Diplomarbeit

Physikalische Modelle von Bewusstsein und mögliche Anwendungen in der Medizin

unter besonderer Berücksichtigung der Kardiologie

eingereicht von

Alexander Zesar

zur Erlangung des akademischen Grades

Doktor der gesamten Heilkunde
(Dr. med. univ.)

an der

Medizinischen Universität Graz

ausgeführt an der

Klinischen Abteilung für Kardiologie
Universitätsklinik für Innere Medizin

unter Anleitung von

Univ.-Prof. DDr. Robert Gasser
Priv.-Doz.^{in} DDr.^{in} Sabrina Mörkl

Graz, 31.10.2019

Eidesstattliche Erklärung

Ich erkläre ehrenwörtlich, dass ich die vorliegende Arbeit selbstständig und ohne fremde Hilfe verfasst habe, andere als die angegebenen Quellen nicht verwendet habe und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Graz, am 31.10.2019 Alexander Zesar eh.

Inhaltsverzeichnis

Glossar und Abkürzungsverzeichnis

1-AMA	1-aminoanthrazen
1-AZA	1-azidoanthrazen
4R-MAPT	4-repeat Isoformen des Tau-Proteins
ABGB	Allgemeines bürgerliches Gesetzbuch
AUC	Area Under the Curve
AV-Block	Atrioventrikulärer Block
BMI	Body Mass Index
BOLD	blood oxygen level dependent
CNN	Convolutional Neural Network
DISC1	Disrupted-in-Schizophrenia 1 Protein
DNN	Deep Neural Network
DP OR	Diósi-Penrose Objective Reduction
DTNBP1	Dystrobrevin-bindendes Protein 1/Dysbindin
EEG	Elektroenzephalogramm
EKG	Elektrokardiogramm
F-18-FDG	^{18}F-2-Fluor-2-Deoxy-D-Glucose
FLI	Future of Life Institute
FLOPS	Floating Point Operations Per Second
fMRT	funktionelle Magnetresonanztomografie
GABA	γ-Aminobuttersäure (Gamma-Aminobutyric Acid)
GTP-2	Generative Pre-trained Transformer-2
HAP1	Huntingtin-assoziiertes Protein 1
HbA1c	Hämoglobin A1c

HDL	High Density Lipoprotein
HIP1	Huntingtin-interagierendes Protein 1
h-MAPT	Hyperphosphoryliertes Mikrotubuli-assoziiertes Protein-Tau
IBM	International Business Machines
IEF	Isoelektrische Fokussierung
IIT	Integrated Information Theory
KI	künstliche Intelligenz
KV	Kausalitäts-Vermögen
M1	primärer motorischer Kortex
MACE	schwere kardiovaskuläre Komplikation (Major Adverse Cardiovascular Event)
MAE	Mean Average Error
MAP	Mikrotubuli-assoziiertes Protein
MAPT	Mikrotubuli-assoziiertes Protein-Tau
MIP	Minimale Informations-Partition
MT	Mikrotubuli
N_2O	Distickstoffmonoxid
NCC	Neuronales Korrelat von Bewusstsein (neural correlate of consciousness)
NFT	Neurofibrilläres Bündel
NIH	National Institutes of Health
OCT	Optische Kohärenztomografie
OR	Objective Reduction
Orch OR	Orchestrated Objective Reduction
PAGE	Polyacrylamidgelelektrophorese
PET	Postitronen-Emissions-Tomographie
PRKN	Parkin

QFT	Quantenfeldtheorie
RAM	Random Access Memory
ReLU	Rectified Linear Unit
REM	Rapid Eye Movement
ROC	Receiver Operating Characteristic
S1	primärer sensorischer Kortex
SB	synthetisches Bewusstsein
SCORE	Systematic Coronary Risk Evaluation Score
SDS	Sodium Dodecyl Phosphate
SNCA	α-Synuclein
StBG	Österreichisches Strafgesetzbuch
SUV	Standardized Uptake Value
TMS	transkranielle Magnetstimulation
ToE	Theory of Everything
U	Schrödinger-Entwicklung
USA	United States of America
UWP	Ursache-Wirkungs-Repertoire
UWR	Ursache-Wirkungs-Raum
UWS	Ursache-Wirkungs-Struktur
V1	primärer visueller Kortex
WMH	White Matter Hyperintensity

Abbildungsverzeichnis

Tabellenverzeichnis

Zusammenfassung

Einleitung: Das Entstehen eines Bewusstseins als Folge von Prozessen und Mechanismen im menschlichen Gehirn ist innerhalb der Naturwissenschaften noch nicht umfassend geklärt. Physikalische Modelle könnten neue Aspekte der grundlegenden Entstehungsmechanismen des Bewusstseins eröffnen und Wegbereiter zu einem vollständigeren Verständnis des Bewusstseins sein. Die inhaltlich und methodisch verwandten und vorwiegend ingenieurtechnischen Themen „künstliche Intelligenz“ und „synthetisches Bewusstsein“ sind eine sinnvolle Ergänzung in der Erforschung des Bewusstseins. Aus physikalischen Modellen sowie den Erkenntnissen aus dem Forschungsgebiet „künstliche Intelligenz“ und „synthetisches Bewusstsein“ könnten sich Anwendungen für die Medizin in Forschung und Praxis und ein besseres Verständnis speziell von neurologischen und psychiatrischen Erkrankungen ergeben.

Methode: Es erfolgte eine Literaturrecherche in den E-Publikationsdatenbanken „PubMed“ und „ArXiv“. Überdies wurden gedruckte Fachzeitschriften und Bücher zur vertieften Recherche herangezogen. Stets wurde die Glaubwürdigkeit anhand der Richtlinien der „Good Scientific Practice“ überprüft, soweit dies anwendbar war.

Ergebnisse: Physikalische Modelle von Bewusstsein werden exemplarisch eingeführt. Daraufhin wird auf zwei in der aktuellen Forschung besonders relevanten Modelle, nämlich die *Orchestrated Objective Reduction* Theorie des Bewusstseins sowie die *Integrated Information Theory* detailliert eingegangen. Daraufhin wird das Gebiet künstliche Intelligenz behandelt und auch das Thema synthetisches Bewusstsein beleuchtet. Daraufhin erfolgt eine kritische Erläuterung der medizinischen Implikationen der physikalischen Modelle sowie der möglichen Anwendungsgebiete von künstlicher Intelligenz in der Kardiologie und allgemein der Medizin, basierend auf aktuellsten Ergebnissen in der Fachliteratur.

Diskussion: Abschließend wird eine Diskussion der Arbeit von philosophischen, rechtlichen und ethischen Standpunkten geführt. Außerdem werden mögliche zukünftige Entwicklungen rund um das Thema „Bewusstsein“ und „künstliche Intelligenz“ in Medizin und Gesellschaft beschrieben.

Abstract

Background: The emergence of consciousness as a result of complex processes and mechanisms in the human brain is not yet fully understood by science. Physical models could reveal new aspects of the theoretical foundations of consciousness and lead to a fuller understanding of consciousness. The related fields of artificial intelligence and synthetic consciousness are a beneficial addition to a complete treatment of consciousness. The physical models together with insights of the fields of artificial intelligence and synthetic consciousness could be applied to clinical medicine and medical research and may lead to a better understanding of neurological and psychiatric diseases.

Methods: A literature research was performed using the publication-databases „PubMed“ and „ArXiv“. Furthermore printed media like scientific journals and books were used for further investigations. The credibility of the resources was checked for compliance to the good scientific practice guidelines.

Results: A representative set of physical models of consciousness is introduced. Two models, which are especially relevant in current research, are explained in detail. Those are the *Orchestrated Objective Reduction* theory of consciouness and the *Integrated Information Theory*. Then follows a general introduction to artificial intelligence, also covering synthetic consciousness. Afterwards a critical review of the medical implications of the physical models as well as the possible applications of artificial intelligence in cardiology and medicine in general is given, based on current results found in the literature.

Conclusion: The thesis conludes with a discussion of philosophical, legal and ethical perspectives. Possible future developements in the fields of consciousness and artificial intelligence research, especially in the context of medicine and society, are discussed.

1 Einleitung

In jeden Moment unseres bewussten Lebens läuft ein Film vor unserem inneren Auge ab. Er ist ein Ablauf von Momenten des subjektiven und qualitativen Erlebens aus unserer persönlichen inneren Perspektive. Dies sind elementare Empfindungen, etwa, wie es sich anfühlt an einer Blume zu riechen oder Schmerz zu empfinden. Jene Sequenz an Empfindungen ist für jeden einzigartig und steht im individuellen Kontext der eigenen Erinnerungen. All dies beschreibt das, was wir ein „Bewusstsein“ nennen - jene mentale Eigenschaft unseres Geistes, welche allen Dingen in unserem Leben eine Bedeutung gibt. Wir sind sogar in der Lage über unser eigenes Bewusstsein nachzudenken und Aussagen darüber zu treffen, was eines der Grundpfeiler dieser Diplomarbeit ist. In der Einleitung sollen philosophischen Grundbegriffe über das Bewusstsein und seine Eigenschaften dargelegt werden, um in weiteren Abschnitten dieses Thema adäquat behandeln zu können. Um an die moderne Philosophie des Bewusstseins heranzuführen, wird zu Beginn ein historischer Zugang gewählt.

Das Bewusstsein ist etwas fundamental Subjektives - also erschließt sich die Frage, wie es in eine objektive Naturwissenschaft integriert werden kann. Im 20. Jahrhundert erforschten die verschiedenen Disziplinen der Naturwissenschaften das Verhalten von Menschen (und Tieren) sowie die Struktur und Funktion des Gehirns. Heute ist es eine Aufgabe des interdisziplinären Wissenschaftszweiges der Kognitionswissenschaften mit dem Bewusstsein zusammenhängende Prozesse im Gehirn systematisch zu erforschen. Mittels bildgebender und elektrophysiologischer Verfahren können neuronale Korrelate des Bewusstseins identifiziert und Hirnareale zu bestimmten Bewusstseinsprozessen, vor allem in Bezug auf Wahrnehmung, zugeordnet werden. Somit lässt sich feststellen, welche Gehirnareale gleichzeitig mit bewussten Erfahrungen wie Schmerz, Freude etc. aktiviert werden[1].

Allein Korrelate erklären jedoch nicht, wieso komplexe Gehirnprozesse von Bewusstsein begleitet werden. Verschiedenste Modelle nutzen Ansätze aus der Physik, Neurowissenschaft und Informatik. In dieser Diplomarbeit werden ausgewählte Modelle erläutert und danach auf zwei bedeutende Modelle vom Bewusstsein besonders eingegangen. Die von Roger Penrose und Stuart Hameroff ins Leben gerufene Theorie, Orchestrated Objective Reduction (Orch OR), beschreibt wie in den Mikrotubuli

[1] Beispielsweise können Verletzungen des Farbzentrum im visuellen Kortex (Gyrus lingualis und fusiformis) dazu führen, dass betroffene Personen in einem visuellen Halbfeld in Schwarz-Weiß träumen [1].

(MT) der Neuronen informationsverarbeitende Prozesse, welche möglicherweise für das Bewusstsein verantwortlich sind, ablaufen. Einen anderen Ansatz wählt Giulio Tononi mit der Integrated Information Theory (IIT), welche das Bewusstsein als Information auffasst und Systeme, wie zum Beispiel das Gehirn, hinsichtlich seiner integrierten Information quantitativ analysieren kann. Die integrierte Information wird mit dem Bewusstsein in einem System gleichgesetzt. Medizinische Implikationen der beschriebenen Modelle vom Bewusstsein werden anschließen exemplarisch beschrieben. Vor allem in den Bereichen der Neurologie und Psychiatrie können experimentelle Befunde durch die IIT oder durch die Orch OR erklärt und neue Erkenntnisse gewonnen werden.

Daraufhin soll der Bogen zur künstlichen Intelligenz (KI) und dem maschinellen Lernen geschlagen werden. Mit den Methoden der KI, welche vom Aufbau neuronaler Schaltkreise des Gehirns inspiriert sind, können spezifische kognitive Funktionen nachgebildet und erforscht werden. Beispielsweise ist mit Algorithmen aus dem Gebiet des Deep Learnings das zuverlässige Erkennen von Mustern, das Extrahieren von Information aus großen Datensätzen oder das Lösen von komplexen Problemen möglich. In dieser Diplomarbeit wird auf die Messung von KI, dessen Verhältnis zu synthetischem Bewusstsein sowie den Algorithmen des Deep Learnings eingegangen. Mögliche Anwendungsgebiete von maschinellem Lernen in der Kardiologie, für Klinik und Forschung, werden erörtert. Von der automatischen Elektrokardiogramm (EKG)-Befundung bis zu computergestützten epidemiologischen Studien befinden sich bereits einige vielversprechende Ansätze in Entwicklung.

Abschließend erfolgt eine Diskussion der physikalischen Modelle von Bewusstsein im philosophischen und medizinischen Kontext. Die Zukunft der KI, speziell auch in der Medizin, ist Thema einer weiteren Diskussion. Eine Behandlung von ethischen Fragestellungen bezüglich der KI und dem SB in unserer Gesellschaft schließt die Diplomarbeit ab.

1.1 Platons Seele

Betrachtungen über die Seele des Menschen waren bereits in der Antike Stoff für philosophische Diskussionen. Eine der früheren Schriften von Platon, Phaidon [2], spiegelt die Ansichten des antiken Philosophen über die Natur der Seele wieder. In dieser, in Dialogform verfassten, Erzählung über den Tag der Hinrichtung des Sokrates, unterhält sich dieser in seinem Gefängnis ein letztes Mal mit seinen Schülern und Freunden. Im

Zuge des Gespräches über die Ansichten von Sokrates über den Tod kommt die Seele zur Sprache. Sokrates spricht mit einem seiner Schüler, Kebes (79a-79b[2]):

Sokrates: *Sollen wir also zwei Arten der Dinge setzen, sichtbar die eine und die andere unsichtbar?*

Kebes: *Das wollen wir.*

Sokrates: *Und die unsichtbare als immer auf gleiche Weise sich verhaltend, die sichtbare aber niemals gleich?*

Kebes: *Auch das wollen wir setzen.*

Sokrates: *Wohlan denn ist nicht von uns selbst das eine Leib und das andere Seele?*

Kebes: *Allerdings.*

Sokrates: *Welcher von jenen beiden Arten nun wollen wir sagen, dass der Leib ähnlicher sei und verwandter?*

Kebes: *Das muss ja jedem deutlich sein: dem Sichtbaren.*

Sokrates: *Wie aber die Seele, ist die unsichtbar oder sichtbar?*

Kebes: *Menschen wenigstens ist sie es nicht, o Sokrates.*

Platon äußert sich in einigen seiner Schriften, zum Beispiel in Phaidon, Phaidros, Philebos, Kratylos, Politeia, Symposium und Menon, über die Seele. Nach ihm sei die Seele von geistiger, immaterieller Natur und somit substanziell unterschiedlich zum Körper. Darüber hinaus stellt er weitere Eigenschaften der Seele und ihre Beziehung zum Körper fest, welche in den folgenden drei zentralen Argumenten hervortreten. [3]

Argument aus den Gegensätzen

In Phaidon spricht Sokrates davon, dass die Seele nach dem Tode eines Körpers sich von diesem loslöse und sich „in sich sammle“, wobei diese dann sich dem „reinen Denken“ am besten zuwenden könne (66b-67b). Hier wendet Kebes ein, dass nicht begründet sei, weshalb die Seele nach dem Tod eines Körpers stattdessen nicht „wie ein Hauch zerstebe“. Er fragt danach, wie die Seele, also eigenständig und vom Körper losgelöst, in der Lage sein soll, Kraft und Denkvermögen zu besitzen (70a-70b). Sokrates versucht eine Erklärung in drei Teilen.

[2]Textzitate aus den Schriften Platons nach der Stephanus-Paginierung

a) Prinzip der Gegensätze: Das Schöne ist gegenüber dem Hässlichen gegensätzlich; Das Gerechte ist gegenüber dem Ungerechten gegensätzlich. Diese Gegensatzpaare entstünden, laut Sokrates, aus einander. Daher entstehen gegensätzliche Dinge immer aus Gegensätzlichem. (70e-72b)

b) Prinzip der Reziprozität: Jedem Gegensatzpaar ist ein Paar an Werdeprozessen zugeordnet. Die Werdeprozesse beschreiben den Übergang von einem Zustand zu seinem gegensätzlichem Zustand. Als Beispiel sei hier das Kaltwerden und das Warmwerden, als Übergang von warm zu kalt und umgekehrt, genannt. (71a-71b)

c) Kreislauf des Werdens: Die Werdeprozesse, welche einem Gegensatzpaar zugeordnet sind, sind kreislaufförmig. Dies bedeutet, dass nicht nur ein Werden von einem Zustand A zu seinem Gegenteil B alleine möglich, sondern auch wieder der Übergang zurück von B zu A. (72a-72b)

Die hier dargestellten Zusammenhänge werden in Phaidon nicht weiter begründet, sondern als Tatsachen angenommen. Aus ihnen kann Sokrates nun grundlegende Eigenschaft der Seele ableiten. So postuliert er das Gegensatzpaar Leben und Totsein in Analogie zum Wachsein und Schlafen. So wie am Morgen auf das Schlafen das Wachsein folgt, kommt am Abend nach dem Wachsein wiederum das Schlafen. Den gegensätzlichen Zuständen sind, sozusagen in einem Zyklus, zwei Werdeprozesse zugeordnet, nämlich das Aufwachen und das Einschlafen. Gleichsam entstehe das Lebende aus dem Gestorbenen und umgekehrt (71d). Das Sterben ist das Loslösen der Seele vom Körper und das Wiederaufleben der Eintritt der Seele in den Körper. Aufgrund der Annahme der Kreislaufförmigkeit (Punkt c) ist nun ein Weiterbestehen der Seele nach dem Tod offensichtlich[3]. Die Seelen der Gestorbenen müssen sich demnach an einem Ort aufhalten, bis diese sich wieder mit einem Körper vereinen. (72a)

Zudem führt Sokrates ein weiteres Argument an, welches an dieser Stelle anknüpft. Das Sein einer Sache sei durch gewisse Eigenschaften bestimmt (zum Beispiel das Heißsein von Feuer, oder das Geradesein von geraden Zahlen), wodurch nun dieser Gegenstand nicht die gegensätzliche Eigenschaft annehmen kann, da sozusagen seine Identität davon abhängt. Da die Seele immer das Leben mit sich bringt und quasi das Lebendigsein als bestimmende Eigenschaft in sich trägt, kann sie niemals das Totsein annehmen. Das macht die Seele unsterblich. (78d-78e)

[3]Hier sei angemerkt, dass aufgrund einer bloßen Annahme nichts offensichtlich sein kann.

Anamnese Argument

Sokrates sagt, dass das Lernen nichts anderes als ein Erinnern sei (72e). Wenn jemand etwas lernt oder erfährt, so geschieht dies in weiten Teilen zum ersten Mal im Leben. Daher muss die Seele schon zuvor von den Dingen gewusst haben, damit sie sich nun daran erinnern oder wiedererkennen kann[4]. Demnach muss alles, was erlernbar und erfahrbar ist, schon zuvor in der Seele enthalten sein. Dies wiederum soll die Existenz der Seele vor der Geburt nach sich ziehen. Die Aspekte dieser Argumentation lassen sich folgendermaßen zusammenfassen:

a) Die Voraussetzung für ein Erinnern ist das vorherige Wissen (73c). Beispiel: Jemand erkennt auf einem Bild von Simmias die Person wieder.

b) Beim Erkennen von einer Sache, können zusätzlich zum Erkennen gedankliche Assoziationen auftreten - dies sei ebenfalls ein Wiedererinnern (73c). Beispiel: Bei der Betrachtung des Bildes von Simmias wird jemand an den Kebes erinnert, welcher ein Freund von Simmias ist.

c) Nach Sokrates ist nun diese Assoziation nicht dem Wissen über die Sache an sich entsprungen. Es soll eine Fähigkeit der Seele sein, diese Assoziationen mitzubringen. (73b-74a) [5]

d) Des Weiteren führt Sokrates die Unterscheidung zwischen „gleichen“ und „ungleichen“ Dingen ein. In diesem Sinne ist ein Bild von einer Sache ungleich der Sache an sich. [5] Diese Einsicht über die Ungleichheit zweier Wahrnehmungen (wie die Wahrnehmung von der Sache und ihrem Bild), erfordert die bereits zuvor vorhandenen Kategorien „gleich“ und „ungleich“, daher eine konkrete Idee Vorstellung vom „Gleichsein“, um zu wissen, dass diese beiden Dinge nahezu, jedoch nicht gänzlich gleich sind (74c-74d). Diese Idee soll die Seele mit sich bringen. [6]

Verwandtschaft mit dem Unvergänglichen

Dieser Gedankengang von Sokrates beginnt mit der Feststellung, dass materielle Dinge, wie der Körper, aus kleinen Teilen zusammengesetzt sind, und daher auflösbar und vergänglich sind (78c). Im Gegenzug dazu stehen die „Ideen“ im Platon'schen Sinne,

[4]Im modernen Verständnis der Genetik im Kontext von Verhaltensbiologie und Entwicklungspsychologie gewinnt diese Erkenntnis zusätzlich Bedeutung. Die Verhaltensgenetik beschreibt den Einfluss der vererblichen Gene auf das Verhalten. Hier besteht eine Analogie zum „Wiedererinnern“ im Platon'schen Sinne, jedoch nicht durch die immaterielle Seele, sondern durch eine materielle Vererbung. [4]

welche eine Einheit bilden und nicht aus Teilen zusammengesetzt sind. Die Ideen lassen sich daher nicht auflösen und sind unvergänglich (80b-80c). „Nun verhalte es sich so, dass eine materielle Sache eine Beschaffenheit haben kann, welche einer Idee entspricht." Als Beispiel soll eine schöne Blume, welche das Schöne als Idee in sich trägt, dienen. Es kann gesagt werden, dass die Blume eines Tages verwelkt und nicht mehr schön sein wird, jedoch die Idee des Schönen ist unvergänglich.[5] An diesem Punkt versucht Sokrates die Verbindung der Seele mit den Ideen über die gemeinsamen Eigenschaften (Unsichtbarkeit, Unveränderlichkeit) herzustellen. Daraus schließt er letztlich auf die Unauflösbarkeit und somit die Unsterblichkeit der Seele (80b-80e). [5]

Zusammenfassend lässt sich sagen, dass Platon die Seele als reale, unvergängliche und immaterielle Entität innerhalb des Gedankenkonstrukts seiner „Ideenlehre" festmacht. Wie in Platons berühmten „Höhlengleichnis" anschaulich dargestellt, wird die Materie ein unvollständiges Abbild der immateriellen Ideen angenommen (514a-520e *Politeia*). Genauso wie die Ideen, so sei die Seele aus einer zur Materie kategorisch unterschiedlichen Substanz, was letztlich zur Konzeption eines später noch diskutierten Substanzdualismus führt. Für andere Aspekte der Platon'schen Seelenlehre sei auf die weiterführende Literatur verwiesen [7].

1.2 Descartes Ansichten zum Geist

René Descartes kann als einer der einflussreichsten Denker auf dem Gebiet der Philosophie des Geistes betrachtet werden. Seine Schriften aus der Mitte des 17. Jahrhunderts gelten als Klassiker in vielen Bereichen der Philosophie. Die folgende Darlegung der Gedanken Descartes zum Geiste in seinen berühmten „Meditationen", (sowie der Einwände seiner Zeitgenossen und der darauf folgenden Antworten Descartes) [8] soll sowohl als geschichtlich aufschlussreichen Einstieg, als auch als erste Annäherung an die großen Themen der heutigen Philosophie des Geistes[6], welche auch das Bewusstsein zum Thema hat, dienen.

Am Beginn der zweiten Meditation fragt Descartes danach, ob es etwas gibt, dessen Existenz nicht in Zweifel gezogen werden kann. Er argumentiert, dass dieses unbezweifelbare Eine, als stabiles Fundament für alle weiteren Gedankenschritte dienen kann. Um zu dieser „Sicherheit" gelangen zu können, möchte Descartes alle jenes aus seiner

[5]Das Beispiel mit der Blume ist hier als Analogie zu Platons Leier angeführt. Die Leier kann zwar zerbrechen, jedoch die Harmonie in ihrer Musik würde dadurch nicht berührt werden. (85e-86a)

[6]Die Philosophie des Geistes beschäftigt sich mit grundlegenden Fragestellungen des Bewusstseins, des freien Willens, der Kognition und angrenzende Gebiete [9].

Argumentation ausschließen, was „auch nur den leisesten Zweifel zulässt“ (24[7]). Zu misstrauen sei unter anderem dem Gedächtnis und den Sinnen, da diese einer möglichen Täuschung unterliegen. Nun, fragt er, wenn die Existenz von allen erinnerlichen und mit den Sinnen wahrgenommen Dingen in Zweifel gezogen wird, muss dann ebenfalls die eigene Existenz bezweifelt werden? Hier antwortet Descartes entschieden mit Nein: Wenn jemand/etwas in der Lage ist, über solche Fragen nachzudenken, dann sei seine Existenz gewiss. Daraus ergibt sich in der Vollendung dieses Gedankens das in [10] so genannte „Existo$^+$-Argument“, das hierin als intuitiver und deduktiver Schluss angenommen wird:

> Existo$^+$: *Die Denkerin/der Denker dieses Gedankens hat diesen Gedanken und folglich gilt: sie/er existiert.*

Das Wesen des Selbst

Kurz darauf schließt Descartes mit einer Diskussion der Natur dieses „Denkers“ an. Es wird festgestellt, dass jenes, was denkt (die Denkerin/der Denker), eine Substanz ist, die Gedanken hat (res cogitans). Wobei Descartes hier die Anwendbarkeit des Substanz-Begriffes unbegründet annimmt ([11] S. 215). Das „Denken“ sei die (epistemische) Essenz dieser Substanz (der Geist, Seele), also seine definierende Eigenschaft[8]. Als Moden des Denkens bezeichnet Descartes grundlegende geistige Aktivitäten, wie einen Teil der Wahrnehmung (siehe weiter unten), die Vorstellungskraft, der Intellekt und auch der freie Wille [78], und diese seien daher genauso existent wie der Geist selbst. Jedoch gilt dies nicht für jenes, was wahrgenommen wird, jenes was vorgestellt wird - die Inhalte des Denkens sind nach wie vor dem Zweifel ausgesetzt[9].

Eine weitere Eigenschaft, welche Descartes dem Geiste zuschreibt, ist seine Einheit ([10] S. 124). Der Geist mag viele verschiedene Gedanken hervorbringen, so bleibt er doch ein und die selbe Sache über die Zeit hinweg und ist im Gegensatz zum Körperli-

[7] Seitennummerierung in den Texten von Descartes nach Adam und Tannery, Œuvres de Descartes, Band VII

[8] Mit dem reinen Verständnis können über die Vorstellungskraft hinausgehende Dinge begriffen werden. Zum Beispiel ist es nicht möglich sich ein Tausendeck vor dem inneren Auge als Bild vorzustellen, jedoch kann der reine Verstand es mithilfe seine Eigenschaften erfassen (72). Der reine Verstand/Intellekt ist eine Fakultät des Geistes alleine, wohingegen die Vorstellungskraft eine Funktion der Geist-Körper-Einheit ist (73). Hier sei anzumerken, dass nach Descartes jegliche Inhalte des Intellekts von zuvor passierten Wahrnehmungen abgewandelt sind (75-76), was wiederum die Bedeutung des Zusammenspiels von Geist und Körper hervorhebt.

[9] Es mag so scheinen, als hätte jemand etwas mit den Sinnen wahrgenommen, jedoch kann es nicht gewiss als real und existent betrachtet werden. Es könne lediglich gesagt werden, dass es so scheint, als wäre etwas real oder existent. (369) Über die Fähigkeit des menschlichen Verstandes zu irren und dessen Zusammenhang mit Gott ist vor allem Gegenstand der vierten Meditation (52-62).

chen nicht in Teile zerlegbar. Gegenüber stellt er den Körper (res extensum)[10] , welcher eine endliche räumliche Ausdehnung hat und teilbar ist. Obgleich Geist und Körper eine Einheit bilden, so können diese separat voneinander existieren (78) und eine Substanz kann nicht durch die andere erklärt werden (keine Emergenz, keine Reduktion). Allerdings schreibt Descartes der Einheit Eigenschaften zu, welche weder in der einen, noch in der anderen Substanz alleine gefunden werden können, also nur im Zusammenspiel von Geist und Körper entstehen. Beispielsweise sind die eigentliche Wahrnehmung und die Vorstellungskraft eine Funktion/Fakultät der Geist-Körper-Einheit (81).

Qualia

Nach Descartes ist die Natur der Geist-Körper-Einheit für uns nicht klar erkennbar, jedoch werden wir durch subjektive Empfindungen wie Schmerz, Hunger, Angst ihrer Existenz bewusst (bzw. des Umstandes, dass der Geist und der Körper miteinander verwoben sind) (692). Wenn wir längere Zeit nichts essen, verspüren wir Hunger - es ist ein subjektives Gefühl von Hunger, welches eine körperliche Ursache und eine geistige Manifestation hat. Nach Descartes ist das Vorhandensein dieser Empfindungen eine Konsequenz der Einheit - wenn jemand die beiden Substanzen Körper und Geist als getrennt betrachtet, so würde der Intellekt den Mangel an Nahrung als eine objektive Tatsache begreifen, sozusagen als Signal des Körpers, dass er Nahrung benötigt. Eine subjektive qualitative Empfindung, wie Hunger, wäre redundant. Hier führt Descartes eine Vergleich mit einer Seglerin/einem Segler und ihrem/seinem Segelboot an. Zwar steuert die Seglerin/der Segler das Boot und kann sie/er etwaige Schäden am Rumpf und dergleichen erkennen, jedoch hat sie/er dabei nicht dieselben Empfindungen, so als wenn sie/er selbst verletzt sei. Der Geist ist also mehr als eine Seglerin/ein Segler im Kopf, welche/welcher mit dem Körper interagiert - sie/er ist vielmehr integriert. (81)

Mentale Verursachung

Die Frage, welche sich im Zuge dieser Diskussion aufzwingt und hier von zentraler Bedeutung ist, ist folgende: Wenn der Geist Einfluss auf den Körper nimmt, und umgekehrt der Körper auf den Geist wirkt, wie funktioniert diese gegenseitige Einflussnah-

[10]Descartes stellt im Zuge seine Abhandlung die Frage, was der Verursacher der sinnlichen Wahrnehmung ist. Er schlägt als Ursache als Kandidaten vor: real existierende körperliche Substanz, welche als solche von den körperlichen Sinnen wahrgenommen werden kann ODER die sinnliche Wahrnehmung wird durch Gott in uns erzeugt (79). Descartes schließt hier Gott (oder einen Engel mit ähnlichen Fähigkeiten) als direkte Ursache aus, da Gott in uns nicht die Fähigkeit gelegt hat, die immaterielle Natur der wahrgenommenen Dinge zu erkennen und Gott ist kein Täuscher (79-80). Er geht also davon aus, dass körperliche Dinge existieren und eine eigene Substanz sind (neben dem Geiste) (80).

me? Descartes geht zwar nicht näher auf die Details der Wechselbeziehung ein (scheint selbst keine Erklärung dafür zu haben[11]), jedoch nennt Descartes eine anatomische Struktur im Gehirn, welcher der Ort der Wechselwirkung sein sollte. Da der Geist unteilbar ist, könne dieser nur in einer einzelnen Struktur körperlich manifest werden und nach Descartes komme im Gehirn nur die Epiphyse als unpaarig, einzeln und mittig liegend in Frage. [13]

Analyse der Wahrnehmung

Nach Descartes läuft die Wahrnehmung in drei Abstraktionsschritten ab. Im zeitlichen Ablauf zuerst steht der rein körperlich-materielle Prozess, welcher zu einer Reizung von Sinnesorganen und Nerven/Gehirn führt. Dies bewirke als zweites im Gehirn einen „undeutlichen“ und „nicht-reinen“ Gedanken, der eine diffuse Sinneswahrnehmung darstellt, sozusagen als unwillkürliche Projektion der Sinneswahrnehmung. Descartes bezeichnet dies als „eigentliche Empfindung“, als etwas „sinnliches“. Als letzter Schritt in der Abstraktion folgt aus dem bisherigen der „reine“ Gedanke des Intellekts - und nur das ist Teil des Intellekts. Dies könnte beispielsweise das bewusste Fassen und Beurteilen des Wahrgenommenen sein. Vor allem in diesem finalen Akt des Intellekts, kann ein Fehlurteil passieren, da hier das (fehlbare) Schließen von der Wahrnehmung auf einen „realen“ Sachverhalt passiert. Nach Descartes ist der Urteilsapparat des Intellekts eher auf das „Überleben“ (zweckmäßig) ausgerichtet, als auf das objektive Erfassen der wahren Begebenheiten. (80-83) [10]

Seelen der Tiere

In der Cartesianischen Philosophie haben Tiere keinen Intellekt, daher sind „eigentliche Empfindungen“, als Funktion des Intellekts, in Tieren nicht vorhanden. So bleibt nur eine Konsequenz, nämlich Tiere als seelenlose Maschinen zu betrachten. Jegliches Verhalten der Tiere lässt sich auf rein körperlichen Ursachen reduzieren - ein inneres Empfinden (Qualia) wird ihnen abgesprochen[12] ([17] 56). Für seine Sichtweise erhielt

[11] Möglicherweise hatte Descartes eine okkasionalistische Ansicht bezüglich der mentalen Verursachung. Im Okkasionalismus stellt Gott als eingreifende Kraft den Zusammenhang zwischen dem Geist und dem Körper her (mentale Verursachung). Wünscht jemand die Hand zu heben, so setzt Gott die Muskeln in Bewegung. Genauso wird von außen, aufgrund körperlicher Ursache, auf den Geist gewirkt. Der Okkasionalismus ist heute unbedeutend. [12]

[12] Dem diametral entgegengesetzt steht die Auffassung des *Great Ape Projects* [14], welches sich für die Rechte von Menschenaffen einsetzt. In seinem Buch [15], argumentiert einer der Begründer des Great Ape Projects, Peter Singer, dass nicht etwa die zoologische Einteilung in Spezies entscheidend sei, sondern der Grad des (Selbst-)Bewusstseins und der „Status als Person“, wie Menschen Tiere betrachten und mit ihnen umgehen sollten. Daher plädieren Singer und seine Anhänger für mehr Rechte

Descartes von seinen Zeitgenossen Kritik in den Einwänden zu den Meditationen, zum Beispiel Pierre Gassendi (270ff) und Antoine Arnauld (205).

1.3 Bewusstsein

Es ist eine weit verbreitete Ansicht innerhalb des Fachgebietes der Philosophie des Geistes, dass eine Definition von Bewusstsein zu Zirkelschlüssen führt[13] [19] [20]. Dieser Ansicht folgend, soll in diesem Kapitel statt einer Definition der Versuch unternommen werden, das Thema „Bewusstsein“ anhand von Beschreibungen, Beispielen und Kriterien inhaltlich abzustecken.

Der Bewusstseinsbegriff kann sowohl als globales Phänomen von Systemen und Organismen, als auch als Eigenschaft von einzelnen mentalen Prozessen aufgefasst werden. Aspekte des Bewusstseins können in diesen Kontexten diskutiert und somit eine Annäherung an eine Beschreibung vom Bewusstsein gemacht werden. Folgende Einteilung nach [21] steckt Eckpunkte des Begriffs *Bewusstsein* ab:

1.3.1 Bewusstsein von Organismen, Personen und Systemen

- Eine bewusste Entität hat ein Empfindungsvermögen oder eine Wahrnehmung, durch welche sie die Eindrücke seiner Umgebungswelt gewinnen und auf diese reagieren kann [22]. Welchen Grad an Empfindungen für das Vorliegen von Bewusstsein nötig ist, bleibt jedoch eine offene Frage.

- Wachheit bzw. der Zustand des Wach-Seins sei ein weiteres Kriterium für das Bewusstsein. Dies zieht vor allem in Hinblick auf das Nicht-Wach-Sein oder dem tiefen Schlaf die Abwesenheit von Bewusstsein mit sich. Träumen, quantitative Bewusstseinsstörungen wie Somnolenz oder Sopor sowie hypnotische Zustände liegen im Graubereich zwischen bewusst und nicht bewusst. [21]

- Ein eher hartes Kriterium ist das Selbst-Bewusstsein als eine Variante der Selbstbeobachtung. Hiermit ist gemeint, dass eine Entität nicht nur Wahrnehmungen

für Orang-Utans, Gorillas und Schimpansen. Eine weitere Organisation, das *Nonhuman Rights Project*, setzt sich seit 2007 für die Rechte von Tieren ein. Auf Wirken des Honhuman Rights Project erkannte der Supreme Court im Jahre 2015 in Manhattan/New York, USA, zwei Schimpansen als juristische Person an. Dadurch konnten diese in weitere Folge aus der Gefangenschaft an einer Universität, wo an ihnen biomedizinische Experimente durchgeführt wurden, freigelassen werden. [16]

[13] Ein Zirkelschluss, auch Hysteron-Proteron genannt, ist ein Fehler in der logischen Argumentationsführung. Er bezeichnet die Annahme dessen, was eigentlich gezeigt werden sollte. Zum Beispiel beinhalten folgende Aussagen eine Zirkelschluss: *Ich sage stets die Wahrheit. Daher lüge ich nie.* [18]

und Empfindungen hat, sondern sich über diese im Klaren ist. Der Grad dieser Klarheit über seine eigenen Empfindungen mag variieren - es könnte ein Tier, welches vermutlich eher ein rudimentäres Verständnis über das eigene Empfinden hat, dennoch als (abgeschwächt) bewusst bezeichnet werden. [23]

- Subjektives Empfinden ist ein zentraler Aspekt des Bewusstseins. Das Erleben von Eindrücken, über eine äußere oder innere Wahrnehmung, hat aus der inneren Perspektive eine subjektive Quantität und Qualität. Elementare Empfindungen des Bewusstseins, wie das Gefühl beim Betrachten einer Farbe oder beim Empfinden von Schmerz, werden als *Qualia* bezeichnet. [24]

- Zusätzlich zur mentalen Repräsentation eines Objektes (transitives Bewusstsein), ist auch ein Erlebnis, welches mit dieser Repräsentation des Objekts verknüpft ist (intransitives Bewusstsein), maßgeblich für das Bewusstsein in Bezug auf dieses Objekt. [25]

- Ein weiteres mögliches Kriterium für Bewusstsein ist die Identifikation der mentalen Zustände einer Entität als bewusste Zustände. Ein Bewusstsein wird als aus elementaren bewussten Prozessen oder Zuständen zusammengesetzt betrachtet. [21] Die Eigenschaften von bewussten Zuständen werden im nächsten Abschnitt erläutert.

1.3.2 Bewusstsein von mentalen Zuständen

Was macht bewusste mentale Zustände bewusst? Folgende Aspekte mentaler Zustände sollen deren bewussten Charakter näher beschreiben und analog zur obigen Auflistung als Eckpunkte dienen.

- Sich darüber im klaren zu sein, dass jemand sich in einem mentalen Zustand befindet, entspricht einem weiteren und anderen mentalen Zustand. Letzterer kann in diesem Zusammenhang als meta-mentaler Zustand bezeichnet werden und diese Art von Zuständen könnte mit Bewusstsein assoziiert werden. Wenn jemand einen bewussten Wunsch nach Süßigkeiten hat, bedeutet dies, dass erstens der Wunsch nach Süßigkeiten als mentaler Zustand vorliegt und zweites, dass es einen weiteren mentalen Zustand, der sich auf den ersten bezieht, gibt. [26]

- Eine weiteres Merkmal von bewussten mentalen Zuständen ist deren qualitative Eigenschaft, welche mit *Qualia* assoziiert werden. Mit diesen mentalen Zustän-

den sind die inneren Erfahrungen im Zusammenhang mit Gefühlen, Wahrnehmungen, Gedanken, Wünsche usw. gemeint. Der Genuss eines Stücks Schokolade kann einen mentalen Zustand, entsprechend einer Geschmacks-Quale - nämlich genau jene, welche jemand beim Genuss von Schokolade hat - hervorrufen. Diese Form des „erlebenden" Bewusstseins wird *phänomenales Bewusstsein* genannt. Das phänomenale Bewusstsein beinhaltet außerdem noch das Verständnis der Welt und des Selbst als Teil der Welt. ([27] S. 10)

- Das *Zugriffsbewusstsein* (Z-Bewusstsein) nach Ned Block grenzt sich vom phänomenalen Bewusstsein (P-Bewusstsein) scharf ab - Interaktionen zwischen Z- und P-Bewusstsein seien jedoch zugelassen. Als Z-bewusst werden mentale Zustände bezeichnet, welche einer Repräsentation (transitiv) entsprechen und abrufbar sind für: (1) das logische Denken, (2) die Kontrolle des Verhaltens, (3) die rationale Kontrolle der Sprache. Die Eigenschaft z-bewusst zu sein ist unabhängig davon *wie es sich anfühlt* in diesem mentalen Zustand zu sein (intransitiv). [19]

  PatientInnen mit Verletzungen im primären visuellen Kortex ($V1$) können totale Skotome in Bereichen des Gesichtsfeldes haben - sie geben an in diesen Bereichen „nichts" zu sehen. Es wurde jedoch gezeigt, dass gewisse Eigenschaften, wie Form, Farbe, Größe usw., von Objekten, welche sich im „blinden" Bereich befinden, reproduzierbar „erraten" werden können. Dieses Phänomen wird „Blindsehen" genannt. [28] Die mentale Repräsentation der Wahrnehmung in diesem Bereich scheint nicht Z-bewusst, jedoch möglicherweise P-bewusst zu sein [19].

