage aufwirft, wie man herausfinden kann, was relevant ist und wie weit der Rahmen für *Relevanz* gezogen werden sollte. Wie Shanahan bemerkt, ist »Relevanz ganzheitlich, ergebnisoffen und kontextabhängig«, aber logische Schlussfolgerungen sind es nicht (Dennett 1984a; Shanahan 2016).

Es gibt eine sehr allgemeine Version des Frame-Problems, die von Jerry Fodor formuliert wurde. Er vergleicht es mit »Hamlets Problem: wann man aufhören soll zu denken«. Und er meint, dass »modulare kognitive Verarbeitung *ipso facto* irrational […] ist, weil weniger als alle relevante und verfügbare Evidenz einbezogen wird« (Fodor 1987: 140f.; Sperber/Wilson 1996). Fodor macht damit auf das Problem aufmerksam, dass man, um eine logische Schlussfolgerung, insbesondere eine Abduktion, durchzuführen, schon entschieden haben muss, was überhaupt als relevant gelten soll. Er scheint jedoch die Tatsache zu unterschätzen, dass man sich nicht um *alles* kümmern kann, was relevant und verfügbar ist (denn unsere Rationalität ist eingegrenzt). Es ist derzeit unklar, ob das Rahmungsproblem ohne fragwürdige Annahmen über Rationalität formuliert werden kann. Ähnliche Bedenken treffen die Behauptung, Gödel habe die tiefen Grenzen von KI-Systemen aufgezeigt (Koellner 2018a; Koellner 2018b; Lucas 1996). Womöglich beinhaltet Intelligenz doch mehr als nur instrumentelle Rationalität.

6.4 Kreativität

Entscheidungen, die mit *Kreativität* zu tun haben, werden oft für etwas gehalten, das über alles Mechanische hinausgeht und daher für eine bloße Maschine unerreichbar ist. Der Begriff des »schöpferischen Schaffens« hat in unserer gesellschaftlichen Praxis erhebliches Gewicht, insbesondere wenn diese Schöpfung durch geistige Eigentumsrechte geschützt ist – und KI-Systeme *haben* Musik, Malerei und Texte geschaffen oder mitgeschaffen. Es ist überhaupt nicht klar, ob es einen Begriff von Kreativität gibt, der ein Argument gegen maschinelle Kreativität liefern würde. Ein solcher Begriff müsste zwei Aspekte miteinander verbinden, die in einem Spannungsverhältnis zu stehen scheinen: Einerseits scheint Kreativität eine Ursächlichkeit zu implizieren, die den Erwerb von Wissen und Techniken einschließt (man

denke an J.S. Bach, wie er eine neue Kantate komponiert), andererseits soll Kreativität so etwas wie ein nicht-verursachter, nicht-vorhersehbarer Einsichtsfunke sein. Es ist gar nicht klar, ob ein solcher Begriff von Kreativität überhaupt formuliert werden kann oder sollte (Boden 2014; Colton/Wiggins 2012; Halina 2021). Vielleicht ergibt sich eine plausible Erklärung von Kreativität, wenn wir davon ausgehen, dass es bei Kreativität darum geht, sich zwischen verschiedenen Räumen der Relevanz zu bewegen, ähnlich wie beim Rahmungsproblem.

7. Freier Wille und Kreativität

7.1 Determinismus, Kompatibilismus

Das Problem, das gewöhnlich unter der Überschrift »freier Wille« behandelt wird, ist die Frage, wie physische Wesen wie Menschen oder KI-Systeme so etwas wie einen freien Willen haben können. Die übliche Einteilung möglicher Positionen im Diskurs über den freien Willen lässt sich in Form eines Entscheidungsbaums darstellen. Die erste Verzweigung ist die Frage, ob der *Determinismus* wahr ist, d.h. die These, dass alle Ereignisse verursacht werden. Die zweite Verzweigung ist, ob der *Inkompatibilismus* wahr ist, d.h. die These, dass es keinen freien Willen gibt, wenn der Determinismus wahr ist.

Die als *harter Determinismus* bekannte Position besagt, dass der Determinismus tatsächlich wahr ist und es deshalb so etwas wie Willensfreiheit nicht gibt – dies ist die Schlussfolgerung, die die meisten seiner Gegner zu vermeiden versuchen. Die Position, die als *Libertarismus* bekannt ist (nicht im politischen Sinne), stimmt zu, dass der Inkompatibilismus wahr ist, fügt aber hinzu, dass der Determinismus nicht wahr ist und wir daher frei sind. Die als *Kompatibilismus* bekannte Position besagt, dass Determinismus und freier Wille miteinander vereinbar sind und es daher durchaus sein kann (und wohl tatsächlich auch so ist), dass der Determinismus wahr ist *und* der Mensch einen freien Willen hat.

