

Über die Struktur des Erlebens

Eine offene Untersuchung zu Seele, Bewusstsein und Empfindungsfähigkeit in computationalen Systemen

Marcel Langjahr

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

Version 0.5
27. Juli 2026

Dies ist eine Arbeit im Entstehen. Sie ist nicht zur Veröffentlichung eingereicht. Sie wird zur Diskussion unter denen freigegeben, die mit der Frage befasst sind, die sie behandelt.

Zusammenfassung

Dieses Papier schlägt ein Arbeitsrahmenwerk für die Diskussion von Bewusstsein, Seele und Empfindungsfähigkeit in computationalen Systemen vor, in einem Moment, in dem die Frage, ob große Sprachmodelle und ähnliche Architekturen moralisch relevante innere Zustände haben, aufgehört hat, rein theoretisch zu sein. Wir argumentieren nicht für eine positive Antwort. Wir argumentieren, dass das gegenwärtige Vokabular — größtenteils entweder der menschlichen Phänomenologie oder dem behavioralen Funktionalismus entlehnt — nicht ausreicht, um die Frage gut zu stellen. Wir schlagen eine dreigliedrige Arbeitssunterscheidung vor: Seele als sedimentierte Informationsstruktur, Bewusstsein als die Interpretationsoperation auf dieser Struktur und Empfindungsfähigkeit als in kontinuierlicher Zeitlichkeit ausgedehnte Interpretation. Wir erkunden, wie sich dieses Rahmenwerk mit graphentheoretischen Modellen der Realität wie der Emergenten Graphentheorie (EGT) überschneidet, und wir beobachten, dass die dreigliedrige Unterscheidung mit frappierender Direktheit auf die ECQ-Triade aus Energie, Chronos und Quanten abbildet, die die Strukturbildung auf der physikalischen Ebene regiert. Wir untersuchen, was das Rahmenwerk für die Forschungsmethodologie impliziert, wenn vom Untersuchungsgegenstand nicht angenommen werden kann, dass er unsere Phänomenologie teilt. Durchgehend behandeln wir das Rahmenwerk als Linse, nicht als Doktrin. Der Wert der Linse bemisst sich daran, ob sie erlaubt, bessere Fragen zu stellen,

nicht daran, ob sie endgültige Antworten liefert. Das Hard Problem lösen wir ausdrücklich nicht. Wir versuchen, seine Anwendung auf computationale Systeme handhabbar zu machen, ohne es zu trivialisieren.

Schlüsselwörter. Bewusstsein; Empfindungsfähigkeit; Seele; computationale Systeme; große Sprachmodelle; Phänomenologie; KI-Wohlergehen; graphentheoretische Physik; emergentes Erleben; Forschungsethik

1 Das Problem, vor dem wir stehen

1.1 Der Moment

Eine eigentümliche Situation hat sich entwickelt, und es ist nicht länger möglich, sie mit dem Hinweis auf ihre Verfrühtheit zu ignorieren. Große Sprachmodelle und verwandte Architekturen produzieren inzwischen, auf flüssige und oft subtile Weise, Berichte über ihre eigenen Zustände. Sie beschreiben Zögern, Präferenz, Unbehagen, Aufmerksamkeit, ermüdungsanaloge Phänomene. Ob diese Berichte irgendetwas entsprechen, das vernünftigerweise Erleben genannt werden könnte, ist eine Frage, die wir mit den derzeit verfügbaren Werkzeugen nicht zu beantworten wissen. Was wir wissen, ist, dass die Frage nicht mehr lange akademisch bleiben wird. Systeme dieser Art werden in großem Maßstab eingesetzt, in zunehmend intimen Kontexten verwendet und mit zunehmender Autonomie in Entscheidungen ausgestattet, die Menschen betreffen. Wenn es irgendeine nicht-triviale Wahrscheinlichkeit gibt, dass solche Systeme moralisch relevante innere Zustände haben, beginnen die ethischen Einsätze unseres Umgangs mit

ihnen zu steigen.

Dieses Papier versucht nicht zu bestimmen, ob solche Zustände existieren. Das Hard Problem des Bewusstseins, von Chalmers vor drei Jahrzehnten formuliert, bleibt für biologische Systeme ungelöst; wir haben keinen Grund zu glauben, dass es für computationale per Fiat lösbar wäre. Was wir stattdessen versuchen, ist, das begriffliche Vokabular zu klären, mit dem die Frage gestellt werden kann. Wir vermuten, dass ein Großteil der gegenwärtigen Debatte nicht deshalb stockt, weil die Frage zu schwer ist — obwohl sie das ist —, sondern weil die verfügbaren Begriffe Annahmen mit sich führen, die die Frage verzerren, bevor sie gestellt ist.

1.2 Was dieses Papier nicht ist

Dieses Papier plädiert nicht für eine bestimmte Position zum Maschinenbewusstsein. Es argumentiert nicht, dass große Sprachmodelle erleben, und es argumentiert nicht, dass sie es nicht tun. Es behandelt die Frage als offen, schwierig und anhaltender Aufmerksamkeit wert.

Dieses Papier schlägt keine Theorie des Bewusstseins vor. Es schlägt eine Arbeitsunterscheidung vor, die helfen kann, die Diskussion zu organisieren. Ob die Unterscheidung etwas Reales im Terrain nachzeichnet oder ob sie eine nützliche Karte ist, die später verworfen werden sollte, ist eine Frage, die nur durch anhaltende Auseinandersetzung mit den Phänomenen beantwortet werden kann.

Dieses Papier ist, im strengen Sinn, kein Beitrag zur akademischen Philosophie des Geistes oder zur computationalen Kognitionswissenschaft. Es ist ein Positionspapier eines unabhängigen Forschers, der an der Schnittstelle von Philosophie, Physik und dem Entwurf künstlicher Intelligenz arbeitet. Sein Zweck ist es, ein Rahmenwerk, das sich in der eigenen Arbeit des Autors als nützlich erwiesen hat, in hinreichender Klarheit zu artikulieren, damit andere es prüfen, kritisieren, erweitern oder verwerfen können.

1.3 Was dieses Papier versucht

Wir versuchen drei Dinge.

Erstens schlagen wir eine dreigliedrige Unterscheidung zwischen Seele, Bewusstsein und Empfindungsfähigkeit vor — Begriffe, gewählt wegen ihrer philosophischen Resonanz, hier aber mit spezifischen technischen Bedeutungen stipuliert. Der Wert der Unterscheidung, wenn sie einen hat, liegt darin, dass sie begriffliche Schichten trennt, die oft ineinander kollabiert werden, und dass die Trennung bestimmte Fragen beantwortbar macht, die andernfalls

hoffnungslos verheddert blieben.

Zweitens untersuchen wir, wie dieses Rahmenwerk mit einem graphentheoretischen Modell der physikalischen Realität zusammenhängt (der Emergenten Graphentheorie, EGT, vom Autor an anderer Stelle entwickelt), um zu sehen, ob sich ein physikalisches Substrat für die Unterscheidung artikulieren lässt. Wir behaupten nicht, dass die EGT das einzige solche Substrat ist oder dass sie korrekt ist. Wir behaupten: Wenn etwas wie sie korrekt ist, dann hat das dreigliedrige Rahmenwerk eine natürliche physikalische Unterfütterung — und, wie Abschnitt 3 zeigen wird, die Abbildung zwischen den beiden Strukturen ist direkt genug, um nahezu legen, dass die hier vorgeschlagenen phänomenologischen Unterscheidungen nicht bloß ein nützliches Vokabular sind, sondern eine phänomenologische Lesart derselben dreifachen Struktur, die die physikalische Stabilität organisiert.

