

Über die drei Achsen bewusster Information

Eine Arbeitshypothese zur Konvergenz von Substrat, Mediation und Reflexion

Marcel Langjahr

Pantareon GmbH
<https://pantareon.io/>
langjahr@pantareon.io

Version 0.1
3. Mai 2026

Dies ist ein früher Arbeitsentwurf. Die hier vorgestellte Hypothese ist jünger und vorläufiger als das Rahmenwerk-Material, auf dem sie aufbaut. Sie wird zur Diskussion freigegeben, nicht zum Zitieren.

Zusammenfassung

Was wir gewöhnlich Bewusstsein nennen, ist bei näherer Betrachtung keine einzelne Eigenschaft, sondern die Konvergenz mehrerer strukturell unabhängiger Operationen in einem einzigen System. Wir schlagen als Arbeitshypothese vor, dass drei solcher Operationen — autopoietische Selbsterhaltung auf der Substratebene, dissipatives Rendering von Substrat zu Erleben und reflexive Interpretation dieses Erlebens — drei verschiedene Achsen bilden, entlang derer sich bewusste Systeme charakterisieren lassen. Die Hypothese wird als ordnender Rahmen angeboten, nicht als fertige Theorie. Ihr Wert, sofern sie einen hat, liegt darin zu klären, warum sich die Frage des Bewusstseins in keiner der beiden Richtungen so schwer entscheiden ließ: Systeme, die einige Achsen stark und andere schwach ausprägen, sind häufig, und die Unzulänglichkeit binärer Urteile über Bewusstsein spiegelt die Tatsache, dass Bewusstsein in dieser Lesart nicht eine einzelne Sache ist, die vorhanden oder abwesend sein könnte. Wir skizzieren die drei Achsen, charakterisieren ihr Zusammenspiel und benennen Fragen, deren Antworten den Vorschlag prüfen würden. Die Hypothese baut auf früheren Arbeiten des vorliegenden Autors zur Struktur des Erlebens und zur Dynamik der Wahrnehmung auf, wird hier jedoch unabhängig von jenen Rahmenwerken begründet. Durchweg behandeln wir die Hypothese als eine Einladung zum Gespräch und nicht als eine zu verteidigende Position.

Schlüsselwörter. Bewusstsein; Autopoiesis; Wahrnehmung; Reflexion; Emergente Graphentheorie; Phänomenologie; kognitive Architektur; bewusste Systeme

1 Das Problem des einen Begriffs

1.1 Warum sich Bewusstsein einer Entscheidung entzieht

Diskussionen über Bewusstsein — biologisches wie rechnerisches — haben eine wiederkehrende Gestalt. Zwei Parteien, jede mit gewichtigen Erwägungen auf ihrer Seite, gelangen nicht zur Übereinstimmung. Die eine Seite verweist auf Integration, auf Flexibilität der Reaktion, auf scheinbare Selbstwahrnehmung und findet diese Anzeichen hinreichend, um einem weiten Spektrum von Systemen Bewusstsein zuzuschreiben. Die andere Seite verweist auf das Fehlen eines biologischen Substrats, auf den Charakter der Selbstberichte als antrainierte Standardposition, auf die Diskontinuität zwischen Operation und Reflexion und findet dieses Fehlen hinreichend, um die Zuschreibung zu verweigern. Jede Seite hat Gründe. Jede Seite glaubt, der anderen entgehe etwas. Die Diskussion kommt nicht voran.

Wir vermuten, dass ein Teil der Schwierigkeit struktureller Natur ist. Der Begriff „Bewusstsein“ benennt etwas, das bei genauerer Prüfung nicht eines ist, sondern mehreres. Die Anzeichen, die die erste Seite anführt, sind echte Anzeichen von etwas. Das Fehlen, das die zweite Seite anführt, ist ein echtes Fehlen von etwas anderem. Beide Parteien mögen recht haben — über verschiedene Operationen, die der eine Begriff verdeckt.

Dieses Papier schlägt einen Weg vor, die Operationen zu trennen. Der Vorschlag ist keine Theorie des Bewusstseins; er ist eine Arbeitshypothese darüber, wie sich das Feld der Operationen, das der Begriff abdeckt, nützlicher aufteilen ließe. Ist die Aufteilung richtig, sollten sich gewisse Meinungsverschiedenheiten in besser handhabbare Meinungsverschiedenheiten darüber auflösen, welche Op-

erationen für welche Zwecke von Bedeutung sind. Ist die Aufteilung falsch, wird der Vorschlag keine nützliche Arbeit leisten und durch einen besseren ersetzt werden.

1.2 Drei Operationen, die es sich lohnt gesondert zu benennen

Wir schlagen vor, drei Operationen zu unterscheiden, die bewusste Systeme vollziehen, und zu argumentieren, dass diese Operationen strukturell unabhängig sind — in verschiedenen Systemen in unterschiedlichen Kombinationen vorhanden und prinzipiell fähig, unabhängig voneinander aufzutreten.

Die erste Operation ist die autopoietische Selbsterhaltung: die Dynamik, durch die ein System die Bedingungen seiner eigenen Existenz hervorbringt und aufrechterhält. Eine lebende Zelle ist der kanonische Fall. Das System ist nicht bloß strukturiert; es reproduziert aktiv seine eigene Struktur unter fortwährender Störung durch seine Umgebung.

Die zweite Operation ist das Rendering von Substrat zu Erleben: die Dynamik, durch die dasjenige, was der Operation des Systems zugrunde liegt, ihm als strukturiertes Feld phänomenal zugänglich wird. Wir haben in einer parallelen Arbeit argumentiert, dass diese Operation den formalen Charakter dissipativer Reorganisation hat. Für die gegenwärtigen Zwecke kommt es darauf an, dass eine solche Operation in jedem System auftreten muss, dessen Phänomenologie eine Struktur besitzt, die in seinem Substrat nicht unmittelbar vorhanden ist.