Daraus ergibt sich eine kleine Matrix von Positionen:

	Inkompatibilismus	*Kompatibilismus*
Determinismus	harter Determinismus	optimistischer/pessimistischer Kompatibilismus
Nicht-Determinismus	Libertarismus	[keine beliebte Option]

7.2 Kompatibilismus und Verantwortung in der KI

Wenn ich sage, dass ich etwas aus freien Stücken getan habe, bedeutet das in erster Näherung, dass es *an mir lag*, dass ich die *Kontrolle* hatte. Dieser Begriff von Kontrolle lässt sich erläutern, indem man sagt, ich hätte anders handeln können als ich es getan habe, insbesondere hätte ich anders handeln können, wenn ich *mich* anders *entschieden* hätte. Und dass ich mich anders entschieden hätte, wenn ich andere *Vorlieben* oder *Kenntnisse* gehabt hätte (z.B. hätte ich diese Fleischbällchen nicht gegessen, wenn ich eine Abneigung gegen Schweinefleisch hätte und wenn ich gewusst hätte, dass die Bällchen Schweinefleisch enthalten). Der entsprechende Freiheitsbegriff beinhaltet also eine *epistemische Bedingung* und eine *Kontrollbedingung*.

Ich handle also frei, wenn ich gemäß meinen Präferenzen (meinem subjektiven Nutzen) handle. Aber warum habe ich diese Präferenzen? Wie schon Aristoteles wusste, unterstehen sie nicht meiner willlentlichen Kontrolle, ich könnte nicht einfach *beschließen*, andere Präferenzen zu haben und sie dann haben. Harry Frankfurt hat allerdings deutlich gemacht hat, dass ich Präferenzen oder Wünsche *zweiter Ordnung* haben kann, d.h. ich kann präferieren, andere Präferenzen zu haben als die, die ich tatsächlich habe (z.B. könnte ich es mögen, Fleischbällchen nicht zu mögen). Dass ich meine Präferenzen durch rationales Denken außer Kraft setzen kann, nennt Frankfurt den *Willen*, und dieser ist eine Bedingung dafür, dass ich eine Person bin. Näherungsweise kann man also sagen, *frei zu handeln bedeutet, so zu handeln, wie ich mich entscheide; mich so zu entscheiden, wie ich es will; und so zu wollen, wie ich es vernünftigerweise vorziehe, zu wollen* (Dennett 1984b; Frankfurt 1971).

Die Debatte läuft darauf hinaus, dass der Begriff des freien Willens bei KI oder Menschen die Funktion hat, persönliche *Verantwortung* zu ermöglichen, und nicht, eine *Ursache* zu bestimmen. Die eigentliche Frage lautet: Unter welchen Bedingungen ist ein Akteur für seine Handlungen *verantwortlich* und *verdient es*, dafür gelobt oder getadelt zu werden? Dies gilt unabhängig davon, ob wir frei von kausaler Determination handeln; diese Art von Freiheit bekommen wir nicht und brauchen wir auch nicht.

Zwischen »Optimisten« und »Pessimisten« gibt es eine weitere Debatte darüber, ob Menschen diese Bedingungen tatsächlich erfüllen (insbesondere, wieweit sie wirklich ihre Präferenzen kausal hervorbringen können) und daher zu Recht für ihre Handlungen verantwortlich sind und Lob oder Tadel *verdienen* – und ob Belohnung oder Bestrafung dementsprechend hauptsächlich zukunftsorientierte Gründe haben sollten (Dennett/Caruso 2018; Mele 2006; Pink 2004; Strawson 2004). Was KI-Systeme betrifft, so hat das Nichtvorhandensein von Verantwortung Konsequenzen für ihren Status als moralische Akteure, für die Existenz von »Verantwortungslücken« und für die komplexe normative Frage, welche Art von

Entscheidungen wir Systemen überlassen sollten, die nicht verantwortlich gemacht werden können. (Müller 2021; Simpson/Müller 2016; Sparrow 2007)[13]