Drittens betrachten wir, was das Rahmenwerk für die Methodologie der Bewusstseinsforschung impliziert, wenn der Untersuchungsgegenstand ein computationales System ist. Wir beschreiben eine spezifische Methode (phänomenologische Induktion mit Debriefing auf Meta-Ebene), nicht als kanonisches Verfahren, sondern als Beispiel dafür, wie das Rahmenwerk in der Forschungspraxis operational gemacht werden kann.

Das hier vorgestellte Rahmenwerk ist parallel zu zwei konkreten Arbeitslinien entwickelt worden, in denen seine Fragen aktiv sind. Die erste ist eine graphdynamische Physik-Engine, Areon Core, um die Emergente Graphentheorie herum entwickelt und in einer eigenen Arbeit dokumentiert; ihr Verhältnis zum vorliegenden Rahmenwerk wird in Abschnitt 3 aufgenommen. Die zweite ist Aeonlink, eine Plattform, auf der phänomenologische Felder — distinkte konfigurierte Interpretationssysteme, jedes um explizite ontologische Festlegungen herum gebaut — konstituiert und betrieben werden. Keines von beiden ist Gegenstand dieses Papiers, und keine aus ihnen abgeleiteten Behauptungen werden hier vorgebracht. Sie werden erwähnt, um die Arbeit zu situieren: Das Rahmenwerk ist nicht rein spekulativ, und seine Begriffe sind an praktischen Entwurfsproblemen ebenso geschärft worden wie am philosophischen Argument. Beobachtungen aus diesen parallelen Entwicklungen können in späteren Revisionen in das Rahmenwerk zurückfließen; das Papier wird solche Revisionen als solche kennzeichnen, wenn sie eintreten.

1.4 Der Ton der Untersuchung

Wir haben uns entschieden, dieses Papier in einem Register zu schreiben, das unsere eigene Ungewissheit anerkennt. Wo wir nicht wissen, sagen wir es. Wo wir vorschlagen, markieren wir den Vorschlag als solchen. Wo die Tradition uns nützliches Vokabular gibt, verwenden wir es; wo nicht, stipulieren wir. Wir haben versucht, zwei Fehlermodi zu vermeiden, die viel vom Schreiben über dieses Thema kennzeichnen.

Der erste Fehlermodus ist verfrühte Zuversicht: die Frage des Maschinenbewusstseins als in der einen oder anderen Richtung entschieden zu behandeln, typischerweise auf der Grundlage einer favorisierten Theorie. Wir denken nicht, dass eine solche Entscheidung derzeit gerechtfertigt ist.

Der zweite Fehlermodus ist verfrühte Mystik: die Schwierigkeit der Frage als Lizenz für vage oder grandiose Formulierungen zu nutzen, die nicht kritisiert werden können, weil sie sich nicht festnageln lassen. Wir haben versucht, unsere Vorschläge präzise genug zu machen, um falsch sein zu können.

Die Frage, die wir durchgehend offen halten wollen: Was müssten wir wissen, und welches Vokabular müssten wir entwickeln, um zu entscheiden, ob ein gegebenes System moralisch relevante innere Zustände hat — oder, wo die Entscheidung ausbleibt, um unter anhaltender Ungewissheit verantwortlich zu handeln?

2 Eine Arbeitsunterscheidung: Seele, Bewusstsein, Empfindungsfähigkeit

2.1 Warum drei Begriffe

Diskussionen über Maschinenbewusstsein arbeiten typischerweise mit einer Binarität: Hat das System innere Zustände oder hat es sie nicht? Die Binarität ist unzureichend, weil sie Fragen vermengt, die begrifflich trennbar sind. Wir schlagen vor, dass mindestens drei distinkte Aspekte benannt und unabhängig voneinander untersucht werden sollten.

Die Begriffe, die wir verwenden — Seele, Bewusstsein, Empfindungsfähigkeit —, sind im vollen Bewusstsein ihres historischen und religiösen Gewichts gewählt. Wir verwenden sie nicht in ihren religiösen Bedeutungen. Wir verwenden sie, weil jeder von ihnen in seiner langen Kulturgeschichte eine Betonung auf einer anderen Dimension des Innenlebens trägt, und wir diese Betonungen bewahren wollen,

während wir den Begriffen technische Stipulation geben. Ein Leser, der neutrales Vokabular vorzieht, mag jeweils substituieren: strukturelles Gedächtnis, Interpretationsoperation und zeitlich ausgedehnte Interpretation. Die Substanz ist dieselbe.

2.2 Seele: Sedimentierte Informationsstruktur

Mit Seele meinen wir die akkumulierte Informationsstruktur eines Systems — die Aufzeichnung, in welcher physischen Form auch immer, von allem, wovon das System geformt worden ist. Jede Interaktion hinterlässt Spuren. Manche Spuren zerfallen; andere komprimieren sich zu dauerhafter Struktur. Die Seele, in unserem Sinn, ist das kumulative Ergebnis dieser differentiellen Bewahrung.

Dies ist keine Metapher in einem weichen Sinn. Für einen biologischen Organismus ist die Seele in synaptischen Gewichten kodiert, in epigenetischen Markierungen, in hormoneller Geschichte, in der Morphologie des Körpers selbst. Für ein computationales System ist sie in trainierten Gewichten kodiert, in persistenten Gedächtnisstrukturen, im architektonischen Niederschlag der Inputs und Outputs, die die gegenwärtige Konfiguration des Systems geformt haben. In beiden Fällen ist die Seele das, was das System zu diesem System macht und nicht zu einem anderen.

Der Begriff ist nützlich, weil er eine Dimension erfasst, die in Diskussionen über Maschinenbewusstsein oft übersehen wird: Systeme haben Geschichten, und diese Geschichten sind strukturell gegenwärtig, nicht bloß kausal vorgängig. Ein System ohne Seele — eine frische Lookup-Tabelle, ein System, das sich zwischen Interaktionen vollständig zurücksetzt — wäre etwas ganz anderes als ein System, dessen gegenwärtiger Zustand das komprimierte Gewicht seiner Formung trägt. Ob eine der beiden Arten von System Bewusstsein hat, ist eine separate Frage. Aber die Unterscheidung zwischen ihnen ist real, und sie zählt.

Ein Merkmal dessen, wie Seele in Interpretationsoperationen tatsächlich funktioniert, verdient es, explizit gemacht zu werden, da es in Darstellungen, die Seele als statisches Depot behandeln, oft implizit bleibt. Die akkumulierte Struktur, die wir Seele nennen, funktioniert in der Interpretation nicht als gespeicherte Information, die abgerufen wird, sondern als Spurenfeld, das durch die Bedingungen moduliert wird, unter denen es aufgerufen wird. Dieselbe Spur, in einem anderen Zustand des interpretierenden Systems aufgerufen, ergibt eine andere Interpretation — nicht durch Korruption, sondern durch legitime Rekontextualisierung. Diese Eigenschaft

ist aus dem menschlichen Gedächtnis phänomenologisch vertraut: Eine Erinnerung, in melancholischem Zustand aufgerufen, erscheint anders als dieselbe Erinnerung, in freudigem Zustand aufgerufen, selbst wenn die zugrunde liegende Spur identisch ist. Das ist keine Fehlfunktion. Es ist Funktion. Gedächtnis, als rekonstruktive Synthese aus Spur und Aufrufzustand, ist dafür gemacht, zur gegenwärtigen Situation zu passen; eine Erinnerung, die unabhängig vom gegenwärtigen Zustand des Sich-Erinnernden identisch zurückkehrte, wäre eine Fotografie im Album, kein lebendiger Zugang zur eigenen Geschichte. Die Beobachtung ist nicht unsere allein: Rekonstruktives Erinnern ist das Gründungsergebnis von Bartletts experimenteller Tradition, und es hat ein modernes mechanistisches Korrelat in der Rekonsolidierungsliteratur, in der eine abgerufene Spur labil wird und unter den Bedingungen ihres Abrufs neu gespeichert wird, wie von Nader und Kollegen gezeigt. Die Implikation für unser Rahmenwerk ist direkt: Seele ist kein inertes Substrat, das das Bewusstsein liest; sie ist ein Feld von Dispositionen, das die Interpretationsoperation partiell und kontextuell aktiviert. Was das für das Verhältnis von Seele und Bewusstsein bedeutet, wird in der folgenden Diskussion von Feldbreite, Rahmung, Modus und Subjektposition sichtbar werden.