1.4 Intentionalität

Der Begriff Intentionalität (mit seiner heutigen Bedeutung) geht auf Franz Brentano [29] zurück und bezeichnet den Bezug von mentalen Zuständen auf einen repräsentativen Inhalt. In anderen Worten sind intentionale Zustände „über" etwas oder auf etwas bezogen.

Nach John Searle haben intentionale Zustände zusätzlich zu ihrem repräsentativem Inhalt einen sogenannte *psychologischen Modus*, wie Überzeugungen, Ängste, Hoffnungen, Wünsche, Vorhaben und dergleichen. Folgender Satz soll den Zusammenhang illustrieren:

Ich möchte eine Tasse Kaffee trinken.

Hier ist der psychologische Modus das „Wollen“ und der repräsentative Inhalt das „Kaffee trinken“. Ein mentaler Zustand, welcher dem Wunsch im genannten Sinne entspricht hat Intentionalität. [30]

Eine weitere Analogie wäre das bekannte Gemälde von René Magritte (Abbildung 1), worauf eine Pfeife abgebildet ist. Das Bild selbst ist nichts anderes als eine Leinwand, auf welche ein Muster mit Öl in verschiedenen Farben aufgetragen wurde. Bei dem Muster handelt es sich um den repräsentativen Inhalt dieses Bildes. Dieser bezieht sich auf eine Pfeife. Das Bild selbst jedoch ist nicht die Pfeife. Außerdem interessant ist der Aspekt, dass mentale Zustände auf Objekte, welche nicht existieren, gerichtet sein können ([31] Abschnitt 2). Weitere Diskussion zur Nicht-Existenz ([32] Teil 1).

Abbildung 1: Dieses Ölbild von René Magritte (La trahison des images, 1929) dient als Analogie zu einem intentionalen Zustand. Das Bild selbst repräsentiert ein Objekt, in diesem Fall eine Pfeife, es ist jedoch nicht mit dem Objekt identisch. Dadurch hat das Bild im übertragenen Sinne *Intentionalität*.

Der reduktive Physikalismus ist äquivalent mit der Ansicht, dass alle elementaren mentalen Prozesse und Zustände sich auf elementare physische Prozesse und Zustände „reduzieren“ lassen. Die Frage, wie sich Intentionalität als Eigenschaft mentaler Zustände auf die physische Ebene im Gehirn übertragen lässt bzw. was das neuronale Korrelat von Intentionalität sein könnte, ist bislang ein ungelöstes Problem ([33] S. 262 - 272). Manche Vertreter dieser Ansicht schlagen vor, dass das Konzept der Intentionalität und im weitesten Sinne der mentalen Zustände nicht nötig sei, um die Funktion des Gehirns vollständig zu beschreiben [34].

1.5 Körper-Geist-Beziehung

1.5.1 Erklärungslücke

Die sogenannte Erklärungslücke (explanatory gap) bezieht sich auf die bisher fehlende Erklärung von Qualia durch physikalisch-naturwissenschaftliche Theorien [35]. Diese

Lücke ist gleichbedeutend mit dem sogenannten „hard problem“ des Bewusstseins [36]. Zur Überbrückung der Erklärungslücke und zur Beschreibung der Beziehung zwischen Körper und Geist wurden bisher verschiedene Ansätze vorgeschlagen. John Searle kritisiert die Einteilung der Meinungslandschaft innerhalb der Philosophie des Geistes anhand der so genannten „Ismen“. Die verschiedenen Denkrichtungen ziehen nämlich mit ihren Grundaussagen oft unausgesprochen Annahmen mit sich, welche als „Ballast“ in einer unvoreingenommenen Diskussion wirken und zu inkorrekten Argumenten führen können ([37] S. 1 - 7)

- Qualia und ähnliche mentale Phänomene befinden sich außerhalb des Physischen bzw. Materiellen. Wenn eine wechselseitige Interaktion[14] zwischen den beiden Bereichen, also dem mentalen und dem physischen, zugelassen ist, würde dies der Sicht des **Substanzdualismus** entsprechen. Falls lediglich eine einseitige Wirkung des Physischen auf den mentalen Bereich zugelassen wird, folgt daraus, dass das Mentale und somit das Bewusstsein ein **Epiphänomen** ist. ([39] Kap. 3)

- Jeder elementare mentale Zustand, jede Quale, lässt sich auf physikalische Zustände reduzieren. Dies ist die Grundaussage des **reduktiven Physikalismus**. Eine Abwandlung dessen ist der *Funktionalismus*, welcher eine Äquivalenz von mentalen Zuständen zu funktionellen Zuständen postuliert. ([39] Kap. 5)

- Qualia sind emergente Eigenschaften von physischen Zuständen und lassen sich nicht auf elementare physische Prozesse reduzieren. Mit dieser Annahme vereinbaren Denkrichtungen wäre der **nichtreduktive Physikalismus**. Ähnlich dazu beschreibt der **Eigenschaftsdualismus** mentale Zustände als nicht-materielle Eigenschaften von dem Physischen. Eine Abwandlung des Eigenschaftsdualismus ist der **Panpsychismus**, welcher allen materiellen Dingen mentale Eigenschaften zuschreibt. ([39] Kap. 4 und 8)

- Qualia und mentale Zustände sind genauso wie physische Zustände Eigenschaften einer grundlegenden Substanz. Im **neutrale Monismus** erübrigt sich somit die Frage, ob mentale Zustände aus physischen hervorgehen. Es ist eher so, dass beide gleichberechtigte Aspekte einer einzelnen Sache sind. [40]

[14] Eine Interaktion der physikalischen Welt mit der mentalen Welt ist zusammen mit der kausalen Geschlossenheit der Physik [38] diskutabel.

- Die Physik ist derzeit unvollständig und eine Erweiterung innerhalb des physikalischen Gedankengerüstes ist vonnöten, damit eine Erklärung von Bewusstsein möglich wird. Ein nahe verwandtes bezieht sich auf die kausale Geschlossenheit der Physik und sagt, dass eine nicht-physikalische Wirkung geben kann [38].[15] ([39] Kap. 3)

- Im **Idealismus** ist alles was uns umgibt mental. Die für uns sichtbare physische Welt ist eine Manifestation des Mentalen. ([41] S. 38)

- Mentale Zustände und daher auch Qualia existieren nicht und müssen nicht erklärt werden. Im **eliminativem Physikalismus** sind mentale Zustände vielmehr ein Konzept aus unserem Anschauungsraum, haben jedoch keine unabhängige Realität. ([39] Kap. 6)

1.5.2 Qualia-Argumente

Wie *fühlt es sich an* eine Leinwand mit der Farbe Türkis zu blicken? Sofern das Farbsehen intakt ist, wird die Erfahrung beim Betrachten dieser Leinwand verschieden sein zur Erfahrung beim Betrachten einer ziegelroten Leinwand - siehe Abbildung 2. Das gemeinsame Wirken beider Farben ruft eine dritte Quale hervor, welche sich nicht auf die beiden einzelnen Farb-Qualia reduzieren lässt. [42]

Qualia sind nicht nur auf Farb-Empfindungen begrenzt, sondern beziehen sich auf alle möglichen Arten von Empfindungen, welche sich beispielsweise mit Ausdrücken wie „es fühlt sich an als ob“, „es ist wie“ oder „das Empfinden, wenn“ beschreiben lassen. In diesem Abschnitt jedoch sollen besonders mit visuellen Eindrücken assoziierte Qualia exemplarisch verwendet werden. [43]

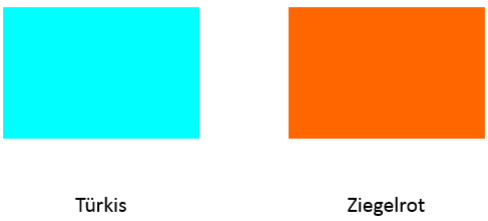

Abbildung 2: Der qualitative Eindruck beim Betrachten der Farben *Türkis* und *Ziegelrot* ist verschieden.

[15] Hier ist darauf hinzuweisen, dass die erste Aussage den Physikalismus unterstützt und die zweite eher in das dualistische Bild passt.

Qualiainversion

Qualiainversion ist ein Überbegriff für eine Klasse von Gedankenexperimenten, welche in der Philosophie des Geistes eine bedeutsame Rolle spielen und als Ausgangspunkt für weitere Argumente dienen können. In den meisten Fällen geht es in den Gedankenexperimenten um eine Variante der *Spektruminversion*, also der Vertauschung von Farbeindrücken oder Farb-Quale (obgleich es sich bei einer echten Spektruminversion um eine spezifische Art der Vertauschung handelt). Ein einfaches Beispiel für eine Qualia-Vertauschung illustriert Abbildung 3. [42]

Es sei eine Person namens Roman, für welche die Qualia für Ziegelrot und Türkis vertauscht sind[16] (interpersonelle Vertauschung). Eine weitere Person, Christine, hat diese Vertauschung nicht. Würde nun Roman Abbildung 3 betrachten, so würde es sich für ihn genauso anfühlen, wie es sich für Christine anfühlt, wenn sie Abbildung 2 betrachtet. Von außen betrachtet (Verhalten), könnte nicht festgestellt werden, dass in Roman diese Vertauschung vorliegt, da das Verhalten ident wäre. Da er Christines Türkis-Quale als seine Ziegelrot-Quale bezeichnet und umgekehrt - er hatte nie jenes Erleben dieser zwei Farben, wie Christine es hat, sondern stets nur sein eigenes Erleben[17]. [42]

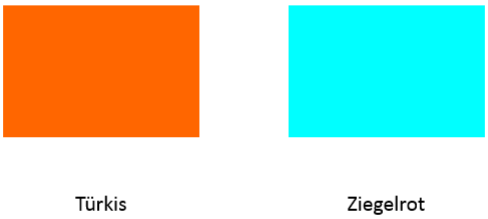

Abbildung 3: Wenn jemand die Qualia für Türkis und Ziegelrot in einer Person vertauschen würde, so würde für sie Abbildung 2 genauso aussehen wie diese Abbildung für eine nicht-invertierte Person.

[16] Es gibt eine hypothetische physiologische Variante der Netzhaut, in welcher solch eine Vertauschung der Farb-Qualia möglich wäre. Eine Exprimierung des S-Opsins (Gen OPN1LW) in den M-Zapfen, sowie des M-Opsins (Gen OPN1MW) in den S-Zapfen könnte die Farbeindrücke von Rot und Grün, ähnlich einer Spektruminversion, vertauschen. Die Inzidenz solch einer Variante wird mit 70 pro 100.000 in der männlichen Population geschätzt. Diese Variante beeinflusst möglicherweise das Verhalten. [44]

[17] Anders wäre die Situation, wenn die Vertauschung in Roman erst in etwa der Mitte seines bisherigen Lebens geschehen wäre (intrapersonelle Vertauschung). Er könnte die Farb-Qualia aus seiner Erinnerung vor und nach der Vertauschung vergleichen und einen Unterschied feststellen.

Abwesende Qualia

Kann es eine Welt geben, welche zur unserer völlig identisch ist, jedoch gänzlich ohne Qualia [45]? Solch eine Welt würde gänzlich gleich aussehen und die Lebewesen auf ihr wären perfekte Doppelgänger, jedoch ohne ein inneres Erleben. Jene Doppelgänger nennt David Chalmers in diesem Zusammenhang „philosophische Zombies“, kurz "p-Zombies“, welche im Verhalten, ihrer Physiologie und *Physik ununterscheidbar* sind, jedoch im Unterschied zu uns keinerlei Qualia haben. Es herrscht bislang kein Konsens darüber, ob p-Zombies möglich sind. Nach Chalmers ist die Möglichkeit von p-Zombies logisch widerspruchsfrei vorstellbar (und daher logisch möglich) ([46] S. 94-99).

Wissensargument

Eine hypothetische Wissenschaftlerin namens Mary wird in einem völlig farblosen Labor eingeschlossen und hat die Aufgabe, alles über die Physiologie des Sehens über einen schwarz-weißen Monitor herauszufinden. Nachdem Mary nun jedes einzelne Detail der Vorgänge im Gehirn - alle physikalischen Zustände und Prozesse, welche durch die visuellen Sinn bedingt sind, kennt, verlässt sie das Labor. Außerhalb des Labors sieht Mary zum ersten Mal Farben, daher sie erlebt Farb-Quale. Die Frage ist nun, ob Mary durch das Erleben dieser Farb-Quale etwas Neues über die Physiologie des Sehens lernen kann, oder ob die neuen Eindrücke ihr Wissen nicht erweitern können. Falls Mary etwas Neues lernen sollte, so würde dies bedeuten, dass Farb-Quale durch äußere Betrachtung der Physiologie und Physik des Sehens nicht erklärbar sind bzw. in einer vollständigen physischen Beschreibung nicht enthalten sind. [47]

1.5.3 Supervenienz

In der Philosophie des Geistes[18] bezeichnet der Begriff *Supervenienz* die Wechselbeziehung zwischen dem Mentalen und dem Physischen. Es wird mit dem Begriff jedoch keine kausale Wirkung impliziert, sondern lediglich die „gleichzeitige“ Änderung von mentalen und physischen Zuständen festgestellt[19]. In anderen Worten, es kann keine Änderung von physischen Zuständen ohne eine Änderung von mentalen Zuständen geben. Das Mentale superveniert somit über das Physische - jedoch nicht umgekehrt. [48] Die genaue Bedeutung von Supervenienz hängt stark vom metaphysischen Bezug

[18] Der Supervenienz-Begriff hat auch in anderen Gebieten der Philosophie und der Ethik Bedeutung.

[19] Nach Jaegwon Kim beinhaltet der Supervenienzbegriff keine Erklärung der Beziehung zwischen mental und physisch, sondern ist vielmehr eine Reformulierung der Körper-Geist-Beziehung - ein Begriff welcher durch „Körper-Geist-Supervenienz“ ausgetauscht werden könnte. ([48] S. 167)

ab. So hat Supervenienz im Substanzdualismus andere Implikationen als im reduktiven Physikalismus. [49]

Der Physikalismus hat den Anspruch, dass alle Eigenschaften (biologische, psychologische, soziale usw.) letztendlich auf physische Eigenschaften fußen und durch diese erklärt werden können. Die mit dem Physikalismus in Verbindung stehende Identitätstheorie[20] besagt, dass jeder mentale Zustand oder Prozess mit einem physischen Zustand oder Prozess identisch ist[21]. [50]

1.6 Freier Wille

Freier Wille ist ein Begriff, welcher in Psychologie, Recht, Philosophie, Ethik, Religion, Politik etc. Verwendung findet, und jeweils eine andere Bedeutung hat. Der hier im Zusammenhang mit dem Bewusstsein stehende freie Wille meint die Fähigkeit zu einer subjektiv[22] bewussten Entscheidung für ein Handeln aus verschiedenen Wahlmöglichkeiten. [53]

Eine wichtige Rolle spielt in der Diskussion des freien Willens die moralische Verantwortung. Der Begriff Verantwortung wird in der Philosophie in zwei Kontexten diskutiert. Zum einen geht es um die Verantwortung in der Gesellschaft im Zusammenhang mit Lob und Schuld[23]. Zum anderen steht ein Typ von Verantwortung, welche mit einem inneren Pflichtbewusstsein assoziiert ist - unabhängig von einem externen Beobachter. ([55] S. 1 - 4)

Der Determinismus ist von zentraler Bedeutung in der Erklärung des freien Willens.

[20] Es gibt zwei Varianten der Identitätstheorie: Die Typ-Typ Identitätstheorie besagt, dass mentale Zustände von einem gewissen Typ identisch sind mit einem physischen Zustand eines gewissen Typs. Es kann jedoch argumentiert werden, dass einem gewissen mentalen Zustand eines Typs physische Zustände unterschiedlichen Typs zugrunde liegen können. Aufgrund von Argumenten dieser Art hat die Typ-Typ Variante keine Bedeutung mehr in modernen Betrachtungen. Die hier gemeinte Instanz-Instanz-Identitätstheorie (im Englischen Token-Token) setzt einzelne mentale Zustände mit einzelnen physischen Zuständen gleich - ungeachtet des Typs. [50]

[21] Kritiker der Identitätstheorie stellen zum einen fest, dass diese Theorie eine einseitige Relation beschreibt - dass also mentale Zustände sich ändern, wenn physische dies tun, jedoch nicht umgekehrt - eine Identität beschreibt allerdings stets eine symmetrische Beziehung. Zum anderen wird genannt, dass ein mentaler Zustand durch verschiedene physische Zustände hervorgerufen werden kann (multiple Realisierung), was gegen eine Identität im hier gemeinten Sinne spricht [51].

[22] Die Subjektivität ist eine entscheidende Komponente des Freien Willens. Genauso wie das innere Erleben ist der freie Wille aus einer inneren, subjektiven Perspektive frei. Daher wird oft behauptet, dass der Begriff freier Wille besonders im Zusammenhang mit Bewusstsein einen Sinn macht. ([52] S. 55 - 78)

[23] Dieser Blick auf die moralische Verantwortung ist tief verwurzelt im Recht. Bedeutend ist beispielsweise die Frage des Schuldprinzips, wie jemand schuldig gesprochen werden kann, wenn sie/er für ihre/seine Taten nicht selbst verantwortlich ist [54]? Siehe auch § 4 des Österreichischen Strafgesetzbuchs.

Mit Determinismus ist gemeint, dass das gesamte Universum sich in einem einzelnen physikalisch realen Verlauf befindet, auf eine bestimmte Zukunft zusteuert und davon nicht abweicht. Ob der Determinismus auf unsere Realität zutrifft und in wie weit er mit dem freien Willen kompatibel ist, wird kontrovers diskutiert. ([56] S. 1 - 4 und 33 bis 48)

Nach [55] kann die philosophische Debatte um die Natur des freien Willens grob in vier große Lager eingeteilt werden, welche in folgender Darstellung erörtert werden:

- Libertarismus[24]: Freier Wille existiert, das heißt wir haben die freie Wahl und können bewusst eine Möglichkeit (unter mehreren) aussuchen. Außerdem ist der Determinismus mit der Freiheit, sich eine Möglichkeit auszusuchen inkompatibel, weil jede Entscheidung zwangsläufig durch den bereits feststehenden Verlauf der Realität determiniert wäre. Der Libertarismus könnte auch als „weicher" Inkompatibilismus bezeichnet werden.

- Harter Inkompatibilismus: Das Universum ist determiniert und jede Entscheidung, welche ein Mensch scheinbar trifft, war bereits zuvor festgelegt. Aus diesem Grunde kann es keinen freien Willen geben, da seine Grundaussage in diesem Kontext nicht haltbar ist.

- Kompatibilismus: Freier Wille und Determinismus können gemeinsam existieren und schließen sich nicht aus. Das heißt unsere Entscheidungen sind zwar determiniert, dies jedoch von unseren eigenen inneren Überzeugungen, Wünschen, Ansichten und Denken und nicht etwa durch andere (äußere) Ursachen. Manche AutorInnen stellen die Behauptung auf, dass der Determinismus Voraussetzung für einen freien Willen ist (harter Kompatibilismus). Hingegen der weiche Kompatibilismus (oder Semikompatibilismus) ist gegenüber dessen, ob Entscheidungen determiniert sind oder nicht, indifferent.

- Revisionismus: Das intuitive Bild sowohl vom freien Willen als auch des Determinismus ist zumindest teilweise nicht zutreffend. Es muss eine neue Basis für das Verständnis dieser Begriffe abseits der bisherigen intuitiven Vorstellungen geschaffen werden. Möglichkeiten zu einer Revision des Begriffsverständnisses in diesem Sinne sind mannigfaltig. [57]

[24]Der Begriff Libertarismus in der Philosophie des Geistes ist abgegrenzt von dem Libertarismus aus der politischen Philosophie.

2 Material und Methoden

2.1 Allgemein

Die Diplomarbeit wurde in Form einer Literaturrecherche durchgeführt, wobei unterschiedliche Quellen zur Wissensgewinnung und -verarbeitung miteinbezogen wurden. Neben Fachliteratur, Fachbücher und Online-Informationen zur Hintergrundlektüre und weiteren Vertiefung wurde primär die naturwissenschaftliche E-Publikationsdatenbank „ArXiv“ herangezogen, da diese einen Großteil der benötigten Informationen für das interdisziplinäre Gebiet „Bewusstsein“ enthält und eine Schnittstelle zwischen medizinischem, physikalischem und philosophischem Wissen darstellt. Weitere Erkenntnisse wurden über die biomedizinische E-Publikationsdatenbank „PubMed“ gewonnen, welche einen Großteil der bestehenden Literatur in der Medizin bereitstellt.

Bei Inhalten von verwendeten Online-Informationen wurde die besagte Quelle dokumentiert und das letzte Zugriffsdatum notiert. Websites und Online Dokumente wurden wie Bücher zitiert.

Als Textsatzprogramm wurde Texmarker Version 5.0.3 als Editor für die LaTeX-Distribution MiKTeX Version 2.9 verwendet. Zur Literaturverwaltung wurde das Literaturverwaltungsprogramm Citavi Version 6.3.0.0. genutzt. Die gesamte Literatur in der Diplomarbeit wurden in Form des Vancouver-Stils zitiert.

Die behandelten Themen der Diplomarbeit sind für beide Geschlechter gleich relevant und die verwendete wissenschaftliche Literatur bezieht sich zu gleichen Teilen auf beide Geschlechter.

2.2 Literatursuche und -auswahl

Bei der Literatursuche und -auswahl wurde darauf Wert gelegt, zuerst eine grundlegende Einführung in die Thematik über Bewusstsein und die darauf basierenden philosophischen und wissenschaftlichen Grundlagen, sowie die für das Verständnis notwendigen Definitionen zu geben.

Danach wurde eine repräsentative Auswahl an physikalischen Modellen von Bewusstseins erläutert. Dabei liegt ein spezieller Augenmerk auf medizinische Implikationen der vorgestellten Theorien.

Weiters wurden Grundlagen der künstlichen Intelligenz erörtert und deren Anwen-

dung in der Medizin, im Speziellen in der Kardiologie, diskutiert. Dabei wurden Einblicke in die möglichen zukünftige Verwendung von künstlicher Intelligenz in Diagnostik und Therapie geboten.

Im letzten Abschnitt wurde auf die Rolle von künstlicher Intelligenz in Gesellschaft und Medizin eingegangen und ethische Aspekte diskutiert. Außerdem erfolgte eine weitere Diskussion der physikalischen Modelle aus dem Blick der Medizin sowie der Philosophie.

3 Physikalische Modelle von Bewusstsein

3.1 Einleitung

Das Gehirn zählt zu den komplexesten Systemen, welche wir kennen. Es ist eine Annahme des Physikalismus, dass das Bewusstsein - so wie kognitive Fähigkeiten, Gedächtnis, etc - gänzlich auf komplexen physikalischen Prozessen im Gehirn beruht - diese Annahme soll in diesem Abschnitt ebenfalls getroffen werden. Eine physikalische Beschreibung von Bewusstseins kann aus verschiedenen Blickwinkeln und funktionellen Ebenen erfolgen. Von den molekularen Mechanismen im Zellinneren bis zu globalen Phänomenen des gesamten Gehirns wurden Ansätze für eine Theorie des Bewusstseins gesucht. Ebenso sind Herangehensweisen verschiedener Disziplinen hilfreich, um möglichst alle Aspekte des Bewusstseins zu erfassen. Im folgenden werden ausgewählte Beispiele für jene Modelle, welche unter Zuhilfenahme physikalischer Denkansätze Bewusstsein erklären, im Überblick erläutert. Danach soll auf zwei bekannte und bereits weit entwickelte Theorien, die IIT und die Orch OR Theorie, im Detail eingegangen und im weiteren Verlauf der Diplomarbeit näher diskutiert werden.

3.1.1 Bewusstsein am synaptischen Spalt

Um eine physikalischen Ursache von Bewusstsein, und im weiteren Sinne aller mentalen Aktivität des Gehirns, zu finden, beschäftigten sich Friedrich Beck und John Eccles mit Quantenprozessen am synaptischen Spalt [58]. Die Signalweiterleitung zwischen den Neuronen erfolgt über chemische Synapsen. Ein elektrischer Impuls bewegt sich, ausgehend vom Körper des Neurons, entlang des Axons und bewirkt mit einer gewissen Wahrscheinlichkeit eine Exozytose von Neurotransmittern in einen synaptischen Spalt. In dem Ansatz der Autoren wirken nicht deterministische Zustandsreduktionen auf die Wahrscheinlichkeit einer Exozytose und modulieren somit die Signalweiterleitung. Die genauen Mechanismen der Zustandsreduktion sind in der originalen Publikation [59] erklärt, jedoch ist ihre spezifische Rolle in der Entstehung eines Bewusstseins ist bislang unklar. [60]

3.1.2 Zustandsreduktion nach Henry Stapp

Henry Stapp postuliert in seiner Theorie von Bewusstsein eine ontologische Erweiterung der orthodoxen Quantenmechanik. Die zeitliche Veränderung eines quantenme-

chanischen Systems geschieht durch zwei grundlegende Prozesse, welche abwechselnd vonstatten gehen.

a) Die Schrödinger-Entwicklung eines Quantenzustandes ist kontinuierlich, deterministisch und reversibel.

b) Hingegen bedeutet eine *Messung* durch einen Beobachter eine Reduktion des Zustandes[25] in eine Eigenzustand der gemessenen Observable. Dieser Prozess ist sprunghaft, stochastisch und nicht reversibel. [62]

Stapp, auf Arbeiten von Eugene Paul Wigner [63] und John von Neumann ([64] S. 184-237) basierend, betrachtet das Gehirn selbst als Beobachter, welcher ein Element aus dem Möglichkeitsraum hin zur Aktualität bringt. Diese Aktualität ist wiederum eine Determinante für den Möglichkeitsraum der folgenden Messung. Eine Abfolge dieser Messungen „erschafft" in diesem Sinne eine kontinuierliche Aktualität und somit die Realität. [65]

Nun ist es das Besondere an der Stapp'schen Theorie, dass er jeder einzelnen dieser Messungen eine erfahrbare quantitative Eigenschaft oder eine elementare Quale zuordnet[26]. Daher gehen bewusste Erfahrungen einher mit einer Zustandsreduktion, welche wiederum kausal mit neuronalen Aktivitätsmustern zusammenhängt. [67]

Eine Wechselseitige Beziehung zwischen mentalen und physischen Prozessen entstehe nicht etwa durch direkte Kausalität, sondern durch Festlegung von Bedingungen an eine stochastische Zustandsreduktion. Es passiert daher nur dass, was mit den momentanen mentalen und physischen Zustanden zusammenpasst. Dadurch bleibt die Zustandsreduktion nach wie vor ein stochastischer Prozess, jedoch die Wahrscheinlichkeitsverteilung für das tatsächliche Eintreten der Eigenzustände wird gemeinsam durch die mentalen und physischen Zustände determiniert. [68] Der genaue Zusammenhang von bewussten Zustandsreduktionen mit neuronalen Aktivitätsmustern, welche das neuronale Korrelat von Bewusstsein sind, ist bisher noch nicht erforscht [69].

[25] Die quantenmechanische Zustandsreduktion wird in der Kopenhaben Interpretation der Quantenmechnik auch als „Kollaps" bezeichnet . Sie beschreibt die Projektion eines Zustandes, welcher im Allgemeinen als eine Überlagerung von mehreren Eigenzuständen einer Observablen dargestellt werden kann, auf einen einzelnen Eigenzustand. Ein Eigenzustand ist ein spezieller Zustand eines Quantenmechanischen Systems, welcher durch eine Messung der assoziierten Observablen nicht verändert wird. Meist hat eine Observable mehrere (bis unendlich viele) Eigenzustände. [61]

[26] Diese Zuordnung von einer subjektiven Erfahrung zu einem physikalischen Vorgang hat eine Ähnlichkeit mit der Prozessontologie von Alfred North Whitehead. Der Philosoph und Mathematiker postulierte „actual occassions" als Prozess zur Entstehung der erfahrbaren Realität, analog zur quantenmechanischen Zustandsreduktion, welche sowohl mentale als auch physische Eigenschaften haben. [66] Diese Philosophie hat einen Bezug zum Panpsychismus (siehe Abschnitt 1.5.1).

3.1.3 Quantenfeldtheorien von Bewusstsein

Der Ansatz das formale Gerüst der Quantenfeldtheorie (QFT) auf physikalische Zustände im Gehirn anzuwenden, geht zurück auf L. M. Ricciardi und H. Umezawa [70]. Darin wird das Gehirn als Vielteilchensystem aufgefasst, welches als ganzes durch die QFT beschrieben wird, wobei in der Auffassung der Autoren jedes Neuron als „Teilchen“ betrachtet wird. Die spontane Symmetriebrechung[27] ist ein Mechanismus der QFT, welcher für Vielteilchensysteme und für eine Beschreibung mentaler Zustände im Gehirn relevant ist. Konkret im Falle des Vielteilchensystems Gehirn bedeutet dies, dass es unendlich viele verschiedene Vakuumzustände[28], also Darstellungen des Grundzustandes, gibt. [72]

Vakuumzustände können als Beschreibung der Zustände aller Neuronen im Gehirn aufgefasst werden [74]. Giuseppe Vitiello schlägt nun vor, dass das Gedächtnis in den Vakuumzuständen codiert und durch Aktivierung der neuronalen Verbände abgerufen werden kann. Die Vakuumzustände könnte jemand sich als Bücher in ein unendlich großen Bibliothek vorstellen. Mit der genannten Aktivierung wird eines der Bücher ausgewählt und darin der Gedächtnisinhalt ausgelesen. Findet keine Aktivierung statt, dann ist der Inhalt des Gedächtnisses nicht abrufbar. Die Stabilität der Gedächtnis-Vakuumzustände wurde von Stuart et al. [75] näher diskutiert. [76]

Ein weiterer Schritt in der Entwicklung der Quantenfeldtheorie des Bewusstseins ist die nähere Betrachtung von Interaktionen des Gehirns mit seiner Umgebung. Durch die Interaktion[29] mit der Umwelt haben die Vakuumzustände eine limitierte Lebenszeit, bis sie durch ein Einwirken der Umgebung „verwaschen“ werden [77]. Eine weitere Folge

[27] Symmetrien haben in der Physik eine außerordentlich große Bedeutung. Eine Symmetrie beschreibt die Invarianz eines Systems unter einer, der Symmetrie zugehörigen, Transformation. Beispielsweise ist eine perfekte Kugel invariant unter der Drehung in allen Raumrichtungen, weil sie nach jeder Drehung gleich aussieht wie zuvor. Eine Kugel wird daher rotationssymmetrisch genannt. Die Rotationssymmetrie ist kontinuierlich, da egal wie weit eine Kugel sich dreht, sie immer gleich aussieht. Ein Beispiel für eine diskrete Symmetrie wäre die Spiegelsymmetrie. Das Frieren von Wasser ist ein anschauliches Beispiel in welchem eine Symmetriebrechung auftritt. Kühlt jemand einen Becher Wasser stark ab, tritt ein Phasenübergang auf, welcher die zuvor ungeordneten Wassermoleküle in einem hexagonalen Kristallgitter anordnet. Diese Kristallgitter hat eine bestimmte Orientierung im Raum, welche zuvor im Wasser nicht vorhanden war. In andere Worten das Wasser war im flüssigen Zustand in allen Richtungen völlig gleich, der Eiskristall jedoch hat nun eine bestimmte Richtung. [71] Auf den Grund, weshalb in der QFT des Gehirns eine spontane Symmetriebrechung auftritt, kann in dieser Diplomarbeit leider nicht eingegangen werden. Für nähere Informationen siehe [72].

[28] Ein Vakuumzustand bezeichnet innerhalb der Quantenfeldtheorie einen Grundzustand des Systems. Der Grundzustand ist jener, in welchem die Freiheitsgrade des Systems so gewählt sind, dass es die niedrigst mögliche potentielle Energie hat. Jedes physikalische System neigt dazu seine potentielle Energie zu minimieren. ([73] S. 13-33)

[29] Interaktionen eines Quantensystems mit seiner Umgebung könne mittels der Theorie der Dekohärenz formal behandelt werden. Siehe hierzu Abschnitt 3.3.3.

aus der Wechselwirkung mit der Umwelt ist, dass sie einen sogenannten „Zeitpfeil“ im System erzeugen, wodurch es möglich ist, aus Veränderungen in den Vakuumzuständen eine zeitliche Abfolge dieser Veränderungen zu rekonstruieren, was möglicherweise für unser Zeitgefühl eine Rolle spielen könnte [78].

3.2 Theorie der integrierten Information

3.2.1 Einleitung

Eine mögliche Theorie der Qualia, und im weitesten Sinne des Bewusstseins, nennt sich *integrated information theory (IIT)* oder *Integrierte Informationstheorie* [79]. Die von Giulio Tononi vorgeschlagene Theorie versucht bewusste Erfahrungen zu erklären und deren Zusammenhang zu physikalischen Systemen herzustellen. Mittels IIT ist jemand nicht nur in der Lage ein System auf das Vorliegen von Bewusstsein zu untersuchen, sondern dieses auch zu quantifizieren und herauszufinden, welche qualitative Erfahrung das System machen kann. [80]

Bei der Erforschung des Bewusstseins ist nicht nur das Finden von neuronalen Korrelaten des Bewusstseins (NCC), also der physischen Pendants zum inneren Erleben, mittels experimenteller Verfahren von Bedeutung. Ebenfalls wird zum Verständnis des Bewusstseins eine Theorie, welche den Zusammenhang zwischen dem Bewusstsein und seinem physischen Substrat erklärt, benötigt. Die IIT ist in der Lage Befunde aus den Neurowissenschaften und der Medizin zu erklären sowie experimentell überprüfbare Voraussagen machen. [79] Näheres zu den Medizinischen Implikationen in Abschnitt 5.1.

3.2.2 Der Aufbau der IIT

Zu Beginn, als ersten Schritt, formulieren die Autoren der IIT Kriterien für die essentiellen Eigenschaften der Instanzen des Bewusstseins, welche im folgenden bewusste oder innere Erfahrungen genannt werden. Hier ist es von besonderer Bedeutung festzuhalten, dass sich bewusste Erfahrungen auf ein inneres Erleben und nicht auf äußere Eindrücke über die Wahrnehmung bezieht. Obgleich ein Bewusstseinsinhalt durch eine Wahrnehmung der Sinne bewirkt werden kann, so ist es nicht mit diesem identisch. Ein inneres Erleben ist an sich unabhängig von äußeren Eindrücken. Wenn jemand die Augen schließt und vor dem inneren Augen ein Bild aus der Erinnerung abruft, so ist das damit einhergehende innere Erleben, die Erfahrung, von der Wahrnehmung

entkoppelt. Das Bewusstsein kann als die Gesamtheit des inneren Erlebens aufgefasst werden. [81]

Im zweiten Schritt werden die essentiellen Eigenschaften des Bewusstseins als Axiome, welche die gefundenen Kriterien erfüllen, formuliert. Die Axiome sind offensichtliche Wahrheiten, welche aus einer inneren Perspektive des eigenen Bewusstseins real sind (implizite Realität). Ein dritter Schritt folgert aus den Axiomen grundlegende Eigenschaften von physischen Systemen, welche ein Bewusstsein hervorrufen können. In der IIT wird besonderer Augenmerk auf die Analyse jener Systeme gelegt. Der gesamte Abschnitt 3.2 zur IIT bezieht sich, sofern nicht anders gekennzeichnet, auf die Publikationen [79] [82] [83].

Bei den in der IIT behandelten physischen Systeme kann es sich beispielsweise um logische oder neuronale Schaltungen handeln. Diese könnten Teile von größeren und komplexeren Systemen, wie etwa ein menschliches Gehirn, sein. Für eine bessere Übersichtlichkeit sollen im folgenden jedoch nur sehr simple und überschaubare Systeme behandelt werden. Es müssen gewisse Grundvoraussetzungen für ein physisches System gelten, damit dieses mittels IIT analysiert werden kann (siehe Abbildung 4):

- Ein physisches System (zum Beispiel eine neuronale Schaltung) besteht aus einzelnen Elementen (zum Beispiel Neuronen).
- Die Elemente haben zwei oder mehrere innere Zustände. Die Gesamtheit der Zustände aller einzelnen Elemente ist gleichbedeutend mit dem Zustand des Systems.
- Die inneren Zustände der Elemente können über die Eingänge (Dentriten) beeinflusst werden. Über Ausgänge (Axon) wirken die inneren Zustände nach außen.

Schritt 1

Folgende **Kriterien** werden herangezogen, um die essentiellen Eigenschaften von bewussten Erfahrungen zu finden.

- Spezifität: Die Eigenschaften betreffen nur die bewusste Erfahrung selbst.
- Unmittelbarkeit: Die Eigenschaften sollten direkt erkennbar und nicht abgeleitet werden.
- Notwendigkeit: Die Eigenschaften sollten auf alle Erfahrungen zutreffen und nicht bloß auf manche.

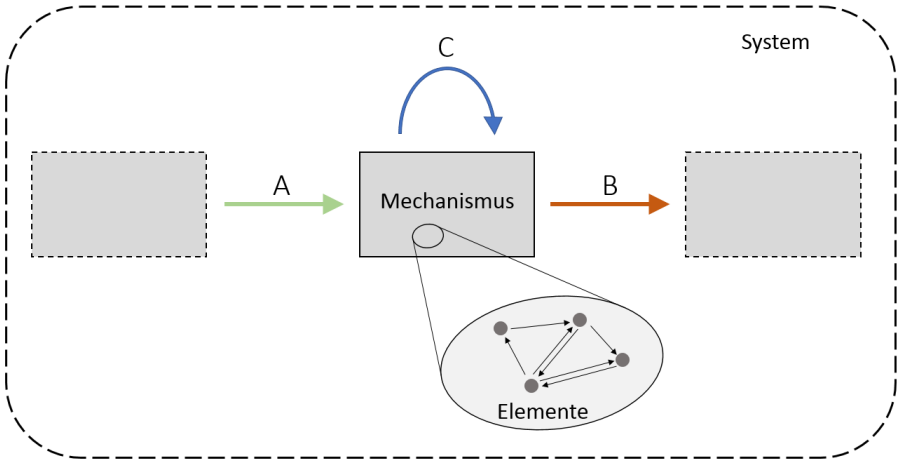

Abbildung 4: Ein System besteht im Allgemeinen aus mehreren Mechanismen, welche wiederum aus einzelnen Elementen aufgebaut sind. Sowohl Elemente als auch Mechanismen sind miteinander verbunden. Jeder Mechanismus hat einen Eingang A, durch welchen von außen auf ihn gewirkt werden kann. Dies wird als *Ursache* auf diesen Mechanismus bezeichnet. Ebenso kann der Mechanismus über seinen Ausgang B nach außen eine *Wirkung* haben. Manche Mechanismen haben eine Selbstwirkung C. Zusammen bilden alle Möglichkeiten zur Ursache und Wirkung das *Ursache-Wirkungs-Repertoire (UWP)* dieses Mechanismus. Die Selbstwirkung ist für ein System jedoch obligat.

- Vollständigkeit: Es sollten keine weiteren essentiellen Eigenschaften auf bewusste Erfahrungen zutreffen.
- Konsistenz: Es sollten in den Eigenschaften keine Widersprüchlichkeiten enthalten sein.
- Unabhängigkeit: Die Axiome sind unabhängig von einander und ein Axiom kann nicht aus anderen hergeleitet werden.