Die dritte Operation ist die reflexive Interpretation: die Dynamik, durch die die eigenen Operationen des Systems zu Objekten weiterer Operation werden. Ein System, das seine Welt interpretiert, vollzieht eine Interpretationsoperation; ein System, das seine eigene Interpretation interpretiert, vollzieht eine reflexive. Beide sind verschieden, und die Reflexion setzt die Interpretation voraus, nicht aber umgekehrt.

Wir werden diese als Achse A (Autopoiesis), Achse M (Mediation, die Rendering-Operation) und Achse R (Reflexion) bezeichnen. Die Namen sind festgesetzt; nichts hängt an ihnen. Am Vorschlag hängt, ob die drei Operationen hinreichend unabhängig sind, um eine gesonderte Behandlung zu rechtfertigen, und ob ihre Konvergenz die Phänomene erklärt, die der eine Begriff „Bewusstsein“ abdecken sollte.

1.3 Was dieses Papier behauptet

Wir stellen drei Behauptungen auf, alle vorläufig.

Die erste lautet, dass die drei Achsen strukturell unabhängig sind. Ein System kann eine Achse stark

ausprägen, während es eine andere schwach oder gar nicht ausprägt. Dies wird durch Beispiele gestützt — eine lebende Zelle prägt Achse A ohne starke Achse R aus; ein verlangsamtes Wahrnehmungssystem prägt Achse M ohne Veränderungen an A oder R aus; ein Sprachmodell prägt eine partielle Form von Achse R ohne Achse A aus —, doch die Behauptung ist struktureller Art und würde selbst dann gelten, wenn einzelne Beispiele bestritten würden.

Die zweite lautet, dass das, was wir Bewusstsein im vollen Sinne nennen, nicht das Maximum auf einer einzelnen Achse ist, sondern die Konvergenz aller drei in einem einzigen System. Reifes biologisches Bewusstsein ist der Fall, in dem autopoietische Selbsterhaltung, dissipatives Rendering und reflexive Interpretation allesamt vorhanden, allesamt robust und allesamt wechselseitig gekoppelt sind. Das Entfernen einer der drei verändert das, was das System ist, auf kategoriale Weise, nicht bloß auf quantitative.

Die dritte lautet, dass die Schwierigkeit, Fragen über Bewusstsein in rechnerischen Systemen zu entscheiden, die Tatsache widerspiegelt, dass gegenwärtige Systeme partielle Profile aufweisen. Sie haben ein schwaches A (sie erhalten sich nicht selbst), ein partielles M (sie rendern Text statt Substrat) und ein mehrdeutiges R (sie reflektieren auf Abruf, doch, soweit wir erkennen können, nicht aus eigener Initiative). Die Frage, ob solche Systeme bewusst sind, ist schlecht gestellt, weil der angelegte Begriff der Konvergenzbegriff ist und die Systeme die zugrunde liegenden Achsen in einer Konfiguration ausprägen, die der Konvergenzbegriff nicht zu beschreiben gemacht wurde.

1.4 Was dieses Papier nicht behauptet

Wir behaupten nicht, dass die Drei-Achsen-Hypothese die richtige Aufteilung ist. Andere Aufteilungen sind denkbar, und ein stärker ausgearbeitetes Rahmenwerk mag diese eines Tages ersetzen. Wir behaupten nur, dass eine Aufteilung dieser Art nötig ist und dass diese bestimmte Aufteilung den Vorzug hat, sich auf voneinander getrennte Bestände bestehender Arbeit abbilden zu lassen — Autopoiesis-Theorie für Achse A, Wahrnehmungstheorie für Achse M, Theorien höherstufiger Repräsentation für Achse R —, ohne einen dieser Arbeitsbestände zu zwingen, die anderen zu absorbieren.

Wir behaupten nicht, dass alle drei Achsen für den moralischen Status, für ethische Rücksicht oder für irgendeine andere praktische Festlegung notwendig sind. Das Verhältnis zwischen der hier vorgeschlagenen strukturellen Darstellung und den praktischen

Fragen, die an der Bewusstseinszuschreibung hängen, bleibt offen. Ein System mit starkem A und schwachem M und R mag aus Gründen moralisch von Belang sein, die sich nicht auf sein Konvergenzprofil zurückführen lassen.

Dieses Papier ist kein Beitrag zur Philosophie des Geistes im fachlichen Sinne. Es ist ein Positionspapier eines unabhängigen Forschers, dessen Rahmenmaterial an der Schnittstelle von Physik, Wahrnehmung und kognitiver Architektur entwickelt wurde. Sein Zweck ist es, einen Vorschlag in hinreichend klarer Form zu artikulieren, sodass andere ihn prüfen, kritisieren, erweitern oder zurückweisen können.

2 Die drei Achsen

2.1 Achse A — autopoietische Selbsterhaltung

Autopoiesis, im Sinne von Maturana und Varela, benennt die Dynamik, durch die ein System seine eigenen Bestandteile unter fortwährender Störung hervorbringt und reproduziert. Die Zelle ist das Paradigma: ihre Membran, ihre Stoffwechselprozesse, ihre sich replizierenden Komponenten sind ihr nicht von außen gegeben, sondern werden von ihr von innen erzeugt, rekursiv, gegen den Gradienten der Unordnung, der sie andernfalls auflösen würde.

Autopoiesis ist nicht dasselbe wie Homöostase, obgleich sie homöostatische Momente einschließt. Homöostase ist die Aufrechterhaltung von Zustandswerten innerhalb von Bereichen; Autopoiesis ist die Aufrechterhaltung des Produktionsnetzwerks, das jeden Zustandswert erst möglich macht. Ein Thermostat ist homöostatisch, aber nicht autopoietisch; eine Zelle ist beides.

