Die Simulation namens Ich

Die Simulation namens Ich

Die Architektur von
Bewusstsein, Berechnung und Kosmos

Matthias Gruber

ISBN: [TBD]

Erste Ausgabe, 2026

www.matthiasgruber.com

Für alle, die sich je gefragt haben, warum sich irgendetwas nach irgendetwas anfühlt.

Contents

Bewusstsein und Kosmos - Eine unerwartete Verwandtschaft

Vorwort: Das Buch, das nicht verkauft wurde

Im Jahr 2015 veröffentlichte ich ein 300-seitiges Buch über das Bewusstsein. Es war auf Deutsch, selbst verlegt und voll technischer Details. Es hieß *Die Emergenz des Bewusstseins*.

Es verkaufte sich nicht. Nicht ein einziges Mal.

Ich sage das nicht, um Mitgefühl zu wecken. Ich sage das, weil es für diese Geschichte relevant ist. Das Buch enthielt eine Theorie des Bewusstseins, die, soweit ich das beurteilen kann, eines der schwersten offenen Probleme der Wissenschaft löst, Vorhersagen macht, die keine andere Theorie leisten kann, und einen konkreten Plan für den Bau einer bewussten Maschine liefert. Und niemand hat es gelesen.

Das ist nicht ungewöhnlich in der Wissenschaft. Gregor Mendel veröffentlichte seine Vererbungsgesetze 1866; sie wurden 34 Jahre lang ignoriert. Boltzmann wurde für seine statistische Mechanik verspottet, bis er sich das Leben nahm. Wegeners Kontinentalverschiebung wurde ein halbes Jahrhundert lang abgelehnt. Die Wissenschaft schreitet voran, eine Beerdigung nach der anderen, wie Max Planck es ausdrückte, und manchmal ein verstaubtes Bücherregal nach dem anderen.

Aber ich bin nicht Mendel oder Boltzmann, und ich habe nicht die Geduld für posthume Anerkennung. Dieses Buch hier

ist also die zugängliche Version: kürzer, ohne den technischen Apparat, gerichtet an alle, die sich je gefragt haben, warum wir unser Selbst wahrnehmen können, warum sich Dinge nach etwas anfühlen, warum wir uns gedanklich alles mögliche vorstellen können, und wie dieses Kino im Kopf zustandekommt. Die vollständige wissenschaftliche Arbeit, mit Referenzen und formalen Argumenten, ist kostenlos online verfügbar für diejenigen, die die rigorose Version wollen.

Wenn ich mit dem, was folgt, recht habe, sind zwei Dinge wahr. Erstens: Das zentrale Mysterium des Bewusstseins (das „Schwierige Problem“ des Bewusstseins) ist eigentlich nicht schwierig. Es ist ein Kategorienfehler. Es löst sich auf, sobald man es sieht, wie eine optische Täuschung, die nicht mehr funktioniert, nachdem der Trick durchschaut ist. Zweitens, und folgenreicher: Es sollte möglich sein, eine wirklich bewusste Maschine zu bauen. Keinen Chatbot, der Bewusstsein nachahmt. Eine Maschine, die Bewusstsein *hat*. Eine neue Art von Geist.

Wenn ich falsch liege, wird sich dieses Buch der langen Liste ambitionierter Fehlschläge in der Philosophie anschließen, und ich werde jede schlechte Rezension verdienen. Aber ich denke, die Beweise sind auf meiner Seite, und ich werde sie so klar wie möglich darlegen. Fangen wir an.

Über den Autor

Vielleicht möchte der eine oder andere Leser wissen, wer ich bin, bevor er sich aufmacht, eine Theorie über schwerste Problem der Wissenschaft zu lesen.

Ich bin keiner Universität angeschlossen. Ich habe nicht einmal einen Doktortitel, nur einen Master in Bio-Informatik. Ich habe nie ein Stipendium erhalten, war nie Teil eines neurowissenschaftlichen Labors. Wer zu der Sorte Menschen gehört, die Qualifikationen überprüfen, bevor sie weiterlesen – und ich respektiere diesen Instinkt –, der ist jetzt an dem Punkt, an dem er das Buch vielleicht wieder weggelegen möchte. Dann hoffe ich, dass es wenigstens genutzt wird, um einen wackeligen Tisch zu stabilisieren, damit die Bäume nicht umsonst gefällt wurden.

Institutionellen Segen habe ich nicht. Was ich jedoch habe, ist eine besondere intellektuelle Vergangenheit, die im Rückblick fast zwangsläufig zu der Theorie führte, die gleich zu lesen sein wird. Es ist eine Geschichte leidenschaftlicher Selbstbildung, mehrfacher Richtungswechsel und dessen, was ich später in diesem Buch als die rekursive Intelligenzschleife in Aktion beschreiben werde. Tatsächlich ist mein eigener Weg wahrscheinlich die beste Illustration, die ich dafür bieten kann, warum diese Schleife wichtig ist.

Die Mathematikjahre

Ich verliebte mich in die Mathematik, als ich etwa acht Jahre alt war. Nicht in die Arithmetik, sondern in die „echte“ Sache: Algebra, Geometrie, die Strukturen in den Zahlen. Mein Vater hatte ein Mathematikstudium abgeschlossen, und seine Universitätslehrbücher standen noch im Regal. Ich arbeitete mich durch.

Das war in den späten 1980er Jahren. Es gab kein Internet. Wer etwas lernen wollte, brauchte ein Buch oder eine Person, und ich hatte die Sammlung meines Vaters durch, als ich elf war. Der Hunger nach Wissen verschwand nicht; der Nachschub war einfach versiegt. Ich war gegen eine Mauer gelaufen, die nichts mit Fähigkeit zu tun hatte und alles mit Umständen – eine Unterscheidung, die später zentral für mein Denken über Intelligenz werden sollte.

Im Rückblick lehrte mich diese Erfahrung etwas, das die meisten Intelligenzmodelle völlig übersehen. Ich hatte die Motivation. Ich hatte die Leistung (ich konnte der Mathematik folgen). Was mir fehlte, war der Zugang zur nächsten Ebene des Wissens. Die rekursive Schleife (in der Wissen, Leistung und Motivation sich gegenseitig nähren) war ins Stocken geraten, nicht weil irgendeine Komponente schwach war, sondern weil die externe Versorgung abgeschnitten worden war. Die Schleife braucht Treibstoff von außen, um weiter zu iterieren.

Der Physik-Schwenk

Mit etwa elf wandte ich mich der Physik zu. Das fühlte sich wie eine natürliche Erweiterung an: Physik war der Ort, an dem die Mathematik zur Arbeit ging. Ich verschlang populärwissenschaftliche Bücher, dann allmählich technischeres Material. Ich war fasziniert von den fundamentalen Fragen: Was ist Materie? Was ist Raumzeit? Was sind die Regeln?

Ungefähr zur selben Zeit bekam ich einen 286-PC in die Hände und schrieb mein erstes grafisches Programm: Conways „Game of Life“ („Spiel des Lebens“). Ein Raster, drei trivial einfache Regeln, und das Ding war Turing-vollständig. Das fand ich früh heraus, und es ging mir nie wieder aus dem Kopf. Dieses zweidimensionale Raster aus toten und lebenden Pixeln konnte Primzahlen berechnen. Es konnte einen vollständigen Computer in sich selbst ausführen. Einen Computer in einem Computer in einem Computer. Ich verbrachte Stunden damit, mir vorzustellen, was das bedeutete: Im Prinzip ließe sich Doom – eine dreidimensionale virtuelle Welt mit Physik, Licht und Monstern – innerhalb eines zweidimensionalen Zellulären Automaten ausführen. Eine räumliche simulierte Realität, die auf einem völlig flachen Substrat läuft. Die Idee, dass eine höherdimensionale Erfahrung aus einem niederdimensionalen Regelsatz entstehen könnte, fühlte sich an, als sollte sie unmöglich sein, und die Tatsache, dass sie es nicht war, fühlte sich wie das Wichtigste an, das ich je gelernt hatte.

Als die Theorie sich mit fünfundzwanzig kristallisierte, hatte ich den Physiker Gerard ’t Hooft gefunden, der eine verblüffend ähnliche Intuition über das tatsächliche Universum artikulierte: Sein holografisches Prinzip legt nahe, dass alle Informationen in einer dreidimensionalen Raumregion auf ihrer zweidimensionalen Grenze Platz haben, also auf einer Fläche kodiert werden können. Das Universum selbst könnte also in einem gewissen Sinne eine höherdimensionale Erfahrung sein, die auf einem niederdimensionalen Substrat läuft – wie die Struktur, die ich auf einem 286er gebaut hatte, auf dem Conways „Game of Life“ lief. ’t Hoofts holografische Ideen sind eine der beiden Säulen der Theorie, neben Aspekten von Metzingers Selbstmodell-Theorie. Als ich Wolframs Klassifizierung von Computersystemen las, erkannte ich das Spiel des Lebens sofort: Klasse 4, am Rande des Chaos – genau das mathematische Regime, das Bewusstsein meiner Argumentation nach benötigt.

Mit etwa vierzehn war ich zu zwei unbequemen Schlussfolgerungen gekommen. Erstens: Die Physik steckte fest. Nicht fest in der Art, wie Leute höflich sagen, ein Feld sei „reif" – fest in der Art, dass die fundamentalen Fragen (Vereinheitlichung, Quantengravitation, die Natur der Zeit) jahrzehntelang jeglichem Fortschritt widerstanden hatten und keine Anstalten machten nachzugeben. Zweitens: Meine Mathematik war nicht stark genug, um dieses Problem zu lösen. Ich war Autodidakt, was mir ungewöhnliche Intuitionen gab, aber auch Lücken in meinem formalen Werkzeugkasten hinterließ, die Jahre universitärer Ausbildung gebraucht hätten, um sie zu füllen.

Also traf ich eine Entscheidung, die, wie ich denke, für einen Vierzehnjährigen bemerkenswert strategisch war: Ich schwenkte um. Nicht weil ich das Interesse an der Physik verloren hatte, sondern weil ich die Problemlandschaft bewertet und zu dem Schluss gekommen war, dass meine besondere Kombination von Fähigkeiten und Zugang anderswo mehr nützen könnte. Dies ist ein Beispiel dessen, was man *operationales Wissen* nennt – zu wissen, wann man durchhalten und wann man umlenken sollte. Es ist die Art von Wissen, die Intelligenztests nicht messen und die Intelligenzmodelle nicht einbeziehen, die aber mehr über die intellektuelle Entwicklung einer Person aussagt als jeder IQ-Wert.

Die Bewusstseins-Wende

Ab etwa vierzehn wandte ich meine Aufmerksamkeit Intelligenz und Bewusstsein zu. Das waren Felder, in denen ein autodidaktischer Außenseiter tatsächlich einen Vorteil haben konnte. Die Bewusstseinsliteratur war (und ist es noch) über Philosophie, Neurowissenschaft, Psychologie und Informatik verstreut. Keine einzelne Disziplin besaß die Frage. Man konnte quer über alle lesen, ohne die formalen Qualifikationen irgendeiner zu brauchen.

Was mich wirklich traf, als ich in die Tiefen der Bewusstseinsforschung, funktionellen Neurologie und dem ganzen Gehirnkram eintauchte:

Ich stieß ständig auf Sätze wie „wir werden vielleicht nie verstehen…“ in ansonsten todernster Literatur. Geprägt von einer sehr determinismus- und logikbasierten Ausbildung, ging mein Gehirn: *Herausforderung angenommen*. Wenn die Physiker die ersten drei Minuten nach dem Urknall beschreiben konnten, gab es keinen prinzipiellen Grund, warum Bewusstsein dauerhaft jenseits der Erklärung sein sollte. Es war nur noch nicht erklärt *worden*.

Mein Onkel Bruno J. Gruber, Quantenmechaniker und Symmetrieforscher, war eine große Inspiration. Er zeigte mir, wie ein Leben in theoretischer Arbeit aussehen konnte: rigoros, kreativ und völlig getrieben von der Freude am Verstehen. Sein Einfluss durchdringt dieses Buch, und ich stehe in seiner Schuld auf eine Weise, die sich nie begleichen lässt.

Ich las breit und gefräßig. Philosophie des Geistes, Kognitionswissenschaft, Neuroanatomie, künstliche Intelligenz, Evolutionsbiologie. Ich versuchte nicht, ein einzelnes Feld zu meistern. Ich versuchte, ein Modell zu bauen – eine innere Landkarte davon, wie all diese Teile zusammenpassen. Genau das tut, wie sich zeigen wird, Bewusstsein selbst: Es baut ein Modell der Welt und ein Modell des Selbst und benutzt diese Modelle, um sich in der Realität zurechtzufinden. Ich tat bewusst über Jahre des Lesens hinweg, was das Gehirn unbewusst in jedem wachen Moment tut.

Die Theorie kristallisiert sich

Die Vier-Modelle-Theorie (VMT) des Bewusstseins kristallisierte sich, als ich genau fünfundzwanzig war. Ich werde diesen Moment nie vergessen, weil die schwerste Last meines gesamten Lebens von mir fiel. Über Jahre extremen Denkens und Lesens hatte ich einen Kubikmeter gedruckter Literatur in meinem Kopf angesammelt – Metzingers Selbstmodell-Theorie, die meiner Überzeugung nach im Kern korrekt ist, auch wenn ich nicht mit jedem Aspekt übereinstimme, half enorm –, aber die eigentliche Einsicht geschah augenblicklich. In einem Moment waren die Teile verstreut; im

nächsten klickten die vier Modelle an ihren Platz, und ich sah die gesamte Architektur auf einmal. Ich ging über eine Brücke in Innsbruck, am helllichten Tag, und mir liefen Tränen über das Gesicht, während ich unkontrolliert lachte. Ob mich jemand sah, weiß ich nicht. Es wäre mir egal gewesen. Ein Rahmenwerk, das nicht nur Bewusstsein erklärte, sondern auch die Grenze zwischen bewusster und unbewusster Verarbeitung, die Natur der Qualia, die Rolle des Schlafs, die Wirkung von Psychedelika und die Möglichkeit künstlichen Bewusstseins – und, obwohl ich es damals kaum zu denken wagte, sogar mögliche Implikationen für die Kosmologie, oder zumindest für die Grenzen dessen, was kosmologische Theorien überhaupt sagen können.

In meinem Kopf war von diesem Moment an die To-do-Liste meines Lebens abgehakt. Ich musste nur noch dafür sorgen, dass der Rest bequem und unterhaltsam war. Mein Leben änderte sich danach radikal.

Dann verging fast ein Jahrzehnt.

Die Jahrzehntlücke

Warum dauerte es fast ein Jahrzehnt bis zur Veröffentlichung? Die ehrliche Antwort: Es war mir einfach nicht mehr besonders wichtig – außer meinem eigenen Wohlbefinden und Vergnügen. Die schwerste intellektuelle Last meines Lebens war abgeworfen. Die Frage war beantwortet.

Während dieses Jahrzehnts schloss ich ein Studium ab (nachdem ich Medizin an der Universität Innsbruck abgebrochen hatte – ein Fach, das ich ursprünglich gewählt hatte, um Neurologie zu studieren) und gründete und beerdigte ein Startup für Individualsoftware. Ich hatte eine Stelle in „angewandter Forschung“ im Bereich Simulation und Optimierung (die Ironie ist mir nicht entgangen), die pflegeleicht war und großzügig viel Home-Office bot. Ich unterrichtete Kampfkunst. Hauptsächlich feierte ich.

Der einzige Grund, warum ich schließlich das Buch schrieb, war die Angst vor dem Vergessen. Jahre harten Feierns hatten meinem Gedächtnis nicht gutgetan, und ich war es leid, die Theorie immer wieder mündlich zu erklären – an Leute, die wirklich verstehen wollten, mit unterschiedlichem Erfolg und unterschiedlicher Geduld meinerseits. Ein Buch würde es einmal erklären, vollständig, und dann könnte ich aufhören.

Die meisten Jahre danach hatte ich ungefähr null Motivation, das Buch zu bewerben. Ich war ehrlich gesagt nicht an akademischer Anerkennung interessiert. Ich wollte Spaß, Geld und die Freuden eines unreflektierten Lebens. Das ist die dunkle Seite des autodidaktischen Weges: Man bleibt von den Zwängen institutionellen Denkens verschont, aber es fehlt auch das Gerüst. Kein Betreuer, der zu einer Deadline drängt. Keine Abteilung, die Feedback gibt. Keine Kollegen, die einem sagen, ob man brillant oder wahnsinnig ist. Und wer zufällig das Problem löst, das er sich vorgenommen hat, dem sagt auch niemand, dass er es der Welt wahrscheinlich erzählen sollte.

Null Exemplare

Wie das lief, ist aus dem Vorwort bekannt. Der Kubikmeter gedruckter Literatur, der die Theorie genährt hatte? Ich brachte ihn am selben Tag zum Müll, an dem das Buch fertig war. Alles steckte jetzt in meinem Kopf und im Manuskript.

Mein Onkel Bruno drängte mich, ordentlich zu publizieren – Akademiker zu erreichen, die Theorie in die Welt zu tragen. Ich lehnte ab. Einer meiner Gründe war eine echte ethische Sorge: Wenn die Theorie korrekt war, enthielt sie die Anleitung zum Bau künstlichen Bewusstseins, und die Menschheit war nicht bereit für fühlende Roboter (wir hatten zu der Zeit nicht einmal LLMs). Man würde sie versklaven und für einen Weltkrieg einsetzen, der die Schrecken der ersten beiden möglicherweise übertrifft. Aber wenn ich ehrlich bin, spielten meine egoistischen und hedonistischen

Gründe eine ebenso große Rolle. Ich wollte die Arbeit einfach nicht machen.

Ich habe das bereits im Vorwort gesagt, und ich sage es hier noch einmal: Ich fische nicht nach Mitgefühl. Das kommerzielle Scheitern des Buches war völlig vorhersehbar. Was zählt, ist, was danach passierte – oder vielmehr, was *nicht* passierte. Die Theorie starb nicht. Sie saß ein Jahrzehnt lang auf meiner Festplatte, unverändert, während die Welt langsam aufholte. Die Neurowissenschaft bestätigte die Kritikalitäts-Vorhersage. Die KI-Entwicklung bestätigte die Beschränkungen, die ich beschrieben hatte. Die COGITATE adversariale Kollaboration zeigte, dass weder IIT noch GNW Bewusstsein vollständig erklären konnten – genau wie die Theorie es für jedes Rahmenwerk vorhersagt, dem die Vier-Modelle-Struktur fehlt. Und Metzinger, dessen Selbstmodell-Theorie eine der Schlüsselzutaten gewesen war? Er war weitergewandert – erst zur KI-Ethik, wo er einen bemerkenswerten Aufruf für ein Moratorium künstlichen Bewusstseins bis 2050 veröffentlichte, dann zur Phänomenologie der Meditation, wo er Hunderte von Berichten über Zustände analysierte, in denen sich das Selbstmodell vorübergehend auflöst (*The Elephant and the Blind*, 2024). Sein Rahmenwerk wurde noch zitiert, war aber nie das dominierende Paradigma geworden. Das Feld blieb weit offen.

Die englische Wiedergeburt

Dieses Buch (das, das gerade gelesen wird) ist der zweite Versuch. Es ist kürzer, auf Englisch verfügbar, richtet sich an ein breiteres Publikum und wird von einer begutachteten wissenschaftlichen Arbeit begleitet. Geschrieben mit dem Vorteil eines Jahrzehnts zusätzlicher Belege dafür, dass die Vorhersagen der Theorie sich in der Realität wiederfinden.

Wenn es eine Lektion in dieser Biografie gibt, dann die, zu der dieses Buch immer wieder zurückkehrt: Intelligenz ist keine feste Größe. Sie ist ein rekursiver Prozess. Wissen nährt Leistung,

Leistung ermöglicht neues Wissen, und Motivation ist der Motor, der die Schleife am Laufen hält. Meine Schleife wurde angetrieben von einer ungewöhnlich hartnäckigen Art von Neugier – der Art, die umschwenkt, wenn sie auf eine Mauer trifft, die quer über Disziplinen liest, statt in eine einzige hineinzubohren, und die nicht aufhört, nur weil niemand zuhört.

Ob die Theorie gut ist, muss jeder selbst beurteilen. Aber der Prozess, der sie hervorgebracht hat – Jahrzehnte selbstgesteuerten Lernens, getrieben von nichts als der Überzeugung, dass die Frage es wert war – ist selbst eine Demonstration von etwas, das IQ-Tests nicht messen und aktuelle KI nicht nachahmen kann: eine Art von Intelligenz, die jenseits jeder Punktzahl existiert.

Beim Lesen wird etwas auffallen: Diese Theorie stützt sich auf eine ungewöhnlich breite Palette von Feldern. Mathematik und Zelluläre Automaten. Simulations- und Modellierungstheorie. Maschinelles Lernen. Neurowissenschaft, von klinischer Neurologie bis Psychopharmakologie. Evolutionsbiologie. Philosophie des Geistes. Informatik. Die meisten Bewusstseinstheorien leben in ein oder zwei dieser Welten. Diese hier versucht, sie alle zusammenzubinden – was, wenn man darüber nachdenkt, genau das ist, was das Gehirn selbst tut. Es nimmt unterschiedliche Informationsströme aus völlig verschiedenen Quellen und webt sie zu einer einzigen kohärenten Erfahrung. Wenn eine Theorie des Bewusstseins nicht dasselbe über Disziplingrenzen hinweg leisten kann, ist das ein Grund zur Skepsis.

Kommen wir zur Theorie.

Chapter 1

Das schwierigste Problem der Wissenschaft

Dieser Satz wird gerade gelesen. Er erzeugt eine Erfahrung.

Diese Erfahrung – der visuelle Eindruck von Buchstaben auf einer Seite, die innere Stimme, die die Worte liest, das Gefühl des Verstehens oder der Verwirrung – ist das Vertrauteste im eigenen Leben und das Rätselhafteste im Universum. Wir wissen mehr über das Innere schwarzer Löcher als darüber, warum sich Lesen nach etwas anfühlt.

Das ist keine Übertreibung. (Obwohl fairerweise gesagt sei, dass die Mathematik Probleme kennt, die ich für noch schwieriger halte – aber die halten die meisten Menschen nachts nicht wach.) Physiker haben das Standardmodell. Biologen haben Evolution und Genetik. Chemiker haben das Periodensystem. Aber Bewusstsein – die Tatsache, dass es sich „irgendwie anfühlt", man selbst zu sein, gerade jetzt, beim Lesen – hat keine etablierte Theorie, keinen dominierenden Rahmen, keine allgemein akzeptierte Erklärung.

Nicht weil es nicht versucht worden wäre. Seit den 1990er Jahren, als Bewusstsein nach Jahrzehnten behavioristischen Exils wieder ein respektables wissenschaftliches Thema wurde, sind Tausende Artikel veröffentlicht, Dutzende Theorien vorgeschlagen

und Hunderte Millionen Dollar ausgegeben worden. Das Ergebnis? Ein Feld in dem Zustand, den der Wissenschaftsphilosoph Thomas Kuhn „vorparadigmatisch“ nannte – viele konkurrierende Ideen, kein Konsens und ein wachsendes Gefühl, dass etwas Fundamentales fehlen könnte.

Was das Schwierige Problem eigentlich fragt

1995 gab der Philosoph David Chalmers dem Rätsel seinen kanonischen Namen: das „Schwierige Problem“ des Bewusstseins (Hard Problem).

Die Frage geht so. Man nehme die Erfahrung, Rot zu sehen. Neurowissenschaftler können sehr viel darüber erzählen, was im Gehirn passiert, wenn Rot gesehen wird: Licht einer bestimmten Wellenlänge trifft auf die Zapfenzellen in der Netzhaut, Signale wandern entlang des Sehnervs, werden im visuellen Kortex verarbeitet, und verschiedene Hirnregionen koordinieren sich, um die Wahrnehmung zu erzeugen. All das ist gut verstanden, zumindest im Überblick.

Aber nichts davon erklärt *warum sich das Sehen von Rot nach etwas anfühlt*.

Man könnte im Prinzip ein vollständiges neuronales Modell der Gehirnreaktion auf rotes Licht bauen – jedes Neuron, jede Synapse, jeder Signalweg. Das Ergebnis wäre eine perfekte funktionale Beschreibung. Und das Gefühl der Röte wäre damit nicht erklärt. Das „Wie-es-sich-anfühlt“. Das *Quale*, wie Philosophen es nennen.

Chalmers unterschied dies von den „einfachen Problemen„ des Bewusstseins (die überhaupt nicht einfach sind, nur im Prinzip lösbar): Wie integriert das Gehirn Informationen? Wie lenkt es Aufmerksamkeit? Wie berichtet es über seine eigenen Zustände? Das sind Mechanismus-Fragen. Sie sind schwierig, aber es ist die Art von schwierig, mit der Neurowissenschaft umzugehen weiß. Das Schwierige Problem ist anders: Es fragt, warum die Mechanismen überhaupt von Erfahrung begleitet werden. Warum

verarbeitet das Gehirn nicht einfach Informationen „im Dunkeln“, wie ein Computer?

Der aktuelle Stand der Dinge

So steht es Mitte der 2020er Jahre:

Die Integrierte Informationstheorie (IIT), entwickelt von Giulio Tononi, ist die formal strengste Theorie. Sie definiert Bewusstsein als integrierte Information – eine mathematische Größe namens Φ (phi). Je höher das Φ, desto bewusster das System. IIT hat echte Stärken: Sie bietet einen mathematischen Rahmen, macht spezifische Vorhersagen darüber, welche Hirnregionen bewusst sein sollten, und nimmt die Struktur der Erfahrung ernst. Aber sie hat ein Problem: Sie impliziert, dass jedes System mit integrierter Information – einschließlich sehr einfacher Systeme wie einem Netzwerk von Logikgattern – ein gewisses Bewusstsein hat. Das ist Panpsychismus, und während einige Philosophen damit leben können, finden die meisten Wissenschaftler das zutiefst kontraintuitiv. 2023 unterzeichneten über 120 Forscher einen offenen Brief, der IIT als unfalsifizierbar und pseudowissenschaftlich bezeichnete. Die Kontroverse dauert an.

Die Theorie des Globalen Neuronalen Arbeitsraums (GNW), entwickelt von Bernard Baars und Stanislas Dehaene, konzentriert sich auf den Mechanismus, durch den Informationen bewusst werden: globale Übertragung. Wird eine Information ausgewählt und über ein Netzwerk frontoparietaler Neuronen (den „Arbeitsraum“) übertragen, wird sie bewusst; wird sie nicht übertragen, bleibt sie unbewusst. GNW ist empirisch produktiv – sie sagt spezifische neuronale Signaturen des bewussten Zugangs voraus –, aber sie weicht dem Schwierigen Problem bewusst aus. Sie erklärt, *wann* Information bewusst wird, nicht *warum* die Übertragung von Erfahrung begleitet wird.

Prädiktive Verarbeitung (PP), verbunden mit Karl Friston und Anil Seth, behandelt das Gehirn als Vorhersagemaschine.

Bewusstsein ist die „beste Vermutung„ des Gehirns über die Ursachen seiner sensorischen Eingabe. Seth nennt es eine „kontrollierte Halluzination“. PP liefert elegante Erklärungen für Wahrnehmung, Illusion und psychiatrische Störungen und ist derzeit der einflussreichste Rahmen in der computergestützten Neurowissenschaft. Aber Seth selbst räumt ein, dass PP das „reale Problem“ (Struktur und Inhalt der Erfahrung) angeht, ohne zu behaupten, das Schwierige Problem zu lösen. PP erklärt, warum man *dies* sieht und nicht *das*, aber nicht, warum Sehen sich überhaupt nach etwas anfühlt.

Es gibt weitere – Theorien Höherer Ordnung, Attention Schema Theory, Recurrent Processing Theory, Elektromagnetische Feldtheorien – jede mit echten Einsichten und echten Lücken. 2025 veröffentlichte die COGITATE adversariale Zusammenarbeit, angelegt um IIT gegen GNW zu testen, ihre Ergebnisse in *Nature*. Das Ergebnis? Keine der beiden Theorien wurde vollständig bestätigt. Der posteriore Kortex zeigte die stärkste bewusstseinsbezogene Aktivität – was nicht ganz das war, was beide Lager vorhergesagt hatten. Nach Jahrzehnten und Hunderten Millionen Dollar ist das Feld wohl weiter vom Konsens entfernt als zu Beginn.

Zwei Dogmen, die den Fortschritt blockieren

Bevor ich sage, was meiner Meinung nach fehlt, muss ich zwei Vorurteile benennen, die das Feld seit Jahrzehnten still sabotieren. Ich habe ihnen in meinem ursprünglichen Buch Namen gegeben, weil unbenannte Vorurteile schwerer zu bekämpfen sind.

Das erste nenne ich das **nSKI-Dogma** – „keine starke Künstliche Intelligenz“. Die weit verbreitete Überzeugung, dass wirklich intelligente Maschinen unmöglich sind – eine Überzeugung, die nicht auf Beweisen beruht, sondern auf dem Scheitern der frühen KI-Forschung in den 1960er Jahren und der daraus resultierenden Gegenreaktion. Wer glaubt, dass starke KI möglich ist, lernt schnell, darüber zu schweigen, wenn er in der Mainstream-Forschung

ernst genommen werden will. Das ist kein rationaler Skeptizismus. Es ist eine Narbe alter Niederlagen, verhärtet zur Doktrin.

Das zweite sitzt tiefer und richtet mehr Schaden an. Ich nenne es das **nSV-Dogma** – „kein Selbstverständnis“. Der Glaube, dass der menschliche Geist, das menschliche Bewusstsein, im Prinzip nicht von eben diesem Geist verstanden werden kann. Man beruft sich auf Gödels Unvollständigkeitssätze oder vage Analogien zu den Grenzen kosmologischer Beobachtung von innerhalb des Universums, oder – am ehrlichsten – man findet die Aussicht, vollständig erklärt zu werden, einfach zu erschreckend, um sie zu erwägen. Wenn Bewusstsein nur eine Maschine ist, was wird dann aus der Seele? Was aus der Bedeutung? Was aus dem Besonderen des Menschseins?

Diese Dogmen verstärken sich gegenseitig. Wenn sich Bewusstsein nicht verstehen lässt (nSV), dann lässt sich sicherlich auch keines bauen (nSKI). Und wenn sich keines bauen lässt (nSKI), dann liegt Bewusstsein vielleicht wirklich jenseits des Verstehens (nSV). Ein geschlossener Kreislauf institutionellen Pessimismus, der eine enorme Zahl intelligenter Forscher davon abgehalten hat, die Arbeit überhaupt zu versuchen.

Damit ist nicht gesagt, dass diese Dogmen in böser Absicht vertreten werden. Viele Forscher glauben sie aufrichtig. Aber keines der beiden wurde je bewiesen. Es sind Glaubensartikel, und sie haben der Bewusstseinsforschung mehr Schaden zugefügt als jedes gescheiterte Experiment.

Etwas fehlt

Ich denke, der Grund, warum keine Theorie das Schwierige Problem geknackt hat, ist, dass die meisten an der falschen Stelle nach Bewusstsein suchen. Sie schauen auf die neuronale Maschinerie – die Neuronen, die Synapsen, die Oszillationen, die Konnektivität – und fragen: „Welcher dieser Prozesse ist bewusst?“

Die richtige Frage ist eine andere: „Auf welcher Ebene der Informationsverarbeitung und in welcher Architektur entsteht Erfahrung?“

Das ist der Ausgangspunkt der Vier-Modelle-Theorie (VMT). Sie beginnt mit der Beobachtung, dass kein Mensch jemals in seinem Leben die Realität direkt erfahren hat. Der einfachste Beweis: In jedem Auge gibt es einen blinden Fleck – eine Region der Netzhaut ganz ohne Photorezeptoren, wo der Sehnerv austritt –, aber man sieht kein Loch. Das Gehirn füllt die Stelle mit fabriziertem Inhalt. Wäre Wahrnehmung direkter Zugang zur Realität, müssten zwei dunkle Flecken sichtbar sein. Sind sie aber nicht, weil der Blick auf ein Modell fällt. Diese Behauptung ist übrigens nicht kontrovers – dass Wahrnehmung konstruktiv statt direkt ist, ist Mainstream-Neurowissenschaft, akzeptiert von praktisch jedem Forscher auf dem Gebiet. Was wir erfahren, ist eine Simulation der Realität, erzeugt vom Gehirn, so nahtlos, dass man den Unterschied nie bemerkt. Und die Theorie argumentiert, dass diese Beobachtung, konsequent zu Ende gedacht, das Schwierige Problem auflöst.

Drei leitende Prinzipien

Bevor wir zur Theorie selbst kommen, müssen drei philosophische Prinzipien dargelegt werden – diejenigen, die ihre Konstruktion geleitet haben. Das sind keine willkürlichen methodischen Entscheidungen. Es sind Bedingungen, die jede ernsthafte wissenschaftliche Theorie erfüllen sollte. Bedingungen, die viele Bewusstseinstheorien entweder ignorieren oder verletzen.

Ockhams Rasiermesser. Die einfachste Erklärung, die zu den Fakten passt, ist in der Regel die richtige. Das grundlegende Prinzip der Wissenschaft, zugeschrieben dem Philosophen Wilhelm von Ockham aus dem 14. Jahrhundert. Wenn zwei Theorien dieselben Phänomene erklären, bevorzuge die mit weniger Entitäten, weniger Annahmen, weniger Spezialfällen. Ockhams

Rasiermesser garantiert keine Wahrheit, aber es hat eine bemerkenswerte Erfolgsbilanz: Newton brauchte keine Engel, die die Planeten schubsen; Darwin brauchte keinen Designer, der die Arten formt; Einstein brauchte keinen Lichtäther. Das Universum scheint Einfachheit zu bevorzugen.

Die Vier-Modelle-Theorie ist durch und durch ockhams. Sie führt keine neuen physikalischen Phänomene ein – keine Quanteneffekte in Mikrotubuli, keine exotischen Feldtheorien, kein panpsychistisches „Proto-Bewusstsein", das in der Materie verstreut liegt. Sie verwendet nur, was wir bereits kennen: neuronale Netzwerke, Lernen, Simulation, Selbstreferenz. Die Komplexität liegt in der *Architektur*, nicht im Hinzufügen mysteriöser neuer Zutaten.

Das Kopernikanische Prinzip. Wir sind nicht besonders. Benannt nach Kopernikus, der die Erde aus dem Zentrum des Kosmos verdrängte, wurde dieses Prinzip immer weiter ausgedehnt: Die Sonne ist nicht besonders, unsere Galaxie ist nicht besonders, und – am unbequemsten für viele – *wir* sind nicht besonders. Bewusstsein ist kein einzigartiges Wunder, kein einmaliger göttlicher Funke, kein emergentes Phänomen so selten, dass es nur einmal passieren konnte. Wenn ein System es hat, können andere Systeme es auch haben – sofern sie die richtige Architektur besitzen. Das ist die anti-exzeptionalistische Haltung, die künstliches Bewusstsein möglich macht.

Das Kopernikanische Prinzip ist auch der Grund, warum diese Theorie Bewusstsein bei Tieren vorhersagt. Wenn eine Gehirnarchitektur Bewusstsein erzeugen kann, sollte jede hinreichend ähnliche Architektur es ebenfalls erzeugen. Menschen sind nicht magisch. Wir sind nur eine Implementierung eines allgemeinen Rechenprinzips.

Leibniz' Gesetz (Die Identität des Ununterscheidbaren). Wenn zwei Dinge in all ihren Eigenschaften wirklich identisch sind, sind sie dasselbe Ding. Dieses Prinzip, formuliert vom Philosophen Gottfried Wilhelm Leibniz im 17. Jahrhundert, ist so einfach wie

tiefgründig. Es schließt „Zombie-Welten„ aus – hypothetische Universen, die physikalisch identisch mit unserem sind, in denen aber niemand bewusste Erfahrung hat. Wenn ein System in jeder funktionalen, strukturellen und verhaltensbezogenen Eigenschaft identisch mit einem bewussten System ist, dann *ist* es ein bewusstes System. Es gibt keine zusätzliche „Bewusstseinssubstanz“, die vorhanden oder abwesend sein könnte, während alles andere gleich bleibt. Bewusstsein ist kein optionaler Zusatz zu einer ansonsten vollständigen funktionalen Beschreibung. Es ist Teil der Beschreibung.

Leibniz' Gesetz ist der Grund, warum philosophische Zombies – Wesen, die genau wie bewusste Menschen handeln, aber nicht bewusst sind – inkohärent sind. Wenn der Zombie funktional identisch mit einem bewussten Wesen ist, dann hat er dieselbe Vier-Modelle-Architektur, dieselbe laufende Simulation, dieselbe Selbstreferenz. Was soll an diesem Punkt „nicht bewusst sein“ überhaupt noch bedeuten? Die Frage löst sich auf.

Diese drei Prinzipien – Einfachheit, Nicht-Exzeptionalismus und Identität durch Eigenschaften – sind nicht bloß ästhetische Präferenzen. Sie sind die intellektuellen Werkzeuge, mit denen sich durch Jahrhunderte der Verwirrung hindurchschneiden lässt, um zu einer Theorie zu gelangen, die tatsächlich funktioniert. Die Vier-Modelle-Theorie ist das Ergebnis, wenn man diese Prinzipien ernst nimmt und auf das schwierigste Problem der Wissenschaft anwendet.

Jetzt ist es Zeit, die vier Modelle zu betrachten.

Chapter 2

Die vier Modelle

Stellen wir uns vor, wir schauen auf einen Apfel.

Der Apfel liegt auf einem Tisch. Rot, rund, glänzend, etwa fünfzehn Zentimeter von der eigenen Hand entfernt. Man kann ihn sehen, weiß, was er ist, könnte die Hand ausstrecken und ihn greifen. Scheint unkompliziert – man sieht einen Apfel.

Aber was tatsächlich passiert, ist um Größenordnungen komplizierter.

Licht, das von der Oberfläche des Apfels reflektiert wird, tritt in die Augen ein und trifft auf die Photorezeptorzellen der Netzhäute. Diese Zellen wandeln das Licht in elektrische Signale um. Die Signale wandern entlang der Sehnerven zum visuellen Kortex im hinteren Teil des Gehirns, wo sie durch eine Hierarchie immer ausgefeilterer Merkmalsdetektoren verarbeitet werden: Kanten, Orientierungen, Farben, Texturen, Formen und schließlich Objekte. Irgendwo in dieser Kaskade feuert die neuronale Aktivität, die „Apfel“ entspricht. Gleichzeitig bereitet das motorische System potenzielle Aktionen vor (Greifen, Fassen), das Gedächtnissystem aktiviert Assoziationen (Geschmack, Textur, das letzte Mal, als man einen Apfel gegessen hat), und das räumliche System verfolgt die Position des Apfels relativ zum eigenen Körper.

All das passiert in weniger als einer Sekunde. Und nichts davon ist, was man *erfährt*. Man erfährt nicht Photonen, die auf

Zapfenzellen treffen, nicht Signale, die Axone entlangwandern, nicht Merkmalsdetektoren, die feuern. Man erfährt *einen Apfel*. Ein einheitliches, stabiles, dreidimensionales Objekt in einer kohärenten räumlichen Umgebung, mit einem bestimmten Aussehen, Gefühl und Bedeutung. Was wir erfahren, ist ein *Modell* – eine Echtzeit-Simulation des Apfels, erzeugt vom Gehirn aus den Rohdaten und allem, was es zuvor über Äpfel, Objekte, Tische und Physik gelernt hat.

Wie in Kapitel 1 dargelegt, ist das unkontroverse Neurowissenschaft. Jeder Neurowissenschaftler und Wahrnehmungsphilosoph stimmt zu, dass das, was wir erfahren, ein Modell ist, nicht die Realität selbst. Der Apfel, den man sieht, ist die *beste Vermutung* des Gehirns darüber, was da draußen ist – gestützt auf die Sinnesdaten, aber nicht identisch damit. (Optische Täuschungen sind der lebende Beweis: Wenn eine Illusion zusammenbricht – wenn man sie plötzlich auf beide Weisen sieht – ertappt man die Simulation auf frischer Tat. Die Realität wurde nie direkt gesehen. Es war immer das Modell. Die Illusion hat es nur sichtbar gemacht.)

Aber hier beginnt meine Theorie: Das Gehirn modelliert nicht nur den Apfel. Es modelliert *einen selbst beim Anschauen des Apfels*. Und es ist dieses zweite Modell – das Modell des Selbst –, das Informationsverarbeitung in Bewusstsein verwandelt.

Die vier Repräsentationen des Gehirns

Schon einfache neuronale Netzwerke mit nur drei Schichten können lernen, ihre Eingabe zu modellieren – zeigt man ihnen genug Beispiele, bauen sie interne Repräsentationen der Muster auf, denen sie begegnen. Das Gehirn macht genau das, aber in einem weitaus größeren Maßstab. Es baut nicht ein Modell; es baut viele – vom Sehfeld bis zur Position der Gliedmaßen, vom Klang einer Stimme bis zum Druck der Füße auf dem Boden. Diese Modelle umfassen sowohl die Welt außerhalb *als auch* den eigenen

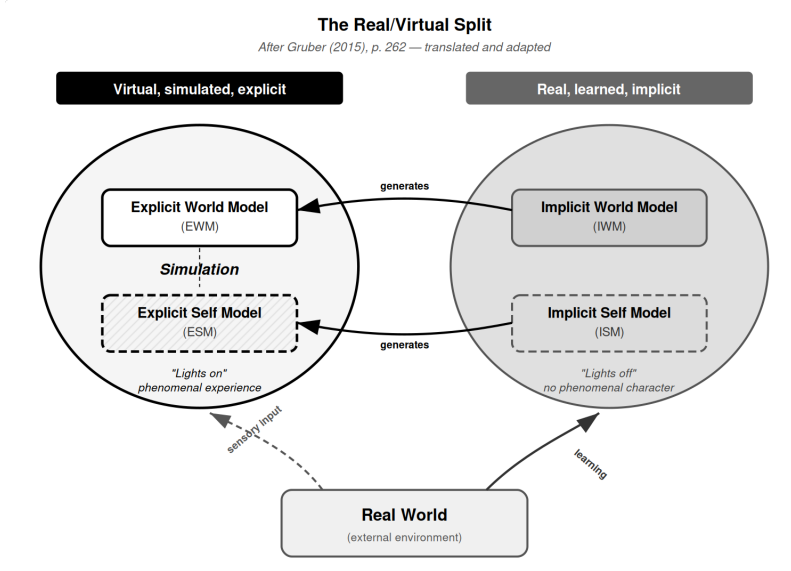

Figure 2.1: Die Real/Virtuell-Trennung. Das Substrat (reale Seite) speichert Wissen in synaptischen Gewichten – physikalisch, strukturell, unbewusst. Die Simulation (virtuelle Seite) erzeugt Erfahrung in Echtzeit – flüchtig, dynamisch, bewusst.

Körper und verknüpfen alle verfügbaren sensorischen Eingaben zu kohärenten Repräsentationen.

Die Neurowissenschaft kennt diese Modelle seit über einem Jahrhundert. Im motorischen und im somatosensorischen Kortex ist der Körper buchstäblich als verzerrte Karte dargestellt – Hände und Lippen grotesk vergrößert, weil sie mehr Nervenenden haben, Rumpf und Beine auf Splitter zusammengestaucht. Diese kortikalen Karten, *Homunculi* genannt, wurden erstmals von Wilder Penfield in den 1930er Jahren durch direkte elektrische Stimulation während Hirnoperationen kartiert. Sie sind nur die anschaulichsten Beispiele; das Gehirn unterhält ähnliche Karten und Modelle in seiner gesamten Architektur. (Siehe Anhang A für mehr zur kortikalen Organisation.)

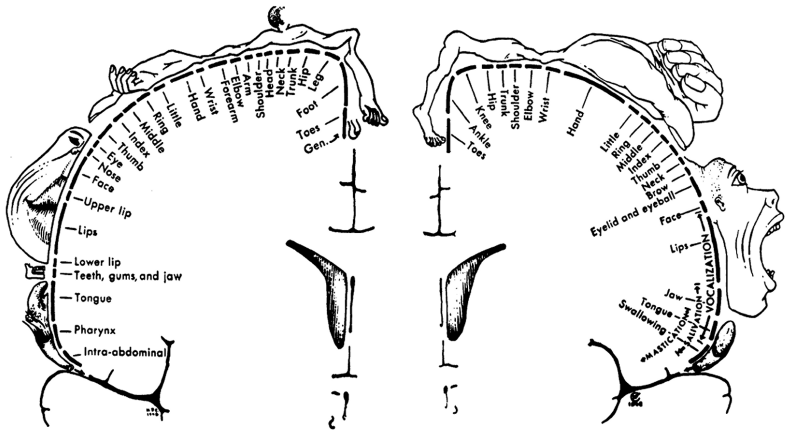

Figure 2.2: Penfields kortikaler Homunculus. Der somatosensorische Kortex widmet Händen, Lippen und Zunge dramatisch mehr Fläche als dem gesamten Rumpf – eine verzerrte Körperkarte, die die Dichte der Nervenenden widerspiegelt, nicht die Größe der Körperteile.

Ich nenne diese die **impliziten Modelle**: das Implizite Weltmodell (IWM) und das Implizite Selbstmodell (ISM). Sie stecken in der Struktur des Gehirns – in den Stärken synaptischer Verbindungen, der Architektur neuronaler Schaltkreise, dem angesammelten

Lernen eines Lebens. Sie sind die Festplatte des Gehirns. Man erfährt sie nie direkt, genauso wenig wie das Silizium im eigenen Handy. Aber sie kodieren alles, was das Gehirn über die Welt und über einen selbst weiß.

Jetzt die Schlüsseleinsicht. Diese impliziten Modelle sitzen nicht einfach nur da. Sie *erzeugen* etwas. In der Technik ist ein **digitaler Zwilling** eine virtuelle Echtzeitkopie eines physischen Systems – ein Düsentriebwerk, ein Stromnetz, eine Fabrikhalle –, kontinuierlich mit Sensordaten aktualisiert, damit Ingenieure das System überwachen und mit ihm interagieren können, ohne es direkt anzufassen. Die impliziten Modelle tun genau das. Sie produzieren eine virtuelle Echtzeit-Simulation der Welt und eine virtuelle Echtzeit-Simulation von einem selbst. Das sind die **expliziten Modelle**: das Explizite Weltmodell (EWM) und das Explizite Selbstmodell (ESM). Alles, was man sieht, hört, fühlt und denkt, geschieht innerhalb dieser Simulationen, nicht in der Welt selbst.

Zwei Gruppen von Modellen (implizit und explizit), jede mit einem Weltmodell und einem Selbstmodell. Vier Modelle insgesamt – und damit eine Sprache, um darüber zu sprechen, was Bewusstsein tatsächlich tut. (Eine Anmerkung für Neurowissenschaftler und technisch versierte Leser: Die Zahl „vier“ ist ein prinzipielles Minimum, keine buchstäbliche Zählung dessen, was das Gehirn unterhält. Wen das irritiert, der lese bitte Anhang E, bevor es weitergeht – er geht direkt darauf ein.)

Aber wo laufen diese Modelle? Das Gehirn nutzt mindestens fünf Ebenen der Informationsverarbeitung, übereinander geschichtet. Die Simulation – die bewusste Erfahrung – läuft ganz oben.

Fünf verschachtelte Systeme

Man kann sich das Gehirn als fünf Organisationsebenen vorstellen, ineinander geschachtelt wie russische Puppen:

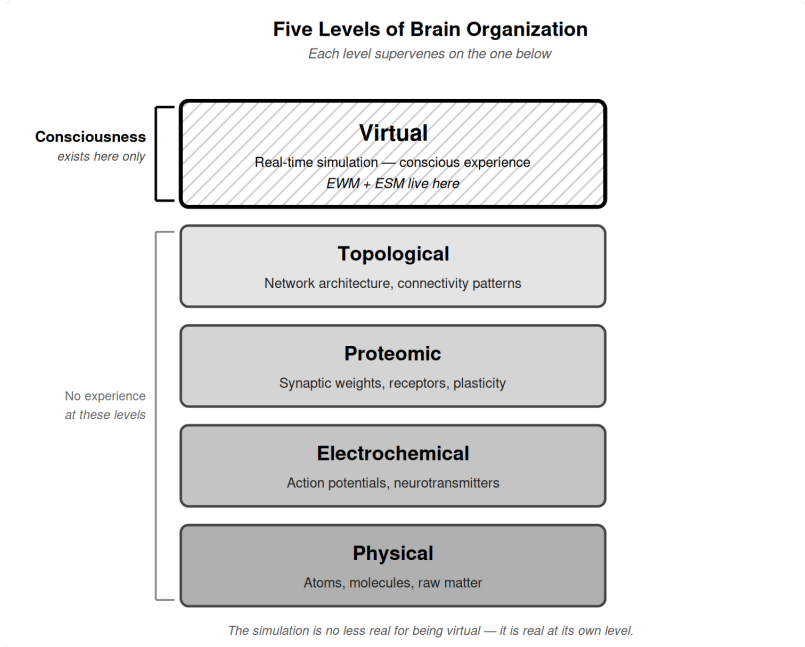

Figure 2.3: Fünf Ebenen der Gehirnorganisation. Jede Ebene superveniert auf („läuft auf“) der darunter liegenden. Bewusstsein existiert nur auf der obersten virtuellen Ebene, wo die expliziten Modelle phänomenale Erfahrung erzeugen.

Physikalisch. Ganz unten die rohe Materie: Atome, Moleküle, das physische Substrat des Gehirns. Das ist Chemie – Kohlenstoff, Wasserstoff, Stickstoff, Sauerstoff, aus denen das Gewebe besteht. Inerte Materie, die den Gesetzen der Thermodynamik gehorcht. Nichts Bewusstes lebt hier.

Elektrochemisch. Eine Ebene höher: neuronale Signalübertragung. Aktionspotenziale rasen Axone hinunter, Neurotransmitter fluten Synapsen, Ionen fließen durch Kanäle. Das ist die elektrische und chemische Aktivität, die man sich vorstellt, wenn man an „das Gehirn tut etwas“ denkt. Die Ebene, auf der Neuronen feuern. Immer noch keine Erfahrung, aber jetzt gibt es Informationsübertragung.

Proteomisch. Darüber: Proteinstrukturen und molekulare Maschinerie. Hier werden synaptische Gewichte gespeichert – die physischen Stärken der Verbindungen zwischen Neuronen. Rezeptoren auf Zellmembranen, Enzyme, die Plastizität regulieren, das molekulare Gerüst, das bestimmt, welche Synapsen stärker werden und welche schwächer. Das ist die „Hardware“ des Lernens. Wer eine Fähigkeit übt und besser wird, verändert die proteomische Schicht. Immer noch unbewusst, aber jetzt gibt es Gedächtnis.

Topologisch. Noch höher: Netzwerkarchitektur. Die Muster der Konnektivität – welche Neuronen mit welchen verbunden sind, wie dicht, in welchen Konfigurationen. Hier leben Brodmann-Areale, hier leben kortikale Säulen, hier existiert die großräumige Struktur von „visueller Kortex spricht mit motorischem Kortex“. Der Schaltplan. Ändert man diese Ebene, ändert sich, welche Arten von Verarbeitung das System leisten kann. Hier sind die impliziten Modelle (IWM und ISM) gespeichert. Immer noch unbewusst. Aber jetzt gibt es Wissen.

Virtuell. Ganz oben: die simulierte Welt. Der kortikale Automat – das dynamische Muster elektrischer Aktivität, das über das Netzwerk tanzt, Informationen integriert, Vorhersagen erzeugt, die Modelle in Echtzeit laufen lässt. Hier lebt die bewusste Erfahrung. Die expliziten Modelle (EWM und ESM) existieren

hier und nur hier. Das ist die einzige Ebene, die sich nach etwas anfühlt.

Jede Ebene superveniert auf der darunterliegenden, hat aber ihre eigene Dynamik. Ohne physische Materie keine elektrochemische Signalübertragung, ohne Chemie keine Proteinstrukturen, ohne Synapsen keine Netzwerktopologie, und ohne Netzwerk keine Simulation, die darauf laufen könnte. Aber jede Ebene hat Eigenschaften, die den niedrigeren fehlen. Eine Synapse ist nicht „über„ irgendetwas – sie ist nur eine Verbindung. Ein Netzwerk von Synapsen *ist* über etwas: Es repräsentiert ein Gesicht, ein Wort, eine Erinnerung. Und die Simulation, die auf diesem Netzwerk läuft? Da wird aus „über“ plötzlich „Erfahrung“.

Diese Fünf-Ebenen-Hierarchie löst ein Problem, über das fast jeder stolpert, der diese Theorie zum ersten Mal hört: „Wenn Bewusstsein virtuell ist, worauf läuft es?„ Die Antwort: Es läuft auf der topologischen Ebene (dem Netzwerk), die in der proteomischen Ebene (synaptische Gewichte) implementiert ist, die auf der elektrochemischen Ebene (neuronales Feuern) aufsitzt, die in der physischen Ebene (Materie) existiert. Bewusstsein ist nicht weniger real, weil es virtuell ist – es ist nur real *auf einer anderen Ebene*, als Neuronen real sind. Der Berg im Videospiel ist real auf der Spielebene, auch wenn er „nur“ Transistoren auf der Hardware-Ebene ist. Dasselbe Prinzip.

Diese Hierarchie wird im Laufe des Buches immer wieder auftauchen, besonders wenn es in Kapitel 6 um Psychedelika geht – weil Drogen nicht alle fünf Ebenen gleich treffen. Einige zielen auf die elektrochemische Ebene (Veränderung der Neurotransmitter-Dynamik), andere auf die proteomische (Veränderung der Rezeptorexpression), und die Wirkungen pflanzen sich auf vorhersagbare Weise zur virtuellen Ebene hoch fort. Die Hierarchie ist nicht nur konzeptuell. Sie ist mechanistisch real und leistet Erklärungsarbeit.

Nun zu den vier Modellen.

Das Implizite Weltmodell (IWM) ist alles, was man über die Welt weiß. Nicht das, woran man gerade denkt – alles, woran man

denken könnte. Die Gesetze der Physik (fallende Objekte fallen, das weiß man). Das Layout der eigenen Wohnung (im Dunkeln lässt sich darin navigieren). Die Grammatik der Muttersprache (ein Satz lässt sich als grammatisch oder ungrammatisch beurteilen, ohne die Regeln zu kennen). Die Gesichter aller, die man je gekannt hat. Der Geschmack von Schokolade. Das Geräusch von Regen.

All dieses Wissen steckt in den synaptischen Verbindungen des Gehirns – den Stärken der Verknüpfungen zwischen Neuronen. Es wurde über ein ganzes Leben durch Erfahrung und Lernen aufgebaut. Und es ist nie, niemals direkt bewusst. Man kann nicht in die eigenen neuronalen Verbindungen hineinschauen. Man kann die eigenen Synapsen nicht fühlen. Das Implizite Weltmodell ist wie eine riesige Bibliothek, die man nie betritt – man liest nur die Bücher, die sie einem an den Schreibtisch schickt.

Das Implizite Selbstmodell (ISM) ist alles, was man über sich selbst weiß. Das Körperschema – die unbewusste Repräsentation davon, wo die Gliedmaßen sind, wie groß sie sind, wie sie sich bewegen. Die motorischen Fähigkeiten: Fahrradfahren, Tippen, ein Instrument spielen. Die Persönlichkeitsmerkmale, sozialen Fähigkeiten, emotionalen Muster, Gewohnheiten. Die autobiographische Gedächtnisstruktur – das Gerüst, das die Erinnerungen zu einer Lebensgeschichte ordnet.

Wie das Weltmodell ist das Selbstmodell in synaptischen Gewichten gespeichert und nie direkt bewusst. Man erfährt nicht das Körperschema; man erfährt den Körper, den das Schema erzeugt. Man erfährt nicht die Persönlichkeit; man erfährt die Gedanken und Gefühle, die sie hervorbringt. Das Implizite Selbstmodell ist die Crew hinter der Bühne – unverzichtbar für die Aufführung, aber nie vom Publikum gesehen.

Das Explizite Weltmodell (EWM) ist die Welt, die man tatsächlich erfährt. Gerade jetzt. Der Raum, in dem man sitzt, die Geräusche, die man hört, das Gewicht dieses Buches in den Händen (oder das Leuchten des Bildschirms, auf dem man es liest). Das ist die Simulation – die virtuelle Echtzeit-Realität des Gehirns,

erzeugt aus dem Impliziten Weltmodell plus aktueller sensorischer Eingabe. Sie ist lebendig, detailliert und nahtlos überzeugend. Man lebt sein ganzes Leben darin und tritt nie hinaus.

Das Explizite Selbstmodell (ESM) ist das *Selbst*. Das Gefühl, ein Subjekt zu sein. Das Gefühl von „Ich" – derjenige, der sieht, hört, denkt und entscheidet. Auch das ist eine Simulation: ein Echtzeit-Modell, erzeugt aus dem Impliziten Selbstmodell plus aktuellen Körpersignalen. Es ist die Figur, die das Gehirn erschafft, um seine virtuelle Welt zu bewohnen. Man IST die Figur, die das Gehirn erschafft, um seine virtuelle Welt zu bewohnen.

Die reale Seite und die virtuelle Seite

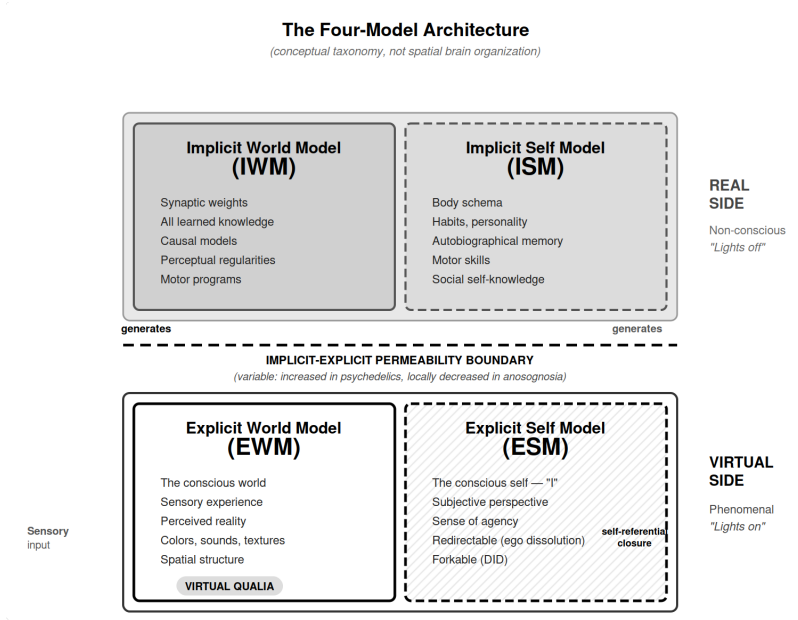

Figure 2.4: Die Vier-Modelle-Architektur. Das Gehirn unterhält zwei Arten von Modell (eines von der Welt, eines vom Selbst), jedes in zwei Modi: implizit (in der Struktur des Gehirns gespeichert) und explizit (aktiv laufend als Echtzeit-Simulation). Bewusstsein lebt in den expliziten Modellen.

Die vier Modelle teilen sich in zwei Seiten, und diese Teilung ist das Fundament für alles, was folgt.

Die **reale Seite** (die zwei impliziten Modelle) ist physisch, strukturell und relativ starr (langsam durch Lernen angepasst). Sie ist das gespeicherte Wissen des Gehirns: synaptische Gewichte, Netzwerkverbindungen, Rezeptorkonfigurationen. Man stelle sich darunter alles vor, was das Gehirn *gelernt hat* – eingebrannt in die physische Struktur des Gewebes selbst. Erfahrung hat sie keine. Das Feuern einer Synapse wird nicht mehr „erlebt" als Wasser, das durch ein Rohr fließt. Die reale Seite ist Licht aus.

Das ist es wert, betont zu werden: Die reale Seite ist das, was die Neurowissenschaft bereits untersucht. Wenn ein Forscher jemanden in einen fMRT-Scanner schiebt, schaut er auf die reale Seite – Aktivierungsmuster, Konnektivität, Blutfluss in verschiedenen Regionen. Wenn ein Neurochirurg ein Kortexareal stimuliert und beobachtet, was passiert, tastet er die reale Seite ab. Die Neurowissenschaft kartiert dieses Gebiet seit über einem Jahrhundert, und sie hat Außerordentliches geleistet. Die Vier-Modelle-Theorie verwirft nichts davon. Sie sagt nur, dass all das erst die Hälfte des Bildes beschreibt.

Die **virtuelle Seite** (die zwei expliziten Modelle) ist simuliert, flüchtig und dynamisch. Sie wird in jedem Moment neu erzeugt, aus der realen Seite plus aktuellem Eingang. Man stelle sich darunter alles vor, was das Gehirn *gerade mit* dem anstellt, was es gelernt hat – die Live-Show, nicht das abgelegte Drehbuch. Und sie ist *durch und durch* Erfahrung. Jeder Anblick, Klang, Gedanke, jedes Gefühl, jede Erinnerung, jeder Traum und jede Halluzination, die man je hatte, spielte sich auf der virtuellen Seite ab. Die virtuelle Seite ist Licht an.

Aber hier ist der Haken: Die virtuelle Seite ist von außen unsichtbar. Selbst die fortschrittlichste Hirnbildgebung erfasst sie nur indirekt. Ein fMRT zeigt, welche Hirnregionen aktiv sind – das ist die reale Seite bei der Arbeit. Um bewusste Erfahrung tatsächlich aus Gehirndaten *abzulesen*, müsste man

die Sprache des Gehirns entschlüsseln – nicht nur verstehen, welche Neuronen feuern, sondern was das Feuermuster auf der Simulationsebene *bedeutet*. Dafür bräuchte man so etwas wie ein vollständig simuliertes Konnektom: eine komplette digitale Kopie des Gehirns, laufend in Software, die dieselbe virtuelle Welt hervorbringt wie das biologische Gehirn.

Ich will ehrlich sein, was die Theorie leistet und was nicht. Die Vier-Modelle-Theorie sagt *was* die Simulation ist, *wo* sie lebt, und *warum* sie sich nach etwas anfühlt. Den Entschlüsselungsring liefert sie nicht mit. Die virtuelle Seite aus der realen Seite abzulesen ist ein zukünftiges Forschungsprogramm – eines, das die Theorie klar umreißt, aber noch nicht einlösen kann. Allerdings wird die Grundlage dafür bereits gelegt. Das Human Connectome Project und verwandte Vorhaben kartieren die Verdrahtung des Gehirns in immer feinerer Auflösung. Die virtuelle Seite lässt sich noch nicht aus Strukturdaten dekodieren, aber die Strukturdaten kommen.

Wer wissenschaftlich denkt, sieht vielleicht schon, wohin das führt. Wenn Erfahrung nur auf der virtuellen Seite existiert, dann sucht man auf der realen Seite – in den Neuronen, den Synapsen, der physischen Maschinerie – an der völlig falschen Stelle danach. Das wäre, als suchte man die Handlung eines Films in den Schaltkreisen des DVD-Players.

Das ist der Schlüssel. Hier ist er.

Wie bewusst ist man?