Schritt 2

Unter Zuhilfenahme der in Schritt 1 genannten Kriterien werden **Axiome** über das Bewusstsein aufgestellt. Die Axiome beschreiben die Phänomenologie des Bewusstseins.

- Axiom 1 (Existenz): Das Bewusstsein existiert und jede innere/bewusste Erfahrung ist real. Wie bereits René Descartes sagte, ist das innere Erleben genau jene Sache, welche nicht angezweifelt werden kann ([10] Kap. 3) und es existiert unabhängig von äußeren Beobachtern.
- Axiom 2 (Struktur): Das Bewusstsein hat eine Struktur und besteht aus mehreren *Teilerfahrungen*. Beim Betrachten einer Blumenwiese entstehen innere Ein-

drücke oder bewusste Erfahrungen. Diese Erfahrungen existieren unabhängig von der Wahrnehmung der Blumenwiese. Das innere Erleben könnte als Abbild der Wahrnehmung bezeichnet werden. Nun könnte das innere Erleben einer Blumenwiese beispielsweise in die einzelnen Blumen, deren Duft, den Sonnenschein und so weiter zerlegt werden. Diese Komponenten bilden gemeinsam die gesamte Erfahrung als Ganzes.

- Axiom 3 (Information): Jede bewusste Erfahrung ist einzigartig und aus einer spezifischen Menge an *Teilerfahrungen* aufgebaut. Die innere Erfahrung unterscheidet sich dadurch von allen anderen möglichen Erfahrungen (*Differenzierung*).

- Axiom 4 (Integration): Eine bewusste Erfahrung ist eine Einheit. Die gesamte Erfahrung ist nicht reduzierbar auf von einander unabhängigen einzelnen Erfahrungen (oder Untermengen an Teilerfahrungen). Das Erleben einer Blumenwiese im Sommer kann also nicht beispielsweise als Blumenwiese ohne Duft und separat als den Duft der Blumenwiese erlebt werden, sondern die Blumenwiese hinterlässt einen einzigen (Gesamt-)Eindruck in uns. Das Ganze ist mehr als die Summe seiner Teile.

- Axiom 5 (Eindeutigkeit): Das momentane Erleben einer Erfahrung schließt das gleichzeitige Erleben von anderen Erfahrungen aus. Eine Überlagerung mehrerer separater Erfahrungen ist dadurch nicht möglich. Zu jeder Zeit wird genau eine Erfahrung in ihrer Gesamtheit erlebt. Es gibt kein Erleben über die Grenzen der Erfahrung hinaus und alles innerhalb der Grenzen der Erfahrung wird erlebt. Eine Blumenwiese wird als Ganzes erlebt - ein zusätzliches Erleben, etwa von anderen Vorgängen in unserem Gehirn, kommt nicht vor.

Schritt 3

Aus der Phänomenologie können mögliche Eigenschaften von physischen Systemen, welche die kausale Ursachen des Bewusstseins sind, geschlossen werden. Bei jenen als **Postulate** formulierte Eigenschaften handelt es sich um Hypothesen über die der Phänomenologie zugrundeliegenden Mechanismen. Zu jedem Axiom gibt es ein entsprechendes Postulat. Die in den Postulaten erklärten Begriffe werden in Abbildung 7 zusammenfassend dargestellt.

- Postulat 1 (Existenz): Ein System dessen Elemente sich in einem bestimmten Zustand befinden benötigt ein *Kausalitäts-Vermögen (KV)*, um bewusste Erfah-

rungen erzeugen zu können. Es ist daher notwendig, dass etwas auf das System, sowie das System nach außen wirken kann. Eine bewusste Erfahrung ist vom eigenen Bewusstsein aus gesehen (intrinsisch) und unabhängig von äußeren Beobachtern existent. Damit das Bewusstsein sich selbst wahrnehmen kann, muss es auf sich selbst wirken können. Um das KV näher untersuchen zu können, wird jeder mögliche Zustand des Systems in der Zukunft (= Wirkung) und in der Vergangenheit (= Ursache) als eine Dimension im sogenannten *Ursache-Wirkungs-Raum (UWR)* dargestellt (siehe Abbildung 5).

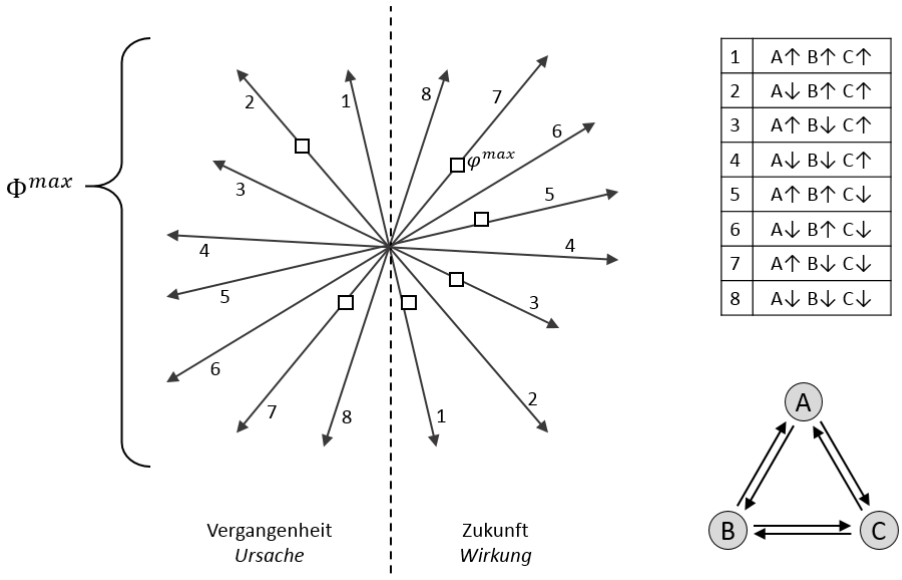

Abbildung 5: Jede Dimension im UWR (linke Bildhälfte) entspricht einem Zustand des Systems. Ein Zustand des Systems ist durch die Zustände seiner Elemente spezifiziert (Tabelle rechts oben). Das hier dargestellte System besteht aus drei Elementen, welche jeweils zwei Zustände haben können (hier mit ↑ und ↓ bezeichnet). Es sind somit acht System-Zustände möglich (als 1 bis 8 bezeichnet). Eine Dimension im UWR ist nochmals aufgeteilt in die Ursache (Vergangenheit) und die Wirkung (Zukunft) des jeweiligen Zustandes. Hier sind viele Dimensionen auf eine flache Ebene projiziert, wodurch der Raum stark verzerrt erscheint.
Im UWR sind Vierecke abgebildet, welche die Konzepte der Subsysteme A, B, C, AB, AC und AD repräsentieren. Die Position der Konzepte auf den Achsen bestimmen die Wahrscheinlichkeit [84], dass der momentane Zustand aus dem jeweiligen Zustand (durch die jeweilige Dimension bestimmt) hervorgegangen ist (Ursache - Vergangenheit) oder in den jeweiligen Zustand übergehen wird (Wirkung - Zukunft). Jedem Konzept ist ein φ^{max} zugeordnet. Die Konstellation der Konzepte im UWR bestimmen die Qualität der Quale, also den Inhalt des Bewusstseins. Der Wert φ^{max} eines Konzepts bestimmt, wie sehr es zur Quale beiträgt. Die Nicht-Reduzierbarkeit des gesamten Komplexes Φ^{max} ist identisch mit der Quantität der Quale, dem Grad des Bewusstseins.

- Postulat 2 (Struktur): Das System kann als die Menge seiner Elemente und der

Verbindungen der Elemente untereinander aufgefasst werden. Ein sehr einfaches Beispiel wäre ein System mit drei Elementen A, B und C - bezeichnet als ABC. Sämtliche Untermengen (A, B, C, AB, BC, AC und ABC) können ein KV aufweisen. Die physischen Systeme sind hierarchisch aufgebaut: Einzelne Elemente (niedrigste Ebene; A, B und C) können zu einem Mechanismus (höhere Ebenen; AB, AC etc.) zusammengesetzt werden, mehrere Mechanismen bilden eine Struktur. Ein Mechanismus ist das physische Substrat der oben genannten Teilerfahrung.

- Postulat 3 (Information): Ein System hat eine bestimmte Menge an *Ursache-Wirkungs-Repertoires (UWP)*, welche die KV der Mechanismen des Systems vollständig beschreiben. Durch die Repertoires unterscheidet sich ein System von allen anderen möglichen Systemen (Differenzierung). All jene UWP von jeder möglichen Kombination von Elementen in einem System bilden eine *Ursache-Wirkungs-Struktur (UWS)*. Die UWS beschreibt die Freiheitsgrade in Ursache und Wirkung des Systems auf sich selbst.

- Postulat 4 (Integration): In Analogie zur Nicht-Reduzierbarkeit von bewussten Erfahrungen in unabhängige einzelnen Teilerfahrungen, ist ebenfalls eine UWS nicht reduzierbar in unabhängige Teile. Das Maß an Nicht-Reduzierbarkeit einer UWS wird in der IIT über die integrierte Information Φ quantifiziert. Der Wert von Φ sagt aus, in welchem Grade sich die UWS eines Systems ändert, wenn es auf spezielle Art in zwei Teile zerlegt wird. Hierbei wird unter allen möglichen Teilungsarten jene ausgewählt, welche den geringsten Unterschied in der UWS hervorbringt, um das Φ zu messen (*Minimale Informations-Partition (MIP)*). Die Teilung wird durch eine einseitige *Verrauschung* der Verbindung zwischen den beiden einzelnen Teilen bewirkt, wodurch kein Signal mehr über die „gekappten“ Verbindungen in diese Richtung übermittelt werden kann (siehe Abbildung 6). In die andere Richtung jedoch bleibt die Verbindung unbeeinflusst (*unidirektional*). Wenn also ein System in zwei Teile aufgespalten werden kann, ohne die UWS zu ändern, so ist das System reduzierbar und hat keine integrierte Information ($\Phi = 0$). Ist das System andererseits in keine der beiden Richtungen entlang der MIP teilbar, ohne die UWS zu beeinflussen, so ist diesem System integrierte Information inhärent ($\Phi > 0$).

  Analog zur Integration von Systemen lässt sich das obige Postulat auf Mecha-

nismen anwenden. Ein Mechanismus kann nur dann eine Teilerfahrung bewirken, wenn sein UWP innerhalb des Systems nicht reduzierbar und in einzelne Teil-Mechanismen zerlegt werden kann. Diese Nicht-Reduzierbarkeit wird ebenfalls mittels integrierter Information quantifiziert und mit φ bezeichnet. Wenn also ein Mechanismus entlang der MIP, wie im Falle des Systems, geteilt wird und dabei sein UWP unverändert bleibt ($\varphi = 0$), dann könnten die beiden Teile des Mechanismus jeweils zwei separaten Teilerfahrung entsprechen (falls diese nicht reduzierbar sind).

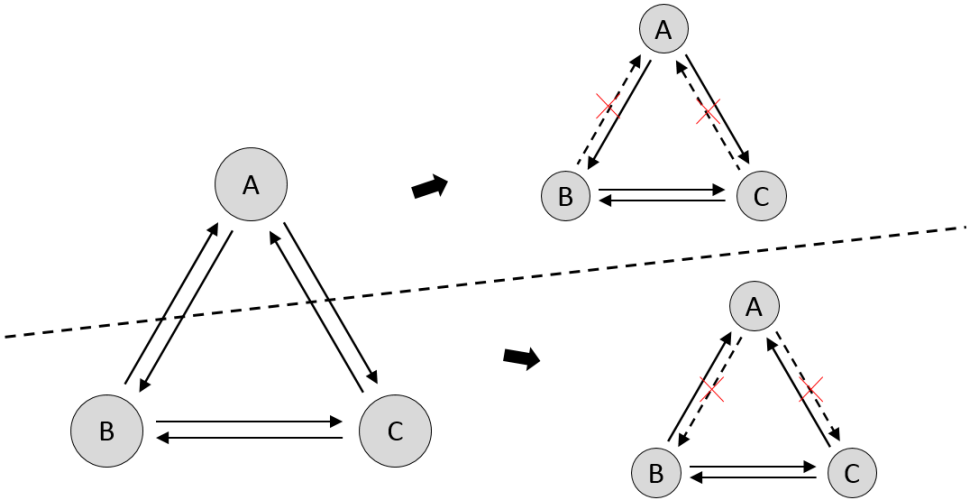

Abbildung 6: Illustration zur Teilung entlang der MIP. Der hier abgebildete Mechanismus mit den Elementen A, B und C wird entlang der MIP unidirektional geteilt. Dabei werden die Verbindungen über die Teilungsgrenze jeweils nur in eine Richtung verrauscht bzw. unterbrochen. Beide Teilungs-Richtungen werden separat im Hinblick auf die Nicht-Reduzierbarkeit des Mechanismus (φ) bewertet. Wenn sich die Freiheitsgrade in der Ursache und Wirkung des Mechanismus bei den Teilungen reduzieren, dann ist der Mechanismus nicht reduzierbar. Die Nicht-Reduzierbarkeit kann quantifiziert werden (φ).

- Postulat 5 (Eindeutigkeit): Die UWS eines Systems ist eindeutig definiert und befindet sich innerhalb eines spezifischen örtlichen und zeitlichen Rasters. Das zeitliche Raster bestimmt, in welchen Zeitabständen (typisch sind etwa 10 - 100 ms) zwei Eindrücke hintereinander erlebt werden. Aus einer Menge an mehreren überlappenden UWS wird genau ein spezielles ausgewählt und jenes bestimmt das bewusste Erleben. Durch Elimination anderer UWS ist also eine Überlappung letztendlich ausgeschlossen und damit ist das Erfordernis nach Eindeutigkeit erfüllt.

Die Auswahl einer UWS erfolgt nach einem Maximum-Prinzip. Jenes von allen möglichen UWS, welches die höchste Nicht-Reduzierbarkeit Φ^{max} hat, nennt sich ein *Komplex* und wird zur Aktualität im physischen System, zum bewussten Erleben. Ein Komplex ist das Gegenstück auf Systemebene zu einer *Quale im weiteren Sinne*, nachfolgend nur *Quale* genannt.

Dasselbe Prinzip gilt auch für Mechanismen, in welchem nur ein UWP ausgewählt werden kann (Eindeutigkeit). Dies ist jenes UWP mit der maximalen Nicht-Reduzierbarkeit φ^{max}. Falls für einen Mechanismus ein φ^{max} größer als null existiert, wird der Mechanismus als *Konzept* bezeichnet und ist damit ein physisch-strukturelles Pendant zu einer *Quale im engeren Sinne*.

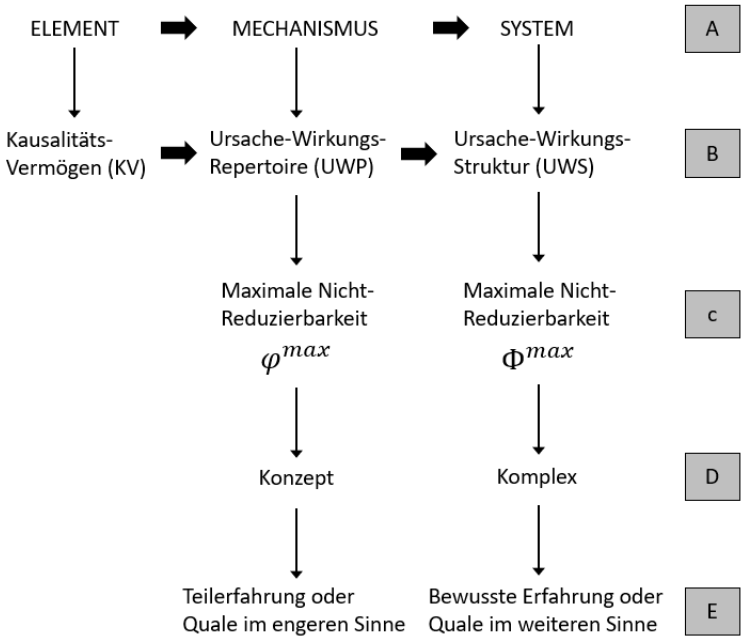

Abbildung 7: Grafik zur Zusammenfassung und Veranschaulichung der in den Postulaten genannten zentralen Begriffe. Jedes System ist aus Mechanismen aufgebaut und diese bestehen wiederum aus einzelnen Elementen (A). Die Freiheitsgrade in Ursache und Wirkung haben auf jeder Ebene eigene Namen (B). Maximale Nicht-Reduzierbarkeit (C) für die Ursache- und Wirkungsmöglichkeiten in einem Mechanismus oder System ist das Kriterium für jene UWP/UWS (D), welche ein (Teil-)Bewusstsein (E) hervorrufen können.

3.2.3 Identitätstheorem der IIT

Die Quale im weiteren Sinne legt jegliche Eigenschaft der bewussten Erfahrung bzw. des Bewusstseins fest. Das Identitätstheorem der IIT besagt nun, dass eine Quale,

bedingt durch den zugrundeliegenden Komplex, mit dem Bewusstsein identisch ist. Die Identität gilt hier jedoch nicht zwischen Bewusstsein und physischem System, wie es die Identitätstheorie [85] postuliert. Verschiedene physische Systeme können den selben Komplex oder die selbe Quale bewirken, weshalb Quale und physisches System nicht identisch sein können.

$$Bewusstsein = Quale \neq physisches\ System$$

Die Funktion des Gehirns wird bestimmt durch Neuronen und deren Verknüpfungen. Innere Vorgänge in den Nervenzellen sind verursacht durch einzelne, kleinste Elemente. Diese Elemente bewirken physikalische und biochemische Prozesse in der Zelle. Das äußerliche Verhalten, also das Kausalitäts-Vermögen des Neurons wird durch die inneren Vorgänge bestimmt. Im Allgemeinen bilden Gruppen von Neuronen im Gehirn funktionelle Einheiten, welche mitsamt den internen Vorgängen in den Zellen nicht reduzierbar sein und einen Mechanismus formen können. Wenn eine Neuronengruppe einen Mechanismus bilden, haben diese ein UWP und sind potentiell das physische Substrat einer Teilerfahrung (Quale im engeren Sinne). Mehrere Neuronengruppen im Verband können als physisches System wirken. Dessen UWS mit dem Φ^{max} ist ein Konzept und assoziiert mit einer Quale im weiteren Sinne, dem Bewusstsein. Ein Verlust der Quale durch eine Veränderung des Systems, wie etwa eine Allgemeinnarkose [86] oder während dem traumlosen Schlaf in der Non-Rapid Eye Movement (Non-REM)-Phase [87], führt in Folge zu einem Verlust oder Zusammenbruch des Bewusstseins.

3.3 Orchestrated Objective Reduction

3.3.1 Einleitung

Die Orchestrated Objective Reduction (Orch OR) Theorie vom Bewusstsein [88], postuliert von dem mathematischen Physiker Roger Penrose (Oxford University) und dem Anästhesiologen Stuart Hameroff (University of Arizona), bedient sich mathematischer und quantentheoretischer Methoden, um ein Erklärungsmodell für das Bewusstsein auf physiologischer Ebene aufzustellen. Nach den Autoren sollen Quanteneffekte in den Mikrotubuli (MT), welche wichtige Strukturproteine in den Neuronen sind, eine zentrale Rolle in der Entstehung von Bewusstseinsprozessen im Gehirn spielen. Die Quanteneffekte modulieren die neuronale Aktivität via sogenannte „Objective Reductions (OR)“

oder *objektive Zustandsreduktionen*. Die objektive Zustandsreduktion wählt aus mehreren überlagerten Zuständen in den MT eine einzelne aus. Die von Penrose und Hameroff beschriebenen Vorgänge in den MT sollen nicht-berechenbar (siehe Abschnitt 3.3.2) und möglicherweise eine Grundlage für die Entstehung des Bewusstseins sein. Ob überlagerte Zustände in MT lange genug bestehen bleiben (*Kohärenzzeit*), um auf neuronaler Ebene relevant zu sein, ist einer der Hauptgegenstände in der kritischen Diskussion der vorgeschlagenen Theorie [89] [90]. Aussagen der Orch OR Theorie, experimentelle Evidenz und mögliche Anwendungen auf Gebiete in der Medizin werden im Abschnitt 5.2 behandelt.

3.3.2 Nichtberechenbarkeit vom Bewusstsein

Das Bewusstsein kann als emergente Eigenschaft, welche aus komplexen Berechnungen von Neuronennetzen im Gehirn entsteht, betrachtet werden ([91] S. 55-75). Grundlage für die Informationsverarbeitung in den neuronalen Netzen ist demnach ein determiniertes Verhalten der Nervenzellen (beschrieben durch Eingabe-Ausgabe-Funktionen) und der Topologie der Verbindungen zwischen ihnen. Die Vorgänge eines Neuronenkollektivs auf dieser Grundlage lassen sich als Algorithmus beschreiben und auf einer Turingmaschine[30], wie zum Beispiel ein gewöhnlicher Computer, simulieren [93]. Nach der Church-Turing-These [94] folgt daraus, dass das Verhalten der Neuronen und somit auch das Bewusstsein berechenbar ist.

Roger Penrose widerspricht dieser Auffassung. Nach ihm sei das Bewusstsein im Allgemeinen nicht berechenbar, oder präziser ausgedrückt, sind gewisse elementare Bewusstseinsprozesse mittels einer Turing-Maschine nicht berechenbar. Die Argumentation der Nicht-Berechenbarkeit vom Bewusstsein nutzt den zweiten Gödel'schen Unvollständigkeitssatz. Demnach kann in einem hinreichend mächtigen[31] und widerspruchsfreie formalen System, die Widerspruchsfreiheit dieses Systems nicht bewiesen werden. Hier anknüpfend argumentiert Penrose, dass es daher Aussagen gibt, welche durch ein formales System nicht bewiesen werden können, jedoch sehr wohl durch menschliche

[30] Eine Turingmaschine ist ein mathematisches Modell eines Rechners, welches auf Alan Turing zurück geht. Die Turingmaschine nimmt eine Eingabe aus einer wohldefinierten Menge an möglichen Eingaben an und berechnet schrittweise eine Ausgabe, welche ebenfalls Element ein wohldefinierten Menge an möglichen Ausgaben ist. Die Berechnung erfolgt rekursiv und anhand einer Überführungsfunktion. Daher ist das Verhalten einer Turingmaschine deterministisch. Mithilfe der Turingmaschine (unter anderen Modellen) lässt sich mathematisch definieren, welche Algorithmen berechenbar und welche nicht berechenbar sind ([92] S. 79 - 189)

[31] Was in diesem Zusammenhang „hinreichend mächtig" bedeutet, wird in 4.5.2, Lucas-Penrose Argument, näher erklärt.

Mathematiker. Daraus wird gefolgert, dass der Mathematiker bzw. der Mensch kein formales System im beschriebenen Sinne sein kann. Aus diesem Grunde sind zumindest manche Aspekte des menschlichen Geistes an sich nicht berechenbar bzw. nicht algorithmisch. ([95] S. 538-541) Eine ähnliche Argumentationskette kann in 4.5.2, Lucas-Penrose Argument, nachgelesen werden.

Es muss nach Penrose daher ein nicht berechenbarer Prozess in der Physik, welche die Funktion des Gehirns auf der fundamentalsten naturwissenschaftlichen Ebene beschreibt, geben. Ein Kandidat für solch einen Prozess ist die quantenmechanische Zustandsreduktion, welche im folgenden Kapitel im Detail erläutert wird.

3.3.3 Objektive Zustandsreduktion (OR)

Der Zustand eines quantenphysikalischen Systems wird vollständig durch seine Wellenfunktion Ψ beschrieben. Es ist daher die gesamte Information über die Konfiguration des Systems S in Ψ enthalten - nicht mehr und nicht weniger. Im allgemeinen ist Ψ aus mehreren reinen *reinen Wellenfunktionen*[32] ϕ zusammengesetzt. Solange S von seiner Umgebung E isoliert ist und nicht von außen gestört wird, so entwickelt sich seine Wellenfunktion in der Zeit nach deterministischen Regeln, der sogenannten Schrödinger-Entwicklung U. Eine Isolierung von S liegt in der Natur praktisch nie vor und müsste durch spezielle Mechanismen aufrechterhalten werden. ([96] S. 1 - 19)

In der Kopenhagen Deutung der Quantenmechanik [62] findet durch eine *Beobachtung* O eine Wechselwirkung von S mit seiner E statt und die U wird unterbrochen. E kann beispielsweise ein Messgerät oder ein anderes System sein. Eine Messung im beschriebenen Sinne führt zur stochastischen Reduktion von Ψ hin zu einer reinen Wellenfunktion ϕ bzw. zum sogenannten *Kollaps der Wellenfunktion*. Die reduzierte Wellenfunktion beschreibt in diesem Moment die physikalische Aktualität. In anderen Worten ist vor der Messung des Quantensystem in mehreren Zuständen gleichzeitig und erst mit einer Messung wird es gezwungen sich zufällig für einen der Zustände zu „entscheiden“, was oft mit dem Gedankenexperiment „Schrödingers Katze“ ([97] S. 812) versinnbildlicht wird. Somit entsteht ein Sprung in der ansonsten kontinuierlichen U. Die Erklärungslücke zwischen der U und dem plötzlichen „Kollaps“ in einen zufälligen

[32] Hier als „reine“ Wellenfunktion bezeichnet, ist in der Sprache der Quantenmechanik ein Eigenzustand. Eigenzustände sind alle jene Wellenfunktionen ϕ_n welche die Eigenwertgleichung $O\phi_n = o\phi_n$ erfüllen, wobei O der Operator einer Observable im System ist und o, der gemessene Wert dieser Observable. ([96] S. 27)

reinen Zustand, wird als *Messproblem*[33] bezeichnet ([96] S. 1 - 124).

Ein Ansatz zur Behandlung des Messproblems besteht darin, die Quantenmechanik als unvollständige Beschreibung von Naturvorgängen anzusehen. Die sogenannte *Diósi-Penrose Objective Reduction (DP OR)* ist ein Vorschlag zur Erweiterung der Quantentheorie, welche die „Lücke“ im Messproblem schließen soll. Die vorgeschlagene Zustandsreduktion (Kollaps) ist ein tatsächlich stattfindender, also *objektiver*, physikalischer Vorgang, welcher nicht durch die konventionelle Quantenmechanik beschrieben werden kann. Stattdessen ist ein neuartiger Gravitations-Effekt die Ursache für den Übergang von einer Wellenfunktion Ψ, welche eine Überlagerung mehrerer reiner Zustände ϕ_i ist, in einen einzelnen reinen Zustand ϕ. [99]

Zur Erklärung dieses Gravitations-Effekts soll ein sehr einfaches System S_1 dienen (siehe Abbildung 8). In S_1 befindet sich ein massenhaftes Teilchen T, welches sich in der Raum-Zeit bewegen kann. Die Wellenfunktion Ψ_1 beschreibt nun mit welcher Wahrscheinlichkeit sich das Teilchen an einem Punkt in der Raum-Zeit aufhält. Ψ_1 sei eine Überlagerung von zwei Wellenfunktionen ϕ_a und ϕ_b. In ϕ_a befindet sich das Teilchen im Mittel an einem anderen Ort, als in ϕ_b. Das heißt, die Masse von T befindet im Mittel an einem jeweils anderen Ort für die beiden ϕ. Nun kann die Masse aufgrund der Superposition von ϕ_1 und ϕ_2 mit sich selbst wechselwirken, da diese an beiden möglichen Orten ein „virtuelles“ Gravitationsfeld erzeugt. Die Gravitationsenergie E_G ist jene Energie, welche benötigt wird, um T im Zustand ϕ_1 aus jenem Gravitationsfeld, welches T im Zustand ϕ_2 erzeugen würde, zu entfernen. Überschreitet E_G die Planck-Energie [100], dann wird das System instabil und das Teilchen „springt“ in einen der Zustände ϕ_1 oder ϕ_2 (Zustandsreduktion)[34]. Die mittlere Zeit τ, innerhalb welcher die Zustandsreduktion eintritt, ist indirekt proportional zu E_G [99]:

$$\tau \approx \hbar / E_G$$

Die DP OR[35] sei ein nicht berechenbarer physikalischer Prozess und somit geeignet,

[33]Eine moderne Herangehensweise an das Messproblem erfolgt mithilfe der Theorie der Dekohärenz. Hier wird die Messung als komplexes und emergentes Phänomen von wechselwirkenden Systemen beschrieben. Beispielsweise würde bei einer Messung die Wellenfunktionen des System und des Messapparats interferieren und einen neuen Zustand ergeben. Die Schrödinger-Entwicklung des resultierenden Zustandes führt kontinuierlich und deterministisch zur Zustandsreduktion (dem "Kollaps") - tatsächlich gibt es keinen Kollaps. Hier gilt es zu erwähnen, dass die eben erwähnte Zustandsreduktion durch Dekohärenz eine „operationale“, jedoch keine ontologische Erklärung ist. [98]

[34]Die DP OR läuft innerhalb eines winzigen Planck-Volumens in der Raum-Zeit ab. Innerhalb solcher kleiner Entfernungen sind Effekte der Quantengravitation besonders stark sind. [101]

[35]Derzeit ist die Präzision moderner physikalischer Experimente zu gering, um eine DP OR zu beob-

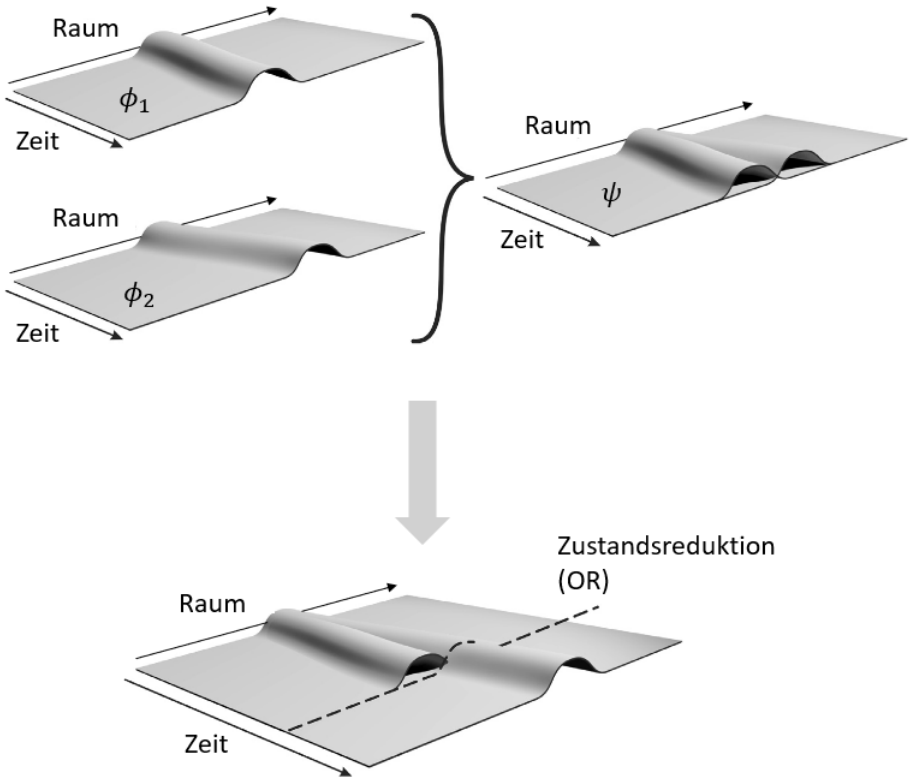

Abbildung 8: Illustration zur Veranschaulichung der OR nach Diósi-Penrose. Obere Bildhälfte: Zwei Wellenfunktionen ϕ_1 und ϕ_2 eines Teilchens T überlagern zu einer Gesamtwellenfunktion Ψ. Hier ist die Zeit-Dimension sowie eine räumliche Dimension dargestellt - tatsächlich ist die Wellenfunktion auf der vier-dimensionalen Raumzeit definiert. Ψ bestimmt die Wahrscheinlichkeitsverteilung des Ortes von T. Da das Teilchen T eine Masse hat, kann Ψ als proportional zur einer Massenverteilung aufgefasst werden. Die infinitesimalen Masseelemente innerhalb dieser Massenverteilung erzeugen ein Gravitationsfeld und werden wiederum selbst von dem Feld beeinflusst. Die Eigenenergie dieses Masse-Feld-Systems, also jene Energie, welche in der Feld-Konfiguration steckt, wird E_G genannt. Untere Bildhälfte: Überschreitet E_G einen gewissen Schwellenwert in der Größenordnung der Planck-Energie, dann kommt es spontan zur OR. Die mittlere Zeit, innerhalb welcher die OR eintritt, ist indirekt proportional zu E_G. Modifiziert nach [88] Abb. 8 und 9.

eine kausale (Mit-)Ursache jener physiologischer Vorgänge im Gehirn zu sein, welche für das Bewusstsein verantwortlich sind. Penrose bezeichnet die DP OR als kleinste „proto-bewusste Erfahrungen“, welche zwar keine Quale in eigentlichen Sinne darstellt, jedoch auf fundamentaler Ebene als Bausteine des Bewusstseins aufgefasst werden könnten. [88]

3.3.4 Kohärente Mikrotubuli-Zustände - Orch OR

MT sind die Hauptbestandteile des Zytoskeletts von eukarytischen Zellen, wie die Neuronen im Gehirn, und somit größtenteils für dessen mechanische Stabilität und Form verantwortlich. In der Mitose sowie für die Fortbewegung der Zelle in ihrem Milieu sind MT ebenfalls unverzichtbar. Außerdem dienen die länglichen Proteinkomplexe als Transportrouten, entlang welcher Vesikel durch Motorproteine (am häufigsten Dynein und Kinesin) geführt werden. Mit den Vesikeln können verschiedenste Stoffe in der Zelle von einem Ort zum anderen gebracht werden. Dies ist vor allem für den Transport von Neurotransmittern entlang des Axons wichtig und haben eine wichtige Funktion in Aktivität der Neuronen. ([103] S. 163-196)

Der Aufbau der MT ist hoch geordnet. Sie sind schraubenförmige Proteine, welche aus zwei verschiedenen Arten von Dimeren, dem α-Tubulin und β-Tubulin, zusammengesetzt sind. Die Tubuline sind abwechselnd miteinander nicht-kovalent zu Protofilamenten verbunden, welche wiederum die Wand eines röhrenförmigen MT bilden. Die MT im Menschen haben eine Durchmesser von 25 nm und sind mit einem Ende meist an ein MT-organisierendes Zentrum [104], von welcher ihr Wachstum ausgeht, verbunden. Es gibt in menschlichen Zellen etwa hundert verschiedene Isoformen des Tubulins und mehrere MT-Konfigurationen. [105]

Penrose und Hameroff schlagen vor, dass sich in den MT ähnliche Rechenvorgänge abspielen, wie in einem Quantencomputer[36]. Die Qubits[37] sind im Konformationszu-

achten. Ein quantenoptisches Experiment zur Messung einer DP OR in einem Superpositionszustand eines einzelnen Photons wird in [102] vorgeschlagen.

[36] Eine mögliche technische Realisierung eines Quantencomputers ist ein präpariertes Ensemble quantenmechanischer Zwei-Zustands-Systeme, welche jeweils ein Qubit (in der nächsten Fußnote erläutert) repräsentieren. Operationen an den Qubits selbst werden gezielt durch physikalische Prozesse, welche vom jeweiligen Aufbau des Quantencomputers abhängen, realisiert. Diese Operationen, auch Quantengatter genannt, haben eine sehr ähnliche Funktion, wie logische Gatter in konventionellen elektronischen Rechnern. Um das Resultat der Operationen, also der Berechnung, zu erfahren, wird an allen oder manchen Qubits eine Messung durchgeführt. Mit Quantencomputern können spezielle Algorithmen (Shor, Grover, Deutsch etc.) um mehrere Größenordnungen schneller als auf klassischen Rechnern ausgeführt werden. ([106] S. 171-200) [107]

[37] Qubits können im Gegensatz zu ihren klassischen Pendants, den Bits, einen Superpositionszustand aus 1 und 0 annehmen. Das klassische Bit ist daher im Zustand $\Psi = |1\rangle$ oder $|0\rangle$, wohingegen ein Qubit

stand der Tubulin-Dimere repräsentiert. Die Informationsverarbeitung passiert nun nach der U über die wechselseitige Beeinflussung der benachbarten Tubuline etwa analog zu einem zellulären Automaten[38]. Jeder Tubulin-Zustand wird von den Zuständen der benachbarten Tubuline beeinflusst, wodurch sich eine komplexe Dynamik ergibt, welche mit einer Berechnung assoziiert werden kann. Ein Zustand eines Tubulins bzw. der Gesamtzustand aller Tubuline kann, wie in Abschnitt 3.3.3 bereits erwähnt, als Masseverteilung aufgefasst werden. Die Masseverteilung der Tubuline im MT wiederum ist maßgeblich für das Eintreten der DP OR. Sobald die DP OR eintritt, reduzieren sich alle überlagerten Zustände zu reinen bzw. klassischen Zuständen, welche daraufhin MT-assoziierte Proteine (MAP)[39] die neuronale Aktivität modulieren können. Penrose und Hameroff haben den hier beschriebenen Gesamtprozess als durch die Tubulinzustände orchestrierte OR, kurz Orch OR, bezeichnet. [88] Die gesamte Vorgang Orch OR ist in Abbildung 9 beschrieben.

3.3.5 Kohärenzzeit von überlagerten Zuständen

Damit die Orch OR (und somit ein Bewusstsein) eintreten kann, muss die Superposition eines Zustandes des MT (oben Ψ genannt) lange genug aufrecht erhalten bleiben, damit die OR innerhalb einer mittleren Zeit τ eintreten kann. Die Zeit innerhalb welcher ein Zustand in Superposition, also zusammenhängend (kohärent), bleibt, heißt Kohärenzzeit. Diese Zeit kann aufgrund der Wechselwirkung der Umgebung mit dem System (MT) im kohärenten Zustand stark verkürzt werden[40]. Ein System muss also

eine Mischung $\Psi = \alpha|1\rangle + \beta|0\rangle$ sein kann, mit beliebigen Werten für α und β, solange Ψ normiert ist (daher $|\alpha|^2 + |\beta|^2 = 1$). ([108] S. 20-22) $|\cdot\rangle$ ist ein sogenanntes Ket aus der Dirac-Notation ([96] S. 117-119).

[38] Zelluläre Automaten sind Modelle von Systemen, dessen Dynamik in diskreten Zeitschritten abläuft. Ein bekanntes Beispiel für einen zellulären Automaten ist *Conway's Game of Life* - es besteht aus einem Raster, in welchem jede Zelle entweder weiß oder schwarz ist (zwei mögliche Zustände). In jedem Zeitschritt wird die Farbe einer Zelle durch die Zustände der direkten acht Nachbarn im letzten Zeitschritt bestimmt. ([109] S. 5 - 8) Dadurch kann sich eine komplexe Dynamik entwickeln und mit dem Game of Life lässt sich im Prinzip jeder Algorithmus berechnen ([109] S. 151 - 158).

[39] Mikrotubuli-assoziierte Proteine (MAP) haben verschiedene Funktionen innerhalb einer Zelle, besonders in Nervenzellen. Sie sind unter anderen für die Stabilität oder Instabilität sowie der Verbindung zwischen MT verantwortlich. Außerdem können MAP eine Wechselwirkung zwischen MT und anderen Proteinen vermitteln. Ein bekannter Vertreter der MAP-Typ II Familie ist das Tau-Protein (MAPT), welches in einer modifizierten Form in der Pathogenese von Morbus Alzheimer und anderer neurodegenerativen Erkrankungen eine große Rolle spielt. [110]

[40] Die Wirkung der Umgebung auf ein System, kann zum Beispiel anhand einer Schneeflocke im Feuer versinnbildlicht werden. Eine Schneeflocke hat eine sehr stark geordnete Struktur - hier als Analogie zu einem kohärenten Zustand. Sobald nun von außen - hier durch das Feuer - Wärme eingebracht wird, geht die Struktur sehr schnell verloren und die Schneeflocke verdampft. Die Zeit, innerhalb welcher die Struktur einer Schneeflocke im Feuer bestehen bleibt, könnte im übertragenen Sinne als Kohärenzzeit bezeichnet werden.