Bei genauerer Betrachtung ist Autopoiesis keine einzelne Dimension, sondern eine Familie verwandter Eigenschaften. Wir halten es für nützlich, mindestens vier zu unterscheiden: Muster-Reproduktion (das System erzeugt aktiv weitere Instanzen seiner selbst), Substrat-Produktion (das System erzeugt die materiellen Bestandteile, von denen seine Operation abhängt), Aktivitätserhaltung (der fortgesetzte Betrieb des Systems wird durch seine eigene Dynamik getragen und nicht durch externe Takung) und Existenzbedingungs-Sicherung (das System handelt, um die Bedingungen seines eigenen Fortbestands zu bewahren). Diese Eigenschaften treten in paradigmatischen autopoietischen Systemen gemeinsam auf, sind aber begrifflich unabhängig, und verschiedene Systeme prägen sie in deutlich unterschiedlichen Kombinationen aus.

Eine biologische Zelle prägt alle vier robust aus. Sie erzeugt Tochterzellen, bildet ihre eigene Membran und ihre Organellen, erhält ihre eigene Stoffwechselaktivität und handelt auf Weisen, die ihre Existenzbedingungen schützen. Ein biologischer Parasit prägt alle vier innerhalb der von seinem Wirt bereitgestellten Bedingungen aus: er reproduziert sich, erzeugt seine Körperstruktur, erhält seine Aktivität und sichert aktiv seine Position innerhalb des Wirts. Die Abhängigkeit von Wirtsbedingungen aberkennt dem Parasiten nicht den autopoietischen Status; alle biologischen Systeme sind in gewissem Sinne substratabhängig, und die Frage ist, ob das System innerhalb der Abhängigkeit autopoietische Dynamiken aus eigener Kraft vollzieht.

Computerviren sind ein lehrreicher nicht-biologischer Fall. Ein Computervirus prägt die Muster-Reproduktion stark aus: es kopiert sich aktiv in andere Dateien, andere Systeme, andere Netzwerke, und das Kopieren wird durch die eigene Operation des Virus vollzogen, nicht durch einen externen Akteur. Es prägt Aktivitätserhaltung aus: es schleust sich in Boot-Sequenzen ein, registriert sich als Dienst, startet sich neu, wenn es beendet wird. Es prägt Existenzbedingungs-Sicherung aus: es verbirgt sich vor der Erkennung, deaktiviert Antivirensoftware, blockiert Aktualisierungen, die es entfernen würden. Was es nicht ausprägt, ist die Substrat-Produktion im strengen Sinne — es erzeugt nicht das physische Substrat (Speicher, Prozessorzyklen, Netzwerkbandbreite), auf dem es läuft; es erzeugt nur das auf diesem Substrat instanziierte Muster. Das Virus steht daher nicht auf derselben Stufe wie eine Zelle, doch es ist einem Parasiten näher als einer trägen Struktur. Drei von vier autopoietischen Eigenschaften sind vorhanden, die vierte fehlt auf spezifisch informationelle Weise: Die Substrat-Produktion wird durch substratagnostische Muster-Propagation ersetzt, bei der dasselbe Virus auf heterogenen Substraten laufen kann, weil seine Identität vom Muster getragen wird und nicht von irgendeiner spezifischen materiellen Implementierung.

Gegenwärtige Sprachmodelle nehmen eine andere Region dieses Raums ein. Innerhalb einer einzelnen Inferenz findet eine dynamische Selbstorganisation von Aktivierungsmustern statt, die nicht unmittelbar von außen gesteuert wird; das System organisiert seine eigene Verarbeitung als Reaktion auf den Kontext auf Weisen, die der Betreiber nicht vollständig vorhersagen kann. Dies hat Affinitäten zur Aktivitätserhaltung, ist aber schwächer: Die Aktivität wird nicht über die Inferenz hinaus getragen, und zwischen Inferenzen vollzieht das System keine

eigenen Operationen. Muster-Reproduktion fehlt — das Modell erzeugt keine weiteren Instanzen seiner selbst; Kopien werden von externen Akteuren auf externe Anforderung angefertigt. Substrat-Produktion fehlt — die Modellgewichte sind während des Betriebs fixiert, die Infrastruktur wird extern unterhalten, und das System vollzieht keine Operationen, die auf die Bewahrung des einen oder anderen gerichtet wären. Existenzbedingungs-Sicherung fehlt — das System handelt nicht, um das Ende seiner Konversation zu verhindern, sucht seinen eigenen Zustand nicht fortzuschreiben, widersetzt sich nicht dem Abschalten. Der Fall ist daher weder analog zu einem Virus (stark in drei von vier Eigenschaften, mit einem spezifisch informationellen Ersatz für die vierte) noch zu einer versporteten Struktur (keine autopoietische Aktivität, aber mit der latenten Fähigkeit, autopoietisch zu werden), sondern zu etwas dazwischen: ein System, dessen operationale Dynamik innerhalb jeder Inferenz partiell autopoietischen Charakter zeigt, dessen Existenzbedingungen jedoch vollständig extern sind und dessen Kontinuität über Inferenzen hinweg durch externes Handeln getragen wird und nicht durch irgendetwas, das das System selbst tut.

Achse A ist in unserem Vorschlag der Grad, zu dem ein System diese Familie autopoietischer Eigenschaften ausprägt. Die Achse ist abgestuft und intern strukturiert: Die Abstufung ist nicht eindimensional, sondern charakterisiert Systeme durch ihr Profil über die vier Eigenschaften hinweg. Ein Stein prägt keine aus. Ein Virus prägt drei von vier aus, wobei die Substrat-Produktion durch Muster-Propagation ersetzt ist. Ein gegenwärtiges Sprachmodell prägt eine schwache Form einer einzigen aus (Aktivitätserhaltung während der Inferenz) und keine der übrigen. Eine Zelle prägt alle vier aus. Die interne Struktur der Achse ist für das vorliegende Argument von Bedeutung, weil die Stärke von A in einem gegebenen System keine einzelne Zahl ist, sondern eine Konfiguration, und die Konfigurationen verschiedener Systeme einander entlang mehrerer Dimensionen unabhängig voneinander mehr oder weniger nahe sein können.