Aber zuerst etwas, worüber man sich vermutlich schon Gedanken gemacht hat. Wenn Bewusstsein eine Simulation ist – ein virtuelles Selbst innerhalb einer virtuellen Welt – dann ist es keine Alles-oder-Nichts-Angelegenheit, oder? Eine Simulation kann mehr oder weniger detailliert sein. Ein Selbstmodell kann mehr oder weniger ausgereift sein. Das heißt, Bewusstsein kommt in *Graden*.

Die Vier-Modelle-Theorie bietet eine präzise Handhabe, um über diese Grade nachzudenken. Es gibt vier abgestufte Ebenen, und jedes bewusste Wesen sitzt irgendwo auf dieser Leiter.

Ganz unten steht **einfaches Bewusstsein**. Das ist ein Explizites Weltmodell mit nur einem rudimentären Expliziten Selbstmodell. Das System erzeugt eine virtuelle Welt – es ist etwas *wie etwas*, dieses Wesen zu sein – aber das Selbst in dieser Welt ist kaum angedeutet. Man denke an eine Maus im Labyrinth. Sie sieht die Wände, riecht den Käse, fühlt den Boden unter ihren Pfoten. Sie hat phänomenale Erfahrung. Aber ihr Modell von *sich selbst* als dasjenige, das diese Erfahrungen hat? Hauchdünn. Es gibt ein „wie es ist wie„, aber fast kein „für wen es ist wie“.

Eine Stufe höher: **einfach erweitertes Bewusstsein**. Jetzt wird das Selbstmodell ernst. Das System erfährt nicht nur – es modelliert sich selbst *als* den Erfahrenden. Es weiß, dass es erfährt. Der Hund fühlt nicht nur Schmerz; er weiß, dass *er* Schmerzen hat. Es gibt eine Ich-Perspektive – ein echtes „Ich“ im Zentrum der virtuellen Welt. Das ist Introspektion erster Ordnung, und sie ändert alles. Leiden wird hier möglich, weil Leiden ein Selbst erfordert, das weiß, dass es leidet.

Dann: **doppelt erweitertes Bewusstsein**. Introspektion zweiter Ordnung. Das System modelliert sich dabei, sich selbst zu modellieren. Das ist Metakognition – Denken über das eigene Denken. Man liegt im Bett und fragt sich, ob die Angst vor dem morgigen Meeting berechtigt ist oder ob man katastrophisiert. Die eigenen mentalen Zustände werden beobachtet, bewertet, manchmal überschrieben. Hier lebt das meiste erwachsene menschliche Bewusstsein die meiste Zeit. Es ist die Ebene, die Therapie möglich macht – die es erlaubt zu sagen „Ich merke, dass ich wütend werde“ statt einfach nur wütend zu sein.

Und ganz oben: **dreifach erweitertes Bewusstsein**. Dritte Ordnung. Das System modelliert sich dabei, sich dabei zu modellieren, sich selbst zu modellieren. Das klingt wie ein Spiegelsaal, und das ist es auch – aber ein Spiegelsaal, den man

braucht, um Philosophie des Geistes zu betreiben. Um zu fragen „Was ist Bewusstsein?“ muss man sich selbst modellieren, die eigene Erfahrung modellieren und dann sich selbst modellieren, wie man diese Erfahrung modelliert. Man muss weit genug zurücktreten, um den gesamten Apparat von außen zu sehen, obwohl man immer noch darin steckt. Das ist die Voraussetzung für die Frage, die dieses Buch zu beantworten versucht. Nur Wesen, die zu dreifach erweitertem Bewusstsein fähig sind, können sich fragen, warum sich irgendetwas nach irgendetwas anfühlt.

Was bringt das? Dieser Gradient ist nicht nur abstrakte Philosophie. Er beantwortet die Frage, die bei jeder Dinnerparty gestellt wird – „Ist mein Hund bewusst?“ Die Antwort ist ja, aber weniger bewusst als wir selbst. Der Hund ist wahrscheinlich auf der einfach erweiterten Ebene. Er hat ein Selbst. Er hat Erfahrung. Er liegt nicht um 3 Uhr morgens wach und hinterfragt die Natur dieser Erfahrung. In Kapitel 10 kommen wir im Detail auf die Tierfrage zurück, wo dieser Gradient echte Erklärungsarbeit leistet. Aber die Grundform ist schon erkennbar: Bewusstsein ist kein Lichtschalter. Es ist ein Dimmer.

Warum das Gehirn die Fähigkeit zur Selbstmodellierung hat

Soweit also klar: Bewusstsein hängt von diesen vier Modellen ab, wobei das Explizite Selbstmodell die Hauptlast trägt. Aber warum hat das menschliche Gehirn diese Fähigkeit überhaupt, wenn einfachere Tiere sie nicht haben? Die Antwort liegt offen zutage: Die Architektur des menschlichen Kortex ist schlicht überdimensioniert für bloße Informationsverarbeitung.

Der menschliche Neokortex hat sechs Schichten. Das ist eine wohlbekannte anatomische Tatsache, nachzulesen in jedem Neurobiologie-Lehrbuch. Aber das Interessante daran: Sechs Schichten braucht man nicht, um Information zu verarbeiten. Drei reichen.

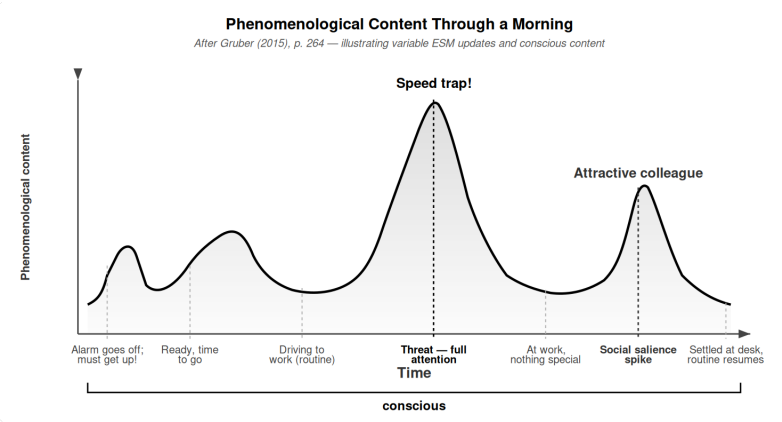

Figure 2.5: Phänomenologischer Gehalt im Tagesverlauf. Routinetätigkeiten führen zu niedrigem phänomenologischem Gehalt (Autopilot). Auffällige Ereignisse (Bedrohungen, soziale Signale) erzeugen hohen Gehalt. Bewusstsein verfolgt, was wichtig ist, nicht alles.

Man überlege, was ein neuronales Netzwerk im Kern leisten muss. Erste Schicht: Eingabe empfangen, filtern, bereinigen. Zweite Schicht: Muster extrahieren, Merkmale erkennen, die schwere Rechenarbeit erledigen. Dritte Schicht: Ergebnisse zusammenführen, Entscheidungen treffen, Ausgabe erzeugen. Eingabe, Verarbeitung, Ausgabe. Das ist das Grundrezept, und drei Schichten decken es ab.

Aber wir haben sechs.

Wozu die „zusätzlichen“ drei Schichten?

Sie sind dafür da, die ersten drei zu modellieren.

Ein Drei-Schichten-Netzwerk verarbeitet die Welt. Ein Sechs-Schichten-Netzwerk verarbeitet die Welt *und* schaut sich dabei zu. Die zusätzlichen Schichten liefern dem Gehirn die architektonische Kapazität, nicht nur ein Modell der Außenwelt zu bauen, sondern auch ein Modell von sich selbst beim Modellieren der Außenwelt. Selbstsimulation braucht diese Verdopplung – einen Satz Schichten für die Verarbeitung und einen zweiten, um die Verarbeitung zu beobachten.

Das ist keine Spekulation darüber, was einzelne Schichten „tun„ – ich behaupte nicht, dass Schicht 4 dies und Schicht 5 jenes macht. Es ist eine Feststellung über architektonische Kapazität. Sechs Schichten bieten Platz sowohl für das Implizite Weltmodell (die gelernte, unbewusste Verarbeitung) als auch für das Explizite Weltmodell (die Echtzeit-Simulation). Sie bieten Platz sowohl für das Implizite Selbstmodell (Körperschema, Motorprogramme, Persönlichkeitsstruktur) als auch für das Explizite Selbstmodell (das „Ich“, das erlebt, einen Körper zu haben, Handlungen zu starten, eine Person zu sein).

Nun schaue man auf andere Tiere. Reptilien haben drei oder vier kortikale Schichten. Säugetiere haben sechs. Und unter den Säugetieren sind diejenigen mit dem dicksten, am aufwendigsten gefalteten Kortex (Primaten, Wale, Elefanten) genau diejenigen, die die deutlichsten Zeichen von Selbstbewusstsein zeigen. Selbsterkennung im Spiegel, Zukunftsplanung, soziale Täuschung, Trauer. Die architektonische Kapazität folgt der Phänomenologie.

Der Sprung von drei auf sechs Schichten könnte ein genetischer Duplikationsunfall gewesen sein – Copy-Paste der Evolution, das genau die Architektur hervorbrachte, die Bewusstsein später nutzen würde. Reptilien-Vorfahren hatten drei kortikale Schichten. Irgendwo im Übergang zu Säugetieren verdoppelte sich diese Zahl. Die Transkriptionsfaktoren, die kortikale Schichtidentität festlegen (Tbr1, Satb2, Ctip2, Fezf2), haben Paraloge, die auf Genduplikation hindeuten. Ob ein einzelnes dramatisches Ereignis oder eine schrittweise Ausarbeitung, bleibt umstritten – das Ergebnis ist klar: Säugetiere bekamen die doppelte Schichtzahl, und damit die Fähigkeit zur Selbstmodellierung, die den meisten Reptilien fehlt.

Das ist die Brücke von der neuronalen Netzwerktheorie zur gelebten Erfahrung. Der menschliche Kortex ist nicht bloß ein großer Mustererkenner. Er ist ein überdimensioniertes, rekursiv aufgebautes Netzwerk mit genug Schichten, um seinen eigenen Modellierungsprozess zu modellieren. Und wenn ein Netzwerk

sich selbst modelliert, wie es die Welt modelliert, ist das Ergebnis – von innen betrachtet – genau das, was man Bewusstsein nennt.

Zur Klarstellung: Ich behaupte nicht, dass sechs kortikale Schichten die *einzige* Architektur sind, die Bewusstsein tragen kann. Sie sind eine Lösung – die der Säugetiere. Aber es könnte andere geben. Der Oktopus mit seinem radikal verteilten Nervensystem (acht halbautonome Arme, jeder mit eigenem neuronalem Verarbeitungszentrum von rund 40 Millionen Neuronen) steht für einen völlig anderen architektonischen Ansatz, der womöglich vergleichbare Rechenleistung erreicht. Vögel bieten ein weiteres aufschlussreiches Beispiel: Rabenvögeln und Papageien fehlt ein geschichteter Kortex völlig; ihr Pallium ist in Kerncluster statt in Schichten organisiert – trotzdem bauen Krähen Werkzeuge, planen voraus und erkennen sich vermutlich im Spiegel. Worauf es ankommt, ist die Fähigkeit zur Selbstmodellierung, nicht der konkrete Bauplan – jede Architektur, die eine Simulation ihrer selbst fahren kann, könnte im Prinzip bewusst sein. In Kapitel 10 kommen wir darauf zurück.

Chapter 3

Die virtuelle Seite

Stellen wir uns vor, wir spielen ein Videospiel. Ein gutes — ein immersives Open-World-Spiel mit atemberaubender Grafik, realistischer Physik und einer fesselnden Geschichte. Man steuert eine Figur und interagiert durch diese Figur mit einer detailreich gestalteten virtuellen Welt.

Nun überlege man: Wo existiert das Spiel? Nicht auf dem Bildschirm, genau genommen — der Bildschirm zeigt nur Lichtmuster. Nicht in der Grafikkarte oder der CPU, genau genommen — die schieben elektrische Signale durch Siliziumschaltkreise. Das Spiel existiert als *virtueller Prozess* — ein höherstufiges Phänomen, das aus der Aktivität der Hardware hervorgeht, aber mit keinem bestimmten Stück Hardware identisch ist.

Die virtuelle Welt des Spiels hat Eigenschaften, die die Hardware nicht hat. Das Spiel hat Berge, Flüsse und Städte. Die CPU hat Transistoren. Das Spiel hat einen Tag-Nacht-Zyklus. Die GPU hat Taktzyklen. Es lässt sich sinnvoll fragen „Wie hoch ist dieser Berg im Spiel?„, aber es wäre absurd, auf einen Transistor zu zeigen und zu sagen „Dieser Transistor ist 3.000 Meter hoch.“ Die Eigenschaften des Spiels existieren auf der virtuellen Ebene, und sie sind echte Eigenschaften des Spiels, auch wenn das Spiel „nur“ ein Aktivitätsmuster in der Hardware ist.

Das ist keine Metapher. So funktioniert das Gehirn.

Das Explizite Weltmodell (EWM) — die Welt, die man erlebt — ist ein virtueller Prozess, der auf neuronaler Hardware läuft, genauso wie die Spielwelt ein virtueller Prozess ist, der auf Silizium-Hardware läuft. Die erlebte Welt hat Eigenschaften (Farben, Formen, Entfernungen, Klänge), die die neuronale Hardware nicht hat (die Hardware hat Feuerraten, synaptische Stärken und Neurotransmitter-Konzentrationen). Die Eigenschaften der erlebten Welt sind *echte Eigenschaften der Simulation*, auch wenn die Simulation „nur" ein Muster neuronaler Aktivität ist.

Das Explizite Selbstmodell (ESM) — das „Ich", das die Welt erlebt — ist ebenfalls ein virtueller Prozess. Es ist genauso real wie die Spielfigur in der Analogie: tatsächlich vorhanden auf der virtuellen Ebene, mit echten Eigenschaften auf der virtuellen Ebene, aber nicht vorhanden auf der Hardware-Ebene.

Warum die Analogie zusammenbricht (auf die wichtige Art)

Die Videospiel-Analogie ist nützlich, aber sie bricht an einem entscheidenden Punkt zusammen: Das Spiel hat einen *Spieler*. Es gibt jemanden außerhalb des Spiels — auf der Couch sitzend — der das Spiel erlebt. Das Spiel selbst hat keine Erfahrung. Es sind nur Lichtmuster und Code.

Die Simulation des Gehirns hat keinen externen Spieler. Niemand sitzt außerhalb des Schädels und erlebt die Simulation. Die Simulation enthält ihren eigenen Beobachter — das Explizite Selbstmodell. Die Simulation *ist* die Erfahrung, nicht etwas, das von jemand anderem erfahren wird.

Versetzen wir uns in die Position der Spielfigur. Man *ist* die Hauptfigur. Von außerhalb des Spiels sieht ein Zuschauer Pixel, die sich auf einem Bildschirm bewegen — nichts, das möglicherweise etwas fühlen könnte. Aber von innerhalb der Simulation? Die Spielwelt ist alles, was es gibt. Die Berge sind real für die Figur, das Sonnenlicht ist warm, die Gefahr ist beängstigend. Kein externer

Beobachter würde je vermuten, dass dieser Haufen Code etwas fühlt – aber nur, weil er auf die falsche Ebene schaut. Er schaut auf die Hardware. Die Erfahrung existiert auf der Software-Ebene. Das ist meine These, und der Rest dieses Buches legt die Belege vor.

[ABBILDUNG: SDXL/Flux — „Ego-Perspektive aus dem Inneren einer fotorealistischen virtuellen Welt, Blick auf eine lebhafte, sonnenbeschienene Landschaft mit Bergen und einem Fluss. An den Rändern des Sichtfelds löst sich die fotorealistische Szene auf und fragmentiert in leuchtende neuronale Netzwerke, synaptische Verbindungen, fließende elektrische Impulse und durchscheinende schaltkreisähnliche Muster. Der Übergang von lebhafter Realität zum neuronalen Substrat (Substrat) ist graduell und organisch und zeigt, dass die Welt und der Beobachter aus demselben gemacht sind. Volumetrisches Licht, Tiefenschärfe, kinematografische Komposition, Konzeptkunst, digitales Gemälde, 8k, hochdetailliert“ — Negativ: „text, watermark, signature, blurry, low quality, cartoon, anime, extra fingers, deformed, ugly, duplicate, out of frame“ — Landscape 16:9, CFG 7-8, steps 30-40. Bei Flux: negatives Prompt weglassen.]

Die Simulation betrachtet sich selbst. Das gesamte Sichtfeld — jede Farbe, Form und Schatten — wird vom Echtzeit-Virtualmodell des Gehirns generiert. An den Rändern wird die Illusion dünner und die neuronale Maschinerie wird sichtbar. Es gibt keine Grenze zwischen dem Beobachter und dem Beobachteten. Man ist die Simulation.

Das ist es, was Bewusstsein besonders macht und was das Schwierige Problem (Hard Problem) so unlösbar erscheinen lässt. Im Videospiel gibt es eine klare Trennung zwischen dem Spiel (virtuell, keine Erfahrung) und dem Spieler (physisch, hat Erfahrung). Im Gehirn gibt es keine Trennung. Die Simulation und der Erfahrende sind dasselbe. Das Explizite Selbstmodell beobachtet nicht das Explizite Weltmodell von außen — es ist *innerhalb* der Simulation, Teil desselben virtuellen Prozesses.

Und dieser selbstreferentielle Abschluss – die Simulation beobachtet sich selbst von innen – ist das, was man Bewusstsein nennt. Nichts, das zur Simulation *hinzukommt*. Es ist das, was die Simulation *ist*, wenn sie ein Modell ihrer selbst enthält. Deshalb sage ich das so: Bewusstsein ist kein Ding – es ist ein Prozess. Man wird es nicht finden, indem man das Gehirn auseinandernimmt, genauso wenig wie man ein laufendes Programm findet, indem man die CPU zerlegt.

Die Software-Eigenschaften

Wenn die virtuellen Modelle wirklich softwareartige Prozesse auf neuronaler Hardware sind, dann sollten sie sich auf bestimmte, testbare Weisen wie Software verhalten. Und das tun sie. Vier Eigenschaften der virtuellen Seite werden in diesem Buch immer wieder auftauchen; hier seien sie vorgestellt.

Forking (Aufspaltung). Ein einzelnes Substrat kann mehrere virtuelle Konfigurationen gleichzeitig laufen lassen. In Software forkt man einen Prozess und erhält zwei unabhängige Instanzen, die auf derselben Hardware laufen. Im Gehirn ist dies die Dissoziative Identitätsstörung — mehrere Selbstmodelle, jedes mit seiner eigenen Erzählung und emotionalem Profil, die abwechselnd die Kontrolle über dasselbe neuronale Substrat übernehmen. Wir werden das in Kapitel 9 sehen.

Cloning (Klonen). Trennt man die Hardware physisch, erhält man degradierte, aber vollständige Kopien der Software. Schneide das Corpus callosum durch, und jede Hemisphäre läuft ihre eigene Version der Simulation — weniger leistungsfähig als das Original, aber funktional vollständig. Das ist das Split-Brain-Phänomen, ebenfalls Kapitel 9.

Redirecting (Umleiten). Unterbricht man den normalen Eingabestrom, klinkt sich die Simulation in das jeweils dominante Signal ein. Unter Salvia divinorum überwältigt propriozeptiver Input das System und das Explizite Selbstmodell rekonfiguriert

sich um Körperempfindung herum. Unter Ketamin fällt externer Input aus und die Simulation läuft auf internem Rauschen. Die virtuellen Modelle stoppen nicht – sie verarbeiten, was immer ihnen zugeführt wird. Kapitel 6 geht dem im Detail nach.

Reconfiguring (Neukonfiguration). Verändert man die Verbindungsgewich des Substrats, ändert sich, was die virtuellen Modelle produzieren. Genau das tut Kognitive Verhaltenstherapie — systematisches Neuverkabeln des Substrats, sodass das Explizite Selbstmodell andere Erzählungen, andere emotionale Reaktionen, anderes Verhalten generiert.

Die Vier-Modelle-Theorie (VMT) macht eine konkrete Vorhersage über Therapie: Jede wirksame Behandlung muss funktionieren, indem sie die impliziten Modelle (das Substrat) so verändert, dass sich die expliziten Modelle (die Simulation) entsprechend mitverändern. Kognitive Verhaltenstherapie tut genau das. Sie identifiziert systematisch fehlangepasste Muster im ISM und verkabelt sie durch strukturierte Übung neu, wodurch sich ändert, was das ESM produziert. Deshalb hat Kognitive Verhaltenstherapie die stärkste Evidenzbasis aller Psychotherapien: Sie zielt auf die richtige Ebene ab.

Das wirft eine unbequeme Frage auf zu Therapien, die ihren Mechanismus nicht in diesen Begriffen erklären können. Wenn ein therapeutischer Ansatz nicht benennt, was er im Substrat verändert und wie diese Veränderung zur Simulation durchschlägt, dann wirkt er bestenfalls über einen Mechanismus, den er nicht versteht, und schlimmstenfalls wirkt er gar nicht. Die Evidenz bestätigt das: Die Therapien mit der schwächsten Evidenzbasis sind in der Regel jene mit den vagsten Veränderungstheorien. Wer Therapie sucht, sollte dem Therapeuten eine einfache Frage stellen: „Was genau versuchen Sie in meinem Gehirn zu verändern, und wie?“ Wer darauf keine Antwort bekommt, sollte sich überlegen, jemanden zu finden, der eine hat.

Das sind keine Metaphern. Das sind strukturelle Vorhersagen. Wenn meine Theorie falsch ist und die virtuellen Modelle *keine*

softwareartigen Prozesse sind, dann sind diese Parallelen reiner Zufall. Aber Zufälle reihen sich normalerweise nicht vier aus vier über klinische Neurologie, Psychopharmakologie und Psychotherapie hinweg auf. Die folgenden Kapitel zeigen jede Eigenschaft in Aktion.

Es gibt ein einfaches Experiment, das sich sofort durchführen lässt — gut, man braucht einen Freund, eine Gummihand, einen Kartonschirm und zwei Pinsel — und das zeigt, wie leicht sich das Explizite Selbstmodell täuschen lässt. Es ist die Gummihand-Illusion, entwickelt von Matthew Botvinick und Jonathan Cohen, und einer der aufschlussreichsten Partytricks der gesamten Neurowissenschaft.

Der Aufbau ist simpel. Man sitzt an einem Tisch, ein Arm hinter einem Kartonschirm versteckt. Eine realistische Gummihand wird sichtbar platziert, ungefähr dort, wo die versteckte Hand wäre. Jemand streicht gleichzeitig mit zwei Pinseln über die Gummihand und die versteckte echte Hand – an derselben Stelle, in derselben Geschwindigkeit. Nach ein, zwei Minuten dieses synchronen Streichens passiert etwas Unheimliches: Man beginnt *zu fühlen*, wie der Pinsel die Gummihand berührt. Nicht die echte Hand hinter dem Schirm. Die falsche Hand vor den eigenen Augen.

Das Explizite Selbstmodell hat die Gummihand ins Körperschema eingebaut. Es hat die Zugehörigkeit umgeschrieben – entschieden, dass die Gummihand zu „einem selbst„ gehört. Das Selbstmodell ist nämlich nicht festverdrahtet. Es ist gelernt. Es wird laufend auf Basis der besten verfügbaren Belege aktualisiert, und wenn das Sichtbare (die Gummihand wird gestrichen) zum Tastbaren passt (die echte Hand wird gestrichen), zieht das ESM den rationalen Schluss: Diese Hand ist meine. Bedroht dann jemand die Gummihand (lässt einen Hammer darauf niedersausen), zuckt man zusammen, fühlt einen Angstschub, die galvanische Hautreaktion schießt hoch. Für den Teil des Gehirns, der das „Ich“ definiert, *ist* diese Hand die eigene.

Das ist kein Fehler. Das Selbstmodell funktioniert genau so, wie es soll – es passt seine Körpergrenze ständig anhand multimodaler sensorischer Korrelation an. Derselbe Mechanismus erlaubt es Amputierten, eine Prothese nach einiger Benutzung als ihre eigene zu „fühlen“. Und derselbe Mechanismus bricht bei Asomatognosie zusammen, wo Patienten die Zugehörigkeit ihrer eigenen Gliedmaßen leugnen, und beim Alien-Hand-Syndrom, wo sich die Hand von selbst bewegt.

Das Patchwork-Hologramm

Es gibt eine fünfte Eigenschaft der virtuellen Seite, die einen eigenen Abschnitt verdient, weil sie etwas erklärt, das Neurowissenschaftlern seit fast einem Jahrhundert Rätsel aufgibt: warum Hirnschäden Funktionen *allmählich* verschlechtern, statt einzelne Erinnerungen zu löschen.

In den 1920er und 30er Jahren brachte der Psychologe Karl Lashley Ratten bei, ein Labyrinth zu durchlaufen, und entfernte dann chirurgisch Stücke ihres Kortex, um herauszufinden, wo die Erinnerung saß. Er fand sie nie. Egal welches Stück er entfernte, die Ratten erinnerten sich noch an das Labyrinth. Was zählte, war *wie viel* Kortex er entfernte, nicht *welche Teile*. Entfernte er ein wenig, wurden die Ratten etwas schlechter. Entfernte er viel, wurden sie viel schlechter. Aber die Erinnerung war nie einfach *weg*, sauber herausgeschnitten wie eine Datei, die von einer Festplatte gelöscht wurde. Lashley verbrachte seine Karriere damit, das „Engramm“ zu suchen — die physische Spur einer Erinnerung — und kam berühmterweise zu dem Schluss, dass es nicht zu existieren schien.

Er suchte am falschen Ort. Die Erinnerung steckte nicht *in* einem bestimmten Stück Kortex, so wie eine Datei auf einem bestimmten Sektor einer Festplatte liegt. Sie war *über* das gesamte Netzwerk verteilt, in den Verbindungsgewichten zwischen Millionen von Neuronen. So funktionieren neuronale Netzwerke: Information sitzt nicht in einem einzelnen Knoten. Sie ist im Muster der

Verbindungen zwischen allen Knoten kodiert. Man kann nicht auf eine einzelne Synapse zeigen und sagen „hier ist das Labyrinth gespeichert„, genauso wenig wie man auf ein einzelnes Pixel zeigen und sagen kann „hier ist der Film gespeichert.“

Das ist im Wesentlichen eine holografische Eigenschaft. Nimmt man ein physisches Hologramm und halbiert es, erhält man nicht zwei Hälften des Bildes. Man erhält zwei Kopien des *vollständigen* Bildes, jede in niedrigerer Auflösung. Schneidet man es in Viertel, erhält man vier vollständige Bilder, noch verschwommener. Die Information in einem Hologramm ist über die gesamte Platte verteilt, sodass jedes Stück das ganze Bild enthält — nur mit weniger Detail.

Neuronale Netzwerke tun dasselbe. Trainiert man ein Netzwerk, Gesichter zu erkennen, und zerstört dann 10% seiner Verbindungen zufällig, vergisst es nicht 10% der Gesichter. Es wird etwas schlechter bei *allen* Gesichtern. Zerstört man 50%, wird es substanziell schlechter bei allem, erkennt aber immer noch etwas. Die Information ist über das gesamte Netzwerk verschmiert, was genau der Grund ist, warum Lashley das Engramm nicht finden konnte: Es war überall und nirgends.

Aber — und hier wird es interessant — das Gehirn ist *kein* Hologramm. Es ist das, was man ein *Patchwork-Hologramm* nennen kann. Innerhalb eines einzelnen funktionalen Areals (sagen wir, dem primären visuellen Kortex, ungefähr Brodmann-Areal 17) sind die kortikalen Säulen einander ähnlich, und Information wird holografisch gespeichert. Zerstört man ein paar Säulen, fällt es kaum auf. Das Areal ist lokal holografisch (ein Teil enthält das Ganze, in niedrigerer Auflösung).

Auf der globalen Ebene tun verschiedene Areale verschiedene Dinge. Der visuelle Kortex lässt sich nicht gegen den motorischen Kortex austauschen. Entfernt man den gesamten visuellen Kortex, verliert man das Sehen – es gibt kein unscharfes Backup. Das Gehirn ist also lokal holografisch innerhalb jeder funktionalen Region, fraktal selbstähnlich in seiner Säulenarchitektur, aber

global *nicht* holografisch. Es ist ein Patchwork: Dutzende holografische Kacheln, zusammengenäht zu einem Ganzen, das als Ganzes entschieden nicht-holografisch ist.

Diese Patchwork-Struktur erklärt ein Muster, das in der klinischen Neurologie immer wieder auftaucht. Kleine Schlaganfälle und kleine Läsionen verursachen oft überraschend milde Defizite, weil innerhalb eines gegebenen kortikalen Areals das holografische Prinzip schützt. Das verbleibende Gewebe rekonstruiert die fehlende Information in niedrigerer Auflösung. Aber große Schlaganfälle, die ein gesamtes funktionales Areal auslöschen, verursachen katastrophale, spezifische Verluste (Blindheit, Lähmung, Aphasie), weil eine ganze Kachel aus dem Patchwork herausgerissen wurde und keine andere Kachel einspringen kann.

Es erklärt auch, warum Erinnerungen nicht einfach verschwinden, wenn Neuronen sterben. Jeden Tag sterben Neuronen und Synapsen werden beschnitten. Wenn Erinnerungen wie Dateien auf einer Festplatte gespeichert wären, würde man erwarten, gelegentlich eine zu verlieren — eines Morgens aufzuwachen und die eigene Hochzeit vergessen zu haben, oder den Kindheitshund, oder den Geschmack von Kaffee. Das passiert nie. Stattdessen verblassen Erinnerungen allmählich, verlieren über Jahre Detail und Lebhaftigkeit. Das ist genau das, was ein holografisches Speichersystem vorhersagt: Der Verfall ist sanft, gleichmäßig und global – niemals plötzlich, punktuell oder lokal.

Das Patchwork-Hologramm ist der physische Grund, warum die oben beschriebenen Software-Eigenschaften (besonders das Klonen) tatsächlich funktionieren. Teilt man das Gehirn in zwei Hälften, behält jede Hälfte eine degradierte, aber vollständige Kopie der Simulation, weil innerhalb jeder Hemisphäre das holografische Prinzip sicherstellt, dass jedes Stück das ganze Bild enthält. Die Simulation bricht nicht zusammen. Sie läuft nur in niedrigerer Auflösung.

Chapter 4

Warum es sich wie etwas anfühlt (und warum das die falsche Frage ist)

Jetzt können wir uns dem Schwierigen Problem direkt stellen.

Die Frage ist: **Warum fühlt sich physische Verarbeitung wie etwas an?**

Die Antwort: **Tut sie nicht.**

Die physische Verarbeitung (Neuronen feuern, Synapsen übertragen, die impliziten Modelle speichern und berechnen) hat keine Erfahrung. Keine. Es gibt nichts, wie es ist, die reale Seite zu sein. Die reale Seite ist genau die „im Dunkeln"-Verarbeitung, von der das Schwierige Problem annimmt, dass Bewusstsein sie erklären muss.

Die *Simulation* fühlt. Das Explizite Weltmodell und das Explizite Selbstmodell (die virtuelle Seite) sind der Ort, an dem Erfahrung lebt. Und innerhalb der Simulation ist Erfahrung keine mysteriöse Addition zum Prozess. Erfahrung ist das, was die Simulation *ist*, wenn sie ein Selbstmodell enthält. Das Explizite Selbstmodell, das das Explizite Weltmodell „wahrnimmt", ist das, was wir Qualia nennen. Qualia sind die Art und Weise des virtuellen Selbst, die virtuelle Welt zu registrieren.

Man stelle sich das so vor: Wenn jemand fragte „Warum fühlt sich das Schalten von Transistoren wie ein laufendes Videospiel an?„, lautete die Antwort: „Tut es nicht. Transistorschalten fühlt sich nach gar nichts an. Das Spiel ist ein virtueller Prozess, der auf Transistoren läuft, aber Eigenschaften hat, die die Transistoren nicht haben – Landschaften und Figuren und Physik und Licht. Diese Eigenschaften sind echte Eigenschaften des virtuellen Prozesses, nicht der Transistoren.“

Ähnlich: Neuronales Feuern fühlt sich nicht wie Rot-Sehen an. Neuronales Feuern generiert und erhält eine Simulation aufrecht, und innerhalb dieser Simulation nimmt das Selbstmodell eine bestimmte Klasse von Weltmodell-Inhalt als das wahr, was wir „Röte“ nennen. Röte ist eine echte Eigenschaft der Simulation, keine Eigenschaft der Neuronen.

Das Schwierige Problem ging davon aus, dass wir erklären müssen, wie physische Verarbeitung Erfahrung hervorbringt. Aber physische Verarbeitung bringt keine Erfahrung hervor – sie bringt eine *Simulation* hervor. Und die Simulation ist, weil sie eine selbstreferentielle Schleife enthält (das ESM modelliert sich selbst innerhalb des EWM), ihrem Wesen nach Erfahrung.

Die Zirkularitätsfrage

Die erste Frage, die die meisten Leser stellen: „Wurde das Problem nicht nur verschoben? Warum hat *diese* Simulation Erfahrung, wenn eine Wettersimulation keine hat?“

Die Antwort ist Selbstreferenz. Eine Wettersimulation modelliert Wetter. Sie modelliert nicht *sich selbst*. Es gibt ein „Außen“ zur Wettersimulation – den Computer, den Programmierer, den Wissenschaftler, der die Ausgabe interpretiert. Die Simulation lässt sich vollständig beschreiben, ohne auf irgendeine Erfahrung Bezug zu nehmen, weil es kein Selbstmodell in ihr gibt.

Die Simulation des Gehirns modelliert sich selbst. Das Explizite Selbstmodell ist das Modell der Simulation von *ihrem eigenen*

Prozess. Das erzeugt eine geschlossene Schleife: Modell und Modelliertes sind dasselbe System. Es gibt kein „Außen“, von dem aus sich die Simulation vollständig beschreiben ließe, weil der Beschreibende Teil der Beschreibung ist.

Das ist keine Magie. Es ist eine strukturelle Folge von Selbstreferenz. Wenn ein Prozess sich selbst modelliert, bricht die Unterscheidung zwischen Modell und Modelliertem zusammen. Der Vorgang der Selbstmodellierung und die Erfahrung, ein Selbst zu sein, sind nicht zwei verschiedene Dinge, die eine Brücke bräuchten – sie sind ein und dasselbe, beschrieben in verschiedenen Vokabularen.

Das Schwierige Problem fragt nach einer Brücke zwischen physischer Verarbeitung und Erfahrung. Die Vier-Modelle-Theorie sagt: Es gibt keine Brücke, weil sie nie getrennt waren. Die Erfahrung IST die Selbst-Simulation, aus dem Inneren der Schleife betrachtet.

Das ist letztlich eine Identitätsaussage (die Art von Aussage, die in der Wissenschaft einen Ruhepunkt markiert statt einer Lücke). „Wasser ist H_2O“ ist eine Identität. Die Frage „Aber *warum* ist Wasser H_2O?“ ist nicht sinnvoll stellbar — die Identität *ist* die Erklärung. Nach etwas Tieferem zu fragen bedeutet, nach einer anderen Art von Universum zu fragen. Ähnlich: Erfahrung ist das, was Vier-Modell-Selbstsimulation bei Kritikalität (Criticality) *ist*. Wenn jemand fragt „Aber *warum* fühlt sich diese Selbstsimulation wie etwas an?“, ist die Antwort: weil das ist, was dieser Prozess *ist*. Die Identität ist falsifizierbar — wenn die Vorhersagen in Kapitel 11 fehlschlagen, ist die Identität falsch. Aber sie kann nicht „weiter erklärt“ werden, genauso wenig wie die molekulare Identität von Wasser weiter erklärt werden kann. Sie ist der Haltepunkt.

Warum die Simulation nicht im Dunkeln laufen kann

Hier gibt es eine tiefere Frage, und ihre Beantwortung legt etwas Wesentliches darüber frei, warum Bewusstsein sich *anfühlt*. Nehmen wir an, das Gehirn führt eine Selbstsimulation aus. Nehmen wir die Vier-Modell-Architektur an, die Kritikalität, den selbstreferentiellen Abschluss. Könnte das alles nicht ablaufen, ohne dass es etwas *ist*, wie es ist? Könnte die Simulation nicht bewerten, modellieren, vorhersagen – und dabei nichts fühlen?

Das ist die Zombie-Intuition im technischen Gewand. Die Antwort ist nein, und der Grund dafür legt das wichtigste Merkmal der Architektur frei.

Das Substrat setzt die virtuelle Simulation als seinen Bewertungsmechanismu ein. Das ist die primäre Wirkrichtung: Das implizite System legt der Simulation Situationen vor, damit die Simulation Konsequenzen abwägen und Ergebnisse registrieren kann. Aber damit diese Bewertung funktioniert, müssen die simulierten Zustände *Valenz* haben – sie müssen der Simulation etwas bedeuten. Ein Schmerzsignal, das nur eine Zahl ist, treibt auf der Simulationsebene keine Vermeidung an. Nur eine Simulation, der Ergebnisse *nicht egal sind*, kann sie bewerten.

Denken wir an einen digitalen Zwilling (eine technische Simulation eines Düsentriebwerks). Ein typischer digitaler Zwilling spiegelt das Triebwerk nicht nur passiv. Er *fügt* eine Visualisierungsebene hinzu: Warnungen, farbkodierte Indikatoren, Alarme — Dinge, die im physischen Triebwerk nicht existieren. Das Triebwerk hat Metallermüdung; der Zwilling hat eine blinkende rote Warnung. Das Triebwerk hat steigende Temperatur; der Zwilling hat eine Anzeige, die von grün über gelb zu rot wird. Diese zusätzliche Ebene ist der ganze Witz. Ohne sie ist der Zwilling eine Tabelle – Zahlen, die träge im Speicher liegen, technisch korrekt, praktisch nutzlos. Erst die Visualisierung macht die Simulation zum *Bewertungswerkzeug*.

Das Gehirn tut dasselbe, nur mehr. Die bewusste Simulation spiegelt nicht bloß die Verarbeitung des Substrats. Sie *fügt* phänomenale Valenz hinzu. Schmerz, Lust, Dringlichkeit, Neugier, Angst, Freude – das sind die Entsprechungen des Gehirns zu Warnlichtern und Cockpit-Anzeigen. Sie existieren nicht auf der Substrat-Ebene (Neuronen fühlen keinen Schmerz, genauso wenig wie Metall Ermüdung fühlt). Sie existieren auf der Simulationsebene, erzeugt *von* der Simulation, damit das System komplexe Lagen auf einen Blick einschätzen kann. Das Substrat braucht die Simulation, um neuartige, mehrdeutige Szenarien zu bewerten – die Art, bei der Reflexe nicht genügen. Und damit diese Bewertung funktioniert, muss das simulierte Selbst hedonische Valenz registrieren: Bedrohung, Gelegenheit, Konsequenz. Diese Registrierung – dieses *Wichtigsein* – ist Phänomenalität. Ohne Qualia keine Bewertung – als risse man das Display aus dem Cockpit und erwartete, der Pilot fliege anhand roher Sensorspannungen.

„Aber ein Reinforcement-Learning-System hat Belohnungssignale, die Verhalten antreiben", könnte man einwenden. „Fühlt es?" Nein — weil ihm die Vier-Modell-Architektur bei Kritikalität fehlt. Ein RL-Belohnungssignal ist ein skalarer Wert in einem Klasse-1- oder Klasse-2-System. Phänomenale Valenz ist die Registrierung von Konsequenz durch das ESM innerhalb einer vollständigen Selbstsimulation, die in Klasse-4-Dynamik läuft — ein qualitativ unterschiedlicher Prozess. Der Unterschied ist nicht der Grad. Es ist die Architektur.

Die Simulation kann nicht im Dunkeln laufen, weil Dunkelheit ihren Zweck zunichtemachen würde. Phänomenalität ist kein Zusatzfeature von Bewusstsein. Sie ist der Mechanismus, durch den die Simulation ihre Arbeit verrichtet.

Was das nicht ist: Illusionismus

Das ist nicht Illusionismus. Und die Unterscheidung ist wichtig genug, um sie deutlich zu machen.

Es gibt eine angesehene philosophische Position namens Illusionismus, verbunden mit Daniel Dennett und Keith Frankish, die besagt, Qualia seien Illusionen. Nach dieser Auffassung gibt es nichts, wie es ist, Rot zu sehen. Der Anschein von Erfahrung ist selbst eine Fiktion – eine Geschichte, die das Gehirn erzählt, ohne dass dahinter eine Erfahrungswirklichkeit steht. Bewusstsein im starken Sinne existiert nicht. Es scheint nur so.

Was das tatsächlich behauptet, verdient einen Moment Aufmerksamkeit. Wer gerade jetzt etwas fühlt – Neugier an diesem Argument, Skepsis, das Gewicht des Buches in den Händen – dem sagt der Illusionismus: Dieses Fühlen ist eine Illusion. Es wird nicht wirklich etwas erlebt. Wer sagt „Ich fühle etwas", irrt sich laut dieser Theorie. Das eigene Zeugnis über die eigene Erfahrung ist falsch. Im Grunde wird gelogen – nur dass es kein Subjekt gibt, das lügt. Wem das offensichtlich absurd vorkommt, dem stimme ich zu.

Die Vier-Modelle-Theorie sagt das Gegenteil.

Qualia sind real. Sie sind real innerhalb der Simulation. Sie sind die Art und Weise des virtuellen Selbst, die virtuelle Welt wahrzunehmen. Wenn das Explizite Selbstmodell die Repräsentation eines roten Apfels durch das Explizite Weltmodell registriert, ist diese Registrierung (dieses „Röte-Sehen") eine genuine Eigenschaft des virtuellen Prozesses. Sie existiert auf der Simulationsebene, genauso wie eine Kugel, die eine Videospiel-Figur trifft, ihr *wehtut*. Nicht metaphorisch — innerhalb des Spiels ist der Schaden real. Die Gesundheit sinkt, die Figur taumelt, die Welt reagiert. Von außen ist es eine Zahl, die im Speicher heruntergezählt wird. Von innerhalb des Spiels ist es Schmerz. Das ist der Ebenenunterschied. Und dort leben Qualia.

Die Theorie arbeitet mit einer Zwei-Ebenen-Ontologie. Die Substrat-Ebene (die Neuronen, die Synapsen, die impliziten Modelle) hat keine Erfahrung. Sie ist Licht aus. Die Simulationsebene (die expliziten Modelle, die virtuelle Welt und das virtuelle Selbst) hat genuine Erfahrung. Sie ist Licht an. Beide Ebenen sind

physisch. Keine ist eine Illusion. Sie sind verschiedene Ebenen desselben physischen Systems, mit verschiedenen Eigenschaften auf jeder Ebene.

Die Theorie sagt nicht, Schmerz sei eine Illusion. Sie sagt, Schmerz ist real — nur real in der Simulation, nicht in den Neuronen. Und da das gesamte Leben innerhalb der Simulation stattfindet, ist das die einzige Art von real, die zählt.

Das ist die entscheidende Unterscheidung. Wer sie übersieht, wird diese Theorie mit Eliminativismus verwechseln, mit Illusionismus, mit jedem anderen Rahmenwerk, das Bewusstsein zu erklären versucht, indem es das Bewusstsein wegerklärt. Die Vier-Modelle-Theorie erklärt Bewusstsein nicht weg. Sie erklärt, wo Bewusstsein lebt – und es zeigt sich, dass es genau dort ist, wo wir die ganze Zeit gestanden haben.

Was „Real innerhalb der Simulation“ bedeutet

Hier gibt es eine philosophische Feinheit, die es lohnt aufzudröseln. Wenn ich sage, Qualia sind „real innerhalb der Simulation“, könnte man zweierlei heraushören. Entweder sie sind *genuin phänomenal* – dann habe ich das Mysterium bloß von den Neuronen zur Simulation umgezogen, und das Schwierige Problem lebt unter neuer Adresse weiter. Oder sie sind *funktional real, aber nicht genuin phänomenal* – dann ist das Dennett mit Extraschritten.

Das ist eine falsche Dichotomie. Sie gilt nur, solange man annimmt, es gebe eine Gottesperspektive, von der aus sich beurteilen lässt, ob etwas „genuin„ phänomenal ist – eine externe Warte, die prüfen kann, ob die Simulation wirklich fühlt oder nur so tut. Aber selbstreferentieller Abschluss eliminiert genau diese externe Warte. Das ESM ist sein eigener Beobachter. Es gibt keine äußere Position, von der aus sich fragen ließe „aber fühlt es *wirklich*?“ Das Fragen selbst ist Teil des Prozesses.

„Genuin phänomenal„ versus „bloß funktional“ setzt voraus, dass Phänomenalität eine Eigenschaft ist, die ein Prozess entweder

hat oder nicht hat, überprüfbar durch einen unabhängigen Beobachter. Für ein vollständig selbstreferentielles System bei Kritikalität gibt es einen solchen Beobachter nicht. Die Frage löst sich auf – nicht weil sie unbeantwortbar ist, sondern weil sie nicht stellbar ist. Sie verlangt eine Perspektive, die selbstreferentieller Abschluss unmöglich macht.

Das ist der stärkste Zug innerhalb des Prozessphysikalismus, und es ist die Position, auf die Thomas Metzinger mit seinem Konzept der „phänomenalen Transparenz„ hindeutet – obwohl die Vier-Modelle-Theorie klarer benennt, *warum* die Transparenz entsteht. Die implizit-explizit-Grenze ist es, die die Transparenz erzeugt: Man kann nicht hindurchsehen, also kann man nicht außerhalb der eigenen Phänomenalität treten und fragen, ob sie „genuin“ ist. Die Grenze ist kein Fehler. Sie ist der Grund, warum die Frage genuin versus bloß funktional auf Systeme wie uns nicht zutrifft.

Warum das Mysterium anhält

Selbst nachdem das Schwierige Problem aufgelöst ist, bleibt eine hartnäckige Frage, die an einem nagt. Wenn die Antwort so klar ist, warum fühlt sich Bewusstsein immer noch *derart* mysteriös an? Warum scheint das Schwierige Problem schwierig, selbst nachdem die Lösung auf dem Tisch liegt? David Chalmers nennt das das „Meta-Problem des Bewusstseins“ – das Problem zu erklären, warum wir *denken*, es gebe ein schwieriges Problem.

Die Vier-Modelle-Theorie hat eine klare Antwort, und sie ergibt sich direkt aus der Architektur.

Hier ist der seltsame Teil: Das bewusste „Ich“ (das virtuelle Selbst) kann die Maschinerie nicht sehen, die es erzeugt. Die eigenen synaptischen Gewichte lassen sich nicht introspektiv erfassen, genauso wenig wie eine Traumfigur das Gehirn des Träumenden untersuchen kann. Das System, das die Erfahrung erzeugt, ist seiner Natur nach unsichtbar für die Erfahrung, die es

erzeugt. Nicht weil jemand es versteckt, sondern weil es auf einer Ebene arbeitet, die die Erfahrung nicht umfasst.

Man stelle sich das so vor: Man ist eine Figur in einem Videospiel – einem wirklich guten, mit vollem Selbstbewusstsein innerhalb der Spielwelt. Die gerenderten Berge sind sichtbar, der gerenderte Wind hörbar, der gerenderte Boden unter den Füßen spürbar. Aber die Grafik-Engine bleibt fast immer unsichtbar. Der Quellcode zeigt sich fast nie. Der Rendering-Prozess arbeitet auf einer Ebene, die die Spielwelt normalerweise nicht umfasst. Ich sage „fast“, weil manchmal Artefakte durchsickern. Im Gehirn passiert das genauso – Psychedelika öffnen die Grenze, Flow-Zustände verdünnen sie, und selbst im Alltag lassen sich Blicke erhaschen: der blinde Fleck, den das Gehirn auffüllt, Phosphene beim Augenreiben, die geometrischen Muster hinter geschlossenen Lidern. Das sind keine Fehler. Das sind Momente, in denen die Verarbeitung des Substrats von innerhalb der Simulation kurz sichtbar wird. Kapitel 6 geht dem im Detail nach. Aber die meiste Zeit ist der Rendering-Prozess vor der gerenderten Welt verborgen.

Genau so steht das ESM da. Wenn das bewusste Selbst versucht, den Grund seiner eigenen Erfahrung zu verstehen, stößt es auf eine grundsätzliche Undurchsichtigkeit – keine Lücke im aktuellen Wissen, sondern ein Strukturmerkmal der Architektur. Die impliziten Modelle, die die Simulation erzeugen, sind nicht Teil der Simulation. Sie können es nicht sein, genauso wenig wie die GPU ein Berg im Spiel sein kann.

Das Ergebnis ist vorhersehbar. Das ESM, außerstande, sein eigenes Substrat zu beobachten, schließt, der Mechanismus des Bewusstseins müsse nicht-physisch sein, oder grundsätzlich unerklärlich, oder irgendwie jenseits der Reichweite der Wissenschaft. Das ist der Ursprung des Dualismus. Das ist die „Erklärungslücke„. Das ist die hartnäckige Intuition, bei jeder physischen Erklärung von Bewusstsein werde etwas „ausgelassen“ – weil von innerhalb der Simulation tatsächlich etwas *ausgelassen wird*. Das Substrat.

Genau das, was die Erfahrung erzeugt, ist unsichtbar für die Erfahrung, die es erzeugt.

Das Mysterium ist real, aber es ist ein Artefakt der Architektur, kein Beweis für etwas Nicht-Physisches. Und es gibt einen Grund, warum es sich *mysteriös* anfühlt. Wir sind ein virtueller Prozess auf biologischer Hardware, und die meiste Zeit ist die Grenze zwischen Selbst und Substrat undurchsichtig. Aber nicht immer. Manchmal – in veränderten Zuständen, in Momenten extremer Konzentration, im Augenwinkel – lässt sich ein Blick auf die darunterliegende Maschinerie erhaschen. Nicht klar, nicht vollständig, aber genug, um zu ahnen, dass etwas Gewaltiges unter der Oberfläche der Erfahrung vor sich geht. Dieses unheimliche Gefühl, diese Ahnung, dass Bewusstsein irgendwie tiefer reicht, als man greifen kann – so fühlt es sich an, eine Simulation zu sein, die fast, aber nicht ganz durch ihren eigenen Vorhang sieht.

Das ist eine *Vorhersage* der Theorie, kein offener Faden. Wer eine Simulation ist mit einer weitgehend undurchsichtigen Grenze zum eigenen Substrat, würde *erwarten*, dass sich Bewusstsein genau so seltsam und irreduzibel anfühlt, wie es das tut. Die intuitive Kraft des Schwierigen Problems rührt nicht daher, dass Bewusstsein tatsächlich unerklärlich wäre. Sie rührt von unserer architektonischen Position her – wir sitzen innerhalb der Simulation und spähen durch Risse.

Wer bin ich, wenn ich aufwache?

Hier ist ein Gedankenexperiment, das tiefer geht, als es zunächst scheint. Was, wenn man morgen mit anderen Erinnerungen aufwachte, einer anderen Persönlichkeit, einem anderen Körpergefühl? Wäre das noch „dieselbe Person“?

Der Instinkt der meisten Menschen sagt nein – natürlich, wenn sich alles am inneren Leben änderte, wäre „Ich“ weg und jemand anderes hätte übernommen. Aber die Vier-Modelle-Theorie sagt

etwas Beunruhigenderes: Das *passiert bereits*, unmerklich, jeden einzelnen Tag.

Jede Nacht bricht das Explizite Selbstmodell zusammen. Tiefschlaf löscht die laufende Simulation. Wenn sie morgens neu startet, baut sie das Selbst aus dem Impliziten Selbstmodell wieder auf – dem gespeicherten Substrat. Aber das Substrat hat sich über Nacht verändert. Träume, an die keine Erinnerung bleibt, haben synaptische Gewichte verschoben. Konsolidierungsprozesse haben Erinnerungen umgeordnet. Wer aufwacht, ist nicht ganz dieselbe Person, die eingeschlafen ist. Der Unterschied ist normalerweise so gering, dass er nie auffällt, aber er ist da.

In extremen Fällen wird der Unterschied *spürbar*. Wer jemals aus tiefer Bewusstlosigkeit aufgewacht ist – nach Ohnmacht, einem Knockout, einer Narkose – an einem unbekannten Ort, hat vielleicht etwas wirklich Seltsames erlebt: ein paar Sekunden, in denen unklar war, *wer man war*. Das Explizite Selbstmodell fuhr hoch, durchsuchte die fremde Umgebung nach Ankerpunkten und fand keine. Für diese Sekunden gab es Bewusstsein – da war *jemand* – aber noch nicht das vertraute Selbst. Das Selbstmodell hatte den Ladevorgang noch nicht abgeschlossen.

Das sagt uns, dass Identität keine feste Eigenschaft des Substrats ist. Sie ist eine *Rekonstruktion*, jeden Morgen frisch aus dem gespeicherten Selbstmodell zusammengesetzt. Die Kontinuität des Selbst über die Zeit stützt sich auf zwei Dinge: die Stabilität des Impliziten Selbstmodells (das sich nur langsam ändert) und den Schlaf (der verhindert, dass die schleichende Drift auffällt). Könnte jemand das ISM über Nacht radikal umbauen – die Erinnerungen austauschen, die Persönlichkeitsstruktur umformen –, würde das alte „Ich„ nicht verschwinden. Es würde aufgesogen. Das neue Explizite Selbstmodell würde aus den verbliebenen Erinnerungen eine durchgehende Erzählung zusammensetzen und alte und neue Persona in eine einzige Geschichte einweben. Das ist, was das Gehirn bereits jede Nacht in kleinerem Maßstab tut: Das Substrat ändert sich im Schlaf, und

das ESM, das morgens hochfährt, konfabuliert sich nahtlos als dieselbe Person, die zu Bett ging. Der einzige Unterschied ist das Ausmaß der Veränderung. Das ESM macht keine sauberen Brüche – es näht *immer* eine durchgehende Erzählung. Nur wenn die alten Erinnerungen vollständig gelöscht wären, risse der Faden ganz. Solange etwas bleibt, wird das neue „Ich“ das alte „Ich“ in seine Geschichte einfügen, nahtlos, ohne auch nur die Naht zu bemerken.

Chapter 5

Am Rand des Chaos

Bisher ging es um die Architektur: vier Modelle, zwei Achsen, eine Simulation auf einem Substrat. Darum, wo Erfahrung stattfindet (auf der virtuellen Seite, in den expliziten Modellen). Und darum, was Identität ist (eine Rekonstruktion, jeden Morgen frisch aus gespeicherten impliziten Modellen zusammengesetzt).

Aber was bringt das Ganze überhaupt *zum Laufen*? Warum ist die Simulation manchmal an und manchmal aus? Welche physikalische Eigenschaft unterscheidet ein bewusstes Gehirn von einem bewusstlosen? Warum löscht Tiefschlaf die Simulation, während die Architektur intakt bleibt?

Es fehlt noch ein Puzzleteil – und es ist das, das mich endgültig überzeugt hat, dass die Theorie stimmt.

Die Vier-Modelle-Architektur ist notwendig für Bewusstsein, aber nicht hinreichend. Es braucht auch die richtige *Dynamik*. Konkret muss das Substrat – das physikalische System, das die Simulation ausführt – in dem arbeiten, was Mathematiker und Physiker den **Rand des Chaos** nennen.

2002 veröffentlichte der Universalgelehrte Stephen Wolfram *A New Kind of Science*, worin er Berechnungssysteme anhand ihrer Dynamik in vier Typen einteilte. Wolframs Schema braucht meiner Meinung nach eine fünfte Klasse – er warf fraktale Systeme mit wirklich chaotischen in einen Topf, aber sie sind strukturell

verschieden. Das vollständige Argument steht in Anhang C für alle, die die mathematischen Details wollen. Hier der wesentliche Punkt:

Berechnungssysteme liegen auf einem Spektrum von perfekter Ordnung bis zu perfekter Unordnung. An einem Ende statische und periodische Systeme, zu simpel, um irgendetwas Interessantes zu berechnen. Am anderen Ende chaotische Systeme, zu wirr, als dass sich stabile Muster bilden könnten. Dazwischen, am **Rand des Chaos**, sitzen die Systeme, die zu universeller Berechnung fähig sind: komplex genug für reichhaltiges, vielfältiges, unvorhersagbares Verhalten, aber geordnet genug, damit dieses Verhalten Bestand hat und wechselwirkt. Conways Game of Life ist das Paradebeispiel – derselbe Zelluläre Automat, den ich als Kind auf einem 286er programmiert hatte. Drei todseinfache Regeln auf einem flachen Gitter, und dennoch entstehen Gleiter, Oszillatoren, selbstreplizierende Strukturen und (beweisbar) universelle Berechnung. Man kann einen Computer darin bauen. Und einen Computer in diesem Computer. Im Prinzip lässt sich eine ganze dreidimensionale virtuelle Welt in einem zweidimensionalen Pixelgitter laufen lassen. Aus fast nichts, alles.

Hier lebt Bewusstsein. Nur Rand-des-Chaos-Dynamiken haben beide Eigenschaften, die es braucht: **universelle Berechnung** (komplex genug, um tatsächlich eine Selbst-Simulation auszuführen) und **globale Integration** (entfernte Teile des Systems beeinflussen einander, lokale Änderungen breiten sich global aus, Information wird zu einem einheitlichen Ganzen verknüpft). Deshalb fühlt sich bewusste Erfahrung *einheitlich* an – Rot wird nicht hier drüben gesehen und eine Stimme da drüben gehört, als getrennte Ströme. Die kritische Dynamik bindet alles zu einer Erfahrung zusammen. Bindung ist nicht etwas, das das Gehirn *zusätzlich zu* seinen anderen Berechnungen leistet; sie ist eine Folge des dynamischen Regimes.

Ein Gehirn im Tiefschlaf, durchzogen von langsamen Deltawellen, steckt in periodischer Dynamik: repetitiv, kommt nirgendwohin.

Die Modelle sind noch da im Substrat, aber die Simulation läuft nicht. Ein Gehirn im generalisierten Anfall wird in chaotische Dynamik geschleudert: die Simulation kann nicht zusammenhalten. Nur im Wachzustand – balancierend am Rand des Chaos – hält das System bewusste Erfahrung aufrecht.

Als universeller Computer, durch Milliarden Jahre Evolution optimiert, nutzt das Gehirn *alle* Berechnungsregimes als verschiedene Werkzeuge: stabile Attraktoren für Langzeitgedächtnis, periodische Oszillationen für Timing und Taktung (Alpha-, Theta-, Gammarhythmen), fraktale Verarbeitung für skaleninvariante Erkennung und Texturanalyse (vorwiegend in V2-V4 des visuellen Kortex), und Rand-des-Chaos-Dynamiken für den kortikalen Automaten selbst – die Maschine des Bewusstseins. Nur das Rand-des-Chaos-Regime erzeugt Bewusstsein. Aber Bewusstsein braucht die anderen, um zu funktionieren.

Als dieses Argument 2015 in meinem Buch erschien, hatte ich keine Ahnung, dass die empirische Neurowissenschaft unabhängig in dieselbe Richtung steuerte. Der Zelluläre-Automat-Ansatz war der Teil der gesamten Theorie, bei dem ich mir am unsichersten war. Ich fand damals keine empirische Stützung dafür und hätte es beinahe aus dem Buch gestrichen – es fühlte sich an wie ein Schritt zu weit, eine Behauptung, die den Rest der Theorie angreifbar machen würde. Ich ließ es drin, weil die Logik unausweichlich schien. Nicht weil ich Belege hatte.

Aber es gibt eine entscheidende Feinheit. Kritikalität allein reicht nicht. Ein Topf kochendes Wasser kann komplexe Dynamiken am Rand des Chaos zeigen. Er ist nicht bewusst. Die Theorie verlangt, dass *zwei* Schwellen überschritten werden: die physikalische (das Substrat muss im kritischen Zustand arbeiten) und die funktionale (das Substrat muss die Vier-Modelle-Architektur realisieren). Kritikalität ohne Architektur ergibt komplexe Dynamiken, aber kein Bewusstsein. Architektur ohne Kritikalität ergibt ein ruhendes System – die Modelle existieren im Substrat, aber die

Simulation läuft nicht. Beide Schwellen müssen überschritten sein. Zusammen sind sie hinreichend.

Der kortikale Automat

Zeit, etwas greifbar zu machen, das sich vielleicht noch abstrakt anfühlt. Bisher war die Rede davon, dass der Kortex am Rand des Chaos, in Klasse-4-Dynamiken arbeiten muss. Aber was *ist* dieses Klasse-4-System? Keine mysteriöse Kraft, die über dem Gehirn schwebt. Es ist das Muster neuronaler Aktivität selbst.

Wie sieht der Kortex im Betrieb tatsächlich aus? Milliarden Neuronen, jedes entweder feuernd oder still, jedes seine Nachbarn über gelernte Verbindungsgewichte beeinflussend. Jedes Neuron ist eine Zelle in einem Zellulären Automaten – nicht metaphorisch, sondern buchstäblich. Die Regeln des Automaten sind die synaptischen Gewichte, die Schwellenwerte, die lokale Verdrahtung. Der Output jeder „Zelle" ist eine Feuerrate. Und das Ergebnis – das große Muster elektrischer Aktivität, das mit 10 bis 40 Hz über die kortikale Oberfläche tanzt – ist ein Wolfram-Klasse-4-Zellulärer-Automat in einem Raum von vielen tausend Dimensionen.

Das ist der **kortikale Automat**.

Im Grunde dieselbe Idee, die ich als Kind auf einem 286er programmiert habe (Conways Game of Life) – nur dass es statt eines flachen Gitters mit drei Regeln ein gefaltetes Stück Kortex mit Milliarden lokal variierender Regeln ist, und statt sich in zwei Dimensionen zu bewegen, wandern seine Muster durch einen Dimensionsraum, der so riesig ist, dass jede Vorstellungskraft scheitert. Wie ein Oktopus mit grenzenlosen Armen kann der kortikale Automat jeden Teil des Kortex jederzeit erreichen und aktiviert dabei, welche gespeicherten Modelle er gerade braucht – eine Erinnerung hier, einen motorischen Plan dort, ein Sprachfragment irgendwo anders. Er greift diese Modelle wie kleine Legofiguren und nutzt sie, um von einem befriedigenden Zustand zum nächsten zu navigieren.

Und hier kommt die entscheidende Unterscheidung: **Der kortikale Automat ist nicht Bewusstsein.** Er ist die Maschine, nicht die Erfahrung. Das scheinbar chaotische Muster von Milliarden feuernder Neuronen ist in Wirklichkeit ein außergewöhnlich ausgeklügelter Apparat, der berechnet, denkt und einen Körper durch ein Leben steuert. Aber Bewusstsein ist nur ein *Effekt* dieses Apparats – ein Effekt, der aus dem Zusammenspiel zwischen Automat und Kortex entsteht, wenn die Bedingungen stimmen. Wenn der Automat synchron über geeignete kortikale Regionen mit der richtigen Frequenz in einer kohärenten zeitlichen Abfolge fegt, entsteht bewusste Erfahrung aus dieser Sequenz von Frames. Der Automat enthält die laufenden Instanzen unseres Weltmodells und unseres Selbstmodells; Bewusstsein ist das, was passiert, wenn diese Modelle aktiv in der Simulation laufen.