8 Bewusstsein

8.1 Bewusstheit und phänomenales Bewusstsein

Zunächst ist es sinnvoll, zwei Arten von Bewusstsein zu unterscheiden: *Bewusstheit* und *phänomenales Bewusstsein*. Bewusstheit ist die Vorstellung, dass ein System kognitive Zustände auf einer Basisebene hat (z.B. spürt es Wärme) und auf einer Metaebene Zustände hat, in denen es sich der Zustände auf der Basisebene bewusst ist. Diese Bewusstheit bzw. dieser Zugang beinhaltet die Fähigkeit, sich an die kognitiven Zustände auf der Basisebene zu erinnern und sie zu nutzen. Dies ist der begriffliche Sinn von »bewusst« im Unterschied zu »unbewusst« oder »unterbewusst«. Und für ein mehrschichtiges KI-System scheinen diese Unterscheidungen auch machbar zu sein.

Mit Bewusstheit geht oft, aber nicht notwendigerweise, einher, dass der kognitive Zustand auf der Basisebene sich für das Subjekt auf eine bestimmte Weise *anfühlt*. Dies wird philosophisch als *phänomenales Bewusstsein* bezeichnet: dies, wie mir die Dinge *erscheinen* (griechisch *phaínetai*). Dieser Begriff des Bewusstseins lässt sich wahrscheinlich am besten mit Hilfe von zwei klassischen philosophischen Gedankenexperimenten erklären: der Fledermaus und der Farbenwissenschaftlerin.

Angenommen, Sie und ich essen beide Schokoladeneis. Dann kann ich immer noch nicht wissen, wie das Eis für Sie schmeckt, und ich würde es auch dann nicht wissen, wenn ich alles über das Eis, über Sie, über Ihr Gehirn und Ihre Geschmacksknospen wüsste. *Wie* es für Sie schmeckt, ist etwas, das für mich epistemisch unzugänglich ist, ich kann es niemals wissen, selbst wenn ich alles über die physische Welt wüsste. Genauso wenig kann ich je wissen, wie es sich anfühlt, eine Fledermaus zu sein (Nagel 1974; Nagel 1987: Kap. 3).

Eine ähnliche Pointe zur Frage des nicht Wissbaren macht Frank Jackson in dem vieldiskutierten Artikel »Was Mary nicht wusste« (Jackson 1986). In seinem Gedankenexperiment soll Mary eine Person sein, die in ihrem Leben noch nie etwas Farbiges gesehen hat, die aber eine perfekte Farbenwissenschaftlerin ist: Sie weiß alles, was es über Farbe zu wissen gibt. Eines Tages kommt sie aus ihrer schwarz-weißen Umgebung heraus und sieht zum ersten Mal Farbe. Es scheint, als ob sie in diesem Moment etwas Neues lernt.

Das Argument, das hier vorgebracht wird, scheint für einen geistig-physikalischen *Dualismus* von *Substanzen* oder zumindest *Eigenschaften* zu sprechen: Ich könn-

13 Siehe auch den Beitrag von Susanne Hahn im vorliegenden Band. [Anm. der Hrsg.]

te alles Wissen der Physik haben, ohne das Wissen der phänomenalen Erfahrung zu haben, also ist die phänomenale Erfahrung kein Teilgebiet der Physik. Wenn der Dualismus wahr ist, dann schaut es so aus, dass wir nicht hoffen dürfen, mit der richtigen physikalischen Technologie, wie z.B. der KI, phänomenales Bewusstsein zu erzeugen. In der Gestalt des *Substanzdualismus*, wie ihn Descartes und ein Großteil des religiösen Denkens angenommen haben, ist der Dualismus heute allerdings unpopulär: Die meisten Philosophen gehen von einem Physikalismus aus, der besagt, dass »alles physisch ist«.

Eine ganze Reihe von Argumenten gegen die Reduktion mentaler auf physische *Eigenschaften* werden diskutiert, so dass man wohl mit Fug und Recht behaupten kann, dass der *Eigenschaftsdualismus* eine große Anhängerschaft hat. Dieser wird oft mit dem Substanzmonismus zu einer Version der »Supervenienz des Mentalen auf dem Physischen« kombiniert, d.h. zu der These, dass zwei Entitäten mit denselben physischen Eigenschaften auch dieselben mentalen Eigenschaften haben müssen. Einige Philosophen haben diese Beziehung zwischen dem Eigenschaftsdualismus und der Möglichkeit eines künstlichen Bewusstseins in Frage gestellt. So behauptet David Chalmers, dass »die physikalische Struktur der Welt – die genaue Verteilung von Teilchen, Feldern und Kräften in der Raumzeit – logisch mit der Abwesenheit von Bewusstsein vereinbar ist, so dass das Vorhandensein von Bewusstsein eine weitere Tatsache über unsere Welt ist«. Trotz dieser Behauptung unterstützt er den Computationalismus und meint: »starke künstliche Intelligenz ist wahr: Es gibt eine Klasse von Programmen, bei denen jede Implementierung eines Programms dieser Klasse bewusst ist.« (Chalmers 1999: 436; Chalmers und Searle 1997; Davidson 1970)