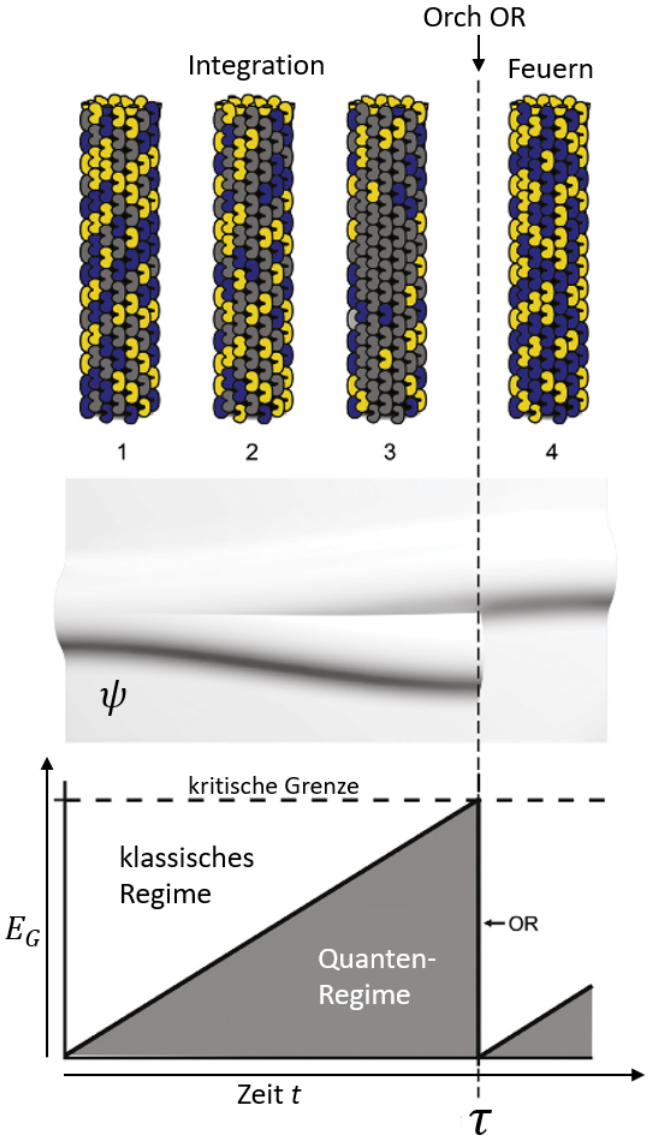

Abbildung 9: Es sind drei verschiedene Konfigurationen in den Tubulin-Zuständen in einem MT dargestellt (1,2 und 3). Die gelben Tubuline sind im reinen Zustand $|0\rangle$, die blauen in $|1\rangle$ und die grauen Tubuline sind in einem überlagerten Zustand $\alpha|1\rangle + \beta|0\rangle$. Durch die U der Zustände kommt es, mit der Zeit, ausgehend von reinen Zuständen zu einer zunehmende Überlagerung reiner Zustände Ψ in den Tubulinen (mehr graue Tubuline). Sobald die Eigenenergie des Gravitationsfeldes E_G eine kritische Grenze erreicht, tritt die DP OR ein, womit alle überlagerten Zustände wiederum in reine Zustände (4) übergehen. Die DP OR wird mit einem „elementaren Bewusstseinsmoment“ oder einer „proto-bewusste Erfahrung“ assoziiert. Modifiziert nach [88] Abb. 10.

von der Umgebung abgeschirmt werden, damit die Kohärenzzeit über τ liegt und im System eine spontane DP OR eintreten kann. ([111] Kapitel 2)

Max Tegmark schätzt die Kohärenzzeit von überlagerten Tubulin-Zuständen auf 10^{-13} s, welche zu kurz ist, damit Orch OR in neurophysiologischen Prozessen, welche sich auf einer Zeitskala von etwa 10^{-3} - 10^{-1} s abspielen, von Relevanz sein kann. Daraus folgert Tegmark, dass im Gehirn keine Quantenprozesse für die Leistungen des Gehirns, inklusive dem Bewusstsein, verantwortlich sind. Vielmehr lasse sich die Funktion des Gehirns auf eine klassische Informationsverarbeitung zurückführen. [89]

Hameroff et al. führen Argumente an, weshalb die Berechnungen von Tegmark ungültig seien. Zum einen sei das gewählte physikalische Modell nicht auf Mikrotubuli zutreffend, sondern ist ein starke Vereinfachung der tatsächlichen physikalischen Gegebenheiten. So sei beispielsweise nur ein überlagerter Zustand von Solitonen[41] in einem MT betrachtet worden, anstatt überlagerte Zustände von Tubulin-Konformationen zu berücksichtigen. Nach eigenen Berechnungen mit einem korrigierten Modell kommen die Autoren auf Kohärenzzeiten von 10^{-5} - 10^{-4} s. Überdies könnten Aktinfilamente das Wasser um die MT „ordnen“, wodurch dessen Entropie sinkt und die thermisch bedingte Dekohärenz abnimmt. In Anbetracht dessen würde sich die Kohärenzzeit von überlagerten Tubulin-Zuständen auf etwa 10^{-2} - 10^{-1} s belaufen, wodurch Orch OR einen Einfluss auf neurophysiologische Prozesse nehmen kann. [90]

[41] Ein Soliton ist ein Wellenpaket, welches sich durch ein Medium bewegt, ohne dabei seine Form zu verändern. Solch ein Wellenpaket kann in der Quantenmechanik als ein Teilchen aufgefasst werden und wird dann als „Quasiteilchen“ bezeichnet. ([112] S. 5 - 18)

4 Künstliche Intelligenz

4.1 Einleitung

Die Fortschritte der letzten Jahre in den Informationstechnologien gibt dem Forschungsgebiet der künstlichen Intelligenz (KI) neue Relevanz. Durch schnellere und immer besser verfügbare Computer sowie die rasche Entwicklung speziell von künstlichen neuronalen Netzen, nehmen KI-Techniken Einzug in beinahe jeden Bereich unserer Gesellschaft. Angefangen von selbst fahrenden Autos, über das Page-Rank-System von Internetsuchmaschinen bis hin zu Mustererkennung, welche Algorithmen mittlerweile besser beherrschen als der Mensch, sind die Anwendungsgebiete scheinbar grenzenlos.

Ein Indikator für die Relevanz der KI sind jüngste Investitionen und Erfolge von verschiedensten Forscherteams, Organisationen, Institutionen und Unternehmen weltweit. Folgender Auflistung soll ein Überblick über die aktuellsten Geschehnisse in der Welt der KI-Forschung vermitteln und exemplarisch die momentane Entwicklung aufzeigen:

- Das von Elon Musk mit ins Leben gerufene gemeinnützige Projekt OpenAI[42] hat mittlerweile ein Investitionsvolumen von etwa zwei Milliarden US-Dollar. Über 100 Mitarbeiter beschäftigen sich mit der Entwicklung von Anwendungen der KI zum Nutzen der Gesellschaft. Ihr Ziel ist die Entwicklung einer starken KI. [113] Unter den jüngsten Erfolgen sind ist der KI-Algorithmus „Generative Pre-trained Transformer - 2 (GTP-2)“, welcher aus einem 40 Gigabyte großen Korpus aus englischen Texten ein maschinelles Sprachverständnis erlernen konnte. GTP2 ist in der Lage einen Satz oder Absatz zu einen zusammenhängenden und sinnvollen Text fortzusetzen. Außerdem kann das Programm Texte zusammenfassen, Fragen zum Textverständnis beantworten und in andere Sprachen übersetzen. [114]Ein weiterer Erfolg ist das KI-Modell „OpenAI Five“, welches die Rolle der fünf Spieler eines Teams im E-Sport[43] Spiel *Dota 2* völlig autonom übernehmen kann. Erst kürzlich hat OpenAI Five die amtierenden Dota 2 Weltmeister „OG“ in zwei Spielen besiegt. [116]

[42] „artificial intelligence“ (AI) ist die englische Übersetzung von „künstliche Intelligenz“ (KI). In dieser Diplomarbeit wird, wenn möglich, die Abkürzung „KI“ verwendet.

[43] E-Sport bezeichnet das wettbewerbsmäßige Spielen von Mutiplayer-Computerspielen. Der jährlich an verschiedenen Orten der Welt durchgeführte Dota 2 Wettbewerb „The International“ fand zuletzt im August 2019 in der Mercedes-Benz Arena in Shanghai statt und war mit über 15 Millionen US-Dollar für den ersten Platz dotiert. Das Team „OG“ gewann den Bewerb bereits zum zweiten Mal. [115]

- Im Jahre 2018 hat die Kommission der Europäischen Union eine „Strategie für künstliche Intelligenz“ ein Investitionsvolumen von 20 Milliarden Euro für die Förderung der Entwicklung und Implementierung von KI-Techniken festgelegt. Besondere Schwerpunkte sollen auf a) die Bereitstellung von großen Datenmengen für das Training von KI-Modellen, b) die Förderung von künftigen Fachkräften und Ausbildungsprogrammen auf dem Gebiet der KI sowie c) die Entwicklung eines *Europäischen Ethik-Ansatzes in der KI* gelegt werden. Eine europaweite Zusammenarbeit zur Erreichung dieser Ziele wird angestrebt. [117]

  Nach Schätzung des Experten für Informationstechnologie Thomas Davenport wird die Volksrepublik China bis 2030 ein Gesamtinvestitionsvolumen von 30 Milliarden US-Dollar in die Entwicklung von KI und benachbarte Gebiete für staatliche Firmen bereitstellen. In den USA ist „AI Next“ mit einem Budget von 2 Milliarden US-Dollar das größte staatliche Förderprojekt auf dem Gebiet der KI. Jedoch im privaten Sektor werden die Investitionen spezifisch für die KI-Forschung von Apple, Amazon, Facebook, Google, Microsoft und International Business Machines (IBM) auf insgesamt 54 Milliarden US-Dollar geschätzt. [118]

- Das Londoner Unternehmen DeepMind, einer der Tochterfirmen von Google Inc., beschäftigt sich seit 2010 mit KI. Im Jahre 2016 konnte das bei DeepMind entwickelte künstliche neuronale Netz, genannt AlphaGo, im Brettspiel *Go* einen der besten Spieler der Welt, Lee Sedol, 4 zu 1 besiegen. Obgleich bereits in 1997 der Schachcomputer Deep Blue den damals amtierenden Weltmeister Garri Kasparow schlagen konnte ([119] S. 3-8), wurde Go bis zu diesem Zeitpunkt als zu komplex erachtet, um von einem Computer auf professionellem Niveau gespielt zu werden. [120] Eine Weiterentwicklung von AlphaGo ist AlphaZero. Das Besondere an diesem neuronalen Netz ist, dass es innerhalb weniger Tage die Spiele Go, Schach und Shogi[44] eigenständig lernte. Dabei spielte AlphaZero gegen ältere Versionen seiner selbst und wurde durch bestärkendes Lernen (reinforcement learning) so sehr rasch immer besser. Derzeit führt AlphaZero die Go Weltrangliste an. [122]

  Weitere Projekte von DeepMind sind AlphaFold, welches die Faltungsstruktur von Proteinen schneller und präziser als jeder andere Algorithmus (2019) vorhersagen kann [123]. AlphaStar ist ein tiefes neuronales Netz welches im Dezember 2018 einen der besten Spieler des Strategie Computerspiels StarCraft II in einer

[44] Shogi ist ein beliebtes japanisches Brettspiel, welches dem Schach sehr ähnlich ist [121].

Partie 5 zu 0 besiegte [124]. Außerdem wurde ein neuronales Netz entwickelt, welches Daten auf dieselbe Art abspeichern soll, wie das Gehirn Gedächtnisinhalte ablege. Diese Technik wird „Neural Turing Machine“ genannt. [125]

4.2 Was ist künstliche Intelligenz?

Eine exakte Definition von (künstlicher) Intelligenz ist ähnlich herausfordernd wie eine Definition des Bewusstseins. Vermutlich ist es sinnvoller, auch im Hinblick auf unser unvollständiges Verständnis der physischen und mentalen Prozesse im Gehirn, das Thema zu umschreiben.

Nach Russell und Norvig kann das Ziel der KI in zwei Dimensionen abgesteckt werden. Tabelle 2 zeigt, welche Zugänge es bezüglich dessen, was die KI erreichen soll, gibt. Zum einen sollte unterschieden werden, ob die KI einem Menschen nachempfunden wird oder rein auf eine abstrakte Rationalität[45] fokussiert. Andererseits kann die KI auf das *Denken* oder auf das *Verhalten* bezogen werden. Dies repräsentiert einen Großteil des Meinungsspektrums in der Fachliteratur zur KI und angrenzender Gebiete [127].

Tabelle 2: Möglichkeiten zur Charakterisierung des Ziels der KI in zwei Dimensionen. Adaptiert nach [126], Abbildung 1.1.

	Mensch	Rationalität
Denken	(A) KI denkt wie ein Mensch	(B) KI denkt rational
Verhalten	(C) KI verhaltet sich wie ein Mensch	(D) KI verhaltet sich rational

(A) KI denkt wie ein Mensch

Dieser Ansatz deckt sich mit dem Zugang der theoretischen Neurowissenschaften und der Kognitionswissenschaften, welche durch Computermodelle versuchen die neuronalen Prozesse im Gehirn abzubilden. Sollte jemand dazu in der Lage sein, wäre gleichzeitig eine KI als Abbild des Menschen geschaffen. Bisherige Modelle vom Gehirn, welche auf den Erkenntnissen von kognitiven Psychologie beruhen, werden *kognitive Architekturen* genannt. [128]

[45] Hier ist nicht gemeint, dass der Mensch nicht rational sein könnte. Vielmehr ist hier eine Unterscheidung von einem auch fehlerbehafteten und gelegentlich irrationalen, eben menschlichen Verhalten oder Denken zu einer rein idealen Vorstellung von Rationalität und Logik relevant. ([126] S. 2)

(B) KI denkt rational

Das logische Denken, abseits der Fehlbarkeit des menschlichen Intuition, basiert auf korrekten Schlussfolgerungen. Eine rationale KI, könnte *im Prinzip* jedes Problem, welches sich in der Sprache der Logik ausdrücken lässt, lösen. Dies lässt sich mithilfe von symbolischer Manipulationen nach definierten Regeln in einem Algorithmus umsetzen. Erste Erfolge gab es zu diesem Ansatz bereits in den 60er Jahren, als demonstriert wurde, wie Computerprogramme automatisch Beweise für mathematische Sätze[46] generieren konnten - eine Aufgabe, welche für gewöhnlich ein Mensch erst nach einer mehrjährigen Ausbildung in der Mathematik bewältigen kann. ([126] S. 4)

Eine nach dem Prinzip der Logik arbeitenden KI hat zwei Hauptkomponenten, eine *Wissensdatenbank* und eine *Inferenzmaschine*. Die Wissensdatenbank enthält das Wissen, welches als Prämisse für die logische Schlussfolgerung herangezogen wird. Jener Algorithmus, welcher nach den definierten Regeln der formalen Logik Wissen verarbeitet nennt sich Inferenzmaschine. Beim Lösen von allgemeinen Problemen stößt die rationale KI auf zwei fundamentale Schwierigkeiten ([130] S. 20-71):

- Das Wissen in der Wissensdatenbank muss zu Verwendung in der Inferenzmaschine in der Sprache der Logik abgelegt werden. Dies ist jedoch im Allgemeinen - spezielle bei informellem und unsicherem Wissen - äußerst schwierig oder gar unmöglich. Das Genannte ist eines der Hauptprobleme der Wissensrepräsentation. ([131] S. 1 - 11)

- In der Praxis stößt eine auf formaler Logik basierenden KI an die Grenzen der Rechenkapazität von heutigen Computern. Wo eingegrenzte Aufgaben, mit einer kleinen Menge an Lösungsmöglichkeiten, noch relativ leicht bewältigbar sind, ist es aus prinzipiellen Gründen nicht möglich eine Inferenzmaschine, welche allgemeine Probleme mit Logik löst, real auf einem Rechner auszuführen. Ohne Einsatz von heuristischen[47] Methoden und anderen erweiterten Konzepten aus

[46] Der erste bedeutende Beweis mithilfe eines Computers wurde 1976 durchgeführt. Das bis dahin nicht beweisbare Vier-Farben Theorem besagt, dass jedes Land in einer Landkarte mit vier verschiedenen Farben eingefärbt werden kann, sodass keine zwei Länder mit derselben Farbe aneinander grenzen. Das analoge Fünf-Farben Theorem konnte bereits im 19. Jahrhundert bewiesen werden. ([129] S. 1-11)

[47] Heuristische Methoden sind in der modernen Informatik ein Oberbegriff für jene Techniken, welche eine bestmögliche Lösung in der gleichzeitig bestmöglichen Zeit finden können. Dabei wird nicht jede Lösungsmöglichkeit untersucht, sondern nur jene, welche zuvor einem geschickt gewählten Kriterium genügen. Dies kann mittels verschiedenster Algorithmen passieren, zum Beispiel durch simuliertes Abkühlen, Fuzzylogik oder dem A*-Algorithmus. Dadurch wird die nötige Zahl der Rechenschritte entscheidend reduziert. Allerdings ist die produzierte Lösung im Allgemeinen nicht die beste, sondern lediglich eine der „guten“. [132]

der Informationstheorie würde es auf Grund der „kombinatorischen Explosion“[48] zu lange dauern, bis der Algorithmus eine Lösung zu einem beliebigen gestellten Problem gefunden hätte. [133]

(C) KI verhält sich wie ein Mensch

Da derzeit keine definitive und allgemeine Definition von KI vorliegt, ist es naheliegend sie mit den kognitiven Fähigkeiten eines Menschen zu vergleichen. Eine Maschine, welche sich gleich wie ein Mensch verhalten könnte, könnte sicher als „intelligent“ bezeichnet werden. Speziell in Hinblick auf die Anwendung einer allgemeinen KI in der Praxis scheint es unbedeutend, welche Abläufe zu einem menschlichen Verhalten geführt haben, entgegen dem Ziel (A). Im Kapitel 4.5 wird unter anderem der Turing Test erläutert, welcher als ein rein auf das Verhalten einer Maschine fokussierter Intelligenztest entworfen wurde. ([126] S. 2-3) Eine Kritik am Turing Test und Vorschläge für Testverfahren, welche ebenfalls eine KI mit den anderen beschriebenen Zielen (A, B, D) überprüfen, werden danach ebenfalls erläutert.

(D) KI verhaltet sich rational

Russel und Norvig vertreten die Position, dass die KI sich rational verhalten sollte (D). Für sie ist es nicht von Bedeutung, ob die KI ein menschenähnliches Verhalten oder gar Denken hat. Vor allem für die Praxis sei lediglich das zum Ziel führende Verhalten von Relevanz. Dieser Standpunkt ist jener, welcher von den meisten großen und erfolgreichen Unternehmungen (siehe Abschnitt 4.1) eingenommen wird. Eine KI sollte daher in jeder Situation, in welcher sie sich befindet, die beste mögliche Handlung vollführen. Dies entspricht dem Modell der *intelligenten Agenten*, welche mit ihrer Umgebung interagieren können. Sie bekommen Informationen, zum Beispiel über Sensoren, Datenschnittstellen etc., verarbeiten diese und handeln in Folge. ([126] S. 34-63) Abbildung 10 illustriert dieses Modell.

4.3 Starke und schwache KI

Eine feinere Unterteilung, welche die Einteilung in Tabelle 2 ergänzt, kann durch die Kategorien „starke“ und „schwache“ KI erfolgen. Auch hier handelt es sich um eine Spezifikation des Ziels der KI, welche, aufgrund der modernen Entwicklung auf dem

[48] Die kombinatorische Explosion war in den Anfängen der KI-Forschung ein starkes Gegenargument zu einer KI für allgemeine Probleme. Sie bezeichnet das exponentielle Anwachsen der nötigen Rechenschritte als Funktion der Lösungsmöglichkeiten zu einem allgemeinen Problem. [133]

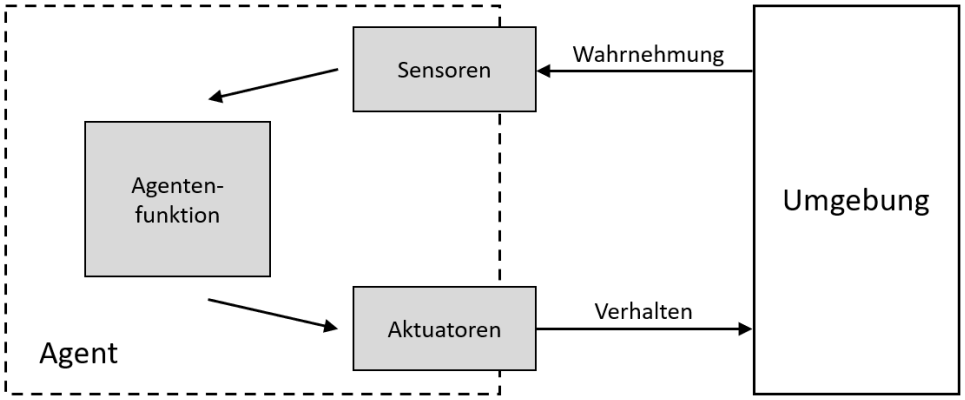

Abbildung 10: Ein intelligenter Agent zeigt ein an die jeweilige Umgebung angepasstes rationales Verhalten. Das Verhalten wird aus den Eindrücken des Agenten über seine Sensoren bestimmt. Eine Agentenfunktion verarbeitet diese Eindrücke deterministisch und bestimmt daraufhin das Verhalten des Agenten. Die Funktion im Agenten kann auf eine spezifische Aufgabe hin ausgerichtet oder trainiert werden. Adaptiert nach [126], Abbildung 2.1.

Forschungsgebiet den Einzug facettenreicher Anwendungen der KI in vielen Bereichen unseres Lebens, immer mehr an Relevanz gewinnt. Im folgenden sollen die zwei Begriffe näher definiert werden. [127]

- Eine KI wird als *stark* bezeichnet, wenn sie über die geistigen Fähigkeiten eines Menschen oder darüber hinaus verfügt. Ob ein phänomenales Bewusstsein ebenfalls ein Element der genannten Liste sein soll/wird ist Gegenstand kontroverser Diskussionen. Eine starke KI muss nicht notwendigerweise menschlich oder dem Menschen nachempfunden sein. Es geht eher darum, dass eine starke KI Fähigkeiten wie einen eigenen Antrieb, freien Willen, Lernfähigkeit, Anpassungsvermögen, sprachliche Kommunikation etc. hat und daher zur Lösung von allgemeinen Problemen befähigt ist. Derzeit ist die Realisation einer starken KI noch nicht möglich. Ihre Existenz hätte jedoch weitreichende ethische, rechtliche und gesellschaftliche Konsequenzen (siehe Abschnitt 7.4). [134]

- Wenn eine KI auf die Lösung von spezifischen Problemen ausgerichtet ist, jedoch über keine Anwendbarkeit über ein begrenztes Gebiet hinaus verfügt, wird von *schwacher* KI gesprochen. Zwar gibt es heute Algorithmen, welche manche Probleme besser lösen können, als ein Mensch, jedoch wird ihnen ein tieferes Verständnis abgesprochen. Alle derzeit existierenden KI-Systeme und Algorithmen

sind in diesem Sinne schwach, allerdings wird die Grenze des Möglichen immer weiter nach hinten versetzt. ([126] S. 1020-1021)

4.4 Beziehung (künstliche) Intelligenz und (synthetisches) Bewusstsein

Wir können mit Sicherheit davon ausgehen, dass der Mensch im Vergleich zu anderen Lebewesen auf der Erde einen hohen Grad an Intelligenz hat. Es wird vermutet, dass dies ebenfalls auf sein Bewusstsein zutrifft und dass Bewusstsein und Intelligenz in allen Lebenwesen mit einem Gehirn korreliert sind. [135] In diesem Sinne sind die Tiere in Abbildung 11, welches einen 2-dimensionalen Raum aus dem Bewusstseins- und Intelligenzgrad aufspannt, auf einer Geraden angeordnet. Es ist ebenfalls bekannt, dass die Intelligenz eines Tieres mit der Neuronenzahl in seinem zerebralen Kortex korreliert [136] [137]. Dieser Trend ist in der Grafik ebenfalls erkennbar. Abseits der biologischen Linie sind Computer, welche mithilfe von maschinellen Lernen usw. in speziellen kognitiven Fähigkeiten immer besser werden.

Nach der IIT (siehe Abschnitt 3.2) kann einem konventionellen Computer aufgrund der parallel laufenden und vorwärts gerichteten Informationsverarbeitung maximal ein rudimentäres Bewusstsein zugeordnet werden. Selbst wenn auf einem Supercomputer ein gesamtes menschliches Gehirn simuliert werden würde, dann würde kein Bewusstsein vorliegen [140].

Andererseits können physikalisch realisierte Schaltkreise eine Struktur aufweisen, welche neuronalen Netzen nachempfunden sind. Neuromorphe Chips bilden künstliche neuronale Netze direkt in Transistorschaltungen ab, wodurch sich diese stark von konventionellen Mikrochips unterscheiden. Obgleich derzeitige Realisationen, wie zum Beispiel „Loihi“ von Intel Corporation [139], nicht annähernd die Rechenkapazität eines gewöhnlichen Computers erreichen können und daher ihre Intelligenz limitiert ist, kann ihnen möglicherweise ein höherer Bewusstseinsgrad zugeordnet werden [140].

Ob in Zukunft eine KI mit einem synthetischen Bewusstsein verbunden wird und welche Folgen dies hat, ist Gegenstand der aktuellen Forschung. Eine Diskussion dieses Themas erfolgt in den Abschnitten 7.3 und 7.4.

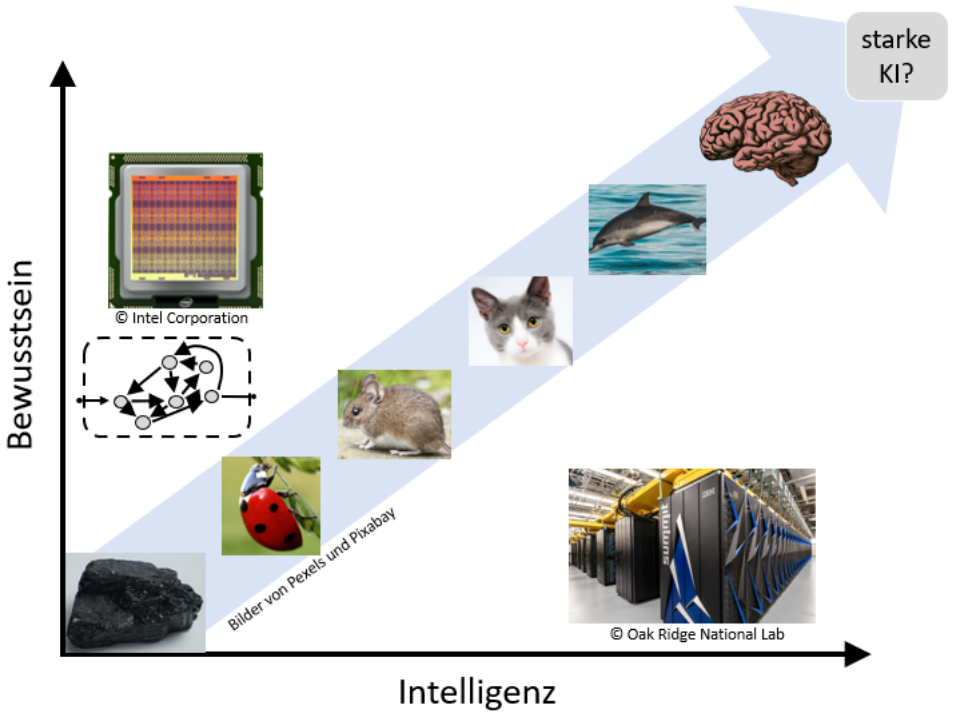

Abbildung 11: Biologische und künstliche Systeme können anhand ihrer Grade an Intelligenz sowie an Bewusstsein in einer 2-dimensionalen Ebene angeordnet werden. Über dem blauen Pfeil befinden sich die biologischen Systeme, hier in erster Linie Säugetiere, für welche eine Korrelation von Intelligenz und Bewusstsein vermutet wird. Wo etwa Einzeller oder Pflanzen ihren Platz in dieser Darstellung haben, bleibt bislang eine offene Frage. Rechts unten befindet sich Summit, der derzeit (2019) schnellste Supercomputer der Welt [138]. Auf der linken oberen Seite könnten rekurrente Systeme angeordnet werden, beispielsweise der neuromorphe Chip „Loihi“ von Intel, in welchem sich *in silico* neuronale Netze befinden [139]. Diese Darstellung ist an [135] angelehnt.

4.5 Die Messung von (künstlicher) Intelligenz

Wie lässt sich die Intelligenz einer Maschine messen? Wie lässt sich feststellen ob sie etwas versteht und richtige Schlüsse ziehen kann, also intelligent ist? Der Versuch solch einer Testung steht schon von Beginn an vor großen Schwierigkeiten. Für das Vorliegen von Intelligenz stehen keine einheitlichen, eindeutig zu testenden Kriterien in einer Maschine zur Verfügung stehen. Dieser Frage nachzugehen ist eines der Anliegen dieser Diplomarbeit. Turing versuchte diese Probleme zu umgehen und schlug einen behaviouristischen Test vor, in welcher ein spezieller Aspekt des Verhaltens, nämlich die sprachliche Kommunikation der Maschine mit jener des Menschen verglichen wird (Abschnitt 4.5.1). Einwände gegen dieses Testverfahren und weitere Entwicklungen werden in Abschnitt 4.5.2 behandelt.

4.5.1 Imitationsspiel

Der „Turing-Test“ im heutigen Sprachgebrauch bezieht sich auf ein im Jahre 1950 von Alan Turing vorgeschlagenes Experiment [141], welches intelligente Maschinen auf den Prüfstand stellen soll. Dabei wird die Frage, ob eine Maschine denken kann, auf folgende reduziert: Kann eine Maschine eine Unterhaltung führen - auf gleiche Weise wie ein Mensch es könnte? Als Mittel diese Fragen zu beantworten beschreibt Turing das sogenannte „Imitationsspiel“, bei welchem sich eine menschliche Befragerin/ein menschlicher Befrager, eine menschliche Vergleichsperson und eine Maschine in abgetrennten Räumen befinden. Die Befragerin/der Befrager kann mittels Textnachrichten mit seinen/ihren beiden unbekannten GesprächspartnerInnen kommunizieren - sie/er stellt eine Frage und das Gegenüber muss antworten - andere Möglichkeiten zum Informationsaustausch sollen unterbunden werden. Alle drei Parteien kennen den Versuchsaufbau und haben jeweils eine andere Aufgabe:

1. Die Befragerin/der Befrager weiß, dass sie/er ein Gespräch mit einem Menschen und einer Maschine führt. Sie/er soll durch gezielte Fragen entscheiden, wer Mensch und wer Maschine ist.

2. Die Maschine soll ein menschliches Verhalten imitieren und die Befragerin/den Befrager durch ihre Antworten davon überzeugen, dass sie ein Mensch sei.

3. Die menschliche Vergleichsperson hat die Aufgabe, die Befragerin/den Befrager zu unterstützen und sie/ihn zu den richtigen Schlüssen zu führen (dass die Ver-

gleichsperson der Mensch ist).

Der Turing-Test in seiner operationalistischen Interpretation umgeht die Notwendigkeit einer Definition von Intelligenz, Verstand und Bewusstsein und macht sich stattdessen die Tatsache zunutze, dass diese Eigenschaften für einen Menschen als solche erkannt werden können. Wenn etwas den Turing-Test besteht, ist es per Definition intelligent. Es gibt bei dieser „Definition“ jedoch einen Aspekt, der beachtet werden muss. Beim Turing-Test handelt es sich um ein induktives Verfahren, welches aufgrund der Fehlbarkeit der involvierten Menschen eine begrenzte Spezifität hat. Es ist daher notwendig eine Maschine wiederholt zu testen, um die Wahrscheinlichkeit einer richtigen Beurteilung - ob die Maschine intelligent ist oder nicht - zu erhöhen. Letztendlich bleibt dennoch jede Aussage, die der Turing-Test in der Lage ist zu tätigen, von probabilistischer Natur und kann daher nicht absolut (im Sinne einer Definition) gültig sein. [142]

4.5.2 Kritik an und Einwände gegen den Turing-Test als Kriterium für Intelligenz

Lucas-Penrose Argument

Die Unvollständigkeitssätze von Kurt Gödel zählen zu den wichtigsten Resultaten der mathematischen Logik und machen Aussagen über ein formales System, welches im folgenden mit S bezeichnet wird. Folgende Bedingungen muss das formale System erfüllen, damit die Sätze von Gödel angewandt werden können:

i) S ist hinreichend mächtig. Es muss ein Mindestmaß an Arithmetik in S enthalten sein. Zum Beispiel wäre ein System mit natürlichen Zahlen mitsamt einer Definition von Multiplikation, Addition und grundlegenden logischen Operationen ein hinreichend mächtiges System (Äquivalent: S erfüllt den Satz von Löb[49])

ii) S ist ausreichend konsistent/widerspruchsfrei. Es darf daher aus S nicht auf eine Aussage und gleichzeitig auf ihre Negation geschlossen werden.

iii) S ist rekursiv aufzählbar, also mechanisch. Das heißt, dass es einen Algorithmus gibt, welcher alle möglichen Aussagen, welche im System S formulierbar sind, aufzählen kann.

[49] Der Satz von Löb ist ein Resultat aus der mathematischen Logik. Der Satz kann in Worten wie folgt beschrieben werden: i) Eine Aussage x ist in einem formalen System S (enthält Peano-Arithmetik) beweisbar $\Rightarrow$ x ist wahr, ii) Satz von Löb: Ist i) beweisbar in S $\Rightarrow$ x ist beweisbar in S.

Erster Satz: *In einem formalen System S, welches die Bedingungen i) bis iii) erfüllt, existiert zumindest eine in S formulierbare, aber nicht beweisbare/entscheidbare Aussage x (Gödelsatz).*

Zweiter Satz: *Ein formales System S, welches die Bedingungen i) bis iii) erfüllt, kann seine eigene Konsistenz ii) nicht beweisen.*

Durch die Existenz einer nicht entscheidbaren Aussage x ist das formale System unvollständig. ([143] S. 1 - 53)

Nun könnte solch ein formales System in einer Rechenmaschine implementiert sein. Der Rechenmaschine kann eine Aussage eingegeben werden, woraufhin diese errechnet, ob es sich um eine wahre oder falsche Aussage, im Sinne des formalen Systems, handelt. Das in der Diskussion des Turing-Tests besonders relevante Lucas-Penrose Argument ([144] und [145] S. 113-115) läuft sinngemäß auf folgenden Gedankengang hinaus:

1. A ist eine Rechenmaschine, in welcher ein formales System im Gödel'schen Sinne S implementiert ist.

2. Daher gibt es eine Aussage x welche in S zwar formuliert, aber nicht entschieden werden kann (A unterliegt der Lucas-Penrose-Bedingung).

3. Für einen menschlichen Geist M sind genau drei Fälle möglich:

   (a) M ist nicht mechanisch, daher dürfen die Sätze von Gödel nicht angewandt werden, da dies der Bedingung iii) widerspricht.

   (b) M ist mechanisch jedoch inkonsistent, daher dürfen die Sätze von Gödel nicht angewandt werden, da dies der Bedingung ii) widerspricht.

   (c) M ist mechanisch und konsistent, daher existiert eine Aussage y, welche von einem Menschen nicht entschieden werden kann (analog zu 1 und 2).

4. M unterliegt nicht derselben Restriktion wie eine Maschine (M unterliegt nicht der Lucas-Penrose-Bedingung) und M kann jede Aussage (einschließlich x) entscheiden.

5. M ist konsistent.

6. Aus 4 und 5 folgt, dass 3a richtig sein muss: M ist nicht-mechanisch.

Die Bedingung i) kann aufgrund der Komplexität der hier in Frage kommenden formalen Systemen stets angenommen werden. Aus den genannten Argumenten (1 bis 5) kann geschlossen werden, dass eine Maschine prinzipiell den Turing-Test nicht tatsächlich bestehen kann. Etwaigen Täuschungen einer Maschine, zum Beispiel durch offene Formulierungen oder geschickt gewählte Aussagen, kann somit mit Sicherheit entgegnet werden. Um mit Sicherheit einen Mensch von einer Maschine unterscheiden zu können, braucht die/der menschliche BefragerIn bloß nach einem Beweis für die in 2 genannte Aussage x fragen. Die Maschine A kann diesen nicht erbringen - ein Mensch M jedoch schon. Wenn jemand die Argumentation etwas weiterführt (Argument 6), führt dies zu dem Schluss, dass der menschliche Geist nicht-mechanisch ist [146] und daher auch nicht berechenbar ist.

Es sind zahlreiche Kritiken zum Lucas-Penrose Argument in der Literatur zu finden [147] [148] [149], welche besonders die Aussagen 4 und 5 anfechten. In [150] konnte gezeigt werden, dass das Lucas-Penrose Argument zu Zirkularität oder Inkonsistenz führt und somit eigentlich ungültig ist.

Chauvinismus Argument

Wir messen Intelligenz mit unseren eigenen Maßstäben, die von uns Menschen abgeleitet sind. Mit dem Turing-Test vergleichen wir also eine, möglicherweise intelligente, Entität mit uns selbst und bewerten eine zu geringe Ähnlichkeit mit uns selbst als „nicht intelligent“. So sei es denkbar, dass ein andersartiges Wesen, Maschine oder sonstiges, sich in anderen Belangen, als seine Intelligenz, von Menschen unterscheidet und daher durch sein „Nicht-Mensch-Sein“ alleine am Turing-Test scheitert, obwohl seine Intelligenz wohl vorliegt. Eine intelligente Entität könnte beispielsweise nicht die menschlichen Gepflogenheiten kennen, könnte eventuell moralische Einwände gegen das Täuschen von Menschen haben, könnte eine völlig andere Sprache zur Kommunikation nutzen oder ähnliches. In solchen Fällen wäre der Turing-Test in seiner ursprünglichen Version nicht anwendbar. [151]

Daraus folgt, dass eine Anpassung der Interpretation des Turing-Test notwendig ist. Eine Möglichkeit wäre es, ihn als ein hinreichendes, jedoch nicht notwendiges Kriterium für das Vorhandensein von Intelligenz aufzufassen. Das heißt er kann das Vorhandensein von Intelligenz bestätigen, jedoch im Umkehrschluss kann ein negatives Testresultat nicht als “nicht intelligent“ gewertet werden. Eine weitere Strategie, um den die Sensitivität des Turing Tests zu erhöhen, bestünde darin die Strategie der Befragerin/des

Befragers zu ändern. Sie/Er könnte etwa, anstatt die Maschinen mit dem Menschen in ihrem „Mensch-Sein“ zu vergleichen und versuchen die Identität der Maschine zu erkennen, sondern bloß die Intelligenz der Maschine an sich bewerten, eventuell mit dem Menschen vergleichend. Ein solches Vorgehen könnte eine Quantifizierung, also eine abgestufte Bewertung, der Intelligenz ermöglichen. Versuche, die dieses Prinzip umsetzen, finden in Abschnitt 4.5.3 Erwähnung.

Determinismus Argument

Ein anderer Einwand beruht auf der Annahme, dass es vorstellbar ist, dass ein Satz an Regeln, die das Verhalten eines Menschen in jeder erdenklichen und möglichen Situation vorhersagen, existiert. Voraussetzung für die Gültigkeit dieser Annahme ist der Umstand, dass das Verhalten eines Menschen determiniert ist, was vor allem im Bezug auf die Diskussion zum Freien Willen besonders relevant ist. Außerdem nötig wäre es, dass das Verhalten eines Menschen sich in einem Regelwerk formulieren lässt (siehe Lucas-Penrose Argument). Falls dies gilt ist es a priori vorstellbar, dass eine Maschine, welche diese Regeln kennt, im Verhalten von einem Menschen ununterscheidbar ist.

Folgendes Gedankenexperiment von Ned Block [152] soll den Turing-Test als hinreichendes Kriterium für Intelligenz widerlegen: Es sei ein einem Menschen völlig ebenbürtiges Wesen, sein tatsächliches und potentielles Verhalten ist identisch (Disposition) - in jeder erdenklichen und möglichen Situation würde die Reaktion des Wesens gleich ausfallen, wie jene dieses einen Menschen. Jedoch sei ein Unterschied zugelassen, nämlich dass das Wesen weder Intelligenz noch Bewusstsein besitzt - keine mentalen Zustände aufweist, welche Bedeutung, Intentionalität oder Qualia entsprechen könnten, sondern das menschliche Verhalten lediglich „simuliert“. Eine mögliche Realisation eines solchen Wesens nennt sich „Blockhead“. Blockhead ist ein Wesen, welches völlig gleich aussieht wie ein Mensch, dessen Verhalten jedoch zur Gänze von einer Tabelle, in welcher eine Verhaltens-Antwort auf jede mögliche Sinneswahrnehmung bzw. Abfolge von Sinneswahrnehmungen von Blockhead verzeichnet ist, bestimmt wird. Da Blockhead prinzipiell in der Lage ist den Turing-Test zu bestehen, ohne jedoch intelligent zu sein, spricht gegen die Auffassung des Turing-Tests als hinreichendes Kriterium für Intelligenz.