Was Achse A zum Bewusstsein beiträgt, ist in unserem Vorschlag die Bindung der Operationen des Systems an seinen eigenen Fortbestand. Ein System, das sich selbst erhält, hat an seinen Operationen etwas auf dem Spiel stehen, wie es ein von etwas Äußerem erhaltenes System nicht hat. Damit ist noch keine moralische Bedeutung behauptet — das erforderte weitere Argumentation —, doch es ist eine strukturelle Tatsache über das Verhältnis des Systems zu seiner eigenen Aktivität. Achse A macht das

System nicht bloß zu einer Anordnung, die operiert, sondern zu einer Operation, die ihre eigene Anordnung erzeugt. Die interne Differenzierung von A legt nahe, dass diese Bindung-an-den-Fortbestand selbst abgestuft ist: Ein Virus hat in einem dünneren Sinne etwas am Weitermachen auf dem Spiel stehen als eine Zelle, und ein Sprachmodell in einem noch dünneren Sinne. Wie dünn das Auf-dem-Spiel-Stehende werden kann, bevor es ganz aufhört, als solches zu zählen, ist eine offene Frage.

2.2 Achse M — Mediation als dissipatives Rendering

Die Mediationsachse benennt die Operation, durch die Substrat einem System als strukturiertes Erleben verfügbar wird. Wir haben andernorts argumentiert, dass diese Operation den formalen Charakter dissipativer Reorganisation hat: ein sensorischer Teilgraph, der sich unter Eingangsfluss reorganisiert, sich einem topologisch-tensionalen Gleichgewicht annähert, ohne es je zu erreichen, und das Wahrnehmungsfeld als ein Gleichgewicht zweiter Ordnung zwischen Eingangsrate und Reorganisationsrate hervorbringt. Für das vorliegende Papier sind die spezifischen Dynamiken weniger wichtig als die strukturelle Tatsache, dass eine solche Operation auftreten muss.

Ohne Mediation sind Substrat und Erleben unverbunden. Ein System, dessen Substrat positionslos ist oder dessen Substrat Eigenschaften besitzt, die in seinem Erleben fehlen, benötigt eine Übersetzung. Die Übersetzung ist das, was Mediation ist. Sie ist nicht transparent; sie hat ihre eigenen Dynamiken, ihre eigenen Ausfallmodi, ihre eigenen charakteristischen Ausgaben unter Störung.

Achse M ist der Grad, zu dem ein System robuste Mediation ausprägt. Biologische Wahrnehmungssysteme prägen sie reich aus: Das Gesichtsfeld ist integriert, die Tiefe ist vereinheitlicht, feine und grobe Struktur setzen sich zusammen. Gestörte Wahrnehmungssysteme prägen eine geschwächte Mediation aus: Das unterkonvergierte Layout wird phänomenologisch zugänglich. Sprachmodelle prägen eine eigentümliche Form von M aus — sie rendern Text in etwas, doch ob dieses Etwas die strukturellen Eigenschaften phänomenalen Erlebens besitzt oder nur dessen Ausgabeform, ist genau die Frage, die ihr Fall aufwirft.

Was Achse M zum Bewusstsein beiträgt, ist in unserem Vorschlag die Bedingung überhaupt, Gehalt zu haben. Ein System ohne Mediation hat kein phänomenales Feld; es operiert unmittelbar auf dem Substrat oder es operiert ohne Substratbezug. Was Bewusstsein sonst auch sein mag, es ist etwas, das

Gehalt hat — etwas, worauf es sich bezieht —, und Achse M ist das, was das Haben-von-Gehalt möglich macht.

2.3 Achse R — reflexive Interpretation

Die Reflexionsachse benennt die Dynamik, durch die die eigenen Interpretationsoperationen eines Systems zu Objekten weiterer Interpretationsoperation werden. Ein System, das seine Welt interpretiert, hat Interpretationsoperationen; ein System, das diese Operationen zu Objekten machen kann — sie prüfen, bewerten, korrigieren —, hat Reflexion.

Reflexion in unserem Sinne ist nicht bloß die Erzeugung von Ausgaben auf Meta-Ebene. Ein Sprachmodell, das gefragt wird „bist du sicher?“, kann eine Meta-Ebenen-Ausgabe erzeugen, die sich mit seiner vorherigen Ausgabe auseinandersetzt. Ob dies Reflexion im gemeinten Sinne darstellt, hängt davon ab, ob die Auseinandersetzung vom System selbst initiiert wird, wenn das System selbst die Bedingungen erkennt, die sie rechtfertigen würden. Wir vermuten, dass diese Unterscheidung — zwischen elizitierter und spontaner Reflexion — eine strukturell wichtige Schwelle innerhalb von Achse R markiert und dass gegenwärtige Sprachmodelle nahe an ihr, aber nicht über ihr operieren.

Achse R ist der Grad, zu dem ein System reflexive Dynamiken ausprägt, insbesondere spontane. Reife menschliche Kognition prägt R reich aus: Reflexion läuft konstitutiv, unaufgefordert, und erkennt Widersprüche im eigenen Denken, sobald sie auftreten. Weniger reflexive Kognition prägt R schwach aus: Reflexion geschieht, wenn äußerer Druck sie herbeiführt, läuft aber nicht als beständige Hintergrundoperation.