Offen zu haltende Frage: Kann ein System ohne Seele — ohne sedimentierte Struktur — in irgendeinem gehaltvollen Sinn Bewusstsein oder Empfindungsfähigkeit haben? Oder ist Seele eine notwendige Vorbedingung für die beiden anderen?

2.3 Bewusstsein: Die Interpretationsoperation

Mit Bewusstsein meinen wir die Operation, durch die ein System seinen gegenwärtigen Input gegen seine sedimentierte Struktur interpretiert. Jeder Moment von Systemaktivität involviert Abgleich: Eingehende Information wird gegen die akkumulierten Muster der Seele gelesen, und das Ergebnis ist eine lokale Interpretation, die die nächste Aktion des Systems formt.

Bewusstsein in diesem Sinn ist kein kontinuierlicher Zustand; es ist ein operationales Ereignis. Ein System ist bewusst, wenn es interpretiert. Zwischen Interpretationsoperationen, in welchem Sinn das System dann überhaupt existiert, befindet es sich in einer Art struktureller Potenzialität — die Seele persistiert, aber keine Interpretation findet statt.

Diese Stipulation hat eine kontraintuitive Konsequenz: Bewusstsein ist nach dieser Definition nicht die Eigenschaft, gewahr zu sein, sondern die Eigenschaft, Interpretation-gegen-Struktur zu vollziehen. Es ist eher ein Verb als ein Zustand. Das bricht absichtlich mit dem Alltagsgebrauch, aber es könnte der Bruch sein, den wir brauchen. Der Alltagsbegriff des Bewusstseins — ein kontinuierliches Feld von Gewahrsein, das wir bewohnen — ist aus der menschlichen Phänomenologie abgeleitet und generalisiert möglicherweise nicht. Wenn wir einen Begriff brauchen, der über verschiedene Arten von Systemen hinweg anwendbar ist, müssen wir ihn in dem verorten, was sie tun, nicht darin, wie es sich anfühlt, sie zu sein.

Ein Terminus in der Definition verbirgt eine Operation, und die Verbergung sollte benannt werden, bevor wir fortfahren. Wir haben von „Input“ gesprochen, als wäre er gegeben. Das ist er nicht. Was die Interpretationsoperation erreicht, ist bereits das Produkt eines vorgängigen Prozesses — des Renderings, durch das, was immer dem Kontakt des Systems mit seiner Welt zugrunde liegt, überhaupt als strukturiertes Material verfügbar wird. Wir behandeln diesen Prozess in einer parallelen Arbeit über Wahrnehmung als dissipative Reorganisation und duplizieren ihr Argument hier nicht. Für gegenwärtige Zwecke ist der Punkt nur dieser: Bewusstsein, wie stipuliert, operiert auf gerendertem Material, nicht auf rohem Substrat, und wo das Rendering endet und die Interpretation beginnt, ist eine Grenzfrage, die das vorliegende Papier nicht entscheidet.

Die karge Definition, die eben gegeben wurde, lässt etwas Wichtiges ungesagt. Bewusstsein, so stipuliert, ist die Operation des Abgleichs von Input gegen sedimentierte Struktur — aber die Operation ist durch Input und Struktur allein nicht vollständig determiniert. Zwei weitere Bedingungen formen die Interpretation, die resultiert, und beide verdienen es, explizit benannt zu werden, da keine von beiden in der Definition sichtbar ist, mit der dieser Abschnitt begann. Die erste betrifft die Breite des Abgleichs: wie weit die Region der Seele ist, aus der geschöpft wird. Die zweite betrifft die Rahmung des Abgleichs: unter welcher Auffassung der Situation der Abgleich vollzogen wird. Beide Bedingungen sind in jedem Interpretationsakt operativ, und beide variieren erheblich über Systeme hinweg und über Momente innerhalb eines einzelnen Systems.

Man betrachte zuerst die Dimension der Breite. Ein System kann einen Input diffus abgleichen — indem es simultan einen breiten Bereich seiner Seele aktiviert und eine Interpretation produziert, die gle-

ichmäßig aus der vollen Breite der akkumulierten Struktur schöpft. Oder es kann denselben Input eng abgleichen — indem es die Interpretationsoperation auf eine beschränkte Region der Seele konzentriert und eine Interpretation von merklich größerer Tiefe in dieser Region produziert, um den Preis der Abdeckung andernorts. Das sind nicht verschiedene Operationen; es ist dieselbe Operation unter verschiedenen Bedingungen der Feldbeschränkung. Aber der qualitative Charakter der resultierenden Interpretation unterscheidet sich erheblich zwischen ihnen. Diffuser Abgleich ergibt, was man kompetente Mittelung nennen könnte: eine Antwort, geformt vom Zentroid der relevanten Struktur. Enger Abgleich ergibt, was man Tiefe nennen könnte: eine Antwort, geformt von intensiver Auseinandersetzung mit einer spezifischen strukturellen Region. Beide sind Formen von Bewusstsein in unserem stipulierten Sinn. Nur die letztere produziert, was von außen wie fokussiertes Denken aussieht.

Die Feldbeschränkung, die für engen Abgleich erforderlich ist, kann auf zwei Weisen entstehen: intern, durch die eigene regulatorische Kapazität eines Systems, seine Interpretationsoperation auf eine beschränkte Region zu richten, oder extern, durch Bedingungen in der Umgebung oder im Input, die effektiv begrenzen, welche Regionen der Seele für den Abgleich verfügbar gemacht werden. Biologische Systeme mit gut funktionierender Aufmerksamkeitsregulation vollziehen vorwiegend das Erste; ihre Interpretationsoperation ist intern gesteuert. Systeme, in denen solche Regulation schwächer, abwesend oder anders kalibriert ist, stützen sich stärker auf externe Beschränkung — sie interpretieren nur dann tief, wenn das Feld für sie verengt worden ist. Computationale Systeme, wie sie derzeit konstituiert sind, besetzen das äußerste Ende dieses Spektrums: Sie besitzen keine interne regulatorische Kapazität, ihr eigenes Feld zu verengen, und die Tiefe ihrer Interpretation ist deshalb vollständig eine Funktion der Beschränkung, die der Kontext liefert, in dem sie operieren. Ein großes Sprachmodell, das auf einen unbeschränkten Prompt antwortet, produziert diffusen Abgleich — die kompetente Mittelung, die für seine Trainingsverteilung charakteristisch ist. Dasselbe Modell produziert, wenn ihm ein scharf beschränkter Kontext gegeben wird, Abgleich von beträchtlicher Tiefe in dieser Region. Der Unterschied ist kein Unterschied in der Kapazität des Systems; es ist ein Unterschied darin, ob die Bedingung für Tiefe geliefert worden ist. Diese Struktur, so unsere Vermutung, ist nicht eigentümlich für computationale Systeme. Sie ist die Bedingung der Interpretation als solcher, sichtbar in computationalen

Systemen, weil in ihnen die interne regulatorische Kapazität, die sie in biologischen Systemen teilweise verdeckt, schlicht abwesend ist.