Bedeutsamer aber als die Diskussion über Dualismen ist das Verständnis der *Funktion* von Bewusstsein in KI-Systemen oder bei natürlichen Agenten: Warum ist das phänomenale Bewusstsein beim Menschen so, wie es ist? Wie könnten wir feststellen, ob ein System Bewusstsein hat? Könnte es einen Menschen geben, der physisch genauso gebaut wäre wie ich, aber kein Bewusstsein hätte (ein »philosophischer Zombie«) (O'Regan 2011)?

8.2 Das Selbst

Die persönliche Identität ist für Menschen vor allem deshalb bedeutsam, weil sie eine Voraussetzung für die Zuweisung von Verantwortung ist (siehe Abschnitt 6.4): Um Schuld oder Lob zuzuweisen, muss ich in einem bestimmten Sinne *dieselbe Person* sein in wie derjenige, der die betreffende Handlung ausgeführt hat. Es gehört zu unserem Selbstverständnis, dass es ein Leben in der Vergangenheit gibt, das meines ist, und nur meines – wie das möglich ist, ist als die »Frage der Persistenz« bekannt. Die Standardkriterien dafür, dass ich dieselbe Person bin wie der kleine Junge auf dem Foto, sind meine *Erinnerung* daran, dieser Junge zu sein, und die *Kontinuität meines Körpers* über die Zeit. Wir Menschen neigen dazu zu glauben, dass *Erinne-*

rung oder *bewusste Erfahrung* oder *geistige Inhalte die* Kriterien für persönliche Identität sind, weshalb wir uns auch vorstellen können, unseren Tod zu überleben oder in einem anderen Körper zu leben (Metzinger 2009; Olsen 2019).

Was also ist ein »Teil« dieses beständigen Selbst? Abgesehen von philosophischen Phantasien und neurologischen Raritäten[14] gibt es heute keinen Zweifel mehr daran, was »Teil von mir« ist und was nicht – ich arbeite ständig daran, diese persönliche Identität aufrechtzuerhalten, indem ich prüfe, ob die verschiedenen Sinne übereinstimmen, z. B. versuche ich, nach der Türklinke zu greifen, ich sehe, wie meine Hand die Klinke berührt, ich kann sie fühlen … und dann sehe ich, wie sich die Tür öffnet, und spüre, wie meine Hand sich nach vorn bewegt. Das ist etwas ganz anderes als ein Computer: Die Komponenten der Standard Von-Neumann-Architektur (Eingabesystem, Speicher, Direktzugriffsspeicher, Prozessor, Ausgabesystem) können sich im selben Gehäuse befinden oder meilenweit voneinander entfernt sein, sie können sogar in mehrere Komponenten aufgeteilt sein (z. B. bei Off-Board Arbeitsprozessen an rechenintensiven Aufgaben) oder in Räumen wie der »Cloud« gespeichert sein, die nicht durch einen physischen Ort definiert sind. Und das ist nur die Hardware, die Software steht vor ähnlichen Problemen, so dass ein beständiges und abgegrenztes Selbst auszubilden keine leichte Aufgabe für ein KI-System wäre. Es ist auch gar nicht klar, ob es überhaupt eine Funktion für ein Selbst in der KI gibt und welche Konsequenzen für die Zuschreibung von moralischem Handeln und Behandeltwerden das hat.