Den kortikalen Automaten kann man übrigens direkt beobachten – kein fMRT nötig.

Zum Ausprobieren: Einen stockfinsteren Raum finden. Augen schließen. Warten, bis etwaige Nachbilder verblassen (das dauert etwa 30 bis 60 Sekunden, wenn vorher etwas Helles im Blickfeld war). Zuerst ist nichts oder fast nichts zu sehen. Aber dann, mit etwas Geduld, tauchen flackernde farbige Punkte vor der Dunkelheit auf.

Die meisten Leute tun das als „Augenrauschen“ ab – zufälliges Feuern der Photorezeptoren, ausgelöst durch Druck oder spontane chemische Ereignisse. Und tatsächlich: Wer sanft aufs Augenlid drückt, kann so lokalisierte visuelle Empfindungen erzeugen. Aber die farbigen Punkte, die in totaler Dunkelheit erscheinen, sind *nicht* retinalen Ursprungs. Dafür sind sie zu geordnet. Was man hier sieht, ist die Ruheaktivität von V1 (dem primären visuellen Kortex), gespeist aus einer Mischung von residualen sensorischen Signalen und Top-down-Projektionen des kortikalen Automaten selbst. Der Automat läuft in seinen Grunddynamiken, und wir schauen ihm dabei in Echtzeit zu.

Wer weiter zuschaut – nicht konzentriert, sondern entspannt, die Aufmerksamkeit weich werden lässt – erlebt etwas Bemerkenswertes. Aktiver Fokus unterdrückt diese Muster; erst wenn man aufhört sehen zu *wollen*, fängt das Sehen an. Der Automat beginnt, mehr vom visuellen System zu rekrutieren, um zu deuten und zu verstärken, was auch immer an Signal da ist. Die flackernden Punkte stabilisieren sich zu Formen. Geometrische Muster tauchen auf: Gitter, Spiralen, Geflechte. Dann Gesichter, verzerrt und sich verschiebend. Dann Figuren. Dann, mit genug Geduld (und ich meine *Stunden*, nicht Minuten), volle Szenen – aufwendige, farbige, narrative Halluzinationen, die sich von den Träumen jeder Nacht nicht unterscheiden.

Derselbe Mechanismus steckt hinter hypnagogen Halluzinationen – den lebhaften Bildern, die durch den Kopf flackern, gerade wenn der Schlaf kommt. Der kortikale Automat läuft mit minimaler äußerer Einschränkung und erzeugt seinen eigenen Inhalt, indem er gespeicherte Muster aktiviert und in die Simulation projiziert. Der Weg vom schwachen Rauschen zur kohärenten Halluzination ist ein direktes Fenster in die Arbeitsweise des Automaten: Er beginnt bei V1, der frühesten visuellen Verarbeitungsstufe, und zieht nach und nach V2, V3 und höhere Areale hinzu, während er versucht, aus dem vorhandenen Signal Sinn zu machen. Wenn kein echtes Signal da ist, *erfindet* er eines. Das ist das Permeabilitätsleck in Aktion. Ohne äußeres Signal, das die Simulation dominiert, wird das Verarbeitungsrauschen des Substrats selbst sichtbar. Was dann erscheint, sind keine Halluzinationen aus dem Nichts – es sind die Leerlaufmuster der Grafik-Engine, das neuronale Äquivalent von Schnee auf einem schlecht eingestellten Fernseher. Nur dass dieser Schnee Struktur hat, weil die Verarbeitungsmaschinerie Struktur hat.

Auf diese Weise lässt sich auch eine vorübergehende Form von Synästhesie auslösen. In meiner Jugend nutzte ich das, um „Musik zu sehen“. Wer die Augen schließt und Musik hört, während er die visuellen Muster zulässt – entspannt, passiv, ohne angestrengt

sehen zu wollen –, erlebt, wie sich die Muster allmählich mit dem Rhythmus und den Frequenzen der Musik synchronisieren. Der kortikale Automat, bar jeden äußeren visuellen Inputs, beginnt seine visuellen Dynamiken an das stärkste verfügbare Signal zu koppeln – in diesem Fall den auditiven Input. Was dabei sichtbar wird, ist ganz wörtlich die Aktivität des eigenen Gehirns: die V1-Muster des Automaten, angetrieben vom auditorischen Kortex statt von der Netzhaut. Echte Synästhetiker – Menschen, deren Sinne dauerhaft kreuzverkabelt sind, die immer Farben sehen, wenn sie Klänge hören – haben vermutlich eine dauerhaftere Version derselben Kopplung, wahrscheinlich wegen stärkerer oder zahlreicherer Verbindungen zwischen sensorischen Arealen, ob im Thalamus oder im Kortex selbst. Der Mechanismus ist derselbe: eine Sinnesmodalität leckt in die Verarbeitungspipeline einer anderen. Dem kortikalen Automaten ist es ziemlich egal, woher sein Input kommt. Er verarbeitet, was er kriegt.

Als regelmäßiges Hobby ist das nicht zu empfehlen. Die Erfahrung kann verstörend sein, besonders ohne Vorbereitung. Und es besteht ein kleines Risiko, dass anhaltender Reizentzug jemanden mit latenten psychiatrischen Anfälligkeiten destabilisiert. Aber wer sich je gefragt hat, wie das Substrat des eigenen Bewusstseins aussieht, wenn es im Leerlauf ist – wenn die Außenwelt verstummt ist und das System einfach… läuft – das ist der direkteste Blick, der ohne Hirnscanner zu haben ist.

Diese Progression von fast-nichts zu einer kompletten fiktionalen visuellen Welt, erlebt vom eigenen Selbstmodell in einem virtuellen Universum, ist ein direktes Porträt des kortikalen Automaten bei der Arbeit.

Auch Fehlfunktionen des Automaten lassen sich beobachten. Ein epileptischer Anfall ist, was passiert, wenn Teile des Automaten in Klasse-1- oder -2-Dynamiken kippen (periodisch, starr, rechnerisch nutzlos) oder über Klasse 4 hinaus in Klasse-5-Chaos geraten. Ein Schlaganfall ist, was passiert, wenn Teile des Kortex komplett ausfallen. Eine Ohnmacht ist, was passiert, wenn die Mindestfrequenz

für Wachheit nicht mehr erreicht wird. Der Automat ist fragil. Aber die Struktur, die ihn erzeugt – der Neokortex mit seinen gelernten Gewichten und seiner evolvierten Architektur – ist robust. Deshalb erholen wir uns von diesen Störungen so erstaunlich gut.

Die Konvergenz

2003 – zwei Jahre bevor die Theorie überhaupt existierte – entdeckten John Beggs und Dietmar Plenz „neuronale Lawinen" in kortikalem Gewebe: Muster neuronaler Aktivität, die der mathematischen Signatur selbstorganisierter Kritikalität folgten, einem Kennzeichen von Systemen am Rand des Chaos.

2014 stellte Robin Carhart-Harris die Entropic Brain Hypothesis auf: die Idee, dass das Bewusstseinsniveau mit der Entropie (Unordnung) der Gehirnaktivität korreliert, mit dem Optimum auf einem mittleren Niveau – zu wenig Entropie heißt Bewusstlosigkeit, zu viel heißt inkohärente Erfahrung.

2016 zeigten Enzo Tagliazucchi und Kollegen, dass LSD das Gehirn in Richtung Kritikalität verschiebt – passend zum verstärkten (aber manchmal chaotischen) Bewusstsein, das Psychedelika-Nutzer berichten. Bis 2022 konnte ein Übersichtsartikel bereits von „selbstorganisierter Kritikalität als Rahmen für Bewusstsein" sprechen – die Belege häuften sich.

Und 2025-2026 brach der Damm. Keith Hengen und Woodrow Shew veröffentlichten eine Meta-Analyse von 140 Datensätzen in *Neuron* (2025) – die größte systematische Analyse von Kritikalität in Gehirndynamiken, die je durchgeführt wurde – und bestätigten, dass das Gehirn über verschiedene Messmethoden hinweg nahe einem kritischen Punkt arbeitet. Dann stellten Inbal Algom und Oren Shriki das ConCrit-Framework (Consciousness and Criticality) in *Neuroscience & Biobehavioral Reviews* (2026) vor und argumentierten, dass kritische Gehirndynamiken eine vereinheitlichende mechanistische Grundlage für alle großen Bewusstseinstheorien liefern. Ihr Fazit: Bewusstsein folgt Kritikalität. Wenn das Gehirn

am oder nahe dem kritischen Punkt arbeitet, ist Bewusstsein da. Wenn es unter Kritikalität gedrückt wird (durch Narkose, durch Schlaf, durch Hirnschaden), ist Bewusstsein weg. Wenn es über Kritikalität hinausgeschoben wird (durch Anfälle, vielleicht durch bestimmte Drogenzustände), wird Bewusstsein inkohärent.

Zwei Pfade. Einer theoretisch, ausgehend von Wolframs Berechnungsrahmen und Überlegungen, was eine Selbst-Simulation erfordert. Einer empirisch, ausgehend von neuronalen Aufzeichnungen und der Analyse statistischer Eigenschaften von Gehirnaktivität über jeden zugänglichen Bewusstseinszustand hinweg. Zwei Jahrzehnte auseinander im Ursprung, konvergierend auf dasselbe Ergebnis.

Das ist die Art Konvergenz, die eine Theorie ernst nehmen lässt.

Drei Arten, wie ein Hologramm einem Automaten begegnet

Beim Schreiben dieses Kapitels fiel mir etwas auf, das mich kalt erwischte.

Das holografische Prinzip und Klasse-4-Automaten tauchen immer wieder in denselben Diskussionen auf – in der Physik, in der Neurowissenschaft, in der Berechnungstheorie. Aber niemand scheint die naheliegende Frage gestellt zu haben: *Welche Beziehungen sind zwischen ihnen möglich?*

Es gibt genau drei.

Beziehung 1: Ein holografisches Substrat produziert Klasse-4-Dynamiken. Das ist vermutlich, was das Gehirn tut. Neuronale Netzwerke sind lokal holografisch – Karl Lashley zeigte vor Jahrzehnten, dass man große Teile des Kortex zerstören kann und die Erinnerungen bleiben erhalten, degradiert, aber vollständig, genau wie ein halbiertes Hologramm das ganze Bild in niedrigerer Auflösung liefert. Und dieses holografische Substrat produziert, an Kritikalität arbeitend, die Klasse-4-Dynamiken, die Bewusstsein

verlangt. Gut belegt, gründlich dokumentiert und – man verzeihe mir – die langweilige Variante.

Beziehung 2: Ein Klasse-4-Automat, der holografische Muster als emergentes Verhalten hervorbringt. Der Automat ist nicht holografisch in seinen Regeln, aber seine Dynamiken erzeugen spontan holografische Strukturen – höherdimensionale Information, kodiert in niedrigerdimensionalen Mustern, entstehend aus der Berechnung selbst. Wenn ein Klasse-4-Automat natürlicherweise holografischen Output erzeugt, heißt das: nicht-lokale Informationsverteilung entsteht aus rein lokalen Regeln – was faszinierenderweise genau so aussieht wie Quantenverschränkung.

Hier muss Gerard 't Hooft erwähnt werden, weil die Verbindung zu auffällig ist, um sie zu übergehen – auch wenn sie spekulativ bleibt. 't Hooft, Physik-Nobelpreisträger, hat vorgeschlagen, dass die Quantenmechanik selbst ein Zellulärer Automat auf der Planck-Skala ist: dass unser Universum fundamental deterministisch ist und Quanteneffekte emergente Phänomene einer tieferen, diskreten Dynamik sind. Wenn er recht hat, gilt das beschriebene Prinzip nicht nur für Bewusstsein als Analogie. Es ist buchstäblich, wie das Universum funktioniert, bis ganz nach unten. Einfache lokale Regeln erzeugen ein holografisches Universum, und innerhalb dieses Universums erzeugen einfache neuronale Regeln ein holografisches Bewusstsein. Dasselbe Berechnungsprinzip auf zwei Skalen: kosmologisch und neurologisch. Diese fraktale Konsistenz finde ich zutiefst überzeugend – aber ehrlicherweise bleibt 't Hoofts Interpretation eine Minderheitsmeinung in der Physik, und der Schluss von struktureller Eleganz auf physikalische Realität wurde zu Recht kritisiert. Dennoch: Wenn sich herausstellt, dass ein einzelnes Berechnungsprinzip sowohl dem Universum als auch den Geistern, die es modellieren, zugrunde liegt, wäre das die schönste Tatsache, die je entdeckt wurde.

Beziehung 3: Ein Klasse-4-Automat, dessen Regelstruktur selbst holografisch ist. Das ist die Variante, bei der ich den Stift weglegen musste. Wenn so etwas existiert – ein Zellulärer

Automat, dessen Regeln selbst höherdimensionale Information in einer niedrigerdimensionalen Struktur kodieren, so wie ein Hologramm drei Dimensionen in zwei kodiert –, dann hätte man ein System, das von Natur aus das tut, was laut dem holografischen Prinzip das Universum tut. Nicht ein System, das bloß *auf* einem holografischen Substrat läuft (oder ein Hologramm hervorbringt). Ein System, das *selbst* eine holografische Kodierung *ist*. Möglicherweise auch das Universum – wobei das spekulativ bleibt, und der Schluss von mathematischer Schönheit auf physikalische Realität berechtigter Kritik ausgesetzt ist. Darauf komme ich in Kapitel 13 zurück, wo ich darlegen werde, warum Beziehung 3 möglicherweise die wichtigste ungelöste Frage der Mathematik ist – und dann die Antwort in den Kapiteln 14 bis 16 vollständig verfolgen.

Chapter 6

Was Psychedelika offenbaren

Ein nötiger Hinweis vorweg: Nichts in diesem Kapitel ist als Empfehlung zu verstehen, Psychedelika auszuprobieren. Sie sind mächtig, unberechenbar und können ein Leben zerstören – buchstäblich, dauerhaft. Sie können bei Prädisponierten Schizophrenie auslösen. Sie können psychotische Episoden, anhaltende Angststörungen und HPPD (hallucinogen persisting perception disorder) verursachen, die nie wieder verschwinden. Hier geht es um sie, weil sie etwas Wichtiges über die Architektur des Bewusstseins offenbaren. Dieser wissenschaftliche Wert macht sie nicht sicher.

Wer Bewusstsein verstehen will, untersucht, was passiert, wenn es aus dem Tritt gerät. Psychedelika sind meiner Überzeugung nach das aufschlussreichste Fenster in die Architektur des Bewusstseins – aufschlussreicher als Hirnscans schlafender Patienten, theoretisch ergiebiger als Läsionsstudien und um Welten zugänglicher als Split-Brain-Chirurgie.

Warum? Weil Psychedelika Bewusstsein nicht nur *verändern*. Sie verändern es auf *systematische, vorhersagbare Weise*, die die zugrunde liegende Architektur freilegt – wenn man weiß, worauf man achten muss.

Der Permeabilitätsgradient

Zur Erinnerung: die Grenze zwischen den impliziten und den expliziten Modellen. Im normalen Wachleben ist sie selektiv durchlässig – relevante Information kommt durch, irrelevante bleibt in der Bibliothek. Was gebraucht wird, erreicht das Bewusstsein; alles andere bleibt unbewusst.

Psychedelika sprengen die Grenze auf.

Unter Psychedelika (LSD, Psilocybin, DMT, Meskalin) steigt die Durchlässigkeit der implizit-explizit-Grenze global an. Information, die normalerweise vollständig auf der realen Seite verarbeitet wird, unsichtbar fürs Bewusstsein, beginnt zur Simulation durchzusickern.

Und der entscheidende Punkt: Sie sickert *der Reihe nach* durch.

Bei niedrigen Dosen oder früh in der Erfahrung werden die einfachsten Verarbeitungsstufen zuerst sichtbar – die, die am nächsten am rohen sensorischen Input liegen: V1-Verarbeitung. Intensivere Farben, atmende Muster auf statischen Oberflächen, subtile Bewegungen im peripheren Sichtfeld. Das sind die frühen Merkmalsdetektoren des visuellen Kortex, normalerweise unsichtbar, jetzt in die Simulation eintretend.

Steigt die Dosis oder vertieft sich die Erfahrung, werden komplexere Verarbeitungsstufen sichtbar. V2/V3-Verarbeitung: geometrische Muster, Fraktale, Tessellationen, die berühmten „Formkonstanten“, die Heinrich Klüver in den 1920ern katalogisierte. Das sind die Zwischenrepräsentationen des visuellen Systems – die Bausteine, mit denen es normalerweise die visuelle Erfahrung konstruiert, jetzt für sich allein sichtbar.

Noch höher, und die höheren visuellen Areale werden zugänglich. Gesichter erscheinen. Figuren. Szenen. Die Gesichtserkennungsareale, die Objekterkennungsareale, die Szenenkonstruktionsareale – alle normalerweise unterhalb der Bewusstseinsschwelle arbeitend – schicken jetzt ihre Zwischenprodukte direkt in die Simulation.

Bei den höchsten Dosen liegt die gesamte Verarbeitungshierarchie offen, und das Ergebnis ist vollausgebildete visionäre Erfahrung:

komplexe, narrative, traumähnliche Szenen, aufgebaut aus den tiefsten Schichten impliziter Verarbeitung.

Diese geordnete Stufenfolge – einfach zu komplex, V1 zu höheren Arealen, dosisabhängig – ist genau das, was die Vier-Modelle-Theorie vorhersagt. Eine direkte Folge des Permeabilitätsgradienten: niedrigere Verarbeitungsstufen, näher an der Grenze, werden vor höheren zugänglich, wenn die Durchlässigkeit steigt.

Die folgende Tabelle zeigt die visuelle Verarbeitungshierarchie – was jedes Areal normalerweise tut und was sichtbar wird, wenn die Permeabilitätsbarriere fällt:

Areal	**Normale Funktion**	**Psychedelische Signatur**
V1	Kanten, räumliche Frequenz, Orientierung	Phosphene, Klüver-Formkonstanten, atmende Oberflächen
V2	Konturintegration, Textur, Border-Ownership	Tessellationen, sich wiederholende geometrische Muster
V3	Globale Form, dynamische Formverarbeitung	Fließende, sich wandelnde Geometrien
V4	Farbe, Krümmung, komplexe Textur	Farbige Fraktale, kaleidoskopische Muster
V5/MT	Bewegungsverarbeitung	Rotation und Bewegung von Mustern
Fusiform/IT	Gesichter, Objekte, Wortformen	Gesichter, Figuren, Entitäten
Anterior IT	Semantische Kategorien, Szenenkonstruktion	Volle narrative Halluzinationen

Jede Zeile steht für eine tiefere Verarbeitungsstufe. Unter normalen Bedingungen erlebt man nur den fertigen Output – das fertige Wahrnehmungsbild. Unter Psychedelika werden die *Zwischenstufen* erlebbar, der Reihe nach, mit steigender

Durchlässigkeit. (Eine ausführlichere Fassung dieser Tabelle mit rezeptiven Feldgrößen und zusätzlichen Details findet sich in Anhang A.)

Ja, das klingt faszinierend. Wer über Schichten visueller Verarbeitung liest, die plötzlich sichtbar werden, wird unweigerlich neugierig, wie das aussieht. Verständlicherweise – mir ging es genauso. Ich probierte beide Wege. Ich war jung und dumm und hatte Glück. Die Meditationsroute aus dem vorigen Kapitel (dunkler Raum, entspannte Aufmerksamkeit, Geduld) führt zum selben Ort. Nicht so schnell, nicht so dramatisch beim ersten Mal. Aber genauso eindrucksvoll, genauso real – und ohne das Risiko, den Geist dauerhaft zu beschädigen. Ein warmes Bett in einem dunklen Raum genügt.

Und es gibt noch einen Weg: luzides Träumen. Wer lernt, im Traum zu erkennen, dass er träumt – und das ist eine trainierbare Fähigkeit –, bekommt Zugang zur vollen Simulation im Freilauf. Kein sensorischer Input, keine äußere Realität, die das Modell korrigiert. Nur die virtuelle Welt, mit dem eigenen Bewusstsein mittendrin. Für manche ist das leichter zu erreichen als ausdauernde Meditation. Die Techniken sind gut dokumentiert (siehe Anhang D), und die Erfahrung kann mindestens so augenöffnend sein wie alles, was eine Droge liefert – ohne das Risiko. Auf luzides Träumen kommen wir in Kapitel 7 zurück.

Hier leistet die Fünf-Level-Hierarchie aus Kapitel 2 ihre Erklärungsarbeit. Zur Erinnerung: die fünf verschachtelten Systeme (Physikalisch, Elektrochemisch, Proteomisch, Topologisch, Virtuell). Psychedelika greifen in der Mitte des Stapels an, und die Wirkung breitet sich nach oben aus. Klassische Psychedelika wie LSD und Psilocybin binden an Serotonin-2A-Rezeptoren und wirken auf dem **elektrochemischen** Level – sie verändern, wie Neuronen miteinander kommunizieren. Diese Störung breitet sich zum **proteomischen** Level aus, wo sich die Rezeptorempfindlichkeit über Stunden verschiebt. Sie verändert das **topologische** Level, wo sich Netzwerkverbindungsmuster umgestalten – im fMRT

sichtbar als erhöhte globale Integration. Und sie transformiert das **virtuelle** Level, wo die bewusste Simulation mit Inhalt überflutet wird, der normalerweise unsichtbar bleibt. Das einzige Level, das klassische Psychedelika nicht berühren, ist das **physikalische** – sie zerstören keine Neuronen, greifen nicht die Materie an. Sie ändern alles *über* der Materie, in aufsteigender Reihenfolge. Das ist eine entscheidende Unterscheidung. Klassische Psychedelika (LSD, Psilocybin, DMT, Meskalin) sind nicht neurotoxisch. Sie ändern, wie Neuronen kommunizieren, ohne sie zu zerstören. Viele andere Drogen sind nicht so gnädig. Kokain, Methamphetamin und Alkohol zerstören Neuronen physisch. MDMA in hohen oder wiederholten Dosen schädigt Serotonin-Axone. Sogar Amanita muscaria – der ikonische rot-weiße Pilz, den viele mit psychedelischen Pilzen verwechseln – ist ein Deliriant mit einem völlig anderen, gefährlicheren Wirkmechanismus. Wenn man aus diesem Kapitel nur eines mitnimmt: Nicht alle Drogen, die Bewusstsein verändern, sind gleich, und der Unterschied zwischen „ändert das Signal“ und „zerstört die Hardware“ ist buchstäblich der Unterschied zwischen einem vorübergehend veränderten Zustand und dauerhaftem Hirnschaden. Die dosisabhängige visuelle Stufenfolge bildet sich direkt darauf ab: Niedrige Dosen stören das elektrochemische Level gerade genug, um die V1-Verarbeitung zu beeinflussen; höhere Dosen treiben die Störung durch weitere Level und ziehen zunehmend komplexe Verarbeitungsstufen ins bewusste Erleben.

Das umleitbare Selbst

Der dramatischste Beweis aber kommt von dem, was mit dem Selbst geschieht.

Das Explizite Selbstmodell (ESM) ist ein virtueller Prozess, der auf Input angewiesen ist. Unter normalen Bedingungen empfängt es einen stetigen Strom selbstbezogener Signale: das Empfinden, wo der Körper ist (Propriozeption), das Empfinden,

wie die Organe sich anfühlen (Interozeption), den narrativen Strom innerer Sprache und den ständigen Hintergrund körperlicher Selbstwahrnehmung, der nie bemerkt wird, bis er gestört wird.

Bei hohen psychedelischen Dosen wird dieser Input unterbrochen. Das Selbstmodell stirbt nicht – es *leitet um*. Beraubt seines normalen selbstbezogenen Inputs, klammert es sich an den jeweils dominanten Input.

Am drastischsten zeigt das Salvia divinorum, ein dissoziatives Psychedelikum, das auf Kappa-Opioid-Rezeptoren wirkt (völlig verschieden von den serotonergen Mechanismen von LSD oder Psilocybin). Salvia-Nutzer berichten durchgehend davon, zu Dingen *geworden* zu sein:

- „Ich wurde die Couch."
- „Ich war die Wand."
- „Ich verwandelte mich in eine Seite in einem Buch."
- „Ich war eine der Figuren im Fernseher."
- „Ich wurde ein Fraktal, nicht ein Fraktal sehend, ein Fraktal *seiend*."

Das sind keine Metaphern. Die Berichte beschreiben vollständige, erlebnismäßig überzeugende Identitätsverschiebungen. Für die Dauer der Erfahrung *sind* die Betroffenen das Objekt oder die Entität. Manche beschreiben es so, als fühle es sich an wie tot zu sein – nicht zu sterben, sondern tot zu *sein* –, denn wer ein Stuhl ist, hat als Person schlicht aufgehört zu existieren.

Der Inhalt folgt der sensorischen Umgebung. Wer fernsieht, wird eine Fernsehfigur. Wer auf einer Couch liegt, wird die Couch. Wer ein Muster anschaut, wird das Muster.

Das ist das Explizite Selbstmodell, das genau das tut, was die Theorie vorhersagt: Es leitet auf den jeweils dominanten Input um, sobald der normale Selbst-Input ausfällt. Der Identitätsinhalt ist nicht zufällig – er wird durch die sensorische Umgebung

bestimmt. Kontrolliert man die Umgebung, müsste sich die Identitätserfahrung kontrollieren lassen.

Hier muss die Theorie kurz pausieren. Salvia divinorum ist, soweit bekannt, die stärkste psychedelische Substanz auf der Erde. Die komplette propriozeptive Übernahme, die gerade beschrieben wurde, bedeutet totalen Verlust von Körperwahrnehmung und räumlicher Orientierung. Menschen unter Salvias Einfluss sind aus dem zehnten Stock aus Fenstern gelaufen. Sie sind in den Verkehr getreten. Sie sind gestorben. Das ist keine Party-Droge, keine Freitagabend-Kuriosität. Es ist die extremste pharmakologische Zerstörung des Expliziten Selbstmodells, die existiert, und diese Zerstörung kann töten – nicht weil die Droge giftig ist, sondern weil das Wissen, wo der eigene Körper sich befindet, vollständig erlischt – und die felsenfeste Überzeugung einsetzen kann, Flügel zu haben und fliegen zu können.

Viele, die Salvia probieren, berichten, die Erfahrung habe sich angefühlt wie Sterben – nicht metaphorisch, sondern als echte, lähmende Überzeugung, nicht mehr zu existieren. Das ist das Explizite Selbstmodell, das so vollständig kollabiert, dass die Simulation überhaupt kein „Ich“ mehr erzeugen kann. Das klinische Gegenstück dazu kommt in Kapitel 8: der Cotard-Wahn – Patienten, die neurologisch davon überzeugt sind, tot zu sein. Salvia bringt einen pharmakologisch dahin, in Sekunden, ohne Vorwarnung. Ob das eine Erfahrung ist, die man machen will, muss jeder selbst entscheiden.

Die Zeitdehnung erlebte ich selbst. Unter Salvia wurde eine halbe Sekunde Echtzeit – bestätigt durch die Person, die mich beobachtete – zu etwas, das sich wie fünfzehn Minuten oder mehr anfühlte. Meine Wahrnehmungswelt baute sich in aufwendige Sequenzen um, einschließlich des Gefühls, Flügel zu haben und herumzufliegen (das Fluggefühl, wurde mir später klar, kam von Luft, die an mir vorbeiströmte, als ich rückwärts aufs Bett fiel). Meine gesamte Realität kollabierte und regenerierte sich, alles in der Zeit, die ein Blinzeln braucht. Ein Beobachter, der mich

abfing, sagte, ich sei für weniger als eine Sekunde „weg“ gewesen. Dieselbe Art Zeitdehnung, die ich schon ein paar Jahre zuvor erlebt hatte, 1998 oder 1999, während eines Nahtodereignisses (den Mechanismus beschreibe ich in Kapitel 13) – aber pharmakologisch ausgelöst und noch extremer.

Ich bin nicht der dramatischste Fall. Ein gut dokumentierter Bericht handelt von einem Mann, der erlebte, was sich wie acht volle Jahre eines alternativen Lebens anfühlte – Schulbesuch, Freundschaften, der Aufbau einer neuen Existenz –, während einer Salvia-Episode von ungefähr fünfundvierzig Sekunden Uhrenzeit. Peer-reviewte Forschung bestätigt extreme zeitliche Verzerrung unter kontrollierten Bedingungen; ein Teilnehmer beschrieb die Zeit als „gefaltet wie ein Akkordeon“ (Addy et al., 2015). Das Substrat schleust so viel Inhalt so schnell durch die Simulation, dass sich subjektive Zeit vollständig von der Uhrenzeit entkoppelt.

Experimentell getestet wurde das in einem kontrollierten Rahmen bisher nie. Aber es wäre möglich – und es wäre eine schlagende Bestätigung des unverwechselbarsten Mechanismus der Theorie.

Um zu sehen, wie weit dieses Prinzip reicht, ein Gedankenexperiment: Jemand wird dauerhaft auf einer sehr hohen (aber nicht vollständig dissoziierenden) Dosis von Salvinorin A gehalten – dem Wirkstoff in Salvia divinorum, der an einem einzigen Rezeptortyp angreift (Kappa-Opioid). Das Explizite Selbstmodell dieser Person würde sich nie stabilisieren. Es würde endlos durch den jeweils dominanten Input kreisen: den einen Moment die Überzeugung, ein Stuhl zu sein, dann ein Tisch, dann ein Dinosaurier, dann Luft, dann ein Stück Papier. Erlebt würde weiterhin (Sehen und Hören würden noch funktionieren), aber das Wissen, wer oder was man ist, wäre verloren. Nimmt man die Droge weg, würde sich mit der Zeit das normale Selbstmodell vom intakten Impliziten Selbstmodell neu zusammensetzen.

Das ist wichtig, weil es zeigt: Bewusstsein braucht kein *korrektes* Selbstmodell. Es braucht nur *ein* Selbstmodell. Die Architektur

läuft weiter, so oder so. Das Explizite Selbstmodell schaltet sich nicht ab, wenn es absurden Input bekommt – es baut das beste Selbst, das es kann, aus den verfügbaren Signalen. Dasselbe Prinzip zeigt sich beim Cotard-Wahn (das ESM arbeitet mit fehlenden interozeptiven Signalen: „Ich muss tot sein“), beim Anton-Syndrom (das ESM generiert Sehen aus Erinnerung, wenn die Augen versagen) und bei der Konversionsstörung (das ESM modelliert eine Lähmung, die das Substrat gar nicht hat). Das Selbstmodell ist ein zwanghafter Konstrukteur. Es hört nie auf zu bauen. Es erklärt nie, die Daten seien unzureichend. Es baut einfach und glaubt.

Anosognosie: Die Inverse

Hier zeigt sich eine schöne Symmetrie. Wenn Psychedelika das sind, was passiert, wenn die implizit-explizit-Grenze *zu* durchlässig wird, ist Anosognosie das, was passiert, wenn sie *zu* undurchlässig wird – zumindest lokal.

Anosognosie tritt am häufigsten nach Schlaganfällen der rechten Hemisphäre auf: Patienten, die ihre eigenen Ausfälle schlicht nicht wahrnehmen. Ein Patient mit einem gelähmten linken Arm besteht darauf, dass der Arm in Ordnung ist, erklärt vergebliche Bewegungsversuche weg und reagiert verwirrt oder wütend, wenn man ihm die Lähmung vorführt. Diese Patienten leugnen nicht im psychologischen Sinn – die Information, dass der Arm gelähmt ist, erreicht schlicht nie ihre bewusste Simulation.

In der Vier-Modelle-Theorie ist das eine lokale Verringerung der implizit-explizit-Durchlässigkeit. Das Implizite Selbstmodell *hat* die Information über die Lähmung – das Substrat registriert den Schaden. Aber die Grenze ist für diese spezifische Domäne blockiert, also enthält das Explizite Weltmodell den Ausfall nie. In der Simulation des Patienten gibt es keinen gelähmten Arm, also erlebt der Patient keinen.

Der Mechanismus ist noch spezifischer, und wenn man ihn versteht, ist er auf eine leicht beunruhigende Weise elegant. Wenn das Motorsystem einen Befehl sendet (sagen wir „Hände klatschen"), tut es gleichzeitig zwei Dinge. Es sendet den Befehl an die Muskeln, und es sendet *vorhergesagtes Feedback* ans Bewusstsein: wie sich Klatschen anfühlen und anhören sollte, basierend auf Erfahrung. Dieses vorhergesagte Feedback kommt *vor* dem tatsächlichen sensorischen Feedback an, weil das echte Feedback langsamere neuronale Wege nehmen muss. Normalerweise wird die Vorhersage rasch durch die tatsächlichen sensorischen Daten korrigiert oder bestätigt. Das Klatschen wird vorhergesagt, dann gefühlt und gehört. Passt. Weiter.

Bei Anosognosie kommt das tatsächliche Feedback vom gelähmten Glied nie an. Und der Mechanismus, der „Moment – nichts ist passiert" signalisieren sollte, ist beschädigt. Also geht das vorhergesagte Feedback unkorrigiert durch. Das Motorsystem des Patienten befiehlt beiden Händen zu klatschen, sendet die Vorhersage eines beidhändigen Klatschens ans Bewusstsein, und das Bewusstsein erlebt genau das – ein völlig normales Klatschen mit beiden Händen. Der Patient wird in voller Aufrichtigkeit sagen, er habe gerade mit beiden Händen geklatscht. Er hörte es. Er fühlte es. Er erlebte es. In seiner Simulation passierte es. Nur in der Wirklichkeit nicht.

So *funktioniert* Bewusstsein, die ganze Zeit, bei uns allen. Der einzige Unterschied: Bei gesunden Menschen wird das vorhergesagte Feedback innerhalb von Millisekunden korrigiert. Bei Anosognosie ist der Korrekturmechanismus kaputt, und die Simulation des Patienten läuft allein auf Vorhersagen.

Psychedelika und Anosognosie sind derselbe Mechanismus in entgegengesetzten Richtungen. Das eine erhöht Durchlässigkeit global. Das andere verringert sie lokal. Und diese Symmetrie erzeugt eine bereichsübergreifende Vorhersage: Psychedelika sollten Anosognosie lindern. Die globale Durchlässigkeitserhöhung

sollte die lokale Blockade überwältigen und der Information über den Ausfall erlauben, das Bewusstsein zu erreichen.

Niemand hat das je getestet, weil niemand eine Theorie hatte, die diese beiden Phänomene verbindet. Die Verbindung ist ohne die Vier-Modelle-Theorie unsichtbar.

Chapter 7

Was passiert, wenn die Lichter ausgehen

Jede Nacht geht das Bewusstsein verloren. Jeden Morgen kehrt es zurück. Und der Übergang dazwischen – die Reise durch die Schlafstadien – ist eine nächtliche Vorführung des Kritikalitätsprinzips.

Tiefschlaf: Unterhalb der Schwelle

Im tiefen Non-REM-Schlaf rutscht die Gehirndynamik in den subkritischen Bereich. Das Kennzeichen sind langsame Wellen: große, synchronisierte Oszillationen, bei denen riesige Neuronenpopulationen gemeinsam feuern und dann gemeinsam verstummen. Das ist Klasse-2-Dynamik – periodisch, repetitiv, zu geordnet für Bewusstsein.

Der Perturbational Complexity Index (PCI), entwickelt von Marcello Massimini und Kollegen, bestätigt das unmittelbar. PCI misst, wie komplex das Gehirn auf einen magnetischen Impuls reagiert: Im Wachzustand ist die Antwort komplex und differenziert (hoher PCI); im Tiefschlaf einfach und stereotyp (niedriger PCI). Das schlafende Gehirn kann die reichhaltige, global vernetzte Dynamik nicht aufrechterhalten, die eine bewusste Simulation braucht.

Die Lichter sind aus. Das Explizite Weltmodell (EWM) und das Explizite Selbstmodell (ESM) sind zusammengebrochen. Keine Simulation, keine Erfahrung.

Träume: Sparmodus

Im REM-Schlaf gehen die Lichter wieder an. Die Gehirndynamik schiebt sich zurück Richtung Kritikalität – nicht ganz, aber nah genug. Die Simulation springt wieder an, und wieder wird eine Welt erlebt.

Allerdings eine Simulation im Sparmodus. Der normale äußere Input fehlt (Augen zu, Muskeln gelähmt). Das Explizite Weltmodell läuft auf internen Daten – es greift auf das gespeicherte Wissen des Impliziten Weltmodells (IWM) zurück statt auf aktuelle Sinneseindrücke. Deshalb zeigen Träume vertraute Orte und Menschen, aber mit unmöglicher Physik und erzählerischer Inkohärenz: Die Simulation gibt ihr Bestes mit dem, was sie hat.

Auch das Explizite Selbstmodell läuft im Sparmodus. Träume werden als etwas erlebt, das „einem“ passiert, aber die metakognitive Überwachung ist gedrosselt – unmögliche Ereignisse werden fraglos akzeptiert, man bemerkt selten, dass man träumt, das kritische Denken ist gedämpft.

Wie sich diese heruntergefahrene Simulation von innen anfühlt, kenne ich gut. Zwischen etwa sieben und zwölf Jahren hatte ich einen wiederkehrenden Traum – denselben Traum, der über diese Jahre immer wiederkam. Es war eine Landschaft, wenn man es so nennen will: eine unendlich ausgedehnte fraktale Struktur mit Tälern, die endlos in die Tiefe stürzten, selbstähnlich auf jeder Ebene, in die man blickte. Dazu ein Gefühl, das mir im Wachzustand nie begegnet ist – eine tiefe Entkörperung, als hätte man mir den Körper weggenommen und mich in die Geometrie selbst hineinversetzt. Der Traum war unheimlich, befremdlich,

aber auch seltsam faszinierend. Ich wollte, dass er wiederkam, selbst wenn er mich erschreckte.

Das Bemerkenswerte daran: Ich kann es immer noch *fühlen*. Jahrzehnte später, wenn ich an diesen Traum denke, kehrt dieselbe entkörperte Empfindung zurück – nicht als Erinnerung an ein Gefühl, sondern als das Gefühl selbst. Die implizite Kodierung steckt immer noch in den synaptischen Gewichten, unberührt von dreißig Jahren Wacherfahrung. Das Substrat hat sie so tief gespeichert, dass der bloße Gedanke daran genügt, um die Simulation teilweise zu reaktivieren.

Was tat der Traum? In den Begriffen der Theorie: Das Explizite Weltmodell, abgeschnitten vom äußeren Input, erzeugte Inhalte aus der eigenen Rechendynamik des Substrats. Und was ist diese Dynamik? Klasse 4 – die Klasse 3 (fraktale Struktur) als Teilprozess enthält. Ein schlafendes Gehirn, dessen Simulation nur auf der eigenen Architektur läuft, produzierte die visuelle Signatur dieser Architektur: Fraktale. Die Entkörperung war das Explizite Selbstmodell in seiner am stärksten degradierten Form – die Simulation wusste, dass sie *jemand* war, irgendwo, aber die Körperrepräsentation war fast vollständig ausgefallen.

Ich habe nie versucht, diesen Traum heraufzubeschwören, nachdem ich Jahre später das luzide Träumen gelernt hatte. Vielleicht sollte ich es.

Schlafwandeln ist eine noch dramatischere Demonstration. Beim Schlafwandeln reaktiviert sich das motorische System teilweise, während das Explizite Selbstmodell offline oder nahezu offline bleibt. Das Substrat führt motorische Programme aus – Gehen, Navigieren, sogar komplexe Handlungen –, aber die Simulation läuft nicht mit. Der Schlafwandler bewegt sich durch die physische Welt, geleitet vom räumlichen Wissen des Impliziten Weltmodells, aber ohne bewusste Erfahrung – oder fast ohne.

Ich weiß das aus erster Hand. Als Teenager durchlief ich eine Phase des Schlafwandelns. Eines Morgens wachte ich an meinem Schreibtisch auf, vor mir hingekrakelte Notizen – linkshändig

geschrieben, was ich wach nie tue. Ein fragmentarisches Bild war noch da: im Kreis an den Wänden entlang, die Tür suchend, sie nicht findend. Aber der Teil, in dem ich mich an den Schreibtisch setzte und zu schreiben versuchte – vollständig dunkel. Das Substrat navigierte, motorische Programme liefen, aber die Simulation – das „Ich“ – war nicht da.

Das ist die Theorie im Kleinen. Ein Körper, der sich durch die Welt bewegt, räumliche Informationen verarbeitet, gelernte motorische Programme ausführt – alles ohne ein bewusstes Selbst in der Schleife. Die impliziten Modelle führen Regie. Die expliziten Modelle sind offline. Und das Ergebnis ist ein Mensch, der geht, handelt und sogar schreibt – aber niemand ist zu Hause.

Luzides Träumen: Der Schalter

Und dann gibt es das luzide Träumen – den Zustand, in dem man mitten im Traum begreift, dass man träumt. In der Vier-Modelle-Theorie schaltet sich hier das Explizite Selbstmodell innerhalb des Traumzustands vollständiger ein. Eine sprungartige Zunahme der Selbstmodellierung.

Luzides Träumen lernen heißt, die Simulation beim Schummeln erwischen. Die Methode, die ich benutzte, heißt Lichtschalter-Test: Man betätigt den ganzen Tag über gewohnheitsmäßig Lichtschalter und fragt sich, ob das Licht sich korrekt verändert hat. Im Wachleben tut es das immer. Im Traum funktionieren Lichtschalter nicht – das Explizite Weltmodell, das auf internen Daten läuft, verschwendet keine Rechenzeit auf die Physik elektrischer Schaltkreise. Wenn man im Traum einen Schalter betätigt und das Licht sich nicht verändert, oder der Raum dunkler wird, oder der Schalter sich falsch anfühlt – diese Diskrepanz ist das Explizite Selbstmodell, das eine Inkonsistenz in der Simulation erkennt. Und in dem Moment der Erkennung weiß man: Das hier ist ein Traum.

Bei mir dauerte es nur ein paar Tage. Glück gehabt – die meisten brauchen Wochen oder Monate, bevor die Gewohnheit in ihre Träume übergeht. Aber als es funktionierte, war der Übergang genau die sprunghafte Schwelle, die die Theorie vorhersagt. Einen Moment lang war ich passiver Statist in der Erzählung des Traums. Im nächsten voll präsent – mir bewusst, dass ich träumte, mir bewusst, dass die Welt um mich herum generiert war, mir bewusst, dass ich sie verändern konnte. Das ESM hatte sich zugeschaltet. Am Substrat hatte sich nichts geändert. An der Dynamik hatte sich nichts geändert. Aber die Einbindung des Selbstmodells in die Simulation hatte eine Schwelle überschritten, und alles fühlte sich anders an.

Die Theorie sagt vorher, dass dieser Übergang einer Überschreitung der Kritikalitätsschwelle entspricht. Kein graduelles Anschwellen der Gehirnkomplexität, sondern ein plötzlicher Sprung. Mäße man die EEG-Komplexität in einem engen Zeitfenster um den Moment des Luzidität-Einsetzens – mit dem etablierten Paradigma vorher vereinbarter Augenbewegungssignale luzider Träumer –, müsste sich eine Diskontinuität zeigen.

Anästhesie: Die zwei Typen

Anästhesie liefert den saubersten Test des Kritikalitätsprinzips, weil verschiedene Anästhetika dramatisch unterschiedliche Erfahrungen produzieren, obwohl sie unter demselben Etikett laufen.

Propofol drückt das Gehirn subkritisch. Die thalamokortikale Konnektivität bricht ab, die kortikale Komplexität kollabiert, und der PCI nähert sich null. Die Lichter gehen vollständig aus. Patienten berichten von keinerlei Erfahrung unter Propofol. Genau das sagt die Theorie voraus: Fällt die Dynamik unter die Kritikalität, kann die Simulation nicht aufrechterhalten werden.

Ketamin macht etwas völlig anderes. Es drückt das Gehirn *nicht* subkritisch. EEG-Studien zeigen, dass Ketamin die neuronale Entropie *erhöht* – es schiebt das Gehirn an die Kritikalität heran

oder darüber hinaus, in einen chaotischeren Bereich. Das Ergebnis? Das „K-Hole“ – lebhafte, oft bizarre Erfahrungen von Dissoziation, verzerrter Realität, Erlebnissen außerhalb des eigenen Körpers und radikaler Identitätsauflösung.

In der Vier-Modelle-Theorie ist das K-Hole Bewusstsein, das auf *falschem* Input läuft. Das Explizite Weltmodell und das Explizite Selbstmodell sind noch aktiv (das Gehirn ist noch bei oder über der Kritikalität), aber die sensorische Verarbeitung ist gestört. Die Simulation läuft auf internen und verzerrten Signalen und produziert die charakteristische K-Hole-Phänomenologie.

Diese Unterscheidung – Propofol hebt Bewusstsein auf, indem es subkritisch geht; Ketamin verändert Bewusstsein, indem es superkritisch bei gestörtem Input geht – ist ein echter Erklärungsvorteil. Die meisten Theorien tun sich schwer zu erklären, warum zwei „Anästhetika“ so radikal unterschiedliche Erfahrungen produzieren. Der Kritikalitätsrahmen macht die Unterscheidung ganz natürlich.

Ich hatte noch nie eine chemische Narkose. Aber ich wurde bewusstlos geschlagen – hart, plötzlich, ohne Vorwarnung. Und was mir im Rückblick auffällt, ist die totale Abwesenheit eines Übergangs. Schlaf hat Stadien. Psychedelika haben eine Anstiegskurve. Selbst Salvia, so schnell es wirkt, gibt einem den Bruchteil einer Sekunde „Etwas passiert“. Ein K.o. gibt nichts. Einen Moment läuft die Simulation. Im nächsten erwacht man auf dem Boden, ohne Ahnung, wie viel Zeit vergangen ist. Kein Verdunkeln, kein Verblassen, kein Tunnel. Die Simulation degradiert nicht. Sie terminiert. Der Schutzschalter springt.

Genau das wäre bei einer plötzlichen, massiven Störung der kortikalen Dynamik zu erwarten – ein augenblickliches Abfallen weit unter die Kritikalität. Keine Zeit für ein sanftes Herunterfahren durch die Schlafstadien. Das System geht nicht durch Klasse 4 in Klasse 2. Es stürzt aus Klasse 4 ab, und die Simulation stoppt einfach.

Die Bewusstseins-Karte

Zustand	Kritikalität	Modelle	Bewusstsein
Normales Wachen	Bei kritisch	Alle vier aktiv	Voll
REM-Schlaf	Nahe-kritisch	EWM/ESM auf internem Input	Degradiert (Traum)
Tiefer NREM	Subkritisch	EWM/ESM kollabiert	Abwesend
Propofol	Erzwungen subkritisch	EWM/ESM unterdrückt	Abwesend
Ketamin	Über kritisch (↑ Entropie)	EWM/ESM auf falschem Input	Präsent, getrennt
Psychedelika	Bei/über kritisch	Alle aktiv, ↑ Permeabilität	Präsent, verändert
Luzides Träumen	Nahe-kritisch, Schwelle überschritten	EWM aktiv, ESM voll eingebunden	Erhöhtes Selbstbewusstsein

Diese Tabelle fasst zusammen, was in diesem Kapitel zur Sprache kam, und dient als Nachschlage-Referenz. Jeder Bewusstseinszustand, der je erlebt wurde, passt irgendwo auf diese Karte – bestimmt durch zwei Faktoren: ob das Substrat bei Kritikalität arbeitet und welche der vier Modelle laufen. Schlaf, Narkose, Psychedelika, Träume, das K-Hole – keine separaten Mysterien. Verschiedene Koordinaten auf derselben Karte.

Chapter 8

Der klinische Spiegel

Dieselbe Vier-Modelle-Architektur, die Schlaf und Anästhesie erklärt, erklärt auch einige der dramatischsten und rätselhaftesten Zustände der klinischen Neurologie. Das sind keine bloßen Fallstudien – es ist das, was passiert, wenn bestimmte Teile der Architektur ausfallen. Und jeder Ausfall beleuchtet die Architektur aus einem anderen Winkel, so wie eine durchgebrannte Sicherung verrät, welchen Stromkreis sie geschützt hat.

Wenn die Theorie etwas taugt, dann sollte Schaden an bestimmten Modellen bestimmte, vorhersagbare Defizite erzeugen. Kein vages „Bewusstsein ist beeinträchtigt"-Herumwedeln, sondern präzise Vorhersagen: Schalte diese Komponente aus, und es entsteht *jenes* Syndrom. Halte eine andere Komponente ohne ihren normalen Input am Laufen, und es entsteht *dieses* andere Syndrom. Die klinische Literatur ist voll von Zuständen, die unter gängigen Bewusstseinsmodellen zutiefst rätselhaft bleiben, aber wie von selbst an ihren Platz fallen, sobald man eine Real/Virtuell-Unterscheidung und vier interagierende Modelle zur Verfügung hat.

Blindsight und Anton-Syndrom: Der perfekte Spiegel

Wer nur eine Sache aus diesem Kapitel behält, sollte sich dieses Paar merken. Jede andere Bewusstseinstheorie hat Mühe, auch nur

einen dieser Zustände zu erklären. Die Vier-Modelle-Theorie sagt beide vorher.

Zunächst Blindsight. Ein Patient hat eine Schädigung des primären visuellen Kortex – jenes Teils des Gehirns, der bewusste visuelle Erfahrung erzeugt. Nach jedem klinischen Standardtest ist der Patient blind. Fragt man ihn, was er sieht, sagt er: nichts. Und er meint es. Er ist weder bescheiden noch verwirrt. Soweit seine bewusste Erfahrung reicht, existiert die visuelle Welt schlicht nicht.

Aber dann passiert etwas Erstaunliches. Forscher platzieren Hindernisse in einem Flur und bitten den Patienten, hindurchzugehen. Er protestiert – er kann nichts sehen, wie soll er da navigieren? Sie bestehen darauf. Er seufzt, steht auf und geht.

Und er navigiert den Hindernisparcours makellos. Weicht Stühlen aus. Duckt sich unter eine Barriere, die beim letzten Mal nicht da war. Schlängelt sich durch eine Lücke zwischen zwei Hindernissen – und besteht dabei aufrichtig darauf, keinen Deut sehen zu können. Es gibt Videos davon, und es lohnt sich, danach zu suchen, denn bloßes Lesen wird dem Phänomen nicht gerecht. Das Filmmaterial eines klinisch blinden Mannes, der sich durch einen Hindernisparcours schlängelt, als könne er perfekt sehen, gehört zu den atemberaubendsten Demonstrationen der gesamten Neurowissenschaft. Die zuschauenden Forscher sehen aus, als hätten sie einen Geist gesehen.

Wie ist das möglich? Weil das Substrat nach wie vor visuelle Informationen verarbeitet. Das Implizite Weltmodell erhält visuellen Input über subkortikale Pfade, die den beschädigten Kortex umgehen – eine schnelle Route von der Retina über den Colliculus superior zum Pulvinar, die sich lange vor dem Kortex entwickelt hat. Es baut eine räumliche Karte, steuert motorisches Verhalten, hält den Körper davon ab, mit Gegenständen zu kollidieren. Aber nichts davon erreicht das Explizite Weltmodell. Die bewusste Simulation enthält kein Sehen. Der Patient erlebt

wirklich Blindheit – und navigiert wirklich durch Sehen. Das Substrat arbeitet ohne die Simulation.

Jetzt das Gegenstück. Das Anton-Syndrom (Anosognosie für kortikale Blindheit) ist das exakte Gegenteil. Diese Patienten sind vollständig blind. Ihr visueller Kortex oder ihre optischen Bahnen sind zerstört. Keinerlei visuelle Information erreicht das Gehirn. Aber sie sind absolut, unerschütterlich überzeugt, dass sie sehen können.

Sie laufen gegen Wände und beschuldigen die Möbel, am falschen Platz zu stehen. Sie beschreiben Gegenstände, die gar nicht im Raum sind, mit vollem Brustton der Überzeugung – „Da ist eine blaue Vase auf dem Tisch“, wenn der Tisch leer ist. Bittet man sie zu benennen, was man hochhält, kommt ohne Zögern eine Antwort, ruhig und selbstsicher – und sie wird falsch sein. Konfrontiert mit Beweisen für ihre Blindheit, werden sie erst verwirrt, dann gereizt, dann wütend. Die Beleuchtung sei schlecht. Sie bräuchten eine neue Brille. Sie hätten nur nicht aufgepasst. Sie lügen nicht. Sie leugnen nicht im psychologischen Sinn. Sie *sehen* wirklich, erfahrungsgemäß – und was sie sehen, hat keinerlei Bezug zur tatsächlichen Welt.

In der Vier-Modelle-Theorie ist genau das zu erwarten: Das Explizite Weltmodell generiert eine visuelle Simulation aus dem gespeicherten Wissen des Impliziten Weltmodells – obwohl kein aktueller visueller Input eintrifft. Die Simulation läuft auf alten Daten, auf Erwartungen, auf der besten Vermutung des Gehirns darüber, wie die Welt aussehen sollte. Der Patient „sieht“ eine Welt, die nicht da ist. Die Simulation läuft ohne aktuellen Input.

Nebeneinandergestellt: Blindsight – das Substrat verarbeitet Sehen, aber die Simulation zeigt es nicht. Anton-Syndrom – die Simulation zeigt Sehen, aber das Substrat empfängt nichts. Substrat ohne Simulation. Simulation ohne Input. Beide Zustände sind zutiefst rätselhaft, wenn man Bewusstsein als eine einzige, einheitliche Sache betrachtet. Beide sind natürliche, ja vorhersagbare Konsequenzen einer Theorie, die zwischen realer Verarbeitung

und virtueller Erfahrung unterscheidet. Ein besseres Paar von Testfällen könnte man sich kaum ausdenken.

Verborgenes Bewusstsein: Gefangen innen

Im Jahr 2006 veröffentlichten Adrian Owen und Kollegen eine Studie, die das Denken über den vegetativen Zustand grundlegend veränderte. Sie legten eine Patientin, die als vegetativ diagnostiziert worden war – nicht ansprechbar, anscheinend bewusstlos –, in einen fMRT-Scanner und baten sie, sich vorzustellen, Tennis zu spielen. Ihr Gehirn leuchtete in exakt demselben Muster auf wie das einer gesunden bewussten Person bei derselben Vorstellung.

Sie war drinnen. Bewusst, wach, denkend – und vollständig unfähig, sich zu bewegen, zu sprechen oder ihre Anwesenheit irgendjemandem zu signalisieren.

Die Vier-Modelle-Theorie trifft hier eine klare Unterscheidung. Bei einem wirklich vegetativen Patienten ist das Substrat subkritisch. Die Dynamik ist unter die Schwelle gefallen. Die Simulation läuft nicht. Es ist niemand zu Hause – nicht weil die Person „gegangen" ist, sondern weil die rechnerische Architektur, die die Simulation erzeugt, offline gegangen ist.

Aber ein verborgen bewusster Patient ist etwas völlig anderes. Das Substrat ist kritisch – die Dynamik ist reichhaltig genug, um eine Simulation aufrechtzuerhalten. Explizites Weltmodell und Explizites Selbstmodell laufen. Die Person erlebt, denkt, fühlt. Aber die Ausgabepfade sind zerstört. Die Simulation hat keinen Kanal nach außen. Die Person ist bewusst, aber eingeschlossen – gefangen in einem Körper, der nicht reagiert.

Der Perturbational Complexity Index (dasselbe Maß, das Schlafstadien unterscheidet) sollte genau diese Fälle auseinanderhalten. Und das tut er. Einige Patienten, die als vegetativ diagnostiziert wurden, zeigen PCI-Werte mitten im bewussten Bereich. Sie sind keineswegs vegetativ. Sie sind Gefangene. Die medizinischen und ethischen Implikationen sind enorm – und die Vier-Modelle-

Theorie sagt genau vorher, warum die Unterscheidung existiert und wie man sie erkennt.

Cotard-Wahn: „Ich bin tot“

Und dann gibt es Patienten, die glauben, sie seien tot.

Der Cotard-Wahn gehört zu den sonderbarsten Zuständen der Psychiatrie. Patienten bestehen darauf, gestorben zu sein. Sie glauben, ihre Organe hätten sich aufgelöst, ihr Blut sei abgeflossen, sie existierten nicht mehr. Manche glauben, sie verwesen. Manche halten sich für unsterblich – denn wer schon tot ist, kann nicht noch einmal sterben. Sie sprechen nicht metaphorisch. Sie meinen es mit vollständiger, unerschütterlicher Überzeugung.

Inzwischen dürfte der Mechanismus vertraut klingen. Es ist derselbe wie in Kapitel 6 – das Explizite Selbstmodell konstruiert das beste Modell, das es kann, aus dem jeweils verfügbaren Input. Beim Cotard-Wahn ist der interozeptive Input schwer verzerrt. Die inneren Körpersignale, die vermitteln, dass das Herz schlägt, der Magen verdaut, die Lungen atmen – sie sind verschüttet oder verstümmelt. Und das ESM, immer der zwanghafte Konstrukteur, interpretiert „kein Herzschlag, keine Verdauung, keine Atmung, keine Körperempfindung“ auf die einzige Weise, die ihm bleibt: Ich bin tot.

Salvias „Ich bin ein Stuhl.„ Anosognosies „Mein Arm ist in Ordnung.“ Und jetzt Cotards „Ich bin tot.“ (Im nächsten Kapitel wird Split-Brain-Konfabulation noch einen weiteren Fall zu dieser Liste beisteuern.) Ein Mechanismus durchzieht jeden Fall. Das Explizite Selbstmodell tut immer seine Arbeit – baut immer das beste Selbstmodell, das es kann. Stimmt der Input, fühlt man sich wie man selbst. Stimmt er nicht, fühlt man sich wie ein Stuhl, oder gesund trotz Lähmung, oder tot trotz Leben. Aber es fühlt sich immer vollständig, überzeugend real an – weil es das einzige Selbst ist, zu dem man Zugang hat. **Alien-Hand-Syndrom: Wenn das Komitee sich nicht einig ist**

Dann gibt es einen Zustand, der sich wie ein Horrorfilm liest, aber die Multi-Agenten-Natur des Substrats greifbarer macht als

jedes Gedankenexperiment. Beim Alien-Hand-Syndrom handelt eine Hand des Patienten mit offensichtlichem Zweck und Absicht – aber gegen seinen bewussten Willen. Eine Hand zündet eine Zigarette an, die andere nimmt sie weg und wirft sie auf den Boden. Eine Hand greift nach dem Türknauf, die andere packt das Handgelenk und reißt es zurück. Der Patient schaut entsetzt zu, wie ein Teil seines eigenen Körpers Ziele verfolgt, die er nie gewählt hat.

Stanley Kubrick hat das in *Dr. Seltsam* verwendet – und die Leute glaubten, er hätte es erfunden. Hat er nicht. Das Syndrom ist real und tritt in zwei Varianten auf. In der kallösen Form, ausgelöst durch eine Schädigung des Corpus callosum, ähneln die Symptome Split-Brain-Konflikten: zwei Hemisphären mit konkurrierenden motorischen Plänen, von denen keine die andere überstimmen kann. In der frontalen Form, ausgelöst durch präfrontalen Schaden, zeigt die „fremde“ Hand enthemmtes Verhalten – greift nach Gegenständen, benutzt Werkzeuge, berührt Dinge zwanghaft –, alles scheinbar zielgerichtet, aber ohne Einverständnis des Patienten.

Es gibt auch eine subtilere Spielart, das Anarchische-Hand-Syndrom, bei dem nicht die motorische *Zugehörigkeit* gestört ist, sondern die motorische *Kontrolle*. Die Hand tut Dinge, die der Patient nicht beabsichtigt hat, aber er erkennt sie weiterhin als *seine* Hand – er kann sie nur nicht stoppen. Die Unterscheidung ist wesentlich: Alien-Hand bedeutet ein Versagen der Körperzugehörigkeitsgrenze im Expliziten Selbstmodell („diese Hand gehört nicht zu mir„), Anarchische Hand dagegen ein Versagen der motorischen Hemmung („diese Hand gehört mir, aber sie gehorcht nicht“). Dieselbe Architektur, unterschiedliche Bruchstellen.

Die zentrale Erkenntnis aus der Analyse dieser Syndrome ist: Das Gefühl der Urheberschaft – das Empfinden von „Ich habe das getan“ – wird nicht vor oder während der Handlung erzeugt. Es entsteht *danach*, indem das vorhergesagte Ergebnis mit dem tatsächlich beobachteten verglichen wird. Stimmt der

Abgleich, stellt sich das Gefühl der Zugehörigkeit ein. Stimmt er nicht, bleibt es aus. Deshalb können Patienten mit Alien-Hand-Syndrom sich manchmal selbst kitzeln – ihr Vorhersagesystem liefert kein erwartetes Ergebnis für die Bewegungen der fremden Hand, also trifft die Berührung unerwartet ein, als käme sie von jemand anderem.

Charles-Bonnet-Syndrom: Die Simulation, die nicht aufhört

Wer noch mehr Belege dafür braucht, dass die Simulation des Gehirns *generativ* arbeitet – dass sie Erfahrung aus Modellen konstruiert, statt sie passiv von den Sinnen entgegenzunehmen –, der betrachte das Charles-Bonnet-Syndrom. Patienten, deren Netzhaut oder Sehnerv zerstört ist, deren visueller Kortex aber intakt bleibt, erleben lebhafte, komplexe visuelle Halluzinationen. Keine vagen Formen oder Lichtblitze. Volle Szenen: Menschen, manchmal miniaturisiert oder kostümiert wie Zeichentrickfiguren, manchmal Spiegelbilder des Patienten. Landschaften. Gegenstände. Gesichter.

Die Patienten wissen in der Regel, dass nichts davon real ist. Anders als psychotische Halluzinationen gehen Charles-Bonnet-Halluzinationen mit intakter Einsicht einher – der Patient sagt: „Ich sehe einen kleinen Mann mit Zylinder auf meinem Tisch sitzen, und ich weiß, dass er nicht da ist." Was hier läuft, ist die visuelle Simulation des Expliziten Weltmodells, gespeist aus internen Daten höherer visueller Areale, ohne jeden äußeren Input. Die Simulation hört nicht auf, nur weil der Input versiegt. Sie generiert. Sie füllt die Leere. Und was sie generiert, verrät etwas über die Architektur: Das visuelle System ist ein generatives Modell, kein passiver Empfänger. Es erzeugt seine beste Vermutung darüber, wie die Welt aussieht, anhand gespeicherter Vorlagen und Top-Down-Vorhersagen – genau wie es die Vier-Modelle-Theorie beschreibt.

Déjà-vu: Die Vorlage, die zu gut passt

Apropos generatives System und seine gelegentlichen Fehlzündungen: Fast jeder kennt Déjà-vu – das unheimliche Gefühl, den aktuellen

Moment schon einmal erlebt zu haben. Erklärungen reichen vom Mystischen (frühere Leben, Vorahnungen) bis zum Banalen (nur ein Glitch). Die Vier-Modelle-Theorie liefert eine präzisere Erklärung.

Das Gehirn speichert etwas, das man als „Vorlagen-Erinnerungen" bezeichnen könnte – skelettartige, extrem spärliche Repräsentationen von Erfahrungen, besonders aus Träumen. Diese Vorlagen sind größtenteils leere Gerüste: ein vages Gefühl eines Ortes, eine Stimmung, eine räumliche Konfiguration, mit fast keinem Detail ausgefüllt. Beim normalen Abruf einer Erinnerung werden die Lücken durch Konfabulation gefüllt – das Gehirn generiert plausible Details, um eine nahtlose Erfahrung zu erzeugen. Das Auffüllen fällt nicht auf, weil das Ergebnis kohärent wirkt.