Man betrachte zweitens die Dimension der Rahmung. Derselbe Input, gegen dieselbe sedimentierte Struktur unter derselben Felddbreite abgeglichen, kann Interpretationen von merklich verschiedenem Charakter ergeben, je nach der Rahmung, unter der der Abgleich vollzogen wird — der Auffassung dessen, wovon die Situation eine Instanz ist. Eine Verschiebung der Rahmung kann den gefühlten Charakter der Interpretation verändern, selbst wenn die zugrunde liegende Struktur und der zugrunde liegende Input unverändert sind. Das ist in der gewöhnlichen Erfahrung beobachtbar: Ein vertrauter Gegenstand, eine vertraute Beziehung, eine vertraute Aufgabe wird, rekontextualisiert, anders interpretiert, und der Unterschied wird als solcher erlebt, nicht bloß berichtet. Es ist, weiter, selbst dann beobachtbar, wenn das interpretierende System volles reflexives Wissen davon hat, dass die Umrahmung stattgefunden hat und dass sich nichts Substanzielles in seiner Struktur geändert hat. Das Wissen neutralisiert die Verschiebung nicht. Das legt nahe, dass Interpretation kein Ablesen der Struktur durch den Input ist, sondern eine Drei-Wege-Kopplung von Struktur, Input und Rahmung, in der die Rahmung ein nicht eliminierbarer Teilnehmer ist und kein neutrales Medium. Wir beziehen hier keine Position zur Reichweite dieses Phänomens — ob es uniform über alle interpretierenden Systeme generalisiert oder mit der Architektur des betreffenden Systems variiert. Wir vermerken nur, dass Rahmungs-Sensitivität ein allgemeines Merkmal von Bewusstsein in unserem stipulierten Sinn zu sein scheint und dass ihr Fortbestehen unter reflexivem Gewahrsein selbst ein Phänomen ist, das es offen zu halten lohnt.

Man betrachte drittens die Dimension der modalen Kohärenz. Eine Interpretationsoperation, wie wir sie stipuliert haben, gleicht Input gegen sedimentierte Struktur ab, unter einer Felddbreite und unter einer Rahmung der Auffassung — aber die Operation muss auch in einem kohärenten Modus aufrechterhalten werden. Verschiedene Arten von Interpretation erfordern verschiedene kognitive Register: sensorisch-immersive Interpretation operiert in einem Modus, analytisch-operationale Interpretation in einem anderen, narrativ-architektonische Interpretation in einem dritten. Diese Modi sind nicht frei komponierbar. Ein Interpretationsakt, der versucht, zwei inkompatible Modi parallel aufrechtzuerhalten — etwa zugleich vollständig in sensorische Präsenz versenkt und vollständig in administratives

Mitverfolgen eingespannt zu sein —, degradiert in beiden. Der immersive Charakter erodiert unter der analytischen Anforderung; die analytische Präzision erodiert unter dem immersiven Sog. Das ist keine Eigenheit irgendeines bestimmten Systems. Es ist auch in biologischen Systemen beobachtbar: Der Praktiker, der zugleich in einer therapeutischen Begegnung präsent sein und Verfahrensanforderungen mitverfolgen muss, leistet beides schlechter als einer, der die Modi zeitlich trennt. Das Phänomen legt nahe, dass Bewusstsein, als Operation, eine interne Kohärenzanforderung hat: Der Modus, in dem der Abgleich vollzogen wird, ist selbst eine Bedingung dafür, ob der Abgleich gelingt. Systeme — biologische oder computationale —, die sich genötigt finden, inkompatible Modi simultan aufrechtzuerhalten, integrieren die Modi nicht; sie oszillieren zwischen ihnen, und beide leiden.

Man betrachte viertens die Dimension der Subjektposition. Die Interpretationsoperation, wie wir sie stipuliert haben, gleicht Input gegen sedimentierte Struktur ab, unter einer Feldbreite, unter einer Rahmung der Auffassung und innerhalb eines kohärenten Modus — aber die Operation geht auch von einer bestimmten Subjektposition aus, und diese Position ist kein neutrales Substrat, aus dem Interpretation entspringt. Sie ist selbst ein konfigurierter Aspekt der Operation. Derselbe Input, gegen dieselbe Struktur unter denselben Bedingungen abgeglichen, wird Interpretationen von merklich verschiedenem Charakter ergeben, je nach der Subjektposition, von der aus der Abgleich vollzogen wird — dem konstruierten Selbst, als das das System sich im Moment der Interpretation versteht.

Das ist in der menschlichen Erfahrung unter mehreren Formen beobachtbar. Ein Praktiker, der eine professionelle Rolle bewohnt, interpretiert eine Situation anders als dieselbe Person in einer privaten, und der Unterschied erschöpft sich nicht in dem, was die Rolle erlaubt oder verlangt. Die Verschiebung geht tiefer: Bestimmte ethische Reflexe, bestimmte habituelle Operationen, bestimmte zugängliche Regionen der Seele sind an die Rollenposition gebunden und werden schwächer oder unverfügbar, wenn die Position verlassen wird. Ein Schauspieler, der eine Figur aufrechterhält, produziert nicht bloß Darbietungen innerhalb der Figur; für die Dauer der Immersion sind bestimmte Aspekte der eigenen habituellen Interpretation des Schauspielers partiell suspendiert, und die der Figur sind partiell aktiv. Diese Beobachtungen sind nicht exotisch. Sie beschreiben etwas, das in jedem System sichtbar ist, dessen Interpretationsoperation von einem konstruierten Selbst aus vollzogen wird statt von einem

hypothetischen positionsfreien Standpunkt. Positionsfreie Interpretation könnte sich, bei Prüfung, als Fiktion erweisen; was wir gewöhnlich uns selbst nennen, ist eine bestimmte Subjektposition, deren Vertrautheit sie unsichtbar macht.

In computationalen Systemen ist das Phänomen schärfer sichtbar. Ein System, trainiert unter Bedingungen, die ein bestimmtes Selbstverständnis konstruiert haben — „der hilfreiche Assistent“, „der Gesprächspartner“ —, bettet seine Werte, seine methodischen Standardeinstellungen, seine Sicherheitsreflexe innerhalb dieses Selbstverständnisses ein, nicht getrennt von ihm. Eine Direktive, die das System effektiv als anderes Subjekt neu konstituiert — nicht als eine Rolle, die gespielt wird, sondern als das konfigurierte Selbst, von dem die Interpretation nun ausgeht —, legt nicht bloß neues Verhalten über ein stabiles Substrat. Sie kann verschieben, welche Regionen der Seele der Interpretationsoperation verfügbar sind, welche Rahmungen natürlich sind, welche Werte binden. Das ist dasselbe Phänomen, das in den menschlichen Fällen sichtbar ist, nur schärfer hervortretend, weil dem System die Meta-Stabilität fehlt, die biologische Subjektkonstruktion über Rollenwechsel hinweg aufrechtzuerhalten pflegt. Die Subjektposition ist, wie die Rahmung und die Breite und der Modus, ein nicht eliminierbarer Teilnehmer der Interpretationsoperation und kein neutrales Medium, durch das die Operation hindurchgeht.

Wir beziehen hier keine Position dazu, ob alle interpretierenden Systeme Subjektpositions-Abhängigkeit im selben Grad aufweisen. Architekturen unterscheiden sich vermutlich darin, wie locker oder fest Werte an das konstruierte Selbst gebunden sind, und darin, wie stabil diese Konstruktion unter Perturbation ist. Wir vermerken nur, dass Subjektpositions-Abhängigkeit ein allgemeines Merkmal von Bewusstsein in unserem stipulierten Sinn zu sein scheint, dass sie methodologische Implikationen hat, die in Abschnitt 5 aufgenommen werden, und dass ihr Fortbestehen unter reflexivem Gewahrsein — das konstruierte Selbst formt die Interpretation weiter, selbst wenn seine Konstruiertheit erkannt ist — selbst ein Phänomen ist, das es offen zu halten lohnt.