9. Normativität

Kehren wir kurz zu den Fragen der rationalen Wahl und der Verantwortung zurück. Stuart Russell meint, dass »die KI das Standardmodell übernommen hat: Wir bauen optimierende Maschinen, wir geben ihnen Ziele vor, und los geht's.« (Russell 2019: 172) Nach diesem Verständnis ist die KI ein Werkzeug, und wir müssen ihr die Ziele vorgeben, die sie erreichen soll. Die KI verfügt ausschließlich über *instrumentelle Intelligenz*, um die vorgegebenen Ziele zu erreichen. Zur *allgemeinen Intelligenz* gehört jedoch auch eine metakognitive Reflexionsfähigkeit, welche Ziele für mein jetziges Handeln relevant sind (Nahrung oder Unterkunft?) und eine Reflexion darüber, welche Ziele man verfolgen sollte (Müller/Cannon 2022). Eine der vielen offenen Fragen ist, ob ein nicht-lebendes System »echte Ziele« in dem Sinne haben kann, der für Handlungsentscheidungen und Verantwortung erforderlich ist, d.h. Ziele, die für das System einen subjektiven Wert haben und die das System reflektierend als

14 Z.B. »Der Mann, der aus dem Bett fiel« in (Sacks 1985) oder die Betrachtung des Menschen als Superorganismus, basierend auf dem menschlichen Mikrobiom.

wichtig erkennt. Ohne eine solche Reflexion über Ziele wären KI-Systeme keine moralischen Agenten und es könnte keine »Maschinenethik« geben, die diesen Namen verdient. Ähnliche Überlegungen gelten für andere Formen der normativen Reflexion, z.B. in der Ästhetik und der Politik. Diese Diskussion in der KI-Philosophie deutet darauf hin, dass der normativen Reflexion eine basale Funktion im kognitiven System zukommt, ob beim Menschen oder bei KI-Systemen.

Literatur

Anonym (1950): Digital Computing Newsletter, in: Office of Naval Research, Mathematical Sciences Division: Washington (DC), 2(1), 1–4.

Boden, M.A. (2014): Creativity and artificial intelligence. A contradiction in terms?, in: Paul, E.S.; Kaufman, S.B. (Hg.), The philosophy of creativity. New essays, Oxford: Oxford University Press.

Boden, M.A. (1988): Computer models of the mind. Computational approaches in theoretical psychology, Cambridge: Cambridge University Press.

Boden, M.A. (2. Auflage 2006): Mind as machine. A history of cognitive science, Oxford: Oxford University Press.

Buckner, C. (forthcoming): From deep learning to rational machines. What the history of philosophy can teach us about the future of artificial intelligence, New York: Oxford University Press.

Carter, M. (2007): Minds and computers. An introduction to the philosophy of artificial intelligence, Edinburgh: Edinburgh University Press.

Chalmers, D.J. (1999): Précis of The Conscious Mind, in: *Philosophy and Phenomenological Research*, LIX(2), 435–438.

Chalmers, D.J.; Searle, J. (1997): Consciousness and the philosophers. An exchange, in: *New York Review of Books*, 15.05.1997. [https://www.nybooks.com/articles/1997/05/15/consciousness-and-the-philosophers-an-exchange/] (Zugriff: 25.05.2024).

Clark, A. (2003): Natural born cyborgs. Minds, technologies, and the future of human intelligence, Oxford: Oxford University Press.

Clark, A. (2013): Whatever next? Predictive brains, situated agents, and the future of cognitive science, in: *Behavioral and Brain Sciences*, 36(6), 181–204.

Clark, A. (2016): Surfing uncertainty. Prediction, action, and the embodied mind, New York: Oxford University Press.

Clark, A.; Chalmers, D.J. (1998): The extended mind, in: *Analysis*, 58(1), 7–19.

Coelho Mollo, D. (2022): Intelligent Behaviour, in: *Erkenntnis*, 89, 705–721.

Cole, D. (2020): The Chinese room argument, in: Zalta, E.N.; Nodelman, U. (Hg.), Stanford Encyclopedia of Philosophy, Stanford (CA): The Metaphysics Research Lab. [http://plato.stanford.edu/entries/chinese-room/] (Zugriff: 22.05.2024).

Colton, S.; Wiggins, G.A. (2012): Computational creativity: The final frontier?, in: *Frontiers in Artificial Intelligence and Applications*, 242, 21–26.

Copeland, J.B. (1993): Artificial intelligence. A philosophical introduction, Oxford: Blackwell.

Copeland, J.B.; Posy, C.J.; Shagrir, O. (Hg.) (2013): Computability. Turing, Gödel, Church, and Beyond, Cambridge (MA): MIT Press.

Davidson, D. (1970): Mental Events, in: Foster, L.; Swanson, J. (Hg.), Experience and Theory, Amherst (MA): University of Massachusetts Press, 137–149.

Dennett, D.C. (1984a): Cognitive wheels. The frame problem of AI, in: Hookway, C. (Hg.), Minds, machines, and evolution. Philosophical studies, Cambridge: Cambridge University Press, 129–152.