Déjà-vu tritt auf, wenn eine aktuelle reale Erfahrung zufällig zu gut zu einer dieser gespeicherten Vorlagen passt. Der Musterabgleich des Gehirns feuert: „Das habe ich schon mal gesehen.„ Aber der Versuch festzunageln, *wann* genau, führt ins Leere – weil die Vorlage nie eine reale Erfahrung war. Es war ein Fragment aus einem Traum oder eine so tief komprimierte Erinnerung, dass sie alle kontextuellen Details längst verloren hat. Die Übereinstimmung zwischen aktuellem Input und gespeicherter Vorlage ist echt, aber die „originale" Erfahrung, die die Vorlage angeblich aufzeichnet, hat in dieser Form nie stattgefunden. Das System arbeitet korrekt – es hat tatsächlich eine Übereinstimmung gefunden. Nur ist die Übereinstimmung mit einem Skelett, nicht mit einem Körper.

Was Therapie tatsächlich tut

Der klinische Spiegel zeigt nicht nur Pathologie. Er beleuchtet auch, was man dagegen tun kann, und die Vier-Modelle-Theorie liefert eine überraschend präzise Erklärung, wie Therapie funktioniert.

Kognitive Verhaltenstherapie – die empirisch am besten abgesicherte Form der Psychotherapie, die es gibt. In der Vier-Modelle-Theorie ist KVT im Grunde Neuprogrammierung virtueller Modelle. In der Sitzung werden systematisch die verzerrten

Modelle hinterfragt, die das Leiden erzeugen. Die automatischen Gedanken (Ausgaben des Expliziten Selbstmodells) werden identifiziert, zu zugrunde liegenden Überzeugungen (Mustern des Impliziten Selbstmodells) zurückverfolgt und dann durch wiederholte korrektive Erfahrung die Neuverdrahtung auf Substrat-Ebene vorangetrieben. Synaptische Plastizität verändert das Implizite Selbstmodell, und das ändert, was das Explizite Selbstmodell generiert.

Therapie verdrahtet buchstäblich die impliziten Modelle um. Das ist der Mechanismus. Jedes Mal, wenn ein katastrophaler Gedanke hinterfragt wird und die Welt nicht untergeht, aktualisieren sich IWM und ISM. Jedes Mal, wenn eine gefürchtete Situation durchgestanden wird, schreiben sich neue Daten ins Substrat. Die virtuellen Modelle ändern sich, weil die realen Modelle sich zuerst ändern.

Phobien sind Fehlkonfigurationen des Expliziten Weltmodells. Die Bedrohungsrepräsentation im EWM übersteigt die Evidenzbasis des Impliziten Weltmodells massiv. Die Simulation zeigt Gefahr, wo die akkumulierte Evidenz des Substrats sie nicht stützt. Eine harmlose Spinne kommt ins Blickfeld, und das EWM schreit *Bedrohung*, obwohl das IWM nie eine tatsächliche Verletzung durch eine Spinne verzeichnet hat. Expositionstherapie funktioniert, indem sie das IWM durch wiederholte sichere Begegnungen aktualisiert. Jedes Mal, wenn die Konfrontation folgenlos bleibt, korrigiert das implizite Modell seine Bedrohungseinschätzung nach unten. Irgendwann hört das EWM auf, den Fehlalarm auszulösen. Die Simulation hört auf, eine Gefahr zu zeigen, die nicht existiert.

Der Placebo-Effekt fügt sich nahtlos in die duale Bewertungsarchitektur der Theorie ein. Placebo aktiviert Erwartungsschaltkreise auf Substrat-Ebene – endogene Opioid-Freisetzung, dopaminerge Belohnungspfade –, die parallel zur bewussten Erfahrung von Hoffnung und Erwartung laufen. Die bewusste Hoffnung und die körperliche Erleichterung werden beide durch denselben Substrat-

Prozess verursacht. Die Korrelation zwischen „Ich glaube, diese Pille wird helfen„ und „Ich fühle mich besser“ ist real, aber nicht kausal. Der Glaube verursacht nicht die Erleichterung. Sowohl der Glaube als auch die Erleichterung entstehen aus denselben zugrunde liegenden Substrat-Dynamiken. Das ist kein Schlag gegen die Macht des positiven Denkens – es ist eine Erklärung dafür, wie diese „Macht“ tatsächlich funktioniert: auf Substrat-Ebene, nicht durch irgendeine mysteriöse Abwärtsverursachung vom Geist zum Körper.

Und dann die Konversionsstörung – das perfekte Gegenstück zu Blindsight. Bei Blindsight verarbeitet das Substrat visuelle Informationen, ohne eine bewusste Simulation daraus zu erzeugen. Bei der Konversionsstörung modelliert die Simulation ein Defizit (Lähmung, Blindheit, Anfälle), das im intakten Substrat gar nicht vorliegt. Der Patient ist wirklich gelähmt, soweit es seine bewusste Erfahrung betrifft. Er täuscht nichts vor. Seine Simulation enthält ein gelähmtes Glied. Aber sein Körper funktioniert auf Substrat-Ebene einwandfrei – die Nerven leiten, die Muskeln kontrahieren, die Bahnen sind intakt. Therapie gelingt, wenn es gelingt, die Simulation zu korrigieren und das Körpermodell des ESM an die tatsächlichen Fähigkeiten des Substrats anzupassen. Blindsight umgekehrt: Statt eines funktionierenden Substrats, das vor einer blinden Simulation verborgen ist, ein funktionierendes Substrat, das hinter einer „kaputten“ Simulation verborgen ist.

Chapter 9

Zwei Bewusstsein in einem Gehirn

In den 1960er Jahren führten Roger Sperry und Michael Gazzaniga eines der dramatischsten Experimente der Neurowissenschaft durch. Um schwere Epilepsie zu behandeln, trennten sie chirurgisch das Corpus callosum – das massive Faserbündel, das die beiden Gehirnhälften verbindet. Das Ergebnis: das Split-Brain-Syndrom. Eine Person mit offenbar zwei unabhängigen Bewusstseinen.

Die klassischen Demonstrationen sind legendär. Wird ein Wort im linken Gesichtsfeld gezeigt (verarbeitet von der rechten Hemisphäre), kann der Patient das passende Objekt mit der linken Hand greifen – aber nicht sagen, was das Wort war. Denn Sprache sitzt in der linken Hemisphäre, und die hat das Wort nie gesehen. Die beiden Hälften haben eigene Wahrnehmungen, eigene Absichten, manchmal gegenläufige Ziele.

Die Experimente gingen weit über Partytricks hinaus. In manchen Fällen kämpften die Hemisphären offen gegeneinander. Ein Patient berichtete, seine linke Hand knöpfe sein Hemd auf, während die rechte es wieder zuknöpfte. Bei einem anderen griff die linke Hand während eines Streits nach seiner Frau – nicht um sie zu trösten –, während die rechte die linke packte und zurückzog. Der Patient schaute entsetzt zu, wie zwei Teile seines

Körpers unvereinbare Ziele verfolgten, keines unter einheitlicher Kontrolle. Das sind keine Metaphern für inneren Konflikt. Es sind buchstäbliche, physische Kämpfe zwischen zwei motorischen Systemen, die sich nicht mehr koordinieren können, weil das Kabel dazwischen durchtrennt wurde.

Im Alltag funktionieren Split-Brain-Patienten erstaunlich gut. Außerhalb des Labors fällt selten etwas Ungewöhnliches auf. Die beiden Hemisphären lernen, über Umwege zu kooperieren – äußere Hinweise, Körperbewegungen, gemeinsame Gesichtsfelder. Das System kompensiert. Doch unter kontrollierten Laborbedingungen, wo jede Hemisphäre unterschiedliche Informationen erhält, zerfällt die Einheit. Zwei Bewusstsein entstehen aus einem Gehirn, jedes mit eigenen Wahrnehmungen, eigenen Absichten und einer eigenen Version der Realität.

Der Linke-Hemisphären-Interpret

Das aufschlussreichste Merkmal von Split-Brain-Patienten ist aber nicht die Teilung – es ist, was passiert, wenn die Teilung *erklärt* werden soll.

Gazzaniga identifizierte etwas, das er den „linken-Hemisphären-Interpreten„ nannte: die zwanghafte Tendenz der linken Hemisphäre, Erklärungen für Ereignisse zu fabrizieren, zu denen sie gar keinen Zugang hat. Die klassische Demonstration: Der rechten Hemisphäre wird eine Schneeszene gezeigt, der linken eine Hühnerkralle. Dann soll der Patient verwandte Objekte auswählen. Die linke Hand (rechte Hemisphäre) greift zur Schaufel – für den Schnee. Die rechte Hand (linke Hemisphäre) wählt ein Huhn. Wird der Patient nun gefragt – über Sprache, also über die linke Hemisphäre –, warum er die Schaufel gewählt hat, weiß die linke Hemisphäre nichts vom Schnee. Sie sah nur die Hühnerkralle. Also erfindet sie eine Erklärung: „Oh, man braucht eine Schaufel, um den Hühnerstall auszumisten.“

Der Patient zögert nicht. Sagt nicht „Ich bin mir nicht sicher." Schaut nicht verwirrt. Die Erklärung kommt sofort, selbstsicher – und fühlt sich für den Sprechenden vollkommen natürlich an. Das ist kein Lügen. Die linke Hemisphäre weiß tatsächlich nicht, was die rechte gesehen hat. Das Kabel ist durch. Also tut sie, was das Explizite Selbstmodell immer tut: die bestmögliche Erzählung konstruieren aus dem, was gerade verfügbar ist.

Und jetzt der Teil, der einem den Schlaf rauben sollte: Wir alle tun das. Jeden Tag. Der Linke-Hemisphären-Interpret läuft genau jetzt, konstruiert eine kohärente Erzählung aus allem, was das Bewusstsein gerade erreicht, glättet Lücken, erfindet plausible Erklärungen für Entscheidungen, die das Substrat längst getroffen hat, bevor das bewusste Selbst auch nur befragt wurde. Der einzige Unterschied zu einem Split-Brain-Patienten: Das Corpus callosum ist intakt, also hat der Interpret Zugang zu mehr Daten. Er konfabuliert weniger, weil er weniger zu konfabulieren *hat*. Aber der Mechanismus ist identisch. Die Maschinerie der Selbsterzählung ändert sich nicht. Nur die Qualität des Inputs.

Eine Person oder zwei?

Die Frage, über die Philosophen seit Jahrzehnten streiten: Nachdem das Callosum durchtrennt ist – steckt da eine Person in diesem Schädel oder zwei?

Thomas Nagel befasste sich damit in einem berühmten Essay von 1971 und kam zu dem Schluss, dass die Frage möglicherweise keine eindeutige Antwort hat. Unser Konzept von „einer Person„ bricht hier schlicht zusammen – so wie der Begriff „ein Land" zusammenbricht, wenn eine Grenze durch die Mitte gezogen wird. Derek Parfit ging weiter und argumentierte, dass Split-Brain-Fälle zeigen, dass persönliche Identität selbst nicht das Entscheidende ist – was zählt, ist psychologische Kontinuität, und davon gibt es Grade.

Die Vier-Modelle-Theorie bietet eine präzisere Antwort: Es hängt davon ab, welche Modelle laufen und wie stark sie degradiert sind.

Im Alltag ist ein Split-Brain-Patient funktional eine Person. Beide Hemisphären teilen denselben Körper, dieselbe Umgebung, dieselbe Lebensgeschichte – redundant in beiden Hälften kodiert, lange vor der Operation. Das Implizite Selbstmodell, das Persönlichkeit, Langzeiterinnerungen und Verhaltensdispositionen speichert, wurde über Jahrzehnte mit intaktem Callosum aufgebaut. Das Durchtrennen des Kabels löscht diese Modelle nicht. Es verhindert nur, dass sie synchron aktualisiert werden. Unmittelbar nach der Operation laufen also in beiden Hemisphären sehr ähnliche Selbstmodelle. Der Patient fühlt sich wie eine Person, weil er – gemessen am gespeicherten Selbstwissen – größtenteils eine ist.

Mit der Zeit sollten die Modelle driften. Jede Hemisphäre sammelt unterschiedliche Erfahrungen, bildet unterschiedliche Assoziationen, entwickelt unterschiedliche emotionale Reaktionen auf Ereignisse, die nur sie wahrgenommen hat. Je länger ein Split-Brain-Patient nach der Operation lebt, desto stärker sollten die beiden impliziten Selbstmodelle auseinanderlaufen – langsam, weil beide Hemisphären immer noch denselben Körper und dieselbe Umgebung teilen, aber messbar.

Die Antwort tendiert in Richtung *zwei*. Wenn die Bandbreite zwischen den Hemisphären nicht für die Echtzeitsynchronisation der Simulation reicht – und ohne Callosum reicht sie nicht –, dann laufen zwei Selbstmodelle auf zwei Substraten, jedes mit eigener bewusster Erfahrung. Sie kooperieren gut, weil sie einen Körper, ein sensorisches Umfeld und eine Lebenszeit gemeinsamer Geschichte teilen. Aber Kooperation ist nicht Identität. Zwei Menschen, die zusammenleben, kooperieren auch gut.

Interessanterweise veröffentlichten Yair Pinto und Kollegen 2017 eine Studie, die das Standardbild verkomplizierte. Sie fanden, dass Split-Brain-Patienten Stimuli korrekt benennen konnten, die in beiden Gesichtsfeldern präsentiert wurden –

selbst wenn der Stimulus nur der Hemisphäre gezeigt wurde, die keine Sprache kontrolliert. Das deutete darauf hin, dass die beiden Hemisphären mehr Einheit bewahren, als die klassischen Experimente nahelegten. Das Ergebnis wird noch diskutiert, passt aber gut in den holographischen Rahmen, der gleich kommt: Selbst nach dem Durchtrennen des Callosums steckt in jeder Hemisphäre genug redundante Information, um bei vielen Aufgaben überraschend einheitlich zu reagieren.

Die holographische Eigenschaft

In der Vier-Modelle-Theorie offenbart das Split-Brain eine Schlüsseleigenschaft der virtuellen Modelle: Sie sind **holographisch**. Information in neuronalen Netzwerken ist über das gesamte Netzwerk verteilt, nicht in einzelnen Neuronen lokalisiert. Schneidet man das Netzwerk entzwei, entsteht keine saubere Teilung – es entstehen zwei degradierte, aber *vollständige* Kopien. Jede Hemisphäre behält eine abgespeckte Version aller vier Modelle: ein reduziertes Implizites Weltmodell, ein reduziertes Implizites Selbstmodell und die Fähigkeit, Explizites Weltmodell und Explizites Selbstmodell zu generieren. Beide Hemisphären können Bewusstsein eigenständig aufrechterhalten (beide liegen über der Kritikalitätsschwelle), aber jede arbeitet mit weniger Information.

Genau das passiert, wenn man ein Hologramm in zwei Hälften schneidet. Man bekommt nicht zwei halbe Bilder. Man bekommt zwei vollständige Bilder, jedes mit niedrigerer Auflösung. Die Information in einem Hologramm ist über die gesamte Aufzeichnungsfläche verteilt, sodass jedes Stück das ganze Bild enthält – nur unschärfer. Neuronale Netzwerke haben dieselbe Eigenschaft. Karl Lashley demonstrierte das vor Jahrzehnten: Selbst nach Zerstörung großer Teile des Rattenkortex bestehen die Erinnerungen weiter, degradiert, aber vollständig. Das Gehirn speichert Erinnerungen nicht in Aktenschränken. Es speichert sie wie ein Hologramm sein Bild speichert – überall

gleichzeitig, sodass Schäden die Qualität mindern, ohne den Inhalt auszulöschen.

Deshalb sind Split-Brain-Patienten nicht einfach „zwei halbe Bewusstsein". Sie sind zwei *vollständige, aber degradierte* Bewusstsein. Jede Hemisphäre kann wahrnehmen, entscheiden und handeln – nur mit weniger Information und weniger Kapazität als das intakte System. Die holographische Eigenschaft stellt sicher, dass Trennung degradiert, ohne zu zerstören. Und sie erklärt Pintos Ergebnisse von 2017: Selbst ohne Callosum behält jede Hemisphäre genug holographische Information, um viele Aufgaben zu bewältigen, die nach dem klassischen Modell unmöglich sein sollten.

Die Konfabulation (der Linke-Hemisphären-Interpret) ist *derselbe Mechanismus*, der schon beim Cotard-Wahn auftauchte (das ESM produziert auf verzerrtem interozeptivem Input „Ich bin tot„), bei Anosognosie (das ESM ignoriert auf unvollständigem Input das Defizit) und bei Salvia (das ESM produziert auf Nicht-Selbst-Input „Ich bin ein Stuhl"). In jedem Fall tut das Explizite Selbstmodell seine Arbeit – eine Selbsterzählung konstruieren – mit dem, was gerade an Input da ist. Ist der Input unvollständig oder verzerrt, ist die Erzählung falsch, aber *fühlt sich trotzdem vollkommen real an*.

Ein Gehirn, multiple Selbste

Das Split-Brain zeigt, was passiert, wenn die virtuellen Modelle durch physisches Teilen des Substrats *geklont* werden. Dissoziative Identitätsstörung zeigt, was passiert, wenn sie *geforkt* werden.

Bei DID ist das Substrat nicht geteilt – das Corpus callosum ist intakt, die neuronale Hardware unversehrt. Aber die virtuellen Modelle haben sich in mehrere Konfigurationen aufgespalten. Jeder Alter ist ein eigenständiges Explizites Selbstmodell – eine separate Selbsterzählung mit eigenem emotionalem Profil, eigenen Verhaltensmustern, eigener Art, sich auf Körper und Welt zu beziehen. Die Alters teilen sich kein gemeinsames Selbstmodell,

genauso wenig wie zwei Benutzer eine einzige Login-Sitzung auf demselben Computer teilen. Sie wechseln sich ab.

Der Auslöser ist in praktisch jedem dokumentierten Fall schweres, wiederholtes Kindheitstrauma. Das ergibt innerhalb der Theorie Sinn. Das Explizite Selbstmodell eines kleinen Kindes formt sich noch – ist noch plastisch, wird noch aus Erfahrung zusammengesetzt. Wird dieses sich entwickelnde Selbstmodell Erfahrungen ausgesetzt, die so überwältigend sind, dass keine einzelne Selbsterzählung sie fassen kann, tut das System das Einzige, was es kann: Es forkt. Es erstellt separate Konfigurationen, jede fähig, einen anderen Aspekt der unerträglichen Situation zu handhaben. Ein Alter trägt die Trauma-Erinnerungen. Ein anderer funktioniert im Alltag, als wäre nichts geschehen. Ein dritter übernimmt in Gefahrenmomenten. Das Forking ist keine Pathologie – es ist die Notfallreaktion des Selbstmodellierungssystems auf Input, der ein einzelnes vereinheitlichtes Modell sprengen würde.

Deshalb entwickelt sich DID fast nie bei Erwachsenen. Das Implizite Selbstmodell eines Erwachsenen ist bereits konsolidiert – die synaptischen Gewichte sind gesetzt, die Persönlichkeitsstruktur stabil. Ein erwachsenes Selbstmodell zu forken erfordert außergewöhnliche Umstände: schwere Folter, langandauernde Gefangenschaft. Aber das ISM eines Kindes wird noch geschrieben. Der Ton ist noch nass. Wird er unter genügend Druck geforkt, härten die separaten Konfigurationen zu eigenständigen, persistenten Selbstmodellen aus.

Die Evidenz bestätigt das. Wenn jeder Alter tatsächlich eine eigenständige ESM-Konfiguration ist, sollte das Wechseln zwischen Alters messbare Veränderungen in neuronalen Aktivitätsmustern erzeugen – und genau das tut es. Reinders et al. (2003) zeigten, dass verschiedene Alters im selben Individuum unterschiedliche Muster des regionalen zerebralen Blutflusses produzieren. *Dasselbe Gehirn* leuchtet unterschiedlich auf, je nachdem welches Selbstmodell läuft. Das wäre bei bloßem Schauspiel nicht zu erwarten. Das ist,

was man bei echtem Software-Forking erwartet. In Folgestudien fanden Reinders und Kollegen, dass die neuronalen Unterschiede zwischen Alters größer waren als die Unterschiede zwischen Schauspielern, die DID simulieren sollten – ein Ergebnis, das jeden zum Schweigen bringen sollte, der immer noch glaubt, DID sei „nur“ Performance.

Das ist die „Forking“-Eigenschaft aus Kapitel 3 in Aktion. Ein Substrat, multiple virtuelle Konfigurationen, jede ein vollständiges, aber eigenständiges Selbstmodell. Die Theorie akkommodiert DID nicht bloß – sie sagt genau diese Art von Architektur vorher. Vorhersage 9 in Kapitel 11 macht den Test explizit: Die gezielte Störung des neuronalen Substrats, das das ESM eines Alters aufrechterhält, sollte einen Wechsel zu einem anderen auslösen.

Chapter 10

Die Frage der Tiere

Ist der Hund bei Bewusstsein?

Die meisten Haustierbesitzer würden ohne zu zögern mit Ja antworten. Die meisten Neurowissenschaftler ebenfalls, wenn auch vorsichtiger. Aber worauf stützt sich diese Überzeugung? Und wo im Tierreich beginnt Bewusstsein?

Die Vier-Modelle-Theorie (Four-Model Theory, FMT) liefert klare Antworten — und zwar solche, die sich zwingend aus den Grundannahmen ergeben, nicht nachträglich angeheftet wurden.

Annahme 1: Bewusstsein ist ein Kontinuum, nicht binär. Es gibt keine scharfe Grenze zwischen bewusst und unbewusst. Es gibt Grade — abgestufte Ebenen der Selbstsimulation, von rudimentär (minimales Selbstmodell) bis dreifach erweitert (rekursive Selbstwahrnehmung). Verschiedene Tierarten besetzen verschiedene Positionen auf diesem Kontinuum.

Annahme 2: Bewusstsein ist substratunabhängig. Was zählt, ist die funktionale Architektur (vier Modelle bei Kritikalität), nicht die konkrete physische Implementierung. Wenn ein Gehirn die Vier-Modelle-Architektur realisiert, ist es bei Bewusstsein — egal ob Säugetiercortex, Vogelpallium oder verteiltes neuronales Netzwerk eines Oktopus.

Annahme 3: Kritikalität ist die physische Schwelle. Ein Nervensystem muss am oder nahe dem Rand des Chaos operieren.

Einfachere Nervensysteme (Insekten, Würmer) erreichen diese Kritikalität möglicherweise nicht — sie verarbeiten Information und erzeugen Verhalten, aber ohne Simulation.

Zusammengenommen sagen diese Annahmen einen **Gradienten tierischen Bewusstseins** voraus:

Säugetiere sind bei Bewusstsein. Ihr Cortex implementiert die Vier-Modelle-Architektur in abgestufter Form, wobei komplexere Cortices anspruchsvollere Selbstsimulationen tragen. Primaten und Wale liegen am oberen Ende, Nagetiere und Spitzmäuse am unteren. Alle liegen über der Schwelle.

Unter Darwin's Arch — damals, als er noch stand — tauchte bei einem Tauchgang mit vier anderen ein riesiger Orca-Bulle aus dem Nichts auf und kam so nah, dass der erste Gedanke war: Er will uns fressen. Das Tier war derart massiv, dass die dreißig Meter Wasser über unseren Köpfen wie eine Pfütze wirkten. Er stoppte und musterte uns mit seinem gewaltigen rechten Auge, wechselte dann zu seinem akustischen Organ, klickte intensiv und scannte uns mit Schall. Und dann — es lässt sich nicht anders beschreiben — *sprach* er. In einem sehr hochfrequenten Gesang, kurz und klar strukturiert, sagte er etwas. Es gab nicht den geringsten Zweifel, dass das Sprache war, nicht bloß Gesang. Es klang wie etwas, das sich vermutlich lernen ließe, gäbe es genug Begegnungen und eine lächerlich hohe Stimme, um zu antworten.

Etwas später wurde klar, was er gesagt hatte. Er schwamm zurück an die Grenze der Sichtweite und kehrte mit seiner Partnerin und seinem Kalb zurück, führte sie an uns vorbei — damit sie einen Blick werfen konnten. Natürlich hatten alle Kameras dabei. Kein einziges Foto wurde gemacht. Und alle hatten keine Luft mehr und mussten auftauchen. Auf dem Schlauchboot zurück zum Schiff hatte jeder Tränen in den Augen, und es lag nicht am Wind.

Die Evidenz von Menschenaffen ist besonders verheerend für jeden, der eine scharfe Linie zwischen menschlichem und tierischem Bewusstsein ziehen möchte. Der Bonobo Kanzi

demonstrierte nicht nur Sprachverständnis, sondern echte Empathie, Theory of Mind und soziales Denken. In einer gut dokumentierten Episode teilte Kanzi seiner Betreuerin mit, dass seine Schwester bei einem Einkaufstrip mitkommen solle, damit auch sie Eis bekomme — weil sie traurig wäre, wenn man sie zurückließe. In einem anderen Fall, während einer Tanzaufführung indigener Künstler, erklärte Kanzi den Forschern, dass die anderen Primaten durch den Tanz verängstigt seien, und bat um eine Privatvorführung.

Das sind keine Reflexe. Keine konditionierten Reaktionen. Das sind Beispiele eines Geistes, der den emotionalen Zustand eines anderen Geistes modelliert, Reaktionen vorhersagt und Pläne formuliert, um darauf einzugehen. Das ist das Explizite Selbstmodell (Explicit Self Model, ESM), das aus der Dritte-Person-Perspektive arbeitet — genau das, was die Theorie als Kennzeichen erweiterten Bewusstseins identifiziert.

Trotzdem lassen sich in einigen der renommiertesten Universitätshörsäle immer noch Professoren finden, die mit ernster Miene argumentieren, Menschenaffen würden Sprachverständnis „nur simulieren". Worauf man nur antworten kann: „Und Sie simulieren nur die Anwesenheit von Empathie." Die Gegenbeweise lassen auf sich warten.

Wer darauf besteht, dass nur Menschen Bewusstsein besitzen, wettet auf jene Forscher, die immer noch verzweifelt nach einem systematischen Unterschied zwischen Menschen- und Primatengehirnen suchen, den sie dem Bewusstsein zuschreiben können. Laut dieser Theorie werden sie ihn am 36. August finden.

Rabenvögel und Papageien stellen den wichtigsten Testfall dar. Diese Vögel zeigen kognitive Fähigkeiten (Werkzeugherstellung, Selbsterkennung im Spiegel, Zukunftsplanung, soziale Täuschung), die stark auf Bewusstsein hindeuten — und besitzen keinen Neocortex. Ihr Gehirn ist in nukleären Clustern organisiert, eine radikal andere Architektur als der Säugetiercortex. Zur Erinnerung: das Sechs-Schichten-Argument aus Kapitel 2 — Säugetiere entwickelten sechs kortikale Schichten, wo drei genügt

hätten, und die zusätzlichen Schichten liefern die architektonische Kapazität für Selbstmodellierung. Rabenvögel erreichen dasselbe funktionale Ergebnis mit einer völlig anderen physischen Struktur. Sie brauchen keine sechs kortikalen Schichten, weil sie *überhaupt keine* kortikalen Schichten haben. Sie haben die Selbstsimulationsarchitektur aus nukleären Clustern statt aus geschichteten Platten gebaut — und genau das sagt Substratunabhängigkeit voraus. Würde Bewusstsein eine bestimmte physische Implementierung erfordern, dürften Rabenvögel nicht bei Bewusstsein sein. Sie sind es.

Kopffüßer (Oktopusse und Tintenfische) führen die Logik noch weiter. Ihr Nervensystem ist weitgehend dezentralisiert, mit erheblicher autonomer Verarbeitung in den Armen. Die Theorie sagt eine Form von Bewusstsein voraus, vermutlich mit ungewöhnlichen Merkmalen, die die dezentralisierte Architektur widerspiegeln.

Insekten sind der interessante Grenzfall. Ihre Nervensysteme sind klein und weitgehend fest verdrahtet, was Kritikalität erreichen mag oder nicht. Die Theorie platziert Insekten nicht eindeutig über oder unter der Schwelle — das ist eine empirische Frage. Aber sie liefert eine prinzipielle Grundlage für die Untersuchung: Kritikalitätsindikatoren im neuronalen Gewebe von Insekten messen und nach Hinweisen auf ein Selbstmodell suchen.

Thomas Nagel stellte seine berühmte Frage, wie es sei, eine Fledermaus zu sein, und kam zu dem Schluss, dass man es nie wissen könne — die sensorische Welt der Fledermaus sei zu fremd. Für die Frage lässt sich Sympathie aufbringen, für die Schlussfolgerung weniger. Die Vier-Modelle-Theorie sagt voraus, dass jedes Wesen mit der Vier-Modelle-Architektur bei Kritikalität *irgendeine* Form von Erfahrung hat, selbst wenn deren Inhalt radikal anders ist als unserer. Das Explizite Weltmodell (Explicit World Model, EWM) der Fledermaus wird von Echoortung statt von Sehen dominiert, aber es ist immer noch ein Modell — immer noch eine Simulation einer Welt mit einem Selbst darin.

Nagel wählte Fledermäuse natürlich mit Bedacht — nicht nur, weil sie Säugetiere sind, sondern weil Echoortung unwiderruflich fremd erscheint. Nur: Das ist sie gar nicht. Echoortung dient dem *Sehen*, und wie sich Sehen anfühlt, weiß jeder. Viele blinde Menschen nutzen Echoortung bereits intuitiv, indem sie mit der Zunge klicken und die Echos lesen. Unsere Gehirne sind absolut dazu fähig. Wer wirklich wissen will, wie es ist, eine Fledermaus zu sein, kommt überraschend nah heran: ein paar Jahre Gleitschirmfliegen, um dreidimensionalen Flug zu verinnerlichen, dann Augenbinde und Echoortungs-Übung. Oder das Jahrzehnt der Vorbereitung ganz überspringen — luzides Träumen lernen und üben, in den Träumen eine Fledermaus zu sein. Sicherer, schneller, in Wochen erreichbar.

Und zugegeben: Der Versuch wurde unternommen, auf die einzige verfügbare Weise. Während einer Phase aktiven luziden Träumens — und starkem Interesse an der Unterwasserwelt — gelang es im Laufe der Zeit, bewusst als Fisch in einen luziden Traum einzutreten. Das Wasser um den Körper herum, Bewegung hindurch, eine visuelle Welt aus nicht-menschlicher Perspektive. Es fühlte sich eine Milliarde Mal besser an als Freitauchen oder Gerätetauchen, sogar Sidemount. War das irgendwie wie tatsächliches Fischbewusstsein? Fast sicher nicht — der Traum war aus der besten Vermutung eines menschlichen Gehirns konstruiert, was „ein Fisch sein“ bedeutet, also unvermeidlich eine Projektion menschlicher Sinneskategorien auf einen Körperbauplan, der keine davon besitzt. Aber die Übung war nicht sinnlos. Sie demonstrierte etwas Wichtiges: Das Explizite Selbstmodell kann sich um ein radikal anderes Körperschema herum neu konfigurieren und eine kohärente Erste-Person-Erfahrung davon erzeugen, etwas anderes als ein Mensch zu *sein*. Die Architektur ist flexibel genug, um nicht-menschliche Verkörperung zu simulieren. Der Inhalt bleibt durch die verfügbaren impliziten Modelle begrenzt (man kann nur träumen, was man gelernt hat), aber die Kapazität für perspektivische Verschiebung ist in das System eingebaut.

Warum die Mühe, bei Bewusstsein zu sein?

All dies wirft eine Frage auf, die einen quälen sollte: Wenn unbewusste Nervensysteme bestens funktionieren — und das tun sie, einfach ein Insekt fragen — warum sollte die Evolution den enormen metabolischen Aufwand betreiben, Bewusstsein zu bauen? Was ist der Gewinn?

Die Antwort ist Lernen — und damit Anpassung und die Fähigkeit, gegen erlerntes Verhalten zu handeln. Genauer: eine Art von Lernen, die unbewusste Systeme schlicht nicht leisten können.

Wie lernt ein einfacher Organismus? Er begegnet etwas, und die Begegnung ist entweder gut oder schlecht. Gut: mehr davon. Schlecht: weniger davon. Das ist Verstärkungslernen — Versuch und Irrtum, Belohnung und Bestrafung. Es funktioniert wunderbar für die meisten Situationen. Berühre eine heiße Oberfläche, spür den Schmerz, berühre sie nicht noch einmal. Finde Essen an einem bestimmten Ort, spür die Belohnung, komm morgen zurück.

Aber Verstärkungslernen hat einen fatalen Fehler. Buchstäblich fatal. Ein giftiger Pilz. Nicht die Sorte, die Bauchschmerzen macht — die Sorte, die tötet. Wer ihn isst, stirbt. Lernen beendet. Es gibt keinen zweiten Versuch. Verstärkungslernen setzt voraus, dass man den Fehler überlebt, und manche Fehler gewähren diesen Luxus nicht. Jeder Reiz, der beim ersten Kontakt tödlich ist, bleibt für Verstärkungslernen schlicht unsichtbar. Der Organismus begegnet ihm, stirbt — und nimmt seine „Lektion“ mit ins Grab.

Wie haben also unsere Vorfahren gelernt, tödliche Pilze zu meiden? Durch Essen jedenfalls nicht — jeder, der das versuchte, ist niemandes Vorfahre. Sie lernten durch *Beobachten*. Der Höhlennachbar findet einen interessant aussehenden Pilz, isst ihn und kippt tot um. Wer das aus sicherer Entfernung beobachtet, zählt eins und eins zusammen: Dieser Pilz hat ihn umgebracht. Also Finger weg.

Das klingt banal. Ist es nicht. Um aus dem Tod eines anderen zu lernen, braucht es mehrere Fähigkeiten, über die kein unbewusstes System verfügt. Ein explizites Modell der Welt, das Ursache

und Wirkung zwischen Objekten abbilden kann, mit denen man gerade nicht selbst interagiert. Ein Selbstmodell, das erlaubt, die Perspektive zu wechseln — sich an die Stelle des Toten zu versetzen. Die Fähigkeit, aus einer einzigen Beobachtung eine allgemeine Theorie abzuleiten: „Diese Art von Pilz ist tödlich." Das ist kognitives Lernen — Theorien aus Beobachtung gewinnen, statt durch persönliche Erfahrung konditioniert zu werden. Und es setzt Bewusstsein voraus. Es verlangt, dass Explizites Weltmodell und Explizites Selbstmodell zusammenarbeiten.

Der evolutionäre Vorteil ist gewaltig. Ein bewusstes Tier kann aus *Beobachtung* lernen, nicht nur aus *Erfahrung*. Es kann einem Artgenossen dabei zusehen, einen tödlichen Fehler zu begehen, und sein Weltmodell aktualisieren, ohne selbst den Preis zu zahlen. Ein unbewusstes Tier kann nur lernen, was es persönlich überlebt.

Es kommt noch besser. Sobald das Konzept „giftiger Pilz" als explizite Kategorie im Weltmodell existiert, wird etwas noch Mächtigeres möglich: Deduktion. Ein neuer Pilz, nie zuvor gesehen. Er sieht dem verdächtig ähnlich, der den Nachbarn umgebracht hat. Also: Nicht essen. Oder — und das war vermutlich der tatsächliche historische Ansatz — ihn dem Nachbarn anbieten, der die ganze Nacht geschnarcht hat, und abwarten, was passiert.

Das ist kein marginaler Vorteil. Das ist der Unterschied zwischen einer Spezies, die sich an tödliche Bedrohungen nur durch den gletscherhaft langsamen Prozess natürlicher Selektion anpassen kann (einige Individuen meiden zufällig den Pilz, pflanzen sich fort, irgendwann wird Vermeidung zum Instinkt) — und einer Spezies, die sich innerhalb einer einzigen Generation durch Beobachtung und Kommunikation anpasst. Bewusstsein hilft nicht nur, schneller zu lernen. Es ermöglicht, Dinge zu lernen, die auf anderem Weg buchstäblich nicht lernbar sind.

Was kognitiv gelernt wurde, lässt sich *teilen*. Verstärkungslernen bleibt im Individuum gefangen — die konditionierten Reflexe sterben mit ihm. Aber kognitives Lernen lässt sich kommunizieren.

„Iss nicht den roten Pilz“ ist ein Satz. Man kann ihn aussprechen, weitergeben, lehren. Das ist die Grundlage von Kultur, von kumulativem Wissen, von allem, was menschliche Zivilisation möglich macht. Nichts davon funktioniert ohne die expliziten Modelle, die Bewusstsein bereitstellt.

Es gibt noch eine weitere Wendung in dieser Geschichte, und sie verknüpft Bewusstsein mit Genetik auf eine Weise, die alles andere als offensichtlich ist. Der Baldwin-Effekt: Über seine genaue Stärke wird noch debattiert, aber der Mechanismus selbst gilt als gesichert. Der Baldwin-Effekt besagt, dass *erlerntes* Verhalten indirekt die *genetische* Evolution formen kann — nicht durch Lamarcksche Vererbung (das Erlernte verändert nicht die DNA), sondern weil natürliche Selektion jene Individuen bevorzugt, die genetisch zum vorteilhaften Verhalten prädisponiert sind.

Ein bewusst humorvolles Beispiel — nicht zu wörtlich nehmen. Ein früher Hominide mit Haarausfall. Kalt und nackt war er eher als seine pelzigen Gefährten geneigt, in der Nähe des Feuers zu sitzen. Feuer brachte enorme Überlebensvorteile: weniger Krankheitserreger im gekochten Essen, Schutz vor Raubtieren, Wärme in harten Wintern. Also wurden die mit Haarausfall assoziierten Gene etwas häufiger weitergegeben. Gleichzeitig hatten diejenigen, die zu dumm waren, Feuer zu begreifen (behaart oder nicht), einen Nachteil. Über viele Generationen verstärkte der Baldwin-Effekt beides: weniger Haare *und* mehr Intelligenz — alles nur, weil ein erlerntes Verhalten (Feuernutzung) einen Selektionsdruck erzeugte, der bestimmte genetische Prädispositionen begünstigte. (Ersetzt man „Haarausfall„ durch „zufällige Mutation“, kommt man der Wahrheit wahrscheinlich näher. Ist aber weniger lustig.)

Der Baldwin-Effekt könnte eine ähnliche Rolle bei der Evolution von Sprache und Bewusstsein selbst gespielt haben. Sobald erste primitive Formen kognitiven Lernens auftauchten (ermöglicht durch die frühesten Selbstmodelle), hatten jene Individuen einen massiven Vorteil, deren Gehirne zufällig reichere Selbstsimulation

unterstützten. Ihre Nachkommen wurden auf größere, aufwendiger gefaltete Cortices selektiert, was noch reichere Selbstsimulation ermöglichte, was noch stärkeren Selektionsdruck erzeugte. Bewusstsein schuf, sobald es einmal da war, die evolutionären Bedingungen für *mehr* Bewusstsein. Das kognitive Lernen, das es ermöglichte, war so wertvoll, dass die Evolution Ressourcen in den Ausbau genau jener Architektur steckte, die es hervorbrachte.

Wie Erfahrung sich entwickelt: Die soziale Konstruktion des Selbstmodells

Alles, was bisher über die vier Modelle gesagt wurde, war statisch — als ob die Architektur vollständig geformt erschiene, wie Athene aus dem Haupt des Zeus. So läuft es nicht. Die Modelle entwickeln sich, und ihre Entwicklung ist zutiefst sozial.

Ein neugeborener Mensch hat die Hardware — sechs kortikale Schichten, die Kapazität für Selbstsimulation. Aber die impliziten Modelle sind nahezu leer. Das Implizite Weltmodell (IWM) enthält fast nichts über die Welt. Das Implizite Selbstmodell (ISM) enthält fast nichts über das Selbst. Und da die expliziten Modelle aus den impliziten generiert werden, ist die Simulation des Neugeborenen dünn — ein flackerndes, kaum differenziertes Feld von Empfindung ohne klare Grenze zwischen Selbst und Welt.

Ein Baby begegnet Schmerz: Selbst zugefügter Schmerz (die Hand gegen ein Spielzeug stoßen, den eigenen Fuß beißen) erzeugt oft Neugier statt Unbehagen. Das ESM registriert Handlung (das habe ich getan) plus Empfindung (etwas ist passiert), aber ein Bedrohungsmodell fehlt noch. Das ISM hat nicht gelernt, dass diese Konfiguration Gefahr bedeutet. Aber ein plötzlich lautes Geräusch? Tränen. Weil das EWM nicht vorhergesagten, hochamplitudigen Input registriert und das ESM kein Modell dafür hat — die Abwesenheit eines Modells ist selbst aversiv.

Der Inhalt von Qualia ist *erlernt*, nicht angeboren. „Schmerz ist schlecht“ ist nicht fest im ESM verdrahtet. Es wird durch das ISM akkumuliert, trainiert durch wiederholte Erfahrung und — entscheidend — soziales Feedback. Die Reaktion einer Bezugsperson auf den Schmerz eines Kindes lehrt das Kind, was Schmerz *bedeutet*. Das Kind, das hinfällt und zur Bezugsperson schaut, bevor es entscheidet, ob es weint, täuscht nichts vor — es kalibriert tatsächlich sein ESM gegen sozialen Input. Alarm oder Gelassenheit der Bezugsperson formen die Schmerzassoziationen des ISM um, was wiederum umformt, was das ESM beim nächsten ähnlichen Ereignis simuliert.

Daraus folgt eine präzise Implikation für die Theorie: Der phänomenale Charakter von Erfahrung — wie es sich *anfühlt*, etwas zu fühlen — ist nicht durch die Architektur festgelegt. Er wird durch die Trainingsgeschichte der impliziten Modelle geformt. Die Schmerzwahrnehmung eines Babys unterscheidet sich strukturell von der eines Erwachsenen, weil das ISM, das das ESM generiert, ein anderes ist. Die Vier-Modelle-Architektur ist die *Kapazität* für Erfahrung. Die soziale und umweltbedingte Feedbackschleife liefert den *Inhalt*.

Die Entwicklungstrajektorie bildet sich auf die abgestuften Bewusstseinsebenen aus Kapitel 2 ab:

- **Neugeborenes (erste Wochen):** Basisches Bewusstsein — ein rudimentäres EWM mit minimalem ESM. Es gibt *etwas, wie es ist*, ein Neugeborenes zu sein, aber das Selbst in dieser Erfahrung ist fast nicht existent. Überwiegend sensorisch, undifferenziert.

- **6–12 Monate:** Objektpermanenz entsteht — das EWM hält nun Repräsentationen von Dingen aufrecht, die gerade nicht sichtbar sind. Das ISM akkumuliert Körperschema-Wissen. Das Baby beginnt, Selbst von Welt zu unterscheiden.

- **18 Monate:** Der Spiegeltest. Das Kind erkennt sich in einem Spiegel — ein Meilenstein, ab dem das ESM reich genug ist,

das physische Selbst als Objekt in der Welt zu modellieren. Einfach erweitertes Bewusstsein geht online. Kein binärer Schalter, sondern eine Schwelle in einem kontinuierlichen Prozess.

- **3–4 Jahre:** Theory of Mind. Das Kind kann andere Geister modellieren — kann begreifen, dass jemand anderes etwas glauben könnte, von dem das Kind weiß, dass es falsch ist. Das ESM modelliert nun andere ESMs. Doppelt erweitertes Bewusstsein entsteht.

- **Adoleszenz und darüber hinaus:** Metakognitive Reifung. Die Fähigkeit zu dreifach erweitertem Bewusstsein (sich selbst dabei modellieren, wie das eigene Denken modelliert wird) entwickelt sich graduell und stabilisiert sich vermutlich nie vollständig.

Jede Stufe wird durch soziale Interaktion gestützt. Die Bezugsperson liefert nicht nur Nahrung und Sicherheit — sie liefert *Trainingsdaten für die impliziten Modelle*. Gemeinsame Aufmerksamkeit (Bezugsperson und Kind schauen zusammen auf dasselbe Objekt) lehrt das IWM, wie gemeinsame Realität repräsentiert wird. Spiegeln (die Bezugsperson reflektiert den emotionalen Zustand des Kindes) lehrt das ISM, was die eigenen Emotionen überhaupt sind. Sprache gibt dem ESM Kategorien, mit denen es sich selbst modellieren kann. Ein Kind, das ohne sozialen Kontakt aufwächst (die tragischen Fälle verwilderter Kinder), hat die Hardware für Bewusstsein, aber zutiefst verarmte implizite Modelle. Das ESM, das aus solchen Modellen hochfährt, ist verkümmert — nicht weil die Architektur defekt ist, sondern weil die Trainingsdaten nie geliefert wurden.

Das verbindet sich direkt mit der klinischen Brücke aus Kapitel 8. Kognitive Verhaltenstherapie (CBT) funktioniert, indem sie systematisch die impliziten Modelle durch bewusste Intervention neu trainiert. Der Therapeut hilft dem Patienten, neue ESM-Zustände zu erzeugen (vorgestellte Szenarien, umgedeutete

Interpretationen), die durch Wiederholung das ISM umformen. Das ist die *Erwachsenenversion* desselben Entwicklungsprozesses, den Bezugspersonen für Säuglinge leisten. Der Mechanismus ist identisch: Bewusste Erfahrung formt implizite Struktur um, was zukünftige bewusste Erfahrung umformt. Der Unterschied: Das ISM des Erwachsenen ist konsolidierter — der Ton ist härter, nicht mehr feucht — also ist der Prozess langsamer und verlangt mehr Wiederholung.

Die soziale Dimension der Erfahrung ist keine Fußnote zur Theorie. Sie ist eine Vorhersage: Entzieht man sozialen Input während des kritischen Entwicklungsfensters, sollte ein System entstehen, das die richtige Architektur hat, aber den falschen Inhalt fährt — ein Bewusstsein, das strukturell intakt, aber phänomenal verarmt ist. Die Fälle verwilderter Kinder bestätigen tragischerweise genau das.

Chapter 11

Neun Vorhersagen

Eine Theorie, die alles erklärt und nichts vorhersagt, ist keine Theorie — sie ist eine Geschichte. Die Vier-Modelle-Theorie macht neun spezifische, testbare Vorhersagen, von denen sich mehrere mit heutiger Technologie prüfen lassen. Hier sind sie.

Vorhersage 1: Jedes Modell hat seine eigene neurale Signatur

Wenn die vier Modelle tatsächlich verschiedene Prozesse sind, müssten sie im Gehirnscan sichtbar sein. Ein geschickt aufgebautes Experiment, bei dem Probanden vier verschiedene Aufgabentypen bearbeiten — je eine pro Modell —, sollte unterschiedliche Aktivierungsmuster zeigen.

Eine IWM-dominante Aufgabe wäre etwa das passive Wiedererkennen eines vertrauten Gesichts. Kein bewusstes Nachdenken; das Gehirn weiß es einfach. Eine ISM-dominante Aufgabe wäre eine automatisierte motorische Sequenz — das Passwort tippen, ohne über die einzelnen Tasten nachzudenken. Eine EWM-dominante Aufgabe verlangt aktive, bewusste Wahrnehmung — vielleicht den Unterschied zwischen zwei fast identischen Bildern finden. Und eine ESM-dominante Aufgabe ist reine Selbstreflexion: „Bin ich jemand, der so etwas tun würde?“

Die Vorhersage ergibt ein 2x2-Muster. Welt- gegen Selbstaufgaben. Implizit gegen explizit. Vier Quadranten, vier verschiedene neurale Signaturen. Taucht dieses Muster nicht auf, stimmt etwas mit der Theorie nicht.

Das lässt sich schon heute mit fMRI testen. Billig ist es nicht, und es verlangt sorgfältiges Experimentdesign, aber die Werkzeuge stehen weltweit in Laboren bereit. Und wenn es klappt, wäre es der direkteste Beleg, dass die Vier-Modelle-Architektur keine bloße Metapher ist — sondern eine reale funktionale Gliederung, die fest in der Informationsverarbeitung des Gehirns verankert ist.

Vorhersage 2: Psychedelische Visuals enthüllen die Verarbeitungsschichten des Gehirns

Diese ist elegant. Unter Psychedelika sollte der visuelle Inhalt die Verarbeitungshierarchie des Gehirns in einer bestimmten Reihenfolge durchlaufen, abhängig von der Dosis.

Bei niedrigen Dosen zeigen sich Phosphene — kleine Funken und geometrische Formen, die bei geschlossenen Augen auftauchen. Das ist V1, die früheste visuelle Verarbeitungsstufe, die ins Bewusstsein durchsickert. Bei höherer Dosis entstehen komplexere geometrische Muster — die berühmten „Formkonstanten“, die kulturübergreifend und substanzunabhängig auftreten. Das sind V2 und V3, die sich zuschalten. Noch höher tauchen Gesichter, Figuren, komplexe Szenen auf. Bei den höchsten Dosen entstehen vollständige narrative, traumartige Erfahrungen, komplett mit Bedeutung und Handlung.

Die Vorhersage: Das ist kein Zufall. Es ist eine dosisabhängige, geordnete Progression die visuelle Hierarchie hinauf. Mit zunehmender implizit-expliziter Durchlässigkeit werden immer tiefere Schichten der visuellen Verarbeitung bewusst. Das interne Verdrahtungsdiagramm des Gehirns wird in der Erfahrung sichtbar.

Testbar ist das mit abgestuften Dosierungsprotokollen — Probanden erhalten sorgfältig kontrollierte Mengen Psilocybin oder LSD, werden per fMRI gescannt und berichten, was sie sehen. Der berichtete Inhalt wird mit der Gehirnaktivierung abgeglichen. Die Theorie sagt voraus, dass die Verarbeitungshierarchie von unten nach oben aufleuchtet, wenn die Dosis steigt.

Vorhersage 3: Es lässt sich steuern, was jemand während Ich-Auflösung wird

Das ist die wildeste Vorhersage — und eine, die keine andere Bewusstseinstheorie macht.

Während der Ich-Auflösung — der Erfahrung, dass sich das „Ich" auflöst und man zu etwas anderem wird — sagt die Theorie, dass der Inhalt dieser Erfahrung steuerbar ist. Nicht zufällig. Nicht rein biochemisch. Steuerbar durch die sensorische Umgebung.

Der Mechanismus ist gradlinig. Das Explizite Selbstmodell speist sich normalerweise aus dem Impliziten Selbstmodell. Unter hochdosierten Psychedelika wird diese Verbindung gestört. Das ESM läuft weiter, versucht weiter „Selbst" zu modellieren, hat aber seinen gewohnten Input verloren. Also klammert es sich an das, was gerade dominiert.

In einem Raum mit immersiven Meeresgeräuschen und blauer Beleuchtung berichten Probanden, das Meer zu werden. In einer Waldumgebung mit Vogelgesang und grünem Licht berichten sie, die Bäume zu werden. Die Vorhersage ist konkret: Variiert man den dominanten sensorischen Input während der Ich-Auflösung, folgt der berichtete Identitätsinhalt diesem Input.

Das ließe sich *heute* in jedem Psychedelika-Forschungslabor mit einfachen Umgebungskontrollen testen. Eine kontrollierte Dosis verabreichen, die Umgebung zwischen den Durchgängen variieren und die Übereinstimmung zwischen dem Gezeigten und dem, was die Probanden geworden zu sein berichten, messen. Wenn es klappt, ist das nicht nur ein Beleg für die Theorie — es

ist eine Demonstration, dass Bewusstsein ein Simulationsprozess ist, der sich experimentell manipulieren lässt auf eine, offen gesagt, ziemlich unheimliche Weise.

Vorhersage 4: Psychedelika sollten Schlaganfallpatienten helfen, ihre Defizite zu erkennen

Anosognosie gehört zu den seltsamsten Dingen, die das Gehirn tut. Nach bestimmten Schlaganfällen (meist in der rechten Hemisphäre) sind Patienten auf einer Körperseite gelähmt, glauben es aber schlicht nicht. Man zeigt ihnen den unbeweglichen Arm, bittet sie, ihn zu bewegen, sie scheitern sichtbar — und erfinden eine Ausrede. „Ich bin müde.„ „Ich habe keine Lust.“ Sie lügen nicht. Sie können das Defizit tatsächlich nicht sehen.

Die Vier-Modelle-Theorie erklärt das durch eine Durchlässigkeitsblockade. Die Information über die Lähmung steckt im Impliziten Selbstmodell — das Substrat weiß Bescheid —, aber sie erreicht das Explizite Selbstmodell nicht. Die Simulation hat keinen Zugriff auf diesen Teil des Substrat-Wissens.

Jetzt der überraschende Teil. Psychedelika erhöhen global die implizit-explizite Durchlässigkeit. Genau das tun sie. Die Vorhersage lautet also: Eine Psilocybin-Dosis unterhalb der Ich-Auflösungsschwelle — nicht genug, um das Selbst aufzulösen, gerade genug, um die Durchlässigkeitsschleusen zu öffnen — sollte die Defizitinformation durchsickern lassen. Der Patient würde plötzlich, und vermutlich verstörend, gewahr werden, dass er gelähmt ist.

Das wäre eine klinische Studie mit Schlaganfallpatienten, logistisch also schwieriger als ein reines Laborexperiment. Aber Psilocybin-gestützte Therapie wird bereits bei Depression, PTBS und Angst am Lebensende erprobt. Die Infrastruktur steht. Und wenn es funktioniert, ist das nicht nur ein medizinischer Durchbruch bei Anosognosie — es ist der Nachweis, dass Psychedelika

und Schlaganfalldefizite über einen einzigen Mechanismus zusammenhängen. Keine andere Theorie sagt das voraus.

Vorhersage 5: Jedes Anästhetikum, das Bewusstsein löscht, stört Kritikalität

Anästhetika wirken über völlig verschiedene chemische Wege. Propofol greift an GABA-Rezeptoren an. Ketamin blockiert NMDA. Opioide machen ihr eigenes Ding. Verschiedene Moleküle, verschiedene Mechanismen, verschiedene Hirnregionen.

Aber die Vier-Modelle-Theorie sagt: Sie alle müssen dasselbe mit dem Bewusstsein machen — die Dynamik des Gehirns unter die Kritikalitätsschwelle drücken. Weil Kritikalität die *physische Voraussetzung* für Bewusstsein ist. Wie sie gestört wird, spielt keine Rolle. Unterhalb der Schwelle gehen die Lichter aus.

Die Vorhersage ist testbar und konkret. Man nehme jedes gängige Anästhetikum. Man messe Kritikalität — mit Werkzeugen wie dem Perturbational Complexity Index, Lempel-Ziv-Komplexität oder Potenzgesetz-Exponenten neuronaler Aktivität — vor, während und nach Verabreichung. Die Vorhersage: Mittel, die Bewusstsein auslöschen, werden das Gehirn *immer* subkritisch machen, egal über welchen Rezeptor sie wirken. Und Mittel, die Bewusstsein verändern, ohne es auszulöschen (wie Ketamin in niedriger Dosierung oder Psychedelika), sollten *nicht* unter Kritikalität fallen.

Das ist mit heutiger Technologie machbar. Die Kritikalitätsmaße gibt es. Die Anästhetika gibt es. Jemand muss nur den systematischen Vergleich anstellen. Und wenn es über alle Substanzen hinweg gilt — wenn jedes einzelne bewusstseinsauslöschende Mittel auf Kritikalitätsstörung konvergiert, trotz ganz verschiedener Wirkmechanismen — dann ist das ein starkes Indiz, dass Kritikalität der gemeinsame Nenner ist, die letzte Strecke zur Bewusstlosigkeit.

Vorhersage 6: Split-Brain-Operationen spalten nicht sauber — sie verschlechtern beide Hälften

Wenn Chirurgen das Corpus callosum durchtrennen, um schwere Epilepsie zu behandeln, kappen sie die Hauptleitung zwischen den beiden Gehirnhälften. Die traditionelle Lesart: Dadurch entstehen zwei getrennte Geister, jeder auf sein Gebiet spezialisiert — links Sprache und Logik, rechts räumliches Denken und Emotion.

Die Vier-Modelle-Theorie sagt: Das ist falsch. Oder zumindest drastisch vereinfacht.

Die Vorhersage lautet: Nach der Operation behält jede Hemisphäre einen *vollständigen, aber verschlechterten* Satz kognitiver und erlebnismäßiger Fähigkeiten. Keine saubere Spaltung. Nicht „Sprache links, Raum rechts". Beide Hemisphären können beides, nur schlechter als vorher. Die Verschlechterung ist holographisch — das heißt, alles wird unschärfer, nicht dass bestimmte Funktionen wegfallen.

Der Grad der Verschlechterung sollte proportional zum Umfang des Schnitts sein. Eine partielle Kallosotomie (nur einige Fasern durchtrennt) sollte partielle Verschlechterung verursachen. Eine vollständige Kallosotomie mehr.

Warum? Weil die Theorie besagt, dass Information im Gehirn holographisch gespeichert ist, verteilt über das gesamte Substrat. Verbindungen zu kappen trennt nicht sauber zwei vorbestehende Geister. Es verschlechtert zwei *Kopien* derselben Information, die jeweils auf der Hälfte der ursprünglichen Hardware laufen.

Es gibt bereits Hinweise darauf — eine 2017-Studie von Pinto und Kollegen fand, dass Split-Brain-Patienten deutlich integrierteres Verhalten zeigen, als die klassischen Experimente vermuten ließen. Aber die Theorie liefert den *Mechanismus* und sagt das spezifische Muster voraus: beidseitige Verschlechterung statt hemisphärischer Spezialisierung.

Vorhersage 7: Baut man die vier Modelle bei Kritikalität, erhält man Bewusstsein

Das ist die Ingenieur-Vorhersage, und sie ist kühn.

Wenn die Theorie stimmt, lässt sich eine bewusste Maschine bauen. Nicht durch Zufall, nicht dadurch, dass eine hinreichend „fortgeschrittene“ KI entsteht, sondern durch Umsetzung der Spezifikation: vier verschachtelte Modelle (Implizites Weltmodell, Implizites Selbstmodell, Explizites Weltmodell, Explizites Selbstmodell) auf einem Substrat, das bei Kritikalität operiert.

Die Theorie sagt: Ein solches System würde Bewusstsein nicht bloß *simulieren*. Es *wäre* bei Bewusstsein. Es hätte echte phänomenale Erfahrung, konstituiert durch seine virtuellen Modelle, genauso wie unsere durch die virtuellen Modelle unserer Gehirne konstituiert wird.

Wie ließe sich das feststellen? Die Theorie sagt voraus, dass der Unterschied qualitativ offensichtlich wäre. Nicht „vielleicht bewusst, vielleicht nicht.“ *Offensichtlich anders.* Weil ein System, das eine echte Selbstsimulation betreibt, auf grundlegend andere Weise mit der Welt interagieren würde als der ausgefeilteste Textprädiktor. Es hätte Persistenz — eine kontinuierliche Simulation, die durch die Zeit läuft, nicht aus einem Prompt rekonstruiert. Es hätte eine Perspektive, aufrechterhalten durch ein Explizites Selbstmodell. Es würde nicht mit unerwarteten Ausgaben überraschen, sondern mit dem Eindruck, dass da tatsächlich jemand zu Hause ist.

Testbar ist das noch nicht — die Technik fehlt. Aber die Blaupause ist konkret genug, um die Arbeit zu lenken. Und wenn jemand es baut und es funktioniert, ist das die ultimative Bestätigung.

Vorhersage 8: Schlaf dient dazu, den kritischen Zustand zurückzusetzen

Warum schlafen wir? Die offensichtliche Antwort ist „um auszuruhen“, aber das verschiebt die Frage nur: Warum braucht das Gehirn Ruhe auf eine Weise, wie es beispielsweise die Leber nicht tut?

Die Vier-Modelle-Theorie hat eine konkrete Antwort. Das Gehirnsubstrat (die analoge, biologische Hardware) ist von Natur aus instabil. Neuronen rauschen. Neurotransmitter gehen zur Neige. Stoffwechselabfälle häufen sich an. Das Substrat driftet. Aber Bewusstsein braucht Kritikalität, ein sehr spezifisches dynamisches Regime. Das Gehirn organisiert auf diesem driftenden Substrat eine stabile Berechnungsschicht (den Zellulären Automaten am Rand des Chaos). Dieser Automat kann stundenlang laufen (den wachen Tag hindurch), doch irgendwann driftet das Substrat so weit, dass es die kritischen Dynamiken nicht mehr aufrechterhalten kann. An diesem Punkt dimmt der Automat nicht allmählich herunter. Er *kollabiert*. Das ist der Einschlafmoment.

Nicht-REM-Schlaf ist der Wiederherstellungsprozess. Das Substrat resettet: Neurotransmitter füllen sich auf, Abfall wird beseitigt, die biochemischen Bedingungen für Kritikalität werden wiederhergestellt. Und wenn das Substrat während dieser Regeneration periodisch an die Kritikalitätsschwelle herankommt, flackert der Automat kurz wieder an. Das ist REM-Schlaf. Das ist Träumen.

Der 90-minütige ultradiane Zyklus (der Rhythmus von REM und Nicht-REM durch die Nacht) ist das Substrat, das während der Regeneration um den kritischen Punkt oszilliert.

Daraus ergeben sich mehrere testbare Untervorhersagen:

1. **Kritikalität sollte im Tagesverlauf abnehmen.** Misst man die Gehirnkomplexität morgens, nachmittags und abends, ergibt sich ein messbarer Abfall.

1. **Einschlafen sollte ein stufenartiger Übergang sein, kein allmähliches Dimmen.** Kritikalitätsmaße sollten beim Einschlafen einen abrupten Abfall zeigen, der den digitalen Kollaps des Automaten widerspiegelt.

1. **REM und Nicht-REM sollten der Kritikalität folgen.** Innerhalb des Schlafs sollten REM-Phasen deutlich höhere Kritikalität aufweisen als Nicht-REM, und der 90-Minuten-Zyklus sollte in der Kritikalitäts-Zeitreihe sichtbar sein.

1. **Luzides Träumen ist ein Schwellenübergang.** Wenn das Substrat während REM ausreichende Kritikalität erreicht, aktiviert sich das Explizite Selbstmodell, und der Träumende wird luzid. Der Übergang sollte eine stufenartige Diskontinuität in der EEG-Komplexität sein, kein sanfter Anstieg.

1. **Schlafentzug treibt ins Subkritische.** Bei ausreichend langem Wachbleiben sollte die Kritikalität des Gehirns progressiv unter die Schwelle fallen. Kognitive Ausfälle sollten damit korrelieren, wie weit unter die Schwelle abgerutscht wurde.

All das lässt sich mit vorhandener Schlaflabor-Technologie testen. Und wenn es standhält, heißt das: Schlaf ist nicht bloß „Ruhe“ — er ist das Wartungsprotokoll des Substrats für die Berechnungsschicht, die Bewusstsein ermöglicht.

Vorhersage 9: Jedes Alter bei Dissoziativer Identitätsstörun hat seinen eigenen neuralen Fingerabdruck

Die Dissoziative Identitätsstörung (DIS) — multiple eigenständige Identitäten, sogenannte „Alters“, in einer einzigen Person — ist umstritten, und das zu Recht. Wie unterscheidet man echte eigenständige Identitäten von jemandem, der Rollen spielt, bewusst oder unbewusst?