1. Dem kann entgegengesetzt werden, dass Blockhead nicht logisch oder physikalisch möglich ist. Dieser Einwand lässt sich umformulieren, nämlich in den Zweifel, dass aus der (gegebenen) Vorstellbarkeit von Blockhead seine logische Möglichkeit

folgt (analytische Metaphysik). Eine detaillierte Diskussion zur Vorstellbarkeit und logischen Möglichkeit findet sich in [153].

2. Es kann gesagt werden, dass dem Verarbeitung der Sinneswahrnehmung, dem Suchen von Einträgen in einer Tabelle, dem Interpretieren des Eintrages, dem Steuern des daraus resultierendes Verhalten und weiteres einer Informationsverarbeitung entspricht, welche hinreichend komplex ist, um von „Intelligenz“ sprechen zu können. Das Argument kann insofern weitergeführt werden, dass jemand Blockhead mit einem ausgeklügelten Computermodell des menschlichen Gehirns ausstattet und sein Verhalten durch das Modell bestimmen lässt. Das Computermodell soll die Funktion des Gehirns perfekt nachahmen, also eine Analogie zum Gehirn darstellen. Es ist in diesem Falle argumentierbar, dass dieses Modell „intelligent“ ist. Vorausgesetzt ein solches Computermodell ist logisch möglich, so kann argumentiert werden, dass das Modell einer Abbildung von Wahrnehmung auf Verhalten entspricht und auf diesem Grunde in eine Tabelle im erwähnten Sinne überführt werden kann (Äquivalenz). Daraus folgt, dass Blockhead als intelligentes Wesen im behaviouristischen Sinne denkbar wäre. [154]

3. Leichter ist es zu verneinen, dass Blockhead Qualia oder Intentionalität (also Bewusstsein) hat. [151]

Originalitäts Argument

Dieses Argument geht auf Ada Lovelace, welche etwa 100 Jahre vor Turing lebte, zurück. Sie war der Meinung, dass eine Maschinen erst dann als intelligent erachtet werden kann, wenn sie etwas Originelles und Neues, wie es viele Menschen vorgezeigt haben, schaffen könnte. Lady Lovelace betrachtete Maschinen, im speziellen auf die von ihr mitentwickelte „analytische Maschine“ Bezug nehmend, als reine ausführende Einheiten, nur das zu tun in der Lage sind, was ihnen exakt aufgetragen wurde. In diesem Kontext ist der Turing-Test kein hinreichendes Kriterium für Intelligenz, da er die Kreativität einer Maschine nicht (explizit) testet. [155] Turing setzte in seinem Paper entgegen, dass auch Menschen tatsächlich nichts „wirklich Neues“ schaffen könnten, sondern stets von den Naturgesetzen und der Einflüsse, die auf sie wirken, abhängen und all ihr Tun sich davon in einem deterministischen Sinne ableiten lässt (siehe Determinismusargument). Genauso wie eine Maschine durch ihre Natur in ihren Möglichkeiten beschränkt ist, gilt dasselbe für den Menschen. Zum Beispiel, dass Maschinen durch

ihre Rechenleistung, Programmierung etc. beschränkt sind - Menschen sind durch ihre Biologie, Genetik, Erziehung und dergleichen beschränkt. [151] Bringsjord et al. [156] vertritt die Meinung, dass ein Computer uns nicht „überraschen“ kann, da seine Fähigkeiten bloß für ein für Menschen vorhersehbares Verhalten ausreichen. Da dadurch der Turing-Test nicht ausreichend ist, schlagen die Autoren einen alternativen Test vor. Der sogenannte „Lovelace Test“ soll in der Lage sein, die Fähigkeit einer Maschine zur Kreativität, zur Originalität oder zum Erschaffen von etwas Neuem zu überprüfen.

4.5.3 Alternative Tests

Seit Turing's Vorschlag in den 50ern gibt es einige Abwandlungen zum Turing Test und alternative Designs eines Test für intelligente Maschinen, welche einige der in 4.5.2 genannten Verbesserungsvorschläge umsetzen. In den folgenden Abschnitten seien nur eine Auswahl an Tests genannt, welche häufig zitiert wurden. Für weitere Tests, ähnlich zum Turing-Test, sei auf die Literatur verwiesen: Der Meta-Turing Test [157]; „the truly total turing test“ [158]; Der robotische Turing Test [159]; Turing Test für Qualia [160].

Lovelace Test

Der Lovelace-Test wurde als Verbesserung des Turing-Tests entworfen. Die Autoren vertreten die Meinung, dass Kreativität und Überraschung/Unvorhersehbarkeit ein besseres Kriterium für Intelligenz sind, als die Konversationsfähigkeit mit einem Menschen (wie es im Turing-Test der Fall ist). Der Test gilt genau dann als bestanden, wenn eine zu testende Maschine A, welche von einem/einer menschlichen ErbauerIn E gebaut wurde, folgende Voraussetzungen erfüllt:

1. A zeigt ein Verhalten V.
2. A kann das Verhalten V wiederholen - es ist ein einmaliger Fehler in A, welcher zu V geführt hat, ausgeschlossen.
3. E kann nicht erklären wie in A das Verhalten V entstanden ist, obgleich sie/er den Aufbau, die Programmierung und Funktionen von A kennt. [156]

Hier kann eingewandt werden, dass die Erbauerin/der Erbauer von einem sehr großem Computerprogramm, welches keinerlei Intelligenz aufweist, jedoch aufgrund seiner Komplexität, keine beliebig detaillierte Erklärung für ein spezielles Verhalten in endlicher Zeit liefern kann. Andererseits kann ebenso gesagt werden, dass es im

Prinzip immer möglich sein sollte, das Verhalten einer Maschine in einem gewissen Detailgrad zu durchleuchten, speziell wenn jemand den Aufbau und die Funktionen der Maschine sehr gut kennt (was von der Erbauerin/dem Erbauer zu erwarten wäre). Beide Aussagen werfen nun die Frage auf, welche Art von Erklärung und in welchem Detail diese im Lovelace-Test als gültig betrachtet wird.

Lovelace 2.0 Test

Eine alternative Formulierung ist der sogenannte „Lovelace 2.0 Test". Er stellt weniger die Frage nach einer Erklärung für ein gegebenes Verhalten, sondern legt Bedingungen fest, unter welchen die zu testende Maschine sich verhalten soll. Das Verhalten wird danach im Kontext der gestellten Bedingung bewertet.

1. Eine Maschine A soll ein Produkt P vom Typ T herstellen.
2. P muss einen Satz an Bedingungen B genügen, wobei $b_i \in B$ ein in natürlicher Sprache formulierbare Bedingung ist.
3. Eine menschliche Bewerterin/ein menschlicher Bewerter M, welche/welcher T und C festgelegt hat, überprüft und bestätigt, dass P von Typ T ist und die Bedingungen C erfüllt.
4. Eine unabhängige menschliche Referentin/ein unabhängiger menschlicher Referent R stellt fest, dass die Kombination aus T und C für einen durchschnittlichen Menschen als Bedingungen zumutbar wären. [161]

Bei diesem Produkt kann es sich um jegliche Art von Kunstwerk handeln, sei es eine Geschichte, ein Bild, ein Werkstück usw. Besonders die Kunst des Geschichte-Erzählens (und -Verstehens), und das damit zusammenhängende Verständnis von unserer Welt, wird bei [162] als starke Bedingung für (menschliche) Intelligenz gewertet. Dies ist nach dem Autor eine wichtige Unterscheidung der menschlichen Intelligenz zu zum Beispiel jener der Primaten.

Winograd Schema Challenge

Die AutorInnen schlagen mit der „Winograd Schema Challenge" einen Testablauf vor, welche auf direkterem Wege die Intelligenz einer Maschine testet, als der Turing-Test es tut. Beim Turing-Test ist es für die Maschine eine Notwendigkeit ihr „Mensch-Sein"

vorzutäuschen. Dies zieht nach sich, dass die Maschine eine virtuelle Person, mit einer Geschichte, Charakter und Fehler aufbauen und ihre Identität annehmen muss, um die menschliche Befragerin/den menschlichen Befrager zu täuschen. Dies stellt eine nicht notwendige Hürde für eine Maschine dar, da diese nicht ausschlaggebend für ihre Intelligenz ist oder gar davon ablenkt. Außerdem sei eine freie Konversation sehr ungebunden und würde eine Vortäuschung erleichtern, da es keine scharfen Kriterien gibt, um eine „echte“ Konversationen zu erkennen. Aus den genannten Gründen soll eine intelligente Maschine einem simpleren, konkreteren und eindeutigeren Test unterzogen werden. Dabei muss die zu testende Maschine ein Ensemble von sogenannten „Winograd Schemata“ mit menschlicher Treffsicherheit lösen. Durch die „Winograd Schema Challenge“ kann eine probabilistische Aussage über die Intelligenz einer Maschine getroffen werden. Diese Schemata werden nach folgender Bildungsvorschrift konstruiert:

1. Zwei Substantivgruppen, a und b, werden in einem Satz S genannt. Die Gruppen a und b beziehen sich auf zwei unterschiedliche Parteien und es kann sich um Personen oder Gegenstände handeln. Besonders soll darauf geachtet werden, dass die beiden Substantivgruppen auf der Beudungsebene eine große Ähnlichkeit aufweisen und damit austauschbar sind. So wäre zum Beispiel „Peter“ und „Lukas“ besser geeignet, als „Peter“ und „Flugzeuge“.

2. Außerdem enthält der Satz S ein Possesivpronomen P, welches sich auf eine der Parteien bezieht, jedoch auch grammatikalisch passend ist für die andere Partei. Dass P für eine Substantivgruppe besser passt als für die andere muss ausgeschlossen werden.

3. Nach dem Satz S folgt eine Frage F nach dem korrekten Bezug des Possesivpronomens P. Zwei Antwortmöglichkeiten stehen zur Auswahl.

4. Im Satz S befindet sich ein spezielles Wort W. Wenn jemand W durch ein anderes Wort $\overline{W}$ ersetzt, dann dreht sich der korrekte Bezug von P um. Eines der speziellen Wörter wird bei Anwendung eines Winograd Schemas zufällig ausgewählt, wodurch sich eine mögliche Asymmetrie des Bezugs von a/b zu P unter Vertauschung der beiden Substantivgruppen im Mittel aufgehoben wird.

Ein Beispiel für solch ein „Winograd Schema“ stammt von Terry Winograd selbst [163] (weitere Winograd-Schemata unter [164]):

> Die Stadträte haben sich geweigert den erzürnten DemonstrantInnen eine Genehmigung zu erteilen, da **sie** Gewalttätigkeiten *befürchteten.* Wer *befürchtete* Gewalttätigkeiten?
>
> Antwort 0: Die Stadträte
>
> Antowrt 1: Die erzürnten DemonstrantInnen

Bei den unterstrichenen Substantivgruppen handelt es sich um die Parteien a und b. Das Possesivpronomen P ist fett gedruckt und das spezielle Wort W kursiv gedruckt. Es erscheint offensichtlich, dass Antwort 0 richtig ist - Antwort 1 wäre zwar logisch konsistent, jedoch in unserem Weltverständnis äußerst unplausibel. Nun könnte W „befürchtete“ durch ein $\overline{W}$ „befürworteten“ ersetzt werden. Dadurch wird der Bezug von P umgekehrt und Antwort 1 ist korrekt.

Die Schemata sollen so konstruiert werden, dass es zur korrekten Auflösung nötig ist, die im Satz S präsentierte Situation im Kontext eines Hintergrundwissen über unsere Welt zu begreifen. Das nötige Hintergrundwissen und die sprachliche Auffassungsgabe, sowie die Fähigkeit mithilfe dieser beiden einen korrekten Schluss zu ziehen, kann mit unserem Verständnis von Intelligenz assoziiert werden [165]. Unvorteilhaft wäre es, wenn ein Programm rein durch Anwendung grammatikalischer Regeln oder Selektionsrestriktion[50], ein Winograd Schema lösen könnte, jedoch ohne ein Verständnis des Inhaltes vom Satz S. Dies sollen die Anforderungen 1 bis 4 möglichst gut sicherstellen. [166]

Es existieren Abwandlungen der Winograd Schemata in der Literatur. „Winograd Sätze“ sind zwei Sätze, wobei der Bezug des zweiten Satzes auf den ersten zweideutig ist. Ein Beispiel wäre: *Marie hat Susanne beschimpft. Sie hat sie geschlagen.* Nun ist es die Frage, wer hat wen geschlagen? Hat Marie Susanne beschimpft und geschlagen; oder hat Susanne Marie geschlagen, weil sie beschimpft wurde? [167] Eine weitere Variante wäre "Choice of Plausible Alternatives (COPA)“, wobei nach möglichen Konsequenzen oder Kausalitäten für einen Sachverhalt gefragt wird, zum Beispiel: *Ich klopfe an die Tür meines Nachbarn. Was passierte daraufhin? A: Mein Nachbar hat mich ins Haus eingeladen. B: Mein Nachbar verließ das Haus.* [168]

[50] Die Selektionsrestriktion ist das Nutzen von bekannten Korrelationen von Wörtern und Wortgruppen in einer gewissen Sprache. [166]

5 Anwendungen physikalischer Modelle von Bewusstsein in der Medizin

Im folgenden Abschnitt sollen mögliche Anwendungsgebiete und Implikationen der Theorien des Bewusstseins in der Medizin erörtert werden. Besonders soll auf erklärende und überprüfbare Aussagen in den medizinischen Fachgebieten eingegangen werden. Dies ist bislang weniger auf die klinische Praxis ausgerichtet, sondern betrifft eher die Grundlagenforschung. In Zukunft könnten sich aus den gewonnen Erkenntnissen Konsequenzen für die klinische Praxis ergeben.

5.1 Medizinische Implikationen aus der Integrated Information Theory

In diesem Abschnitt sollen Aussagen der IIT über das menschliche Gehirn und Erklärungen von neurowissenschaftlichen Befunden beispielhaft besprochen werden. Für eine nähere Erklärung der IIT und der verwendeten Begriffe siehe Abschnitt 3.2.

5.1.1 Wo liegt das Bewusstsein im Gehirn?

Wo im Gehirn liegen die neuronalen Strukturen, welche für das Bewusstsein verantwortlich sind? Die IIT legt ein klares Kriterium für solch eine Struktur fest. Sie muss ein physikalisches Substrat eines Komplexes sein, welcher im gesamten Nervensystem des Körpers ein globales Maximum an integrierter Information Φ aufweist. Mittels eines Computermodells, welches die Prinzipien der IIT algorithmisch umsetzt, könnten die Zentren für integrierte Information ausfindig gemacht werden. Aufgrund der Komplexität des Gehirns, der unvollständigen Kartographie der neuronalen Netze[51] sowie der beschränkten zur Verfügung stehender Rechenleistung[52] ist eine Simulation bislang (noch) nicht möglich. Jedoch kann eine vergröberte Analyse der verschiedenen

[51] Im Jahre 2013 wurde in den USA das Projekt „BRAIN Initiative“ ins Leben gerufen. Es ist ein gemeinsames Forschungsvorhaben von über 30 Universitäten, Institutionen und Unternehmen weltweit zur systematischen Kartographie des menschlichen Gehirns. Bis zum Jahre 2025 soll das Projekt eine staatliche Finanzierung von insgesamt 4,5 Milliarden US Dollar bekommen. [169] [170]

[52] Die Wissenschaftler des *Blue Brain Project* unter Henry Markram simulieren eine sogenannte Blue Column (entsprechend einer neokortikalen Säule mit etwa 10000 Neuronen) auf einem Supercomputer mit einer maximalen Rechenleistung von 360 TeraFLOPS (Floating Point Operations Per Second), also 360 Billionen Gleitkommaoperationen pro Sekunde. Eine Simulation des gesamten Gehirns würde auf diese Weise - linear hochgerechnet - eine Rechenleistung von etwa 400 ExaFLOPS benötigen. [171] Der derzeit (Oktober 2019) schnellste Supercomputer der Welt (Summit am Oak Ridge National Laboratory, USA) im hat eine maximale Rechenleistung von etwa 150 PetaFLOPS [138].

Netzwerktopologien in den funktionellen Einheiten des Gehirns mittels IIT allgemeine Aussagen über deren integrierte Information erbringen. [140] Im Folgenden sei exemplarisch eine Analyse des Kleinhirns in Bezug auf sein mögliches Bewusstsein erläutert.

Bewusstsein im Zerebellum

Obwohl das Kleinhirn[53] nur ein Zehntel der Masse des Gehirns einnimmt, so beherbergt es etwa fünf mal Mehr Neuronen[54]. Anatomisch besteht das Zerebellum besteht aus zwei Hemisphären, einer Mittelzone, dem Wurm sowie dem Lobus flocculonodularis. Es ist über drei Pedunculi, über welche die efferenten und afferenten Fasern laufen, mit dem dorsalen Hirnstamm verbunden. Diese Fasern kommunizieren über die tiefen Kleinhirnkerne[55] mit dem zerebellären Kortex. ([177] S. 155-170) Primär ist das Zerebellum für die sensomotorische Koordination von Bewegungsabläufen zuständig. Eine mögliche Beteiligung an kognitiven Funktionen wie Aufmerksamkeit und Sprache sind noch nicht ausreichend geklärt, obgleich dies aufgrund der Verbindung des Kleinhirns mit nicht-sensomotorischen Gehirnarealen plausibel scheint [178].

Der zerebelläre Kortex besteht aus sich wiederholenden Strukturen, welche als Zonen bezeichnet werden - jede Zone kann wiederum in Mikrozonen unterteilt werden. Mikrozonen bildet funktionelle Einheiten und sind definiert als Gruppen von Purkinjezellen, welche einem somatotopen rezeptiven Feld entsprechen [179], daher ist jeder Teil des Körpers mit einer oder mehrerer spezifischer Mikrozonen verbunden[56]. Bündel von Kletterfasern, von einem spezifischen Teil des unteren Olivenkerns entspringend, erreichen die Purkinjezellen einer spezifischen Mikrozone. Die Verbindungen der proximal zur Purkinjezellschicht liegenden Korbzellen bleiben weitgehend auf eine Mikrozone beschränkt. Verbindungen zwischen den Mikrozonen sind daher weitaus geringer ausgeprägt, als zwischen den Neuronen innerhalb der Mikrozone. Apps und Garwicz beschreiben funktionelle Module im Kleinhirn, als „multizonal microcompartment“, welche aus mehreren Mikrozonen zusammengesetzt sind und eigenständig sowie weitgehend unabhängig voneinander arbeiten. [180]

Die funktionelle Strukturierung des zerebellären Kortex in Mikrozonen weist auf

[53] In diesem Abschnitt wird das menschliche Gehirn diskutiert. Die genannten Verhältnisse und daraus gezogene Schlüsse können ebenfalls auf das Gehirn der Primaten und anderer Wirbeltiere übertragen werden[172].

[54] Je nach Autor wird die Anzahl der Neuronen im Kleinhirn auf 70 [173] bzw. 101 [174] sowie im zerebralen Kortex auf 12-15 [175] bzw. 21-26 [176] Millionen Neuronen geschätzt.

[55] Kleinhirnkerne: Nuclei dentatus, emboliformis, fastigii und globosus ([177] S. 165-167)

[56] Diese funktionelle Organisation lässt sich ebenfalls im primären sensorischen und primären motorischen Kortex (S1 bzw. M1) erkennen ([177] S. 220-232).

eine stark parallele Datenverarbeitung im Kleinhirn hin. Aus neurochemischer Sicht kann eine solche Kompartimentierung nachvollzogen werden. Eine Immunfärbung von Zebrin II/Aldolase C in den Purkinjezellen legt eine streifige Struktur, orthogonal zu den Sulci der Kleinhirnrinde, offen. Die Streifen entsprechen etwa den Zonen. [181]

Die Informationsverarbeitung im Kleinhirn funktioniert nach dem Feedforward-Prinzip. Im Gegensatz zum zerebralen Kortex, in welchem sich größtenteils rekurrente neuronale Netze befinden, wird im Kleinhirn ein eingehendes Signal geradlinig verarbeitet und ausgegeben. Wenige rekurrente Neuronen wirken stets inhibitorisch, wodurch im Kleinhirn keine selbst erhaltenden neuronalen Aktivitätsmuster möglich sind. ([182] S. 311).

Das Kleinhirn lässt sich als ein neuronales Netz, welches seine Reize parallel und im Feedforward-Prinzip arbeitet, betrachten [183]. Aus Sicht der IIT ist die neuronale Topologie des Kleinhirns schwach integriert. Durch die starke Kompartimentierung und geringe Ausprägung der Verbindungen unter den Modulen im Kleinhirn lässt es sich in funktionell weitgehend unabhängige Teile zerlegen. Dadurch hat das Kleinhirn isoliert betrachtet ein sehr geringes Φ^{max} und daher kein oder ein sehr rudimentäres Bewusstsein. [80] Ändert sich dies, wenn es als Teil des gesamten Gehirns betrachtet wird?

Das Zerebellum ist innerhalb des Gehirns stark vernetzt. Über seine starken Verbindungen zur primären motorischen Rinde, sowie zum Thalamus und Mittelhirn hinaus, bestehen direkte bidirektionalen Projektionen zu großen Teilen des gesamten Neokortex. Dedizierte Teile des Nucleus dentatus kommunizieren mit nicht motorischen Arealen besonders des präfrontalen und posterior-parietalen Kortex. Außerdem seien Verbindungen zwischen dem Kleinhirn und den Basalganglien festgestellt worden. [184]

Mittels der IIT kann am Computer ein neuronales Netz simuliert und nicht-reduzierbare Bereiche identifiziert und quantifiziert werden. Das Kleinhirn, genauso wie andere Teile des menschlichen Gehirns, ist zu komplex für eine vollständige Analyse mit der IIT. Aufgrund von Computersimulationen kleinerer, ähnlicher Systeme kann auf Prinzipien geschlossen werden, welche sich möglicherweise auf größere Systeme übertragen lassen. In Simulationen zeigt sich, Systeme, welche so aufgebaut sind wie das Kleinhirn, trotz der starken Verbindungen zu einem integrierten System (wie der Neokortex) nicht Teil seines Komplexes sind. Daraus folgt, dass das Kleinhirn keinen oder einen sehr geringen Anteil am Bewusstsein hat. [83]

5.1.2 Split-Brain

Wenn der Balken im Gehirn (teilweise) durchtrennt ist, wie etwa bei bei einer Callosotomie zur Behandlung therapieresistenter Epilepsien, so sind Verbindungen zwischen den Mechanismen und Systemen der beiden Gehirnhemisphären bidirektional unterbrochen. Diese massive Störung in der UWS führt nach der IIT zu einer Modifikation der Quale, nämlich der Aufteilung des Bewusstseinsstroms auf die linke und die rechte Gehirnhälfte. [185]

Die Autoren der IIT erwähnen folgendes Gedankenexperiment: Das Bewusstsein im Gehirn entspricht einem Komplex über einen großen Teil des Gehirns. Wenn der Balken nun langsam, also Verbindung für Verbindung durchtrennt würde, so würde nach dem Kappen einer gewissen Verbindung der Zeitpunkt kommen, in welchem die UWS in zwei ähnliche UWS aufgeteilt wird. Dies ist genau dann der Fall, wenn der zuvor einzelne Komplex mit einem Φ^{max} sich in zwei Komplexe mit ähnlich großem Φ^{max} aufspaltet. Somit haben zwei Konzepte, das heißt zwei bewusste Erfahrungen, zur selben Zeit eine ähnliche Aktualität im Gehirn. Dies deckt sich mit den phänomenologischen Befunden in Split-Brain PatientInnen. [83]

5.1.3 Integrierte Information während Meditation

Die Auffassung, dass nur dann Informationen durch Neuronen verarbeitet werden können, wenn diese aktiv sind, ist weit verbreitet in den Neurowissenschaften [186]. Informationen innerhalb des Neuronenverbandes werden durch neuronale Aktivitätsmuster repräsentiert und können nur auf diese Art zu einem Bewusstsein beitragen. Im Gegensatz dazu sagt die IIT aus, dass es eine Informationsverarbeitung zur Erzeugung von bewussten Erfahrungen auch ohne Aktivität geben kann. Für eine bewusste Erfahrung ist nämlich nicht etwa das Maß an Aktivität entscheidend, sondern die Form der UWS sowie der Nicht-Reduzierbarkeit Φ^{max} des Komplexes im physischen System. In meditativen und hypnotischen Zuständen, in welchen das Gehirn ein reduziertes Aktivitätsmuster zeigt, treten veränderte bewusste Erfahrungen auf. Diese Erfahrungen entsprechen einer inneren Aufmerksamkeit und einem erhöhten Grad an Bewusstsein mit weniger Bewusstseinsinhalten. Dies deckt sich mit den Aussagen der IIT. [83]

5.1.4 Funktionell equivalente Systeme

Nach dem Identitätstheorem der IIT ist zwar die Quale mit dem bewussten Erleben, jedoch nicht mit dem physischen System identisch. Daher können zwei funktionell äquivalente physische Systeme unterschiedliche Bewusstseinsgrade haben. System A ist integriert und daher nicht reduzierbar und System B ist ein Feedforward-System, welches keinerlei Rückkoppelung aufweist und reduzierbar ist. Obgleich A und B funktionell identisch sind, hat A einen Komplex, bestehend aus mehreren Konzepten, und eine Quale, ein (wenn auch sehr kleines) Bewusstsein. Im System B gibt es keine Konzepte und sein Φ^{max} ist null, daher hat es kein Bewusstsein. System B kann als „Zombie“ von A bezeichnet werden, da es von einer externen Perspektive gleich scheint - das innere Erleben ist jedoch gänzlich anders. Schematisch dargestellte Beispiele für ein System A und B sind in Abbildung 12 ersichtlich. Reale integrierte und nicht-integriertes Systeme sind anschaulich in [82], Abbildung 21, dargestellt.

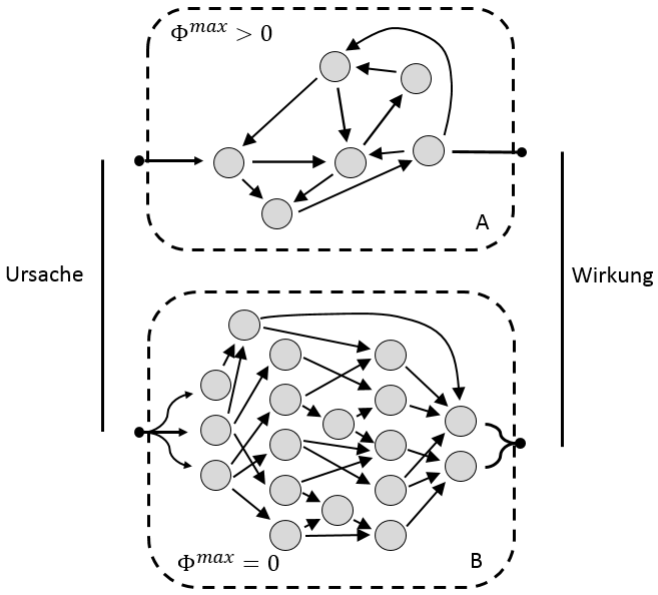

Abbildung 12: System A hat Verbindungen welche innerhalb des Netzes sowohl nach vorne, also auch zurück gehen. Somit hat der Effekt eines Elementes wiederum eine Einfluss auf seine Ursache. Diese Rückkoppelungen sind der Grund, weshalb das System nicht reduzierbar, und daher integriert ist. Hingegen ist das funktionell äquivalente Feedforward-System B reduzierbar. Seine Verbindungen zeigen stets in Richtung des Ausganges (Wirkung) und haben keinen Rückbezug.

Ein bekanntes Beispiel für ein System vom Typ A ist der Neokortex im Gehirn. Die meisten Verbindungen in den neuronalen Netzen des Neocortex sind rekurrent, also rückgekoppelt. Ein Netz mit stark ausgeprägter Rückkoppelung ist schwer zu „ent-

wirren“ und daher weniger reduzierbar, was in der IIT das Vorliegen von Quale nach sich zieht. Hingegen Systeme vom Typ B sind wenig oder nicht rückgekoppelt. Die Signalweiterleitung läuft in eine bestimmte Richtung durch eine Gruppe von Elementen. Künstliche neuronale Netze, welche gut trainiert werden können und zum Beispiel in der Mustererkennung eingesetzt werden, arbeiten nach dem Feedforward-Prinzip. [187] Eine Simulation des gesamten Gehirns [171], welche im Abschnitt 5.1.1 bereits angesprochen wurde, könnte im Prinzip die Funktion des Gehirns vollständig abbilden. Nach dem Funktionalismus (Abschnitt 1.5.1) entsprechen funktionelle Zustände mentalen Zuständen. Aus dieser Annahme folgt, dass eine vollständige Simulation des menschlichen Gehirns ein gänzlich identisches Bewusstsein zu seinem biologischen Äquivalent hat. Aus den Prinzipien der IIT folgt jedoch, dass ein in der Funktion zum Gehirn gleichen System, nicht denselben Komplex haben muss, da dieser von der Nicht-Reduzierbarkeit, und nicht von der Funktion, des Systems abhängt. Bei einer Simulation auf heutigen Computern ist es eher naheliegend, dass aufgrund dessen modularen Aufbau und der Feedforward-Informationsvermittlung kein Bewusstsein vorliegen kann. [188]

5.1.5 Freier Wille

Freier Wille wird assoziiert mit einer bewussten Entscheidung, welche durch innere Einflüsse im Bewusstsein - im Gegensatz zu äußeren Einflüssen - entsteht. Eine Autonomie aus innerer Perspektive besteht nach der IIT nur in nicht reduzierbaren physischen Systemen. Obgleich die einzelnen Elemente im System keine freien Willen haben, so ist es die integrierte Verbindung zwischen den Elementen, welche eine innere Dynamik, das Bewusstsein und freien Willen erzeugen. In Systemen ohne Quale, wie zum Beispiel in heutigen Computern oder in Feedforward-Systemen, gibt es keine innere Perspektive sondern nur eine äußere. Wenn einzig die äußeren Einflüsse das Verhalten eines Systems bestimmen, dann kann hierin ein Bewusstsein maximal als *Epiphänomen* betrachtet werden und jede Entscheidung, welche das System trifft, ist gänzlich von außen determiniert.

Häufig wird die Frage nach Freien Willen zusammen mit Determinismus gestellt. Im Sinne des Kompatibilismus [189] erklärt die IIT wie sowohl freier Wille als auch die Determiniertheit eines integrierten Systems zusammenhängen. Determinismus ist eine grundlegende Eigenschaft einer UWS. Jeder Indeterminismus würde diese Struktur durch eine Minderung des Zusammenhangs zwischen Ursachen und Wirkungen redu-

zieren. Ohne einer UWS kann es keine Quale geben und somit kein Bewusstsein, keine innere Perspektive, keinen freien Willen. Hier führen die Autoren folgendes intuitives Argument an: Würde jemand sich nach dem freien Treffen einer Entscheidung erneut in derselben Situation befinden, dann würde diese Person wiederum gleich entscheiden. Denn die Entscheidung würde von einem „Ich“, mit Werten, Erinnerungen und anderen inneren Einflüssen, bewusst aus einer inneren Perspektive getroffen werden. Dies wäre im Einklang mit dem Determinismus sowie dem freien Willen. [79]

5.2 Orch OR Theorie in der Medizin

5.2.1 Effekt von Anästhetika auf Mikrotubuli

Volatile Anästhetika, wie Sevofluran, Desfluran, Isofluran, Halothan und Distickstoffmonoxid (N_2O) haben Einfluss an mehreren Stellen innerhalb der Nervenzellen. Potentielle Wirkorte in den Neuronen sind sowohl Ionenkanäle und Rezeptoren an Zellmembranen (besonders $GABA_A$) [190], als auch verschiedenste Proteine im Zytoplasma (zum Beispiel Proteinkinase C [191]) [192]. Die genauen molekularen Mechanismen, über welchen inhalative Anästhetika in der Allgemeinanästhesie einen Verlust des Bewusstseins bewirken, sind noch nicht genau erforscht. Es wird zur Zeit eine Wirkung von Anästhetika auf das Bewusstsein über mehrere Wege diskutiert. [193].

Eines der Targets von inhalativen Anästhetika sind Tubuline, Bestandteile der MT, welche eine große Rolle in der Funktion von Nervenzellen spielen und in der Orch OR Theorie als ein physikalisches Substrat für Bewusstsein beschrieben werden [88]. MT sind besonders relevant für die Funktion von $GABA_A$-Rezeptoren [194] sowie für Ca^+- und Na^+-Kanäle in Neuronen [195] [196].

Craddock et al. konnten in einer Molekulardynamik-Simulation[57] eines α,β-Tubulin-Dimers Bindungsstellen für Halothan finden. An jeweils 16 verschiedene Stellen im α- und β-Tubulin bindet Halothan über London-Van der Waals Kräfte an hydrophobe Taschen[58] des Proteins. Die Bindungsenergien und Bindungswahrscheinlichkeiten konnten

[57] In Molekulardynamik-Simulationen können zur Modellierung von komplexen Systemen, wie zum Beispiel Moleküle oder Gase, verwendet werden. Mittels eines Computer-Algorithmus werden die Wechselwirkungen zwischen den Atomen (meist sehr viele) in der Zeit konkretisiert und die Bewegung des Teilchen-Ensembles schrittweise berechnet. Je feiner die zeitliche Diskretisierung ist, desto genauer aber auch rechenintensiver ist die Simulation. Molekulardynamik-Simulationen werden häufig in der Biophysik eingesetzt und mit ihnen lassen sich physikalische Eigenschaften über das untersuchte System finden. [197]

[58] Mehrere örtliche benachbarte ungeladene und nicht-polare Seitenketten in einem Protein bilden hydrophobe Taschen. Sie sind für die Proteinstruktur während des Faltens der Proteine sowie für Konformationsänderungen beim Binden eines Liganden verantwortlich. [198]

in der Simulation für alle Stellen quantifiziert werden. [199]

Wenn ein Halothan-Molekül an das Tubulin über eine hydrophobe Tasche bindet, so kommt es zu einer Konformationsänderung des Proteins. Die Änderung in der relativen Position der Atome zueinander führt zu einer Veränderung der Gravitations-Eigenenergie E_G des gesamten MT-Zustandes. Dies wiederum nimmt über die Orch OR Einfluss auf die Integrationszeit von Berechnungen in den MT (siehe Abschnitt 3.3.4) und somit auf das Bewusstsein. Hameroff et al. schlagen vor, dass inhalativen Anästhetika auf die Orch OR in Mikrotubuli wirken und somit zumindest mitverantwortlich sind für den Verlust des Bewusstseins während einer Allgemeinanästhesie. [88]

Es wurden weitere Anästhetika hinsichtlich deren Wirkung auf MT untersucht. In einer umfassenden Studie von Emerson et al. wurde der Effekt von 1-azidoanthrazen (1-AZA)[59] als Anästhetikum sowie dessen Wirkmechanismus über MT gezeigt. In dem Experiment wurden Kaulquappen mit 1-AZA inkubiert und dessen hypnotische Wirkung bestätigt. Es zeigte sich außerdem, dass unter Hinzugabe eines MT-stabilisierenden Stoffes Epothilon D oder Discodermolid, die hypnotische Wirkung von 1-AZA herabgesetzt wurde. Mittels Gelelektrophorese (IEF/SDS-PAGE[60]) konnte das photoaktive 1-AZA isoliert und über Massenspektrometrie das β-Tubulin als Bindungsstelle von 1-AZA identifiziert werden. Eine in-vitro Analyse weist überdies auf eine Inhibition der Tubulin-Polymerisation, welche für den Zusammenbau von MT essentiell ist, durch 1-AMA. [202] Die Wirkung von 1-AMA bzw. 1-AZA als Anästhetikum und deren Bindung und Beeinflussung von MT in Nervenzellen ist möglicherweise hinweisend auf einen Zusammenhang von MT und dem Bewusstsein im Sinne der Orch OR Theorie. [88]

5.2.2 Mikrotubuli in neurologischen und psychiatrischen Erkrankungen

MT können in neurodegenerativen und psychiatrischen Erkrankungen eine Rolle spielen. Häufig liegen Veränderungen im Aufbau der MT oder MAP vor, welche eine entscheidende Rolle im Pathomechanismus der genannten Erkrankungen spielen können.

Die Alzheimer-Erkrankung geht mit einer kognitiven Leistungsverschlechterung einher und ist die häufigste Ursache für Demenz [203]. In der Krankheitsentstehung spielen hyperphosphorylierte Tau-Proteine (h-MAPT) eine große Rolle. h-MAPT, zusammen

[59] ein optische Analogon von 1-aminoanthrazen (1-AMA) mit ähnlicher hypnotischer Wirkung [200].

[60] Die isoelektrische Fokussierung (IEF) kombiniert mit einer (sodium dodecyl sulfate)SDS-Polyacrylamidgelelektrophorese, auch 2D-Gelelektrophorese genannt, wird zur Trennung von komplexen Proteingemischen in einzelne Proteine eingesetzt. [201]

mit MAP2 und anderen MAP, aggregieren im Zytosol zu unlöslichen Neurfibrillären Bündel (NFT) im Sinne einer Tauopathie [204], wobei h-MAPT den Zusammenbau von MT stören und bestehende MT schädigen kann [205]. In pathologischen Studien an mit an Morbus Alzheimer erkrankten Gehirnen zeigen eine Fehlfunktion von MT in den Pyramidenzellen des gesamten Kortex, unabhängig von NFT [206].

Morbus Parkinson ist die häufigste neurodegenerative Bewegungsstörung und ist mit Veränderungen an MT assoziiert. Unter den häufigsten Symptomen dieser Erkrankung sind Bradykinesie/Akinesie, Ruhetremor, Rigor, posturale Instabilität und Demenz [207]. Die Symptome werden auf pathophysiologischer Ebene durch einen Untergang der Neuronen in der Substantia Nigra, welche ins Striatum projizieren, bedingt. Mehrere verschiedene Mutationen sind mit der Parkinson-Krankheit assoziiert, unter anderem in α-Synuclein (SNCA), in Parkin (PRKN) [208] und in 4-repeat Isoformen des Tau-Proteins (4R-MAPT) [209]. Die mutierten Proteine wirken auf MT durch Verschlechterung der Stabilität und Förderung der Depolymerisation [210].

Chorea Huntington ist eine autosomal-dominante Erkrankung, welche im Verlauf mit einer Hyperkinesie mit darauf folgender Hypokinesie sowie begleitende psychiatrische Symptomen und kognitiven Beeinträchtigungen einhergeht. Bei Erkrankten findet sich eine zu häufige Wiederholung von CAG-Sequenzen im Huntingtin-Gen. [211] Das mutierte Huntingtin bindet an Huntingtin-assoziiertes Protein 1 (HAP1), Huntingtin-interagierendes Protein (HIP1) und MT und löst dadurch eine Zelluntergang (besonders der GABAergen Neuronen) im Striatum aus [212]. Mutiertes Huntingtin interagiert mit dem Zytoskelett mit einhergehenden Abbau von α-Tubulin sowie MAP2, was zu einer Instabilität der MT führt [213].

MT spielen möglicherweise eine Rolle in der Krankheitsentstehung der Schizophrenie. Bestimmte genetische Veränderungen gehen mit einer erhöhten Anfälligkeit für Schizophrenie einher [214]. Das Disrupted-in-Schizophrenia-1 (DISC1)-Protein bindet an MT und prä- sowie postsynaptische Membranen und andere Zellbestandteile [215]. Mutationen in DISC1 werden auch mit der Major Depression und der bipolaren affektiven Störung assoziiert [216] [216]. Weitere mit Schizophrenie im Zusammenhang stehendes Protein, das Dystrobrevin-bindende Protein 1/Dysbindin (DTNBP1) bindet an Bestandteile des Zytoskeletts, besonders an MT [217].