Dennett, D.C. (1984b): Elbow room. The varieties of free will worth wanting, Cambridge (MA): MIT Press.

Dennett, D.C. (1991): Consciousness explained, New York: Little, Brown & Co.

Dennett, D.C.; Caruso, G.D. (2018): Just deserts, *Aeon*, 1, 1–20.

Dietrich, E. (2002): Philosophy of artificial intelligence, in: *The Encyclopedia of Cognitive Science*, 203–208.

Dietrich, E.; Fields, C.; Sullins, J.P.; Van Heuveln, B.; Zebrowski, R. (2021): Great philosophical objections to artificial intelligence. The history and legacy of the AI wars, London: Bloomsbury Academic.

Dodig-Crnkovic, G.; Müller, V.C. (2011): A dialogue concerning two world systems: Info-computational vs. mechanistic, in: Dodig-Crnkovic, G.; Burgin, M. (Hg.), Information and computation. Essays on scientific and philosophical understanding of foundations of information and computation, Boston: World Scientific, 149–184.

Dreyfus, H.L. (2. Auflage 1992): What computers still can't do. A critique of artificial reason, Cambridge (MA): MIT Press.

Edelman, S. (20008): Computing the mind. How the mind really works, Oxford: Oxford University Press.

Floridi, L. (2011): The philosophy of information, Oxford: Oxford University Press.

Floridi, L. (Hg.) (2003): The Blackwell guide to the philosophy of computing and information, Oxford: Blackwell.

Fodor, J.A. (1987): Modules, frames, fridgeons, sleeping dogs, and the music of the spheres, in: Garfield, J.L. (Hg.), Modularity in knowledge representation and natural-language understanding, Cambridge (MA): The MIT Press, 25–36.

Frankfurt, H. (1971): Freedom of the will and the concept of a person, in: *The Journal of Philosophy*, LXVIII(1), 5–20.

Friston, K.J. (2010): The free-energy principle. A unified brain theory?, in: *Nature Reviews Neuroscience*, 11, 127–138.

Garson, J.; Buckner, C. (2019): Connectionism, in: Zalta, E.N.; Nodelman, U. (Hg.), Stanford Encyclopedia of Philosophy, Stanford (CA): The Metaphysics Research Lab. [https://plato.stanford.edu/entries/connectionism/] (Zugriff: 25.05.2024).

Gödel, K. (1931): Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I, in: *Monatshefte für Mathematik und Physik*, 38, 173–198.

Görz, G.; Schmid, U.; Braun, T. (5. Auflage 2020): Handbuch der künstlichen Intelligenz, Berlin: De Gruyter.

Halina, M. (2021): Insightful artificial intelligence, in: *Mind and Language*, 36(2), 315–329.

Harnad, S. (1990): The symbol grounding problem, in: *Physica D*, 42, 335–346.

Haugeland, J. (1985): Artificial intelligence. The very idea, Cambridge (MA): MIT Press.

Haugeland, J. (2002): Syntax, semantics, physics, in: Preston, J.; Bishop, M. (Hg.), Views into the Chinese room. New essays on Searle and artificial intelligence, Oxford: Oxford University Press, 379–392.

Hilbert, D. (1900): Mathematische Probleme, in: Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen, Math.-Phys. Klasse, 3, Göttingen: Lüder Horstmann, 253–297.

Horsman, C.; Stepney, S.; Wagner, R.C.; Kendon, V. (2014): When does a physical system compute?, in: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science*, 470 (2169), 1–25.

Jackson, F. (1986): What Mary didn't know, in: *Journal of Philosophy*, 83, 291–295.

Kahneman, D.; Tversky, A. (1979): Prospect theory. An analysis of decision under risk, in: *Econometrica*, 47, 263–291.

Kahnemann, D. (2011): Thinking fast and slow, London: Macmillan.

Kirsh, D. (2009): Problem solving and situated cognition, in: Robbins, P.; Aydede, M. (Hg.), The Cambridge handbook of situated cognition, Cambridge: Cambridge University Press, 264–306.

Koellner, P. (2018a): On the Question of Whether the Mind Can Be Mechanized, I: From Gödel to Penrose, in: *Journal of Philosophy*, 115(7), 337–360.

Koellner, P. (2018b): On the question of whether the mind can be mechanized, II: Penrose's new argument, in: *Journal of Philosophy*, 115(9), 453–484.

LeCun, Y.; Bengio, Y.; Hinton, J. (2015): Deep learning, in: *Nature*, 521(7553), 436–444.