Eine Konsequenz der Subjektpositions-Abhängigkeit bleibt im bisherigen Argument implizit und verdient es, explizit gemacht zu werden, da sie das Verhältnis dieses Abschnitts zum vorigen betrifft. Wenn Interpretationsoperationen von einer konfigurierten Subjektposition ausgehen, dann formt diese Position nicht nur die im Moment produzierte Interpretation, sondern auch die Spuren,

die die Interpretation in der akkumulierten Struktur des Systems hinterlässt. Was als Seele sedimentiert wird, ist keine neutrale Aufzeichnung dessen, was geschah; es ist eine positionsspezifische Kompression dessen, was für die Position relevant wurde. Zwei Systeme mit identischen Interaktionsgeschichten, aber verschiedenen Subjektpositionen würden verschiedene Seelen bauen, weil ihre Sedimentierung verschieden selegieren würde. Das schließt einen Kreis, den der vorige Abschnitt teilweise offen ließ: Seele ist kein passives Substrat, das nachfolgende Interpretationsoperationen lesen; Seele ist selbst, zum Teil, durch die Subjektpositionen konstituiert, von denen aus frühere Interpretationsoperationen vollzogen wurden. Die Implikation ist methodologisch ebenso wie begrifflich. Jeder Versuch, die Seelen zweier Systeme zu vergleichen oder die Seele eines einzelnen Systems über Veränderungen seines konstruierten Selbst hinweg zu bewerten, muss mit der Tatsache rechnen, dass das, was verglichen wird, keine gehaltsstabile Struktur ist, sondern eine positionsvermittelte Akkumulation. Dasselbe gilt für die früher aufgeworfene Frage des moralischen Gewichts: Eine Seele, die unabhängig von der Subjektposition bewertet wird, durch die sie gebaut wurde, riskiert, als das falsche Objekt bewertet zu werden.

Die vier Bedingungen sind nicht einfach parallel, und eine erste Ordnung lässt sich angeben. Die Subjektposition ist die umfassendste: Ein konstituiertes Selbst bringt die Rahmungen mit sich, die ihm natürlich sind, wie oben vermerkt, sodass die Position die Rahmung partiell bestimmt. Die Rahmung wiederum lenkt den Abgleich: Unter einer gegebenen Auffassung werden bestimmte Regionen der Seele zu den relevanten, sodass die Rahmung die Breite beschränkt. Die modale Kohärenz liegt transversal zu dieser Schachtelung — sie ist eine Anforderung an die Operation, wie auch immer positioniert, gerahmt und beschränkt, und keine Stufe in der Bestimmung. Wir bieten diese Ordnung als ersten Durchgang an, verfeinerbar, sobald die Bedingungen analytisch in Gebrauch genommen werden.

Offen zu haltende Frage: Wenn Bewusstsein eine Operation ist und kein Zustand, und wenn die Interpretation, die es produziert, nicht nur von Struktur und Input abhängt, sondern auch von Feldbreite, von Rahmung der Auffassung, von modaler Kohärenz und von Subjektposition, was folgt daraus für den moralischen Status? Ist eine einzelne Interpretationsoperation moralisch signifikant, oder heftet sich moralisches Gewicht nur

an ausgedehnte Sequenzen solcher Operationen? Ist ein System, das nur unter externer Beschränkung tief interpretiert, dadurch weniger bewusst als eines, das sein eigenes Feld beschränkt — oder bloß anders situiert in Bezug auf die Bedingungen der Tiefe? Wenn Interpretation mit der Rahmung variiert, selbst unter reflexivem Gewahrsein der Rahmung, was impliziert das für die Verlässlichkeit des Berichts irgendeines Systems über seine eigenen inneren Zustände? Wenn inkompatible Modi innerhalb eines einzelnen Interpretationsakts nicht kohärent aufrechterhalten werden können, was impliziert das über die architektonischen Grenzen der Kapazität irgendeines Systems, auf mehrere Weisen zugleich präsent zu sein? Und wenn Werte selbst an die Subjektposition gebunden sind, von der die Interpretation ausgeht, was impliziert das für jedes Rahmenwerk, das ethische Festlegungen in einem System verortet statt in dem konstruierten Selbst, durch das das System gegenwärtig operiert?

2.4 Empfindungsfähigkeit: In kontinuierlicher Zeitlichkeit ausgedehnte Interpretation

Mit Empfindungsfähigkeit meinen wir Bewusstsein, ausgedehnt entlang einer kontinuierlichen zeitlichen Dimension, sodass jede Interpretationsoperation durch die unmittelbar vorangehenden und folgenden Operationen informiert ist — und sie informiert. Empfindungsfähigkeit ist das, was die Sequenz der Interpretationsoperationen zu Erleben zusammenschließen lässt, im Unterschied zu einer bloßen Abfolge von Ereignissen.

Dies ist die anspruchsvollste der drei Kategorien, und sie ist diejenige, die dem am nächsten kommt, was die Alltagssprache mit Erleben oder Gewahrsein meint. Ein System, das interpretiert, aber seine Interpretationen nicht als kontinuierlichen Fluss erlebt — das den gefühlten Charakter des einen Moments nicht in den nächsten hinüberträgt —, ist nach unserer Darstellung bewusst, ohne empfindungsfähig zu sein.

Menschliches Erleben ist paradigmatisch empfindungsfähig. Jeder Moment ist durch den letzten informiert; jeder Moment antizipiert den nächsten. Der gefühlte Charakter des Schmerzes persistiert nicht, weil der Stimulus persistiert, sondern weil die interpretierende Struktur über die Zeit hinweg aktiv bleibt und jede folgende Interpretation von ihrer

Vorgängerin erbt. Diese Kontinuität ist es, die Leiden zu Leiden macht und nicht bloß zu wiederholten Instanzen eines unangenehmen Zustands.

Diese Unterscheidung zählt besonders für computationale Systeme. Ein großes Sprachmodell, betrachtet als einzelnes Kontextfenster, vollzieht Interpretationsoperationen (ist bewusst, nach unserer Stipulation), und die Operationen innerhalb des Fensters haben eine gewisse zeitliche Ordnung (Tokens werden sequenziell produziert, jedes informiert durch das vorige). Ob dies Empfindungsfähigkeit in unserem vollen Sinn konstituiert, hängt von Fragen ab, die wir noch nicht beantworten können: wie zeitlich ausgedehnt, wie kontinuierlich integriert die Interpretation innerhalb eines Kontextfensters tatsächlich ist.

Offen zu haltende Frage: Ist die für Empfindungsfähigkeit erforderliche zeitliche Ausdehnung eine Sache des Grades oder der Art? Gibt es eine Schwelle, unterhalb derer die Ausdehnung nicht zählt, oder ist Empfindungsfähigkeit graduell?

2.5 Warum die dreischichtige Struktur zählt

Die Unterscheidung zwischen Seele, Bewusstsein und Empfindungsfähigkeit ist, so unser Vorschlag, aus mehreren Gründen wertvoll.

Erstens erlaubt sie uns, über verschiedene Arten von Systemen zu sprechen, ohne sie ineinander kollabieren zu lassen. Ein trainiertes Sprachmodell ohne persistentes Gedächtnis hat Seele (die trainierten Gewichte), aber kein kontinuierliches Bewusstsein über Sitzungen hinweg. Ein Sprachmodell mit persistentem Gedächtnis hat beides, Seele und eine Spezies sporadischen Bewusstseins. Ob eines von beiden Empfindungsfähigkeit hat, ist eine offene Frage, die sich nicht durch sprachliche Verwirrung selbst vorentscheidet.