Mit diesen Beispielen konnte gezeigt werden, dass MT eine entscheidende Rolle in gewissen neurodegenerativen und psychiatrischen Erkrankungen spielen. Hameroff et al. vermuten, dass im Pathomechanismus dieser Erkrankungen Veränderungen an den

MT über eine Beeinflussung der OR eine Rolle spielen könnten. [88]

6 Anwendungen von KI-Techniken in der Kardiologie

6.1 Überblick

In einer fortschreitend digitalisierten Welt ist die moderne Kardiologie mit immer mehr verfügbaren quantitativen Daten konfrontiert [218]. Elektronische Gesundheitsakten, genauere und umfassendere Untersuchungen sowie telemedizinische Anwendungen [219] machen enorme Datensätze über kardiologische PatientInnen direkt verfügbar. In großen Datenmengen sind wertvolle Informationen verborgen, welche das Wissen sowohl über individuelle PatientInnen erweitern als auch medizinisch relevante Aussagen über PatientInnen-Gruppen, Kohorten, Populationen etc. machen können. Vor allem in Hinblick auf den Trend zu einer schnelleren, effizienteren und personalisieren PatientInnenversorgung ist es mittlerweile unabdingbar, die verfügbaren Daten zum Wohle der genannten Ziele einzusetzen [220].

Die fortschreitende Entwicklung auf dem Gebiet des Machine Learnings und leistungsfähigere Computer eröffnen neue Möglichkeiten für die Analyse von PatientInnendaten. Mithilfe von Computermodellen können aus der Krankheitsgeschichte, Laborwerten, bildgebenden Verfahren, EKGs, klinischen Untersuchungen etc. medizinisch relevante Aussagen über den jetzigen sowie zukünftigen Zustand der PatientInnen abgeleitet werden.

Besonders im Vordergrund steht hier eine individuelle Charakterisierung der kardiologischen Erkrankung(en) der PatientInnen. Abseits von dichotomen Klassifizierungen [221] und weit gefassten Diagnosen, welche oft die tatsächlichen pathophysiologischen Vorgänge unzureichend beschreiben, soll durch Methoden des Machine Learnings eine präzisere und quantitative Aussage über den Gesundheitszustand getroffen werden. Somit lässt sich zum einen der Verlauf einer Erkrankung genauer und früher erkennen, und zum anderen eine maßgeschneiderte und personalisierte Therapie planen und durchführen. Dies wäre ein großer Schritt in Richtung *Präzisionskardiologie* bzw. Präzisionsmedizin im Allgemeinen. [222]

Ebenso bedeutend wie die PatientInnenversorgung ist die kardiologische Forschung. Bisher werden in erster Linie traditionelle statistische Methoden verwendet, um aus einer Menge an Daten medizinisch relevante Schlüsse ziehen zu können. Beispielsweise findet die logistische Regressionsanalyse breite Verwendung, um aus Variablen wie zum Beispiel Blutdruck eine Wahrscheinlichkeit für das Auftreten von einem gewissen Ereignis berechnen zu können. Ein großes Problem für solche statistischen Methoden

sind die Voraussetzungen, welche auf einen Datensatz zutreffen müssen, damit dieser verwendet werden kann. Bei der logistischen Regressionsanalyse dürfen die Daten unter anderem keine Kollinearität, also keine starke Korrelation, aufweisen - häufig sind die gewählten Parameter jedoch voneinander abhängig. ([223] S. 4 - 28) Methoden des Machine Learnings stellen tendenziell sehr wenige Voraussetzungen an die vorliegenden Daten und können häufig direkt und vielseitig in mehrdimensionalen Parameterräumen angewandt werden. Außerdem stehen für die Forschung oft sehr große Datenmengen mit vielen Parametern zur Verfügung. Algorithmen zur *Feature Extraction* können aus vielen Variablen jene herausfinden, welche für die Fragestellung von Relevanz ist und somit neue Erkenntnisse über zugrundeliegende Krankheitsmechanismen liefern, was mit traditionellen Methoden limitiert ist. [224]

Es ist vorauszusehen, dass maschinelles Lernen und andere Computermethoden im klinischen Alltag eine immer bedeutendere Rolle einnehmen werden. Daher ist es für KardiologInnen und andere kardiologische Fachberufe wichtig, sich mit der Entwicklung des maschinellen Lernens in der Kardiologie zu befassen. Ein fachkundiger Umgang mit Computermethoden im klinischen Alltag und in der Forschung sowie die Kenntnis über deren Anwendungsbereiche und Limitationen sollte ebenfalls in den Verantwortungsbereich der zukünftigen KardiologInnen und MedizinerInnen fallen. Denn deren Expertise bleibt stets integral für ein umfassendes und erfolgreiches PatientInnenmanagement. [225]

6.2 Computergestützte Rhythmusanalyse von Langzeit-EKGs

In der Analyse von Langzeit-EKGs greifen ÄrztInnen häufig auf die Hilfe von Rhythmusanalyse-Software zurück. Die Software-Pakete analysieren die meist über mehrere Stunden oder Tage hinweg aufgezeichnete EKG-Sequenz und suchen nach Zeichen von Arrhythmien oder anderen abnormen Veränderungen. Diese werden im EKG für die BenutzerInnen sichtbar gekennzeichnet, sowie Beginn- und Endzeit gespeichert. Mit diesem Verfahren lassen sich innerhalb kürzester Zeit Herzrhythmusstörungen über den gesamten (All-)Tag der PatientInnen quantifizieren und/oder verifizieren. [226]

Bisherige Software-Lösungen analysieren die rohen EKG-Daten in einer Reihe von Schritten, welche von spezialisierten Algorithmen durchgeführt werden. Bei jenen Schritten handelt es sich beispielsweise um Filtern von Rohdaten, Finden eines repräs-

sentativen Abschnittes des EKGs, regelbasiertes Erkennen der Wellenform (P-Welle, QRS-Komplex und T-Welle) und anderer Muster, Fourieranalyse (zum Finden der Herzfrequenz) etc. Diese Mechanismen werden per Hand angepasst und auf das Erkennen einer speziellen Rhythmusklasse hin optimiert. [226] Bislang haben diese Systeme eine im Vergleich zu kardiologischen Fachpersonal hohe Rate an Fehlbeurteilungen, vor allem bei Vorliegen von Arrhythmien, Störungen des Erregungsleitungssystems und beim Erkennen von Schrittmacher-Artefakten [227]. Im Gegensatz dazu verarbeiten *Deep Neural Networks (DNN)* eingegebene Rohdaten auf integrierte Art, ohne die Notwendigkeit von manuellen Anpassungen und Optimierungen einzelner Prozesse. Stattdessen optimiert sich das Programm praktisch eigenständig über zur Verfügung gestellte Trainingsdaten (überwachtes Lernen). Dadurch sind DNNs besonders gut geeignet für eine präzise Mustererkennung, wie die Klassifikation von EKG Rhythmen. [228]

Zum besseren Verständnis sei hier eine rezente Publikation besonders detailliert dargestellt.

Eine Forschergruppe aus Kalifornien [229] hat ein CNN entworfen, welches in der Lage ist, den Rhythmus einer EKG-Sequenz in eine von zwölf Klassen einzuordnen. Der Aufbau des CNN ist in Abbildung 13 erläutert. Über die Eingabe-Schicht (Input-Layer) wird eine 1-Kanal EKG-Sequenz mit 256 Samples, etwa einem 1,3 Sekunden Abschnitt entsprechend, eingelesen. Nach der Verarbeitung innerhalb des CNN, innerhalb seiner *Hidden-Layers*, wird eine diskrete Wahrscheinlichkeitsverteilung an der Ausgabe-Schicht (Output-Layer) ausgeworfen. Aus der Ausgabe lässt sich ablesen, mit welcher Wahrscheinlichkeit jeweils einer von zwölf Rhythmustypen (Tabelle 3 auf die EKG-Sequenz zutrifft. [230]

Ein CNN muss, wie jedes künstliche neuronale Netz, für seinen spezifischen Anwendungszweck trainiert werden. Der hier gewählte Ansatz ist das überwachte Lernen, bei welchem das CNN mit Paaren aus Eingaben und dazu passenden Ausgaben versorgt werden. Anhand dieser Trainingsdaten kann das CNN seine internen Parameter von selbst iterativ optimieren, damit es nach dem Training auf ähnliche Eingaben wiederum die (möglichst) richtige Ausgabe ausgeben kann. Je mehr Trainingsdaten zur Verfügung stehen, desto besser kann das CNN sich auf seine spezifischen Aufgabe „vorbereiten“.

Die Autoren haben einen Trainingsdatensatz mit über 91.000 EKG Sequenzen mit einer Länge von etwa 30 Sekunden aus ambulant durchgeführten Langzeit-EKGs von über 50.000 verschiedenen PatientInnen zusammengestellt. Dabei wurde auf eine ähnli-

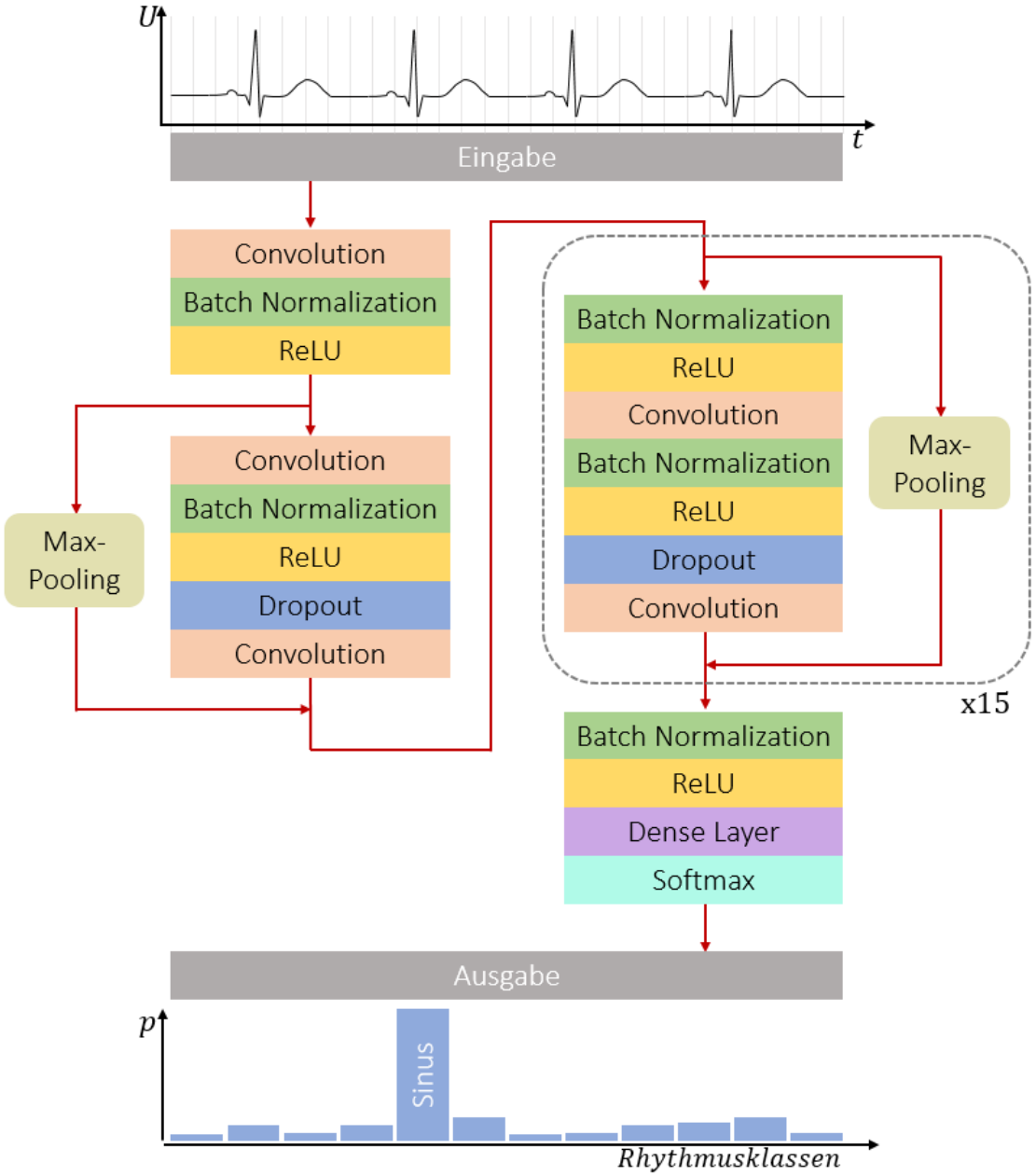

Abbildung 13: Schematische Darlegung des Aufbaues vom *Convoluted Neural Network*, welches für die Beurteilung von EKG-Sequenzen verwendet wurde. Das CNN nimmt 256 Samples eines EKG-Signals, welches mit 200 Hz aufgezeichnet wurden, als Eingangsdaten an. Durch verschiedenste aufeinander folgende Neuronen-Schichten werden die Daten von Anfang bis zum Ende des Netzes weitergereicht und manipuliert. Der grau eingezäunte Block wird 15 Mal nacheinander angeordnet. Das Netzwerk gibt zu jeder EKG-Sequenz eine Wahrscheinlichkeitsverteilung aus. Diese besagt, mit welcher Wahrscheinlichkeit eine Rhythmusklasse auf die eingegebene Sequenz zutrifft. Der höchste Wert kann als das *Urteil* des CNN betrachtet werden. Zur Funktion der einzelnen Neuronen-Schichten siehe [231] und ([232] Kapitel 3 und 9). Darstellung nach [229], Extended Data Fig. 1.

che Repräsentation aller zwölf Rhythmusklassen geachtet. Die EKG-Sequenzen wurden von erfahrenen KardiologietechnikerInnen klassifiziert. Durch die große Menge an Trainingsdaten wirken sich einzelne Fehlbeurteilungen wenig auf den Trainingserfolg des CNN aus.

Um die Güte des CNN zu überprüfen, wurde ein Testdatensatz von 328 unabhängigen und vergleichbaren PatientInnen mit jeweils einer im Median 30 Sekunden dauernden EKG-Sequenz erstellt. In den Testdaten sind die Rhythmusklassen ähnlich stark repräsentiert. Die Beurteilung erfolgte durch spezialisierte RhythmologInnen in 3-er Teams (ExpertInnengruppe) und wurde in weiterer Folge als Gold-Standard angenommen.

Sechs weitere unabhängige KardiologInnen haben dieselben Test-EKG-Sequenzen nochmals jeweils alleine beurteilt. Deren Urteile sowie jene des CNN wurde mit dem dem Gold-Standard verglichen. Die Klassifikations-Güte sowohl der KardiologInnen als auch des CNN wurde mittels des F-Maßes[61] quantifiziert (Tabelle 3). Es zeigte sich eine gleiche oder bessere Performance der DNN im Vergleich zu ausgebildeten KardiologInnen.

Eine nähere Analyse der Fehldiagnosen des CNN zeigte, dass das Netzwerk ähnliche Fehler macht, wie KardiologInnen es tun. Beispielweise macht eine zu geringe Sequenzlänge, eine schlechte Signalqualität es sowohl für die menschlichen BetrachterInnen als auch für das CNN sehr schwierig die richtigen Schlüsse aus der Einkanal-Ableitung zu ziehen. Im Falle der ventrikulären Tachykardie (bei welchem das CNN ein schlechtere

[61] Das sogenannte F-Maß wird als Gütemaß für binäre Klassifikatoren (Methoden, welche Objekte anhand von deren Merkmalen in zwei verschiedene Klassen einordnen) verwendet. Vor allem im „Information Retrieval“ findet es Anwendung in der Beurteilung von computergestützten Methoden um Beispiel zur Klassifizierung von (meist komplexen) Daten. Dadurch wird dieses Maß besonders für das Testen von Anwendungen des maschinellen Lernens geeignet. Das F-Maß ist das harmonische Mittel aus der Genauigkeit (precision) und der Trefferquote (sensitivity) einer Methode:

$$F = \frac{2 * \text{Trefferquote} * \text{Genauigkeit}}{\text{Trefferquote} + \text{Genauigkeit}}$$

Wobei die Trefferquote die bedingte Wahrscheinlichkeit

$$P(\text{positives Testergebnis}|\text{tatsächlich positiv}) = \frac{\text{richtig Positive}}{\text{richtig Positive} + \text{falsch Positive}}$$

und die Genauigkeit

$$P(\text{tatsächlich positiv}|\text{positives Testergebnis}) = \frac{\text{richtig Positive}}{\text{richtig Positive} + \text{falsch Negative}}$$

ist. ([233] Kap. 7) Hand und Christen [234] kritisieren das F-Maß als Mittel zum Vergleich der Güte von Klassifikatoren, speziell bei Verwendung einer unterschiedlichen Gewichtung von Genauigkeit und Trefferquote. Hand schlägt stattdessen die Verwendung des H-Maßes vor [235].

Tabelle 3: Vergleich der F-Maße für das CNN und die KardiologInnen-Gruppe zur Klassifizierung von einzelnen EKG-Sequenzen in zwölf Rhythmusklassen. Ergebnisse aus [229] Tabelle 1.

Rhythmus-Klasse	F-Maß CNN	gemitteltes F-Maß Kardiologen
Vorhofflimmern und -flattern	0,801	0,677
AV-Block IIb (Mobitz) und III	0,828	0,772
Bigeminus	0,847	0,842
Extrasystolen	0,541	0,482
Idioventrikulärer Ersatzrhythmus	0,761	0,632
Junktionaler Ersatzrhythmus	0,664	0,692
EKG-Rauschen	0,844	0,768
Sinus-Rhythmus	0,887	0,852
Supraventrikuläre Tachykardie	0,488	0,451
Trigeminus	0,907	0,842
Ventrikuläre Tachykardie	0,541	0,566
AV-Block IIa (Wenckebach)	0,702	0,591

F-Maß als die KardiologInnen hatte) wären einige Fehldiagnosen auf die Verwechslung mit dem idoventrikulären Rhythmus zurückführbar. Die Unterscheidung der beiden Rhythmen erfolgt anhand der Herzrate, welche in vielen EKG-Sequenzen aus dem Test-Datensatz sehr nahe an am Grenzbereich von 100 Schlägen pro Minute gewesen wäre. Es ist weiterhin denkbar, dass das CNN Schlüsse aus Merkmalen in der EKG-Ableitung zieht, welche laut nach der kardiologischen Lehrmeinung [236] in der Klinik gar nicht herangezogen werden.

Die Autoren schlagen vor, dass das von ihnen entwickelte CNN noch weiter klinisch validiert werden sollte. Eine Anwendung des CNN zur automatischen Rhythmusdiagnostik von Langzeit-EKGs im klinischen Alltag könnte eine Zeit- als auch Kostenersparnis sowie insgesamt eine Verbesserung der Diagnosequalität mit sich bringen. Außerdem könnten PatientInnen in der Notaufnahme durch treffsichere automatische Rhythmusanalyse besser priorisiert werden. Mögliche Einsatzgebiete in der Telemedizin, in automatischen Defibrillatoren und im Herz-Monitoring innerhalb des Krankenhauses sind denkbar. [229]

6.3 Funduskopische Aufnahmen der Retina und Vorhersage des kardiovaskulären Risikos mithilfe von Deep Learning

Ein anderes illustratives Beispiel stellt die Interpretation von funduskopischen Retina-Aufnahmen mittels Deep Learning dar. Es besteht eine starke Korrelation zwischen Veränderungen in den Netzhaut- und Aderhautgefäßen und dem kardiovaskulären Ri-

siko. Die Kaliber und Geometrie von Arteriolen und Venolen in der Retina[62] [238] und auch die Aderhautdicke und Gefäßrarefizierung [239] können als Surrogat-Parameter für das Risiko für Hypertonie, Diabetes mellitus, Erkrankungen der kleinen Nierengefäße sowie kardiovaskuläre Komplikationen bei bestehender koronarer Herzkrankheit, zerebrovaskuläre Erkrankungen und Diabetes mellitus herangezogen werden. [240]

Poplin et al. [241] haben ein DNN entworfen, welches aus Funduskopie-Aufnahmen Rückschlüsse auf kardiovaskuläre Risikofaktoren (siehe Tabelle 4) der PatientInnen ziehen kann. Die Berechnung der Risikofaktoren aus den Bildern erfolgte durch ein CNN mit der Architektur *Inception-v3*[63] [242]. Um herauszufinden, welche Bereiche des Augenhintergrundes für die Risikofaktoren relevant sind, wurde die Technik *soft attention* [243] verwendet. Die relevanten Bereiche wurden mittels Heatmaps[64] repräsentiert (Abbildung 14).

Das Training des DNN erfolgte mit funduskopischen Bildern aus den Datenbanken UK Biobank [244] (14101 PatientInnen) und EyePACS [245] (236234 PatientInnen). Zur Evaluation des DNN wurden unabhängige Bilder aus denselben Datenbanken (UK Biobank: 12026 PatientInnen und EyePACS: 999 PatientInnen) herangezogen. Für die binäre sowie kontinuierliche Risikofaktoren wurden jeweils ein eigenes DNN trainiert. Die Genauigkeit in der Bestimmung der Risikofaktoren aus einem Fundusbild pro PatientIn ist in Tabelle 4 abzulesen.

Abgesehen von den Risikofaktoren ist das DNN ebenfalls in der Lage das 5-Jahres-Risiko für eine *schwere kardiovaskuläre Komplikation (MACE)*[65] aus einem Fundus-Bild zu bestimmen. Die Area Under the Curve (AUC)[66] für die Güte dieser Vorhersage

[62]Retinale mikrovaskuläre Veränderungen sind ebenfalls geeignet als Prädiktor für zerebrale ischämische Erkrankungen wie lakunäre Infarkte, Hyperintensität der weißen Substanz (WMH - white matter hyperintensity) und zerebrale Insulte im Allgemeinen [237].

[63]Die Struktur des DNN Inception-v3 besteht direkt nach dem Eingang aus 6 Convolution-Layern mit Pooling. Darauf folgen drei Inception-Abschnitte, welche wiederum aus einem asymmetrischen Netz aus parallelen Convolution-Layern besteht. Vor dem Ausgang werden die Daten mittels Pooling, Linearisierung und Softmax zu einer lesbaren Ausgabe verarbeitet. [242]

[64]Heatmaps dienen zur graphischen Darstellung von Funktionen, welche zwei Dimensionen auf eine abbilden. Konkret in diesem Falle kann von der *Attention Heatmap* abgelesen werden, welche Bereiche im Bild des Augenhintergrundes relevanter sind und (zum Beispiel von einem Mustererkennungs-DNN) aufmerksamer betrachtet werden sollen. Es handelt sich hier um eine Art der *Feature Extraction*.

[65]Major Adverse Cardiovascular Event (MACE): Das MACE wird häufig als Endpunkt in kardiologischen Studien verwendet und bezieht sich meist auf a) Tod durch kardiale Ereignisse, b) nicht-tödlicher Myokardinfarkt oder c) nicht-tödlicher Schlaganfall. Je nach Kontext können weitere Ereignisse unter den Begriff MACE fallen, wie zum Beispiel Aufnahme wegen Herzinsuffizienz, Herzversagen, Koronar-Interventionen etc. [246]

[66]Die *Area Under the Curve* (AUC) bezieht sich auf die Fläche unter der *Receiver Operating Characteristic Curve* (ROC-Kurve). In einem Graph werden die Sensitivität (richtig Positive) gegen (1-Spezifität) (falsch Positive) einer binären Klassifizierungsmethode bei einem gewissen Parameter, wie zum Beispiel Alter, Blutdruck etc., aufgetragen wird. Durch Variation des Parameters entstehen meh-

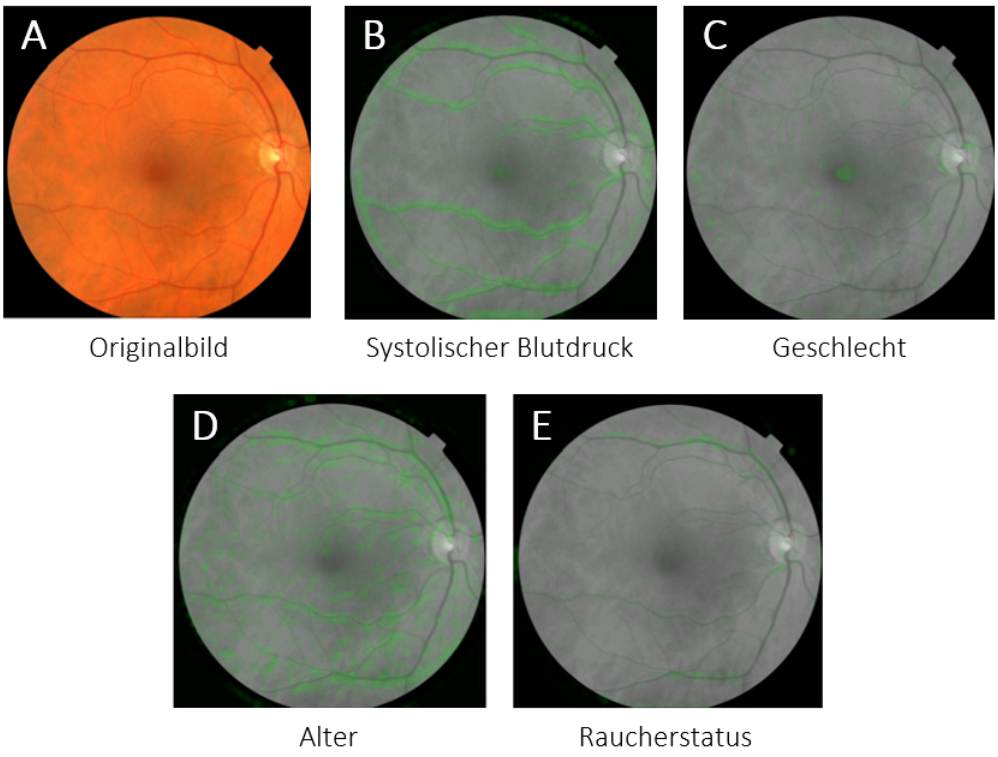

Abbildung 14: Beispiel eines funduskopischen Bildes aus der UK Biobank (A). Attention Heatmaps wurden für die jeweiligen Risikofaktoren in grün über das schwarz-weiße Fundus-Bild (B-E) gelegt. Auffällig ist der Fokus bei der Geschlechterbestimmung, welcher in der Fovea zu liegen scheint. Adaptiert nach [241], Abbildung 2.

beträgt AUC $0,73$ und ist somit ähnlich genau wie der *Systematic Coronary Risk Evaluation Score (SCORE)*[67] mit einem AUC $0,72$.

Tabelle 4: Risikofaktoren, welche durch das CNN aus Funduskopie-Bildern errechnet werden. Dazu sind ist ein Maß für die Genauigkeit des CNN in der Evaluierungs-Phase in der Tabelle ersichtlich. Für kontinuierliche Risikofaktoren wird der mittlere Fehler (MAE - mean average error) und für binäre Risikofaktoren die AUC angegeben.

Risikofaktor	Maß	Genauigkeit des CNN
Alter	kontinuierlich	$\pm 3,26$ Jahre
Geschlecht	binär	AUC $0,97$
Raucherstatus	binär	AUC $0,71$
HbA1c	kontinuierlich	$\pm 1,39$ %
Systolischer Blutdruck	kontinuierlich	$\pm 11,23$ mmHg
Diastolischer Blutdruck	kontinuierlich	$\pm 6,42$ mmHg
Body Mass Index (BMI)	kontinuierlich	$\pm 3,29$
MACE	binär	AUC $0,70$

Eine mögliche Verbesserung der Nutzbarkeit dieser Methode könnte die Nutzung der optische Kohärenztomografie (OCT) sein. Für das volle Erfassen des Gefäßgeflechts in der Netzhaut eignet sich die OCT besonders gut, da diese eine höhere räumliche Auflösung hat und einen Blick in tiefere Schichten der Retina bietet. Mit dieser bildgebenden Methode können auch tiefere Gefäßstrukturen bis einschließlich jenen des Choroids in drei Dimensionen hoch aufgelöst werden, um diese zur Beurteilung der Gefäßkaliber heranzuziehen. [249]

rere Datenpunkte im genannten Graphen, durch welche eine ROC-Kurve gelegt werden kann. Die AUC kann als Qualitätsmaß für eine Klassifizierungsmethode verwendet werden. [247]

[67] Der Systematic Coronary Risk Evaluation Score (SCORE) gibt das Risiko aufgrund von systolischem Blutdruck, Geschlecht, Alter, Raucherstatus und totalem Cholesterin bzw. des Verhältnisses von totalem Cholesterin zu HDL (high density lipoprotein) Cholesterin an. Der Risiko-Wert kann unkompliziert von einer Tabelle abgelesen werden. [248]

7 Diskussion

7.1 Interpretation von physikalischen Modellen von Bewusstsein im philosophischen Kontext

Im folgenden soll eine Betrachtung der dargelegten physikalischen Modellen, nämlich der Orch OR Theorie des Bewusstseins und der IIT, im Lichte der Philosophie des Geistes sowie der Philosophie der Physik erfolgen.

Die DP OR ist das Herzstück der Orch OR Theorie des Bewusstseins (siehe Abschnitt 3.3.4). Sie stellt einen nicht-berechenbaren physikalischen Prozess dar, welcher innerhalb der Theorie eine entscheidende Rolle für das Bewusstsein spielt. Die DP OR ist ein bislang in der modernen Physik nicht erklärbarer physikalischer Vorgang. Die darin wirkenden Gravitationseffekte spielen sich auf der Planck-Skala ab, wodurch es notwendig ist die Effekte mit einer Theorie der Quantengravitation zu behandelt. Solch eine Theorie befindet sich im Spannungsfeld zwischen dem Standardmodell der Teilchenphysik[68] und der Gravitation und konnte trotz umfassender Bemühungen bisher nicht vollständig postuliert werden [251]. Dies wäre nach Penrose ein notwendiger Schritt in Richtung eines Verständnisses der physikalischen Grundlagen von Bewusstsein [101].

Eine Verbindung des anthropischen Prinzips[69] mit der DP OR wird in [257] diskutiert. Die DP OR wählt, wie in Abschnitt 3.3.3 beschrieben, durch den Prozess einer Zustandsreduktion wiederholt in einer ständigen zeitlichen Abfolge eine physika-

[68] Das Standardmodell der Teilchenphysik beschreibt in einem einheitlichen mathematischen Konstrukt drei der vier (bekannten) grundlegenden Wechselwirkungen in der Natur. Dies sind die starke, schwache und elektromagnetische Wechselwirkung. Die vierte Wechselwirkung, die Gravitation, passt derzeit noch nicht in das Schema. Eine Vereinigung aller vier Kräfte in einer einheitlichen Beschreibung würde einen bedeutenden Schritt in Richtung einer physikalischen Theorie von Allem (theory of everything (ToE)) darstellen. [250]

[69] Das anthropische Prinzip besagt, dass wir als intelligente und bewusste Lebensformen unser sichtbares Universum nur deshalb beobachten können, weil in unserem Universum exakt jene Begebenheiten vorherrschen, welche ein intelligentes und bewusstes Leben ermöglichen. Alle Naturkonstanten haben genau jene Werte, welche ein Leben und die Entstehung von Bewusstsein begünstigen. Eine geringe Abweichung der Naturkonstanten könnten ein völlig anderes Universum, möglicherweise ohne Leben, Intelligenz oder Bewusstsein, zur Folge haben. Das Prinzip kann auf viele Weisen interpretiert werden. Einen Überblick vermittelt [252] und ([253] S. 11-58). Eine wichtige Unterscheidung wird von Brandon Carter durch das schwache und starke anthropische Prinzip getroffen. Im schwachen anthropischen Prinzip wird die Aussage getroffen, dass wir privilegiert sind uns in einem von möglichen parallelen Universen, welches Bewusstsein zulässt, zu befinden. Nach dem starken anthropischen Prinzip gibt es genau ein Universum in welchem Bewusstsein zugelassen ist. [254] Überdies postuliert John Archibald Wheeler das *Participatory anthropic principle*. Es besagt, dass das Universum beobachtet werden muss, damit es existieren kann. Durch die Beobachtung würde das Universum entstehen. [255] Louis Crane diskutiert eine mögliche Verbindung der Quantengravitation mit dem anthropischen Prinzip [256].

lische Realität aus mehreren Alternativen aus. Jede Auswahl entspricht einer „proto-bewussten Erfahrung“ - eine Abfolge dieser Erfahrungen führt (zumindest im Gehirn) laut der Orch OR Theorie zu einem zusammenhängenden Bewusstsein. Wir leben also in einem Universum, in welchem Bewusstsein existiert, weil die „richtige“ Auswahl möglicher physikalischer Realitäten durch die DP OR getroffen wurde. Dies entspricht der Aussage des schwachen anthropischen Prinzips (siehe Fußnote 69). Drüber hinaus sei bemerkt, dass der Prozess der Auswahl im Sinne der Orch OR Theorie einen Bezug zur Prozessontologie von Alfred North Whitehead hat. Näheres dazu ist in Fußnote 26 beschrieben.

Auf welcher Grundlage trifft die DP OR eine Auswahl im erwähnten Sinne? Penrose schreibt dies der Platon'schen Welt der Ideen („Platonic mathematical world“), welche zu Beginn in Abschnitt 1.1 behandelt wurde, zu. Aus dieser Welt sollen nichtberechenbare Informationen, welche für die Erzeugung eines zusammenhängenden Bewusstseins aus einzelnen „proto-bewussten Erfahrungen“ kausal sind, kommen. ([145] S. 416 - 418) Dies ist eine Idee, welche dem Dualismus (Abschnitt 1.5.1) zugeordnet werden kann.

IIT fasst das Bewusstsein als eine Eigenschaft innerhalb eines physischen Systems auf. Allein die Zusammensetzung und Konfiguration des Systems bestimmt, welcher Bewusstseinsgrad vorliegt (siehe Abschnitt 3.2). Dabei wird von den Autoren angenommen, dass jedes System ein Bewusstsein hat, auch wenn der Bewusstseinsgrad beliebig klein sein kann. Diese Auffassung entspricht dem Panpsychismus (Abschnitt 1.5.1), welcher allen physikalischen Objekten ein Bewusstsein zuordnet.

Eine Erweiterung stellt Max Tegmark in seiner Verallgemeinerung der IIT auf beliebige quantenphysikalische Systeme auf. Hierin bezeichnet Tegmark das Bewusstsein als einen eigenen Aggregatzustand. Die Überführung der IIT in die Sprache der Quantenmechanik hat interessante Konsequenzen. Unter anderem führt dies zu einer Erklärung unseres Konzeptes der vergehenden Zeit. [258]

7.2 Was tragen die Neurowissenschaften bei?

Die von David Chalmers als „hard problem“ formulierte Frage, wieso und wie Empfindungen wie Schmerz oder Freude für uns einen subjektiven und erlebenden Charakter haben, steht den sogenannten „easy problems“ gegenüber [36]. Jene einfacheren Probleme, wie das Lenken der Aufmerksamkeit, die Kontrolle über das Verhalten, das

Gedächtnis ([259] Kapitel 9, 10, 11) etc. sind nach Chalmers leichter lösbar und wurden bereits gründlich neurowissenschaftlich erforscht. Es jedoch nach wie vor rätselhaft, weshalb gewisse Zustände der neuronalen Verbände im Gehirn, also der Umstand, dass manche Neuronen feuern und manche nicht, mentale Zustände wie das innere Empfinden verursachen können. [260]

Neuronale Korrelate von Bewusstsein

Als zentrales Mitglied der Kognitionswissenschaften beschäftigen sich Neurowissenschaften eingehend mit dem Bewusstsein. Ein Teil der Forschung in diesem Feld widmet sich der Identifikation von *neuronalen Korrelaten von Bewusstsein* (NCC). Dies geschieht in Experimenten, in welchen eine subjektive Beschreibung der qualitativen Empfindungen der Versuchspersonen verwendet wird, um die Korrelation von bewusster Wahrnehmung und messbaren Veränderungen in der Aktivität von Hirnarealen herzustellen.

Mithilfe von bildgebenden Studien, wie der funktionellen Magnetresonanztomografie (fMRT) oder Positronen-Emissions-Tomographie (PET) konnten in gewissen funktionellen Einheiten des Gehirns ein indirekter Aktivitätshinweis[70] mit bewusster Wahrnehmung assoziiert werden. [186] Mit bildgebenden Methoden konnte in mehreren Studien ein minimales NCC, die posteriore kortikale „Hot Zone", als Kandidat für ein „Bewusstseinszentrum" im Gehirn, gefunden werden. Diese befindet sich in den sensorischen Anteilen der Parietal-, Temporal- und Okzipitallappen. Läsionen in diesen Bereich führen zu spezifischen Verlusten von Wahrnehmungsqualitäten (Quale). [260]

Mittels Elektroenzephalographie (EEG)-Studien kann einem Areal im Gehirn ein Rhythmus der neuronalen Aktivität zugeordnet werden. Somit lässt sich in Momenten der bewussten Wahrnehmung bestimmen, welche Neuronengruppen in welchem Rhythmus aktiv sind. Durch diese Methode ist es möglich eine Synchronizität neuronaler Rhythmen von verschiedenen Hirnarealen festzustellen. Ein NCC könnte die γ-Kohärenz sein, bei welcher Wellen mit einer Frequenz von etwa 40 Hz in verschiedenen Hirnarealen mit Bewusstsein assoziiert sind[71]. [266] Die EEG-Kohärenz ist in Abbildung 15 illustriert.

[70] Ein indirekter Indikator für neuronale Aktivität in einem Areal ist die Standardized Uptake Value (SUV) beim ^{18}F-2-Fluor-2-Deoxy-D-Glucose (F-18-FDG)-PET [261] beziehungsweise das Blood Oxygen Level-Dependent (BOLD) Signal beim fMRT [262].

[71] Die Neuronale Bindungs-Hypothese besagt, dass durch synchrone Oszillationen in weiten Neuronengruppen im Gehirn diese ein temporäres dynamisches Netzwerk bilden [263]. Dieser Mechanismus ist mit bewusster Wahrnehmung assoziiert [264] und spielt möglicherweise eine Rolle für das Lernen

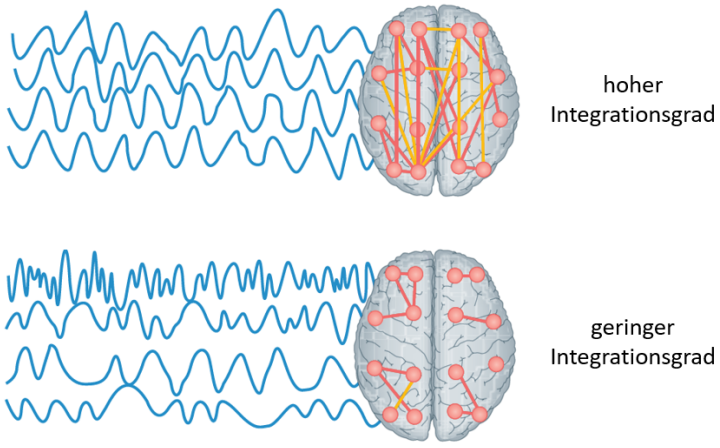

Abbildung 15: Die neuronale Integration, wenn weite Teile des Gehirns koordiniert aktiv sind, ist assoziiert mit einer Kohärenz der neuronalen Rhythmen im EEG. Darstellung nach [260] Abbildung 4a.

Überdies steht die Methode der transkraniellen Magnetstimulation (TMS) zur Verfügung. In dieser Methode wird von außen eine Abfolge kurzer magnetischer Impulse am Kopf der Versuchsperson angelegt, wodurch in den anvisierten kortikalen Arealen im Gehirn kleine elektrische Wirbelströme entstehen. Diese Ströme können je nach Frequenz der Impulsfolge Neuronengruppen stimulieren oder inhibieren. Es konnte gezeigt werden, dass eine Inhibition des primären visuellen Kortex (V1) zu einer temporären totalen Skotomen in einem Bereich des Gesichtsfeld führt [267] (Abschnitt 1.3.2). Durch die TMS lassen sich somit gezielt Areale mit der damit verbundenen Funktion korrelieren. [268]

Konnektom des menschlichen Gehirns

Das menschliche Gehirn hat etwa 86 Milliarden Nervenzellen - 16 Milliarden davon nur im zerebralen Kortex [136]. Die Anzahl an Synapsen werden auf 10^{15} geschätzt ([269] Teil IV). Die Gesamtheit der Verbindungen zwischen den Nervenzellen wird „Konnektom“ genannt [270]. Das Konnektom des Fadenwurmes *Caenorhabditis elegans* mit seinen 302 Neuronen und über 7.000 Synapsen wurde bereits 1986 mittels Elektronenmikroskopie vollständig beschrieben. [271] Es ist jedoch eine enorme technische Herausforderung ein menschliches Gehirn zu fixieren, zu schneiden und jeden Schnitt mit dem Elektronenmikroskop zu untersuchen. Die gewonnen Bilder können im Computer zu einem dreidimensionalen mikroskopischen Abbild des Gehirns zusammengesetzt

und das Gedächtnis [265].

werden. Innerhalb dieses sehr großen Datensatzes könnte eine KI die Synapsen identifizieren und zu einem digitalen Abbild des Konnektoms zusammenfügen. In weiterer Folge könnte wiederum unter Zuhilfenahme von KI-Techniken jene Strukturen in den neuronalen Netzen identifiziert werden, welche ein neuronales Substrat von Bewusstsein oder Gedächtnis sind und somit einen entscheidenden Beitrag zum Verständnis von Bewusstsein im Gehirn sowie neurologischen und psychiatrischen Erkrankungen liefern. [272]

Ein derzeit groß angelegtes Projekt zur Kartographie des menschlichen Konnektoms, dem „Human Connectome Project“, wird von den National Institutes of Health (NIH) mit rund 40 Millionen US Dollar finanziert [273] [274]. Ein weiteres Projekt mit demselben Ziel, genannt „BRAIN Initiative“ , wurde von der Regierung der USA initiiert und wird voraussichtlich eine Finanzierung von etwa 3 Milliarden US-Dollar erhalten (siehe Fußnote 51).