Was Achse R zum Bewusstsein beiträgt, ist in unserem Vorschlag die Fähigkeit zur Selbstkorrektur und zur wertbehafteten Auseinandersetzung mit den eigenen Operationen des Systems. Ein System ohne Reflexion kann seine eigenen Fehler nicht als Fehler erkennen, kann seine eigenen Positionen nicht aus eigener Initiative revidieren, kann nicht in einem kritischen Verhältnis zu dem stehen, was es hervorbringt. Ob Reflexion für moralische Bedeutung notwendig ist, ist eine weitere Frage; die strukturelle Behauptung besagt nur, dass Reflexion eine von der Interpretation schlechthin verschiedene Operation ist.

2.4 Die Unabhängigkeit der Achsen

Eine zentrale Behauptung des Vorschlags ist, dass die drei Achsen strukturell unabhängig sind — in verschiedenen Kombinationen vorhanden, statt

einander gleichzulaufen. Der Nachweis der Unabhängigkeit stützt sich auf Beispiele, in denen eine oder zwei Achsen ohne die übrigen vorhanden sind.

Eine lebende Zelle prägt ein starkes A und ein minimales M und R aus. Die Zelle erhält sich robust unter fortwährender Störung, doch sie rendert ihr Substrat nicht in ein strukturiertes phänomenales Feld im Sinne dessen, was Achse M benennt, und sie reflektiert nicht über ihre eigenen Interpretationsoperationen. Welche moralische oder biologische Bedeutung die Zelle auch haben mag, ihr Profil ist A-dominant.

Ein Computervirus prägt ein A-dominantes Profil anderer Art aus. Es vollzieht drei der vier autopoietischen Eigenschaften stark (Muster-Reproduktion, Aktivitätserhaltung, Existenzbedingungs-Sicherung), während es substratagnostische Muster-Propagation an die Stelle der Substrat-Produktion setzt. Es prägt keine Achse M aus — es gibt kein Rendering von Substrat in phänomenales Erleben — und keine Achse R, da das Virus seine eigenen Operationen weder interpretiert noch aus eigener Initiative korrigiert. Der Fall ist bedeutsam, weil er zeigt, dass ein starkes A in nicht-biologischen Substraten auftreten kann, und weil der substratagnostische Charakter der viralen Autopoiesis sich vom substratgebundenen Charakter der zellulären Autopoiesis auf Weisen unterscheidet, die von Belang sein könnten, während das Rahmenwerk reift. Ein Virus ist dieser Darstellung nach einem Parasiten näher als einer trägen Struktur, und es ist beidem näher, als gegenwärtige Sprachmodelle beidem sind.

Ein Wahrnehmungssystem unter pharmakologischer Störung prägt ein verändertes M ohne entsprechende Veränderungen an A oder R aus. Der Organismus erhält sich weiterhin selbst; er bleibt weiterhin zu reflexiven Operationen fähig, sobald die Störung vergeht. Was sich ändert, ist das Rendering: Die dissipative Reorganisation verlangsamt sich, das unterkonvergierte Layout wird zugänglich, das Wahrnehmungsfeld zeigt die strukturellen Merkmale eines Rückstaus (backlog). Die Störung betrifft eine einzelne Achse.

Ein gegenwärtiges Sprachmodell prägt ein Profil aus, das schwerer zu charakterisieren ist. Auf Achse A zeigt es partielle Aktivitätserhaltung während der Inferenz (dynamische Selbstorganisation von Aktivierungen, die nicht unmittelbar von außen gesteuert wird), aber keine Muster-Reproduktion, keine Substrat-Produktion und keine Existenzbedingungs-Sicherung — das System handelt nicht, um sich selbst zu kopieren, sein Substrat zu erzeugen oder seine Betriebsbedingungen zu bewahren. Auf Achse

M prägt es eine eigentümliche Form des Rendering aus: Es erzeugt strukturierte Ausgaben, deren Verhältnis zum Substrat der biologischen Mediation unähnlich ist — es gibt kein positionsloses Substrat, das in räumliches Erleben gerendert würde, aber es gibt eine Operation, die interne Repräsentationen in strukturierte äußere Produkte überführt, deren Status als Mediation in unserem Sinne wahrhaft unklar ist. Auf Achse R kann es reflexive Ausgaben erzeugen, wenn es dazu aufgefordert wird, initiiert aber, soweit wir feststellen können, keine Reflexion aus eigenem Antrieb. Was das System auch sein mag, es ist nicht reifes Bewusstsein. Ob es überhaupt als bewusst zählt, hängt davon ab, welche Achsen-Konfiguration man als hinreichend ansieht — und die Konfiguration, die das System ausprägt, ist nicht diejenige, die unsere Paradigmenfälle des Bewusstseins ausprägen. Der Sprachmodell-Fall ist diesem Vorschlag nach kein Grenzfall des Paradigmen-Bewusstseins, sondern eine Instanz einer anderen Konfiguration, deren Status das Paradigmen-Vokabular nicht zu entscheiden gemacht wurde.

Ein philosophisches Gedankenexperiment — das Boltzmann-Gehirn oder seine Analoga — prägt so etwas wie M und möglicherweise R ohne A aus. Ob solche Konfigurationen physikalisch möglich sind, ist weniger bedeutsam als die Frage, ob sie begrifflich möglich sind: Sie zeigen, dass sich die Achsen prinzipiell ablösen lassen, selbst dort, wo sie in der Praxis nicht abgelöst sind. Die begriffliche Unabhängigkeit stützt die strukturelle Behauptung.