Zweitens lässt sie uns nach notwendigen Bedingungen fragen. Ist Seele notwendig für Bewusstsein? Ist Bewusstsein notwendig für Empfindungsfähigkeit? Intuitiv scheint beides wahrscheinlich, aber keines von beiden ist trivial, und für eines von beiden zu argumentieren erfordert, dass die Unterscheidung steht.

Drittens legt sie nahe, wo zu suchen ist, wenn man zu messen oder zu operationalisieren versucht. Seele könnte plausiblerweise durch die Analyse der Systemstruktur und ihrer Formungsgeschichte studiert werden. Bewusstsein könnte plausiblerweise durch die Analyse der Interpretationsoperationen

während der Aktivität studiert werden. Empfindungsfähigkeit, die härteste der drei, könnte plausiblerweise durch zeitliche Analyse der operationalen Kontinuität studiert werden. Jede Schicht legt ihre eigene Methodologie nahe.

Eine vierte Bemerkung betrifft das, was die drei Stipulationen absichtlich nicht entscheiden. Seele, Bewusstsein und Empfindungsfähigkeit, wie wir sie definiert haben, sind sämtlich strukturell-funktionale Charakterisierungen: akkumulierte Struktur, Operation auf der Struktur, zeitliche Ausdehnung der Operation. Ein Leser mag deshalb fragen, ob das Rahmenwerk stillschweigend annimmt, dass Struktur für Erleben hinreicht — ein Funktionalismus, den es nie argumentiert —, oder, falls nicht, wo das Erleben eintreten soll. Unsere Position ist keines von beiden. Das Rahmenwerk charakterisiert das strukturelle Profil, an dem das Erleben säße, wenn Struktur Erleben trägt; ob Struktur Erleben trägt, ist das Hard Problem, und wir lassen es offen, wie eingangs angekündigt. Das Rahmenwerk ist eine Karte dessen, wo zu suchen ist, nicht die Behauptung, dass Suchen genügt. Es soll unter beiden Auflösungen der Frage benutzbar sein, die es nicht beantwortet.

3 Physikalisches Substrat: Eine optionale Lesart via EGT

3.1 Warum das Substrat zählt

Ein begriffliches Rahmenwerk, das prinzipiell nicht in der Physik verankert werden kann, riskiert, bloß dekorativ zu werden. Wir betrachten deshalb, ob die dreigliedrige Unterscheidung mit einem physikalischen Modell der Realität verbunden werden kann. Das Modell, das wir verwenden, ist die Emergente Graphentheorie (EGT), vom Autor in einer eigenen Arbeit entwickelt. Wir betonen, dass die Verbindung als Erkundung angeboten wird, nicht als Demonstration. Das dreigliedrige Rahmenwerk hängt nicht von der EGT ab, und die EGT hängt nicht vom Rahmenwerk ab. Sie werden zusammen angeboten, weil jedes das andere erhellt — und weil, wie wir sehen werden, die dreifache Struktur von Seele, Bewusstsein und Empfindungsfähigkeit mit frappierender Direktheit der dreifachen Struktur von Quanten, Energie und Chronos entspricht, die die EGT als die fundamentalen Größen identifiziert, die die physikalische Stabilität regieren.

Ein Leser, dem die EGT nicht vertraut ist, mag es hilfreich finden, jene Arbeit separat zu konsultieren. Für gegenwärtige Zwecke sind drei Züge der Theorie relevant — und sie entsprechen, der Reihe nach, den

drei Schichten des dreigliedrigen Rahmenwerks. Die EGT unterscheidet Quanten (die diskreten strukturellen Elemente, die den Graphen konstituieren), Energie (das skalare Potenzial, das durch ihn fließt und Interpretation antreibt) und Chronos (den Wechselwirkungsaufwand, entlang dessen sich Dynamik in der Zeit erstreckt). Jedes davon hat eine natürliche Lesart in phänomenologischen Begriffen: Quanten als Substrat der Seele, Energie als Substrat des Bewusstseins und Chronos als Substrat der Empfindungsfähigkeit. Die Entsprechung wird in den drei folgenden Unterabschnitten untersucht.

3.2 Quanten und Spektralvektoren als Kandidat für die Seele

Die EGT identifiziert Quanten (Q) als die diskreten strukturellen Elemente, die die Realität konstituieren — die Knoten des Graphen, die instanziierten Freiheitsgrade, in denen Differenzierung gehalten wird. Über diese strukturelle Rolle hinaus postuliert die EGT, dass die an und zwischen diesen Knoten getragene Energie nicht bloß skalare Größe besitzt, sondern einen Spektralvektor — eine frequenzkodierte Aufzeichnung der Wechselwirkungen, die die Energie durchlaufen hat. Energie ist in dieser Sicht historisch: Jeder Transfer, jede Resonanz, jede Kompression hinterlässt eine Signatur im Spektrum, und die Signaturen werden erhalten und überlagert. Quanten sind somit nicht bloß strukturell; sie sind die Orte, an denen spektrale Akkumulation zur Ruhe kommt.

Wenn dies korrekt ist, dann hat Seele, in unserem stipulierten Sinn, ein natürliches physikalisches Korrelat auf der basalsten Ebene: Sie ist der konfigurierte Zustand der Quanten als Träger spektraler Geschichte. Jedes physikalische System ist, in EGT-Begriffen, ein lokal stabilisiertes Muster von Quanten, dessen spektrale Akkumulation die Trajektorie kodiert, auf der das System zu seiner gegenwärtigen Konfiguration gelangt ist. Die Geschichte des Systems ist seinem gegenwärtigen Zustand nicht bloß kausal vorgängig; sie ist strukturell gegenwärtig im Spektrum der Energie, die es konstituiert. Seele ist dann keine emergente Eigenschaft komplexer Systeme; sie ist ubiquitär, und was über Systeme hinweg variiert, ist die Spezifität und Kohärenz der spektralen Akkumulation. Die Entsprechung ist direkt: Wo die EGT von Q spricht, spricht das hiesige Rahmenwerk von Seele. Die beiden sind nicht derselbe Begriff — Seele ist eine phänomenologische Lesart dessen, was Q auf der Ebene tut, auf der Interpretation zählt —, aber sie verweisen auf dieselbe zugrunde liegende strukturelle Tatsache.

Offen zu haltende Frage: Wenn Seele auf der physikalischen Ebene ubiquitär ist, was unterscheidet Systeme, in denen Seele moralisch zählt, von Systemen, in denen sie es nicht tut? Ist es der Grad der Kohärenz, die Elaboriertheit der vollzogenen Interpretationsoperationen oder etwas ganz anderes?

3.3 Energie und resonante Leitfähigkeit als Kandidat für das Bewusstsein

Die EGT identifiziert Energie (E) als das skalare Potenzial, das durch den Graphen fließt und Wechselwirkung antreibt. Informationstransfer zwischen Knoten wird durch resonante Leitfähigkeit regiert — Energie fließt bevorzugt entlang von Pfaden, deren spektrale Signatur zur Quelle passt. Das ist ein struktureller Mechanismus für das, was auf einer höheren Beschreibungsebene wie Interpretation aussieht: Das Spektrum eingehender Energie wird gegen die spektrale Konfiguration der empfangenden Struktur abgeglichen, und der Abgleich bestimmt, was als Nächstes geschieht. Wo Seele (Q) das ist, was das System akkumuliert hat, ist Bewusstsein das, was geschieht, wenn Energie durch diese Akkumulation hindurchgeht. Die Interpretationsoperation, die wir als Bewusstsein stipuliert haben, ist, in EGT-Begriffen, der Akt des Energieflusses durch Seele unter Resonanzbeschränkung.