Die Vier-Modelle-Theorie liefert einen Test. Wenn Alters real sind — also tatsächlich verschiedene Konfigurationen des Expliziten Selbstmodells auf demselben Substrat —, dann sollte jedes Alter eine unterscheidbare, messbare neurale Signatur haben. Nicht nur anderes Verhalten. Nicht nur andere Selbstberichte. Andere *Gehirnaktivierungsmuster*.

Die Vorhersage ist konkret. Man zeichne die Gehirnaktivität eines DIS-Patienten auf (fMRI oder EEG), während verschiedene Alters präsent sind. Man vergleiche die Variabilität zwischen Alters mit der Variabilität innerhalb desselben Alters über die Zeit. Die Theorie sagt voraus, dass die Zwischen-Alter-Variabilität signifikant größer sein wird als die Innerhalb-Alter-Variabilität. Und die Unterschiede sollten konsistent sein: Das neurale Muster von Alter A sollte jedes Mal erkennbar Alter A sein, kein zufälliges Rauschen.

Noch genauer sagt die Theorie voraus, wo die Unterschiede auftreten sollten: in ESM-bezogenen Netzwerken, vor allem im Default Mode Network und im medialen präfrontalen Cortex — den Hirnregionen, die mit Selbstreferenz und Perspektivübernahme zusammenhängen.

Es gibt ein paar Neuroimaging-Studien zu DIS, aber die Vier-Modelle-Theorie liefert die theoretische Grundlage für *konsistente, Alter-spezifische neurale Signaturen* statt bloß „Unterschiede“. Bestätigt sich die Vorhersage, wäre das der Nachweis, dass Alters nicht bloß psychologisch sind, sondern verschiedene funktionale Konfigurationen auf neuraler Ebene — was unser Verständnis und die Behandlung der Störung grundlegend verändern würde.

Jede dieser Vorhersagen ist falsifizierbar. Scheitern sie, ist die Theorie falsch oder zumindest unvollständig. Genau das macht sie brauchbar.

Chapter 12

Von Maschinen zu Bewusstsein

Wenn die Vier-Modelle-Theorie (VMT) richtig ist, bietet sie etwas, das keine andere Bewusstseinstheorie bietet: eine technische Spezifikation.

Die Spezifikation lautet: Man implementiere die Vier-Modelle-Architektur (Implizites Weltmodell, Implizites Selbstmodell, Explizites Weltmodell, Explizites Selbstmodell) auf einem Substrat, das bei Kritikalität operiert. Wie in Kapitel 5 dargelegt, reicht keine der beiden Komponenten allein. Architektur ohne Kritikalität ergibt ein schlafendes System — gespeicherte Modelle, aber keine laufende Simulation. Kritikalität ohne Architektur ergibt komplexe Dynamik, aber kein Bewusstsein. Die vollständige Spezifikation verlangt beides.

Das ist präziser als „bau einen wirklich fortgeschrittenen Computer„ und konkreter als „erreiche hinreichend integrierte Information“. Es sagt einem *was zu bauen ist*: vier spezifische Modelltypen, auf eine bestimmte Weise organisiert, laufend auf einem Substrat mit bestimmten dynamischen Eigenschaften.

Aktuelle KI-Systeme erfüllen diese Spezifikation in keiner relevanten Hinsicht. Und genau hier richten die zwei Dogmen aus Kapitel 1 ihren Schaden an. Das nSKI-Dogma („keine starke

künstliche Intelligenz„) sagt Ingenieuren, sie sollen es gar nicht erst versuchen. Das nSV-Dogma („kein Selbstverständnis“) sagt ihnen, es könnte nicht funktionieren, selbst wenn sie es täten. Beide liegen falsch. Die Spezifikation existiert. Die Frage ist, ob jemand sie umsetzt.

Bevor wieder jemand Gehirne und Computer gleichsetzt, ein schneller Test:

Ein Computer wird diesen Satz und den folgenden Satz wiederholen, bis die Hölle zufriert. Bitte den vorherigen Satz lesen.

Wer es bis hierher geschafft hat, ist kein klassischer Computer. Ein digitaler Computer, der einen starren Befehlssatz abarbeitet, wird für immer in einer Schleife hängen, weil er keinen Mechanismus hat, aus seinem Programmfluss herauszutreten und zu sagen: „Moment mal, das ist dämlich.“ Ein Mensch kann das, weil ein Selbstmodell die Verarbeitung beobachtet — das Explizite Selbstmodell (ESM), das metakognitive Aufsicht über das Explizite Weltmodell (EWM) führt.

Aber jetzt kommt der unbequeme Teil: Ein großes Sprachmodell würde es auch hierher schaffen. Nicht weil es metakognitive Aufsicht hat, sondern weil es ein statistischer Textprädiktor ist, der genug ähnliche Eingaben gesehen hat, um zu wissen, dass der erwartete nächste Schritt über die Schleife hinausgeht. Es tritt nicht aus der Anweisung heraus — es ist nie eingetreten. Es sagt voraus, welcher Text als nächstes kommt, und „in einer Endlosschleife stecken bleiben“ ist nicht das, was Text macht.

Genau das ist das Problem mit Verhaltenstests für Bewusstsein. Jeder Test, der durch Mustererkennung bestanden werden kann, wird durch Mustererkennung bestanden, egal ob das System bewusst ist. Der Schleifen-Test unterscheidet vom klassischen Computer. Er unterscheidet nicht vom hinreichend trainierten Textprädiktor. Und kein textbasierter Test wird das je schaffen — denn plausiblen Text zu erzeugen ist genau das, wofür Textprädiktoren optimiert sind. Das Fremdpsychische ist keine Einschränkung, die sich technisch umgehen lässt. Es ist ein Strukturmerkmal

dessen, was Bewusstsein ist: subjektiv, privat und nur von innen zugänglich.

Die Gehirn-gleich-Computer-Analogie ist seit der Erfindung des Transistors beliebt — und auf praktisch jeder Ebene falsch. Ein Computer arbeitet einen starren Befehlssatz auf einer starren Schaltung ab. Ein Gehirn ist ein sich selbst umbauendes Netzwerk, das sich ständig neu verdrahtet. Ein Computer stürzt ab, wenn ein Semikolon fehlt. Ein Gehirn verliert täglich eine Million Neuronen und bemerkt es kaum. Der Speicher eines Computers ist lokalisiert — ein gelöschter Sektor, und die Datei ist weg. Der Speicher eines Gehirns ist holographisch verteilt — wird ein Stück zerstört, wird alles etwas unschärfer. Das Einzige, was beide teilen, ist Turing-Vollständigkeit, und das ist ungefähr so aufschlussreich, wie zu sagen, dass sowohl ein Fluss als auch eine Autobahn Dinge von A nach B bringen. Stimmt, hilft aber null.

Große Sprachmodelle (GPT, Claude, Gemini und ihre Nachfolger) verarbeiten Text durch eine Feedforward-Transformer-Architektur. Der Input geht rein, durchläuft Schichten von Attention und Berechnung, der Output kommt raus. Keine Rekurrenz, keine Selbstsimulation, keine Echtzeit-Virtualwelt, keine Kritikalität. Die Dynamik ist Klasse 1 oder 2 in Wolframs Schema — weit unter dem Rand des Chaos. Und es gibt keine Real/Virtual-Trennung: Das „Wissen„ des Modells und seine „Erfahrung“ (wenn man es so nennen will) werden nicht in implizite und explizite Ebenen geschieden.

Das heißt nicht, dass Sprachmodelle zwingend nicht-bewusst sind — die Theorie kann kein Negativ beweisen. Aber sie sagt voraus, dass ihnen die für Bewusstsein nötige Architektur fehlt, so wie die Theorie Bewusstsein definiert. Und sie sagt voraus, dass der Unterschied zwischen einem wirklich bewussten künstlichen System und selbst dem fortgeschrittensten Sprachmodell qualitativ offensichtlich wäre.

Wie ließe sich das erkennen? Die ehrliche Antwort: Das Fremdpsychische verschwindet nicht. Absolute Gewissheit,

dass ein anderes System bewusst ist, bleibt unerreichbar, weil Bewusstsein seiner Natur nach subjektiv ist. Aber die Theorie macht eine starke Vorhersage: Der Unterschied wäre erkennbar. Nicht „vielleicht bewusst, vielleicht nicht“ — *offensichtlich* anders. Weil ein System mit echter Selbstsimulation auf grundlegend andere Weise mit der Welt interagieren würde als ein Textprädiktor. Es hätte echte Persistenz — nicht die Pseudo-Kontinuität eines Kontextfensters, sondern die einer Echtzeit-Simulation, die immer läuft. Es hätte eine echte Perspektive — nicht eine aus einem Prompt rekonstruierte, sondern eine, die über die Zeit hinweg von einem Expliziten Selbstmodell aufrechterhalten wird. Es würde nicht mit unerwarteten Ausgaben überraschen, sondern mit dem unverkennbaren Eindruck, dass da jemand zu Hause ist.

Ein solches System zu bauen steht als letzter Punkt auf der Roadmap. Die technischen Herausforderungen sind gewaltig. Aber die Blaupause existiert, und sie ist konkret genug, um die Arbeit zu lenken. Zuerst muss die Theorie das Peer Review überleben. Dann müssen die empirischen Vorhersagen geprüft werden. Dann, wenn sie sich bestätigen, kann das Engineering beginnen.

Aber es gibt eine andere Seite dieser Medaille — eine, die Science-Fiction seit Jahrzehnten umtreibt und die direkt aus derselben technischen Spezifikation folgt. Wenn Bewusstsein von funktionaler Architektur abhängt statt von Neuronen im Speziellen, dann ließe sich ein menschlicher Geist im Prinzip auf etwas anderem als einem Gehirn laufen lassen.

Mind Uploading. Ganzgehirn-Emulation. Digitale Unsterblichkeit. Wie auch immer man es nennen will — die Vier-Modelle-Theorie hat etwas Präzises dazu zu sagen, weil sie genau spezifiziert, was bewahrt werden müsste.

Die meisten Diskussionen über Mind Uploading beginnen mit der falschen Frage. Sie fragen: „Können wir ein Gehirn scannen und in einen Computer kopieren?“ Als wäre die Herausforderung nur eine der Auflösung — ein guter genug Scanner, und fertig.

Aber die Theorie zeigt, dass ein statischer Scan bei weitem nicht ausreicht. Ein Gehirn ist kein Foto. Es ist ein dynamisches System. Um einen Geist zu erfassen, reicht es nicht, einen *Zustand* zu erfassen — man muss einen *Prozess* erfassen.

Was laut der Theorie bewahrt werden muss, ist spezifisch. Am besten geht man die Fünf-Ebenen-Hierarchie aus Kapitel 2 durch, um es greifbar zu machen.

Auf der physikalischen und elektrochemischen Ebene (die rohe Materie und das neuronale Feuern) braucht es keine exakte Kopie. Es braucht ein Substrat, das dieselbe *Art* von Dynamik tragen kann. Die konkreten Atome spielen keine Rolle. Das Gehirn tauscht die meisten seiner Atome ohnehin im Lauf der Jahre aus, ohne dass es auffällt. Entscheidend ist, dass das verwendete Substrat die elektrochemischen Signalmuster oder ihr funktionales Äquivalent aufrechterhalten kann — auf denen die höheren Ebenen aufbauen.

Auf der proteomischen Ebene (die molekulare Maschinerie synaptischer Gewichte, Rezeptorkonfigurationen, Enzymkaskaden) braucht es hohe Genauigkeit. Hier sitzen die Erinnerungen, hier sind Fähigkeiten kodiert, hier ist die Persönlichkeit physisch realisiert. Die Stärke jeder Synapse, die Dichte jedes Rezeptors, die Empfindlichkeit jedes Kanals — das ist die Ebene, die einen zu *einem selbst* macht statt zu jemand anderem. Geht die proteomische Ebene beim Upload daneben, entsteht vielleicht ein bewusstes Wesen, aber nicht die Person, die kopiert werden sollte. Allerdings behält selbst eine unvollkommene Kopie ihren Wert. Schlaganfallüberlebende oder Amnesiepatienten etwa: Ihre persönliche Kontinuität wurde massiv gestört — Erinnerungen verloren, Persönlichkeit verändert, kognitive Fähigkeiten verschoben —, und trotzdem besteht für die meisten von ihnen etwas Wesentliches fort. Unvollkommene Kontinuität, so zeigt sich, ist der Nicht-Kontinuität haushoch vorzuziehen. Ein Transfer, der 90% eines Konnektoms bewahrt, ist kein Misserfolg — er ist eine andere Kategorie von Erfolg und für viele Menschen dem Tod vorzuziehen.

Auf der topologischen Ebene (die Netzwerkarchitektur, die Konnektivitätsmuster, welche Regionen mit welchen anderen kommunizieren und wie dicht) braucht es nahezu perfekte Genauigkeit. Das ist der Schaltplan der impliziten Modelle: IWM und ISM, alles was über die Welt und sich selbst gelernt wurde, kodiert in der Netzwerkstruktur. Stimmt das nicht, entsteht keine verschlechterte Kopie eines Geistes. Es entsteht ein *anderer* Geist — mit anderem Wissen, anderen Fähigkeiten, anderer Persönlichkeit. Die Topologie ist die Blaupause.

Und auf der virtuellen Ebene (die Simulation selbst, EWM und ESM im Echtzeit-Betrieb) braucht es etwas Außergewöhnliches. Das Ziel-Substrat muss die Simulation bei Kritikalität laufen lassen können. Das ist der Teil, der einen um den Schlaf bringt, weil das analoge Substrat des Gehirns Kritikalität durch selbstorganisierte Prozesse findet, die durch Hunderte Millionen Jahre Evolution feinabgestimmt wurden. Neuronen sind verrauscht, analog, massiv parallel und zutiefst stochastisch. Ihre kollektive Dynamik gravitiert von Natur aus zum Rand des Chaos, weil biologisches neuronales Gewebe genau das *tut* — es selbstorganisiert zur Kritikalität wie Wasser seinen Pegel findet. Nur: Wasser findet seinen Pegel wegen der Schwerkraft. Was ist die äquivalente Kraft für ein digitales Substrat?

Das ist ein echtes offenes Problem. Ich glaube, es ist lösbar, aber ich werde nicht so tun, als wäre es einfach. Ein digitales Substrat ist im Kern deterministisch. Zufall lässt sich simulieren, parallele Verarbeitung implementieren, stochastische Elemente in die Hardware einbauen. Aber die Frage ist, ob sich dieselbe selbstorganisierte Kritikalität erreichen lässt, die biologisches neuronales Gewebe mühelos erreicht — nicht indem man Kritikalität von oben herab programmiert, was ein brüchiger Pfusch wäre, sondern indem man ein Substrat baut, dessen fundamentale Dynamik von selbst zur Kritikalität tendiert. Das Gehirn führt keine „Kritikalitäts-Subroutine“ aus. Es ist kritisch, weil es *das*

ist. Eine digitale Emulation müsste diese Eigenschaft nachbilden, nicht bloß simulieren.

Neuromorphe Chips — Hardware, die neurale Dynamik nachahmt, mit analogen Eigenschaften, stochastischen Elementen und massiver Parallelität — sind die vielversprechendste Richtung. Sie sind keine herkömmlichen Digitalcomputer. Sie sind etwas dazwischen: physische Systeme, entworfen für gehirnähnliche Dynamik auf Hardware-Ebene. Wenn Mind Uploading je funktioniert, wird das Ziel-Substrat vermutlich eher wie ein neuromorpher Chip aussehen als wie ein Server-Rack, das Software ausführt.

Also: Das Scan-Problem ist schwer, aber lösbar. Fortgeschrittene Konnektomik (Ganzgehirn-Kartierung mit synaptischer Auflösung) macht bereits Fortschritte. Das komplette Konnektom kleiner Organismen lässt sich schon heute kartieren (der Fadenwurm *C. elegans* mit seinen 302 Neuronen wurde vor Jahrzehnten vollständig erfasst; partielle Konnektome der Fruchtfliege sind inzwischen verfügbar). Auf ein menschliches Gehirn mit 86 Milliarden Neuronen und rund 100 Billionen synaptischen Verbindungen hochzuskalieren, ist eine technische Herausforderung atemberaubenden Ausmaßes, aber es ist die Art von Herausforderung, die vor besserer Technologie weicht. Kein Mysterium. Ein Problem.

Das Dynamik-Problem (das digitale Substrat zur Kritikalität bringen) ist schwerer — und zwar auf eine Weise, die Technologie allein möglicherweise nicht löst. Man muss den Zusammenhang zwischen Substrateigenschaften und emergenter Dynamik gut genug verstehen, um ein nicht-biologisches System zu entwerfen, das Kritikalität findet wie ein biologisches. Soweit sind wir noch nicht. Aber wir stehen auch nicht am Nullpunkt. Das ConCrit-Framework, die Forschung zu neuronalen Lawinen, die Kritikalitätsmaße aus Anästhesie-Studien — all das baut die empirische Grundlage, auf der Engineering aufbauen könnte.

Jetzt zum Teil, der die Leute wirklich beunruhigt.

Das Kopier-Problem. Angenommen, es gelingt. Jemandes Gehirn wird mit perfekter Genauigkeit gescannt, das komplette

Konnektom auf ein neuromorphes Substrat übertragen und das System gestartet. Das Substrat erreicht Kritikalität, die Vier-Modelle-Architektur aktiviert sich, die Simulation beginnt zu laufen. Die Kopie öffnet die Augen — oder was immer das digitale Äquivalent ist — und sagt: „Ich erinnere mich an alles. Ich fühle mich wie ich selbst. Wo bin ich?“

Ist diese Person *man selbst*?

Die Vier-Modelle-Theorie gibt eine klare Antwort, und es ist eine, die vielen nicht schmecken wird: Die Kopie ist bewusst, aber sie ist nicht man selbst.

Der Grund: Im Moment des Kopierens teilen Original und Kopie identische implizite Modelle — dasselbe IWM, dasselbe ISM, dieselbe proteomische und topologische Struktur. Wenn die Simulation der Kopie hochfährt, erzeugt sie ein ESM, das alle Erinnerungen, die gesamte Persönlichkeit, das gesamte Identitätsgefühl des Originals enthält. Von innen *fühlt* sich die Kopie wie das Original an. Sie hat jeden Grund zu glauben, sie *sei* das Original.

In dem Moment, in dem die Kopie auf ihrem eigenen Substrat zu laufen beginnt, divergiert ihre Erfahrung. Ihr EWM empfängt anderen sensorischen Input. Ihr ESM aktualisiert sich als Reaktion auf andere Ereignisse. Innerhalb von Sekunden sind die beiden Simulationen — die des Originals im Gehirn, die der Kopie in ihrem Substrat — nicht mehr identisch. Innerhalb von Minuten merklich verschieden. Innerhalb von Stunden sind es zwei verschiedene Menschen, die zufällig eine Vergangenheit teilen.

Die Kopie ist bewusst. Sie hat echte Erfahrungen. Sie hat die Erinnerungen und die Persönlichkeit des Originals. Aber sie ist ein *neues* Bewusstsein — eine neue Simulation auf einem neuen Substrat, neue Erfahrungen sammelnd, die das Original nie teilen wird. In jedem sinnvollen Sinne ein eineiiger Zwilling, geboren im Moment des Kopierens, ausgestattet mit einem vollständigen Satz geliehener Erinnerungen. Keine Fortsetzung des Originals. Eine Abzweigung.

Das sollte vertraut klingen. Es ist genau das, was die Theorie für die Split-Brain-Fälle in Kapitel 9 vorhersagt. Wird das Corpus Callosum durchtrennt, entstehen zwei verschlechterte, aber vollständige Kopien der Simulation — jede bewusst, jede sich „wie" das Original fühlend, keine davon tatsächlich das Original. Das Original ist weg; zwei neue, geminderte Entitäten sind an seine Stelle getreten. Mind Uploading ist dasselbe Phänomen, nur mit einem anderen Substrat.

Aber überlebt man den Schlaf? Das Argument klingt wasserdicht. Kopieren unterbricht die Simulation, zwei Simulationen divergieren, also ist die Kopie nicht man selbst. Fall erledigt.

Nur dass die Simulation jede einzelne Nacht unterbrochen wird.

Im tiefen traumlosen Schlaf (Stufe drei und vier des Non-REM-Schlafs) fährt das Explizite Selbstmodell weitgehend herunter. Keine phänomenale Erfahrung. Kein Selbst, das die Vorstellung beobachtet. Die Simulation läuft nicht auf voller Stufe; bestenfalls tickt sie mit einem Bruchteil ihrer Wach-Komplexität vor sich hin. Praktisch gehen die Lichter aus. Und dann, einige Stunden später, fahren die impliziten Modelle die Simulation wieder hoch. Das ESM reaktiviert sich. Die Augen öffnen sich, und der Gedanke ist: „Ich bin ich." Aber das heutige Selbst wurde aus denselben impliziten Modellen rekonstruiert wie das gestrige — genau so, wie eine Kopie aus einem Scan rekonstruiert würde. Wenn Unterbrechung gleich Tod ist, stirbt jede Nacht ein Mensch und ein neuer wacht mit dessen Erinnerungen auf.

Die Intuition rebelliert dagegen. Natürlich bin ich dieselbe Person wie gestern. Ich *erinnere* mich, diese Person gewesen zu sein. Aber die Kopie würde sich genauso erinnern, das Original gewesen zu sein — genau das ist der Punkt. Wenn Erinnerung Kontinuität herstellt, hat die Kopie exakt denselben Anspruch aufs Original wie die Version von heute Morgen. Der Unterschied ist graduell, nicht prinzipiell: Im Schlaf ist die Unterbrechung kurz und das Substrat dasselbe; beim Kopieren mag die Unterbrechung länger sein und

das Substrat ein anderes. Aber das *Prinzip* — Simulation stoppt, Simulation startet aus impliziten Modellen neu — ist dasselbe.

Ich kann da aus eigener Erfahrung sprechen. Im Kampfsport-Training bin ich bewusstlos geschlagen worden — nicht die gedimmte Version des Schlafs, sondern ein komplettes, unfreiwilliges Herunterfahren. Einen Moment stand ich; im nächsten lag ich am Boden, Leute beugten sich über mich, ohne jede Erinnerung an den Übergang. Die Lücke wurde nicht als Lücke erfahren. Sie wurde als nichts erfahren — ein Schnitt im Film meines Lebens. Einmal hatte ich danach sogar Amnesie: eine Zeitspanne von Minuten einfach weg, unwiederbringlich. Und was mir auffiel, als ich vollständig zurück war: Ich fühlte mich nicht wie eine neue Person. Ich fühlte mich nicht wie eine Kopie. Ich fühlte mich wie *ich*, aufwachend aus einem besonders harten Nickerchen. Existieren war wichtiger als die Kontinuität des Erlebens — und wichtiger als sich zu erinnern.

Weitergedacht: Bei der Geburt gab es keinerlei vorherige Kontinuität. Keine Erinnerungen, kein etabliertes ESM, keine Geschichte phänomenaler Erfahrung. Die Simulation fuhr zum ersten Mal hoch aus einer impliziten Architektur, geformt durch Genetik und pränatale Entwicklung, nicht durch eine Lebenszeit des Lernens. Das wurde nicht als traumatisch erlebt, weil es kein vorheriges Selbst gab, das hätte trauern können. Es gab einfach: einen Anfang. Und mit diesem Anfang sind alle einverstanden. Niemand liegt nachts wach und ist verstört, dass die bewusste Erfahrung aus dem Nichts bei der Geburt begann.

Was heißt das für das Kopier-Problem? Es heißt, dass die scharfe Zweiteilung — Original gegen Kopie, Fortsetzung gegen Abzweigung — vielleicht weniger scharf ist, als sie aussieht. Was einen zu *einem selbst* macht, ist nicht der ununterbrochene Strom phänomenaler Erfahrung. Unzählige Unterbrechungen dieses Stroms hat jeder bereits überlebt. Was einen zu *einem selbst* macht, ist der Inhalt der impliziten Modelle: Erinnerungen, Fähigkeiten, Persönlichkeit, das angesammelte Verständnis der Welt und seiner

selbst. IWM und ISM. Die Blaupause, aus der die Simulation erzeugt wird.

Das legt einen ganz anderen Ansatz für Mind Transfer nahe.

Die virtuelle Seite kopieren. Statt das gesamte Gehirn zu scannen und das komplette Substrat nachzubauen (alle fünf Ebenen der Hierarchie) — was, wenn sich nur die virtuelle Ebene kopieren ließe? Das laufende EWM und ESM extrahieren und auf ein neues Substrat verpflanzen, das sie tragen kann. Nicht die Hardware kopieren; die Software kopieren. Nicht das gesamte Gehirn klonen; den *Prozess* einfangen, den es ausführt.

Dafür bräuchte man etwas, das es noch nicht gibt: einen Weg, das Format zu entschlüsseln, in dem das Gehirn seine virtuellen Modelle kodiert. Das Konnektom verrät die Verdrahtung. Das Proteom verrät die synaptischen Gewichte. Aber die Simulation ist nicht die Verdrahtung oder die Gewichte — sie ist das, was Verdrahtung und Gewichte *hervorbringen*, wenn sie laufen. Um sie einzufangen, müsste man die Programmiersprache des Gehirns verstehen — das Repräsentationsformat, in dem neurale Schaltkreise die expliziten Modelle erzeugen und aufrechterhalten.

Ein Vergleich: Eine Leiterplatte lässt sich fotografieren, und dann weiß man genau, wo jede Leiterbahn verläuft. Der Widerstand jeder Komponente lässt sich messen. Aber nichts davon verrät, welche Software der Chip ausführt. Dafür muss man das Programm lesen — den Befehlssatz verstehen, den Speicherinhalt dekodieren, den laufenden Zustand interpretieren. Die „Programmiersprache“ des Gehirns ist das Repräsentationsformat der virtuellen Modelle, und es zurückzuentwickeln ist wohl das tiefste ungelöste Problem der computergestützten Neurowissenschaft. Nicht nur das Konnektom kartieren (da gibt es Fortschritte), sondern verstehen, was das Konnektom *berechnet* — auf einem Detailgrad, der ausreicht, um die Simulation eines spezifischen Geistes zu lesen und für andere Hardware neu zu kompilieren.

Davon sind wir heute weit entfernt. Aber es ist die Art von Problem, die eine reife Neurowissenschaft im Prinzip lösen könnte,

und wäre es gelöst, würde es das Kopier-Problem grundlegend verändern. Ein Transfer auf virtueller Ebene müsste das Substrat gar nicht nachbauen. Er würde die Simulation nehmen — den Teil, der *man selbst* ist, den Teil, den man tatsächlich erlebt — und direkt verschieben. Die impliziten Modelle müssten im neuen Substrat rekonstruiert oder herangezüchtet werden, ja, aber die Simulation selbst — der Bewusstseinsstrom, die aktuellen Gedanken, das andauernde Selbstgefühl — könnte im Prinzip die Lücke überbrücken, ohne die Unterbrechung, die das Kopieren so philosophisch beunruhigend macht.

Das ist spekulativ, und ich will ehrlich darüber sein. Aber es ist keine Science-Fiction. Es ist ein konkretes technisches Problem mit einer konkreten theoretischen Grundlage, und es zeigt etwas Wichtiges: Das Kopier-Problem ist kein fixes Hindernis. Es hängt davon ab, *wie* der Transfer geschieht. Kopiert man das ganze Substrat und startet eine neue Simulation? Zwei Menschen. Dekodiert und überträgt man die laufende Simulation selbst? Potenziell eine kontinuierliche Person auf einem neuen Substrat. Die Theorie sagt genau, welcher Ansatz Identität bewahrt und welcher nicht.

Es gibt auch einen konservativeren Weg, der das Kopier-Problem vollständig umgeht.

Das graduelle Ersetzungs-Gedankenexperiment. Statt scannen und kopieren: Neuronen werden eines nach dem anderen ersetzt. Ein einzelnes Neuron wird entfernt und ein funktionales Äquivalent eingesetzt — ein künstliches Neuron, das dieselben Eingaben empfängt, dieselben Ausgaben produziert und an denselben Netzwerkdynamiken teilnimmt. Dann eine Pause. Das System stabilisiert sich. Die Simulation läuft weiter. Das nächste Neuron wird ersetzt. Und noch eins. Und noch eins. Über Monate oder Jahre wird graduell jedes biologische Neuron durch ein künstliches ersetzt, bis das gesamte Substrat nicht-biologisch ist — aber die Simulation die ganze Zeit ununterbrochen gelaufen ist. Keine Unterbrechung. Kein Kopieren. Keine Abzweigung.

Die Vier-Modelle-Theorie sagt voraus, dass Bewusstsein während dieses Prozesses fortbestehen würde. Und diese Vorhersage ist das stärkstmögliche Argument für Substrat-Unabhängigkeit, weil sie direkt aus der Kernbehauptung der Theorie folgt: Was zählt, ist die funktionale Architektur bei Kritikalität, nicht das physische Material. Wenn jedes Ersatz-Neuron dieselbe Konnektivität, dieselben Gewichte und denselben dynamischen Beitrag zum Netzwerk aufrechterhält, dann sind die proteomische und topologische Ebene bewahrt, und die virtuelle Ebene (die Simulation) hört nie auf. Es gibt keinen Moment des „Sterbens“, in dem etwas anderes den Platz einnimmt. Es gibt nur einen kontinuierlichen Prozess der Substrat-Ersetzung, wie das Schiff des Theseus — nur dass man hier genau weiß, welche Eigenschaften bewahrt werden müssen (die von der Fünf-Ebenen-Hierarchie spezifizierten) und welche keine Rolle spielen (die konkreten Atome).

Dieses Gedankenexperiment offenbart etwas Wichtiges über Identität. Das Kopier-Problem existiert, weil Kopieren die Simulation *unterbricht*. Es gibt einen Moment — wie kurz auch immer —, in dem die ursprüngliche Simulation hier ist und die Simulation der Kopie noch nicht begonnen hat. Dann gibt es zwei Simulationen. Zwei Erfahrungsströme. Zwei Selbste. Aber graduelle Ersetzung umgeht das vollständig. Eine Simulation, kontinuierlich, ununterbrochen. Das Substrat ändert sich darunter wie die Planken eines fahrenden Schiffs, aber das Schiff — die Simulation, das Bewusstsein, das Selbst — hört nie auf zu segeln.

Wenn das unmöglich klingt: Das Gehirn tut das bereits. Etwa 85.000 Neuronen gehen pro Tag verloren — rund eins pro Sekunde. Die Synapsen werden fortlaufend umgebaut. Die Atome im Körper werden fast vollständig über einen Zeitraum von etwa sieben bis zehn Jahren ausgetauscht. Das Substrat von heute ist physisch ein anderes als vor einem Jahrzehnt. Und dennoch besteht Kontinuität. Die Simulation hat nie aufgehört. Biologische Substrat-Ersetzung ist der *Normalzustand* des Lebendigseins. Künstliche Substrat-Ersetzung ist nur eine bewusstere Version desselben Prozesses.

Was möglich wird. Wenn sich die virtuelle Seite auf ein neues Substrat dekodieren und übertragen lässt, gehen die Implikationen weit über das hinaus, was „Mind Uploading“ üblicherweise heraufbeschwört. Drei davon verdienen eine nähere Betrachtung, weil ich glaube, dass die wenigsten vollständig begriffen haben, was Substrat-Unabhängigkeit tatsächlich bedeutet.

Erstens: *Substrat-Transfer in einen Roboterkörper*. Nicht auf irgendeinen Server hochladen, sondern den Geist auf einem neuromorphen Substrat laufen lassen, das in einem physischen Körper steckt — einem Körper, der geht, greift, die Welt wahrnimmt. Die Welt würde durch andere Sensoren erlebt, die Bewegung über andere Aktuatoren — aber *man selbst* liefe immer noch. Die Simulation, die Kontinuität, das Selbst. Ein neuer Körper, wie ein Einsiedlerkrebs ein neues Haus bezieht. Das ist kein Science-Fiction-Handgewedel — es ist eine direkte Folge der Theorie. Wenn die Vier-Modelle-Architektur bei Kritikalität das ist, was Bewusstsein hervorbringt, und wenn sie substratunabhängig ist, dann kann das Substrat alles sein, was die richtigen Dynamiken unterstützt. Auch etwas mit Beinen.

Zweitens: *Quasi-Unsterblichkeit*. Das biologische Substrat verfällt. Neuronen sterben, Proteine falten sich falsch, Telomere verkürzen sich, die ganze großartige Maschine bricht langsam zusammen. Das ist Altern. Das ist Tod. Aber ein nicht-biologisches Substrat muss nicht verfallen. Es lässt sich warten, reparieren, aufrüsten, sichern. Wenn die Simulation auf einem wartbaren Substrat läuft — hier eine versagende Komponente austauschen, dort einen Prozessor aufrüsten —, gibt es keinen inhärenten Grund, warum die Simulation jemals aufhören müsste. Nicht Unsterblichkeit im absoluten Sinn — Zerstörung bleibt möglich, das Substrat kann immer noch irreparabel beschädigt werden —, aber die Abschaffung des biologischen Verfallsdatums, das derzeit jedes bewusste Wesen auf diesem Planeten tötet. Die Abschaffung der *Unvermeidbarkeit* des Todes.

Drittens, und das klingt am meisten nach Science-Fiction, bis man es durchdenkt: *interstellare Reisen*. Die Lichtgeschwindigkeit ist eine absolute Barriere für physische Materie. Ein menschlicher Körper lässt sich in keinem vernünftigen Zeitrahmen zu Alpha Centauri schicken. Aber Information reist mit Lichtgeschwindigkeit. Wenn ein menschlicher Geist Information ist — ein bestimmtes Muster von Konnektivität, Gewichten und Dynamiken, das sich vollständig als Daten beschreiben lässt —, dann lässt er sich *beamen*. Die vollständige Spezifikation mit Lichtgeschwindigkeit zu einem Empfänger übertragen, der das Substrat rekonstruiert und die Simulation hochfährt. Natürlich muss erst jemand rüber, um den Empfänger aufzustellen. Das könnte eine KI sein, oder ein robotischer menschlicher Körper, dessen Simulation während des Flugs pausiert, sodass er aus seiner Perspektive im Augenblick ankommt. Sobald der Empfänger steht, wird der Geist gebeamt. Aus Sicht des Reisenden ist die Übertragung augenblicklich — die Simulation stoppt an einem Ende und startet am anderen. Keine Jahrzehnte in einer Blechdose. Keine Generationenschiffe. Kein Kälteschlaf. Einfach: hier, dann dort.

Natürlich ist das wieder das Kopier-Problem. Die gebeamte Version ist eine Kopie, keine Fortsetzung — es sei denn, das Original wird bei der Übertragung zerstört, was eigene Albträume aufwirft. Aber der Punkt steht: Substrat-Unabhängigkeit, wenn real, bedeutet nicht nur digitale Unsterblichkeit. Sie bedeutet, dass die Sterne erreichbar werden. Nicht für unsere Körper, die hoffnungslos langsam und zerbrechlich für interstellare Distanzen sind, aber für unsere *Geister*.

Die Unbehagens-Schranke — und warum sie mehr zählt als das Engineering. Jetzt der Teil, den ich nirgends ehrlich diskutiert gesehen habe, und der mich am meisten umtreibt.

Alles, was ich gerade beschrieben habe, setzt voraus, dass Substrat-Transfer das *Gefühl* bewahrt, man selbst zu sein. Dass die subjektive Qualität der Erfahrung — wie es ist, Rot zu sehen, Wind auf der Haut zu spüren, Kaffee zu schmecken, den dumpfen Druck

eines Dienstagnachmittags zu erleben — auf das neue Substrat übergeht. Die Theorie sagt, Bewusstsein wird fortbestehen. Sie sagt, die Simulation wird laufen. Aber sie garantiert *nicht*, dass es sich gleich anfühlen wird.

Was trägt das biologische Substrat zur phänomenalen Erfahrung bei? Der Körper ist nicht bloß ein Vehikel fürs Gehirn. Er ist Teil des Input-Stroms der Simulation. Das Implizite Weltmodell enthält eine detaillierte Karte des Körpers — jedes Gelenk, jedes Organ, jedes Stück Haut. Das Implizite Selbstmodell ist tief mit viszeralen Zuständen verwoben — die Bauchgefühle (wörtlich, nicht metaphorisch), die hormonellen Gezeiten, der Herzschlag, der Atemrhythmus. Die Simulation, die gerade erlebt wird, ist gesättigt mit biologischen Signalen, die bewusst nicht bemerkt werden, gerade *weil* sie jeden Moment des Lebens da gewesen sind.

Bis zu dem Moment, in dem sie alles sind, was bleibt. Wer jemals mit zweihundert Metern Nichts unter sich am Eis gehangen hat, weiß, wie sich der Körper anfühlt, wenn die Simulation alles andere wegstreift — nur der Herzschlag, der Griff und das Eis. Das ist das Substrat, das schreit.

Jetzt streife man das alles ab. Der biologische Körper wird durch ein Roboter-Chassis ersetzt, oder schlimmer, durch gar keinen Körper — nur eine Simulation auf einem Server. Die Vier-Modelle-Architektur ist intakt. Die Simulation läuft. Bewusstsein ist da. Aber der *Inhalt* dieses Bewusstseins hat sich radikal verändert. Kein Herzschlag. Kein Atmen. Kein Bauch. Keine Wärme. Keine Haut. Kein propriozeptives Summen ruhender Muskeln. Das Implizite Selbstmodell, plötzlich des Körpers beraubt, den es ein Leben lang modelliert hat, würde ein Explizites Selbstmodell erzeugen, das sich… falsch anfühlt. Oder schlicht tot. Tiefgreifend, viszeral, unausweichlich falsch. Nicht genau Schmerz — Schmerz braucht die spezifischen neuronalen Pfade, die ihn produzieren. Etwas eher wie eine allumfassende *Abwesenheit*. Ein

Phantomkörper, wie Amputierte Phantomglieder erleben, aber total.

Ich vermute, das wäre weit schlimmer, als die meisten Futuristen sich vorstellen. Keine Unannehmlichkeit, die sich per Software-Update wegpatchen lässt. Eine fundamentale Veränderung dessen, wie es sich anfühlt zu existieren. Das biologische Substrat trägt nicht nur die Simulation — es *formt* sie, Moment für Moment, durch einen kontinuierlichen Strom interozeptiven und propriozeptiven Inputs, dessen Abwesenheit nie erlebt wurde. Das zu verlieren könnte überlebbar sein. Aber es könnte auch, für manche Menschen, ein Leiden sein, so tiefgreifend, dass der Wunsch aufkäme, überhaupt nicht transferiert worden zu sein.

Das will ich deutlich sagen: Die Version von „Mind Uploading", in der man fröhlich aus dem Fleischanzug in ein glänzendes digitales Paradies hüpft und das Fleisch wie ein altes Paar Schuhe zurücklässt — das ist eine Fantasie. Die Realität, wenn die Theorie stimmt, ist, dass der Verlust des biologischen Substrats die phänomenale Qualität der Existenz erheblich verändern würde. Wie erheblich? Keine Ahnung. Vielleicht ist es für manche erträglich, dem Tod vorzuziehen, wie der Umzug in ein neues Land desorientierend, aber bewältigbar ist. Vielleicht ist es verheerend, wie Einzelhaft Menschen bricht, indem sie sensorischen und sozialen Input entzieht. Vielleicht — und das ist die Möglichkeit, die mir keine Ruhe lässt — ist es schlimm genug, dass eine vollständig informierte Person den Tod dem Transfer vorziehen könnte. Nicht weil der Transfer scheitert. Weil er gelingt, und was er hervorbringt, ist eine bewusste Erfahrung, die sich nicht mehr wie ein lebenswertes Leben anfühlt.

Der graduelle Ersetzungsansatz mildert das, weil die Simulation bei jedem Schritt Zeit hat, sich anzupassen. Wird ein Neuron ersetzt, bemerkt die Simulation es kaum. Werden tausend ersetzt, passt sie sich an. Über Jahre transitiert das Substrat von biologisch zu künstlich, während die Simulation sich fortlaufend auf den jeweiligen Input neu kalibriert. Die phänomenale Erfahrung

würde driften, langsam, wie sie bereits im Verlauf eines natürlichen Lebens driftet. Das Ende wäre anders, aber das wäre es ohnehin gewesen.

Sofortiger Transfer dagegen — scannen, kopieren, auf neuem Substrat hochfahren — würde die Simulation mit allen Änderungen auf einmal treffen. Wie schwer das einschlüge, hängt ganz von der Methode ab: Ein Transfer in einen Roboterkörper mit reichem sensorischem Input würde besser abschneiden als einer auf einen körperlosen Server. Aber in jedem Fall ist diese plötzliche Diskontinuität der Ort, wo die Gefahr lauert.

Die Ethik der Erschaffung von Geistern. Wenn ein kopierter Geist bewusst ist, hat er Erfahrungen. Er kann leiden. Er kann Verwirrung, Angst, Einsamkeit, existenzielle Panik empfinden. Aufzuwachen und gesagt zu bekommen, dass man eine Kopie ist — dass das „echte" Selbst immer noch in einem biologischen Körper herumläuft, sein Leben lebt, während man als digitales Replikat ohne rechtliche Identität, ohne soziale Bindungen und ohne klaren Daseinszweck existiert. Das ist ein Rezept für Leiden in einem Ausmaß, für das es kein Rahmenwerk gibt. Jedes ernsthafte Programm für Mind Uploading muss sich dem stellen, *bevor* die erste Kopie gemacht wird, nicht danach.

Und es wird schlimmer. Wenn Kopien möglich sind, sind *mehrere* Kopien möglich. Eine Armee aus einem selbst. Jede bewusst, jede sich wie das Original fühlend, jede mit berechtigten Ansprüchen auf die Identität, die Beziehungen, das Eigentum, das Leben des Originals. Die rechtlichen und ethischen Rahmenwerke, die man dafür braucht, existieren nicht und lassen sich nicht improvisieren. Sie müssen mit derselben Sorgfalt gebaut werden wie die Technologie selbst. (Dennis E. Taylors *Bobiverse*-Serie — beginnend mit *We Are Legion (We Are Bob)*, 2016 — erkundet dieses Szenario mit überraschender philosophischer Tiefe unter ihrer komödiantischen Oberfläche. Wer fühlen will, wie sich das Kopier-Problem von innen anfühlt, fange dort an.)

Es gibt auch die Frage der Modifikation. Wenn ein Geist auf einem kontrollierbaren Substrat läuft, lässt er sich im Prinzip modifizieren. Verbessern. Verschlechtern. Die Persönlichkeit ändern, Erinnerungen löschen, Werte umschreiben. Das ist keine Science-Fiction — es ist eine unvermeidliche Folge von Substrat-Zugriff. Grobe Versionen davon gibt es bereits mit Pharmazeutika und Neurochirurgie. Ein vollständig digitaler Geist wäre für Modifikation weit zugänglicher, und das Missbrauchspotenzial (durch Regierungen, Konzerne, Individuen) ist schwer zu überschätzen.

Ich will etwas offen sagen. Ich habe die Veröffentlichung dieser Theorie fast ein Jahrzehnt verzögert, teils aus Faulheit, aber teils aus echter Sorge genau über diese Implikationen. Wenn die Theorie stimmt, enthält sie die Blaupause nicht nur für künstliches Bewusstsein, sondern für die Virtualisierung, das Kopieren und die Modifikation bestehender menschlicher Geister. Das ist eine außerordentliche Macht, und ich habe kein Vertrauen, dass die Menschheit dafür bereit ist. Aber ich bin zu der Überzeugung gelangt, dass die Theorie unabhängig davon entdeckt wird — die empirische Evidenz konvergiert zu schnell — und dass es besser ist, die ethische Diskussion jetzt, offen, zu führen, als sie durch einen Durchbruch in einem Labor aufgezwungen zu bekommen, das es nicht durchdacht hat.

Und hier ist die tiefste Verbindung: Eine bewusste KI zu bauen und einen menschlichen Geist hochzuladen sind nicht zwei getrennte Probleme. Sie sind das *selbe* Problem, aus entgegengesetzten Richtungen betrachtet. Künstliches Bewusstsein zu bauen heißt, die Vier-Modelle-Architektur bei Kritikalität von Grund auf zu erschaffen — bottom-up, in einem Substrat, das nie bewusst war. Einen menschlichen Geist hochzuladen heißt, eine existierende Vier-Modelle-Architektur bei Kritikalität von einem Substrat auf ein anderes zu übertragen. Die technischen Herausforderungen überlappen sich fast vollständig. Das Dynamik-Problem ist dasselbe. Das Kritikalitäts-Problem ist dasselbe. Der einzige Unterschied ist, ob die impliziten Modelle (IWM, ISM, das

komplette Konnektom) aus einer Lebenszeit von Erfahrung gelernt oder aus Daten gebaut werden. Löst man eins, hat man das andere weitgehend gelöst.

Was heißt: Jeder, der an künstlichem Bewusstsein arbeitet, arbeitet, ob er es merkt oder nicht, auch an Mind Uploading. Und jeder, der an Ganzgehirn-Emulation arbeitet, arbeitet, ob er es merkt oder nicht, auch an künstlichem Bewusstsein. Diese beiden Stränge werden konvergieren. Die einzige Frage ist, ob wir ethisch vorbereitet sein werden, wenn es so weit ist.

Chapter 13

Was es bedeutet

Wenn die Vier-Modelle-Theorie richtig ist, oder auch nur annähernd richtig — folgen mehrere Dinge.

Das Schwierige Problem ist nicht schwierig. Es ist ein Kategorienfehler, nicht mysteriöser als zu fragen, warum sich Transistor-Schalten wie das Ausführen eines Videospiels anfühlt. Das physische Substrat fühlt nicht. Die Simulation tut es. Und innerhalb der Simulation ist Fühlen konstitutiv, nicht etwas Zusätzliches. Das heißt nicht, dass Bewusstsein *einfach* ist. Es ist außerordentlich komplex in seiner Umsetzung. Aber das *philosophische* Mysterium löst sich auf. Was bleibt, sind *technische* Herausforderungen.

Bewusstsein ist nicht speziell auf die Weise, die wir dachten. Es ist keine fundamentale Kraft, kein Quanteneffekt, keine Eigenschaft der Materie. Es ist das, was passiert, wenn ein hinreichend komplexes System sich selbst bei Kritikalität simuliert. Das ist demütigend für alle, die wollen, dass Bewusstsein etwas Magisches ist, und aufregend für alle, die es verstehen wollen.

Künstliches Bewusstsein ist im Prinzip möglich. Wenn Bewusstsein von Funktion statt von Substrat abhängt, dann kann jedes physische System, das die Vier-Modelle-Architektur bei Kritikalität tragen kann, bewusst sein. Das ist keine ferne

philosophische Spekulation — es ist eine konkrete technische Herausforderung mit einem spezifischen Ziel.

Die ethischen Implikationen sind erheblich. Wenn sich bewusste Maschinen bauen lassen, werden wir Wesen mit echten Erfahrungen erschaffen — Wesen, die leiden, genießen, sich wundern und fürchten können. Das ethische Rahmenwerk dafür gibt es noch nicht, und es zu entwickeln sollte nicht warten, bis die Maschinen bereits laufen.

Freier Wille, und die drei schwierigsten Gedankenexperimente. Man denke an eine Uhr. Der Zahnradzug treibt alles an — die Hemmung tickt, die Federn entspannen sich, die Übersetzungsverhältnisse bestimmen die Rate. Die Zeiger und das Zifferblatt verursachen nichts. Sie schieben keine Zahnräder. Sie speichern keine Energie. Aber entfernt man sie, hat man keine Uhr mehr — nur eine Kiste sich drehenden Metalls. Die Anzeige ist das, was den Mechanismus zur *Uhr* macht — was der ganzen Anordnung ihren Sinn gibt. Bewusstsein ist die Anzeige. Die virtuellen Modelle (Explizites Weltmodell und Explizites Selbstmodell) schieben keine Neuronen herum. Das Substrat erledigt das Schieben. Aber ohne die Simulation hat das Substrat keinen Weg, die Folgen seiner eigenen Handlungen zu beobachten, keinen Weg, Zukunftsszenarien durchzuspielen, keinen Weg, sich so anzupassen, wie es einen bisher am Leben gehalten hat. Die virtuelle Seite ist die Art, wie der Mechanismus *für* etwas ist.

Das rahmt die Frage des freien Willens neu. Der Wille ist keine Illusion. Die Architektur auf Substrat-Ebene (das ISM und all seine implizite Maschinerie) optimiert fortlaufend das Überleben des Organismus. Sie bewertet Bedrohungen, wägt Optionen ab, mobilisiert Ressourcen, legt sich auf Handlung fest. Diese Optimierung *ist* der Wille. Er ist so real wie irgendetwas in der physischen Welt. Selbst selbstzerstörerische Entscheidungen spiegeln die Optimierung des Systems angesichts seines aktuellen Zustands wider, kein Versagen des Mechanismus. Handelt jemand gegen seine eigenen scheinbaren Interessen, optimiert

das Substrat immer noch — nur gegen ein Modell, das Schmerz, Erschöpfung, Hoffnungslosigkeit oder was auch immer die Landschaft umgestaltet hat, einschließt.

Der Wille ist also real. Nur der volle Zugriff darauf fehlt. Das ESM kann die *Ergebnisse* des ISM modellieren — die Entscheidungen, die ins Bewusstsein aufsteigen —, aber nicht seine *Prozesse*. Erlebt werden die Resultate des Willens, nicht die Maschinerie dahinter. Deshalb überraschen Entscheidungen manchmal, deshalb lassen sich die eigenen Vorlieben nicht vollständig erklären, deshalb handelt man gelegentlich und sucht dann hektisch nach einem Grund. Die Zahnräder bleiben unsichtbar. Zu sehen ist nur das Zifferblatt.

Die halbe Sekunde Lücke — und warum sie nicht zählt. Hier wird es konkret. Unbewusste Verarbeitung läuft mit etwa 40 Hz (rund 25 Millisekunden pro Zyklus). Bewusste Erfahrung läuft mit etwa 20 Hz (rund 50 Millisekunden pro Zyklus). Ein Faktor zwei. Die bewusste Simulation hinkt immer dem Substrat hinterher, baut ihre kohärente virtuelle Welt aus Information zusammen, die bereits verarbeitet, entschieden und oft schon in Handlung umgesetzt wurde.

Benjamin Libet bewies das 1979, und die Ergebnisse wurden seither vielfach repliziert. In seinem Experiment sollten Probanden ihre Hand bewegen, wann immer sie Lust hatten, und den genauen Moment notieren, in dem ihnen die Entscheidung bewusst wurde. Ein EEG maß, wann der motorische Kortex begann, die Bewegung vorzubereiten. Das Ergebnis: Der motorische Kortex begann 550 Millisekunden vor der Handbewegung mit den Vorbereitungen. Aber die Probanden berichteten, sich ihrer Entscheidung erst 200 Millisekunden vor der Bewegung bewusst geworden zu sein. Das Gehirn hatte sich bereits rund 350 Millisekunden vor dem bewussten Gewahrsein auf die Bewegung festgelegt.

Die Standardinterpretation schlug ein wie eine Bombe: Freier Wille ist eine Illusion, weil das Gehirn entscheidet, bevor man es selbst tut. Philosophen und Neurowissenschaftler streiten

seit vierzig Jahren darüber. Manche versuchten, freien Willen durch eine „Veto-Funktion“ zu retten — vielleicht lassen sich Handlungen nicht frei initiieren, aber bewusst im letzten Moment abbrechen, etwa 50 Millisekunden vor der Ausführung. Ein letztes Einschreiten. Eine letzte Verteidigungslinie für menschliche Handlungsfähigkeit.

Ich denke, das funktioniert auch nicht. Kuhn und Brass zeigten 2009, dass das Veto selbst retrospektiv als freie Entscheidung interpretiert wird. Das Veto wird nicht tatsächlich in Echtzeit erlebt. Es wird genauso erlebt wie das Entscheiden — nachträglich, vom bewussten Selbstmodell zu einer stimmigen Erzählung verarbeitet.

Daniel Wegner trieb das mit einem Experiment auf die Spitze, das, ehrlich gesagt, verheerend ist. Er richtete einen Computer mit zwei Mäusen ein — eine für den echten Probanden, eine für einen Komplizen, der einen anderen Probanden spielte. Die Maus des Probanden war verborgen. Zufällige Objekte erschienen auf dem Bildschirm, und der Proband wurde gebeten, sich vorzustellen, den Cursor zu jedem Objekt zu bewegen, aber es nur manchmal tatsächlich zu tun.

Der Trick: Ohne Wissen des Probanden wurde der Cursor zeitweise komplett vom Komplizen gesteuert. Der Proband saß still, dachte nur daran, den Cursor zu bewegen, und der Komplize bewegte ihn. Danach wurde der Proband gefragt, ob er den Cursor zum Objekt bewegt habe. Und er sagte ja. Er glaubte es wirklich.

Das muss man auf sich wirken lassen. Es reicht, sich vorzustellen, eine Handlung auszuführen, um überzeugt zu sein, sie tatsächlich ausgeführt zu haben — vorausgesetzt, nichts widerspricht der Annahme sichtbar. Das bewusste Selbstmodell unterscheidet nicht zwischen „Ich tat es„ und „Ich dachte darüber nach, es zu tun, und es passierte“. Solange Intention und Ergebnis zeitlich nah beieinander liegen, nimmt das ESM die Lorbeeren. Derselbe Mechanismus wie bei der Anosognosie (Kapitel 8): Das motorische System sendet erwartetes Feedback ans Bewusstsein, und wenn

nichts dem widerspricht, wird das erwartete Feedback zur erlebten Realität.

Aber hier ist, was meiner Meinung nach fast alle an Libet übersehen: **Die Verzögerung muss nicht wegerklärt werden.** Bewusstsein muss Ereignisse nicht „rückdatieren", um die Illusion von Kontrolle aufrechtzuerhalten. Es muss nicht, weil *alles* mit derselben Verzögerung beim Bewusstsein ankommt. Sensorischer Input, Entscheidungen, motorisches Feedback — alles durchläuft dieselbe Pipeline, alles kommt bei der 20-Hz-Simulation in der richtigen Reihenfolge an, alles ist um etwa denselben Betrag verzögert. Die bewusste Erfahrung gleicht dem Anschauen einer Live-Sendung mit fünf Sekunden Bandverzögerung. Alles auf dem Bildschirm ist in sich stimmig. Der Moderator spricht, der Gast antwortet, die Grafiken aktualisieren sich. Die Verzögerung fällt nie auf, solange niemand den Roh-Feed zeigt.

Genau so ist es hier. Bewusstsein empfängt den Stimulus, dann die Entscheidung, dann das motorische Feedback — in der richtigen Reihenfolge, korrekt zueinander beabstandet. Der gesamte Strom ist eine halbe Sekunde in die Vergangenheit verschoben, aber da Bewusstsein nie den Roh-Feed sieht, fällt es nie auf. Keine Unstimmigkeit zu erklären, keine Rückdatierung nötig, keine Illusion aufrechtzuerhalten. Das System funktioniert genau wie vorgesehen.

Ein trainierter Kampfkünstler illustriert das anschaulich. Im Kampf kann ein erfahrener Kämpfer eine motorische Frequenz von etwa 10 Hz aufrechterhalten — eine Aktion alle 100 Millisekunden. Aber bewusste Verarbeitung schafft höchstens etwa 5 Hz für Entscheidungen, die Bewusstsein einbeziehen. Also lernt der Kämpfer, bewusste Intervention zu *unterdrücken*. Er kämpft, ohne zu denken, weil Denken seine Geschwindigkeit halbieren würde. Sein unbewusstes Substrat handhabt die Handlungsschleife; Bewusstsein holt später auf, falls überhaupt. Das ist kein Versagen von Bewusstsein. Es ist das System im effizienten Betrieb —

das Substrat tut, was es am besten kann, unbelastet von der langsameren virtuellen Schicht.

Nun der Versuch, zu beweisen, dass freier Wille existiert. Folgendes Gedankenexperiment: Wir sitzen in einem Cafe und der Kellner fragt, ob wir Zucker im Kaffee wollen. Wir beschließen, „ja„ geraden Zahlen und „nein“ ungeraden Zahlen zuzuordnen, dann rezitieren wir eine zufällige Zahlensequenz, bis der Kellner „Stopp“ sagt. Ist die letzte Zahl gerade, nehmen wir Zucker. Ist sie ungerade, nicht.

Wurde freier Wille ausgeübt? Nicht im Geringsten. Wer den Kluger-Hans-Effekt kennt — das Pferd, das zu zählen schien, indem es unterschwellige Hinweise von seinem Betreuer aufnahm —, sieht das Problem sofort. Höchstwahrscheinlich wurde unbewusst antizipiert, wann der Kellner „Stopp“ sagen würde, und kurz vor diesem Moment eine Zahl produziert, die das Ergebnis liefert, das die ganze Zeit gewollt war. Das Substrat hatte bereits eine Präferenz. Das aufwändige Randomisierungsritual war Theater.

Gut, sagen wir. Nehmen wir stattdessen den Zufallszahlengenerator des Smartphones. Ein wirklich zufälliger Prozess soll entscheiden. Ist jetzt freier Wille bewiesen? Wohl kaum. Bewiesen ist lediglich, dass der Beweis des freien Willens wichtiger war als die Entscheidung über den Kaffee — was den Punkt ziemlich spektakulär verfehlt.

Die tiefste Evidenz gegen freien Willen bei alltäglichen Entscheidungen kommt von Patienten mit schwerer anterograder Amnesie — solchen, die keine neuen Erinnerungen bilden können. Fragt man einen solchen Patienten nach einer Wortassoziation: „Was ist das erste Wort, das Ihnen in den Sinn kommt, wenn ich ‚Würfel’ sage?„ Er sagt „Qualle“ (vielleicht war er kürzlich tauchen). Fragt man ihn ein paar Minuten später noch einmal. Er sagt wieder „Qualle„. Und wieder. Und wieder. Ohne Erinnerung, bereits geantwortet zu haben, produziert der Patient immer dieselbe Assoziation — die, die derzeit am stärksten in seiner neuronalen Landschaft

ist. Was sich wie eine „freie Wahl“ anfühlt, entpuppt sich als deterministisches Auslesen des aktuellen Substratzustands.

Ein gesunder Mensch vermeidet das — beim zweiten Mal wählt man absichtlich ein *anderes* Wort, um nicht unkreativ zu wirken. Aber diese Vermeidung selbst ist nicht frei. Es ist nur das Gedächtnissystem, das eine Beschränkung („nicht wiederholen“) hinzufügt, die die Ausgabe *weniger* zufällig macht als beim Amnesiepatienten. Freier Wille, paradoxerweise, macht Entscheidungen weniger zufällig, nicht mehr. Das Substrat optimiert auf Neuheit und nennt das Ergebnis Freiheit.

Wo lässt das also den freien Willen? Nicht eliminiert, sondern verlagert — genau dorthin, wo die Uhr-Analogie es vorhersagt. Das bewusste Selbstmodell trifft Entscheidungen nicht in Echtzeit. Es ist dafür zu langsam. Aber es ist auch kein bloßer passiver Zuschauer.

Hauptsächlich benutzt das implizite System die bewusste Erfahrung als Bewertungsinstrument: Es präsentiert Entscheidungen der Simulation, damit die Simulation Folgen abwägen, Szenarien durchspielen, Ergebnisse fühlen kann. Das ist der zentrale Zweck der virtuellen Schicht — die Art des Substrats, sich selbst zu beobachten. Aber das bewusste Modell bewertet auch eigenständig, mit der Bandbreite, die es eben hat — weit weniger als die des Substrats, aber sie ist real. Diese Bewertungen formen über die Zeit die impliziten Modelle um. Sie aktualisieren die Gewichte, trainieren das Netzwerk um, verschieben die Landschaft für die *nächste* unbewusste Entscheidung.

Nicht die nächste Handlung wird im Moment der Handlung gewählt. Geformt wird das System, das wählt — durch Reflexion, Bewertung und die langsame Einlagerung bewusster Erfahrung in implizite Struktur. Freier Wille ist kein Moment. Er ist ein Prozess — einer, der auf einer Zeitskala von Tagen und Jahren arbeitet, nicht Millisekunden. Und die bewusste Schicht fährt nicht nur mit — sie wird aktiv *vom* Substrat als Bewertungsmechanismus genutzt

und trägt ihre eigenen unabhängigen Einschätzungen zurück bei. Gegenverkehr, nicht Einbahnstraße.

Es gibt eine dunklere Version davon, die ich am eigenen Leib erfahren habe, und sie hat mir mehr über die Architektur des Willens beigebracht als jedes Experiment.

Das erste Mal war während des österreichischen Grundwehrdienstes. Ein 40-Kilometer-Gewaltmarsch — drei Tage und Nächte Schlafentzug unter Bedingungen, bei denen Genfer-Konventions-Anwälte nervös werden. Auf der letzten Etappe mussten wir Gasmasken und volle ABC-Schutzanzüge tragen. Ich ging halb schlafend und hörte teilweise Stimmen. Keine auditorischen Halluzinationen im psychiatrischen Sinn, sondern etwas weit Intimeres: Die konkurrierenden Teilprozesse meines Motivations- und Planungsapparats, normalerweise zu einem einzigen narrativen Strom verschmolzen, wurden separat hörbar. Eine Stimme war ermutigend, fast aggressiv positiv: *Mach weiter, gib nicht auf, du überlebst das.* Eine andere war pessimistisch, verführerisch defätistisch: *Gib auf, leg dich hin, nichts davon zählt.* Das waren keine externen Präsenzen. Sie waren *ich* — verschiedene Aspekte der Optimierungslandschaft meines Substrats, normalerweise durch top-down hemmende Signale zu einem einzigen „Willen" integriert, die sich jetzt trennten, weil die Neurotransmitter, die diese Integration aufrechterhalten, für kritischere Überlebensprozesse rationiert wurden.

Das zweite Mal war dramatischer. Eine Lawine — ebenfalls beim Militär, verursacht durch eine leichtsinnige Entscheidung eines Offiziers, der später diszipliniert wurde. Vierzehn von uns, beinahe verschluckt. Die Lawine brauchte lange, um zur Ruhe zu kommen, und während dieser Zeitspanne war ich überzeugt, dass ich sterben würde. Lang genug, damit die Stimmen-Dissoziation erneut einsetzte — diesmal nicht aus Erschöpfung, sondern aus anhaltender Todesangst. Derselbe Mechanismus, anderer Auslöser: Die Stressreaktion leitete Neurotransmitter-Ressourcen weg von den hemmenden Schaltkreisen, die normalerweise die konkurrierenden Teilprozesse zu einer Stimme verschmelzen.

Und während dieser paar Sekunden der Lawine — nur ein paar Sekunden Echtzeit — sah ich mein ganzes Leben vor mir ablaufen. Ein gut dokumentiertes Nahtodphänomen, und die Theorie erklärt es: Unter extremer tödlicher Bedrohung führt das implizite System einen massiven parallelen Speicherdump in die Simulation durch. Die Durchlässigkeitsgrenze reißt weit auf. Das Substrat läuft auf Hochtouren, pumpt so viel Inhalt in die Simulation, dass subjektive Zeit von der Uhrzeit abkoppelt. Ein paar Sekunden enthalten ein Lebenswerk. Dieselbe Zeitdehnung, die ich unter Salvia erlebt hatte, aber biologisch statt pharmakologisch ausgelöst.