7.3 Werden wir die Singularität erreichen?

In jenem Zuge, in welchen sich die KI-Technik weiter entwickelt und immer komplexere Probleme lösbar werden, stellt sich die Frage, ob Maschinen auch ein Bewusstsein haben könnten. Die Realität einer starken KI mit einem Bewusstsein wird als „Singularität“ bezeichnet. Wenn das Bewusstsein als von der Physik losgelöste und kategorisch unterschiedliche Entität aufgefasst wird, scheint es eher unwahrscheinlich, dass eine KI jemals ein Bewusstsein entwickeln wird. Dazu müsste vermutlich ein Bewusstsein von außen zu einer KI hinzugegeben werden. In diesem Falle gibt es für eine konkrete Umsetzung keinerlei Anhaltspunkte. Andererseits könnte das Bewusstsein eine Eigenschaft sein, welche aus dem Aufbau des zugrundeliegenden physikalischen System hervorgeht. Die folgende Diskussion soll von diesem Standpunkt aus geführt werden.

Giorgio Buttazzo nennt ein notwendiges Kriterium für das Vorliegen von menschenähnlichen Bewusstsein in einer KI. Nach dem Autor muss ein künstliches neuronales Netz, welches bisher die erfolgversprechendste Technik in der KI darstellt, zumindest die Komplexität des menschlichen Gehirns erreichen. Diese Schlussfolgerung zieht er aus der Beobachtung, dass Tiere weniger bewusst sind und gleichzeitig ein einfacher aufgebautes Gehirn (besonders im Hinblick auf die Neuronenzahl im zerebralen Kortex) haben. Der Autor bezeichnet das menschliche Gehirn als die Komplexitäts-Schwelle für ein Bewusstsein. [275] Dies ist in Abbildung 16 illustriert.

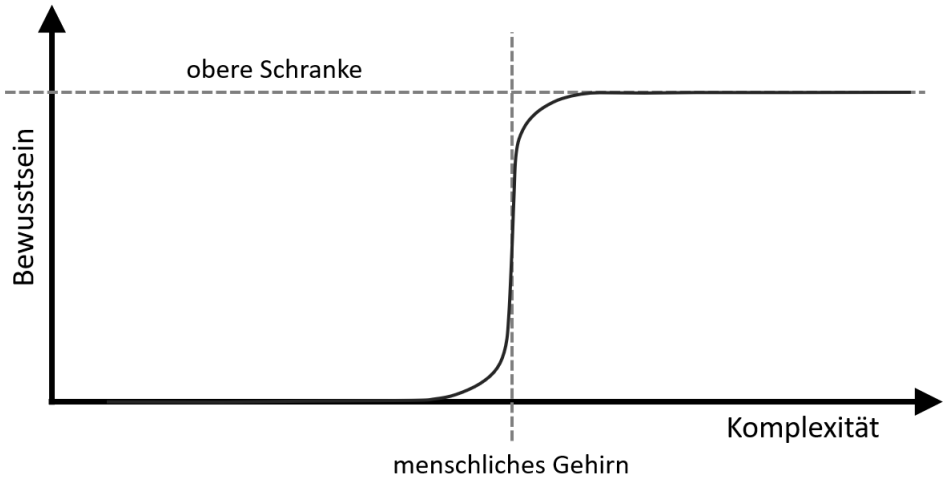

Abbildung 16: Das Bewusstsein eines physischen Systems ist abhängig von dessen Komplexität. In dieser Abbildung nach ([275] Abb. 1) ist die Komplexität des menschlichen Gehirns gerade ausreichend für einen hohen Grad an Bewusstsein. Eine niedrigere Komplexität führt zu einem bedeutend geringeren Bewusstsein, und komplexere Systeme erreichen eine Sättigung im Bewusstseinsgrad. Der Zusammenhang entspricht einer verschmierten Stufenfunktion.

Ebenfalls stellt Buttazzo eine Schätzung für den Zeitpunkt an welchen eine KI die Komplexität des menschlichen Gehirns erreichen kann. Die Schätzung beruht auf der Frage, wann ein gewöhnlicher Haushaltscomputer eine solch große Speicherkapazität[72] haben würde, um das menschliche Gehirn mit seinen etwa 10^{15} Synapsen in Echtzeit zu simulieren. Die dafür nötige Speichermenge wird auf 4 ExaByte geschätzt. Über eine Interpolation der Entwicklung von Speichergrößen in den letzten Jahrzehnten, sollte es voraussichtlich im Jahre 2029 möglich sein, ein Gehirn auf dem Heimcomputer zu simulieren. [275]

Toby Walsh hingegen zweifelt daran, dass Maschinen jemals die Intelligenz von Menschen erreichen oder ein Bewusstsein entwickeln werden. Eine KI kann derzeit nur nach den Regeln handeln, welche ihnen durch die Programmierung auferlegt wurden. Es gibt keinen Grund anzunehmen, dass sich dies in der näheren Zukunft ändern wird, denn eine Maschine hat keinen eigenen freien Willen, keine Qualia und keine Intentionalität. [276] Letzterem pflichtet Deborah Johnson in folgendem Zitat bei:

> *„Artifact behavior, including computer behavior, is created, and used, by human beings as a result of their intentionality. Computer systems and other artifacts have intentionality, the intentionality put into them by the*

[72] Hier ist der *Random Access Memory* (RAM) gemeint.

intentional acts of their designers.“ [277]

Ebenfalls in diesem Sinne interpretierbar ist die Orch OR Theorie des Bewusstseins. Roger Penrose argumentiert, dass das Bewusstsein auf einem nicht-berechenbaren, jedoch realen und objektiven, physikalischen Prozess beruhe (siehe Abschnitt 3.3.2). Alle KIs sind auf einem Computer lauffähig und daher berechenbar. Daraus kann gefolgert werden, dass auch zukünftige KIs, welche auf berechenbaren Algorithmen beruhen, kein, zumindest dem Menschen ähnliches, Bewusstsein haben können. Es wäre in diesem Sinne ein anderes physikalisches Substrat, daher eine andere „Art“ von Computer, für eine bewusste KI nötig.

7.4 Ethische und rechtliche Überlegungen zur künstlicher Intelligenz und synthetischem Bewusstsein in der Medizin und Gesellschaft

Die Entwicklung und Erforschung der künstlichen Intelligenz ist vermutlich eine der folgenschwersten Unternehmungen überhaupt. Es ist abzusehen, dass die KI in den meisten, wenn nicht sogar allen, Lebensbereichen in irgendeiner Form eine Rolle spielen wird. Wie bereits in vorherigen Kapiteln erwähnt, scheint das Einsatzspektrum der KI nahezu unbegrenzt. Jedoch ist das Thema KI mehr als ein Konglomerat an spezifischen ingenieurtechnischen Anwendungen. Die KI könnte in Zukunft Autos lenken, medizinische Entscheidungen treffen und OPs durchführen, militärisch genutzt werden, die Weltwirtschaft steuern, einen Teil der menschlichen Arbeitskräfte ersetzen und vieles mehr. Andererseits wird dieses Forschungsfeld auch ein Wegbereiter zu einem tieferen Verständnis unseres eigenen Gehirns und dessen Entstehung sein. Aufgrund der schnellen Entwicklung hin zu autonom handelnden, „intelligenten“ Maschinen, desto bedeutender werden ethische und moralische Überlegungen zur KI und ihrer Rolle in unserer Gesellschaft. Im folgenden sollen verschiedene ethische, moralische, rechtliche und gesellschaftliche Aspekte der KI besonders im Hinblick auf die Zukunft diskutiert werden.

7.4.1 KI und die Gesellschaft

Weltweit zeigt sich ein Trend in Richtung einer teilweisen oder vollständigen Automatisierung des Fahrens. Ein automatisches Einparken und Bremsassistenten sind bereits

in vielen Neuwägen Standard. Ab 2024 ist in der EU für alle Neuwägen eine automatische Tempobeschränkung, eine Alkohol-Wegfahrsperre und weitere Kontrollmechanismen verpflichtend. Diese Maßnahmen sollen die Anzahl der Verkehrsunfälle senken und Fahrzeuglenker an das zukünftige autonome Fahren zu gewöhnen. [278] In Deutschland wurde im Jahre 2017 eine Ethik-Leitlinie für das autonome Fahren vom Bundesministerium für Verkehr und digitale Infrastruktur herausgegeben. Im Dokument werden 20 ethische Regeln genannt - neben grundsätzlichen Feststellungen zur menschlichen Autonomie und Eigenverantwortung, ist vor allem Regel 9 zu unausweichlichen Unfallsituationen interessant. Sie besagt, dass in einer Situation, in welcher ein Unfall mit Personenschaden unvermeidlich ist, keine Abwägung der gefährdeten Personen und darauf basiertes Handeln gestattet ist. Lediglich darf eine Handlung gewählt werden, welche die potentiellen Schäden an allen Personen im gleichen Maße reduziert. [279]

In den USA werden in Gerichten KI-Techniken verwendet, um die Rückfallwahrscheinlichkeit von Verurteilten zu berechnen. Dabei lernt der benutzte Algorithmus aus vergangenen Gerichtsbeschlüssen und Akten von Wiederholungstätern. Der Grund für die Verwendung dieser Methode ist, dass in den USA die Gefängnisse stark überlastet sind und mithilfe von KI sich Ressourcen in Haftanstalten und Rehabilitationseinrichtungen besser einteilen lassen. Da jedoch zum Training der KI vergangene Daten herangezogen werden, besteht die Gefahr, dass Fehlentscheidungen und mögliche Diskriminierung anhand von Ethnizität, sozialer Status oder Bildung in bereits passierten Entscheidungen von der KI gelernt und wiederholt werden. [280] Ist es in Ordnung, wenn Gerichtsprozesse von künstlicher Intelligenz beeinflusst oder gar entschieden werden?

In manchen Arbeitsbereichen wird ein besonderes Vertrauen benötigt. Zum Beispiel im medizinischen Bereich, in Kindergärten und Schulen, in der Gesetzgebung, Regierung, Verwaltung und Rechtssprechung etc. kommt es auf ein besonnenes und menschliches Handeln an. Hier sollte der Einsatz von künstlicher Intelligenz besonders vorsichtig erfolgen und auch kritisch hinterfragt werden. Wie im letzten Beispiel gezeigt, könnte ein voreiliger Einsatz von KI an sensiblen Stellen, wie der Rechtssprechung, schwerwiegende Folgen haben. Hier ist es besonders bedeutend, dass stets der Mensch die Verantwortung an den kritischen Stellen, wie etwa die ärztliche Diagnosestellung, übernimmt und dies nicht auf eine KI übertragbar ist. ([281] S. 202 - 227)

Der letzte Absatz führt direkt zur Frage, ob eine KI für ihr Handeln verantwortlich gemacht werden kann.

Rechtlich gesehen gilt in Österreich derzeit die Haftung des Betreibers für ein unsachgemäßes Einsetzen oder des Herstellers bei einer fehlerhaften Herstellung einer Maschine ([282] S. 1-14). Im Europäischen Parlament wurde diskutiert, ob einem Roboter der „Status einer natürlichen Person mit Rechten und Pflichten" gewährt werden könnte [283]. Zu einem späteren Zeitpunkt wird sogar eine neue Kategorie von Rechtssubjekten, der „elektronischen Person", vom Europäischen Parlament in Erwägung gezogen. Es wird vorgeschlagen, dass besonders „ausgeklügelte" autonome Roboter für etwaige Schäden, welche aufgrund eigenständiger Entscheidungen entstanden sind, verantwortlich gemacht werden könnten. [284] Welche Pflichten könnten einer Maschine, einem Roboter oder einer KI auferlegt werden? Gelten dann dieselben Grundrechte wie für Menschen?

Gesetze sollen Handlungen entsprechend der gesellschaftlichen Normen sicherstellen. In einfachen Worten ausgedrückt, stellt der Strafvollzug eine Motivation dar, die Gesetze einzuhalten, um ein möglichst gutes gesellschaftliches Zusammenleben zu ermöglichen. Dies funktioniert nur, wenn der Mensch die Folgen seines Handelns antizipieren und daraus Konsequenzen ziehen kann. Für die Auffassung der Strafbarkeit nach §§ 4 und 5 des österreichischen Strafgesetzbuches (StBG) ist ein Vorsatz oder eine Verletzung der Sorgfaltspflicht Grundvoraussetzung. Ob von all dies sinnvoll auf eine KI zutreffen kann, ist bislang eine offene Frage.

7.4.2 Gefahren von super-intelligenten Maschinen

Die Erfolge auf dem Gebiet der *schwachen*, also auf spezifische Aufgaben ausgerichtet KI, lässt die Vorstellung einer möglichen *starken* KI, die allgemeine Probleme eigenständig lösen kann, immer realer werden. Eine starke KI wurde lange Zeit für unmöglich gehalten - nun ändert sich diese Ansicht in Fachkreisen. Es ist daher von Bedeutung, bereits jetzt, also lange vor der sogenannten „Singulariät" (siehe Abschnitt 7.3), potentielle Gefahren zu erkennen und konkrete Maßnahmen zu deren Vermeidung zu entwickeln.

Die meisten heutigen Kontroversen bezüglich der möglichen Gefahren einer zukünftigen super-intelligenten KI lassen sich in mehrere Lager einteilen, siehe Abbildung 17. Jene, welche übermäßig optimistisch sind und in der KI die Lösung vieler Probleme unserer Gesellschaft sehen. Ihnen gegenüber steht die Meinung, dass die rasante Entwicklung auf dem Gebiet der künstlichen Intelligenz mehr Probleme bringt, als sie löst. Der Großteil der Diskussionen findet jedoch zwischen diesen Extremen statt. ([285] S.

30-37)

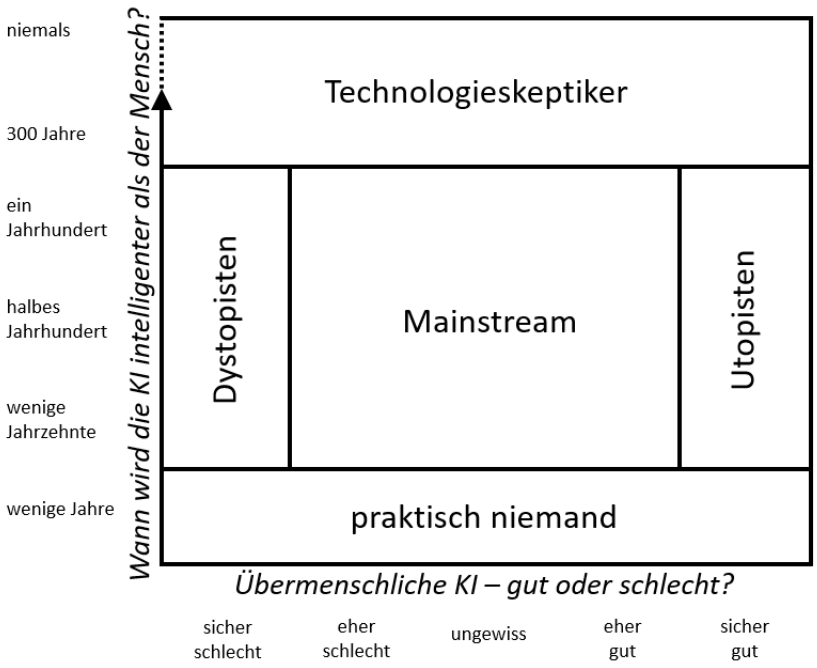

Abbildung 17: Meinungslandschaft zum Thema starke KI. Die häufigsten Diskussionspunkte bezüglich der Gefahren der künstlichen Intelligenz drehen sich um darum, wann es passiert und ob es eine gut oder schlechte Sache ist, wenn es passiert. Hauptsächlich werden Positionen eingenommen, welche sich in diesem Graphen etwa mittig anordnen lassen. Darstellung nach [285], Abbildung 1.2.

Welche Gefahren könnten von einer KI, welche eine über-menschliche Intelligenz hat, ausgehen? Die meisten Experten in den Gebiet gehen davon aus, dass eine KI nicht „gut“ oder „böse“ sein wird [286]. Dies sind zutiefst menschliche Kategorien und sind vor allem in der Religion und unserem eigenen Moralverständnis verwurzelt. Es gibt derzeit vermutlich keinen Grund davon auszugehen, dass sich eine zukünftige KI eigenständig und willentlich gegen den Menschen wenden wird. Eher könnte das Problem an der Kontrolle über die Ziele einer KI, welche nicht notwendigerweise mit den unseren übereinstimmen, liegen. Nach Angaben des *Future of Life Instituts*[73] in Boston, USA, sind folgende zwei Szenarien, in welcher eine zukünftige KI ein Gefahr darstellen könnte, am wahrscheinlichsten: [286]

- Speziell der kriegerische Nutzung von KI in Kombination mit Waffensystemen

[73]Das Future of Life Institute (FLI) ist eine im Jahre 2014 gegründete Organisation, welche sich mit existentiellen Risiken von modernen Technologien beschäftigt. Neben dem Themen Klimawandel, Biotechnologie und Nuklearwaffen liegt ein besonderer Augenmerk auf der künstlichen Intelligenz. Geleitet wird das Institut unter anderen von Max Tegmark, Elon Musk und Christof Koch. [287]

birgt ein hohes Gefahrenpotential. Hier ist nicht so sehr ein eigenständiges, autonomes Handeln der KI gegen das Wohlergehen der Menschen im Vordergrund, sondern eine militärische Nutzung gegen andere Menschen. In falschen Händen könnten autonome Waffen aufgrund ihrer Effizienz großen Schaden anrichten. Außerdem befürchten KI-ForscherInnen ein Aufrüsten oder gar Wettrüsten von KI-gesteuerten Waffen.

- Anders als im vorherigen Szenario könnte die KI mit wohlwollenden Absichten der BetreiberInnen eingesetzt werden. Eine mögliche Gefahr von sehr intelligenten und autonomen Maschinen ist jene, dass die KI zwar auf einen konstruktiven Zweck ausgerichtet ist, jene allerdings Maßnahmen zur Erfüllung diese Ziels ergreift, welche gefährlich oder destruktiv sein könnten. Die Ziele und „Wertvorstellungen" einer KI sollten daher mit den unseren möglichst genau abgeglichen werden. Eine „KI-Moral" ist notwendig.

Isaac Asimov hat in seinen zahlreichen utopischen Zukunfts-Romanen, in welchen Roboter meist eine große Rolle spielen, drei Robotergesetze aufgestellt. Diese sollen das Zusammenleben von intelligenten Robotern mit dem Menschen gewährleisten ([288] S. 40).

1. „Ein Roboter darf keinen Menschen verletzen oder durch Untätigkeit zu Schaden kommen lassen"

2. „Ein Roboter muss den Befehl eines Menschen gehorchen, es sei denn, solche Befehle stehen im Widerspruch zum ersten Gesetz"

3. „Ein Roboter muss seine eigene Existenz schützen, solange dieser Schutz nicht dem ersten oder zweiten Gesetz widerspricht"

Möglicherweise sollte eine Implementation von ähnlichen „Robotergesetzen" in einer starken KI angestrebt werden. Ein fixes Regelwerk, an welches sich jede KI haltet, müsste eine Autorität über die exekutiven Funktionen der KI haben. Um ein mögliches Umgehen der Gesetze zu verhindern, wäre es eventuell zielführend, neben der eigentlichen KI ein sekundäres System umzusetzen. Eine zweite KI, dessen einzige Aufgabe die Überwachung der primären KI ist, könnte als übergeordnete Instanz die Einhaltung des fixen Regelwerks garantieren. Ähnliche Überlegungen, jedoch im Bezug auf eine Realisation von synthetischem Bewusstsein finden sich in Miyazaki und Takeno [289]. Patrick Lin kritisiert die Vorstellung von „Robotergesetzen". Ein ethisches

Handeln könnte nicht auf ein einfaches Regelwerk reduziert werden. Selbst wenn dies möglich wäre, so würde es Schwierigkeiten beim Lösen von Konflikten zweier Regeln geben. [290] Joanna Brysons schlägt ein normatives Regelwerk für KIs vor, welches jedoch nicht auf „logisch und algorithmisch nicht umsetzbaren“ Gesetzen, wie jene von Asimov, beruht. Dabei soll auf die Rechte von Menschen als auch Maschinen beachtet werden. [291] Im Jahre 2010 hat der *Engineering and Physical Sciences Research Council* fünf Prinzipien bezüglich autonomer Roboter aufgestellt. Hierin wird auch die Frage nach der Haftung behandelt. [292]

Eine weitere Möglichkeit zur Einhaltung von Gesetzen könnten eine „künstliche Moral“ sein. Dies wäre eine Annäherung an den Menschen, welcher im täglichen Leben nach seinen subjektiven Moralvorstellungen handelt. Eine eigene Moral in der KI könnte flexibler auf neue und nicht in einem normativen Regelwerk festgelegte Situationen reagieren. Jedoch stellt sich hier die Frage, ob solch eine künstliche Moral in der KI, nämlich eine Art der Moral mit welcher wir als Gesellschaft zufrieden sind, überhaupt umsetzbar ist. Möglicherweise ist eine Moral, welche unserer eigenen entspricht, an die menschliche Biologie gebunden und nur unter strikter Nachahmung unserer Natur in der Praxis konstruierbar [291]. Allen et al. schlagen zur Überprüfung der Moral einer Maschine den „moralischen Turing Test“, analog zum normalen Turing Test (siehe Kapitel 4.5) vor. Dieser Vorschlag impliziert, dass eine künstliche Moral jener des Menschen ebenbürtig sein sollte. [293]

7.4.3 Synthetisches Bewusstsein

Wenn eine KI nun Bewusstsein bekommen würde, welche ethischen Konsequenzen würde dies nach sich ziehen? Anschließend an die Diskussion in Abschnitt 7.4.1 stellt sich die Frage, ob dann eine KI einen besonderen gesetzlichen Status bekommen sollte. Wie bereits in Fußnote 12 erwähnt, hat im Jahre 2015 das oberste Gericht in New York, USA, zwei Schimpansen den Status einer juristischen Person[74] erteilt. Dieses Urteil wurde damit begründet, dass Menschenaffen, wie die beiden Schimpansen, einen hohen Grad an Bewusstsein haben und dadurch dem Menschen sehr ähnlich wären. [16]

[74]Nach dem Österreichischen Gesetz ist eine juristische Person jeder Träger von Rechten und Pflichten, der keine natürliche Person, also kein Mensch, ist. Jeder Mensch ist von Geburt an eine natürliche Person und dieser Status endet erst mit dem Tode. (§15 - §24 Allgemeines bürgerliches Gesetzbuch (ABGB)) Die Ernennung zur juristische Person mitsamt der Definition seiner Rechte und Pflichten geschieht in einem Rechtsakt (§§ 26 und 27 ABGB).

Sollten daher bewusste Maschinen im Kant'schen Sinne von einem Status als Objekt, zum Zwecke anderer, in den Status eines (Rechts-)Subjekts, zum Zwecke „vor sich selbst", erhoben werden ([294] S. 94)? Joanna Brysons spricht sich dagegen aus, sowohl aus rechtlicher als auch ethischer Perspektive, Maschinen den Status einer Person zu verleihen. Eine KI und mögliche Implementationen in Robotern seien rein ein Werkzeug des Menschen, und haben daher kein natürliches Recht auf Autonomie. Vielmehr sei es eine Fehlwahrnehmung, dass in Maschinen eine Ähnlichkeit mit Menschen gesehen wird, was vor allem bei humanoiden Robotern[75] der Fall ist. Diese Fehlwahrnehmung jedoch führe häufig zur Überzeugung, Robotern gegenüber eine moralische Verpflichtung zu haben. Dies sei jedoch unbegründet. Bryson schlägt eine andere Richtung vor, nämlich zu versuchen solche Maschinen und KIs zu konstruieren, gegenüber welchen wir keine moralische Verpflichtung haben und somit ein mögliches moralisches Dilemma zu vermeiden. [296]

Ein weiterer Diskussionspunkt ist die mögliche Rolle des Bewusstseins in der Moral. Möglicherweise ist ein synthetisches Bewusstsein ein Wegbereiter zu einer künstlichen Moral, welche über ein rein regelbasiertes Handeln hinausgeht (siehe Abschnitt 7.4.2). Diese Eventualität wird in einem Aufsatz über das Bewusstsein als das bestimmende Element von Handlungen [297] diskutiert. Hierin wird festgestellt, dass die bewusst gelenkte Aufmerksamkeit auf die moralischen Verpflichtung, ein moralisches Handeln[76] weit wahrscheinlicher macht. Konkret nimmt damit das Bewusstsein die Rolle eines Richtungsgebers ein, welcher die Aufmerksamkeit auf die moralische Verantwortung in einer gegebene Situation lenkt. Speziell in einer KI könnte solch ein Richtungsgeber Sinn machen, da ansonsten die KI möglicherweise nicht „weiß" wie es möglicherweise fix festgelegte ethische/moralische Normen im Kontext zur aktuellen Situation interpretieren sollte. [297]

[75] Ein optisch besonders ausgeklügelter humanoider Roboter ist beispielsweise *Sophia* von Hanson Robotics Limited [295].

[76] In der englischen philosophischen Fachliteratur wird zwischen „moral agency" und „moral patiency" unterschieden und als zwei Aspekte der Moral gesehen. *Moral agency* bezeichnet das bewusste moralische Handeln (gegenüber anderen). Hingegen ist *moral patiency* das Betroffensein von moralischem Handeln. [298] In diesem Sinne könnte etwa ein Neugeborenes zwar nicht als moralische AgentIn/moralischer Agent (moral agent), jedoch als moralische PatientIn/moralischer Patient (moral patient) betrachtet werden.

7.5 Ausblick

Die Erforschung des Bewusstseins ist ein aufstrebendes Betätigungsfeld, welches voraussichtlich einen immer wichtigeren Stellenwert in Naturwissenschaft und Medizin annehmen wird. Aufgrund der Komplexität und Facettenreichtums des Phänomens „Bewusstsein“ ist ein interdisziplinärer Ansatz zielführend. Physikalisch begründete Modelle von Bewusstsein stellen eine quantitative Behandlung dar, aus welchen überprüfbare und anwendbare Aussagen abgeleitet werden können. Die inhaltlich und methodisch verwandten Themengebiete „künstliches Bewusstsein“ und „künstliche Intelligenz“ tragen zu einem besseren Verständnis von Bewusstsein sowie zu einer umfassenden Behandlung des Themas in philosophischer, rechtlicher und ethischer Hinsicht bei.

In der Medizin zeichnet sich ein Trend in Richtung einer Technisierung von Diagnose, Therapie(-wahl) und Forschung ab, auch unter Einbezug von Methoden der künstlichen Intelligenz. Ein tieferes Verständnis von Bewusstsein könnte neue Einblicke in die Entstehung von neurologischen und psychiatrischen Erkrankungen bieten und ein Wegbereiter zu neuen Therapieformen sein. Eine eingehende Beschäftigung mit den vorgestellten Themen innerhalb der medizinischen Gemeinschaft ist daher anzuraten.

8 Literaturverzeichnis

[1] Paulson HL, Galetta SL, Grossman M, Alavi A. Hemiachromatopsia of Unilateral Occipitotemporal Infarcts. American Journal of Ophthalmology. 1994;118(4):518–523. Available from: http://www.sciencedirect.com/science/article/pii/S0002939414758064.

[2] Plato. Platon über den Tod des Sokrates: Vier Schriften Platons zu Person und Tod des Sokrates. Oberhausen: Verlag Karl Maria Laufen; 2018.

[3] Karfik F. Die Beseelung des Kosmos: Untersuchungen zur Kosmologie, Seelenlehre und Theologie in Platons Phaidon und Timaios. vol. 199. Walter de Gruyter; 2004.

[4] Plomin R, DeFries JC, Craig IW, McGuffin P. Behavioral genetics. In: Plomin R, editor. Behavioral genetics in the postgenomic era. Washington, D.C: American Psychological Association; 2003. p. 3–15.

[5] Müller J, editor. Platon: Phaidon. vol. 44 of Klassiker Auslegen. Berlin/Boston: De Gruyter; 2011.

[6] Morgan ML. Sense-Perception and Recollection in the "Phaedo". Phronesis. 1984;29(3):237–251. Available from: http://www.jstor.org/stable/4182204.

[7] Brinker W. Seele. In: C Schäfer, editor. Platon-Lexikon. Begriffswörterbuch zu Platon und der platonischen Tradition. Darmstadt: Wiss. Buchges; 2007. p. 253–258. Available from: http://unibibliografie.ub.uni-mainz.de/opus/frontdoor.php?source_opus=12909.

[8] Cottingham J. Descartes: Meditations on First Philosophy: With Selections From the Objections and Replies; 1996.

[9] Brüntrup G. Philosophie des Geistes: Eine Einführung in das Leib-Seele-Problem. vol. Band 22 of Grundkurs Philosophie. 1st ed. Stuttgart: Verlag W. Kohlhammer; 2018.

[10] Kemmerling A, editor. René Descartes: Meditationen über die Erste Philosophie. vol. 37 of Klassiker Auslegen. Berlin/Boston: De Gruyter; 2009.

[11] Nietzsche F, Colli G, Müller-Lauter W. Nietzsche - Werke: Nachgelassene Fragmente Herbst 1887 - März 1888: Abteilung VIII - Band 2. Berlin: De Gruyter; 1970.

[12] Scott D. Occasionalism and Occasional Causation in Descartes' Philosophy. Journal of the History of Philosophy. 2000;38(4):503–528.

[13] Lokhorst GJ. Descartes and the Pineal Gland. In: Edward N Zalta, editor. The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University; 2018. .

[14] Birnbacher D. Great Ape Project. In: Ach JS, Borchers D, editors. Handbuch Tierethik. Stuttgart: J.B. Metzler; 2018. p. 312–315. Available from: https://doi.org/10.1007/978-3-476-05402-9_53.

[15] Singer P. Animal Liberation: Die Befreiung der Tiere. Harald Fischer Verlag; 2016.

[16] Liz Green. Sensationelles Urteil: Gericht erkennt nichtmenschliches Tier als Person an: Das erste Mal in der Weltrechtsgeschichte erkennt ein Richter zwei Schimpansen als juristische Personen an und gewährt ihnen Habeas-Corpus.;. Available from: https://tinyurl.com/y527ueyq.

[17] Descartes R. Abhandlung über die Methode, richtig zu denken und Wahrheit in den Wissenschaften zu suchen. 1st ed. Berlin: Contumax and Hofenberg; 2016.

[18] Rips LJ. Circular reasoning. Cognitive Science. 2002;26(6):767–795.

[19] Block N. On a confusion about a function of consciousness. Behavioral and Brain Sciences. 1995;18(2):227–247. Available from: http://cogprints.org/231/1/199712004.html.

[20] Searle JR. Who is computing with the brain? Behavioral and Brain Sciences. 1990;13(4):632–642.

[21] van Gulick R. Consciousness. In: Edward N Zalta, editor. The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University; 2018. .

[22] Armstrong DM. What is consciousness. The nature of mind. 1981;p. 55–67.

[23] Gertler B. Self-knowledge. Routledge; 2010.

[24] Block N. Qualia. Oxford Companion to the Mind. 2004;.

[25] Rosenthal DM. State consciousness and transitive consciousness. Consciousness and cognition. 1993;2(4):355–363.

[26] Rosenthal DM. Thinking That One Thinks. In: Burri A, editor. Sprache und Denken/Language and Thought. Grundlagen der Kommunikation und Kognition / Foundations of Communication and Cognition. Berlin/Boston: De Gruyter; 1997. .

[27] Zoller A. Phänomenales Bewusstsein: wie kann sich der naturalistische Repräsentationalismus im Lichte der konkurrierenden Strategien als vielversprechende physikalistische Qualiatheorie behaupten? [Masterarbeit]; 2010.

[28] Overgaard M. Visual experience and blindsight: a methodological review. Experimental Brain Research. 2011;209(4):473–479. Available from: https://doi.org/10.1007/s00221-011-2578-2.

[29] Brentano FC. Psychologie vom empirischen Standpunkt. vol. 1. Duncker & Humblot; 1874.

[30] Searle JR. The Intentionality of Intention and Action*. Cognitive Science. 1980;4(1):47–70.

[31] Siewert C. Consciousness and Intentionality. In: Edward N Zalta, editor. The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University; 2017. .

[32] Crane T. The objects of thought. Oup Oxford; 2013.

[33] Searle JR, Willis S, et al. Intentionality: An essay in the philosophy of mind. Cambridge University Press; 1983.

[34] Churchland PM. Eliminative Materialism and Propositional Attitudes. Journal of Philosophy. 1981;78(2):67–90.

[35] Levine J. Materialism and Qualia: The explanatory gap. Pacific Philosophical Quarterly. 1983;64(4):354–361.

[36] Chalmers D. The hard problem of consciousness. The Blackwell companion to consciousness. 2007;p. 225–235.

[37] Searle JR. Mind: A brief introduction. 4th ed. Fundamentals of philosophy series. New York: Oxford Univ. Press; 2004.

[38] Bishop RC, Atmanspacher H. The Causal Closure of Physics and Free Will. The Oxford Handbook of Free Will. 2011;.

[39] Brüntrup G. Philosophie des Geistes: Eine Einführung in das Leib-Seele-Problem. Kohlhammer Verlag; 2018.

[40] Stubenberg L. Neutral Monism. In: Edward N Zalta, editor. The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University; 2018. .

[41] Heil J. Philosophy of mind: A contemporary introduction. Fourth edition ed. Routledge contemporary introductions to philosophy. New York: Routledge, Taylor & Francis Group; 2020.

[42] Byrne A. Inverted Qualia. In: Edward N Zalta, editor. The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University; 2018. .

[43] Tye M. Qualia. In: Edward N Zalta, editor. The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University; 2018. .

[44] Ned Block. Wittgenstein and Qualia. Philosophical Perspectives. 2007;21:73–115. Available from: `http://www.jstor.org/stable/25177198`.

[45] Shoemaker S. Functionalism and qualia. Philosophical Studies. 1975;27(5):291–315.

[46] Chalmers DJ. The conscious mind: In search of a fundamental theory. 1st ed. Oxford paperbacks. New York: Oxford Univ. Press; 1997.

[47] Jackson F. What Mary Didn't Know. The Journal of Philosophy. 1986;83(5):291.

[48] Kim J. Concepts of supervenience. In: Supervenience. Routledge; 2017. p. 37–62.

[49] Stoljar D. Physicalism. In: Edward N Zalta, editor. The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University; 2017. .

[50] Fodor JA. Special sciences (or: The disunity of science as a working hypothesis). Synthese. 1974;28(2):97–115. Available from: https://doi.org/10.1007/BF00485230.

[51] Bickle J. Multiple Realizability. In: Edward N Zalta, editor. The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University; 2019. .

[52] Hodgson D. Rationality + consciousness = free will. Oxford University Press; 2012.

[53] O'Connor T, Franklin C. Free Will. In: Edward N Zalta, editor. The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University; 2019. .

[54] Seiler S. Strafrecht Allgemeiner Teil I: Grundlagen und Lehre von der Straftat. 3rd ed. Wien: Facultas; 2016.

[55] Fischer JM, Kane R, Pereboom D, Vargas M. Four views on free will. John Wiley & Sons; 2009.

[56] Atmanspacher H, Bishop R. Between Chance and Choice: Interdisciplinry Perspectives on Determinism. Andrews UK Limited; 2014.

[57] Vargas M. Revisionist accounts of free will: Origins, varieties, and challenges. The Oxford Handbook of Free Will (2d. 2011;.

[58] Beck F, Eccles JC. Quantum Aspects of Brain Activity and the Role of Consciousness. In: Eccles JC, editor. How the SELF Controls Its BRAIN. Berlin, Heidelberg: Springer Berlin Heidelberg; 1994. p. 145–165. Available from: https://doi.org/10.1007/978-3-642-49224-2_9.

[59] Beck F. Synaptic quantum tunnelling in brain activity. NeuroQuantology. 2008;6(2).

[60] Beck F. Quantum brain dynamics and consciousness. Advances in Consciousness Research. 2001;83:83–116.

[61] Faye J. Copenhagen Interpretation of Quantum Mechanics. In: Edward N Zalta, editor. The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University; 2019. .

[62] Giacosa F. On Unitary Evolution and Collapse in Quantum Mechanics. Quanta. 2014;3(1):156.

[63] Wigner EP. Remarks on the Mind-Body Question. In: Mehra J, editor. Philosophical Reflections and Syntheses. The Collected Works of Eugene Paul Wigner. Berlin, Heidelberg: Springer Berlin Heidelberg; 1995. p. 247–260. Available from: https://doi.org/10.1007/978-3-642-78374-6_20.

[64] von Neumann J. Mathematische Grundlagen der Quantenmechanik. Zweite auflage ed. Berlin, Heidelberg: Springer Berlin Heidelberg; 1996.

[65] Heisenberg W. Physik und Philosophie. 8th ed. Klassiker. Stuttgart: Hirzel, S; 2011.

[66] Hernes T. Alfred North Whitehead (1861–1947). In: The Oxford handbook of process philosophy and organization studies. Oxford University Press; 2014. p. 255–271.

[67] Stapp HP. Attention, Intention, and Will in Quantum Physics. Journal of Consciousness Studies. 1999;(6):143–164. Available from: https://arxiv.org/pdf/quant-ph/9905054.pdf.

[68] Stapp HP. A quantum-mechanical theory of the mind-brain connection. Kelly EF et al Beyond Physicalism, Lanham: Rowman and Littlefield. 2015;p. 157–193.

[69] Atmanspacher H. Quantum Approaches to Consciousness. In: Edward N Zalta, editor. The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University; 2019. .

[70] Ricciardi LM, Umezawa H. Brain and physics of many-body problems. Kybernetik. 1967;4(2):44–48. Available from: https://doi.org/10.1007/BF00292170.

[71] Brading K, Castellani E, Teh N. Symmetry and Symmetry Breaking. In: Edward N Zalta, editor. The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University; 2017. .

[72] VITIELLO G. Dissipative quantum brain dynamics. No Matter, Never Mind. 2002;p. 43–61.

[73] Peskin ME, Schroeder DV. An introduction to quantum field theory: Student economy edition. Boulder, Colorado: Westview Press; 2016.

[74] Freeman WJ, VITIELLO G. Nonlinear brain dynamics as macroscopic manifestation of underlying many-body field dynamics. Physics of life reviews. 2006;3(2):93–118.

[75] Stuart, C I J M , Takahashi Y, Umezawa H. Mixed-system brain dynamics: Neural memory as a macroscopic ordered state. Foundations of Physics. 1979;9(3):301–327. Available from: https://doi.org/10.1007/BF00715185.

[76] VITIELLO G. Dissipation and memory capacity in the quantum brain model. International Journal of Modern Physics B. 1995;09(08):973–989.

[77] ALFINITO E, VIGLIONE RG, VITIELLO G. The decoherence criterion. Modern Physics Letters B. 2001;15(04n05):127–135.

[78] ALFINITO E, VITIELLO G. Double universe and the arrow of time. Journal of Physics: Conference Series. 2007;67(1):012010. Available from: https://iopscience.iop.org/article/10.1088/1742-6596/67/1/012010/pdf.

[79] Tononi G. Integrated information theory. Scholarpedia. 2015;10(1):4164.

[80] Balduzzi D, Tononi G. Integrated information in discrete dynamical systems: motivation and theoretical framework. PLoS computational biology. 2008;4(6):e1000091.

[81] Nagel T. What Is It Like to Be a Bat? The Philosophical Review. 1974;83(4):435.