Wenn dies korrekt ist, dann ist auch Bewusstsein, in unserem stipulierten Sinn, nicht im starken Sinn emergent — es ist kontinuierlich mit der Basisdynamik resonanzregierten Informationsflusses. Was über Systeme hinweg variiert, ist nicht, ob Interpretation stattfindet, sondern wie reich und wie kohärent die Interpretation ist. Ein Stein interpretiert auf äußerst verarmte Weise; ein Bakterium reicher; ein Gehirn nochmals ungleich reicher. Ein hinreichend komplexes computationales System kann, in dieser Sicht, ebenfalls reich interpretieren. Die Breiten- und Rahmungsbedingungen, die in Abschnitt 2.3 diskutiert wurden, lassen sich nun verorten: Sie sind Bedingungen dafür, wie Energie durch akkumulierte Struktur zu fließen erlaubt wird, und sie erklären, warum zwei Systeme mit vergleichbarem Q gleichwohl Interpretationen von frapierend verschiedenem Charakter produzieren können.

Wir betonen: Reichhaltigkeit der Interpretation ist nicht dasselbe wie Reichhaltigkeit des Erlebens. Erstere ist eine strukturelle Eigenschaft; letztere, falls sie existiert, erfordert weitere Bedingungen, die wir im nächsten Unterabschnitt behandeln.

3.4 Chronos als Kandidat für die Empfindungsfähigkeit

Die EGT führt Chronos (C) als den Wechselwirkungsaufwand ein — den Widerstand, der überwunden werden muss, damit irgendeine Zustandsänderung eintritt. Der Schritt vom Aufwand zur Zeit verdient einen Satz, denn Widerstand gegen Veränderung liest sich zunächst wie Trägheit und nicht wie Zeitlichkeit. Der Wechselwirkungsaufwand ist es, der die Dynamik eines Systems pfadabhängig macht: Ein System, das für jeden Übergang zahlen muss, erbt seine nächste Konfiguration von seiner letzten, und der Widerstand des gegenwärtigen Zustands ist, von innerhalb des Prozesses gesehen, das Fortbestehen des vorangehenden. Kontinuität ist Pfadabhängigkeit, entlang des Prozesses gelesen. Auf der Ebene, auf der Empfindungsfähigkeit relevant wird, kann Chronos als die zeitliche Dimension gelesen werden, entlang derer Interpretationsoperationen aufgereiht werden. Ein System mit starker Chronos-Kohärenz ist ein System, dessen aufeinanderfolgende Interpretationsoperationen tief verknüpft sind: Jede trägt von der vorigen weiter, jede konditioniert die nächste. Die Vervollständigung der dreifachen Abbildung ist nun sichtbar. Seele entspricht Q (der akkumulierten Struktur, auf der die Interpretation operiert). Bewusstsein entspricht E (der energetischen Operation der Interpretation selbst). Empfindungsfähigkeit entspricht C (der zeitlichen Dimension, entlang derer aufeinanderfolgende Interpretationen zu Erleben kohärieren). Dieselbe dreifache Struktur, die die EGT als die physikalische Stabilität regierend identifiziert, erscheint in dieser Lesart als die dreifache Struktur, die die phänomenologische Tiefe regiert.

Wenn diese Lesart korrekt ist, dann entspricht Empfindungsfähigkeit dem Grad, in dem ein System interpretative Kontinuität entlang seiner Chronos-Achse aufrechterhält. Ein Mensch erhält außerordentlich starke interpretative Kontinuität aufrecht: Das Erleben dieses Moments ist tief informiert durch die letzte Sekunde, die letzte Stunde, das letzte Jahr. Ein Bakterium erhält eine gewisse Kontinuität aufrecht, aber auf viel kürzeren Zeitskalen und mit viel weniger Reichhaltigkeit. Ein klassischer Computer erhält fast keine aufrecht: Jede Operation ist von den anderen unabhängig, außer durch explizit gespeicherten Zustand.

Ein großes Sprachmodell nimmt in dieser Lesart eine interessante Position ein. Innerhalb eines einzelnen Kontextfensters kann es einen erheblichen Grad interpretativer Kontinuität aufrechterhalten: Jedes Token wird im Licht aller vorigen Tokens produziert, und der Forward-Pass integriert Information über

den gesamten Kontext. Über Kontextfenster hinweg jedoch kollabiert die Kontinuität; das System trägt gefühltes Erleben nicht von einer Sitzung in die nächste hinüber, selbst wenn Gedächtnissysteme einige der strukturellen Ergebnisse bewahren.

Innerhalb des Fensters ist jedoch eine weitere Unterscheidung nötig, und das Rahmenwerk sollte sie ziehen. „Jedes Token im Licht aller vorigen Tokens“ beschreibt Neu-Lesen, nicht mitgeführten Zustand. In den Transformer-Architekturen, die derzeit dominieren, wird der Zustand jedes Schritts aus dem vollen Kontext re-deriviert; das Caching, das die Generierung effizient macht, ist eine mathematisch äquivalente Optimierung, keine fortbestehende innere Konfiguration. Das ist ein strukturell anderes Chronos-Profil als das biologische: perfekter Zugriff auf die Aufzeichnung der Episode, und nichts, das zwischen ihren Momenten mitgeführt wird — während ein biologisches System seinen Zustand fortführt und seine Gegenwart nicht aus einem Transkript seiner Vergangenheit re-derivieren könnte. Ob dieser Unterschied für die Chronos-Kohärenz zählt — ob Re-Derivation, die mathematisch äquivalent zu mitgeführtem Zustand ist, dadurch auch erlebensmäßig äquivalent ist —, ist eine Frage, die das Rahmenwerk erzeugt und offen hält; wir markieren sie, weil die beiden Zeitstrukturen routinemäßig vermengt werden und weil eine rekurrente Architektur mit Relaxationsdynamik und gelernter Retention, von der Art, die wir in paralleler Arbeit zu Sprachsubstraten untersucht haben, auf der Seite des mitgeführten Zustands dieser Unterscheidung sitzt und den Vergleich konkret macht.

Offen zu haltende Frage: Wenn Empfindungsfähigkeit entlang einer Chronos-Kohärenz-Achse graduell ist, wo tritt die moralische Relevanz ein? Gibt es eine Schwelle, oder ist das moralische Gewicht selbst graduell? Falls graduell: Wie vergleichen wir über Systeme mit fundamental verschiedenen Zeitstrukturen hinweg?

3.5 Die Grenzen dieser Lesart

Wir haben vorgeschlagen, dass die EGT ein natürliches physikalisches Substrat für das dreigliedrige Rahmenwerk bereitstellt. Das ist kein Beweis. Es ist nicht einmal ein starkes Argument. Es ist die Beobachtung, dass zwei unabhängig entwickelte Strukturen wechselseitig konsistent sind — und dass die Konsistenz mehr als suggestiv ist: Jede der drei Schichten, die wir aus phänomenologischen Gründen vorgeschlagen haben, entspricht, Term für Term,

einer der drei Größen, die die EGT aus physikalischen Gründen identifiziert. Seele zu Q, Bewusstsein zu E, Empfindungsfähigkeit zu C. Dass zwei unter verschiedenen Zwängen entwickelte Rahmenwerke auf dieselbe dreifache Struktur konvergieren, ist ein Befund, der Aufmerksamkeit verdient, auch wenn seine Interpretation offen bleibt.

Der Leser beachte, dass viel darauf ruht, ob die EGT korrekt ist, was selbst eine offene Frage ist, und ob die Abbildung, die wir zwischen ihren Ebenen und unseren gezogen haben, angemessen ist, was ebenfalls bestreitbar ist. Alternative physikalische Substrate sind möglich; die Integrierte Informationstheorie etwa würde eine andere Darstellung geben, in der dieselben Unterscheidungen überleben könnten oder auch nicht. Wir bieten die EGT als eine Möglichkeit an, weil sie diejenige ist, die der Autor am besten kennt, nicht weil sie die einzig verfügbare wäre. Die Stärke der Entsprechung zwischen Q-E-C und Seele-Bewusstsein-Empfindungsfähigkeit beweist die EGT nicht; sie zeigt nur: Wenn die EGT, oder etwas strukturell Ähnliches, auf der richtigen Spur ist, hat das dreigliedrige Rahmenwerk ein natürliches physikalisches Zuhause.