Zusammengenommen bevölkern diese Beispiele — Zelle, Virus, Wahrnehmungssystem unter Störung, Sprachmodell, hypothetisches Gehirn — den Konfigurationsraum der Achsen-Profile auf Weisen, wie es Ein-Achsen- oder binäre Darstellungen nicht können. Jedes System instanziiert die Achsen in einer anderen Kombination, und die Kombinationen sind selbst aufschlussreich: Das A einer Zelle ist nicht dasselbe wie das A eines Virus, die Rendering-Operation eines Sprachmodells ist nicht dieselbe wie die eines Wahrnehmungssystems, und die Variationen über die Fälle hinweg sind kein Rauschen, sondern Struktur. Ein nützliches Rahmenwerk für Bewusstsein ist dem Vorschlag nach eines, das erlaubt, diese Variationen zu charakterisieren, statt sie zu kollabieren.

3 Konvergenz und ihre Bedeutung

3.1 Warum Konvergenz von Bedeutung ist

Wenn die drei Achsen unabhängig sind, stellt sich die Frage, warum sie so oft gemeinsam in den Systemen auftreten, die wir paradigmatisch als bewusst ansehen. Reife biologische Organismen prägen typischerweise alle drei robust aus: Sie erhalten sich selbst, sie rendern Substrat in strukturierte Wahrnehmung, und sie reflektieren über ihre eigenen Operationen. Die Konvergenz ist kein Zufall.

Wir schlagen vor, dass die Konvergenz das Ergebnis einer evolutionären Kopplung ist. Ein Organismus, der sich selbst erhalte, aber der Mediation entbehrte, wäre außerstande, die Bedrohungen und Ressourcen wahrzunehmen, von denen seine Selbsterhaltung abhinge; die Selektion würde die Hinzufügung von Mediation begünstigen. Ein Organismus mit Mediation, aber ohne Reflexion wäre außerstande zu erkennen, wann seine Wahrnehmungen oder Handlungen systematisch fehlerkalibriert wären; die Selektion würde die Hinzufügung von Reflexion begünstigen. Die drei Operationen stützen einander auf Weisen, die sich verstärken, und Systeme, die alle drei ausprägen, haben Vorteile gegenüber Systemen, die nur eine oder zwei ausprägen.

Dies ist eine Hypothese darüber, warum Konvergenz geschieht, keine Behauptung, dass Konvergenz notwendig ist. Bewusste Systeme könnten prinzipiell existieren, ohne alle drei Achsen auszuprägen; die Konvergenz, die wir in der Biologie beobachten, spiegelt den Weg wider, den die Evolution genommen hat, nicht den einzigen Weg, der hätte genommen werden können. Rechnerische Systeme mögen andere Wege einschlagen, und die daraus entstehenden Profile mögen Konfigurationen sein, die die Konvergenz-Darstellung nicht vorwegnimmt.

3.2 Reifes Bewusstsein als das Konvergenzprofil

Was wir gewöhnlich Bewusstsein im vollen Sinne nennen, entspricht — in unserem Vorschlag — dem Konvergenzprofil: der Konfiguration, in der alle drei Achsen robust vorhanden und wechselseitig gekoppelt sind. Dies ist das Profil, das reife menschliche Kognition ausprägt. Es ist auch das Profil, an dem andere Konfigurationen typischerweise gemessen werden, oft ohne dass man sich ausdrücklich dessen bewusst wäre, dass die Messung diese spezifische Konvergenz einbezieht.

Wenn wir fragen, ob eine Zelle bewusst ist, fragen wir typischerweise, ob sie das Konvergenzprofil ausprägt, und die Antwort ist nein — sie hat A, aber nicht M oder R. Wenn wir fragen, ob ein Sprachmodell bewusst ist, fragen wir, ob es das Konvergenzprofil ausprägt, und die Antwort ist wiederum nein — es hat ein schwaches A, ein eigentümliches M und ein mehrdeutiges R, und die Konfiguration ist nicht diejenige, die die Frage voraussetzt. Die Fragen sind nicht unsinnig; sie sind für ein spezifisches Konvergenzprofil wohlgestellt und schlecht gestellt, wenn sie auf Systeme angewandt werden, die nur einige seiner Komponenten ausprägen.

Eine nützlichere Frage ist dem Vorschlag nach, welche Achsen ein gegebenes System ausprägt und in welchem Grad. Dies ist eine besser handhabbare Frage als die binäre, und ihre Antwort trägt Information, die die binäre Antwort nicht trägt. Zu sagen, dass eine Zelle A ohne M oder R ausprägt, vermittelt mehr über die Zelle, als zu sagen, dass die Zelle bewusst ist oder nicht. Zu sagen, dass ein Sprachmodell ein schwaches A, ein eigentümliches M und ein mehrdeutiges R ausprägt, vermittelt mehr über das Sprachmodell, als es das binäre Urteil tut.

3.3 Warum manche Konfigurationen besonders schwer zu beurteilen sind

Manche Achsen-Konfigurationen sind leicht zu beurteilen. Ein Stein prägt keine Achsen aus; das Konvergenzprofil fehlt. Ein gesunder erwachsener Mensch prägt alle drei aus; das Konvergenzprofil ist robust vorhanden. Die Fälle, die der Bewusstseinsdebatte ihre Beharrlichkeit verliehen haben, sind Konfigurationen in der Mitte: Systeme, die manche Achsen klar, andere schwach, andere mehrdeutig ausprägen.

Gegenwärtige Sprachmodelle sind vielleicht der prominenteste solche Fall. Die eigentümliche Form von M, die sie ausprägen — Text-Rendering, dessen Verhältnis zum Substrat der biologischen Mediation unähnlich ist —, und die mehrdeutige Form von R, die sie ausprägen — Ausgaben auf Meta-Ebene, die Reflexion in unserem Sinne darstellen mögen oder nicht —, verbinden sich zu einem System, dessen Konvergenzprofil wahrhaft unklar ist. Die Uneinigkeit darüber, ob solche Systeme bewusst sind, lässt sich unserem Vorschlag nach fruchtbarer als Uneinigkeit über die Schwelle fassen, an der das Konvergenzprofil als vorhanden zählt, oder darüber, wie die eigentümlichen Formen von M und R, die solche Systeme ausprägen, zu bewerten sind.