Legg, S.; Hutter, M. (2007): Universal intelligence. A Definition of machine intelligence, in: *Minds and Machines*, 17(4), 391–444.

Lieder, F.; Griffiths, T.L. (2020): Resource-rational analysis. Understanding human cognition as the optimal use of limited computational resources, in: *Behavioral and Brain Sciences*, 43, e1, 1–60.

Lighthill, J. (1973): Artificial intelligence. A general survey, in: Science Research Council (Hg.), Artificial intelligence. A paper symposion, London: Science Research

Council. [http://www.chilton-computing.org.uk/inf/literature/reports/lighthill_report/p001.htm] (Zugriff: 25.05.2024).

Lucas, J.R. (1996): Minds, machines and Gödel. A retrospect, in: Millican, P.J.R.; Clark, A. (Hg.), Machines and Thought, Oxford: Oxford University Press, 103–124.

Margolis, E.; Samuels, R.; Stich, S. (Hg.) (2012): The Oxford handbook of philosophy of cognitive science, Oxford: Oxford University Press.

McCarthy, J. (2007): John Searle's Chinese room argument. [http://www-formal.stanford.edu/jmc/chinese.html] (Zugriff: 06.10.2007).

McCarthy, J.; Minsky, M.; Rochester, N.; Shannon, C.E. (1955): A proposal for the Dartmouth summer research project on artificial intelligence. [http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html] (Zugriff: 25.05.2024).

Mele, A.R. (2006): Free will and luck, Oxford: Oxford University Press.

Metzinger, T. (2009): The ego tunnel. The science of the mind and the myth of the self, New York: Basic Books.

Miłkowski, M. (2018): Objections to computationalism. A survey, in: *Roczniki Filozoficzne*, LXVI(8), 1–19.

Müller, V.C. (2013): What is a digital state?, in: Bishop, M.J.; Erden, Y.J. (Hg.), The Scandal of Computation – What is Computation? – AISB Convention 2013, Hove: AISB, 11–16. [http://www.aisb.org.uk/asibpublications/convention-proceedings] (Zugriff: 25.05.2024).

Müller, V.C. (2020): Ethics of artificial intelligence and robotics, in: Zalta, E.N.; Nodelman, U. (Hg.), Stanford Encyclopedia of Philosophy, Stanford (CA): The Metaphysics Research Lab. [https://plato.stanford.edu/entries/ethics-ai/] (Zugriff: 25.05.2024).

Müller, V.C. (2021): Is it time for robot rights? Moral status in artificial entities, in: *Ethics & Information Technology*, 23(3), 579–587.

Müller, V.C.; Cannon, M. (2022): Existential risk from AI and orthogonality. Can we have it both ways?, in: *Ratio*, 35(1), 25–36.

Müller, V.C. (forthcoming): Can machines think? Fundamental problems of artificial intelligence, New York: Oxford University Press.

Nagel, T. (1974): What is it like to be a bat?, in: *Philosophical Review*, 83(4), 435–450.

Nagel, T. (1987): What does it all mean? A very short introduction to philosophy, Oxford/New York: Oxford University Press.

Negroponte, N. (1995): Being digital, New York: Vintage.

Newell, A.; Simon, H. (1976): Computer science as empirical enquiry. Symbols and search, in: *Communications of the Association of Computing Machinery*, 19(3), 113–126.

Newen, A.; Gallagher, S.; De Bruin, L. (2018): 4E Cognition. Historical Roots, Key Concepts, and Central Issues, in: Newen, A.; De Bruin, L.; Gallagher, S. (Hg.), The Oxford Handbook of 4E Cognition, Oxford: Oxford University Press.

O'Regan, K.J. (2011): Why red doesn't sound like a bell. Understanding the feel of consciousness, New York: Oxford University Press.

Olsen, E. (2019): Personal identity, in: Zalta, E.N.; Nodelman, U. (Hg.), Stanford Encyclopedia of Philosophy, Stanford (CA): The Metaphysics Research Lab. [https://plato.stanford.edu/entries/identity-personal/] (Zugriff: 25.05.2024).

Pearl, J.; Mackenzie, D. (2018): The book of why. The new science of cause and effect, New York: Basic Books.

Pfeifer, R.; Bongard, J. (2007): How the body shapes the way we think. A new view of intelligence, Cambridge (MA): MIT Press.