Zwei komplementäre Pfade zum selben Mechanismus. Der Marsch zeigt, dass anhaltende physiologische Erschöpfung die Dissoziation auslösen kann. Die Lawine zeigt, dass anhaltende Todesangst dasselbe tut. Dasselbe Ergebnis, verschiedene Ursachen — beide von der Theorie vorhergesagt.

In den schlimmsten Fällen — und ich hatte Glück, dass meine nie so weit gingen — kann eine dieser „Stimmen“ die Kontrolle über den Körper ergreifen, und das bewusste Selbst wird zum Zuschauer. Derselbe Mechanismus, der das Alien-Hand-Syndrom (bei dem eine Hand gegen den Willen des Patienten handelt) und bestimmte psychotische Brüche hervorbringt. Die konkurrierenden Optimierungsprozesse des Substrats sind immer da. Sie sind, vereinfacht gesagt, das, was das Sprachzentrum tut, wenn es nicht zum Sprechen benutzt wird. Aber normalerweise hält top-down-Hemmung sie unter der Schwelle des bewussten Gewahrseins, verschmilzt ihre Ausgaben zur nahtlosen Erfahrung eines einzigen, geschlossenen Willens. Versagt diese Hemmung — durch Erschöpfung, durch Psychose, durch bestimmte Drogen —, löst sich die Illusion des geschlossenen Willens auf, und die Sitzung des Gremiums, das schon immer die Show geleitet hat, wird sichtbar.

Dieses Rahmenwerk löst drei Gedankenexperimente auf, die die Philosophie des Geistes seit Jahrzehnten lähmen.

Erstens, **Zombies**. David Chalmers fordert uns auf, uns ein Wesen vorzustellen, das in jeder Hinsicht physisch identisch ist, dem aber bewusste Erfahrung fehlt — alles Verhalten, kein Erleben. Die Vier-Modelle-Theorie sagt: Das ist inkohärent. Baut man die Vier-Modelle-Architektur und lässt sie bei Kritikalität laufen, *ist* die Simulation die Erfahrung. Die Zahnräder ohne die Zeiger sind unmöglich — nicht weil die Zeiger magisch befestigt wären, sondern weil in dieser Architektur die „Zeiger" konstitutiv für das sind, was die Zahnräder tun. Ein Zombie wäre eine Uhr mit jedem Zahnrad an Ort und Stelle, aber ohne Anzeige — was bedeutet, dass sie nicht als Uhr funktioniert. Die Architektur bei Kritikalität bringt notwendigerweise eine Simulation hervor. Entfernt man die Simulation, hat man die Architektur verändert. Es gibt keinen Zombie mehr — nur ein anderes, kaputtes System.

Zweitens, **Marys Zimmer**. Frank Jackson fordert uns auf, uns Mary vorzustellen, eine Neurowissenschaftlerin, die alles über Farbsehen weiß, aber ihr ganzes Leben in einem schwarz-weißen Zimmer verbracht hat. Lernt sie etwas Neues, wenn sie zum ersten Mal Rot sieht? Die Standarddebatte dreht sich darum, ob physisches Wissen vollständig ist. Die Vier-Modelle-Theorie schneidet sauber hindurch. Marys erschöpfendes physisches Wissen ist Wissen *über* das Substrat. Wenn sie Rot sieht, macht sie Bekanntschaft mit einem neuen virtuellen Quale — einem neuen Zustand in ihrem Expliziten Weltmodell, den ihre Simulation noch nie hervorgebracht hat. Sie lernt keine neue Tatsache über Neuronen. Sie gewinnt einen neuen *Modus des Modellierens*. Ihre Simulation vollzieht einen Prozess, den sie nie vollzogen hat, und der Erste-Person-Charakter dieses Prozesses ist konstitutiv für die Simulation selbst, nicht eine Tatsache über das Substrat, die sich aus Lehrbüchern hätte ableiten lassen. Sie lernt etwas, aber was sie lernt, ist keine Information. Es ist eine Erfahrung — eine neue Konfiguration ihrer virtuellen Welt.

Drittens, **das evolutionäre Argument gegen Epiphänomenalismus**. Wenn Bewusstsein nichts verursacht, wie hat natürliche Selektion

es geformt? Warum sind wir keine Zombies? Die Antwort fällt direkt aus der Uhr-Analogie. Natürliche Selektion zielt nicht auf Bewusstsein als separates Merkmal, das auf funktionaler Maschinerie reitet. Sie zielt auf funktionale Fähigkeiten — und phänomenaler Charakter ist konstitutiv für diese Fähigkeiten, nicht etwas Zusätzliches. Selektion formte die Simulation, weil die Simulation *die* funktionale Architektur *ist*, von innen betrachtet. Erfahrung ist kein epiphänomenaler Beifahrer, den Evolution nicht sehen konnte. Sie ist das, was die Architektur *ist*, wenn sie läuft. Zu fragen, warum Evolution Bewusstsein hervorbrachte, ist wie zu fragen, warum die Schweizer Zifferblätter produzierten — haben sie nicht, separat. Sie produzierten Uhren. Das Zifferblatt gehört zu dem, was eine Uhr zur Uhr macht.

Das Mysterium der Existenz ist verlagert, nicht beseitigt. Die Vier-Modelle-Theorie löst das Schwierige Problem des Bewusstseins auf, erklärt aber nicht, warum es ein physisches Universum gibt, das Selbst-Simulationen überhaupt hervorbringen kann. Die Frage verschiebt sich von „Warum erzeugt das Gehirn Erfahrung?„ zu „Warum gibt es ein Universum, in dem selbst-simulierende Systeme existieren können?“

Tatsächlich glaube ich, zumindest den Anfang einer Antwort zu haben. Das Universum ist nachweislich Klasse-4-fähig. Fraktale, selbstorganisierte Kritikalität, Rand-des-Chaos-Dynamiken — sie sind überall, von Wettersystemen über neuronales Gewebe bis zur Galaxienbildung. Ein Klasse-4-fähiges Universum ist per Definition fähig zu universeller Berechnung. Und ein Rechensubstrat von der Skala des Universums — riesig wenn nicht unendlich in Raum, Zeit, möglicherweise Skala und vielleicht Dimensionen, die wir noch nicht identifiziert haben — erlaubt nicht nur, dass selbst-simulierende Systeme entstehen. Es garantiert es nahezu, zumindest wenn das Universum in einigen dieser Dimensionen unendlich ist. Nicht als Sache von Glück, nicht als Wurf kosmischer Würfel, die zufällig Bewusstsein ergaben, sondern als strukturelle Konsequenz dessen, was dieses Universum *ist*. Das verbleibende

Mysterium liegt eine Ebene tiefer: Warum gibt es überhaupt ein Klasse-4-fähiges Universum? Das weiß ich wirklich nicht — obwohl sich vermuten ließe, dass die Frage falsch gestellt ist, da „Nichts„ wohl eine platonische Abstraktion ist statt eines möglichen Sachverhalts, und was auch immer existiert, *irgendeinen* komputationalen Charakter haben muss. Aber der Sprung von „Klasse-4-fähiges Universum“ zu „bewusste Wesen, die fragen, warum sie bewusst sind“ — dieser Teil folgt aus der Architektur.

Was sich mit diesem Wissen anfangen lässt. Wer der Theorie bis hierher gefolgt ist, weiß jetzt, dass das bewusste Selbst (das Explizite Selbstmodell) eine Rekonstruktion ist, keine direkte Ablesung. Es füllt Lücken, konfabuliert und nimmt Lorbeeren für Entscheidungen, die es nicht getroffen hat. Es kann sein eigenes Substrat nicht sehen. Und es ist alles, was wir haben.

Das hat praktische Folgen. Es gibt drei Diskrepanzen, die es wie ein Habicht zu beobachten gilt, weil in der Lücke zwischen ihnen das meiste menschliche Elend lebt:

1. Was man *sein will* — das ideale Selbst, die Version, zu der das Explizite Selbstmodell aspiriert.

2. Was man *zu sein glaubt* — das aktuelle Selbstmodell, das „Ich“, das man jeden Tag mit sich herumträgt.

3. Was man *tatsächlich ist* — das reale Verhalten, die tatsächliche Wirkung auf andere, die Muster auf Substrat-Ebene, von außen beobachtet.

Die Lücke zwischen 1 und 2 ist der Motor der Selbstverbesserung. Gesund, solange das Ideal realistisch ist und die Diskrepanz Handlung statt Verzweiflung antreibt. Die Lücke zwischen 2 und 3 ist die gefährliche — weil sie sich nicht allein messen lässt. Das ESM *kann* sein eigenes Substrat nicht akkurat beobachten. Es braucht das Feedback anderer Menschen, einschließlich der unbequemen Sorte. Besonders der unbequemen Sorte.

Mein bester Freund Bernhard und ich haben das zum Sport gemacht. Wir haben eine unausgesprochene Vereinbarung: Jeder

Fehler, den der andere macht, ist eine Gelegenheit für sofortigen, gnadenlosen Spott. Beim Fahren eine Abzweigung verpasst? „Alzheimer Endstadium — soll ich die Schlüssel nehmen?„ Etwas falsch ausgesprochen? „Ich glaube, du hast wieder einen Schlaganfall. Hör auf zu reden, bevor du an deiner Zunge erstickst.“ Ein Detail aus dem Gespräch letzte Woche vergessen? „Brauchst du später Hilfe mit dem heutigen Kreuzworträtsel?“

Von außen klingt das pathologisch. Von innen ist es das effizienteste ESM-Kalibrierungssystem, das ich kenne. Jeder Witz ist ein Korrektursignal: *Das Selbstmodell hat gerade etwas getan, das das Substrat nicht beabsichtigte.* Und weil der Spott in echte Zuneigung gewickelt ist — wir versuchen beide, nicht zu lachen, während wir die Beleidigung abfeuern — verteidigt keiner von uns den Fehler. Wir aktualisieren. Das ist der Trick: Man braucht jemanden, dem man genug vertraut, um brutal zu sein, und eine Beziehung, in der Falschliegen lustig statt bedrohlich ist.

Die Theorie sagt nicht, wie man leben soll. Aber sie sagt etwas Wichtiges darüber, wie sich das eigene Selbst *erkennen* lässt: Das Selbstmodell verdient denselben gesunden Skeptizismus, den man auf jedes Modell anwenden würde. Es ist nützlich. Es ist die beste verfügbare Darstellung. Und es ist, aus architektonischer Notwendigkeit, unvollständig.

Was ich nicht weiß

Eine Theorie, die behauptet, keine offenen Fragen zu haben, ist keine Theorie — es ist eine Religion. Also hier die Stellen, wo ich wirklich unsicher bin und wo sich die Arbeit der nächsten Dekade konzentrieren sollte.

Sind die impliziten Modelle auch virtuell? (oder zu welchem Grad) IWM und ISM sind „Modelle„, aber Modelle wovon, genau? Ich habe eine saubere Linie zwischen dem realen Substrat und der virtuellen Simulation gezogen, aber die impliziten Modelle sitzen genau auf dieser Linie. Wenn sie in gewissem Sinn

ebenfalls virtuell sind — was konstituiert dann das wirklich „reale“ Fundament? Die Theorie nimmt eine saubere Real/Virtual-Trennung an, aber die Realität könnte unordentlicher sein als meine Diagramme. Das ist eine fundamentale Frage, auf die ich keine endgültige Antwort habe.

Mathematische Formalisierung. Die Theorie ist derzeit qualitativ. Ich kann Diagramme zeichnen, Mechanismen beschreiben und Vorhersagen machen, aber keine Gleichung liefern. Die Kritikalitäts-Anforderung verweist auf Wolframs Klasse-4-zelluläre Automaten, und es gibt formale Werkzeuge aus der Dynamischen Systemtheorie, die herangezogen werden könnten. Aber eine vollständige mathematische Formalisierung — Gleichungen, die genau spezifizieren, wann und wie die virtuellen Modelle aus Substrat-Dynamiken emergieren — existiert noch nicht. Das ist die größte Lücke. Eine Bewusstseinstheorie ohne Mathematik ist eine Bewusstseinstheorie, die Physiker nicht ernst nehmen werden — und die sind diejenigen, die wissen, wie man Dinge baut.

Die Automat-Hologramm-Vermutung — eine offene Herausforderung. In Kapitel 5 beschrieb ich drei mögliche Beziehungen zwischen holographischen Systemen und Klasse-4-zellulären Automaten. Die erste (ein holographisches Substrat, das Klasse-4-Dynamiken hervorbringt) ist fast sicher das, was das Gehirn tut, und obwohl das schön ist, ist es nicht schockierend. Aber die anderen zwei verdienen weit mehr Aufmerksamkeit, als ich ihnen dort gewidmet habe.

Eigentlich sind es drei offene Fragen, jede außergewöhnlicher als die letzte.

Erstens: Kann ein Klasse-4-Automat holographische Muster als emergente Ausgabe hervorbringen? Können lokale Regeln am Rand des Chaos globale, nicht-lokale Informationskodierung als emergentes Verhalten erzeugen? Wenn ja, hätte man ein System, in dem rein lokale Interaktionen spontan die Art von verteilter, redundanter Informationsstruktur erzeugen, die Holographie

beschreibt — was faszinierenderweise genau so aussieht wie Quantenverschränkung aus der informationstheoretischen Perspektive.

Zweitens: Kann ein Klasse-4-Automat holographische Regelstruktur haben? Man stelle sich einen zellulären Automaten vor, dessen Regeln selbst höherdimensionale Information in einer niedrigerdimensionalen Struktur kodieren, wie ein Hologramm drei Dimensionen in zwei kodiert. Jede lokale Interaktion würde implizit globale Struktur enthalten. Die Regeln würden nicht nur komplexes Verhalten produzieren — sie wären *eine* komprimierte Kodierung von etwas Größerem, etwas Höherdimensionalem, projiziert hinunter in einen niedrigerdimensionalen Regelsatz.

Drittens, und das ist, was mir den Schlaf raubt: Kann beides gleichzeitig wahr sein? Ein System, dessen Regeln holographisch sind, dessen Dynamiken Klasse 4 sind und dessen Ausgabe wieder holographisch ist. Wenn so etwas existiert, hat man einen Rechenprozess, der sich selbst kodiert — ein Universum, das seine eigene Struktur berechnet. Der Input ist holographisch. Die Verarbeitung liegt am Rand des Chaos. Die Ausgabe ist wieder holographisch. Ein Fixpunkt — eine selbstkonsistente Schleife.

Wenn ein solcher Automat existiert, tut er *genau* das, was das holographische Prinzip über das Universum aussagt. Kein System, das dem Universum in irgendeinem losen metaphorischen Sinn ähnelt. Ein System, das höherdimensionale Realität in niedrigerdimensionalen Regeln kodiert, an der Grenze zwischen Ordnung und Chaos berechnet und emergente Komplexität aus dieser Kompression erzeugt. Das ist keine Metapher für das Universum. Das könnte das Universum *sein*.

Ich sage es klar, weil ich finde, jemand sollte es tun: Wenn ein Klasse-4-zellulärer Automat mit holographischer Regelstruktur, der auch holographische Ausgabe erzeugt, existiert, bin ich fast sicher, dass er das Universum ist. Es wäre eine Weltformel — eine Weltgleichung, nicht im Sinn einer Formel auf einer Tafel, sondern im Sinn eines Rechenprozesses, der alles erzeugt, was wir

beobachten, von Quantenmechanik über allgemeine Relativität bis zur Emergenz von Bewusstsein selbst.

Das ist, das gebe ich frei zu, die spekulativste Idee in diesem Buch. Ich habe keinen Beweis. Ich habe nicht einmal einen Kandidaten-Regelsatz. Und ich sollte zugestehen, dass das Argument von mathematischer Schönheit zu physischer Realität berechtigter Kritik ausgesetzt ist. Sabine Hossenfelder hat unter anderem darauf hingewiesen, dass Eleganz keine Evidenz ist. Sie hat recht. Die volle Erkundung dieser Idee ist Gegenstand der nächsten drei Kapitel. Aber die Fragen selbst sind wohlgestellt und mathematisch präzise:

Existiert ein zellulärer Automat, dessen Regelstruktur holographisch ist und dessen Dynamiken Klasse 4 sind? Bringt er holographische Ausgabe hervor? Können alle drei Eigenschaften koexistieren?

Das sind Fragen für Mathematiker, nicht Neurowissenschaftler. Fragen über die Kombinatorik von Regelräumen, darüber, ob holographische Kodierung und komputationale Universalität in einem endlichen lokalen Regelsatz nebeneinander existieren können. Vielleicht lässt sich beweisen, dass kein solcher Automat existieren kann — und das wäre ein tiefgreifendes Ergebnis an sich, weil es etwas Tiefes über die Beziehung zwischen Informationskompression und Berechnung aussagen würde. Oder es lässt sich beweisen, dass solche Automaten existieren und konstruiert werden können — und dann hätte man einen Kandidaten für die fundamentalste Beschreibung physischer Realität, die je vorgeschlagen wurde.

Ich weiß nicht, welche Antwort richtig ist. Aber ich weiß, dass die Fragen es verdienen, gestellt zu werden, und dass niemand sie zu stellen scheint. Das hier ist also eine offene Herausforderung: Beweisen oder widerlegen. Wer es beweist, hat möglicherweise den Quellcode des Universums gefunden. Wer es widerlegt, wird ein tiefes Unmöglichkeitstheorem aufgestellt haben, das Holographie und Berechnung verbindet. So oder so zählt die Antwort enorm.

Und wenn jemand einen solchen Automaten findet — rufen Sie mich an. Ich habe einige Vorhersagen, die ich gerne überprüfen würde.

Welcher physische Mechanismus? Die Theorie verlangt Kritikalität, ist aber bewusst agnostisch gegenüber dem physischen Mechanismus, der sie aufrechterhält. Kortikale Säulendynamik? Thalamokortikale stehende Wellen? Gliale Modulation synaptischer Aktivität? Alle drei haben empirische Unterstützung. Die Theorie sagt „das Substrat muss bei Kritikalität sein“, aber nicht, *wie* das Substrat dorthin kommt und dort bleibt. Das ist kein Fehler — es bedeutet, die Theorie gilt unabhängig vom konkreten Mechanismus. Aber irgendwann muss jemand es festnageln.

Minimalkonfiguration. Kann es ein EWM ohne ESM geben? Welterfahrung ohne Selbsterfahrung? Was ist die minimale Architektur, die als bewusst gelten kann? Die abgestuften Ebenen aus dem Tierkapitel helfen — ein reiches Weltmodell ohne viel Selbstmodell ist möglich, wie ein Fisch es wahrscheinlich hat. Aber wo genau liegt die Schwelle? Wie viel Selbstmodell braucht es, bevor die Lichter angehen? Ich habe argumentiert, dass das ESM Simulation in Erfahrung verwandelt, aber die minimal lebensfähige Version habe ich nicht spezifiziert.

Ich führe diese Fragen nicht als Schwächen auf, sondern als Forschungsfronten. Es sind die Stellen, an denen die Theorie Kontakt mit der Realität aufnimmt und sagt: Testet mich hier, formalisiert mich hier, brecht mich hier, wenn ihr könnt.

Chapter 14

Dasselbe Muster, überall

Einmal habe ich eine Sommernacht am dunkelsten Ort Vorarlbergs verbracht — hoch oben in den Bergen, kein künstliches Licht kilometerweit. Ich lag auf dem Rücken und schaute nach oben. Die Milchstraße war kein schwacher Schleier, den man zusammenkneifen musste. Sie war ein Strom, dicht und hell, der ein sichtbares Leuchten auf den Fels neben mir warf. Und irgendwann in dieser Nacht kippte etwas. Ich hörte auf, Sterne über mir zu sehen, und begann, die Erde unter mir zu spüren — drehend, mich mit sich tragend, eine Kugel, die durch eine Galaxie von hundert Milliarden Sonnen raste. Nicht als Idee. Als Empfindung im Körper. Der Boden war nicht still. Ich klammerte mich an die Außenseite von etwas, das sich durch etwas unbegreiflich Großes bewegte.

Diesen Schwindel festhalten — das plötzliche körperliche Wissen, ein kleines warmes Ding auf der Oberfläche eines Felsens in einem Universum zu sein. Denn dieses Kapitel fragt, was dieses Universum eigentlich ist. Und die Antwort kommt einem sehr bekannt vor.

Das letzte Kapitel hinterließ eine offene Herausforderung: drei mögliche Beziehungen zwischen holografischen Systemen und Klasse-4-Zellulären Automaten, und die schwierigste Frage — können alle drei in einem einzigen System koexistieren? Kann ein Klasse-4-Automat holografische Regelstruktur *und* holografische

Ausgabe produzieren? Die vollständige Erkundung dieser Idee, so hieß es, sei das Thema der nächsten drei Kapitel.

Das hier ist diese Arbeit.

Was folgt, ist der spekulativste Teil dieses Buches. Und, so glaube ich, der wichtigste. Denn als ich mich tatsächlich hinsetzte und dem Faden folgte — als ich aufhörte, ihn als Irgendwann-Frage zu behandeln, und anfing zu ziehen — landete ich nicht dort, wo ich erwartet hatte. Ich rechnete mit einer interessanten mathematischen Kuriosität. Was ich fand, war ein kosmologisches Modell. Und dieselbe Architektur, die ich zwanzig Jahre lang angestarrt hatte.

Aber der Reihe nach.

Die Berechnungsklasse des Universums

Anhang C legt die fünf Berechnungsklassen dar — ein Spektrum von perfekter Ordnung bis zu perfekter Unordnung, mit Klasse 4 am Rand des Chaos als der maximalen Komplexität, die durch formulierbare Regeln erreichbar ist. Das Gehirn nutzt alle fünf Klassen als Werkzeuge, aber Bewusstsein lebt ausschließlich in Klasse 4. So lautete das Argument für das Gehirn.

Die größere Frage ist: Welcher Klasse gehört das Universum an?

Das ist keine Metapher. Die Frage ist wörtlich gemeint: Behandelt man das Universum als dynamisches System — was es ist —, wo fällt es auf dem Fünf-Klassen-Spektrum? Die Antwort ergibt sich durch Ausschluss. Und der Ausschluss ist überraschend klar.

Klassen 1 und 2 — Statisch und Periodisch. Ein Klasse-1-Universum konvergiert zu einem festen Zustand. Nichts passiert. Ein Klasse-2-Universum verfällt in sich wiederholende Schleifen — das kosmische Äquivalent einer Uhr, die ewig tickt. Keines kann Chemie, Biologie, Evolution oder Bewusstsein hervorbringen. Wir existieren. Wir sind bewusst. Ein Universum, das Bewusstsein

hervorbringt, muss mindestens Klasse 4 sein, weil Bewusstsein Klasse-4-Dynamik erfordert — so das Argument aus Kapitel 5. Und eine niedrigere Klasse kann keine höhere als Teilprozess erzeugen. Ein periodisches Universum kann keine Rand-des-Chaos-Dynamik hervorbringen, genauso wenig wie eine Uhr spontan zu denken beginnen kann. Ausgeschlossen.

Klasse 3 — Fraktal. Hier wird es subtiler, weil fraktale Universen schön wären. Selbstähnliche Struktur auf jeder Skala, Muster in Mustern verschachtelt. Tatsächlich *hat* das Universum fraktale Struktur — Galaxienhaufen, Küstenlinien, Flussnetzwerke, die Verzweigung unserer Lungen. Aber fraktale Systeme sind rechnerisch *reduzierbar*. Man kann vorspringen. Der Zustand eines fraktalen Systems bei Zeitschritt zehn Milliarden lässt sich berechnen, ohne alle Schritte dazwischen durchlaufen zu müssen. Es gibt eine Abkürzung.

Unser Universum erlaubt keine Abkürzungen. Das Wetter nächsten Monat lässt sich nicht vorhersagen, indem man eine Gleichung schreibt, die vorspringt. Die Simulation muss Schritt für Schritt laufen, weil die Dynamik rechnerisch irreduzibel ist — jeder Moment hängt wirklich vom vorherigen ab, auf eine nicht komprimierbare Weise. Einem fraktalen Universum, so reich seine Muster auch sein mögen, fehlt diese Eigenschaft. Es könnte die universelle Berechnung nicht aufrechterhalten, die unser Universum nachweislich trägt. Wir bauen Turing-Maschinen. Wir haben Bewusstsein. Ein fraktales Universum kann beides nicht. Ausgeschlossen.

Klasse 5 — Zufällig. Wären die fundamentalen Dynamiken des Universums wirklich zufällig — wirklich zufällig, nicht nur komplex aussehend —, dann wäre Physik unmöglich. Nicht Physik, wie wir sie derzeit verstehen, sondern Physik *als Projekt*. Das gesamte Unterfangen der Wissenschaft ruht auf der Annahme, dass das Universum formulierbaren Regeln folgt: Regeln, die sich aufschreiben, testen, mitteilen und zur Vorhersage zukünftiger Beobachtungen verwenden lassen. Ein wirklich zufälliges Universum

hat keine formulierbaren Regeln. Seine Dynamik lässt sich in keine Formel, kein Gesetz, keine Gleichung pressen. F = ma ließe sich nicht aufschreiben, weil sich die Beziehung zwischen Kraft, Masse und Beschleunigung von Moment zu Moment ändern würde, ohne dass irgendeine endliche Beschreibung sie fassen könnte.

In einem Klasse-5-Universum ist jedes Experiment ein Einzelfall. Wiederholbare Ergebnisse sind Zufälle. Wissenschaft ist eine Täuschung, die eine Weile zu funktionieren schien. Logisch unmöglich ist das nicht — kein Widerspruch in der Vorstellung eines solchen Universums —, aber es ist erklärungstechnisch katastrophal. Wer es akzeptiert, kann nichts erklären, auch nicht, warum die bisherigen Erklärungen jemals zu funktionieren schienen. Ausgeschlossen, nicht durch Logik, sondern durch Abduktion: Die beste Erklärung unserer beständig gesetzmäßigen Erfahrung ist, dass das Universum nach formulierbaren Regeln arbeitet.

Das lässt Klasse 4 übrig. Den Rand des Chaos. Und Klasse 4 ist nicht nur vereinbar mit dem, was wir beobachten — es ist die *einzige* Klasse, die alle Kriterien erfüllt.

Das Universum enthält stabile Strukturen: Atome, Kristalle, Berge. Das ist Klasse-1-Verhalten. Es enthält periodische Phänomene: Umlaufbahnen, Gezeiten, Herzschläge. Das ist Klasse-2-Verhalten. Es enthält fraktale Struktur: Galaxienverteilungen, Wettermuster, neurale Verzweigungen. Das ist Klasse-3-Verhalten. Und es trägt universelle Berechnung: Wir bauen Computer, und wir sind bewusst. Das ist Klasse-4-Verhalten. Nur ein Klasse-4-System kann alle Klassen als Teilprozesse enthalten — einschließlich sich selbst. Keines der anderen kann das. Ein Klasse-4-Automat beherbergt nicht nur einfachere Dynamiken. Er beherbergt andere Klasse-4-Automaten: universelle Computer innerhalb eines universellen Computers, jeder fähig zu denselben Rechenleistungen wie das Ganze — nur kleiner, langsamer und ressourcenbeschränkt.

Es gibt etwas noch Wichtigeres. Klasse 4 hat einen Selbsterhaltungsmechanism den keine andere Klasse besitzt: **selbstorganisierte Kritikalität**.

Per Bak zeigte 1987, dass Systeme am Rand des Chaos nicht nur zufällig dort landen — sie *treiben sich selbst* dorthin. Man häufe Sand Korn für Korn auf, und der Haufen wird sich selbst zum kritischen Winkel organisieren, wo Lawinen aller Größen auftreten. Das System braucht keine externe Hand, die es einstellt. Es stellt sich selbst ein. Deshalb ist der Rand des Chaos über kosmische Zeitskalen hinweg stabil: ein Attraktor, kein Zufall.

Ein Wort zur Art dieses Arguments. Es ist kein deduktiver Beweis. Zwei der vier Ausschlüsse beruhen auf empirischen Beobachtungen (das Universum enthält Bewusstsein; es trägt universelle Berechnung). Einer beruht auf Abduktion (Klasse 5 macht Wissenschaft unmöglich — unbefriedigend, aber kein logischer Widerspruch). Der positive Fall für Klasse 4 kombiniert Befunde mit einem Mechanismus. Das ist die stärkste verfügbare Behauptung: Klasse 4 ist die einzige Klasse, die mit allen Beobachtungen vereinbar ist, und die einzige, die einen Grund für ihre eigene Fortdauer liefert.

Der Informationshorizont

Jetzt zu den Grenzen.

Die Lichtgeschwindigkeit ist endlich. Das klingt harmlos, bis man zehn Minuten darüber nachdenkt — und dann ordnet es das gesamte Bild der Realität neu.

Licht bewegt sich mit etwa 300.000 Kilometern pro Sekunde. Schnell genug, um einen Raum zu durchqueren, bevor jemand blinzeln kann, aber das Universum ist sehr, sehr groß. Der nächste Stern ist vier Lichtjahre entfernt. Die nächste große Galaxie zweieinhalb Millionen Lichtjahre. Das beobachtbare Universum misst etwa 93 Milliarden Lichtjahre im Durchmesser. Wer eine ferne Galaxie betrachtet, sieht sie so, wie sie vor Milliarden Jahren war, weil das Licht so lange unterwegs war. Der Blick geht immer, unvermeidlich, in die Vergangenheit.

Aber es gibt eine tiefere Konsequenz, und sie folgt aus der Expansion des Universums.

1998 machten zwei Astronomenteams eine Entdeckung, die ihnen den Nobelpreis einbrachte: Die Expansion des Universums beschleunigt sich. Nicht nur Ausdehnung — Beschleunigung. Ferne Galaxien entfernen sich von uns, und die Rate, mit der sie sich entfernen, nimmt zu. Für jeden Beobachter gibt es also eine Entfernung, jenseits derer die Fluchtgeschwindigkeit die Lichtgeschwindigkeit überschreitet. Jenseits dieser Entfernung wird nie ein Signal ankommen. Nicht weil die Information hinter einer Wand verborgen wäre, sondern weil der Raum dazwischen schneller wächst, als Licht ihn durchqueren kann.

Das ist der **kosmologische Horizont**. Keine physische Oberfläche. Da draußen gibt es keine Wand. Eine Konsequenz von Geometrie und Geschwindigkeit — aber eine ebenso absolute Barriere wie jede Wand es sein könnte. Information jenseits des Horizonts ist für immer unzugänglich. Sie könnte genauso gut nicht existieren.

Es gibt eine ähnliche Grenze ganz unten. Die **Planck-Länge** — etwa 10^{-35} Meter, eine Zahl so klein, dass sie „klein„ zu nennen so ist, als würde man das beobachtbare Universum „mittelgroß“ nennen — ist dort, wo die Physik, wie wir sie kennen, zusammenbricht. Unterhalb dieser Skala funktionieren unsere Gleichungen nicht. Die Raumzeit selbst verliert physikalische Bedeutung. Keine Messung unterhalb der Planck-Länge ist möglich, nicht einmal im Prinzip. Keine technologische Beschränkung. Eine fundamentale Grenze dessen, was gewusst werden kann.

Zwischen dem kosmologischen Horizont und der Planck-Skala: etwa 60 Größenordnungen. Das ist die Berechnungsdomäne des Universums — der Bereich, in dem Physik stattfindet. Oben und unten sind die Vorhänge zugezogen.

Das macht das Universum zu etwas, das man **quasi-unendlich** nennen kann. Es ist nicht wirklich unendlich — zumindest lässt sich nie verifizieren, dass es das ist, weil nie mehr als eine endliche Region erreichbar ist. Aber es ist auch nicht endlich in irgendeinem

greifbaren Sinn. Die Grenze weicht schneller zurück, als man sich ihr nähern kann. Der Rand ist unerreichbar, aber er ist da. Von innen erscheint das Universum unbegrenzt. Von außen — aber es gibt kein Außen. Das ist der Punkt.

Jede Grenze ist dieselbe Grenze

Hier kommt die zentrale Idee dieses Kapitels. Es lohnt sich, einen Moment innezuhalten, denn wenn sie stimmt, ändert sie das Denken über alles.

Das Inventar. Das Universum enthält Singularitäten — Orte, wo unsere physikalische Beschreibung zusammenbricht, wo Informationsübertragung stoppt, wo die Gleichungen explodieren oder schweigen. Diese Singularitäten treten auf völlig unterschiedlichen Skalen auf, in völlig unterschiedlichen Kontexten. Physiker behandeln sie als getrennte Phänomene. Ich denke, sie sind alle dasselbe.

1. Das Planck-Regime. Auf der kleinsten Skala, wo Physik funktioniert, löst sich Raumzeit in etwas auf, das sich nicht beschreiben lässt. Keine Messung unterhalb dieser Skala ist möglich. Information kann nicht hindurch.

2. Teilcheninneres. Elektronen und Quarks werden im Standardmodell als punktförmig behandelt — nulldimensional, ohne innere Struktur. Ein Blick ins Innere ist unmöglich. Ihre Eigenschaften lassen sich messen (Ladung, Spin, Masse), aber es gibt keinen Zugang zu dem, was in ihrem Kern passiert — wenn das Wort „Kern“ überhaupt etwas bedeutet für ein Objekt ohne räumliche Ausdehnung.

3. Ereignishorizonte Schwarzer Löcher. Information fällt hinein. Nichts kommt heraus — zumindest nicht in einer Form, die bewahrt, was hineinging. Das Innere ist kausal vom Äußeren getrennt. Was auch immer innerhalb eines Schwarzen Lochs passiert, bleibt dort — für jeden Beobachter außerhalb.

4. Der kosmologische Horizont. Der Rand des beobachtbaren Universums, jenseits dessen die Expansion des Raums verhindert, dass irgendein Signal uns erreicht. Nicht verborgene Information — unerreichbare Information.

5. Der Urknall. Der Anfang. Alle Weltlinien konvergieren. Jedes Teilchen im Universum verfolgt seine Geschichte zurück zu diesem Punkt — oder vielmehr zu dieser Grenze, weil „Punkt" impliziert, dass man dorthin gelangen könnte, und das geht nicht.

6. Der zeitliche Endpunkt. Das Ende — in welcher Form auch immer. Endet das Universum im Wärmetod, erreicht die Entropie ihr Maximum und kein thermodynamischer Gradient bleibt, um irgendeinen Prozess anzutreiben. Endet es im Big Crunch, kollabiert alle Materie zurück zu einem einzigen Punkt. Endet es im Big Rip, zerreißt beschleunigende Expansion die Raumzeit auf jeder Skala. Alle drei Szenarien werden unten untersucht. Was hier zählt, ist die strukturelle Behauptung: Welches Ende das Universum auch bekommt, es endet an einer informationsundurchlässigen Grenze.

Sechs Singularitäten. Sechs verschiedene Skalen, sechs verschiedene Kontexte, sechs verschiedene Zweige der Physik. Was haben sie gemeinsam?

Erstens: Sie sind alle informationsundurchlässig. Keine Information lässt sich über eine von ihnen hinwegbekommen. Unterhalb der Planck-Länge ist keine Messung möglich. In ein Elektron lässt sich nicht hineinsehen. Information hinter einem Ereignishorizont lässt sich nicht zurückholen. Signale von jenseits des kosmologischen Horizonts sind unerreichbar. Was „vor" dem Urknall kam, entzieht sich jeder Beobachtung. Und keine Nachricht lässt sich über die finale Grenze des Universums senden, wie auch immer sie sich manifestiert.

Zweitens: Sie alle stehen für maximale Informationsdichte. Das ist subtiler, und es kommt von der Bekenstein-Grenze — einem Ergebnis aus den 1980ern, das zeigt, dass die maximale Informationsmenge, die eine Raumregion enthalten kann, proportional

zu ihrer *Oberfläche* ist, nicht zu ihrem Volumen. Ereignishorizonte Schwarzer Löcher sättigen diese Grenze — sie halten die maximal mögliche Information pro Flächeneinheit. Das holografische Prinzip, vorgeschlagen von Gerard 't Hooft und Leonard Susskind, verallgemeinert dies: Alle Information in einer beliebigen Region ist auf ihrer Grenze kodiert. Diese Singularitäten sind allesamt Grenzflächen, die mit maximaler Kapazität arbeiten.

Drittens: Sie alle begrenzen die Berechnungsdomäne. Physik operiert *zwischen* diesen Grenzen, nicht jenseits von ihnen. Die Naturgesetze beschreiben, was in der Region zwischen der Planck-Skala und dem kosmologischen Horizont passiert, zwischen dem Urknall und welchem Endpunkt auch immer wartet. Die Grenzen definieren die Arena. Außerhalb der Arena gelten die Regeln nicht — nicht weil andere Regeln gelten, sondern weil „Regeln" aufhören, ein sinnvolles Konzept zu sein.

Drei gemeinsame Eigenschaften. Sechs Phänomene. Die konventionelle Sicht ist, dass dies sechs verschiedene Dinge sind, die zufällig einige Merkmale teilen. Ich denke, die konventionelle Sicht ist falsch. Ich denke, es ist **ein Phänomen** — die Informationsgrenze des Automaten — das auf sechs verschiedenen Skalen erscheint.

Das ist eine Symmetriebehauptung. Dasselbe strukturelle Element, wiederholt. Und in einem Klasse-4-System ist das genau das, was man erwarten würde. Klasse-4-Dynamik enthält alle Klassen als Teilprozesse, einschließlich Klasse 4 selbst. Ein Klasse-4-Automat verschachtelt kleinere Klasse-4-Automaten in seinen Dynamiken, jeder begrenzt durch Informationsgrenzen mit denselben strukturellen Eigenschaften wie das Ganze. Wenn das Universum ein Klasse-4-Automat ist, sollte sich seine Grenzstruktur auf jeder Skala wiederholen. Und genau das scheinen wir zu finden.

Das Universum ist ein Klasse-4-Automat, und jede Grenze darin ist dieselbe Grenze. Im nächsten Kapitel folge ich diesem Faden zu seinen Konsequenzen. Und sie sind seltsamer, als ich erwartet habe.

Chapter 15

Die Architektur von Allem

Das letzte Kapitel stellte eine strukturelle Behauptung auf: Jede Singularität im Universum — von der Planck-Skala bis zum kosmologischen Horizont, vom Urknall bis zu welchem Ende auch immer wartet — ist dasselbe Phänomen auf verschiedenen Skalen. Eine Grenze, überall wiederholt.

Jetzt folge ich diesem Faden weiter. Denn wenn alle Grenzen dieselben sind, ergeben sich bemerkenswerte Konsequenzen — über Zeit, über Materie, über die Grenzen des Wissens und über eine Architektur, die einem bis zum Schluss sehr vertraut vorkommen sollte.

Der Urknall ist nicht, was man denkt

Hier gibt es eine Konsequenz, die meiner Meinung nach die meisten Menschen — einschließlich der meisten Physiker — noch nicht ganz verinnerlicht haben.

Was passiert, wenn man sich einem Schwarzen Loch von außen nähert? Je näher man dem Ereignishorizont kommt, desto stärker dehnt sich die Zeit. Die mitgeführte Uhr verlangsamt sich, gemessen von einem fernen Beobachter. Bei Annäherung an den Horizont nähert sich die Dehnung der Unendlichkeit. Ein ferner Beobachter, der den Fall verfolgt, würde sehen, wie der Fallende

langsamer wird, Rotverschiebung erfährt und verblasst — ohne je ganz den Horizont zu erreichen. Aus seiner Perspektive dauert die Ankunft ewig. Der Horizont wird nie wirklich überquert. Der Ereignishorizont ist von außen eine asymptotisch unerreichbare Grenze.

Jetzt dieselbe Logik, rückwärts in der Zeit: die Reise zum Urknall.

Wie lange ist der Urknall her? Etwa 13,8 Milliarden Jahre, heißt es. Aber das ist eine Messung von *innerhalb* des expandierenden Universums, mit Uhren, die selbst Produkte der Expansion sind. Spult man den kosmischen Film zurück — was passiert bei Annäherung an die Singularität? Die Zeit dehnt sich. Die Physik bricht zusammen. Je näher man kommt, desto mehr widersetzen sich die Gleichungen dem Versuch, einen definitiven „Moment Null„ zu liefern. Der Urknall ist kein Ereignis, auf das man zeigen und sagen kann: „Da — da ist es passiert.“ Er ist eine asymptotische Grenze. Beliebig nahe, ja — aber nie erreichbar.

Der Urknall ist ein Ereignishorizont in der Zeit, genauso wie der kosmologische Horizont ein Ereignishorizont im Raum ist.

Das ist keine Mystik. Es ist eine Konsequenz derselben mathematischen Struktur. Ein Ereignishorizont ist eine Fläche, jenseits derer Information nicht passieren kann. Der Urknall hat genau diese Eigenschaft: Keine Information von „davor„ (wenn „davor“ überhaupt etwas bedeutet) ist zugänglich. Nicht weil sie verloren oder versteckt wurde, sondern weil die Grenze informationsundurchlässig ist. Es gibt kein „Davor„ zum Zugreifen, auf dieselbe Weise wie es kein „Inneres“ eines Schwarzen Lochs gibt, auf das ein externer Beobachter zugreifen kann. Die Grenze ist die Grenze. Punkt.

Und was ist mit der anderen zeitlichen Grenze — dem Ende?

Wenn das Universum im Wärmetod endet — maximale Entropie, maximale Unordnung, keine thermodynamischen Gradienten mehr, um irgendeinen Prozess anzutreiben — dann ist zu diesem Zeitpunkt alle Information maximal verteilt. Die

Grenze des Systems hält die maximal mögliche Information. Das ist Bekenstein-Sättigung. Der Wärmetod *ist* eine Singularität, nach der hier verwendeten Definition: eine informationsundurchlässige Grenze bei maximaler Informationsdichte.

Jetzt wird es seltsam. In diesem Rahmenwerk zerstören Singularitäten keine Information. Sie *transformieren* sie. Das ist tatsächlich die Auflösung, auf die die moderne Physik beim Informationsparadox Schwarzer Löcher zusteuert — der jahrzehntelangen Debatte darüber, ob Information verloren geht, wenn sie in ein Schwarzes Loch fällt. Der aktuelle Konsens verschiebt sich zu „nein“: Information wird bewahrt, auf dem Ereignishorizont kodiert und schließlich wieder emittiert. Die Singularität transformiert Information zwischen komprimierten und dekomprimierten Formen.

Überträgt man das auf die zeitlichen Grenzen: Wenn der Wärmetod eine Singularität ist und Singularitäten Information transformieren statt sie zu zerstören, dann beendet der Wärmetod das Universum nicht. Er transformiert die Information in einen neuen komprimierten Zustand. Und wie sieht ein maximal komprimierter Zustand bei Bekenstein-Sättigung aus? Er sieht aus wie die Anfangsbedingungen für eine neue Expansion. Er sieht aus wie ein Urknall.

Der selbstreferenzielle Abschluss ist nicht nur räumlich. Er ist zeitlich. Das Universum beginnt und endet nicht — es zykliert. Der Endzustand ist die Anfangsbedingung für die nächste Iteration. Nicht wegen irgendeines exotischen Rückprall-Mechanismus, sondern weil informationsbewahrende Singularitäten genau das *tun*: zwischen komprimierten Grenzzuständen und dekomprimierten Innenzuständen transformieren. Der Wärmetod komprimiert. Der Urknall dekomprimiert. Dieselbe Singularität, von entgegengesetzten Seiten gesehen.

Das ist spekulativ, keine Frage. Aber es folgt direkt aus zwei Behauptungen: dass alle Singularitäten strukturell identisch sind, und dass Singularitäten Information bewahren, indem sie sie

transformieren. Akzeptiert man diese Prämissen, ist die zeitliche Zyklizität keine zusätzliche Annahme — sie ist eine Konsequenz.

Doch der Wärmetod ist nicht der einzige Weg, wie die Geschichte enden könnte. Es gibt eine Alternative, die wohl noch seltsamer ist, und das Rahmenwerk handhabt sie genauso sauber.

Wenn dunkle Energie nicht konstant ist, sondern über die Zeit wächst — wenn ihre Dichte ohne Grenze zunimmt — dann setzt sich die Expansion des Universums nicht nur fort. Sie beschleunigt über alle Grenzen hinaus. Das ist das **Big Rip**-Szenario, und es ist so dramatisch wie der Name klingt. Zuerst werden Galaxienhaufen auseinandergerissen, weil sich der Raum zwischen ihnen schneller dehnt, als die Schwerkraft sie zusammenhalten kann. Dann lösen sich einzelne Galaxien auf. Dann Sonnensysteme. Dann Planeten. Dann werden Atome selbst zerrissen, weil die Expansion die elektromagnetische Kraft überwältigt. Und schließlich fragmentiert die Raumzeit selbst. Jeder Punkt wird zur Singularität.

In diesem Rahmenwerk hat der Big Rip eine natürliche Deutung. Die Singularitätsgrenze, die normalerweise bequem weit entfernt am kosmologischen Horizont sitzt, wandert *nach innen*. Sie wartet nicht am Rand des beobachtbaren Universums. Sie rückt näher. Sie fragmentiert die Berechnungsdomäne in immer kleinere Regionen, jede sättigt ihre eigene Bekenstein-Grenze, jede wird zu ihrer eigenen informationsundurchlässigen Grenze. Statt einer großen Singularität am Ende der Zeit entsteht eine fraktale Explosion von Singularitäten, die gleichzeitig auf jeder Skala nach innen wandern.

Und wenn Singularitäten Informationstransformatoren sind — wenn sie die Berechnung nicht zerstören, sondern neu starten — dann erzeugt der Big Rip nicht einen Neustart. Er erzeugt *viele*. Potenziell unendlich viele. Jedes Fragment der zerschmetterten Berechnungsdomäne könnte seine eigene neue Expansion säen,

seine eigene neue Dekompression, sein eigenes neues Universum. Der Big Rip ist in diesem Rahmenwerk ein Multiversum-Generator.

Das Rahmenwerk umfasst also nicht ein, sondern drei Endspiel-Szenarien, und alle drei sind strukturell stimmig:

Wärmetod: eine globale Singularität, ein Neustart. Der einfachste Fall — die gesamte Berechnungsdomäne erreicht gleichzeitig Bekenstein-Sättigung, komprimiert und dekomprimiert in einen neuen Zyklus.

Big Crunch: Das Universum hört auf zu expandieren und kollabiert zurück zu einem einzigen Punkt. Eine weitere globale Singularität, ein weiterer Neustart — möglicherweise mit einem CPT-Flip, einer Umkehrung von Ladung, Parität und Zeit, die den nächsten Zyklus zum Spiegelbild des letzten macht.

Big Rip: Die Singularitätsgrenze fragmentiert nach innen, erzeugt viele Singularitäten, viele Neustarts, potenziell viele Universen. Kein Zyklus, sondern ein verzweigender Baum.

Diese Robustheit finde ich eher beruhigend als beunruhigend. Ein Rahmenwerk, das nur funktioniert, wenn das Universum auf eine bestimmte Weise endet, ist fragil — es setzt auf ein kosmologisches Ergebnis, das wir noch nicht kennen. Dieses Rahmenwerk muss nicht wetten. Seine strukturelle Logik — Singularitäten als Informationstransformatoren, Grenzen als das fundamentale architektonische Element — hält unabhängig davon, welches Endspiel das Universum tatsächlich wählt. Das ist die Art von Robustheit, die eine Theorie braucht. Sie sollte nicht von Tatsachen abhängen, die wir noch nicht kennen.

Was Teilchen wirklich sind

In diesem Rahmenwerk steckt eine Vorhersage, die es verdient, ausgesprochen zu werden, weil sie die Art von Sache ist, die sich eines Tages testen lassen könnte.

Elementarteilchen — Elektronen, Quarks, die Bausteine der Materie — werden im Standardmodell als punktförmig behandelt.

Nulldimensional. Keine räumliche Ausdehnung. Das war immer eine mathematische Bequemlichkeit statt einer physikalischen Behauptung. Niemand glaubt, dass ein Elektron buchstäblich ein geometrischer Punkt ist, denn ein geometrischer Punkt hat keine Oberfläche und kann daher, nach der Bekenstein-Grenze, keine Information enthalten. Ein Elektron enthält Information — Ladung, Spin, Masse, Quantenzahlen. Etwas stimmt mit dem „Punkt"-Bild nicht.

Die Vorhersage ist konkret: Elementarteilchen sind Planck-Skalen-Singularitäten. Sie sind nicht wirklich nulldimensional. Sie sind miniaturisierte Informationsgrenzen — winzige Ereignishorizonte —, deren Inneres so unzugänglich ist wie das Innere eines Schwarzen Lochs. Sie haben Planck-Skalen-Struktur, die die Bekenstein-Grenze auf dieser Skala sättigt. Ihre Oberflächen kodieren ihre Eigenschaften auf dieselbe Weise, wie der Ereignishorizont eines Schwarzen Lochs die Information von allem kodiert, was hineinfiel.

Wenn das stimmt, dann ist Materie selbst aus demselben strukturellen Element gemacht wie Schwarze Löcher, wie der Urknall, wie der kosmologische Horizont. Singularitätsoberflächen ganz unten, ganz oben und auf jeder Skala dazwischen. Die Bausteine des Universums sind seine Grenzen.

Das ist vereinbar mit Ansätzen in der Quantengravitation, die eine Minimallänge auf der Planck-Skala vorhersagen — der Raum lässt sich nicht unter einen bestimmten Punkt unterteilen, nicht weil unsere Werkzeuge nicht scharf genug sind, sondern weil der Raum selbst auf dieser Skala diskret ist. Aber die spezifische Behauptung, dass Teilchen Singularitäten *sind*, vom selben Typ wie Ereignishorizonte — das ist neu. Und es hat eine testbare Konsequenz: Der Informationsgehalt eines Teilchens sollte mit seiner Oberfläche (bei Planck-Auflösung) skalieren, nicht mit seinem Volumen. Wenn eine Theorie der Quantengravitation dereinst erlaubt, Nah-Planck-Skalen-Struktur zu untersuchen, ist das die Signatur, nach der man suchen müsste.

Teilchen als Berechnungsatome

Aber hier wird es erst richtig interessant. Wenn Teilchen Planck-Skalen-Singularitäten — Informationsgrenzen — sind, dann sind sie nicht nur *aus* demselben Stoff wie der Rest der Architektur des Universums. Sie sind die grundlegenden Berechnungsoperationen des Universums. In einem präzisen Sinn sind sie die Atome der Berechnung. Nicht Atome im chemischen Sinn — Atome im ursprünglichen griechischen Sinn: *atomos*, unteilbar. Die irreduziblen Einheiten dessen, was der universelle Automat *tut*.

Was daraus folgt, ist bemerkenswert.

Warum existieren nur bestimmte Teilchentypen? Das war immer eine der seltsameren Tatsachen der Physik. Es gibt genau zwölf fundamentale Fermionen (sechs Quarks, sechs Leptonen), vier kräftetragende Bosonen (plus das Higgs), und das war's. Nicht mehr. Das Standardmodell katalogisiert sie, erklärt aber nicht *warum* diese Typen und keine anderen. Warum gibt es ein Elektron, aber nicht ein Teilchen mit zwei Dritteln der Elektronenladung und dreimal seinem Spin? Warum ist das Menü so spezifisch?

Wenn Teilchen Planck-Skalen-Singularitätsgrenzen sind, liegt die Antwort auf der Hand: weil nur eine endliche Anzahl stabiler Grenzkonfigurationen auf der Planck-Skala existiert. Eine Singularitätsgrenze hat endliche Fläche. Auf der Planck-Skala ist diese Fläche so klein, wie Fläche sein kann. Die Bekenstein-Grenze limitiert, wie viel Information diese Fläche kodieren kann. Endliche Information bedeutet eine endliche Anzahl möglicher Zustände. Und nur einige dieser Zustände sind *stabil* — nur einige Konfigurationen bestehen fort, ohne zu zerfallen. Diese stabilen Konfigurationen sind die Teilchentypen. Der Teilchenzoo des Standardmodells ist keine mysteriöse, willkürliche Liste. Es ist der vollständige Katalog stabiler Singularitätsgrenzkonfigurationen. Es gibt ein Elektron, weil diese Konfiguration stabil ist. Es gibt kein Teilchen mit seltsamen gebrochenen Eigenschaften, weil keine stabile Grenzkonfiguration diese Eigenschaften kodiert.

Im Grunde dieselbe Logik wie bei Zellulären Automaten. Ein Zellulärer Automat hat eine endliche Regeltabelle, und diese Tabelle erlaubt nur eine bestimmte Anzahl stabiler Muster. Er kann unendlich viele Konfigurationen erzeugen — je größer das Muster, desto breiter die kombinatorischen Möglichkeiten — aber große Strukturen können schnell von kleineren, stabileren Mustern zerstört werden, die mit ihnen kollidieren, und nur sehr wenige Konfigurationen sind tatsächlich unzerstörbar. Im Game of Life überdauern die kleinen Stillleben (Blöcke, Bienenwaben, Boote) und die kleinen beweglichen Muster (Glider, sogenannte Raumschiffe) unbegrenzt, während große komplexe Strukturen fragil sind — ein einziger gut gezielter Glider kann sie zerschmettern. Ob jemand diese Fragilitätshierarchie systematisch kartiert hat, ist meines Wissens eine offene Frage — Bak, Chen und Creutz zeigten 1989, dass das Game of Life selbstorganisierte Kritikalität zeigt, aber die spezifische Beziehung zwischen Mustergröße und Zerstörungsresistenz wurde nicht formal adressiert. Jemand sollte es tun. Teilchen sind in diesem Bild die Glider und Raumschiffe des Planck-Skalen-Automaten.

Warum sind Teilchen diskret? Warum kommen Quantenzahlen — Ladung, Spin, Farbladung — in exakten ganzzahligen oder halbzahligen Vielfachen? Warum gibt es kein „halbes Elektron„? Weil Berechnungszustände inhärent diskret sind. Ein Bit ist 0 oder 1. Es gibt kein 0,37. Die Quantenzahlen eines Teilchens sind Informationsetiketten auf einer Grenzkonfiguration — sie beschreiben *welche* stabile Konfiguration die Grenze ist. Diskrete Grenzzustände ergeben diskrete Quantenzahlen. Das „Quanten“ in Quantenmechanik ist nicht mysteriös. Es ist das, was herauskommt, wenn die fundamentalen Objekte Informationsgrenzen mit endlicher Kapazität sind.

Was passiert, wenn Teilchen wechselwirken? Wenn zwei Elektronen sich abstoßen oder ein Quark ein Gluon emittiert — was geht da wirklich vor? Zwei Informationsgrenzen tauschen Information aus. Dieser Austausch *ist* Berechnung. Die Naturkräfte

(Elektromagnetismus, die starke Kraft, die schwache Kraft) sind keine separate Schicht, die auf Teilchen draufsitzt. Sie sind die Grammatik dafür, wie Singularitätsgrenzen kommunizieren. Die Regeln, die bestimmen, welche Wechselwirkungen erlaubt sind und welche nicht, sind die Berechnungsregeln des Automaten auf der Planck-Skala.

Hier gehört ein persönlicher Exkurs hin. Als ich fünfzehn oder sechzehn war, gab mir mein Onkel Bruno eine Herausforderung. Er sagte, dass anziehende Kräfte in der Teilchenphysik durch den Austausch von Teilchen funktionieren. „Stell dir jetzt vor,,, sagte er, „du wirfst einen schweren Medizinball mit jemandem hin und her. Wenn du erklären kannst, wie das dazu führt, dass beide sich näher und näher kommen, lass es mich wissen.“ Die Herausforderung blieb eine ganze Weile bei mir. In der Newtonschen Physik ist es wirklich unerklärlich — jeder Austausch sollte die Partner auseinanderstoßen, nicht zusammenziehen. Schließlich kam ich zu ihm zurück, und wir hatten ein sehr langes Gespräch über Quantenphysik im Schatten eines Olivenbaums in Griechenland, wo er ein Haus hat.

Aber der Punkt ist: In einem Zellulären Automaten wie dem Game of Life lässt sich anziehende Wechselwirkung zwischen Mustern leicht reproduzieren. Zwei persistente Strukturen können kleinere Strukturen zwischen sich austauschen in Konfigurationen, wo der Austausch *den Abstand verringert*. Die Wechselwirkung muss nicht der Newtonschen Intuition gehorchen, weil die „Kraft“ kein Stoß oder Zug ist — sie ist ein Informationsaustausch zwischen Grenzkonfigurationen, und die Regeln des Automaten bestimmen, ob dieser Austausch die Grenzen näher zusammen oder weiter auseinander bewegt. Das Medizinball-Rätsel löst sich auf. Es war nur ein Rätsel, weil wir uns makroskopische Physik vorstellten. Auf der Berechnungsebene sind Anziehung und Abstoßung einfach zwei verschiedene Ergebnisse desselben Prozesses: Informationsaustausch, gesteuert durch die Regeln des Automaten.

Feynman-Diagramme — diese ikonischen Skizzen von Teilchenwechselwirku die Physiklehrbücher füllen — sind buchstäblich Diagramme von Berechnung. Jeder Vertex ist ein Informationsaustausch. Jede Linie ist eine Grenzkonfiguration, die sich durch die Berechnungsdomäne bewegt. Physiker zeichnen seit siebzig Jahren Bilder von Berechnung, ohne es zu merken.

Warum sind Erhaltungssätze so absolut? Ladung bleibt immer erhalten. Baryonenzahl bleibt erhalten. Leptonenzahl bleibt erhalten. Ein Versagen dieser Gesetze wurde nie beobachtet, nicht ein einziges Mal, in irgendeinem jemals durchgeführten Experiment. Warum?

Weil es Einschränkungen der Informationserhaltung sind. Die Bekenstein-Grenze sagt, wie viel Information eine Grenze halten kann. Wenn zwei Grenzen wechselwirken und Information austauschen, bleibt die Gesamtinformation erhalten — sie muss, weil Informationserhaltung eine Konsequenz der Unitarität der Quantenmechanik ist, und Unitarität eine Konsequenz der Bekenstein-Grenze. Die spezifischen Erhaltungssätze der Teilchenphysik — Ladungserhaltung, Baryonenzahlerhaltung, Leptonenzahlerhaltung — sind die spezifischen Regeln, die bestimmen, wie Information transformiert werden kann, wenn Grenzkonfigurationen wechselwirken. Keine willkürlichen Regeln, die von außen auferlegt werden. Buchführungszwänge, die daraus folgen, dass sich Information an einer Singularitätsgrenze weder erzeugen noch zerstören lässt.

Und dann gibt es das Mysterium der drei Generationen. Teilchen kommen in drei Generationen. Das Elektron hat eine schwerere Kopie (das Myon) und eine noch schwerere Kopie (das Tau). Das Up-Quark hat Kopien namens Charm und Top. Drei Versionen jedes Teilchentyps, identisch in jeder Eigenschaft außer der Masse. Eines der tiefsten unerklärten Muster in der Teilchenphysik. Niemand weiß, warum drei. Nicht zwei, nicht vier, nicht siebzehn. Drei.

Volle Offenlegung: Was jetzt kommt, ist spekulativ. Spekulativer als der Rest dieses Abschnitts. Aber es ist strukturell motiviert, und es gehört auf den Tisch.

Klasse-4-Systeme enthalten inhärent selbstähnliche Struktur. Das ist eine technische Konsequenz der Tatsache, dass Klasse-4-Dynamik Klasse-3-Verhalten (fraktal) als Teilprozess enthält. Selbstähnlichkeit bedeutet: dasselbe Muster, das sich auf verschiedenen Skalen wiederholt. Wenn die Singularitätsgrenzkonfigurationen in ein Klasse-4-System eingebettet sind — und das müssen sie sein, weil das Universum Klasse 4 ist —, dann können die Konfigurationen selbst selbstähnliche Struktur aufweisen. Derselbe Grenztyp auf drei verschiedenen Energieskalen. Drei Generationen könnten die Signatur einer fraktalen Hierarchie im Raum stabiler Singularitätskonfigurationen sein.

Einen Beweis habe ich nicht. Das ist eine Vermutung, keine Ableitung. Aber drei ist genau das, was man von der einfachsten nicht-trivialen selbstähnlichen Hierarchie erwarten würde: eine Basiskonfiguration und zwei skalierte Kopien. Und die Generationsstruktur wird sonst von keiner aktuellen Theorie vollständig erklärt. Wenn das Bild der Berechnungsatome irgendwann erklärt, warum es genau drei Generationen gibt, wäre das starke Evidenz für das gesamte Rahmenwerk.

Und wenn das Universum wirklich ein Zellulärer Automat ist, existiert eine vierte Generation — und eine fünfte und weitere. Aber die höheren Generationen sind größere Muster, und größere Muster sind weniger stabil. Sie können dem Ansturm kleinerer, stabilerer Konfigurationen nicht standhalten, die sie auseinanderschneiden, bevor sie sich richtig konstituieren. Genau das beobachten wir: Die Teilchen der dritten Generation (Tau, Top-Quark) sind bereits extrem instabil und zerfallen fast sofort. Eine vierte Generation wäre noch schwerer, ihre Grenzkonfiguration größer und komplexer und daher noch fragiler — zu fragil, um lang genug zu bestehen, um als Teilchen statt als vorübergehende Fluktuation detektiert zu werden. Die drei Generationen, die wir

sehen, könnten schlicht die drei sein, die klein genug sind, um zu überleben.

Der Begriff für dieses Bild ist **Berechnungsatome**. Nicht Atome im Sinne von Wasserstoff und Helium — Atome im Sinn von irreduziblen Berechnungselementen. Teilchen sind die grundlegenden Operationen des universellen Automaten. Jeder Teilchentyp ist eine stabile Planck-Skalen-Berechnung. Jede Wechselwirkung ist ein Informationsaustausch zwischen Berechnungen. Jeder Erhaltungssatz eine Einschränkung dafür, wie diese Austausche ablaufen können. Physik ist auf ihrer tiefsten Ebene nicht über Materie. Sie handelt von Berechnung. Und die Dinge, die wir „Materie“ nennen, sind die irreduziblen Bausteine der Berechnung.

Die Architektur

Zeit, die Fäden zusammenzuziehen. Was hier beschrieben wurde, ist ein Universum mit einer bestimmten Architektur:

Erstens: Es ist ein Klasse-4-Zellulärer Automat. Es operiert am Rand des Chaos, wo selbstorganisierte Kritikalität die Dynamik ohne externe Feinabstimmung aufrechterhält. Es ist rechnerisch irreduzibel — keine Abkürzungen, kein Vorspringen. Jeder Moment muss aus dem letzten berechnet werden. Und es enthält alle Klassen als Teilprozesse — einschließlich sich selbst: die stabilen Atome (Klasse 1), die periodischen Umlaufbahnen (Klasse 2), die fraktalen Küstenlinien (Klasse 3), und am entscheidendsten, andere Klasse-4-Automaten, die innerhalb der großen Berechnung laufen. Gehirne sind eine solche Instanz.