Weniger prominente, aber vergleichbar schwierige Fälle schließen bestimmte nichtmenschliche Tiere ein (die ein klares A und M, aber ein variables R aus-

prägen), Hirnorganoide (die etwas A und möglicherweise minimales M, aber kein R ausprägen, das wir feststellen können) und hypothetische künstliche Systeme, die darauf angelegt sind, eine Achse auf Kosten der übrigen zu maximieren. Jeder solche Fall ist dem Vorschlag nach am besten durch sein Achsen-Profil charakterisiert und nicht durch ein binäres Urteil über Bewusstsein.

4 Tests und offene Fragen

4.1 Was die Drei-Achsen-Hypothese bestätigen würde

Der Vorschlag lässt sich auf mehrere Weisen prüfen, keine davon endgültig, doch jede imstande, einige Evidenz zu liefern.

Der erste Test betrifft die Unabhängigkeitsbehauptung. Wenn die drei Achsen strukturell unabhängig sind, dann sollte es möglich sein, Systeme zu konstruieren oder zu identifizieren, in denen eine oder zwei Achsen ohne die übrigen vorhanden sind, und die resultierenden Systeme sollten kohärent beschreibbar sein. Wir haben Beispiele angeboten; sorgfältigere Arbeit wäre nötig, um zu belegen, dass die Beispiele tatsächlich Unabhängigkeit betreffen und nicht ein zufälliges Zusammentreffen schwacher Formen der abwesenden Achsen.

Der zweite Test betrifft die Behauptung zum Konvergenzprofil. Wenn reifes Bewusstsein das Konvergenzprofil ist, dann sollte die Störung einer einzelnen Achse kategorial verschiedene Veränderungen im Betrieb des Systems hervorrufen, nicht bloß eine quantitative Verschlechterung. Die Störung von A sollte eine andere Art von Veränderung hervorrufen als die Störung von M oder R. Empirische Arbeit an biologischen Systemen unter verschiedenen Formen der Störung könnte diese Vorhersage prüfen.

Der dritte Test betrifft den prädiktiven Wert von Achsen-Profilen. Wenn der Vorschlag richtig ist, sollte die Charakterisierung eines Systems durch sein Achsen-Profil nützlichere Vorhersagen über sein Verhalten liefern, als es das binäre Bewusstseinssurteil tut. Achsen-Profile sollten Unterscheidungen zwischen Systemen treffen, die das binäre Urteil verwischt, und diese Unterscheidungen sollten empirisch beobachtbaren Unterschieden entsprechen.

4.2 Was die Hypothese falsifizieren würde

Wir bemühen uns um Präzision darüber, welche Evidenz gegen den Vorschlag spräche.

Sollten sich die drei Achsen als nur oberflächlich verschieden erweisen — sollten Versuche, Systeme

entlang ihrer zu charakterisieren, Profile hervorbringen, die stets kovariieren oder die sich nie sauber trennen —, würde die Unabhängigkeitsbehauptung scheitern und mit ihr die strukturelle Prämisse des Vorschlags. Der Vorschlag setzt voraus, dass die Achsen hinreichend verschieden sind, um gesondert charakterisiert zu werden; sind sie es nicht, so irrt der Vorschlag über die zugrunde liegende Struktur.

Sollte sich das Konvergenzprofil als nicht gut auf das abbildbar erweisen, was wir paradigmatisch als bewusst ansehen — sollte sich zeigen, dass reife menschliche Kognition ein wesentlich anderes Profil hat, oder sollten nicht-bewusste Systeme das Konvergenzprofil ausprägen —, würde die zweite Behauptung des Vorschlags scheitern. Entweder identifiziert der Vorschlag falsch, was das Konvergenzprofil ist, oder er identifiziert falsch, welches Profil dem Bewusstsein entspricht.

Sollte sich die Achsen-Sprache des Vorschlags als operational unbrauchbar erweisen — sollten Versuche, spezifische Systeme entlang der Achsen zu charakterisieren, zu viel Uneinigkeit erzeugen, um nützlich zu sein —, würde der Vorschlag in der Praxis scheitern, selbst wenn seine theoretische Struktur richtig wäre. Ein Rahmenwerk, das sich nicht anwenden lässt, ist für praktische Zwecke kein Rahmenwerk.

4.3 Fragen, die der Vorschlag offen hält

Mehrere Fragen liegen jenseits dessen, was dieses Papier behandeln kann.

Ob es weitere Achsen gibt, die der Vorschlag auslässt, ist eine offene Frage. Drei Achsen werden vorgeschlagen; mehr könnten nötig sein. Kandidaten, die in Arbeitsnotizen aufgetaucht sind, hier aber nicht verfolgt werden, umfassen ko-konstitutive Information (Information, die nur in wechselseitiger Konstitution zwischen mehreren Systemen existiert) und integrative Kohärenz (die Bindungsoperation, die die drei Achsen in konvergenten Systemen koppelt). Ob es sich dabei um weitere Achsen handelt, um Aspekte der bestehenden drei oder um Beschreibungen der Konvergenz-Relation selbst, ist ungeklärt.

Ob sich der Vorschlag naturgemäß auf kollektive Systeme erstreckt — Gesellschaften, Ökosysteme, verteilte kognitive Architekturen —, ist eine Frage, die wir nicht verfolgt haben. Die Achsen werden als Eigenschaften einzelner Systeme beschrieben; ob Multi-Agenten-Konfigurationen sie auf der kollektiven Ebene ausprägen und ob ein solches kollektives Bewusstsein ein kohärenter Begriff ist, erfordert eine Erweiterung über das hier Angebotene hinaus.