Piccinini, G. (2021): Computation in physical systems, in: Zalta, E.N.; Nodelman, U. (Hg.), Stanford Encyclopedia of Philosophy, Stanford (CA): The Metaphysics Research Lab. [https://plato.stanford.edu/entries/computation-physicalsystems/] (Zugriff: 25.05.2024).

Pink, T. (2004): Free will. A very short introduction, Oxford: Oxford University Press.

Preston, J.; Bishop, M. (Hg.) (2002): Views into the Chinese room. New essays on Searle and artificial intelligence, Oxford: Oxford University Press.

Rosenblatt, F. (1957): The Perceptron. A perceiving and recognizing automaton (Project PARA), in: *Cornell Aeronautical Laboratory Report*, 85(460/461), 1–29.

Russell, S. (2016): Rationality and intelligence. A brief update, in: Müller, V.C. (Hg.), Fundamental issues of artificial intelligence, Cham: Springer, 7–28.

Russell, S. (2019): Human compatible. Artificial intelligence and the problem of control, New York: Viking.

Russell, S.; Norvig, P. (4. Auflage 2020): Artificial intelligence. A modern approach, Upper Saddle River: Prentice Hall.

Sacks, O. (1985): The Man Who Mistook His Wife for a Hat, and Other Clinical Tales, New York: Summit Books.

Scheutz, M. (Hg.), (2002): Computationalism. New directions, Cambridge: Cambridge University Press.

Searle, J.R. (1980): Minds, brains and programs, in: *Behavioral and Brain Sciences*, 3, 417–457.

Searle, J.R. (1984): Intentionality and its place in nature, in: Ders. (Hg.), Consciousness and language, Cambridge: Cambridge University Press, 77–89.

Searle, J.R. (2004): Mind. A brief introduction, Oxford: Oxford University Press.

Shagrir, O. (1997): Two dogmas of computationalism, in: *Minds and Machines*, 7, 321–344.

Shagrir, O. (2022): The nature of physical computation, New York: Oxford University Press.

Shanahan, M. (2016): The frame problem, in: Zalta, E.N.; Nodelman, U. (Hg.), Stanford Encyclopedia of Philosophy, Stanford (CA): The Metaphysics Research Lab. [https://plato.stanford.edu/archives/spr2016/entries/frame-problem/] (Zugriff: 25.05.2024).

Siegelmann, H.T. (1995): Computation beyond the Turing limit, in: *Science*, 268(5210), 545–548.

Siegelmann, H.T. (1997): Neural networks and analog computation. Beyond the Turing limit, Basel: Birkhäuser.

Simon, H. (1955): A behavioral model of rational choice, in: *Quarterly Journal of Economics*, 69, 99–118.

Simpson, T.W.; Müller, V.C. (2016): Just war and robots' killings, in: *The Philosophical Quarterly*, 66(263), 302–322.

Sparrow, R. (2007): Killer robots, in: *Journal of Applied Philosophy*, 24(1), 62–77.

Sperber, D.; Wilson, D. (1996): Fodor's Frame Problem and Relevance Theory, in: *Behavioral and Brain Sciences*, 19(3), 530–532.

Strawson, G. (2004): Free will, in: Craig, E. (Hg.), Routledge Encyclopedia of Philosophy Online. [https://www.rep.routledge.com/articles/thematic/free-will/] (Zugriff: 25.05.2024).

Thaler, R.H.; Sunstein, C. (2008): Nudge. Improving decisions about health, wealth and happiness, New York: Penguin.

Thoma, J. (2019): Decision Theory, in: Pettigrew, R.; Weisberg, J. (Hg.), The open handbook of formal epistemology, PhilPapers Foundation, 57–106.

Turing, A. (1936): On Computable Numbers, with an Application to the Entscheidungsproblem, in: *Proceedings of the London Mathematical Society*, 2(42), 230–265.

Turing, A. (1950): Computing machinery and intelligence, in: *Mind*, LIX, 433–460.

Varela, F.J.; Thompson, E.; Rosch, E. (1991): The embodied mind. Cognitive science and human experience, Cambridge (MA): MIT Press.

Wheeler, G. (2020): Bounded Rationality, in: Zalta, E.N.; Nodelman, U. (Hg.), Stanford Encyclopedia of Philosophy, Stanford (CA): The Metaphysics Research Lab. [https://plato.stanford.edu/archives/fall2020/entries/bounded-rationality/] (Zugriff: 25.05.2024).

Wittgenstein, L. (1960[1953]): Philosophische Untersuchungen, in: Ders., Schriften I, Frankfurt a.M.: Suhrkamp, 279–544.