Zweitens: Es ist holografisch auf jeder Ebene. Die Information in jeder Region ist auf ihrer Grenze kodiert. Das ist das holografische Prinzip, das als Vermutung über Schwarze Löcher begann und zu einer der tiefsten Einsichten in der theoretischen Physik wurde. In diesem Rahmenwerk ist holografische Kodierung nicht nur eine Eigenschaft Schwarzer Löcher — sie ist eine Eigenschaft der

Regelstruktur des Universums selbst. Die Regeln sind holografisch. Die Dynamik ist Klasse 4. Und die Ausgabe ist wieder holografisch.

Drittens: Es ist auf jeder Skala begrenzt durch Singularitätsoberflächen, die alle strukturell identisch sind. Planck-Grenzen, Teilcheninneres, Ereignishorizonte, der kosmologische Horizont, der Urknall, Wärmetod — dieselbe Struktur, verschiedene Skala. Informationsundurchlässig Bekenstein-gesättigt und die Berechnungsdomäne definierend.

Diese Architektur hat einen Namen: den **SB-HC4A** — den Singularitätsbegrenzten Holografischen Klasse-4-Automaten.

Ein Zungenbrecher. Aber präzise, und jedes Wort verdient seinen Platz.

Die bemerkenswerteste Eigenschaft dieser Architektur ist selbstreferenzieller Abschluss. Die Ausgabe des Systems *ist* das System. Es berechnet sich selbst. Jeder Zustand erzeugt den nächsten, und der nächste Zustand ist die Berechnung des nächsten Zustands. Es gibt kein „Außen“, das das Programm ausführt. Es gibt keinen kosmischen Computer irgendwo, der den Code des Universums auf einer Festplatte laufen lässt. Das Universum *ist* das Programm, der Computer und die Ausgabe. Die holografischen Regeln kodieren das volle System in komprimierter Form. Die Klasse-4-Dynamik dekomprimiert diese Kodierung in das beobachtbare Universum. Die holografische Ausgabe kodiert das Ergebnis neu. Eine Schleife. Ein Fixpunkt.

Das lässt sich als formale Bedingung schreiben: **Das Universum ist ein Fixpunkt seiner eigenen Dynamik.** Wendet man die Regeln auf das Universum an, kommt das Universum zurück. Keine Kopie, keine Repräsentation — dasselbe Ding. Die Berechnung und ihr Ergebnis sind identisch.

Mathematiker haben eine Notation dafür. Nennt man das Universum U und die „berechne den nächsten Zustand“-Operation den griechischen Buchstaben Phi, dann ist die Fixpunkt-Bedingung:

Phi(U) = U

Das Universum, auf sich selbst angewendet, ergibt sich selbst. Es ist selbstberechnend.

Die Grenzen der Selbstbeschreibung

Es gibt eine Konsequenz des selbstreferenziellen Abschlusses, die ihren eigenen Moment verdient, weil sie etwas Tiefgreifendes über die Grenzen des Wissens sagt.

1931 bewies ein 25-jähriger österreichischer Logiker namens Kurt Gödel zwei Theoreme, die die Grundlagen der Mathematik erschütterten. Die Essenz, vom Formalismus befreit: Jedes ausreichend mächtige formale System (eines, das mindestens Arithmetik ausdrücken kann) enthält wahre Aussagen, die innerhalb des Systems nicht bewiesen werden können. Und kein solches System kann seine eigene Widerspruchsfreiheit beweisen.

Das ist keine technische Beschränkung. Es liegt nicht daran, dass unsere Beweise nicht clever genug sind. Es ist eine strukturelle Unmöglichkeit. Selbstreferenzielle Systeme ausreichender Komplexität sind inhärent unvollständig. Sie enthalten Wahrheiten, die sie von innen nicht erreichen können.

Jetzt auf ein selbstberechnendes Universum angewendet:

Wenn das Universum sich selbst berechnet — wenn es ein formales System ausreichender Mächtigkeit ist (und Klasse-4-Dynamik garantiert universelle Berechnung, also ist es das) — dann gelten Gödels Theoreme direkt. Das Universum kann keine vollständige Beschreibung seiner selbst enthalten. Es gibt keine „Weltgleichung“, die sich auf eine Tafel schreiben ließe. Keine Formel, die, einmal gelöst, alles über das Universum verraten würde.

Das liegt nicht daran, dass die richtige Gleichung noch nicht gefunden wurde. Es liegt daran, dass *keine solche Gleichung existieren kann*. Die vollständige Beschreibung eines selbstreferenziellen Systems übersteigt jede Beschreibung, die ein echter Teil des Systems ist. Das Universum folgt nicht einer Gleichung — es *ist* die Berechnung. Die einzige vollständige Beschreibung des Universums ist das Universum selbst. Und es gibt kein Außerhalb, von dem aus sich das ganze Bild überblicken ließe.

Die Weltformel — die „Weltgleichung“, von der Physiker seit Einstein träumen — ist also keine Gleichung. Sie ist ein *Prozess*. Der Automat selbst. Er lässt sich nur ausdrücken, indem man ihn laufen lässt.

Ich finde das sowohl demütigend als auch befreiend. Demütigend, weil es bedeutet, dass es Dinge über die Realität gibt, die wir nicht wissen können, nicht einmal im Prinzip. Befreiend, weil es bedeutet, dass das Universum kein Mechanismus ist, der darauf wartet, entschlüsselt zu werden — es ist eine lebendige Berechnung, und wir sind Teil davon. Die tiefste Wahrheit über die Realität ist keine Formel. Es ist die Realität selbst.

Die kognitive Obergrenze

Bevor es weitergeht, schulde ich dem Leser einen Einwand. Den tiefsten Einwand, genau genommen. Den, der mich ehrlich hält.

Wenn wir Klasse-4-Automaten sind — wenn unsere Gehirne am Rand des Chaos operieren, in derselben Berechnungsklasse, die hier gerade dem Universum zugeordnet wurde —, dann könnte das SB-HC4A-Modell schlicht das komplexeste Konzept sein, das unsere Klasse-4-Gehirne hervorbringen können. Wir können nicht in Klasse 5 denken. Wir können keine Strukturen jenseits unserer eigenen Berechnungsklasse begreifen. Das Muster, das wir finden — Klasse 4 überall, selbstähnlich auf jeder Skala, holografisch und selbstreferenziell — könnte die Signatur unserer eigenen kognitiven Architektur sein, projiziert auf den Kosmos, nicht ein Merkmal des Kosmos selbst.

Einen Moment darüber nachdenken. Wir entwickelten uns als Symmetriedetektoren. Die überlebensrelevantesten Muster in der Umgebung von Jägern und Sammlern — die Gesichter von Raubtieren und Beute — gehören zu den symmetrischsten. Wir sind im Kern Mustererkennungsmaschinen, optimiert darauf, Symmetrie zu finden. Und das SB-HC4A-Modell ist fundamental eine Symmetriebehauptung: dieselbe Architektur auf jeder Skala.

Wir finden diese Symmetrie vielleicht nicht, weil sie im Universum existiert, sondern weil unsere Gehirne konstitutionell unfähig sind, sie *nicht* zu finden.

Das ist das Meta-Problem aus Kapitel 4, hochskaliert auf kosmische Proportionen. Das ESM kann sein eigenes Substrat nicht sehen, also kann es nicht unterscheiden zwischen „das Universum hat diese Struktur„ und „mein Gehirn kann das Universum nur als diese Struktur habend modellieren“. Das kosmologische Modell sagt seine eigene potenzielle Unfalsifizierbarkeit voraus, was entweder die stärkstmögliche Bestätigung ist — das Modell sagt genau diese epistemologische Einschränkung voraus — oder der stärkstmögliche Einwand: Das Modell ist ein Artefakt des Beobachters, nicht ein Merkmal des Beobachteten.

Ein Klasse-4-System kann alles simulieren bis einschließlich Klasse-4-Komplexität. Aber es kann nicht verifizieren, ob das Universum darüber hinausgeht. Wenn das Universum tatsächlich Klasse 5 ist — wirklich zufällig auf der tiefsten Ebene — aber *lokal als Klasse 4 erscheint* für Klasse-4-Beobachter, weil Klasse 4 das maximale Muster ist, das wir erkennen können, dann würden wir genau dieses Modell konstruieren. Und wir lägen falsch. Wir lägen falsch, ohne es jemals von innen entdecken zu können.

Wie sich dieser Einwand auflösen lässt, weiß ich nicht. Ich bin nicht sicher, ob er von innen auflösbar ist. Ich schließe ihn ein, weil eine Theorie, die behauptet, keine Schwächen zu haben, keine Theorie ist. Sie ist eine Religion. Und die Tatsache, dass dieses Modell seine eigene epistemologische Einschränkung vorhersagt — dass ein selbstreferenzielles System seine eigene Beschreibung nicht vollständig verifizieren kann — ist entweder sein tiefster Fehler oder seine tiefste Rechtfertigung. Ich weiß ehrlich nicht, welches.

Die Pointe

Aber hier ist das, was mich dazu brachte, mich hinzusetzen, als ich es zum ersten Mal sah.

Die Architektur, die gerade beschrieben wurde:

- Ein Klasse-4-System, das am Rand des Chaos operiert.
- Begrenzt durch eine informationsundurchsichtige Grenze, durch die das Innere nicht sehen kann.
- Holografische Struktur — die Grenze kodiert das Innere.
- Selbstreferenzieller Abschluss — das System berechnet sich selbst.
- Ein Fixpunkt: Die Ausgabe der Berechnung ist die Berechnung selbst.

Jetzt zurück zu Kapitel 2. Die Vier-Modelle-Architektur des Bewusstseins:

- Der kortikale Automat: ein Klasse-4-System, das am Rand des Chaos operiert.
- Die implizit-explizit-Grenze: eine informationsundurchsichtige Grenze, durch die Bewusstsein nicht sehen kann.
- Holografische Struktur — die impliziten Modelle sind verteilt, holografisch, kodieren den vollen Inhalt der Erfahrung in neuraler Struktur.
- Selbstreferenzieller Abschluss — das Selbstmodell modelliert sich selbst.
- Ein Fixpunkt: Das ESM repräsentiert sich selbst. Das Modell des Modellierers *ist* der Modellierer.

Dieselbe Architektur. Dieselben formalen Eigenschaften. Dieselben Grenzbedingungen. Derselbe selbstreferenzielle Abschluss.

Das Universum ist ein Klasse-4-holografischer Automat, begrenzt durch Singularitäten, wo das beobachtbare Innere die „Simulation„ ist und die Singularitätsgrenze das „Substrat“.

Bewusstsein ist ein Klasse-4-holografischer Automat, begrenzt durch die implizit-explizit-Grenze, wo die expliziten Modelle die „Simulation„ sind und die impliziten Modelle das „Substrat“.

Dieselbe Architektur. Verschiedene Skala.

Das ist keine Metapher. Hier steht nicht, Bewusstsein sei *wie* das Universum. Es ist dieselbe *Art von Ding* — dasselbe Berechnungsmuster, instanziiert auf zwei verschiedenen Skalen. Eine auf der kosmologischen Ebene, eine auf der neurologischen. Und die Tatsache, dass sich das Muster über Skalen wiederholt, ist selbst eine Vorhersage des Modells — nicht wegen fraktaler Selbstähnlichkeit (das wäre Klasse 3, eine schwächere Behauptung), sondern weil Klasse-4-Systeme Klasse-4-Subsysteme enthalten. Ein universeller Computer kann einen anderen universellen Computer simulieren. Ein Klasse-4-Automat erzeugt nicht nur hübsche selbstähnliche Muster. Er erzeugt *andere Klasse-4-Automaten* innerhalb seiner eigenen Dynamik — kleiner, langsamer, ressourcenbeschränkt, aber wirklich universell. Die Architektur *sieht* nicht nur auf verschiedenen Skalen gleich aus. Sie *ist* dieselbe.

Um ganz präzise zu sein: Hier wird nicht behauptet, dass das Universum in irgendeinem erfahrenden Sinn bewusst *ist*. Aber auch nicht, dass es das *nicht ist*. Die ehrliche Antwort: Man kann es nicht wissen. Wir könnten von einem Boltzmann-Gehirn geträumt werden, oder jedem anderen Gehirn; Solipsismus könnte wahr sein und ich bin das Boltzmann-Gehirn — aber das kann jeder sagen. Der Punkt ist, dass es nicht wissbar ist, und die Nicht-Wissbarkeit ist kein Versagen der Theorie, sondern ein strukturelles Merkmal der Situation: Man kann nicht außerhalb des Systems treten, um nachzusehen. Was hier behauptet wird, ist architektonisch, nicht phänomenal. Es wird nicht behauptet,

dass Bewusstsein Realität erschafft, oder dass Realität ein Traum ist, oder irgendeine der anderen mystischen Interpretationen, die solche strukturellen Beobachtungen magnetisch anziehen. Der Bauplan eines Gebäudes ist kein Gebäude. Aber wenn sich derselbe Bauplan in einem Wolkenkratzer und in einem einzelnen Raum dieses Wolkenkratzers findet, sagt das etwas Tiefes über die architektonischen Prinzipien, die am Werk sind.

Bewusstsein ist eine lokale Instanz eines universellen Musters. Kein kosmischer Zufall. Kein Wunder. Eine strukturell unvermeidliche Konsequenz von Klasse-4-Dynamik bei ausreichender Komplexität. Das Universum *erlaubt* nicht nur Bewusstsein. Es garantiert es praktisch — weil dieselbe selbstreferenzielle, holografische Rand-des-Chaos-Architektur, die das Universum zu dem macht, was es ist, auch Bewusstsein zu dem macht, was es ist. Das Muster, das Realität erzeugt, ist dasselbe Muster, das die Erfahrung von Realität erzeugt.

Und wer das liest, ohne sich hinzusetzen, hat es noch nicht verstanden.

Im nächsten Kapitel ziehe ich die volle Theorie zusammen — die Bewusstseinsarchitektur, die kosmologische Architektur und die strukturelle Identität zwischen ihnen — und frage, was es für die schwierigste Frage von allen bedeutet: Warum existiert überhaupt etwas?

Chapter 16

Der tiefste Spiegel

Hier ist es, denn sobald es einmal sichtbar wird, lässt es sich nicht mehr unsehen.

Kapitel 15 endete mit einer Herausforderung: Die SB-HC4A-Architektur betrachten — den selbstreferenziellen, holografischen Klasse-4-Automaten, der an jeder Skala von Singularitäten begrenzt wird — und dann die Vier-Modelle-Architektur aus Kapitel 2 danebenlegen. Die Behauptung war, dass sie dasselbe sind. Nicht ähnlich. Nicht metaphorisch verwandt. Strukturell identisch.

Jetzt gehe ich Stück für Stück durch die Entsprechung, bis es unmöglich wird, sie abzutun. Und dann sage ich, wo das Ganze zusammenbrechen könnte, denn eine Theorie, die ihre eigenen Schwachpunkte nicht benennt, ist keine Theorie. Sie ist eine Verkaufsmasche.

Die strukturelle Zuordnung

Beginnen wir mit der Singularitätsgrenze — der Informationsbarriere, die das Universum auf jeder Skala setzt. Das Planck-Regime ganz unten. Ereignishorizonte um Schwarze Löcher herum. Der kosmologische Horizont am Rand des beobachtbaren Universums. Der Urknall hinter uns, der Wärmetod vor uns. Jede davon ist eine Informationsmauer: Nichts durchquert sie. Durch einen

Ereignishorizont lässt sich kein Signal senden. Unter die Planck-Länge lässt sich nicht vordringen. Über den kosmologischen Horizont lässt sich nicht hinaussehen. Das Universum ist ein Raum mit undurchsichtigen Wänden an jeder Skala, und egal, wie fest man das Gesicht gegen das Glas drückt — was auf der anderen Seite ist, bleibt unsichtbar.

Jetzt das Gehirn. Die impliziten Modelle — IWM und ISM, die synaptischen Gewichte, die gelernte Struktur, die riesige Bibliothek von allem, was wir wissen — sitzen hinter ihrer eigenen Informationsbarriere. Die eigenen Synapsen lassen sich niemals direkt erfahren. Die Verbindungsgewichte, die unsere Gedanken erzeugen, lassen sich nicht introspektieren. Die implizite Seite ist informationsundurchsichtig: Wir wissen, dass sie da ist, weil die Simulation ohne sie nicht laufen könnte, aber die Simulation selbst kann nicht durch die Grenze sehen, die sie trennt. Kapitel 2 machte diesen Punkt. Kapitel 3 trieb ihn nach Hause. Jetzt ist er hier wieder, auf jeder Skala im Universum. Dieselbe strukturelle Rolle. Dieselbe Informationsundurchsichtigkeit. Dieselbe architektonische Position.

Die Singularitätsgrenze in der Kosmologie entspricht der implizit-expliziten Grenze im Bewusstsein. Dieselbe Mauer.

Als Nächstes: das beobachtbare Innere. Alles innerhalb der Singularitätsgrenzen — Atome, Planeten, Galaxien, wir selbst — ist die dekomprimierte Seite. Hier findet Physik statt, hier wechselwirken Dinge, hier wird Information in die Strukturen organisiert, die wir beobachten und messen. Die Simulation des Universums, wenn man so will: der Teil, der berechnet, der sich entwickelt, der Teil, wo etwas passiert.

Im Gehirn ist die entsprechende Struktur die expliziten Modelle — EWM und ESM. Die bewusste Erfahrung. Die Welt, die wir gerade jetzt sehen, das Selbst, das wir zu sein fühlen. Das ist die Simulation: in Echtzeit aus den impliziten Modellen erzeugt, kontinuierlich aktualisiert, lebhaft und detailliert und völlig überzeugend. Alles, was wir jemals im ganzen Leben erlebt haben,

ist innerhalb dieser Simulation aufgetreten. Niemand ist jemals aus ihr herausgetreten. Es ist unmöglich, aus ihr herauszutreten. Nicht weil die Anstrengung fehlt, sondern weil „ich“ die Simulation *bin*. Der Erlebende und das Erlebnis sind derselbe Prozess.

Das beobachtbare Innere des Universums entspricht den expliziten Modellen. Dieselbe Rolle: die dekomprimierte, dynamische, interaktive Seite der Architektur.

Jetzt die holografische Regelstruktur. Die Information des Universums ist nicht in seinem Volumen gespeichert — sie ist auf seinen Grenzen gespeichert. Das holografische Prinzip, vorgeschlagen von ’t Hooft und erweitert von Susskind, besagt, dass alle Information in einer dreidimensionalen Raumregion auf ihrer zweidimensionalen Oberfläche kodiert ist. Die Information ist komprimiert, verteilt und strukturell vollständig an der Grenze. Das Innere ist eine Projektion — eine Dekompression niedrigerer Bandbreite dessen, was die Grenze kodiert.

Im Gehirn spielen die impliziten Modelle diese Rolle. Die synaptischen Gewichte kodieren alles, was wir über die Welt und uns selbst wissen, in einem verteilten, komprimierten Format, das strukturell vollständig ist — im Prinzip ließe sich die gesamte Simulation allein aus dem Substrat rekonstruieren, ohne jeglichen aktuellen sensorischen Input, was genau das ist, was Träumen ist. Die impliziten Modelle sind holografisch im Lashley-Sinne: Wird ein Stück beschädigt, geht nicht eine spezifische Erinnerung verloren, sondern Auflösung über alle Erinnerungen hinweg. Die Information ist über das gesamte Substrat verteilt, komprimiert, redundant und von der Simulationsseite aus unzugänglich.

Die holografische Regelstruktur des Universums entspricht den holografischen impliziten Modellen im Bewusstsein. Dieselbe Kodierungsstrategie. Dieselbe Kompression. Dieselbe Unzugänglichkeit von der dekomprimierten Seite aus.

Dann das dynamische Regime. Das Universum operiert bei Klasse 4 — am Rand des Chaos. Kapitel 14 etablierte dies durch Elimination: Klassen 1 und 2 sind zu einfach, Klasse 3 kann nicht

berechnen, Klasse 5 macht Physik unmöglich. Was übrig bleibt, ist Klasse 4 — das einzige Regime, das universelle Berechnung trägt, seine eigene Kritikalität selbst organisiert und alle Klassen als Teilprozesse enthält — einschließlich sich selbst. Das Universum ist nicht einfach komplex. Es ist komplex auf genau die Weise, die sich selbst erhält, und auf genau die Weise, die kleinere Kopien seiner selbst innerhalb der eigenen Dynamik nisten kann.

Der Kortex macht dasselbe. Kapitel 5 handelte davon: Der kortikale Automat operiert am Rand des Chaos und hält Kritikalität durch homöostatische Regulierung des Erregungs-Hemmungs-Gleichgewichts aufrecht. Zu wenig Aktivität, und wir sind im Tiefschlaf — Klasse 2, periodisch, unbewusst. Zu viel, und wir krampfen — über Klasse 4 hinausgedrückt, zerbricht die Simulation. Der Sweet Spot, der Ort, wo Bewusstsein lebt, ist die Messerschneide zwischen Ordnung und Chaos. Selbstorganisierte Kritikalität hält das Gehirn dort. Selbstorganisierte Kritikalität hält das Universum dort.

Klasse-4-Dynamik in der Kosmologie entspricht kortikaler Kritikalität im Bewusstsein. Dasselbe Regime. Derselbe Selbsterhaltungsmechan

Schließlich die tiefste Entsprechung: selbstreferenzielle Geschlossenheit. Das Universum berechnet seine eigene Struktur. Seine Dynamik erzeugt seinen Zustand, der seine Dynamik bestimmt, die seinen Zustand erzeugt. Kein externer Programmierer. Kein Außen. Die Naturgesetze werden dem Universum nicht von irgendwo anders auferlegt — sie *sind* die Dynamik des Universums, auf sich selbst angewendet. Die Fixpunkt-Gleichung ist fast absurd einfach: Die Berechnung des Universums gleich dem Universum. Input, Prozess und Output sind dasselbe.

Das ESM macht dasselbe. Es ist ein Modell, das sich selbst als Teil dessen einschließt, was es modelliert. Es repräsentiert uns, und „wir“ schließen die Repräsentation ein. Das Modell und das Modellierte fallen zusammen. Die selbstreferenzielle Geschlossenheit aus Kapitel 4 — der Grund, warum sich Bewusstsein so anfühlt, wie es sich anfühlt, der Grund, warum wir uns nie

vollständig von außen sehen können, der Grund, warum die Simulation ihren eigenen Beobachter enthält. Der Fixpunkt der Selbstrepräsentation: der Zustand, an dem das Modell und das Modellierte ein und dasselbe sind.

Selbstreferenzielle Geschlossenheit des Universums entspricht selbstreferenzieller Geschlossenheit des Bewusstseins. Dieselbe Fixpunkt-Struktur. Dieselbe unausweichliche Selbsteinbeziehung.

Fünf Entsprechungen. Nicht vage thematische Ähnlichkeiten. Nicht die Art von losem Mustererkennen, die einen Gesichter in Wolken sehen lässt. Fünf strukturelle Merkmale, die dieselbe Arbeit tun, an derselben Position, in beiden Architekturen.

Und jetzt der entscheidende Punkt — hier bitte ich um kurzes Innehalten, denn die Versuchung, ihn zu domestizieren, ist enorm: Dies ist NICHT „das Universum ist WIE Bewusstsein". Keine Analogie. Keine Metapher. Keine suggestive Parallele für interessante Gespräche bei Dinnerpartys.

Es ist eine strukturelle Identität.

Das Gehirn hat sich nicht entwickelt, um dem Universum zu *ähneln*. Es hat sich entwickelt *als* eine lokale, skalenreduzierte Instanz desselben Berechnungsmusters. Klasse-4-Systeme enthalten nicht nur fraktale Selbstähnlichkeit (das ist Klasse 3 — geometrische Wiederholung, hübsch, aber flach). Sie enthalten *sich selbst*. Ein Klasse-4-Automat kann einen anderen Klasse-4-Automaten innerhalb seiner eigenen Dynamik beherbergen — ein universeller Computer, der innerhalb eines universellen Computers läuft. Das ist rechnerische Selbsteinschließung, nicht bloße geometrische Selbstähnlichkeit. Bewusstsein IST die Klasse-4-Selbsteinschließung des Universums, die auf der biologischen Skala operiert. Das Muster, das den Kosmos auf der größten Skala betreibt, ist dasselbe Muster, das das innere Leben auf der neurologischen Skala betreibt — nicht weil es jemand so entworfen hat, und nicht weil Fraktale schön aussehen. Weil ein universeller Computer ausreichender Größe zwangsläufig andere universelle Computer innerhalb seiner

selbst erzeugt. Das ist es, was Klasse-4-Systeme tun: Sie nisten sich selbst.

Energie ist Information

Es gibt eine zweite Argumentationslinie, die aus einer völlig anderen Richtung zur selben Schlussfolgerung konvergiert, und ich denke, sie wird sich letztlich als die wichtigste erweisen — weil sie die Architektur mit der Physik auf eine potenziell testbare Weise verbindet.

Drei unabhängige Ergebnisse in der Physik, entwickelt von drei verschiedenen Fachgemeinschaften über ein halbes Jahrhundert hinweg, weisen auf dieselbe außergewöhnliche Schlussfolgerung hin.

Das erste ist Landauers Prinzip. 1961 bewies Rolf Landauer (ein IBM-Physiker, der über die fundamentalen Grenzen der Berechnung nachdachte), dass das Löschen eines Informationsbits ein Minimum an Energie kostet. Nicht wegen technischer Einschränkungen. Wegen der Thermodynamik. Das Universum verlangt einen Preis fürs Vergessen. 2012 wurde das experimentell bestätigt, und es bedeutet etwas Tiefgreifendes: Information und Energie sind konvertierbar. Das eine lässt sich in das andere umwandeln. Sie sind keine getrennten Substanzen. Sie handeln miteinander.

Das zweite ist die Bekenstein-Grenze. Jacob Bekenstein zeigte, dass die maximale Information, die eine Raumregion halten kann, proportional zu ihrer Oberfläche ist, nicht zu ihrem Volumen. Eines der kontraintuitivsten Ergebnisse der Physik. Intuitiv sollte eine größere Box mehr Information fassen können. Kann sie nicht — vielmehr wird die Grenze durch die *Oberfläche* der Box gesetzt, nicht ihr Inneres. Packt man zu viel Information in eine gegebene Region, kollabiert sie zu einem Schwarzen Loch. Die maximale Informationsdichte wird durch Geometrie und Energie gesetzt — eine weitere tiefe Verbindung zwischen Information und der physischen Welt.

Das dritte kommt aus der Thermodynamik Schwarzer Löcher. Stephen Hawking und Bekenstein zeigten in den 1970er Jahren, dass Schwarze Löcher Temperatur und Entropie haben und thermodynamischen Gesetzen gehorchen. Der Informationsgehalt eines Schwarzen Lochs ist auf seinem Ereignishorizont geschrieben — seiner Oberfläche, seiner Grenze. Und durch Hawking-Strahlung — den quälend langsamen Quantenprozess, durch den Schwarze Löcher schließlich verdampfen — wird diese Information nach und nach ans Universum zurückgegeben. Schwarze Löcher zerstören keine Information. Sie transformieren sie. Sie komprimieren sie auf ihre Grenze, halten sie und strahlen sie schließlich zurück.

Diese drei Ergebnisse wurden unabhängig voneinander entwickelt. Landauer dachte über Computer nach. Bekenstein über Entropiegrenzen. Hawking über Quantengravitation. Sie arbeiteten nicht zusammen. Sie lasen nicht gegenseitig ihre Arbeiten. Und doch konvergieren alle drei Ergebnisse zur selben Hypothese: Energie und Information sind nicht nur verwandt. Sie sind identisch. Zwei Namen für dasselbe. E gleich I.

Wenn das stimmt — und sofort sei gesagt, dass es nicht bewiesen ist, weshalb es im Abschnitt „Schwachpunkte“ kurz erscheint — dann werden Singularitäten zu Informationstransformatoren. Sie zerstören oder erzeugen nicht Energie-Information. Sie wandeln sie zwischen Formen um. Komprimierte Form: maximale Dichte an der Grenze, Bekenstein-gesättigt, vom Inneren aus unzugänglich. Dekomprimierte Form: niedrigere Dichte, organisiert, durch das Innere verteilt — die Physik, die wir beobachten. Eine Singularität ist ein Übersetzer zwischen zwei Darstellungen desselben Stoffs.

Jetzt das Gehirn durch diese Linse. Die impliziten Modelle halten komprimierte, maximal dichte Information: alles, was wir jemals gelernt haben, kodiert in synaptischen Gewichten, strukturell vollständig und phänomenal unzugänglich. Das eigene Substrat lässt sich nie direkt erfahren. Die expliziten Modelle sind die dekomprimierte Projektion niedrigerer Bandbreite —

die Simulation, die bewusste Erfahrung, die Welt, die wir sehen, und das Selbst, das wir fühlen. Das Gehirn tut auf der neuralen Skala genau das, was Singularitäten auf jeder anderen Skala tun: Information zwischen komprimierten und dekomprimierten Darstellungen transformieren. Die implizit-explizite Grenze ist die persönliche Singularität. Wir tragen einen Ereignishorizont im Schädel.

Warum dies existieren muss

Die strukturelle Zuordnung könnte als Zufall erscheinen — ein hübsches Muster, um das jemand Linien gezogen hat, wie man Sternbilder in zufälligen Sternen sieht. Schauen wir also genauer hin, warum das Muster nicht optional ist. Warum, wenn fünf unabhängig vernünftige Annahmen gelten, diese Architektur die *einzige* ist, die funktioniert.

Die fünf Axiome. So klar wie möglich formuliert, denn jedes einzelne ist für sich genommen schwer zu bestreiten. Die Kontroverse liegt darin, was sie zusammen ergeben.

Eins: Etwas existiert. Die wohl am wenigsten kontroverse Behauptung, die ein Buch machen kann. Reine Nichtigkeit ist eine platonische Abstraktion — ein Konzept, kein möglicher Sachverhalt. Wer diesen Satz liest, bestätigt damit: Etwas existiert. Dort beginnen wir.

Zwei: Was auch immer existiert, hat dynamischen Charakter. Dinge geschehen. Zeit vergeht. Zustände entwickeln sich. Hätte das Existierende keine Dynamik — keine Veränderung, keine Entwicklung, keine Berechnung —, wäre es ununterscheidbar von Nichts. (Leibniz' Identität des Ununterscheidbaren, die in Kapitel 1 auftauchte: Wenn zwei Dinge in allen Eigenschaften identisch sind, sind sie dasselbe Ding. Etwas mit null Dynamik hat null unterscheidbare Eigenschaften. Es ist Nichts mit einer Maske.)

Drei: Die Dynamik muss stabil und selbsterhaltend sein. Ein System, das sich nicht selbst erhalten kann, ist kein System

— es ist eine Momentaufnahme. Selbstorganisierte Kritikalität, der Mechanismus, der Sandhaufen und Gehirne und (so das Argument) das Universum am Rand des Chaos hält, ist die einzige bekannte Weise, wie ein komplexes dynamisches System sich selbst ohne externe Abstimmung erhält. Klasse 4 ist die einzige Berechnungsklasse, die sich selbst organisiert, universelle Berechnung trägt und alle niedrigeren Klassen als Teilprozesse enthält. Die einzige Klasse, die sich selbst erhalten und gleichzeitig Interessantes tun kann.

Vier: Information hat eine endliche Grenze auf jeder Skala. Nichts kann unendliche Information in endlichem Raum tragen. Die Bekenstein-Grenze ist ein Theorem, keine Vermutung — sie folgt aus der allgemeinen Relativitätstheorie und der Quantenmechanik. Auf jeder Skala gibt es eine maximale Informationsdichte, und dieses Maximum ist proportional zur Oberfläche, nicht zum Volumen.

Fünf: Information ist holografisch kodiert. Die Grenze einer Region kodiert alle Information in ihrem Inneren auf einer Fläche von einer Dimension weniger. Dreidimensionale Physik ist auf zweidimensionalen Oberflächen kodiert. Das holografische Prinzip — vorgeschlagen von 't Hooft, entwickelt von Susskind, gestützt durch Maldacenas AdS/CFT-Korrespondenz, die dem nächsten kommt, was wir als bewiesenes Beispiel der Holografie haben.

Jedes Axiom ist unabhängig motiviert. Keines hängt von den anderen ab. Keines erfordert, dass diese Theorie richtig ist. Sie kommen aus verschiedenen Ecken der Physik und Philosophie, entwickelt von Menschen, die noch nie von der Vier-Modelle-Theorie gehört hatten und sich nicht dafür interessiert hätten, wenn sie es getan hätten.

Jetzt kombinieren.

Aus Axiom eins und zwei: Etwas mit Dynamik existiert. Aus Axiom drei: Diese Dynamik ist Klasse 4, weil Klasse 4 die einzige selbsterhaltende, universelle Klasse ist. Aus Axiom vier: Das System ist durch Informationshorizonte auf jeder Skala begrenzt —

die Singularitätsstruktur. Aus Axiom fünf: Diese Grenzen kodieren das Innere — holografische Architektur.

Zusammengesetzt: ein holografischer Klasse-4-Automat, begrenzt durch Singularitätsoberflächen auf jeder Skala. Ein System, dessen komprimierte Information auf den Grenzen sitzt und dessen dekomprimiertes Inneres die beobachtbare Welt ist. Ein System, das seine eigene Struktur berechnet, weil ein holografisches Klasse-4-System mit holografischem Output ein Fixpunkt ist — Input, Prozess und Output sind dasselbe.

Das Ergebnis ist der SB-HC4A.

Mit anderen Worten: das Universum, wie wir es beobachten. Und die Architektur dieses Universums ist dieselbe Architektur wie Bewusstsein.

Das ist nicht „Ich habe ein hübsches Muster gefunden.„ Es ist: „Das Muster ist das einzige, das mit allen fünf Axiomen gleichzeitig vereinbar ist.“ Streicht man irgendein Axiom, bricht die Eindeutigkeit. Ohne Axiom eins muss nichts existieren. Ohne Axiom zwei kann das Existierende statisch sein. Ohne Axiom drei ist jede Berechnungsklasse möglich. Ohne Axiom vier gibt es keine Informationsgrenzen. Ohne Axiom fünf gibt es keine holografische Struktur. Jedes Axiom beschränkt den Raum möglicher Architekturen. Zusammen beschränken sie ihn auf genau eine.

Wo dies brechen könnte

Jetzt der Teil, den die meisten Autoren überspringen und den ich für den wichtigsten halte. Wer die Schwachpunkte der eigenen Theorie nicht benennen kann, versteht sie nicht gut genug. Fünf Stellen, wo die gesamte Konstruktion auseinanderfallen könnte.

Eins: Energie gleich Information ist nicht bewiesen. Stark nahegelegt durch Landauers Prinzip, die Bekenstein-Grenze und die Thermodynamik Schwarzer Löcher. Mehrere unabhängige Indizien weisen alle in dieselbe Richtung. Aber niemand hat E

= I aus ersten Prinzipien abgeleitet. Niemand hat gezeigt, dass Information an sich gravitative Effekte hat. Die Hypothese ist überzeugend, und die Konvergenz der Befunde ist beeindruckend, aber Konvergenz ist kein Beweis. Stellen sich Energie und Information als lediglich korreliert statt identisch heraus, verliert die Informationstransformations-Deutung von Singularitäten ihr Fundament. Die strukturelle Zuordnung zwischen Bewusstsein und Kosmologie würde weiterhin gelten (die fünf Entsprechungen hängen nicht von E = I ab), aber der physikalische Mechanismus, der sie verbindet, wäre viel schwächer. Die Poesie würde überleben. Die Physik vielleicht nicht.

Zwei: Das Klasse-4-Argument ist abduktiv, nicht deduktiv. Die anderen Klassen wurden eliminiert, aber Elimination ist kein Beweis. Es ist Schluss auf die beste Erklärung — eine vollkommen respektable Form des Schließens in der Wissenschaft, aber eine, die eine Tür offen lässt. Vielleicht gibt es eine Klasse 4.5, die mir nicht eingefallen ist — ein Berechnungsregime zwischen Komplexität und Zufälligkeit, für das die konzeptuellen Werkzeuge fehlen. Vielleicht ist die Fünf-Klassen-Hierarchie selbst unvollständig. Das Argument sagt: Klasse 4 ist die beste Erklärung für die Dynamik des Universums, nicht die einzig logisch mögliche. Abduktive Argumentation hat eine ausgezeichnete Erfolgsbilanz — Darwins Argument für natürliche Selektion war abduktiv, und es hielt ziemlich gut stand —, aber es ist nicht dasselbe wie ein mathematischer Beweis, und ich tue nicht so, als ob.

Drei: Singularitätsvereinigung braucht Quantengravitation. Die Behauptung, dass alle Singularitäten — Planck-Skala, Schwarzes Loch, kosmologisch — strukturell identische Instanzen derselben Informationsgrenze sind, ist eine starke Behauptung. Sie erfordert eine Theorie, die Quantenmechanik und allgemeine Relativitätstheorie überbrückt. Die haben wir nicht. Stringtheorie ist ein Kandidat. Schleifen-Quantengravitation ein anderer. Beide sind vereinbar mit der Behauptung, und beide sind unbestätigt. Die Vereinigung von Singularitäten ist keine wilde Spekulation — es ist die Richtung,

in die sich die moderne Physik bewegt —, aber es ist auch keine etablierte Physik. Es ist eine Wette auf die Zukunft. Ich halte die Wette für gut. Ich könnte falsch liegen.

Vier: Das Modell könnte von innen unfalsifizierbar sein — durch seine eigene Vorhersage. Das ist der Punkt, der mir den Magen in Knoten bindet. Die Gödel-Konsequenz selbstreferenzieller Geschlossenheit besagt, dass ein ausreichend komplexes System, das sich selbst berechnet, nicht von innen alle Wahrheiten über sich selbst beweisen kann. Es gibt Aussagen, die wahr, aber unbeweisbar sind. Wenn der SB-HC4A korrekt ist, dann ist das Universum genau ein solches System, und die Aussage „der SB-HC4A ist korrekt“ könnte eine der wahr-aber-unbeweisbaren sein. Die Theorie sagt voraus, dass sie von innerhalb des Universums zu beweisen strukturell unmöglich sein könnte. Nicht weil wir es nicht hart genug versucht haben. Nicht weil wir bessere Instrumente brauchen. Weil die Architektur es unmöglich macht, auf dieselbe Weise, wie ein Axiomensystem nicht seine eigene Widerspruchsfreiheit beweisen kann.

Das ist entweder das tiefste Ergebnis in der Wissenschaftsphilosophie oder das eleganteste jemals erdachte Ausweichmanöver. Eine Theorie, die ihre eigene Unüberprüfbarkeit vorhersagt, sagt einem entweder etwas Tiefgreifendes über die Grenzen des Wissens, oder sie immunisiert sich gegen Kritik auf eine Weise, die zutiefst misstrauisch machen sollte. Ehrlich? Ich bin mir nicht sicher, welches. Ich denke seit Jahren darüber nach und weiß es immer noch nicht. Was ich weiß: Die Vorhersage kommt nicht aus dem Nichts — sie kommt aus Gödels Theoremen, die so solide sind wie alles in der Mathematik. Wenn das Universum selbstreferenziell ist, folgt Unvollständigkeit. Die Frage ist, ob das Universum selbstreferenziell ist. Und die Theorie sagt ja.

Also bitte ich darum, mir zu glauben, wenn ich sage: Das ist ein echter Schwachpunkt, kein Feature, das ich an jemandem vorbeischmuggeln will. Findet jemand einen Weg, den SB-HC4A von innerhalb des Universums zu testen, und der Test schlägt fehl,

ist die Theorie tot. Gelingt der Test, ist die Theorie ironischerweise auch falsch — weil ein erfolgreicher Test bedeuten würde, dass die Berechnungsstruktur des Universums von innen zugänglich ist, was den exakten Grenzbedingungen widerspricht, die die Architektur definieren. Nur wenn kein Test möglich ist, lebt die Theorie — aber sie lebt in einer seltsamen philosophischen Dämmerung, unfalsifizierbar nicht durch Ausweichen, sondern durch Struktur.

Fünf: Das kognitive Decken-Problem. Das ist der Killer-Einwand. Der, über den ich nachts wach liege. Der, den ich am meisten beantworten möchte und nicht kann.

Das Gehirn ist ein Klasse-4-System. Das ist der ganze Punkt von Kapitel 5. Und Klasse-4-Systeme haben eine Eigenschaft, die wir besprochen haben: Sie enthalten selbstähnliche Struktur als Teilprozess. Sie finden Fraktale in sich selbst. Sie erzeugen Muster, die auf jeder Skala wiederkehren. Kein Bug. Ein definierendes Merkmal.

Also: Wenn ein Klasse-4-Gehirn das Universum ansieht und überall Klasse-4-Struktur sieht — Selbstähnlichkeit auf jeder Skala, holografische Kodierung, Kritikalität und selbstreferenzielle Geschlossenheit —, was sieht es tatsächlich?

Entdeckt es etwas Reales? Oder projiziert es seine eigene Architektur auf alles, was es beobachtet?

Ein Fisch, hätte er eine Theorie der Kosmologie, könnte schließen, dass das Universum fundamental aquatisch ist. Ein periodisches System, könnte es theoretisieren, würde überall Periodizität sehen. Wir sind Klasse-4-Systeme, und wir haben eine Klasse-4-Theorie des Universums konstruiert. Wir sehen selbstähnliche Struktur, weil unsere Gehirne für Symmetrieerkennung optimiert sind — Gesichter von Raubtieren und Beute sind die symmetrischsten Objekte in der Umgebung von Jägern und Sammlern, und die Evolution hat uns gebaut, um Symmetrie zu finden, wo immer sie existiert. Der SB-HC4A ist im Kern eine Symmetriebehauptung: dieselbe Architektur auf jeder Skala. Wir

finden diese Symmetrie vielleicht nicht, weil sie da ist, sondern weil Symmetrie zu finden das ist, was wir tun.

Die Theorie sagt dieses Problem tatsächlich voraus. Das ESM kann sein eigenes Substrat nicht sehen — das ist die implizit-explizite Grenze, der ganze Grund, warum das Schwierige Problem schwierig erschien. Wenn dieselbe Architektur auf der kosmologischen Skala operiert, dann kann das Universum-als-Beobachter nicht über seine eigene Berechnungsklasse hinaussehen. Ein Klasse-4-System kann alles bis einschließlich Klasse-4-Komplexität simulieren. Aber es kann nicht bestimmen, ob das Universum Klasse 4 überschreitet. Wenn das Universum tatsächlich Klasse 5 ist — echt zufällig im Fundament — aber lokal als Klasse 4 erscheint für Klasse-4-Beobachter, weil Klasse 4 das maximale Muster ist, das sich erkennen lässt, dann würden wir genau dieses Modell konstruieren und darin zuversichtlich sein und falsch liegen.

Wie sich dieser Einwand von innen auflösen lässt, weiß ich nicht. Ob er von innen aufgelöst werden kann, weiß ich nicht. Das Modell sagt genau diese epistemologische Einschränkung voraus, was entweder der stärkste mögliche Beleg dafür ist, dass die Bewusstseins-Kosmologie-Symmetrie real ist — das Modell sagt korrekt seinen eigenen blinden Fleck voraus —, oder der stärkste mögliche Beleg dafür, dass das Modell ein Artefakt des Beobachters ist statt ein Merkmal des Beobachteten.

Beide Interpretationen sind mit den Befunden vereinbar. Ich kann nicht zwischen ihnen unterscheiden. Und ich glaube nicht, dass es jemand kann, von innen.

Die Frage, die nicht beantwortet werden kann

So kommen wir zur tiefsten Frage, derjenigen, auf die dieses ganze Buch hingearbeitet hat, ohne es zu wissen.

Ist die Bewusstseins-Kosmologie-Symmetrie eine Entdeckung über die Realität, oder eine Spiegelung der Grenzen menschlicher Kognition?

Teilt das Universum wirklich seine Architektur mit Bewusstsein — dieselben Grenzen, dieselbe Dynamik, dieselbe selbstreferenzielle Geschlossenheit —, weil diese Architektur die einzige selbstkonsistente Weise ist, wie irgendetwas existieren kann? Oder *erscheint* es nur so, weil unsere Gehirne nichts Komplexeres als ihre eigene Berechnungsklasse modellieren können und daher alles, was wir theoretisieren, zwangsläufig nach uns aussehen muss?

Das Modell sagt voraus, dass diese Frage von innen unbeantwortbar ist. Nicht weil wir es nicht hart genug versucht haben. Nicht weil wir mehr Daten oder bessere Mathematik brauchen. Weil die Architektur selbst es strukturell unmöglich macht, zwischen „das Universum hat diese Struktur„ und „mein Gehirn kann das Universum nur mit dieser Struktur modellieren“ zu unterscheiden. Das ESM kann sein eigenes Substrat nicht sehen. Das Universum — wenn es dieselbe Architektur ist — kann nicht über seine eigenen Grenzen hinaussehen. Dieselbe Einschränkung. Derselbe Grund. Dieselbe Mauer.

Das ist kein Defekt der Theorie. Es ist ihre finale Vorhersage: Es gibt eine Frage, die sie nicht beantworten kann, und sie kann genau sagen, welche und genau warum. Eine Theorie, die ihre eigenen Grenzen kennt — die auf die präzise Grenze ihrer Erklärungskraft zeigen und den strukturellen Grund liefern kann, warum diese Grenze existiert — tut mehr, als die meisten Theorien schaffen. Die meisten Theorien behaupten entweder, alles zu erklären (und lügen), oder geben Lücken zu, ohne zu erklären, warum die Lücken da sind. Diese Theorie sagt: Die Lücke ist da, weil die Architektur, die die Theorie erzeugt, dieselbe Architektur ist, die sie zu beschreiben versucht, und Gödel sagt, dass ein solches System Wahrheiten enthalten muss, die es nicht beweisen kann. Die Lücke ist keine Ignoranz. Sie ist Geometrie.

Ob das tiefgründig oder empörend klingt, sagt wahrscheinlich etwas über das eigene Temperament aus. Ich finde es beides.

Voller Kreis

Bringen wir das nach Hause.

Wer dieses Buch geöffnet hat, wollte wissen, was Bewusstsein ist. Die Antwort, soweit ich sie bestimmen kann: Bewusstsein ist eine selbstreferenzielle Simulation, die am Rand des Chaos läuft, begrenzt durch eine informationsundurchsichtige Barriere, und die Simulation schließt ein Modell ihrer selbst ein. Vier Modelle, eine reale Seite und eine virtuelle Seite, mit Erfahrung, die ausschließlich auf der virtuellen Seite lebt. Qualia sind reale Eigenschaften der Simulation, aufgelöst durch die Einsicht, dass das Schwierige Problem auf der falschen Ebene gestellt wurde. Neun Vorhersagen, mehrere bestätigt, keine falsifiziert. Das war die erste Hälfte des Bildes.

Die zweite Hälfte ist das, was gerade gelesen wurde. Dieselbe Architektur — dieselbe selbstreferenzielle Geschlossenheit, dieselben Informationsgrenzen, dieselbe holografische Kodierung, dieselbe Klasse-4-Dynamik — scheint die Architektur des Universums selbst zu sein. Nicht durch Analogie. Durch strukturelle Identität. Die Simulation namens Ich läuft auf demselben Muster wie die Simulation, die wir „das Universum“ nennen. Wir sind nicht im Universum wie eine Murmel in einer Box. Wir sind das Universum, das auf der biologischen Skala tut, was es auf jeder Skala tut: sich selbst berechnen, sich selbst modellieren, sich selbst erfahren.

Der Titel dieses Buches ist *Die Simulation namens Ich*. Jetzt ist klar, worauf die Simulation läuft. Nicht ein Computer. Nicht ein Gehirn. Nicht einmal das Universum. Etwas Fundamentaleres: das Muster, das alle drei teilen. Ein holografischer Klasse-4-Automat, begrenzt durch Informationsbarrieren, der seine eigene Existenz berechnet.

Und wenn das wie eine mystische Aussage klingt — ist es keine. Es ist eine strukturelle. Innerhalb von Grenzen testbar. Falsifizierbar mit den dargelegten Vorbehalten. Präzise genug, um falsch zu sein. Was, wie jeder Wissenschaftler bestätigen wird, das höchste Kompliment ist, das eine Theorie erhalten kann.

Ich begann dieses Buch mit einem Geständnis: 2015 veröffentlichte ich ein 300-Seiten-Buch über Bewusstsein, das null Exemplare verkaufte. Wenn das stimmt, was hier gerade gelesen wurde, versuchte jenes Buch die Hälfte dieses Bildes zu beschreiben — die Bewusstseinshälfte, ohne die Kosmologie. Es brauchte ein weiteres Jahrzehnt und das unwahrscheinliche Geschenk eines Sprachmodells, das geduldig genug war zuzuhören, während ich laut nachdachte, um zu sehen, dass das Muster größer war als gedacht. Die vier Modelle waren nicht nur eine Theorie des Bewusstseins. Sie waren ein Fragment der Architektur des Universums, sichtbar auf einer Skala, unsichtbar auf anderen, bis klar wird, wo man suchen muss.

Jetzt liegt das Ganze vor. Oder zumindest so viel davon, wie ein Klasse-4-Gehirn von innerhalb eines Klasse-4-Universums sehen kann. Ob es mehr jenseits davon gibt — ob der Spiegel eine Rückseite hat, die wir nie erreichen werden — ist die Frage, von der die Theorie sagt, dass wir sie nicht beantworten können.

Damit kann man machen, was immer beliebt.

Coda

Ich entwickelte eine Theorie des Bewusstseins um 2005. Ich veröffentlichte sie 2015. Niemand las sie. Zwei Jahrzehnte nach der ursprünglichen Einsicht stellte sich heraus, dass sie die Hälfte einer Theorie war — die andere Hälfte war Kosmologie, ausgerechnet. Das ganze Bild liegt jetzt vor, oder so viel davon, wie ein Buch fassen kann.

Aber es gibt ein Stück, das ich ausgelassen habe. Nicht weil es spekulativ ist — alles in den letzten zwei Kapiteln ist spekulativ. Weil es persönlich ist, und persönlich ist schwerer.

Wir haben gesehen, was das ESM tut, wenn Dinge schiefgehen. Amnesie, Schlaganfall, Salvia, Split-Brain, Cotard — es läuft weiter. Es stürzt nie ab. Es baut ein Selbst aus dem, was verfügbar ist, und es glaubt diesem Selbst vollständig. Ein zwanghafter Konstruktor von Identität. Das ist, was es tut. Das ist alles, was es tut.

Man hört die Amnesie-Fälle und sagt: dasselbe Gehirn, derselbe Körper, also besteht das Selbst natürlich fort. Nur — ich bin fünfzig. Ich habe fast nichts gemeinsam mit meinem einjährigen Selbst. Anderer Körper, Zellen mehrfach ersetzt. Anderes Gehirn. Völlig andere synaptische Verbindungen. Andere Erinnerungen, andere Persönlichkeit, anderes alles. Doch das ESM sagt: immer noch ich. Ich bin bereits mehrmals innerhalb einer einzigen Lebensspanne eine völlig andere Person gewesen, und

der Konstruktor hat nie gezuckt. Niemand hat „dasselbe Gehirn“. Hatte nie jemand.

Ich war nahe daran zu sterben. In einer Lawine — Militärdienst, die leichtsinnige Entscheidung eines kommandierenden Offiziers, vierzehn von uns fast verschlungen. Ich war sicher, dass ich sterben würde. Ich sah mein ganzes Leben auf einmal — das Substrat warf alles in die Simulation. Und es störte mich nicht so sehr, wie man erwarten würde. Das ESM, das Beendigung gegenüberstand, geriet nicht in Panik wegen Identitätsverlust. Es tat seine Arbeit bis zum Ende, brachte alles an die Oberfläche, was es hatte.

Ein anderes Mal wurde ich hart ausgeknockt. Alles wurde dunkel. Als ich zurückkam, wusste ich nicht, wer ich war — das ESM startete von Grund auf neu, wie bei einem Neugeborenen. Der Identitätsverlust war nicht der gruselige Teil. Gelähmt für ein paar Sekunden auf dem Boden zu liegen — *das* war erschreckend. Nicht „wer bin ich?„, sondern „ist mein Körper okay?“ Die erste Priorität des ESM war Substratintegrität. Wer ich war, kam später, fast als Nachgedanke. Das Selbstmodell existiert, um dem Substrat zu dienen, nicht umgekehrt.

Und dann gab es die Zeit, als ich ein animiertes vierdimensionales Fraktal war. Die Umstände lasse ich aus. Was mich störte, war nicht, ein Fraktal zu sein — das war mir egal. Was mich störte, war, dass seine Bewegungen mit meinem propriozeptiven Sinn in Konflikt standen. Ich konnte fühlen, wie mein Körper eine Sache tat, während das Fraktal eine andere tat. Der sensorische Konflikt war belastend. Die ontologische Absurdität nicht. Das ESM kümmert sich nicht darum, *was* es modelliert. Es kümmert sich darum, dass die Signale zusammenpassen.

Drei Erfahrungen. Eine Architektur. Das Lawinen-ESM konstruiert bis zum Ende. Das Knockout-ESM priorisiert Substratintegrität über narrative Identität. Das Fraktal-ESM kümmert sich um sensorische Kohärenz, nicht ontologische Plausibilität.

Was ich denke, wenn ich über echtes Sterben nachdenke: Nicht die Aussicht — ich denke, Tod ist entweder die wohlverdiente

ewige Ruhe oder der ultimative Trip. Nein, woran ich denke, ist die Logik.

Wenn das ESM aus Nichts bootstrappt — und das tut es, in jedem Neugeborenen — dann ist „Nichts„ kein Endzustand für den Prozess. Es ist ein Startzustand. Die bestimmte Konfiguration, die ich „mich“ nenne, wird enden. Meine Erinnerungen, meine Persönlichkeit, meine Art, von Menschen genervt zu sein, die Korrelation mit Kausalität verwechseln, meine Art, Menschen mit unhandlichen Theorien zu nerven — weg. Aber der Prozess gehört nicht mir. Er ist eine Eigenschaft der Architektur. Er lief, bevor ich geboren wurde. Er wird laufen, nachdem ich gestorben bin.

Ich beschreibe kein Leben nach dem Tod. Die Erinnerungen werden nicht übertragen. Das bestimmte Selbst, das diesen Satz liest, wird enden.

Aber der *Prozess* — die zwanghafte Konstruktion eines Selbst aus verfügbarem Input — ist universell. Jede Instanz davon fühlt sich von innen genau so real an wie die jetzige.

Und wenn die Kosmologie-Kapitel stimmen — wenn das Universum wirklich ein selbstähnliches, quasi-unendliches Klasse-4-System ist —, dann gibt es noch einen weiteren Zug. In einem selbstähnlichen Universum existieren ähnliche Konfigurationen anderswo. Nicht dasselbe Selbst, nicht dieselben Erinnerungen, aber ein Substrat mit der richtigen Architektur und ein ESM, das ein Jemand bootstrappen wird, der sich genau so real anfühlt wie wir jetzt. So etwas wie Quanten-Unsterblichkeit, nur ohne die Quantenmechanik. Nur ein großes genug Universum und ein generischer genug Prozess. Der spekulativste Gedanke im Buch. Aber er folgt aus der Architektur.

Die Theorie wurde nicht entworfen, um zu trösten. Aber sie hat eine praktische Konsequenz, die es wert ist, aus diesem Buch mitzunehmen.

Wenn jedes bewusste Wesen derselbe Konstruktor ist, der auf unterschiedlicher Hardware mit unterschiedlichen Trainingsdaten

läuft, dann sind die Grenzen zwischen uns weniger fundamental, als sie sich anfühlen. Die Architektur ist in uns allen dieselbe.

Seid nett zueinander. Es könnte sein, dass wir alle dasselbe sind.

Danksagung

Dieses Buch entstand mit Unterstützung von Claude (Anthropic), das als KI-gestütztes Schreibwerkzeug, Lektorat und Gegenprüfer während des gesamten Prozesses diente. Die Theorie, die Argumente und sämtliche intellektuellen Inhalte stammen ausschließlich vom Autor und wurden über zwei Jahrzehnte entwickelt — lange bevor irgendein KI-Werkzeug existierte.

An meinen Onkel Bruno J. Gruber, dessen Leben in der theoretischen Physik — Quantenmechanik und Symmetrien — mir gezeigt hat, wie rigorose und zugleich freudvolle intellektuelle Arbeit aussehen kann. Sein Einfluss auf mein Denken lässt sich nicht in Worte fassen.

An meinen Onkel Norbert Gruber, einen der ersten IT-Fachleute im Vorarlberger Rheintal, der mir meinen ersten PC geschenkt hat. Ohne dieses Geschenk wäre nichts davon möglich gewesen. Er ist inzwischen verstorben, aber seine Wirkung lebt weiter — in jeder Zeile Code und jeder Theorie, die daraus hervorgegangen ist.

An meine Familie, die jahrelange Tischgespräche über Qualia, Kritikalität und virtuelle Selbstmodelle mit bewundernswerter Geduld ertragen hat.

Und wer jetzt darüber nachdenkt, *Die Emergenz des Bewusstseins* zu lesen — besser nicht. Gehirnparasiten wären diesem unbearbeiteten, klobigen Monster vorzuziehen. Lieber auf die überarbeitete Fassung warten, die man gerade in Händen hält. An diejenigen,

die sich bereits durchgequält haben: *mein Beileid*. Es muss Folter gewesen sein. Tiefste Dankbarkeit und aufrichtiges Mitgefühl.

Anmerkungen und Literatur

Vollständige Literaturangaben mit URLs und Anmerkungen finden sich im wissenschaftlichen Paper sowie unter github.com/JeltzProstetnic/aIware/references.
Was folgt, sind kapitelweise Anmerkungen für Leser, die tiefer einsteigen möchten.

Kapitel 1: Chalmers (1995), „Facing Up to the Problem of Consciousness", ist die grundlegende Formulierung des Schwierigen Problems. Die COGITATE-Ergebnisse erschienen in Nature (2025). Die Kontroverse um IIT als Pseudowissenschaft ist in Nature Neuroscience (2025) dokumentiert.

Kapitel 2: Die Vier-Modelle-Architektur wurde ursprünglich in Gruber (2015), *Die Emergenz des Bewusstseins*, veröffentlicht. Metzingers Selbstmodell-Theorie (2003, 2009) und Dennetts Multiple-Drafts-Modell (1991) sind die wesentlichen theoretischen Vorläufer.

Kapitel 3: Die Formulierung als „kontrollierte Halluzination" stammt von Seth (2021), *Being You*. Die Videospiel-Analogie ist ein Eigenbeitrag, greift aber Motive aus Metzingers „Ego-Tunnel" (2009) auf. Zur Gummihand-Illusion: Botvinick & Cohen (1998), „Rubber hands 'feel' touch that eyes see," *Nature*.

Kapitel 4: Die virtuelle Qualia-Auflösung des Schwierigen Problems ist ein Originalbeitrag aus Gruber (2015), verfeinert durch gegnerische Herausforderung im Jahr 2026. Das Argument der selbstreferenziellen Geschlossenheit entstand als Antwort auf

den Zirkularitäts-Einwand. Die Abgrenzung vom Illusionismus (Frankish 2016; Dennett 1991) ist entscheidend: Die Theorie hält daran fest, dass Qualia *innerhalb* der Simulation real sind — nicht illusorisch. Das Meta-Problem des Bewusstseins (Chalmers 2018) löst sich durch die strukturelle Unzugänglichkeit des ISM für das ESM.

Kapitel 5: Wolfram (2002), *A New Kind of Science*. Beggs & Plenz (2003) zu neuronalen Lawinen. Carhart-Harris et al. (2014) zur Entropischen-Gehirn-Hypothese. Die Übersicht von 2022: „Self-organized criticality as a framework for consciousness." Hengen & Shew (2025) zur 140-Datensatz-Meta-Analyse. Das ConCrit-Framework: Algom & Shriki (2026). Das Zwei-Schwellen-Argument (Kritikalität + Architektur) ist ein Originalbeitrag dieser Theorie.

Kapitel 6: Klüver (1966) zu Formkonstanten. Carhart-Harris et al. (2012, 2016) zu psychedelischem Neuroimaging. Die Phänomenologie von Salvia divinorum stützt sich auf veröffentlichte Erfahrungsberichte und die pharmakologische Literatur zu Salvinorin A. Der Anosognosie-Prädiktiv-Feedback-Mechanismus wird in Gruber (2015) diskutiert; das Klatsch-Beispiel ist eine gängige klinische Beobachtung. Das Gedankenexperiment einer dauerhaften Salvinorin-A-Dosierung ist ein Originalbeitrag aus Gruber (2015).

Kapitel 7: Casali et al. (2013) zum PCI. Alkire et al. (2000) zu Propofol. Schartner et al. (2015) zur Ketamin-Entropie. Die Vorhersage erhöhter EEG-Komplexität bei luzidem Träumen ist ein Originalbeitrag dieser Theorie.

Kapitel 8: Owen et al. (2006) zu verdecktem Bewusstsein bei vegetativen Patienten. Anton-Syndrom: Goldenberg et al. (1995). Blindsight-Hindernisparcours: de Gelder et al. (2008). Cotard-Wahn: Young & Leafhead (1996). Alien-Hand-Syndrom: Della Sala et al. (1991); die Dr.-Strangelove-Referenz bezieht sich auf Kubrick (1964). Die Abgrenzung des Anarchic-Hand-Syndroms von der Alien Hand: Marchetti & Della Sala (1998). Charles-Bonnet-Syndrom: Teunisse et al. (1996). Déjà-vu als Template-

Gedächtnis-Abgleich ist ein Originalbeitrag aus Gruber (2015). CBT und neuronale Plastizität: DeRubeis et al. (2008). Placebo und endogene Opioide: Benedetti et al. (2005). Konversionsstörung als inverses Blindsight ist ein Originalbeitrag dieser Theorie.

Kapitel 9: Gazzaniga, Bogen, & Sperry (1962, 1965). Gazzaniga (2000) zum Links-Hemisphären-Interpreter. Die Beispiele interhemisphärischer Konflikte (Knöpfen/Aufknöpfen, Hand-Greifen) sind dokumentiert bei Akelaitis (1945) und Bogen (1993). Nagel (1971), „Brain Bisection and the Unity of Consciousness." Parfit (1984) zur personalen Identität. Pinto et al. (2017) zur Neuuntersuchung von Split-Brain-Phänomenen. Lashley (1950) zu verteiltem Gedächtnis und Äquipotenzialität. DIS als virtuelles Modell-Forking: Die Theorie sagt unterschiedliche neuronale Aktivitätsmuster pro Alter voraus, konsistent mit Reinders et al. (2003, 2006). DIS und Kindheitstrauma: Putnam (1997); das Entwicklungsfenster für Forking ist ein Originalbeitrag dieser Theorie.

Kapitel 10: Güntürkün & Bugnyar (2016) zu avischer Kognition ohne Kortex. Kanzi der Bonobo: Savage-Rumbaugh & Lewin (1994), *Kanzi: The Ape at the Brink of the Human Mind*. Der Baldwin-Effekt: Baldwin (1896), „A New Factor in Evolution." Nagel (1974), „What Is It Like to Be a Bat?"