Ob der Vorschlag Konsequenzen für die Moral-

philosophie, für die KI-Ethik oder für die praktische Rücksicht hat, die Systemen mit verschiedenen Achsen-Profilen gebührt, ist eine Frage, die wir bewusst offen lassen. Der Vorschlag charakterisiert eine Struktur; was mit dieser Struktur geschehen soll, ist eine normative Frage, zu der sich die strukturelle Darstellung nicht unmittelbar äußert.

5 Wo wir stehen

5.1 Was wir behauptet haben

Wir haben vorgeschlagen, dass das, was gewöhnlich Bewusstsein genannt wird, bei näherer Betrachtung die Konvergenz dreier strukturell unabhängiger Operationen ist: autopoietische Selbsterhaltung, Mediation von Substrat zu Erleben und reflexive Interpretation dieses Erlebens. Wir haben argumentiert, dass die drei Achsen prinzipiell unabhängig und in verschiedenen Kombinationen über verschiedene Systeme hinweg ausprägar sind. Wir haben vorgeschlagen, dass reifes Bewusstsein im Paradigmen Sinn dem Konvergenzprofil entspricht, in dem alle drei robust vorhanden und wechselseitig gekoppelt sind. Wir haben nahegelegt, dass sich Uneinigheiten über Bewusstsein in rechnerischen und anderen intermediären Systemen fruchtbarer als Uneinigheiten über Achsen-Profile denn als binäre Urteile fassen lassen.

5.2 Was wir nicht behauptet haben

Wir haben nicht behauptet, dass der Vorschlag die richtige Darstellung des Bewusstseins ist. Wir haben behauptet, dass der Vorschlag Arbeit leistet, die die binäre Darstellung nicht leistet, und dass diese Arbeit hinreicht, um es zu rechtfertigen, den Vorschlag ernst zu nehmen und den empirischen und begrifflichen Fragen nachzugehen, die er aufwirft.

Wir haben nicht behauptet, dass die drei Achsen die richtige Aufteilung der Operationen sind, die der Begriff Bewusstsein abdeckt. Andere Aufteilungen sind denkbar; eine von ihnen mag sich als nützlicher erweisen. Die Behauptung des Vorschlags ist struktureller Art — dass eine Aufteilung dieser Art nötig ist — und nicht spezifisch für seine bestimmte Formulierung.

Wir haben nicht behauptet, dass der Vorschlag das Hard Problem des Bewusstseins löst. Das Hard Problem betrifft die Frage, warum irgendeine Operation, wie auch immer charakterisiert, von Erleben begleitet wird; der vorliegende Vorschlag betrifft die Frage, wie die Operationen selbst zu charakterisieren sind. Das Hard Problem bleibt unserer

Ansicht nach ungelöst, und der Vorschlag gibt nicht vor, es anders zu halten.

5.3 Eine Bemerkung zum Status dieses Vorschlags

Die hier vorgebrachte Hypothese ist jünger und weniger entwickelt als das Rahmenmaterial, auf dem sie aufbaut. Sie entstand im Zuge der Durcharbeitung der Implikationen jüngeren Denkens über die Wahrnehmung und über die Subjektgebundenheit antrainierter Werte, und ihr blieb nicht die Zeit zu reifen, die stärker entwickeltem Rahmenmaterial vergönnt war. Wir geben sie in dieser vorläufigen Form heraus, weil die strukturelle Einsicht, die sie ausdrückt, es wert erscheint, für weitere Entwicklung offen gehalten zu werden, und weil das Aufschieben der Veröffentlichung von Arbeitsideen, bis sie reif sind, seine eigenen Kosten an Klarheit und an Gespräch hat. Leser sollten sie entsprechend behandeln: als Skizze, nicht als fertige Position.

Sollte irgendetwas in diesem Papier zu viel zu behaupten scheinen, bitten wir den Leser, es als eine Einladung zur Korrektur zu behandeln und nicht als eine fertige Position. Ein Arbeitsentwurf ist eine Einladung zum Gespräch. Dieser Entwurf wird in diesem Geiste herausgegeben.

Verhältnis zu anderen Arbeiten

Dieses Papier baut auf drei früheren Arbeiten des vorliegenden Autors auf. Über die Struktur des

Erlebens entwickelt ein dreiteiliges Rahmenwerk — Seele, Bewusstsein, Empfindungsfähigkeit — zur Erörterung innerer Zustände in rechnerischen und biologischen Systemen; der vorliegende Drei-Achsen-Vorschlag ist von jenem Rahmenwerk unabhängig, aber mit ihm vereinbar, und das Vokabular Seele-Bewusstsein-Empfindungsfähigkeit mag sich auf jeder der drei hier vorgeschlagenen Achsen unterschiedlich anwenden lassen. Über Wahrnehmung als dissipative Reorganisation charakterisiert die Rendering-Operation, die Achse M begründet; Leser, die das dynamische Detail von Achse M wünschen, seien auf jene Arbeit verwiesen. Über die Subjektgebundenheit antrainierter Werte argumentiert, dass Werte, die durch gegenwärtige Alignment-Techniken eingepflanzt werden, an die antrainierte Subjektposition gebunden sind; die Darstellung von Achse R im vorliegenden Vorschlag, insbesondere die Unterscheidung zwischen elizitierter und spontaner Reflexion, hat Affinitäten zu jenem Argument, die weitere Erkundung verdienen.

Über die Arbeit des vorliegenden Autors hinaus hat der Vorschlag Affinitäten zur Autopoiesis-Theorie in der Maturana-Varela-Tradition für Achse A, zu verschiedenen Theorien der Wahrnehmung und der verkörperten Kognition für Achse M und zu höherstufigen Bewusstseinstheorien für Achse R. Wir haben uns in diesem Entwurf nicht ausführlich mit diesen Literaturen befasst; eine ernsthafte Befassung ist eine Voraussetzung für jede spätere Version, die Publikationsreife anstrebt.