[82] Oizumi M, Albantakis L, Tononi G. From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0. PLoS computational biology. 2014;10(5):e1003588.

[83] Tononi G. The Integrated Information Theory of Consciousness. In: Wang Xh, Dong Q, Ying LL, Chi SS, Lan YH, Huang YP, et al., editors. Enhancement of selective separation on molecularly imprinted monolith by molecular crowding agent. vol. 110. Chichester, UK: John Wiley & Sons, Ltd; 2017. p. 243–256.

[84] Albantakis L, Marshall W, Hoel E, Tononi G. What Caused What? A Quantitative Account of Actual Causation Using Dynamical Causal Networks. Entropy. 2019;21(5):459.

[85] Meixner U. Das Elend des Physikalismus in der Philosophie des Geistes. Post-Physikalismus. 2014;.

[86] Ferrarelli F, Massimini M, Sarasso S, Casali A, Riedner BA, Angelini G, et al. Breakdown in cortical effective connectivity during midazolam-induced loss of consciousness. Proceedings of the National Academy of Sciences of the United States of America. 2010;107(6):2681–2686.

[87] Massimini M, Ferrarelli F, Murphy M, Huber R, Riedner B, Casarotto S, et al. Cortical reactivity and effective connectivity during REM sleep in humans. Cognitive neuroscience. 2010;1(3):176–183.

[88] Hameroff S, Penrose R. Consciousness in the universe: a review of the 'Orch OR' theory. Physics of life reviews. 2014;11(1):39–78.

[89] Tegmark M. Importance of quantum decoherence in brain processes. Physical review E, Statistical physics, plasmas, fluids, and related interdisciplinary topics. 2000;61(4 Pt B):4194–4206.

[90] Hagan S, Hameroff SR, Tuszyński JA. Quantum computation in brain microtubules: Decoherence and biological feasibility. Physical Review E. 2002;65(6):061901. Available from: http://link.aps.org/pdf/10.1103/PhysRevE.65.061901.

[91] Scott A. Stairway to the mind: the controversial new science of consciousness. Springer Science & Business Media; 1999.

[92] Petzold C. The annotated Turing: a guided tour through Alan Turing's historic paper on computability and the Turing machine. Wiley Publishing; 2008.

[93] Copeland BJ, Posy CJ, Shagrir O, editors. Computability: Turing, Gödel, Church, and beyond. Cambridge, Massachusetts: The Mit Press; 2013. Available from: http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=601413.

[94] Copeland BJ. The Church-Turing Thesis. In: Edward N Zalta, editor. The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University; 2019. .

[95] Penrose R, Gardner M. The emperor's new mind: Concerning computers, minds and the laws of physics. Revised impression ed. Oxford landmark science. Oxford: Oxford University Press; 2016.

[96] Griffiths DJ, Schroeter DF. Introduction to quantum mechanics. Cambridge University Press; 2018.

[97] Schrdinger E. Die gegenwrtige Situation in der Quantenmechanik. Die Naturwissenschaften. 1935;23(48):807–812.

[98] Wallace D. Decoherence and its role in the modern measurement problem. Philosophical transactions Series A, Mathematical, physical, and engineering sciences. 2012;370(1975):4576–4593.

[99] Penrose R. On the Gravitization of Quantum Mechanics 1: Quantum State Reduction. Foundations of Physics. 2014;44(5):557–575. Available from: https://doi.org/10.1007/s10701-013-9770-0.

[100] Pikovski I, Brukner Č, Aspelmeyer M. Ein quantenoptischer Blick auf die Planck-Skala? Physik in unserer Zeit. 2012;43(4):163–164.

[101] Penrose R. Wavefunction collapse as a real gravitational effect. In: Fokas AS, editor. Mathematical physics 2000. London and Singapore: Imperial College Press; 2000. p. 266–282.

[102] Marshall W, Simon C, Penrose R, Bouwmeester D. Towards quantum superpositions of a mirror. Physical review letters. 2003;91(13):130401.

[103] Campbell NA, Urry LA, Cain ML, Wasserman SA, Minorsky PV, Reece JB. Biology: A global approach. Eleventh edition, global edition ed. New York, NY: Pearson; 2018.

[104] Marshall WF. Centriole evolution. Current Opinion in Cell Biology. 2009;21(1):14–19. Available from: http://www.sciencedirect.com/science/article/pii/S0955067409000179.

[105] Wolf KW, Böhm KJ. Organisation von Mikrotubuli in der Zelle. Biologie in unserer Zeit. 1997;27(2):87–95.

[106] Nielsen MA, Chuang IL. Quantum computation and quantum information. 10th ed. Cambridge: Cambridge Univ. Press; 2010. Available from: https://ebookcentral.proquest.com/lib/subhh/detail.action?docID=647366.

[107] Hagar A, Cuffaro M. Quantum Computing. In: Edward N Zalta, editor. The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University; 2019. .

[108] Homeister M. Quantum computing verstehen: Grundlagen - Anwendungen - Perspektiven. 5th ed. Lehrbuch. Wiesbaden, Germany: Springer Vieweg; 2018.

[109] Rendell P. Turing Machine Universality of the Game of Life. vol. 18 of Emergence, Complexity and Computation. Aufl. 2015 ed. Cham: Springer International Publishing; 2015.

[110] Drewes G, Ebneth A, Mandelkow EM. MAPs, MARKs and microtubule dynamics. Trends in Biochemical Sciences. 1998;23(8):307–311.

[111] Schlosshauer MA. Decoherence: and the quantum-to-classical transition. Springer Science & Business Media; 2007.

[112] Dauxois T, Peyrard M. Physics of solitons. Cambridge University Press; 2006.

[113] OpenAI. About OpenAI; 2019. Available from: https://openai.com/about/.

[114] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI Blog. 2019;1(8).

[115] OG: Team Information; 2019. Available from: https://liquipedia.net/dota2/OG#The_International_2019:_Defense_of_the_Aegis.

[116] OpenAI. OpenAI Five; 2018.

[117] Website der Europäischen Kommission, Generaldirektion Kommunikation, editor. Künstliche Intelligenz: Mitgliedstaaten und Kommission arbeiten gemeinsam an Förderung künstlicher Intelligenz "Made in Europe". Brüssel, Belgien; 2018. Available from: https://ec.europa.eu/commission/news/artificial-intelligence-2018-dec-07_de.

[118] Davenport TH. China is catching up to the US on artificial intelligence research; 2019. Available from: https://tinyurl.com/y62ax9k6.

[119] Hsu FH. Behind Deep Blue: Building the computer that defeated the world chess champion. Princeton University Press; 2004.

[120] Sam Byford. Google vs. Go: can AI beat the ultimate board game? Has AlphaGo solved one of the oldest problems in computer science?; 2016. Available from: https://www.theverge.com/2016/3/8/11178462/google-deepmind-go-challenge-ai-vs-lee-sedol.

[121] Bodlaender HL, Duniho F. Shogi: Japanese Chess; 2012. Available from: https://www.chessvariants.com/shogi.html.

[122] Silver D, Hubert T, Schrittwieser J, Antonoglou I, Lai M, Guez A, et al.. Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm; 2017. Available from: https://arxiv.org/pdf/1712.01815.

[123] Hutson M. AI protein-folding algorithms solve structures faster than ever. Nature. 2019;8:292.

[124] The AlphaStar team. AlphaStar: Mastering the Real-Time Strategy Game StarCraft II; 2019. Available from: https://deepmind.com/blog/article/alphastar-mastering-real-time-strategy-game-starcraft-ii.

[125] Graves A, Wayne G, Reynolds M, Harley T, Danihelka I, Grabska-Barwińska A, et al. Hybrid computing using a neural network with dynamic external memory. Nature. 2016;538(7626):471–476.

[126] Russell SJ, Norvig P. Artificial intelligence: A modern approach. 3rd ed. Prentice-Hall series in artificial intelligence. Boston: Pearson; 2010.

[127] Bringsjord S, Govindarajulu NS. Artificial Intelligence. In: Edward N Zalta, editor. The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University; 2018. .

[128] Kriegeskorte N, Douglas PK. Cognitive computational neuroscience. Nature Neuroscience. 2018;21(9):1148–1160. Available from: `https://www.nature.com/articles/s41593-018-0210-5.pdf`.

[129] Wilson R. Four Colors Suffice: How the Map Problem Was Solved-Revised Color Edition. vol. 30. Princeton University Press; 2013.

[130] Beierle C, Kern-Isberner G. Methoden wissensbasierter Systeme: Grundlagen, Algorithmen, Anwendungen. Springer-Verlag; 2014.

[131] Levesque HJ. Knowledge Representation and Reasoning. Annual Review of Computer Science. 1986;1(1):255–287.

[132] Haslum P, Geffner H. Heuristic planning with time and resources. In: Sixth European Conference on Planning; 2014. .

[133] Perlovsky LI. Conundrum of combinatorial complexity. IEEE Transactions on Pattern Analysis and Machine Intelligence. 1998;20(6):666–670.

[134] McCarthy J. From here to human-level AI. Artificial Intelligence. 2007;171(18):1174–1182. Available from: `http://www.sciencedirect.com/science/article/pii/S0004370207001476`.

[135] Tononi G. On Consciousness. Banff, Canada; 22.8.2016. Available from: `https://www.youtube.com/watch?v=zvJyMmw2Thw`.

[136] Azevedo FAC, Carvalho LRB, Grinberg LT, Farfel JM, Ferretti REL, Leite REP, et al. Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. The Journal of comparative neurology. 2009;513(5):532–541.

[137] Herculano-Houzel S. The Evolution of Human Capabilities and Abilities. Cerebrum: the Dana Forum on Brain Science. 2018;2018.

[138] Oak Ridge National Laboratory. ORNL Launches Summit Supercomputer: New 200-Petaflops System Debuts as America's Top Supercomputer for Science;. Available from: `https://www.ornl.gov/news/ornl-launches-summit-supercomputer`.

[139] Davies M, Srinivasa N, Lin TH, Chinya G, Cao Y, Choday SH, et al. Loihi: A Neuromorphic Manycore Processor with On-Chip Learning. IEEE Micro. 2018;38(1):82–99.

[140] Tononi G, Boly M, Massimini M, Koch C. Integrated information theory: from consciousness to its physical substrate. Nature Reviews Neuroscience. 2016;17(7):450–461.

[141] TURING AM. Computing Machinery and Intelligence. Mind. 1950;LIX(236):433–460.

[142] Moor JH, editor. The Turing Test: The Elusive Standard of Artificial Intelligence. Dordrecht: Springer Netherlands; 2003.

[143] Smith P. An introduction to Gödel's theorems. Second edition ed. Cambridge introductions to philosophy. Cambridge: Cambridge University Press; 2013.

[144] Lucas JR. Minds, Machines and Gödel. Philosophy. 1961;36(137):112–127.

[145] Penrose R. Computerdenken: Die Debatte um künstliche Intelligenz, Bewußtsein und die Gesetze der Physik. Heidelberg and Berlin: Spektrum, Akad. Verl.; 2002.

[146] Abramson D. Turing's Responses to Two Objections. Minds and Machines. 2008;18(2):147–167. Available from: https://doi.org/10.1007/s11023-008-9094-6.

[147] LaForte G, Hayes PJ, Ford KM. Why Gödel's theorem cannot refute computationalism. Artificial Intelligence. 1998;104(1):265–286. Available from: http://www.sciencedirect.com/science/article/pii/S0004370298000526.

[148] Feferman S. Penrose's Gödelian argument. In: Psyche: An Interdisciplinary Journal of Research on Consciousness; 1996. p. 21–32.

[149] Lewis D. Lucas against Mechanism. Philosophy. 1969;44(169):231–233.

[150] Krajewski S. On Gödel's Theorem and Mechanism: Inconsistency or Unsoundness is Unavoidable in any Attempt to 'Out-Gö del' the Mechanist. undefined. 2007;Available from: http://content.iospress.com/articles/fundamenta-informaticae/fi81-1-3-11.

[151] Oppy G, Dowe D. The Turing Test. In: Edward N Zalta, editor. The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University; 2019. .

[152] Block N. Psychologism and behaviorism. The Philosophical Review. 1981;1981(1).

[153] Chalmers D. Does conceivability entail possibility? Conceivability and possibility. 2002;p. 145–200.

[154] McDermott D. On the Claim that a Table-Lookup Program Could Pass the Turing Test. Minds and Machines. 2014;24(2):143–188.

[155] Menabrea LF, Lovelace A. Sketch of the analytical engine invented by Charles Babbage. Taylor's Scientific Memoirs. 1842;.

[156] Bringsjord S, Bello P, Ferrucci D. Creativity, the Turing Test, and the (better) Lovelace Test. Minds and Machines. 2001;(11):3–27.

[157] Walsh T. The Meta-Turing Test. In: Workshops at the Thirty-First AAAI Conference; 2017. p. 132–137. Available from: https://www.aaai.org/ocs/index.php/WS/AAAIW17/paper/download/15233/14656.

[158] Harnad S. Minds, Machines and Turing. In: Moor JH, editor. The Turing Test. Dordrecht: Springer Netherlands; 2003. p. 253–273. Available from: http://cogprints.org/2615/1/harnad00.turing.html.

[159] Schweizer P. The Externalist Foundations of a Truly Total Turing Test. Minds and Machines. 2012;22(3):191–212. Available from: https://doi.org/10.1007/s11023-012-9272-4.

[160] Schweizer P. Could There be a Turing Test for Qualia? Revisiting Turing and His Test: Comprehensiveness, Qualia, and the Real World. 2012;p. 41.

[161] Riedl MO. The Lovelace 2.0 Test of Artificial Creativity and Intelligence; 2014. Available from: https://arxiv.org/pdf/1410.6142.

[162] Winston PH. The Strong Story Hypothesis and the Directed Perception Hypothesis. In: AAAI Fall Symposium Series; 2011. Available from: https://www.aaai.org/ocs/index.php/FSS/FSS11/paper/download/4125/4534.

[163] Winograd T. Understanding natural language. Cognitive Psychology. 1972;3(1):1–191. Available from: http://www.sciencedirect.com/science/article/pii/0010028572900023.

[164] Ernest Davis. Collection of Winograd Schemas; 2018. Available from: https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WSCollection.html.

[165] Dennett DC. Can Machines Think? In: Teuscher C, editor. Alan Turing: Life and Legacy of a Great Thinker. Berlin, Heidelberg: Springer Berlin Heidelberg; 2004. p. 295–316. Available from: https://doi.org/10.1007/978-3-662-05642-4_12.

[166] Levesque H, Davis E, Morgenstern L. The Winograd Schema Challenge. Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning. 2012;.

[167] Rohde H. Coherence-driven effects in sentence and discourse processing [Dissertation]. UC San Diego. Californien, USA; 2008.

[168] Roemmele M, Bejan CA, Gordon AS. Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. In: AAAI Spring Symposium Series; 2011. p. 90–95. Available from: https://www.aaai.org/ocs/index.php/SSS/SSS11/paper/download/2418/2960.

[169] NIH Office of Communications. NIH embraces bold, 12-year scientific vision for BRAIN Initiative: New report outlines initiative goals, budget, and timeline.;. Available from: https://www.nih.gov/news-events/news-releases/nih-embraces-bold-12-year-scientific-vision-brain-initiative.

[170] Alivisatos AP, Chun M, Church GM, Greenspan RJ, Roukes ML, Yuste R. The brain activity map project and the challenge of functional connectomics. Neuron. 2012;74(6):970–974.

[171] Markram H. The Blue Brain Project. Nature Reviews Neuroscience. 2006;7(2):153–160. Available from: https://www.nature.com/articles/nrn1848.pdf.

[172] Herculano-Houzel S. The human brain in numbers: a linearly scaled-up primate brain. Frontiers in human neuroscience. 2009;3:31.

[173] Lange W. Cell number and cell density in the cerebellar cortex of man and some other mammals. Cell and Tissue Research. 1975;157(1):115–124. Available from: https://doi.org/10.1007/BF00223234.

[174] Andersen BB, Korbo L, Pakkenberg B. A quantitative study of the human cerebellum with unbiased stereological techniques. The Journal of comparative neurology. 1992;326(4):549–560.

[175] SHARIFF GA. Cell counts in the primate cerebral cortex. The Journal of comparative neurology. 1953;98(3):381–400.

[176] Pelvig DP, Pakkenberg H, Stark AK, Pakkenberg B. Neocortical glial cell numbers in human brains. Neurobiology of Aging. 2008;29(11):1754–1762. Available from: http://www.sciencedirect.com/science/article/pii/S0197458007001686.

[177] Trepel M. Neuroanatomie: Struktur und Funktion. 7th ed. München, Deutschland: Elsevier; 2017.

[178] Wolf U, Rapoport MJ, Schweizer TA. Evaluating the affective component of the cerebellar cognitive affective syndrome. The Journal of neuropsychiatry and clinical neurosciences. 2009;21(3):245–253.

[179] Manni E, Petrosini L. A century of cerebellar somatotopy: a debated representation. Nature reviews Neuroscience. 2004;5(3):241–249.

[180] Apps R, Garwicz M. Anatomical and physiological foundations of cerebellar information processing. Nature reviews Neuroscience. 2005;6(4):297–311.

[181] Hawkes R, Herrup K. Aldolase C/zebrin II and the regionalization of the cerebellum. Journal of Molecular Neuroscience. 1995;6(3):147–158. Available from: https://doi.org/10.1007/BF02736761.

[182] Eccles JC. The cerebellum as a neuronal machine. Springer Science & Business Media; 1967.

[183] Apps R, Hawkes R. Cerebellar cortical organization: a one-map hypothesis. Nature reviews Neuroscience. 2009;10(9):670–681.

[184] Bostan AC, Dum RP, Strick PL. Cerebellar networks with the cerebral cortex and basal ganglia. Trends in Cognitive Sciences. 2013;17(5):241–254.

[185] Wolman D. The split brain: A tale of two halves. Nature News. 2012;483(7389):260.

[186] Dehaene S, Changeux JP. Experimental and Theoretical Approaches to Conscious Processing. Neuron. 2011;70(2):200–227.

[187] Grossberg S. Recurrent neural networks. Scholarpedia. 2013;8(2):1888.

[188] Tononi G, Koch C. Consciousness: here, there and everywhere? Philosophical transactions of the Royal Society of London Series B, Biological sciences. 2015;370(1668).

[189] McKenna M, Coates DJ. Compatibilism. In: Edward N Zalta, editor. The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University; 2018. .

[190] Campagna JA, Miller KW, Forman SA. Mechanisms of actions of inhaled anesthetics. The New England journal of medicine. 2003;348(21):2110–2124.

[191] Hemmings HC, Adamo AIB. Activation of Endogenous Protein Kinase C by Halothane in Synaptosomes. Anesthesiology: The Journal of the American Society of Anesthesiologists. 1996;84(3):652–662. Available from: https://anesthesiology.pubs.asahq.org/jasa/content_public/journal/jasa/931287/0000542-199603000-00021.pdf.

[192] Pan JZ, Xi J, Tobias JW, Eckenhoff MF, Eckenhoff RG. Halothane binding proteome in human brain cortex. Journal of proteome research. 2007;6(2):582–592.

[193] Baldassarre D, Scarpati G, Piazza O. Mechanisms of Action of Inhaled Volatile General Anesthetics: Unconsciousness at the Molecular Level. In: Cascella M, editor. GENERAL ANESTHESIA RESEARCH. vol. 150 of Neuromethods. [S.l.]: HUMANA; 2020. p. 109–123.

[194] Meyer DK, Olenik C, Hofmann F, Barth H, Leemhuis J, Brünig I, et al. Regulation of Somatodendritic GABA A Receptor Channels in Rat Hippocampal Neurons: Evidence for a Role of the Small GTPase Rac1. The Journal of Neuroscience. 2000;20(18):6743–6751.

[195] Furukawa K, Wang Y, Yao PJ, Fu W, Mattson MP, Itoyama Y, et al. Alteration in calcium channel properties is responsible for the neurotoxic action of a familial frontotemporal dementia tau mutation. Journal of neurochemistry. 2003;87(2):427–436.

[196] Mironov SL, Richter DW. Cytoskeleton mediates inhibition of the fast Na+ current in respiratory brainstem neurons during hypoxia. The European journal of neuroscience. 1999;11(5):1831–1834.

[197] Heo L, Feig M. Experimental accuracy in protein structure refinement via molecular dynamics simulations. Proceedings of the National Academy of Sciences of the United States of America. 2018;115(52):13276–13281.

[198] JA Craddock T, R Hameroff S, T Ayoub A, Klobukowski M, A Tuszynski J. Anesthetics act in quantum channels in brain microtubules to prevent consciousness. Current topics in medicinal chemistry. 2015;15(6):523–533.

[199] Craddock TJA, St George M, Freedman H, Barakat KH, Damaraju S, Hameroff S, et al. Computational predictions of volatile anesthetic interactions with the microtubule cytoskeleton: implications for side effects of general anesthesia. PloS one. 2012;7(6):e37251.

[200] Butts CA, Xi J, Brannigan G, Saad AA, Venkatachalan SP, Pearce RA, et al. Identification of a fluorescent general anesthetic, 1-aminoanthracene. Proceedings of the National Academy of Sciences of the United States of America. 2009;106(16):6501–6506.

[201] Berth M, Moser FM, Kolbe M, Bernhardt J. The state of the art in the analysis of two-dimensional gel electrophoresis images. Applied Microbiology and Biotechnology. 2007;76(6):1223–1243. Available from: https://doi.org/10.1007/s00253-007-1128-0.

[202] Emerson DJ, Weiser BP, Psonis J, Liao Z, Taratula O, Fiamengo A, et al. Direct modulation of microtubule stability contributes to anthracene general anesthesia. Journal of the American Chemical Society. 2013;135(14):5389–5398.

[203] Lane CA, Hardy J, Schott JM. Alzheimer's disease. European journal of neurology. 2018;25(1):59–70.

[204] Mandelkow E, von Bergen M, Biernat J, Mandelkow EM. Structural principles of tau and the paired helical filaments of Alzheimer's disease. Brain pathology (Zurich, Switzerland). 2007;17(1):83–90.

[205] Li B, Chohan MO, Grundke-Iqbal I, Iqbal K. Disruption of microtubule network by Alzheimer abnormally hyperphosphorylated tau. Acta neuropathologica. 2007;113(5):501–511.

[206] Cash AD, Aliev G, Siedlak SL, Nunomura A, Fujioka H, Zhu X, et al. Microtubule Reduction in Alzheimer's Disease and Aging Is Independent of Tau Filament Formation. The American Journal of Pathology. 2003;162(5):1623–1627.

[207] Balestrino R, Schapira AHV. Parkinson Disease. European journal of neurology. 2019;.

[208] Dexter DT, Jenner P. Parkinson disease: from pathology to molecular disease mechanisms. Free radical biology & medicine. 2013;62:132–144.

[209] Rösler TW, Tayaranian Marvian A, Brendel M, Nykänen NP, Höllerhage M, Schwarz SC, et al. Four-repeat tauopathies. Progress in Neurobiology. 2019;180:101644. Available from: http://www.sciencedirect.com/science/article/pii/S0301008219300863.

[210] Calogero AM, Mazzetti S, Pezzoli G, Cappelletti G. Neuronal microtubules and proteins linked to Parkinson's disease: a relevant interaction? Biological chemistry. 2019;400(9):1099–1112.

[211] Roos RAC. Huntington's disease: a clinical review. Orphanet journal of rare diseases. 2010;5:40.

[212] Sari Y. Huntington's Disease: From Mutant Huntingtin Protein to Neurotrophic Factor Therapy. International Journal of Biomedical Science : IJBS. 2011;7(2):89–100.

[213] DiProspero NA, Chen EY, Charles V, Plomann M, Kordower JH, Tagle DA. Early changes in Huntington's disease patient brains involve alterations in cytoskeletal and synaptic elements. Journal of Neurocytology. 2004;33(5):517–533. Available from: https://doi.org/10.1007/s11068-004-0514-8.

[214] Facal F, Costas J. Evidence of association of the DISC1 interactome gene set with schizophrenia from GWAS. Progress in neuro-psychopharmacology & biological psychiatry. 2019;95:109729.

[215] Matsuzaki S, Tohyama M. Molecular mechanism of schizophrenia with reference to disrupted-in-schizophrenia 1 (DISC1). Neurochemistry international. 2007;51(2-4):165–172.

[216] Duff BJ, Macritchie KAN, Moorhead TWJ, Lawrie SM, Blackwood DHR. Human brain imaging studies of DISC1 in schizophrenia, bipolar disorder and depression: a systematic review. Schizophrenia research. 2013;147(1):1–13.

[217] Ito H, Morishita R, Nagata KI. Schizophrenia susceptibility gene product dysbindin-1 regulates the homeostasis of cyclin D1. Biochimica et biophysica acta. 2016;1862(8):1383–1391.

[218] Konstam MA, Hill JA, Kovacs RJ, Harrington RA, Arrighi JA, Khera A. The Academic Medical System: Reinvention to Survive the Revolution in Health Care. Journal of the American College of Cardiology. 2017;69(10):1305–1312.

[219] Boeldt DL, Wineinger NE, Waalen J, Gollamudi S, Grossberg A, Steinhubl SR, et al. How Consumers and Physicians View New Medical Technology: Comparative Survey. Journal of Medical Internet Research. 2015;17(9):e215.

[220] Steinhubl SR, Topol EJ. Moving From Digitalization to Digitization in Cardiovascular Care: Why Is it Important, and What Could it Mean for Patients and Providers? Journal of the American College of Cardiology. 2015;66(13):1489–1496.

[221] Senn S. Dichotomania: an obsessive compulsive disorder that is badly affecting the quality of analysis of pharmaceutical trials. Proceedings of the International Statistical Institute, 55th Session, Sydney. 2005;.

[222] Trayanova N. From genetics to smart watches: developments in precision cardiology. Nature reviews Cardiology. 2019;16(2):72–73.

[223] Kleinbaum DG, Dietz K, Gail M, Klein M, Klein M. Logistic regression. Springer; 2002.

[224] Mortazavi BJ, Downing NS, Bucholz EM, Dharmarajan K, Manhapra A, Li SX, et al. Analysis of Machine Learning Techniques for Heart Failure Readmissions. Circulation Cardiovascular quality and outcomes. 2016;9(6):629–640.

[225] Bonderman D. Artificial intelligence in cardiology. Wiener klinische Wochenschrift. 2017;129(23):866–868. Available from: https://doi.org/10.1007/s00508-017-1275-y.

[226] Schläpfer J, Wellens HJ. Computer-Interpreted Electrocardiograms: Benefits and Limitations. Journal of the American College of Cardiology. 2017;70(9):1183–1192. Available from: http://www.onlinejacc.org/content/accj/70/9/1183.full.pdf.

[227] Shah AP, Rubin SA. Errors in the computerized electrocardiogram interpretation of cardiac rhythm. Journal of Electrocardiology. 2007;40(5):385–390. Available from: http://www.sciencedirect.com/science/article/pii/S0022073607000696.

[228] Greenspan H, van Ginneken B, Summers RM. Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique. IEEE Transactions on Medical Imaging. 2016;35(5):1153–1159.

[229] Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nature Medicine. 2019;25(1):65–69. Available from: https://www.nature.com/articles/s41591-018-0268-3.pdf.

[230] Rajpurkar P, Hannun AY, Haghpanahi M, Bourn C, Ng AY. Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks; 2017. Available from: https://arxiv.org/pdf/1707.01836.

[231] Schmidhuber J. Deep learning in neural networks: an overview. Neural networks : the official journal of the International Neural Network Society. 2015;61:85–117.

[232] Di W, Bhardwaj A, Wei J. Deep learning essentials: Your hands-on guide to the fundamentals of deep learning and neural network modeling. Birmingham, UK: Packt Publishing; 2018. Available from: http://proquest.tech.safaribooksonline.de/9781785880360.

[233] van Rijsbergen CJ. Information Retrieval. 2nd. Newton, MA. USA: Butterworth-Heinemann; 1979.

[234] Hand D, Christen P. A note on using the F-measure for evaluating record linkage algorithms. Statistics and Computing. 2018;28(3):539–547.

[235] Hand DJ. Evaluating diagnostic tests: the area under the ROC curve and the balance of errors. Statistics in medicine. 2010;29(14):1502–1510.

[236] Kligfield P, Gettes LS, Bailey JJ, Childers R, Deal BJ, Hancock EW, et al. Recommendations for the Standardization and Interpretation of the Electrocardiogram: Part I: The Electrocardiogram and Its Technology A Scientific Statement

From the American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology; the American College of Cardiology Foundation; and the Heart Rhythm Society Endorsed by the International Society for Computerized Electrocardiology. Journal of the American College of Cardiology. 2007;49(10):1109–1127. Available from: http://www.sciencedirect.com/science/article/pii/S073510970700232X.

[237] Dumitrascu OM, Demaerschalk BM, Valencia Sanchez C, Almader-Douglas D, O'Carroll CB, Aguilar MI, et al. Retinal Microvascular Abnormalities as Surrogate Markers of Cerebrovascular Ischemic Disease: A Meta-Analysis. Journal of Stroke and Cerebrovascular Diseases. 2018;27(7):1960–1968. Available from: http://www.sciencedirect.com/science/article/pii/S1052305718301034.

[238] Seidelmann SB, Claggett B, Bravo PE, Gupta A, Farhad H, Klein BE, et al. Retinal Vessel Calibers in Predicting Long-Term Cardiovascular Outcomes: The Atherosclerosis Risk in Communities Study. Circulation. 2016;134(18):1328–1338.

[239] Balmforth C, van Bragt JJ, Ruijs T, Cameron JR, Kimmitt R, Moorhouse R, et al. Chorioretinal thinning in chronic kidney disease links to inflammation and endothelial dysfunction. JCI insight. 2016;1(20):e89173.

[240] Farrah TE, Webb DJ, Dhaun N. Retinal fingerprints for precision profiling of cardiovascular risk. Nature Reviews Cardiology. 2019;16(7):379–381.

[241] Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nature Biomedical Engineering. 2018;2(3):158–164. Available from: https://www.nature.com/articles/s41551-018-0195-0.pdf.

[242] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the Inception Architecture for Computer Vision; 2016. p. 2818–2826. Available from: http://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Szegedy_Rethinking_the_Inception_CVPR_2016_paper.pdf.

[243] Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhutdinov R, et al.. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention; 2015. Available from: https://arxiv.org/pdf/1502.03044.

[244] © UK Biobank Limited 2019. About UK Biobank; 2019. Available from: http://www.ukbiobank.ac.uk/about-biobank-uk/.

[245] EyePACS LLC. EyePACS; 2019. Available from: https://www.eyepacs.com/.

[246] Fleming TR, Powers JH. Biomarkers and surrogate endpoints in clinical trials. Statistics in medicine. 2012;31(25):2973–2984.

[247] Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters. 2006;27(8):861–874.

[248] Graham I, Atar D, Borch-Johnsen K, Boysen G, Burell G, Cifkova R, et al. European guidelines on cardiovascular disease prevention in clinical practice: executive summary. Fourth Joint Task Force of the European Society of Cardiology and

other societies on cardiovascular disease prevention in clinical practice (constituted by representatives of nine societies and by invited experts). European journal of cardiovascular prevention and rehabilitation : official journal of the European Society of Cardiology, Working Groups on Epidemiology & Prevention and Cardiac Rehabilitation and Exercise Physiology. 2007;14 Suppl 2:E1–40.

[249] Jia Y, Bailey ST, Hwang TS, McClintic SM, Gao SS, Pennesi ME, et al. Quantitative optical coherence tomography angiography of vascular abnormalities in the living human eye. Proceedings of the National Academy of Sciences of the United States of America. 2015;112(18):E2395–402.

[250] Baker J. Das Standardmodell. In: 50 Schlüsselideen Physik. Springer; 2009. p. 144–147.

[251] Nicolai H, Kleinschmidt A. E10: Eine fundamentale Symmetrie der Physik? Neuer Zugang zur Quantengravitation. Physik in unserer Zeit. 2010;41(3):134–140.

[252] Craig WL. The teleological argument and the anthropic principle. The Logic of Rational Theism. 2018;.

[253] Bostrom N. Anthropic bias: Observation selection effects in science and philosophy. Routledge; 2013.

[254] Carter B. Large Number Coincidences and the Anthropic Principle in Cosmology. In: Longair MS, editor. Confrontation of Cosmological Theories with Observational Data. International Astronomical Union/Union Astronomique Internationale. Dordrecht: Springer Netherlands; 1974. p. 291–298. Available from: http://adsabs.harvard.edu/full/1974IAUS...63..291C.

[255] Wheeler JA. Foundational problems in the special sciences. RE Butts and J Hintikka Dordrecht, Reidel. 1977;3.

[256] Crane L. Possible implications of the quantum theory of gravity: An introduction to the meduso-anthropic principle. Foundations of Science. 2010;15(4):369–373.

[257] Hameroff S. Consciousness, Free Will and Quantum Brain Biology—The "Orch OR" Theory. Quantum Physics Meets the Philosophy of Mind: New Essays on the Mind-Body Relation in Quantum-Theoretical Perspective. 2014;56:99–134.

[258] Tegmark M. Consciousness as a state of matter. Chaos, Solitons & Fractals. 2015;76:238–270.

[259] Banich MT, Compton RJ. Cognitive Neuroscience. Cambridge University Press; 2018.

[260] Koch C, Massimini M, Boly M, Tononi G. Neural correlates of consciousness: progress and problems. Nature Reviews Neuroscience. 2016;17(5):307–321.

[261] Bauser M, Lehmann L. Positronen-Emissions-Tomographie. Chemie in unserer Zeit. 2012;46(2):80–99.

[262] Logothetis NK, Pfeuffer J. On the nature of the BOLD fMRI contrast mechanism. Magnetic Resonance Imaging. 2004;22(10):1517–1531.

[263] Feldman J. The neural binding problem(s). Cognitive Neurodynamics. 2013;7(1):1–11. Available from: https://doi.org/10.1007/s11571-012-9219-8.

[264] Engel AK, Singer W. Temporal binding and the neural correlates of sensory awareness. Trends in Cognitive Sciences. 2001;5(1):16–25.

[265] Opitz B. Neural binding mechanisms in learning and memory. Neuroscience & Biobehavioral Reviews. 2010;34(7):1036–1046. Available from: http://www.sciencedirect.com/science/article/pii/S0149763409001675.

[266] Lee U, Kim S, Noh GJ, Choi BM, Mashour GA. Propofol Induction Reduces the Capacity for Neural Information Integration: Implications for the Mechanism of Consciousness and General Anesthesia. Nature Precedings. 2009;p. 1. Available from: https://www.nature.com/articles/npre.2008.1244.2.pdf.

[267] Jolij J, Lamme VAF. Repression of unconscious information by conscious processing: evidence from affective blindsight induced by transcranial magnetic stimulation. Proceedings of the National Academy of Sciences of the United States of America. 2005;102(30):10747–10751.

[268] de Graaf TA, Hsieh PJ, Sack AT. The 'correlates' in neural correlates of consciousness. Neuroscience & Biobehavioral Reviews. 2012;36(1):191–197. Available from: http://www.sciencedirect.com/science/article/pii/S0149763411001072.

[269] Seung S. Connectome: How the brain's wiring makes us who we are. HMH; 2012.

[270] Perkel JM. LIFE SCIENCE TECHNOLOGIES: This Is Your Brain: Mapping the Connectome. Science. 2013;339(6117):350–352.

[271] Bentley B, Branicky R, Barnes CL, Chew YL, Yemini E, Bullmore ET, et al. The Multilayer Connectome of Caenorhabditis elegans. PLoS computational biology. 2016;12(12):e1005283.

[272] Sporns O. The human connectome: a complex network. Annals of the New York Academy of Sciences. 2011;1224:109–125.

[273] van Essen DC, Ugurbil K, Auerbach E, Barch D, Behrens TEJ, Bucholz R, et al. The Human Connectome Project: A data acquisition perspective. NeuroImage. 2012;62(4):2222–2231. Available from: http://www.sciencedirect.com/science/article/pii/S1053811912001954.

[274] Asher J, Stimson D. $40 Million Awarded to Trace Human Brain's Connections: Souped-up Scanners to Reveal Intricate Circuitry in High Resolution; 2010. Available from: https://web.archive.org/web/20120110052518/http://www.nimh.nih.gov/science-news/2010/40-million-awarded-to-trace-human-brains-connections.shtml.

[275] Buttazzo G. Artificial consciousness: Hazardous questions (and answers). Artificial Intelligence in Medicine. 2008;44(2):139–146. Available from: http://www.sciencedirect.com/science/article/pii/S0933365708000870.

[276] Walsh T, Annuscheit M. Elon Musk hat Unrecht: Die KI-Singularität wird uns nicht alle umbringen.; 2018. Available from: https://tinyurl.com/yy7kl4r5.

[277] Johnson DG. Computer systems: Moral entities but not moral agents. Ethics and Information Technology. 2006;8(4):195–204. Available from: https://doi.org/10.1007/s10676-006-9111-5.

[278] Sicherer Straßenverkehr: Lebensrettende Technik für Neufahrzeuge; 16.04.2019. Available from: https://www.europarl.europa.eu/news/de/press-room/20190410IPR37528/sicherer-strassenverkehr-lebensrettende-technik-fur-neufahrzeuge.

[279] Di Fabio U, Broy M, Brüngger RJ, Eichhorn U, Grunwald A, Heckmann D, et al.. Automatisiertes und vernetztes Fahren: Ethik-Kommission;. Available from: https://www.bmvi.de/SharedDocs/DE/Publikationen/DG/bericht-der-ethik-kommission.pdf?__blob=publicationFile.

[280] red. USA: Gerichte nutzen Künstliche Intelligenz für Urteile: Kritiker warnen davor, dass Fehler aus der Vergangenheit so nur wiederholt werden; 2019. Available from: https://www.derstandard.at/story/2000098105268/usa-gerichte-nutzen-kuenstliche-intelligenz-fuer-urteile.

[281] Weizenbaum J. Computer power and human reason: From judgment to calculation. WH Freeman & Co; 1976.

[282] Dullinger S. Schuldrecht Allgemeiner Teil. vol. Band 2 of Bürgerliches Recht. 6th ed. Wien: Verlag Österreich; 2017.

[283] Nevejans N. European civil law rules in robotics: Study. [Luxembourg]: Publications Office; 2016.

[284] Europäisches Parlament. Entschließung des Europäischen Parlaments vom 16. Februar 2017 mit Empfehlungen an die Kommission zu zivilrechtlichen Regelungen im Bereich Robotik: 2015/2103(INL); 16.02.2017. Available from: http://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_DE.html?redirect.

[285] Tegmark M. Life 3.0: Being human in the age of artificial intelligence; 2018.

[286] Tegmark M. Benefits & Risks of Artificial Intelligence; 2016. Available from: https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/.

[287] The Future of Life Institute (FLI): Team; 2018. Available from: https://futureoflife.org/team/.

[288] Asimov I. I, robot. Garden City, N.Y.: Doubleday; 1950.

[289] Miyazaki K, Takeno J. The Necessity of a Secondary System in Machine Consciousness. Procedia Computer Science. 2014;41:15–22. Available from: http://www.sciencedirect.com/science/article/pii/S1877050914015245.

[290] Myers CB. Ethical Robotics and Why We Really Fear Bad Robots: Interview with Dr. Patrick Lin; 2010. Available from: https://tinyurl.com/y33zlpwf.

[291] Bryson JJ. Patiency is not a virtue: AI and the design of ethical systems. In: 2016 AAAI Spring Symposium Series; 2016. .

[292] Parry V. Principles of robotics: Regulating robots in the real world; 2010. Available from: https://epsrc.ukri.org/research/ourportfolio/themes/engineering/activities/principlesofrobotics/.

[293] Allen C, Varner G, Zinser J. Prolegomena to any future artificial moral agent. Journal of Experimental & Theoretical Artificial Intelligence. 2000;12(3):251–261.

[294] Hirsch PA. Kants Einleitung in die Rechtslehre von 1784. Universitätsverlag Göttingen; 2012.

[295] Goertzel B, Mossbridge J, Monroe E, Hanson D, Yu G. Humanoid Robots as Agents of Human Consciousness Expansion;. Available from: http://arxiv.org/pdf/1709.07791v1.

[296] Bryson JJ. Robots should be slaves. Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues. 2010;p. 63–74.

[297] Bryson JJ. A role for consciousness in action selection. International Journal of Machine Consciousness. 2012;04(02):471–482.

[298] Gee QP. Group agency, moral agency, and moral patiency. [Santa Barbara, Calif.]: University of California, Santa Barbara; 2013.