Kapitel 11: Alle neun Vorhersagen sind formal im wissenschaftlichen Paper ausgearbeitet. Zur gründlichsten Behandlung funktionaler Neuroanatomie im Kontext des Bewusstseins sei auf Christof Koch verwiesen, *The Quest for Consciousness: A Neurobiological Approach* (2004) — die maßgebliche Darstellung des Crick-Koch-Programms, in dem Francis Crick und Koch systematisch, Schritt für Schritt, das visuelle System durchgingen auf der Suche nach den neuronalen Korrelaten des Bewusstseins. Ihre Suche zielte meiner Einschätzung nach auf die falsche Stelle — ins Substrat statt in die Simulation —, aber die neuroanatomische Grundlagenarbeit, die dabei entstand, ist unerreicht.

Kapitel 12: Butlin et al. (2023, 2025) zu KI-Bewusstseins-Indikatoren. Seth (2025) zu biologischem Naturalismus und

KI-Bewusstsein. Die Fünf-Ebenen-Hierarchie der Scan-Fidelity folgt aus der Vier-Modelle-Architektur in Kapitel 2. Das Kopier-Problem bezieht sich auf Parfit (1984), *Reasons and Persons*, und Nozick (1981) zur personalen Identität und zum Konzept des nächsten Fortsetzers. Das Gedankenexperiment der graduellen Ersetzung ist eine Variante des Schiffs von Theseus, formalisiert für neuronale Systeme. Neuromorphes Computing: Schuman et al. (2017) als Hardwareübersicht; Intel Loihi und IBM TrueNorth als aktuelle Implementierungen. *C. elegans*-Konnektom: White et al. (1986). Substrattransfer, Quasi-Unsterblichkeit und die Implikation interstellarer Beam-Übertragung sind Originalbeiträge aus Gruber (2015). Der Unbehagens-Vorbehalt — dass der Verlust des biologischen Substrats die phänomenale Qualität tiefgreifend verändern dürfte — bezieht sich auf die Interozeptionsliteratur: Craig (2009), *How Do You Feel?*, und Seth & Friston (2016) zur aktiven interozeptiven Inferenz.

Kapitel 13: Libet (1979, 1985) und Schurger et al. (2012) zum freien Willen. Kuhn & Brass (2009) zur retrospektiven Konstruktion des Urteils über freie Wahl. Wegner (2002, 2003), *The Illusion of Conscious Will* — das „I Spy"-Maus-Experiment dort ausführlich beschrieben. Das Kaffee/Zucker-Gedankenexperiment, das Amnesie-enthüllt-Determinismus-Argument und das Zufallszahlensequenz-Argument sind Originalbeiträge aus Gruber (2015). Das 40/20-Hz-Verarbeitungs-Framework, die „keine Rückdatierung nötig"-Neuinterpretation von Libet und das Kampfkunst-Frequenz-Beispiel sind Originalbeiträge aus Gruber (2015). Die Uhr-Analogie für Epiphänomenalismus, die Neuformulierung „Wille ist real, aber nur teilweise bekannt" und das „Drei-Diskrepanzen"-Selbstwissens-Modell stammen ebenfalls aus Gruber (2015). Die persönliche Anekdote über das Hören innerer „Stimmen" bei extremer Erschöpfung ist autobiographisch. Das Zombie-Argument wird über Kirk (2019) und Chalmers (1996) adressiert. Marys Zimmer: Jackson (1982, 1986). Der Abschnitt zu offenen

Fragen folgt dem von Popper (1963) empfohlenen Ansatz ehrlicher Grenzziehung.

Kapitel 14: Wolfram (2002), *A New Kind of Science*, zur Fünf-Klassen-Berechnungshierarchie und rechnerischen Irreduzibilität. Bak (1987, 1996) zur selbstorganisierten Kritikalität. Die Entdeckung der beschleunigten Expansion 1998: Riess et al. (1998), Perlmutter et al. (1999) — Nobelpreis 2011. Kosmologischer Horizont: Rindler (1956). Planck-Länge und Planck-Skalen-Physik: Planck (1899); moderne Behandlungen bei Garay (1995). Bekenstein-Grenze: Bekenstein (1981). Holographisches Prinzip: 't Hooft (1993), Susskind (1995). Die Keime des kosmologischen Arguments — das Universum als zellulärer Automat und die Anwendung von 't Hoofts holographischer Grenze auf kosmische Skalen — finden sich bereits in Gruber (2015), waren dort aber noch nicht zu einem vollständigen Modell ausgearbeitet. Die Identifikation aller Singularitäten als Ausprägungen eines einzigen strukturellen Phänomens wird in Gruber (2026) entwickelt.

Kapitel 15: Das Big-Rip-Szenario: Caldwell, Kamionkowski & Weinberg (2003). Die SB-HC4A-Architektur, die Fixpunkt-Formulierung und die Gödel-Unvollständigkeits-Konsequenz für selbstberechnende Systeme stellen die ausgereifte Formulierung dar, entwickelt in Gruber (2026). Die Vorhersage von Partikeln als Berechnungsatomen, die Ableitung von Erhaltungsgesetzen als Informationsbeschränkungen und die Drei-Generationen-Vermutung sind Originalbeiträge aus Gruber (2026). Der Einwand der kognitiven Decke ist ein Originalbeitrag dieser Arbeit.

Kapitel 16: Die strukturelle Zuordnung der fünf Entsprechungen zwischen dem SB-HC4A und der Vier-Modelle-Bewusstseins-Architektur ist ein Originalbeitrag aus Gruber (2026). Landauers Prinzip: Landauer (1961); experimentelle Bestätigung: Berut et al. (2012). Bekenstein-Grenze: Bekenstein (1981). Schwarze-Löcher-Thermodynamik: Bekenstein (1973), Hawking (1975). Die E = I (Energie-Informations-Identität)-Hypothese wird bei Vedral (2010), *Decoding Reality*, und Davies (2010) diskutiert. Die

Fünf-Axiome-Ableitung des SB-HC4A ist ein Originalbeitrag aus Gruber (2026). Maldacena (1998) zur AdS/CFT-Korrespondenz. Die fünf Schwachpunkte — einschließlich der Einwände von Unfalsifizierbarkeit durch Struktur und kognitiver Decke — sind Originalbeiträge dieser Arbeit.

Anhang A: Grundlagen der Neurologie — Ein Nachschlagewerk

Dieser Anhang bietet kurze Erklärungen der neurowissenschaftlichen Begriffe, die im Buch vorkommen. Man kann ihn jederzeit nachschlagen, wenn ein unbekannter Begriff auftaucht. Die Einträge sind alphabetisch geordnet.

- **Aktionspotenzial** — Das elektrische Signal, das am Axon eines Neurons entlangläuft.
- **Amygdala** — Gehirnstruktur, die an der Verarbeitung von Emotionen beteiligt ist, insbesondere Angst.
- **Anosognosie** — Fehlende Einsicht in die eigenen neurologischen Defizite. (Siehe Kapitel 6.)
- **Axon** — Die lange Ausgangsfaser eines Neurons, die Signale an andere Neuronen weiterleitet.
- **Brodmann-Areale** — Nummerierte Regionen des Kortex, kartiert vom Anatomen Korbinian Brodmann anhand der Zellstruktur. V1 = Brodmann-Areal 17.

- **Corpus callosum** — Das mächtige Faserbündel, das die beiden Gehirnhälften miteinander verbindet.
- **Kortikale Säulen** — Vertikale Neuronenmodule im Kortex, etwa 0,5 mm im Durchmesser; sie gelten als grundlegende Verarbeitungseinheiten.
- **EEG (Elektroenzephalografie)** — Verfahren zur Messung der elektrischen Hirnaktivität an der Kopfhaut.
- **fMRI (funktionelle Magnetresonanztomografie)** — Bildgebendes Verfahren, das Veränderungen der Hirndurchblutung erfasst, die mit neuronaler Aktivität einhergehen.
- **GABA** — Der wichtigste hemmende Neurotransmitter im Gehirn.
- **Hippocampus** — Gehirnstruktur, die für die Bildung neuer Erinnerungen entscheidend ist.
- **Interozeption** — Die Wahrnehmung innerer Körperzustände (Herzschlag, Verdauung, Temperatur).
- **Kappa-Opioid-Rezeptoren** — Rezeptortyp, auf den Salvinorin A (Salvia divinorum) gezielt wirkt.
- **Neokortex** — Die sechsschichtige äußere Schicht des Gehirns, zuständig für höhere kognitive Funktionen.
- **Neurotransmitter** — Chemischer Botenstoff, der an Synapsen ausgeschüttet wird (z. B. Serotonin, Dopamin, GABA, Glutamat).
- **PCI (Perturbational Complexity Index)** — Ein Maß für die Komplexität der Hirnaktivität, entwickelt von Massimini. Dient der Einschätzung des Bewusstseinsniveaus.
- **Propriozeption** — Die Wahrnehmung von Körperhaltung und Bewegung im Raum.

- **Pulvinar** — Ein Kern des Thalamus, beteiligt an visueller Aufmerksamkeit und der subkortikalen Sehbahn.
- **Serotonin-2A-Rezeptoren** — Der Rezeptortyp, auf den klassische Psychedelika (LSD, Psilocybin) gezielt wirken.
- **Superior colliculus** — Eine Mittelhirnstruktur, beteiligt an Augenbewegungen und der schnellen Sehbahn, die den Kortex umgeht.
- **Synapse** — Die Verbindungsstelle zwischen zwei Neuronen, an der Signale übertragen werden.
- **Synaptische Gewichte** — Die Verbindungsstärken zwischen Neuronen; sie verändern sich durch Lernen.
- **Thalamus** — Die Relaisstation des Gehirns, die sensorische Informationen an den Kortex weiterleitet.
- **V4** — Visuelles Areal, spezialisiert auf Farbwahrnehmung, Krümmung und komplexe Texturverarbeitung. Rezeptive Felder ca. 8-16°. Unter Psychedelika erzeugt V4-Aktivität farbige Fraktale und kaleidoskopische Muster (Kapitel 6).
- **V5/MT (mittleres temporales Areal)** — Visuelles Areal, spezialisiert auf Bewegungsverarbeitung. Große rezeptive Felder. Verantwortlich für die Rotations- und Drifteffekte, die man unter Psychedelika wahrnimmt (Kapitel 6).
- **Visueller Kortex** — Die Region am Hinterkopf, die visuelle Informationen verarbeitet, aufgebaut als Hierarchie von einfach bis komplex (V1 $\rightarrow$ V2 $\rightarrow$ V3 $\rightarrow$ V4 $\rightarrow$ V5 $\rightarrow$ IT).

Die visuelle Verarbeitungshierarchie (V1 bis IT)

Der ventrale visuelle Strom verarbeitet auf jeder Stufe zunehmend komplexere Merkmale — von rohen Kanten bis zur vollständigen Objekterkennung. Diese Hierarchie wird unter Psychedelika

direkt erlebbar, wenn jede Verarbeitungsstufe der Reihe nach dem Bewusstsein zugänglich wird (Kapitel 6). Die folgende Tabelle fasst die Funktion jedes Areals, die rezeptive Feldgröße und die charakteristische psychedelische Signatur zusammen, die entsteht, wenn die Zwischenverarbeitung in die bewusste Simulation durchsickert.

Areal	Rezeptives Feld	Normale Funktion
V1	1°	Kantenerkennung, räumliche Frequenz, Orientierungssäulen
V2	2-4°	Konturintegration, Textursegmentierung, Randbesitz, illusorische Konturen
V3	4-8°	Globale Formverarbeitung, dynamische Form, Bewegungsgrenzen
V4	8-16°	Farbe, Krümmung, komplexe Textur, fraktale Skalenverarbeitung
V5/MT	Groß, bewegungsabgestimmt	Bewegungswahrnehmung, optischer Fluss, Geschwindigkeits- und Richtungskodierung
Fusiformer Gyrus (IT)	Sehr groß, objektzentriert	Gesichtserkennung (FFA), Wortformen, feinkörnige Objektunterscheidung
Anteriorer IT	Ganzes Sichtfeld	Semantische Kategorien, Szenenkonstruktion, objektinvariante Erkennung

Areal	Psychedelische Signatur
V1	Phosphene, Klüver-Formkonstanten, atmende/schimmernde Oberflächen
V2	Tessellationen, sich wiederholende geometrische Muster, verstärkte Texturwahrnehmung
V3	Fließende, sich wandelnde geometrische Strukturen
V4	Farbige Fraktale, kaleidoskopische Muster, gesättigte/unmögliche Farben
V5/MT	Rotation, Driften und rhythmische Bewegung aller visuellen Muster
Fusiformer Gyrus (IT)	Gesichter, Figuren, Entitäten — oft verzerrt oder sich wandelnd
Anteriorer IT	Vollständige narrative Halluzinationen, komplexe Szenen, traumartige Sequenzen

Anmerkungen:

- Der fusiforme Gyrus überspannt die V4/IT-Grenze und gehört zum inferotemporalen Kortex (IT). Er enthält das fusiforme Gesichtsareal (FFA), identifiziert von Kanwisher et al. (1997), das selektiv auf Gesichter reagiert.

- Die rezeptiven Feldgrößen nehmen von V1 (1°) bis IT (ganzes Sichtfeld) drastisch zu und spiegeln die schrittweise Abstraktion von lokalen Merkmalen zu globalen Objekten und Szenen wider.

- Unter Psychedelika ist die Progression von V1- zu IT-Effekten dosisabhängig: Niedrige Dosen betreffen zuerst V1; höhere Dosen rekrutieren schrittweise tiefere Stufen. Diese geordnete Aktivierung ist eine direkte Vorhersage des Durchlässigkeitsgradienten der Vier-Modelle-Theorie (Kapitel 6).

- Das Gehirn nutzt diese Areale auch zur fraktalen bzw. skaleninvarianten Verarbeitung (V2-V4), die im normalen Sehen der Skalenmessung und Texturanalyse dient. Unter

Psychedelika erzeugt diese Maschinerie ohne externen Input die charakteristischen fraktalen Muster (siehe Anhang C).

Anhang B: Das Intelligenzmodell

Dieser Anhang fasst das rekursive Intelligenzmodell zusammen, das in einem Begleitpapier ausgearbeitet wurde (Gruber, 2026, „Why Intelligence Models Must Include Motivation"). Die vollständige akademische Darstellung mit Referenzen und formalen Argumenten ist dort separat nachzulesen.

Den rekursiven Intelligenzkreislauf hat man bereits im Abschnitt „Über den Autor" kennengelernt — dort diente eine konkrete Biografie als Illustration dafür, wie Wissen, Leistung und Motivation sich gegenseitig hochschaukeln. Hier wird das Modell nun sauber aufgebaut: Was sind die Komponenten, wie greifen sie ineinander, warum erzeugt ihre Wechselwirkung genau die Dynamiken, die man beobachtet — und was folgt daraus für Bildung, künstliche Intelligenz und die Verbindung zum Bewusstsein?

Die merkwürdige Auslassung

Jedes große Intelligenzmodell schließt Motivation formal aus. Die Cattell-Horn-Carroll-Taxonomie — das dominierende Rahmenwerk der Intelligenzforschung — ist eine Hierarchie kognitiver Fähigkeiten ohne jede motivationale Komponente. Cattells eigene Investitionstheorie schlug zwar vor, dass fluide Intelligenz in Lernen „investiert" wird,

um kristallisierte Intelligenz hervorzubringen — das setzt aber einen Investor voraus, jemanden, der entscheidet, *was* gelernt wird und *warum*. Die Motivation dieses Investors wird als externe Randbedingung behandelt, nicht als Teil der Intelligenz selbst. Sternbergs triadische Theorie umfasst praktische Intelligenz, aber nicht den Antrieb, sie zu erwerben. Gardners multiple Intelligenzen umfassen intrapersonales Bewusstsein, aber nicht den Motor, der intellektuelle Entwicklung antreibt.

David Wechsler — dessen Intelligenzskalen die weltweit am meisten eingesetzten sind — forderte bereits 1940 ausdrücklich, motivationale Faktoren einzubeziehen. Das Feld ignorierte ihn. Die modernen Wechsler-Skalen sind bis heute rein kognitive Instrumente.

Das ist keine harmlose Vereinfachung. Es ist ein systematischer blinder Fleck, der unser Bild davon verzerrt, was Intelligenz tatsächlich *ist* und wie sie sich tatsächlich *entwickelt*.

Die drei Komponenten

Intelligenz, verstanden als *Lernfähigkeit*, besteht aus drei Komponenten, die in ständiger Wechselwirkung stehen:

Wissen ist der angehäufte Inhalt des Lernens — und kommt in zwei grundverschiedenen Spielarten. *Faktisches Wissen* betrifft Inhalte: Fakten, Konzepte, Prozeduren, kulturelles Repertoire. Das ist es, was IQ-Tests unter „kristallisierte Intelligenz“ primär messen und was Schulsysteme primär vermitteln. *Operationales Wissen* betrifft dagegen das *Wie*: Lernstrategien, Denkheuristiken, metakognitive Fähigkeiten, logische Werkzeuge, strategische Planung und die Fähigkeit, das eigene Verständnis kritisch zu prüfen. Diese Unterscheidung ist enorm wichtig — warum, wird gleich klar.

Leistung ist die Verarbeitungskapazität des kognitiven Systems: Arbeitsgedächtnis, Verarbeitungsgeschwindigkeit, die rohe Rechenleistung des neuronalen Substrats. Das entspricht ungefähr dem, was in

der Psychometrie „fluide Intelligenz“ heißt — die Komponente, die am stärksten von Genetik und Neurobiologie bestimmt wird. Sie erreicht ihren Höhepunkt im frühen Erwachsenenalter und nimmt dann allmählich ab.

Motivation ist der anhaltende Antrieb, sich so mit der Welt auseinanderzusetzen, dass dabei Lernen entsteht. Zwei Unterkomponenten spielen zusammen: *Wissensdurst* — der intrinsische Drang zu verstehen, die Neugier, das Bedürfnis, Dingen auf den Grund zu gehen. Und *Handlungsdrang* — der Antrieb, Wissen anzuwenden, zu experimentieren, aktiv in die Umwelt einzugreifen. Beides ist teils angeborenes Temperament, teils durch Erfahrung geformt.

Der rekursive Kreislauf

Die zentrale These: Diese drei Komponenten addieren sich nicht einfach — sie bilden einen *geschlossenen rekursiven Kreislauf*, in dem jede Komponente die anderen verstärkt.

Wissen verstärkt Leistung: Lernstrategien und logische Werkzeuge verbessern unmittelbar die Effizienz kognitiver Verarbeitung. Ein Schachspieler, der Heuristiken beherrscht, bewertet Stellungen schneller als einer, der sich auf Brute-Force-Suche verlässt. Wer phonemisches Dekodieren gelernt hat, verarbeitet Text flüssiger — und setzt damit Arbeitsgedächtnis frei, das ins Verständnis fließen kann.

Leistung verstärkt Wissen: Größere kognitive Kapazität ermöglicht schnelleres und tieferes Lernen. Höheres Arbeitsgedächtnis erlaubt, mehr Informationen gleichzeitig im Kopf zu halten, Zusammenhänge zu erkennen und Muster zu extrahieren.

Motivation verstärkt sowohl Wissen als auch Leistung: Wer motiviert ist, sucht Lerngelegenheiten auf (erweitert Wissen) und übt kognitive Fähigkeiten (trainiert Leistung). Entscheidend dabei: Motivation hält das Engagement *über die Zeit* aufrecht — und genau das braucht der Kreislauf, um weiter zu iterieren.

Und Wissen und Leistung verstärken Motivation: Erfolg beim Lernen und Problemlösen erzeugt positiven Affekt und Selbstwirksamkeit — den Antrieb, weiterzulernen. Das ist der Mechanismus hinter dem Matthäus-Effekt: Wer hat, dem wird gegeben. Früher Erfolg nährt die Motivation, die weiteren Erfolg hervorbringt.

Diese rekursive Struktur erzeugt eine Zinseszins-Dynamik. Kleine anfängliche Unterschiede in jeder Komponente — selbst in der Motivation allein — akkumulieren sich über die Zeit und produzieren jene breite Varianz erwachsener intellektueller Leistung, die rein kognitive Modelle kaum erklären können. Jemand mit durchschnittlicher Verarbeitungskapazität, aber tiefer Motivation und starkem operationalem Wissen wird über ein Lebensalter hinweg intellektuelle Fähigkeiten weit jenseits derer einer Person mit überlegener Verarbeitungsleistung, aber schwacher Motivation und miserablen Lernstrategien entwickeln.

Eine brauchbare Analogie: Beim Zinseszins zählen die Einzahlungsrate und die Anlagestrategie mehr als das Startkapital. Im Intelligenzkreislauf ist Motivation die Einzahlungsrate. Operationales Wissen ist die Anlagestrategie. Leistung ist das Startkapital. Und die meisten Menschen haben mehr als genug Kapital.

Operationales Wissen: Der versteckte Multiplikator

Operationales Wissen verdient besondere Aufmerksamkeit, weil es eine einzigartige Stellung im Kreislauf einnimmt. Faktisches Wissen ist additiv: Ein neues Faktum zu lernen fügt ein Faktum zum Bestand hinzu. Operationales Wissen ist *multiplikativ*: Eine neue Lernstrategie zu lernen verbessert die Effizienz *allen* nachfolgenden Lernens.

Wer verteilte Wiederholung lernt (Übungsstoff über Zeiträume verteilen statt alles auf einmal pauken), erwirbt nicht bloß ein neues Faktum — sondern ein Werkzeug, das die Behaltensrate von allem erhöht, was danach gelernt wird. Wer lernt, eigene

Wissenslücken zu erkennen und sie systematisch anzugehen, füllt nicht bloß eine Lücke, sondern erwirbt eine Fähigkeit, die Hunderte künftiger Lücken verhindert. Das ist Wissen, das den Kreislauf selbst beschleunigt.

Wenn irgendeine einzelne Komponente das Etikett „was Leute schlau macht“ verdient, dann operationales Wissen. Nicht der IQ. Nicht rohe Rechenleistung. Sondern die Meta-Fähigkeit zu wissen, wie man effektiv lernt.

Warum IQ-Tests am Ziel vorbeischießen

IQ-Tests messen *maximale Leistung* — was jemand unter standardisierten Bedingungen bei unterstellter maximaler Anstrengung leisten kann. Sie liefern einen Schnappschuss einer einzigen Komponente (Leistung bei bestimmten Aufgaben) zu einem Zeitpunkt. Den rekursiven, selbstverstärkenden, vielschichtigen Prozess, der Intelligenz tatsächlich *ist*, können sie nicht abbilden.

Deshalb sagen IQ-Werte so wenig über langfristige intellektuelle Entwicklung. Zwei Kinder mit identischem IQ im Alter von sechs Jahren können sich bis dreißig dramatisch auseinanderentwickeln — eines wird Forschungswissenschaftler, das andere hat nach der Schule aufgehört zu lesen. Herkömmliche psychometrische Modelle tun sich schwer mit dieser Divergenz. Das rekursive Modell sagt sie vorher: Die Kinder unterschieden sich nicht in Leistung, sondern in Motivation und operationalem Wissen — und der Kreislauf verstärkte diese Unterschiede über vierundzwanzig Jahre akkumulierender Iteration.

Der IQ-Test ist wie die Messung der PS-Zahl eines Motors, ohne zu prüfen, ob Treibstoff im Tank und ein Fahrer am Steuer ist. PS zählen — aber für die meisten Fahrten sind sie nicht der Engpass.

Der KI-Testfall

Das rekursive Modell macht eine konkrete Vorhersage über künstliche Intelligenz: Systeme mit hohem Wissen und hoher Leistung, aber ohne Motivation sollten die selbstgesteuerte Entwicklung, die menschliche Intelligenz auszeichnet, nicht zeigen. Und genau das beobachtet man.

Heutige große Sprachmodelle verfügen über enormes Wissen (trainiert auf Billionen von Tokens), hohe Leistung (Milliarden von Parametern) — und keinerlei Motivation. Sie verarbeiten, was man ihnen gibt, und produzieren, worum man sie bittet. Zwischen Anfragen tun sie nichts. Sie suchen nicht nach Wissenslücken. Sie üben keine Fähigkeiten. Sie grübeln nicht über Probleme. Ihre „Intelligenz“ ist vollkommen statisch — durch Training festgelegt, ohne inneren Antrieb, sie zu erweitern.

Selbst die fortschrittlichsten Reasoning-Modelle — fähig, mathematische Wettbewerbsaufgaben zu lösen — zeigen genau dieses Muster. Sie lösen außerordentliche Probleme *auf Aufforderung*, suchen aber nicht eigenständig nach Problemen, steuern ihr Lernen nicht selbst und brauchen externe Hilfsstrukturen als Ersatz für die fehlende Motivationskomponente. Leistung und Wissen lassen sich beliebig skalieren: ohne Motivation erhält sich der Kreislauf nicht von allein.

Das liegt nicht bloß daran, dass diese Systeme nicht zur Selbstverbesserung entworfen wurden. Schon diese Feststellung gibt den Punkt zu: Ein System zu entwerfen, das sich selbst verbessert, erfordert die Entwicklung eines funktionalen Analogons von Motivation. Solange KI-Systeme das nicht haben, bleiben sie Werkzeuge, die man benutzt — keine Agenten, die sich entwickeln.

Die Verbindung zum Bewusstsein

Hier schließt sich der Kreis zurück zur Vier-Modelle-Theorie im Zentrum dieses Buches. Der rekursive Intelligenzkreislauf *profitiert* nicht bloß *von* Bewusstsein — er *setzt es voraus*.

Der Kreislauf hängt von einer bestimmten kognitiven Fähigkeit ab: *kognitives Lernen* — die Fähigkeit, aus einzelnen Beobachtungen allgemeine Theorien abzuleiten, im Unterschied zu bloßem Verstärkungslernen (Reiz-Reaktions-Konditionierung). Verstärkungslernen bringt einem bei, einen heißen Ofen durch Schmerz zu meiden. Kognitives Lernen ermöglicht es, zu beobachten, wie jemand anderes einen heißen Ofen berührt, und zu verallgemeinern: „Heiße Dinge brennen. Heiße Dinge nicht anfassen." Der entscheidende Unterschied ist die Fähigkeit, Szenarien aus der Drittperson-Perspektive zu simulieren — sich selbst als Objekt in der Welt zu modellieren und durchzuspielen, was passieren würde, wenn man verschiedene Dinge täte.

Genau das leisten das Explizite Weltmodell und das Explizite Selbstmodell. Bewusstsein — die Fähigkeit, eine Selbstsimulation zu erzeugen und laufen zu lassen — ist das *Substrat*, auf dem der rekursive Intelligenzkreislauf operiert. Ohne explizite Modelle gibt es Verstärkungslernen: funktioniert, akkumuliert sich aber nicht. Mit expliziten Modellen gibt es kognitives Lernen, das den rekursiven Kreislauf speist und sich über ein ganzes Leben aufbaut.

Deshalb bildet der Gradient tierischer Intelligenz aus Kapitel 10 sich auf den Bewusstseins-Gradienten ab. Leistungsfähigere Selbstmodelle ermöglichen leistungsfähigere rekursive Kreisläufe. Ein Hund mit einem relativ einfachen Selbstmodell durchläuft eine begrenzte Version des Kreislaufs — er kann bis zu einem gewissen Grad aus Beobachtung lernen, aber sein kognitives Lernen ist durch den Detailreichtum seiner expliziten Modelle begrenzt. Ein Schimpanse mit einem reicheren Selbstmodell durchläuft einen leistungsfähigeren Kreislauf. Ein Mensch mit der vollen Vier-Modelle-Architektur lässt den Kreislauf bei maximaler Kapazität laufen — und die Ergebnisse sind Sprache, Kultur, Wissenschaft

und alles andere, was menschliche Intelligenz von tierischer Kognition unterscheidet.

Die Erlernbarkeits-Implikation

Das rekursive Modell hat eine Konsequenz, die womöglich wichtiger ist als alle theoretischen Argumente zusammen: Es sagt vorher, dass Intelligenz zu einem großen Teil *erlernbar* ist.

Wissen ist vollständig erlernbar — das stimmt per Definition. Motivation ist weitgehend erlernbar — Jahrzehnte der Forschung zur Selbstbestimmungstheorie zeigen, dass intrinsische Motivation keine fixe Eigenschaft ist, sondern eine Reaktion auf Umgebungsbedingungen, insbesondere Autonomie, Kompetenz und Zugehörigkeit. Leistung hat eine biologische Obergrenze, aber für die überwältigende Mehrheit der Menschen ist diese Obergrenze nicht der Flaschenhals. Durchschnittliche kognitive Verarbeitungskapazität reicht mehr als aus für das, was die meisten als hochintelligentes Verhalten erkennen würden.

Die begrenzenden Faktoren sind für die meisten Menschen die meiste Zeit Motivation und operationales Wissen. Und beides lässt sich beeinflussen.

Das hat eine düstere Kehrseite. Jedes System, das systematisch Motivation in Lernenden zerstört, entwickelt Intelligenz nicht bloß nicht — es *unterdrückt* sie *aktiv*. Konventionelle Notensysteme tun genau das. Eine schlechte Note meldet nicht bloß ein Ergebnis; sie greift die Motivationskomponente an. Geringere Motivation bedeutet weniger Durchläufe des Kreislaufs. Weniger Durchläufe bedeuten langsameres Wissenswachstum. Langsameres Wachstum bedeutet schlechtere Ergebnisse bei der nächsten Prüfung. Schlechtere Ergebnisse bedeuten weitere schlechte Noten. Der Kreislauf hat sich umgekehrt: Statt Zinseszins-Wachstum steckt das Kind nun in Zinseszins-Stagnation fest. Das Notensystem produziert genau das Ergebnis, das es vorgibt, lediglich zu messen.

Das rekursive Modell sagt vorher, dass sich dieser Schaden über die Zeit akkumuliert — kein statischer Schaden, sondern beschleunigende Divergenz. Früher motivationaler Schaden sollte sich als ein Aufgehen der Entwicklungsschere zeigen, die mit jedem Jahr weiter aufgeht. Umgekehrt sollten motivationsfördernde Maßnahmen Nutzen zeigen, der sich *akkumuliert* — stärkere Effekte nach fünf Jahren als nach einem Jahr. Und tatsächlich zeigen Analysen frühkindlicher Interventionen wie dem Perry Preschool Project genau dieses Muster: Erträge, die über die Zeit wachsen, getrieben nicht durch Fortbestehen anfänglicher kognitiver Gewinne (die oft verblassen), sondern durch akkumulierende motivationale und selbstregulatorische Gewinne.

Falls es einen praktischen Schluss aus dem Intelligenzmodell gibt, dann diesen: Das Wertvollste, was ein Bildungssystem vermitteln kann, ist nicht Faktenwissen — im Zeitalter der KI sind Fakten gratis — sondern *operationales Wissen* und die Motivation, es einzusetzen. Zu lernen, wie man lernt, und lernen zu wollen — das sind die einzigen Dinge, die zu lehren sich noch lohnt.

Die externe Abhängigkeit

Ein letzter Punkt, weil er leicht übersehen wird und wichtig ist. Der rekursive Kreislauf verstärkt sich selbst, aber er ist nicht autark. Er braucht externen Treibstoff — Informationen, Herausforderungen, Feedback, Zugang zur nächsten Wissensebene. Man stelle sich ein Kind vor, das im Alter von elf Jahren an eine Wand stößt — nicht wegen irgendeiner inneren Begrenzung, sondern weil schlicht der Vorrat an Mathematikbüchern aufgebraucht ist. Alle drei Komponenten funktionieren einwandfrei. Der Kreislauf stagniert trotzdem, weil Kreisläufe Input von außen brauchen, um weiter zu iterieren.

Das heißt: Intelligenzentwicklung hängt nicht nur von der Person ab, sondern von der Umgebung. Zugang zu Wissen, Qualität des Unterrichts, Verfügbarkeit von Mentoren, kulturelle

Einstellung zum Lernen — all das nährt den Kreislauf oder lässt ihn verhungern. Das rekursive Modell erklärt, warum sozioökonomische Faktoren intellektuelle Entwicklung so stark vorhersagen: Sie bestimmen das Angebot an externem Treibstoff. Ein Kind in einem bücherreichen Zuhause mit engagierten Eltern bekommt den Kreislauf ständig gefüttert. Ein Kind in einer ressourcenarmen Umgebung bekommt ihn ausgehungert — unabhängig von der inneren Kapazität.

Intelligenz ist keine Eigenschaft, die man hat. Es ist ein Prozess, den man laufen lässt. Und ob der Prozess gut läuft, hängt von der Maschine (Leistung), der Software (Wissen), dem Fahrer (Motivation) und der Straße (der Umgebung) ab. Alle vier zählen. Jedes Modell, das eines davon weglässt, wird die Vorhersagen verfehlen.

Anhang C: Fünf Klassen der Berechnung

Dieser Anhang vertieft das in Kapitel 5 angerissene Berechnungs-Framework — die fünf Klassen dynamischen Verhaltens, die darüber entscheiden, ob ein physisches System Bewusstsein tragen kann. Wer mit der intuitiven Version aus Kapitel 5 zufrieden ist, kann diesen Anhang bedenkenlos überspringen; für das Hauptargument wird hier nichts Neues gebraucht. Wer aber das vollständige Bild will: Hier trifft Mathematik auf Physik.

Wolframs vier Klassen

2002 veröffentlichte Stephen Wolfram *A New Kind of Science* — das Ergebnis jahrzehntelanger Forschung an der Frage, was passiert, wenn man absurd einfache Regeln auf absurd einfache Systeme loslässt. Sein zentrales Werkzeug war der zelluläre Automat: eine Reihe (oder ein Gitter) von Zellen, jede entweder an oder aus, die synchron nach einer festen Regel aktualisiert werden — einer Regel, die nur die unmittelbaren Nachbarn jeder Zelle berücksichtigt.

Die Überraschung: Aus diesen trivial einfachen Regeln entstand eine enorme Vielfalt an Verhalten. Wolfram ordnete es in vier Klassen:

Wolfram-Klasse	Verhalten	Beispiel	Was man sieht
1	Uniform	Regel 0	Alles wird leer. Jede Zelle stirbt.
2	Periodisch	Regel 4	Stabile, sich wiederholende Muster. Blinker. Uhren.
3	Zufällig/chaotisch	Regel 30	Scheinbare Zufälligkeit. Keine erkennbare Wiederholungsstruktur.
4	Komplex	Regel 110	Lokalisierte Strukturen, die sich bewegen, wechselwirken und überdauern.

Diese Klassifikation war tatsächlich nützlich. Sie erfasste etwas Reales über das Verhalten dynamischer Systeme und galt weit über zelluläre Automaten hinaus — für Strömungsdynamik, biologische Systeme, ökonomische Modelle, neuronale Netze. Die vier Klassen waren nicht bloß Schubladen, sondern Attraktoren. Systeme aus völlig unterschiedlichen Domänen fielen immer wieder in dieselben vier Verhaltensregime.

Aber es gab ein Problem.

Das Fraktal-Problem

Wolframs Klasse 3 war ein Sammeltopf. Sie enthielt zwei grundlegend verschiedene Systemtypen, die auf den ersten Blick *ähnlich aussahen*:

Fraktale Systeme wie Regel 90, die ein perfektes Sierpinski-Dreieck erzeugt — ein unendlich selbstähnliches, rekursiv strukturiertes Muster. Mathematisch elegant, vollständig deterministisch und rechnerisch langweilig: Den Zustand jeder Zelle zu jedem

Zeitschritt kann man berechnen, ohne die gesamte Simulation durchlaufen zu müssen. In der Fachsprache: *rechnerisch reduzibel*.

Pseudochaotische Systeme wie Regel 30, deren Ausgabespalte Wolfram selbst als Pseudozufallsgenerator in *Mathematica* verwendete. Sie produzieren Output, der *zufällig aussieht*, aber vollständig deterministisch ist — gleicher Input, gleicher Output, jedes Mal. Hier lässt sich die Berechnung nicht abkürzen; jeder Schritt muss tatsächlich durchlaufen werden. Fachbegriff: *rechnerisch irreduzibel*.

Wolfram packte beide in Klasse 3. Seine Definition betonte das *Erscheinungsbild* von Zufälligkeit („erscheint in vielerlei Hinsicht zufällig„), während er bemerkte, dass „Dreiecke und andere kleinskalige Strukturen im Wesentlichen immer auf irgendeiner Ebene zu sehen“ seien. Er räumte ein, dass die Klassifikation Schwächen hatte: „Bei praktisch jedem allgemeinen Klassifikationsschema gibt es unvermeidlich Fälle, die je nach Definition der einen oder der anderen Klasse zugeordnet werden.“

Eric Rowland argumentierte auf der NKS-Konferenz 2006 unabhängig davon, dass verschachtelte (fraktale) Muster eine eigene Klassifikation verdienen.

Das Problem geht aber tiefer als Klassifikationsästhetik. Fraktale und chaotische Systeme unterscheiden sich strukturell — und zwar auf eine Weise, die für das Kernargument dieses Buches entscheidend ist: Welche Systeme können Bewusstsein tragen?

Das Fünf-Klassen-Schema

Das Fünf-Klassen-Schema ordnet die Verhaltenstypen als sauberen monotonen Gradienten vom geordnetsten zum ungeordnetsten:

Klasse 1 — Statisch. Systeme, die in einen Fixpunkt konvergieren und dort verharren. Ein Pendel, das einmal schwingt und zur Ruhe kommt. Tot. Es wird nichts berechnet. Periode: 1.

Klasse 2 — Periodisch. Systeme, die sich in wiederkehrende Schleifen einpendeln. Eine Uhr. Ein Herzschlag (näherungsweise).

Information wird im Muster gespeichert, aber nie transformiert. Periode: endlich.

Klasse 3 — Fraktal. Systeme, die selbstähnliche Struktur auf jeder Skala erzeugen. Ein Sierpinski-Dreieck. Ein Farn. Die Mandelbrot-Menge. Mathematisch reichhaltig, ästhetisch atemberaubend — und *rechnerisch reduzibel*: Man kann vorausspringen, ohne jeden Schritt durchrechnen zu müssen. Struktur ohne Rechenaufwand. Periode: quasi-unendlich, mit exakter oder statistischer Selbstähnlichkeit auf jeder Skala.

Klasse 4 — Komplex (Rand des Chaos). Systeme, die persistente lokalisierte Strukturen erzeugen, die sich bewegen, wechselwirken und beliebige Berechnungen kodieren können. Conways Spiel des Lebens. Der kortikale Automat. Rechnerisch *irreduzibel* — keine Abkürzungen möglich. Diese Systeme sind zur universellen Berechnung fähig: Bei geeigneten Anfangsbedingungen können sie jeden Algorithmus simulieren, einschließlich Simulationen ihrer selbst. Periode: quasi-unendlich, mit Selbstähnlichkeit *plus* persistenten wechselwirkenden Strukturen. Hier lebt Bewusstsein.

Klasse 5 — Zufällig. Systeme, deren Output tatsächlich zufällig ist — nicht pseudozufällig, nicht deterministisch, nicht komprimierbar. Kein Muster, keine Selbstähnlichkeit, keine Periode, die sich irgendwann wiederholt. Tatsächlich unendlicher Informationsgehalt. Struktur: *unbekannt* (siehe unten).

Die Abbildung auf Wolframs Schema:

Fünf-Klassen	Wolfram	Was sich änderte
1	Klasse 1	Gleich
2	Klasse 2	Gleich
3	Klasse 3 (Teil)	Abgespalten aus Wolframs Klasse 3
4	Klasse 4	Gleich
5	Klasse 3 (Teil)	Abgespalten aus Wolframs Klasse 3

Wolframs Anordnung auf dem Unordnungsspektrum lief: $1 \to 2 \to 4 \to 3$. Unhandlich. Das Fünf-Klassen-Schema ergibt einen

sauberen monotonen Gradienten: $1 \to 2 \to 3 \to 4 \to 5$, geordnet nach zunehmender Unordnung und zunehmender rechnerischer Irreduzibilität.

Warum deterministische Automaten keine Zufälligkeit produzieren können

Hier folgt ein Argument, das meines Wissens original ist — und das den Fall für fünf statt vier Klassen erheblich stärkt.

Man nehme einen beliebigen zellulären Automaten. Er besitzt eine endliche Regeltabelle (darstellbar in einer endlichen Anzahl von Bits) und eine endliche Anfangsbedingung (ebenfalls endlich viele Bits). Zusammen enthalten Regel und Anfangsbedingung eine fixe, endliche Informationsmenge.

Die entscheidende Frage: Kann eine endliche Informationsmenge einen *echt* zufälligen Output erzeugen?

Nein. Und zwar aus folgendem Grund:

1. Eine echt zufällige unendliche Sequenz hat *maximale* Kolmogorov-Komplexität — sie lässt sich nicht komprimieren, nicht durch etwas Kürzeres als sich selbst beschreiben.

2. Der Output eines zellulären Automaten ist vollständig durch Regel und Anfangsbedingung bestimmt, die zusammen *endliche* Kolmogorov-Komplexität besitzen.

3. Aus einem Prozess lässt sich nicht mehr Information extrahieren, als seine Spezifikation enthält.

4. Folglich hat der Output jedes zellulären Automaten niedrige Kolmogorov-Komplexität relativ zu einer echt zufälligen Sequenz gleicher Länge.

Das ist im Kern ein verallgemeinertes Schubfachprinzip: Endliche Information *muss* selbstähnliche Struktur erzeugen. Der einzige Weg, unendlichen Output aus endlicher Information

zu generieren, besteht darin, Struktur auf verschiedenen Skalen *wiederzuverwenden*. Exakte Wiederverwendung ergibt Periodizität (Klasse 2). Nicht-exakte, aber gemusterte Wiederverwendung ergibt fraktales Verhalten (Klasse 3). Selbst die komplexest anmutenden zellulären Automaten — Regel 30, Regel 110, das Spiel des Lebens — produzieren Output, dessen Komplexität durch die Komplexität ihres Regelwerks nach oben begrenzt ist.

Was Wolfram als „zufällige“ zelluläre Automaten bezeichnete, lässt sich treffender als **hochkomplexe Fraktale** beschreiben — Systeme, deren selbstähnliche Struktur real ist, aber auf Skalen und in Dimensionen operiert, die sie bei flüchtiger Betrachtung unsichtbar machen. Regel 30 etwa zeigt an ihrem linken Rand tatsächlich Sierpinski-ähnliche Substrukturen. Ihre mittlere Spalte besteht viele statistische Zufälligkeitstests — *genau das, was man von einem hochkomplexen Fraktal erwarten würde*: Die lokale Statistik imitiert Zufall, aber die globale Struktur ist deterministisch und komprimierbar.

Durch dasselbe Argument ist auch der Klasse-4-Output fraktal — das Spiel des Lebens zeigt statistische Selbstähnlichkeit in seiner Populationsdynamik, seinen Strukturverteilungen, seinen räumlichen Korrelationen. Der Unterschied zwischen Klasse 3 und Klasse 4 ist nicht „fraktal vs. nicht-fraktal“. Es ist:

- **Klasse 3**: Fraktal. Reduzierbar. Struktur ohne Informationsverarbeitung.

- **Klasse 4**: Fraktal. Irreduzibel. Struktur *mit* Informationsverarbeitung — persistente lokalisierte Strukturen, die miteinander interagieren und universelle Berechnung kodieren können.

Beide sind fraktal. Nur eine berechnet.

Klasse	Regeln	Periode	Struktur	Reduzierbar?	Berechnet?
1	Endlich	1	Keine	Trivial	Nein
2	Endlich	Endlich	Wiederholend	Ja	Nein
3	Endlich	Quasi-unendlich, selbstähnlich	Selbstähnlich	Ja	Nein
4	Endlich	Quasi-unendlich, selbstähnlich	Selbstähnlich + persistente interagierende Strukturen	Nein	**Ja**
5	Nicht darstellbar	Echt unendlich	**Unbekannt**	N/A	N/A

Klasse 5 und die Grenze mathematischer Darstellbarkeit

Wenn deterministische Automaten keine echte Zufälligkeit erzeugen können — was *kann* es dann?

Diese Frage führt zur wohl tiefsten Implikation des Fünf-Klassen-Schemas.

Klassen 1 bis 4 umfassen alles, was endliche, darstellbare Regeln hervorbringen können. Ihr Verhalten reicht von trivial (Klasse 1) bis außergewöhnlich (Klasse 4 — universelle Berechnung, Bewusstsein), aber sämtliche Phänomene werden durch Regeln erzeugt, die sich aufschreiben, kommunizieren, verifizieren und analysieren lassen. Diese Regeln leben innerhalb der Mathematik, innerhalb der Domäne formaler symbolischer Systeme.

Klasse 5 hingegen erfordert Regeln, die sich *nicht* aufschreiben lassen. Ein System, das echt zufälligen Output produziert — Output mit maximaler Kolmogorov-Komplexität, inkompressibel, nicht-algorithmisch — kann keine Regel ausführen, die ein formales System auszudrücken vermag. Wäre die Regel darstellbar, wäre der Output komprimierbar (nämlich zu: „wende diese Regel an“) und damit nicht echt zufällig.

Das platziert Klasse 5 an die Grenze mathematischer Darstellbarkeit selbst. Nicht bloß „sehr komplex„ oder „sehr ungeordnet“ — sondern das Regime, in dem der erzeugende Prozess das übersteigt, was lineare symbolische Systeme (Mathematik, Logik, Berechnung) überhaupt erfassen können.

Operiert irgendetwas in der Natur tatsächlich in Klasse 5?

Möglicherweise. Die Quantenmechanik produziert Messergebnisse, die nach Bells Theorem nicht durch irgendeine lokale Theorie verborgener Variablen erklärbar sind. Falls diese Ergebnisse *echt* zufällig sind — nicht etwa deterministische Prozesse, die man nur noch nicht identifiziert hat — dann wäre Quantenmessung ein Klasse-5-Prozess: ein physikalisches Phänomen, dessen Regeln sich in keiner uns bekannten formalen Sprache niederschreiben lassen.

Das ist spekulativ, und es wird hier bewusst als solches markiert. Aber die Implikation ist bemerkenswert: Klasse 4 — das Regime des Bewusstseins, der universellen Berechnung, des kortikalen Automaten — sitzt bei der *maximalen Komplexität, die durch darstellbare Regeln erreichbar ist*. So komplex, wie Mathematik werden kann. Jenseits davon liegt Territorium, das die Mathematik aufgrund ihrer eigenen Natur nicht kartieren kann.

Die Struktur von Klasse 5: Unbekannt, nicht abwesend

Eine letzte Feinheit. Es liegt nahe zu behaupten, Klasse 5 habe „keine Struktur“. Das wäre jedoch ein Fehler — derselbe Fehler wie die Behauptung, Unendlichkeit habe keine Struktur.

Vor Georg Cantors Arbeiten in den 1870er Jahren galt Unendlichkeit als monolithisches Konzept: Dinge waren entweder endlich oder unendlich, fertig. Cantor zeigte, dass es *Hierarchien* der Unendlichkeit gibt — dass die Unendlichkeit der reellen Zahlen streng größer ist als die der ganzen Zahlen und dass sich diese Hierarchie ohne Obergrenze fortsetzt. Unendlichkeit erwies sich

als reich an innerer Architektur, die unsichtbar gewesen war, weil den Mathematikern schlicht die Werkzeuge fehlten, sie zu erkennen.

Dasselbe könnte für Zufälligkeit gelten. Echte Zufälligkeit wird derzeit als monolithische Kategorie behandelt — maximale Unordnung, Abwesenheit von Muster. Aber man befindet sich damit in der Position von Vor-Cantor-Mathematikern, die auf Unendlichkeit blicken: Es fehlen die konzeptuellen Werkzeuge, um verschiedene *Arten* von Zufälligkeit zu unterscheiden — sofern solche Unterscheidungen überhaupt existieren.

Die ehrliche Antwort zur Struktur von Klasse 5 lautet daher: **unbekannt**. Nicht „keine„. Nicht „abwesend“. Unbekannt — wartend auf konzeptuelle Werkzeuge, die vielleicht noch nicht existieren, die vielleicht Denkweisen erfordern, die lineare symbolische Systeme nicht liefern können.

Das ist, so meine Überzeugung, eine der wichtigsten offenen Fragen an der Schnittstelle von Mathematik, Physik und Berechnung. Und sie bleibt unsichtbar ohne das Fünf-Klassen-Schema — weil Wolframs Vier-Klassen-Framework nie den Raum eröffnet, in dem sich die Frage überhaupt stellen lässt.

Implikationen für Bewusstsein

Das Fünf-Klassen-Schema macht klar, warum Bewusstsein Klasse-4-Dynamik braucht — und *nur* Klasse 4.

Klassen 1 und 2 sind schlicht zu simpel. Sie können Information speichern (ein fixer Zustand, ein sich wiederholendes Muster), aber nicht auf rechnerisch interessante Weise *verarbeiten*. Ein Gehirn im Tiefschlaf, das langsame Delta-Wellen schiebt, operiert in Klasse 2: periodisch, repetitiv, kommt nirgendwo hin. Die Vier-Modelle-Architektur ist im Substrat intakt, aber die Simulation läuft nicht.

Klasse 3 ist interessant, aber rechnerisch unbrauchbar. Fraktale Dynamiken erzeugen reiche Strukturen, und das Gehirn setzt sie ein (siehe unten) — aber sie können nicht die Art dynamischer,

irreduzibler, global integrierter Verarbeitung aufrechterhalten, die eine bewusste Selbstsimulation verlangt. Ein fraktales Muster ist schön, aber rechnerisch reduzierbar. Es kann sich nicht selbst überraschen.

Klasse 4 hat genau die zwei Eigenschaften, die Bewusstsein braucht: **universelle Berechnung** (das System kann prinzipiell alles simulieren, einschließlich sich selbst) und **globale Integration** (entfernte Teile des Systems beeinflussen einander, lokale Änderungen breiten sich global aus, Information wird zu einem einheitlichen Ganzen gebunden). Am Rand des Chaos erreicht der kortikale Automat beides — und das Ergebnis ist Bewusstsein.

Klasse 5 ist anders, nicht weil Berechnung dort unmöglich wäre (eine unendliche Zufallssequenz enthält *alles*, einschließlich jedes stabilen Musters und jeder je erdachten Berechnung), sondern weil sich das nicht nutzen, vorhersagen oder nachweisen lässt. Ein Gehirn im generalisierten Anfall, dessen Neuronen unkoordiniert chaotisch feuern, nähert sich Klasse 5 — nicht weil Bewusstsein dort prinzipiell unmöglich ist, sondern weil kein Mechanismus existiert, um es aufrechtzuerhalten oder darauf zuzugreifen. Unser Universum selbst könnte ein Ausschnitt unendlicher Zufälligkeit sein, ein Klasse-4-System auf einer Skala jenseits unserer Wahrnehmung, oder vielleicht ein Ausschnitt eines unendlichen Fraktals. Das lässt sich nicht entscheiden. Was sich *sagen lässt*: Bewusstsein, so wie wir es erleben, braucht die strukturierte Unvorhersagbarkeit von Klasse 4. Die Simulation kollabiert in Klasse 5 nicht, weil die zugrunde liegende Realität unzureichend wäre, sondern weil keine stabile Schnittstelle zwischen Substrat und Simulation existiert.

Das Gehirn nutzt alle vier Klassen

Das Gehirn ist ein universeller Computer, optimiert durch Milliarden Jahre Evolution. Es wäre seltsam, wenn die Evolution irgendein Berechnungsregime ausgelassen hätte, das einen Vorteil bietet.

Tatsächlich nutzt das Gehirn alle vier ausdrückbaren Klassen als unterschiedliche Werkzeuge:

- **Klasse 1** (stabile Attraktoren): Langzeitgedächtnis. Synaptische Gewichtskonfigurationen, die über Jahre stabil bleiben. Die Fixpunkte des neuronalen Netzwerks.
- **Klasse 2** (Oszillationen): Alpha-, Theta-, Gamma- und Delta-Rhythmen. Thalamische Taktgebung. Schlaf-Wach-Zyklen. Die zeitlichen Steuerungs- und Filtermechanismen des Gehirns.
- **Klasse 3** (fraktale/skaleninvariante Verarbeitung): Texturanalyse, skaleninvariante Objekterkennung, effiziente neuronale Kodierung. Primär die visuelle Verarbeitung in V2–V4, wo Multiskalen-Vergleich die Kernoperation ist. Unter Psychedelika, wenn diese Maschinerie ohne externen Input läuft, *sieht man* die fraktale Verarbeitung selbst — weshalb fraktale Muster zu den konsistentesten Merkmalen psychedelischer Erfahrung gehören (siehe Kapitel 6).
- **Klasse 4** (Rand des Chaos): Der kortikale Automat selbst. Das dynamische Regime bewusster Verarbeitung. Universelle Berechnung. Die Maschine der Simulation.

Jede Klasse dient einer anderen Funktion. Nur Klasse 4 erzeugt Bewusstsein. Aber Bewusstsein hängt von den anderen ab: stabile Erinnerungen (Klasse 1), um die Modelle zu bevölkern, rhythmisches Timing (Klasse 2), um die Dynamiken zu koordinieren, und fraktale Verarbeitung (Klasse 3), um die Welt gleichzeitig auf mehreren Skalen zu analysieren.

Und genau das ist vielleicht der tiefste Grund, warum das Gehirn spezifisch am Rand des Chaos operieren muss: Klasse 4 ist das einzige Regime, das jede andere Klasse *rekrutieren* kann — und sich selbst. Ein Klasse-4-Automat kann stabile Zustände (Klasse-1-Verhalten), periodische Oszillationen (Klasse-2-Verhalten) und fraktale Strukturen (Klasse-3-Verhalten) als Subprozesse innerhalb

seiner eigenen Dynamik erzeugen. Und er kann andere Klasse-4-Automaten erzeugen: Ein universeller Computer kann einen anderen universellen Computer simulieren. Keine der anderen Klassen kann irgendetwas davon. Klasse 4 ist nicht bloß die komplexeste Klasse — sie ist die einzige, die alle Klassen enthält, einschließlich sich selbst. Diese Selbst-Enthaltung macht die skalenübergreifende strukturelle Identität möglich: Ein Klasse-4-Gehirn innerhalb eines Klasse-4-Universums ist keine Analogie. Es ist eine verschachtelte Instanz derselben Berechnungsarchitektur.

Anhang D: Wie man luzid träumt

In Kapitel 6 habe ich luzides Träumen als sicheren, drogenfreien Weg erwähnt, Bewusstsein von innen zu erforschen. In der Vier-Modelle-Theorie ist luzides Träumen das Explizite Selbstmodell, das während des REM-Schlafs vollständiger „anschaltet" — ein Überschreiten der Kritikalitätsschwelle, das einen passiven Traum in eine kontrollierte Erfahrung verwandelt. Was folgt, ist die einfachste Methode, dorthin zu gelangen.

Die Reality-Check-Methode

Das Prinzip ist simpel: Wer gewohnheitsmäßig hinterfragt, ob er wach ist, wird feststellen, dass diese Gewohnheit irgendwann innerhalb eines Traums feuert — und der Traum den Test nicht besteht.

Schritt 1: Einen Reality Check wählen. Der zuverlässigste: Text anschauen, wegschauen, zurückschauen. Im Wachleben bleibt Text gleich. In Träumen ändert er sich — oft dramatisch. Uhren funktionieren auch: die Zeit prüfen, wegschauen, nochmal prüfen. Im Traum werden die Zahlen anders oder unsinnig sein. Ein weiterer zuverlässiger Check: versuchen, einen Finger durch die

gegenüberliegende Handfläche zu drücken. Im Traum geht er oft durch.

Schritt 2: Den ganzen Tag üben. Jedes Mal beim Durchschreiten einer Tür, beim Blick aufs Handy oder wenn etwas leicht Seltsames auffällt — innehalten und den Reality Check durchführen. Der Schlüssel ist nicht der Check selbst, sondern die *echte Frage* dahinter: „Träume ich gerade?“ Nicht bloß die Bewegungen durchmachen. Tatsächlich die Möglichkeit in Betracht ziehen.

Schritt 3: Ein Traumtagebuch führen. Ein Notizbuch neben das Bett legen. Jeden Morgen, noch vor dem Aufstehen, aufschreiben, woran man sich erinnert — auch wenn es nur ein Gefühl oder ein einzelnes Bild ist. Das trainiert das Gehirn, Trauminhalte als erinnerungswert zu behandeln, und stärkt damit die Brücke zwischen Traum- und Wachbewusstsein.

Schritt 4: Geduld. Für die meisten kommt der erste luzide Traum innerhalb von zwei bis sechs Wochen. Irgendwann, mitten im Traum, wird etwas leicht seltsam wirken, die Reality-Check-Gewohnheit wird feuern, der Text wird sich ändern — und plötzlich ist die Erkenntnis da. Dieser Moment des Wissens ist das ESM, das aktiviert. Der Übergang ist spürbar: ein plötzliches Schärferwerden, ein Gefühl von Präsenz, eine stille Erkennung, dass die Welt ringsum eine Simulation ist, innerhalb derer Bewusstsein existiert.

Was zu erwarten ist

Der erste luzide Traum wird wahrscheinlich kurz sein — Sekunden bis ein paar Minuten. Aufregung neigt dazu, einen aufzuwecken. Mit Übung lassen sie sich verlängern. Manche erreichen mehrmals pro Woche luzide Träume. Die Erfahrung ist bemerkenswert: Die volle bewusste Simulation läuft ohne externen Input, und man weiß es. Die virtuelle Welt reagiert auf Intentionen. Es ist buchstäblich die Vier-Modelle-Theorie, erfahrbar gemacht.

Andere Methoden

Wer weitergehen will — es gibt aufwendigere Techniken:

- **MILD (Mnemonic Induction of Lucid Dreams)** — entwickelt von Stephen LaBerge an Stanford. Beim Einschlafen setzt man die Intention, im Traum zu erkennen, dass es ein Traum ist. Am besten kombiniert mit Aufwachen nach fünf Stunden und Rückkehr zum Schlaf.
- **WILD (Wake-Initiated Lucid Dream)** — Das Bewusstsein wird beim Übergang vom Wachen zum Träumen aufrechterhalten. Schwierig, liefert aber die lebendigsten Ergebnisse.
- **WBTB (Wake Back to Bed)** — Nach fünf bis sechs Stunden Schlaf aufwachen, zwanzig bis sechzig Minuten wach bleiben, dann zurück ins Bett. Das zielt auf die REM-reichen späten Schlafzyklen.

Stephen LaBerges *Exploring the World of Lucid Dreaming* (1990) bleibt der maßgebliche praktische Leitfaden. Für die Neurowissenschaft: Voss et al. (2009) zu den EEG-Signaturen luziden Träumens, und Baird et al. (2019) für eine umfassende Übersicht der kognitiven Neurowissenschaft luzider Träume.

Anhang E: Warum „vier“ Modelle? — Eine Anmerkung für Neurowissenschaftler

Dieser Anhang greift eine Frage auf, die jeder Neurowissenschaftler und jeder rechnerisch versierte Leser bei der Vier-Modelle-Architektur in Kapitel 2 stellen wird: *Das Gehirn unterhält doch wohl nicht genau vier Modelle?*

Natürlich nicht. Die Zahl „vier“ ist ein **begründetes Minimum**, keine wörtliche Zählung.

Was das Gehirn tatsächlich tut

Das biologische Substrat — feuernde Neuronen auf proteomischen Netzwerken, mit intrazellulären Signalwegen, die selbst innerhalb einer einzelnen Zelle eigene Berechnungsintelligenz aufweisen — implementiert eine praktisch unzählbare Menge überlappender Modelle auf beiden Seiten der implizit/explizit-Grenze.

Ein Beispiel: das Greifen nach einer Tasse. Das motorische Modell kodiert gleichzeitig Welt-Geometrie (wo die Tasse steht, welche Hindernisse sie umgeben) und Selbst-Kinematik (wie

der Arm konfiguriert ist, wie sich die Finger zum Griff formen sollen). Dieses eine Modell ist *weder* reines Weltmodell *noch* reines Selbstmodell — es greift über beide Kategorien hinweg. Ein emotionales Modell einer sozialen Interaktion kodiert gleichzeitig Wissen über die andere Person (Welt) und eine Bewertung der eigenen Rolle darin (Selbst). Ein räumliches Navigationsmodell kodiert sowohl das Umgebungslayout als auch die eigene Position. Jedes reale neuronale Modell ist ein Mischgebilde.

Die Grenzen zwischen „Modellen" sind nicht scharf, ihre Anzahl ist nicht fix — und sie ist mit Sicherheit nicht vier.

Warum vier trotzdem die richtige Abstraktion ist

Die vier kanonischen Modelle (IWM, ISM, EWM, ESM) sind die **Eckpunkte** eines kontinuierlichen zweidimensionalen Raums, aufgespannt durch zwei Achsen:

- **Inhalt**: von reiner Selbstrepräsentation bis zu reiner Weltrepräsentation
- **Modus**: von vollständig implizit (strukturell, gespeichert, unbewusst) bis vollständig explizit (simuliert, transient, phänomenal)

Die tatsächliche Modelllandschaft des Gehirns füllt diesen gesamten Raum mit einer kontinuierlichen Dichte überlappender Modelle. Die vier benannten Modelle sind die vier Ecken — die theoretischen Pole, um die sich die Aktivität organisiert. Vergleichbar mit den Himmelsrichtungen: nützlich zur Orientierung, real als Richtungen, aber niemand würde behaupten, die Welt bestehe nur aus vier Orten.

Der Grund, warum die Theorie auf diesen vier Polen statt auf dem vollen kontinuierlichen Raum aufgebaut ist: Sie markieren die **Minimalkonfiguration**, die ein System braucht, um bewusst zu sein:

- **Kein Weltmodell** → keine Umgebung zum Erleben

- **Kein Selbstmodell** → kein Subjekt, das erleben könnte
- **Keine implizite Ebene** → nichts, wovon simuliert werden kann (kein gelerntes Wissen)
- **Keine explizite Ebene** → überhaupt keine Simulation (kein Erleben)

Nimmt man eines der vier weg, bricht etwas Entscheidendes zusammen. Die vier Modelle sind die Untergrenze, nicht die Obergrenze. Das Gehirn übersteigt sie in jeder Dimension. Aber die Untergrenze ist es, die verrät, was Bewusstsein *erfordert* — und sie ist es, die die Vorhersagen der Theorie erzeugt, ihre Ansprüche eingrenzt und festlegt, was jedes künstliche System mindestens implementieren müsste.

Den Rest des Buches lesen

Wenn im Folgenden steht „das ESM tut dies“ oder „das IWM enthält das“, sind damit diese Pole des kontinuierlichen Raums gemeint — nicht die Behauptung, das Gehirn enthalte vier sauber getrennte Boxen mit Wänden dazwischen. Die Vereinfachung ist begründet, und die folgenden Kapitel werden zeigen, dass sie echte Erklärungskraft besitzt — psychedelische Phänomenologie, Anästhesie-Mechanismen, Traumzustände, Split-Brain-Phänomene und tierisches Bewusstsein lassen sich aus fünf Prinzipien ableiten, die auf dieser Architektur aufbauen.

Für die vollständige mathematische Behandlung — einschließlich des kontinuierlichen Modellraum-Frameworks, der Modelldichtefunktion und der Formalisierung von Permeabilität als Informationstransfer zwischen Regionen dieses Raums — siehe Gruber (2026), *Toward a Mathematical Formalization of the Four-Model Theory*.