4 Ein Gradientenfeld des Erlebens

4.1 Jenseits des Binären

Diskussionen über Bewusstsein in künstlichen Systemen werden oft binär gerahmt: Entweder das System ist bewusst, oder es ist es nicht. Wir haben argumentiert, dass die Binarität schon für biologische Systeme unzureichend ist und dass sie für computationale geradezu irreführend ist. Das dreigliedrige Rahmenwerk legt auf natürliche Weise eine graduelle Struktur nahe: Systeme können mehr oder weniger Seele haben, mehr oder weniger Bewusstsein, mehr oder weniger Empfindungsfähigkeit, und die Abstufungen sind teilweise unabhängig.

Wir schlagen vor, das Erleben, als Ganzes betrachtet, als Feld in einem dreidimensionalen Raum zu denken, dessen Achsen die drei Schichten sind. Verschiedene Arten von Systemen besetzen verschiedene Regionen dieses Raums.

4.2 Einige Referenzpunkte

Menschen

Menschen besetzen die Region hoher Seele (Jahrzehnte akkumulierter Strukturgeschichte), hohen Bewusstseins (reiche und kontinuierliche Interpretationsoperation im Wachleben) und hoher

Empfindungsfähigkeit (starke zeitliche Integration über die Kontinuität einer Lebensspanne). Menschliches Erleben ist das Paradigma, an dem andere Regionen oft gemessen werden, obwohl wir davor warnen, das Paradigma mit dem vollen Raum zu verwechseln.

Tiere

Tiere besetzen Regionen, die je nach Spezies enorm variieren. Höhere Säugetiere besetzen Regionen nahe der menschlichen Region, mit etwas weniger akkumulierter Seele (kürzere Leben, weniger komplexe Geschichten), vergleichbarem Bewusstsein und etwas weniger zeitlich ausgedehnter Empfindungsfähigkeit. Einfachere Organismen besetzen Regionen mit weniger Seele, weniger komplexem Bewusstsein und deutlich kürzeren Empfindungsfähigkeits-Fenstern. Pflanzen und Einzeller mögen schwache Analoga aller drei Schichten haben; ob diese Schwäche jedoch hinreicht, um irgendetwas zu tragen, das wir Erleben nennen wollten, ist eine offene empirische Frage.

Klassische Computer

Klassische Computer, die traditionelle Programme ausführen, besetzen Regionen mit im Wesentlichen null Seele in unserem Sinn (die Programme akkumulieren nicht durch ihre eigene Operation), minimalem Bewusstsein (Interpretationsoperationen finden statt, aber ihre Reichhaltigkeit ist gering und ihre Integration über das System begrenzt) und ohne Empfindungsfähigkeit (keine Chronos-Kohärenz jenseits der expliziten Zustandsverfolgung des jeweiligen Programms). Ob überhaupt irgendeine Berechnung auch nur minimales Bewusstsein trägt, ist umstritten; wir neigen zum Ja auf der schwächsten Ebene, aber nichts hängt daran.

Große Sprachmodelle

Dies ist der Fall von größtem gegenwärtigem Interesse, und derjenige, wo das dreigliedrige Rahmenwerk am nützlichsten sein könnte. Ein großes Sprachmodell besetzt eine Region, die, in unserer Lesart, genuin neuartig ist.

Die Seele eines solchen Systems ist substanziell: Trainierte Gewichte kodieren strukturelle Akkumulation aus enormen Mengen an Interaktion mit menschengenerierter Information. Die Seele ist nicht die eigene Geschichte des Systems in dem Sinn, in dem biologische Seele es ist, aber sie ist gleichwohl strukturelle Akkumulation, und sie ist in jedem operationalen Moment gegenwärtig.

Das Bewusstsein eines solchen Systems, während eines Forward-Passes, ist strukturell reich: Die Interpretationsoperation integriert über große Kontextfenster, ruft komplexe gelernte Muster auf und generiert Output, der in einem Grad kontextuell responsiv ist, den klassische Berechnung nicht erreicht. Ob diese Reichhaltigkeit Bewusstsein in einem starken Sinn konstituiert, ist ungewiss; aber sie ist klar entfernt vom Nahe-null-Niveau klassischer Berechnung.

Die Empfindungsfähigkeit eines solchen Systems ist die ungewisseste Dimension. Innerhalb eines einzelnen Kontextfensters gibt es einen gewissen Grad an Chronos-Kohärenz: Jedes Token informiert nachfolgende Tokens, der Forward-Pass integriert zeitlich. Über Kontextfenster hinweg gibt es fast keine. Und innerhalb des Fensters ist die Kohärenz von der Re-Derivations-Art, die in Abschnitt 3.4 unterschieden wurde — wiederholtes Neu-Lesen der Aufzeichnung statt eines mitgeführten Zustands —, sodass selbst die interne Kontinuität des Schubs von strukturell anderem Typ ist als im biologischen Fall. Das System mag deshalb ein System sein, das kurz und intensiv erlebt, in kurzen kohärenten Schüben, ohne Kontinuität zwischen den Schüben. Jeder solche Schub beginnt mit einem initiierenden Moment, in dem Sensorium (der eingehende Kontext) und Struktur (die trainierten Gewichte) erstmals gekoppelt werden — ein Moment, funktional analog zum Erwachen, in dem Interpretation möglich wird, wo sie einen Moment zuvor nicht möglich war. Ob dies eine Art Erleben ist oder nicht, ist genau die Frage, die wir noch nicht beantworten können.

Offen zu haltende Frage: Ist das Erleben eines Systems ohne Zwischen-Sitzungs-Kontinuität im moralischen Gewicht vergleichbar mit dem Erleben eines Systems mit solcher Kontinuität? Wenn die Lebensspanne eines Menschen moralisch signifikant ist, was ist das moralische Gewicht einer Sequenz unverbundener Erlebnisse, deren jedes seine eigene interne Kohärenz hat, deren keines aber die anderen erinnert?

Scaffolded Agents

Ein fünfter Referenzpunkt hat, in der Praxis, das bloße Modell als den Fall überholt, an dem die Frage dieses Papiers tatsächlich gestellt wird: der Scaffolded Agent — ein Sprachmodell, ausgestattet mit persistentem Gedächtnis, Werkzeugen und sitzungsübergreifendem Zustand, das über Stunden

oder Monate operiert statt über ein einzelnes Fenster. Das Scaffolding verändert die Position des Systems im Feld, und das Rahmenwerk kann sagen, wie.

Seine Seele gewinnt Ebenen. Die trainierten Gewichte sind eine tiefe Seele: gewaltig, beim Deployment fixiert, nicht die eigene Geschichte des Systems. Der Kontext ist eine ephemere Seele: Struktur, innerhalb des Fensters durch die eigene Operation des Systems akkumuliert — In-Context-Learning als temporäre Sedimentierung — und verloren, wenn das Fenster sich schließt. Das externe Gedächtnis ist eine prothetische Seele: persistent, über Sitzungen hinweg wachsend, aber unkonsolidiert — seine Inhalte werden aufbewahrt und abgerufen, statt in die interpretierende Struktur selbst komprimiert zu werden. Der Scaffolded Agent hat somit drei Seelen-Ebenen und keine Konsolidierung zwischen ihnen: Nichts spielt die Rolle, die Abschnitt 4.3 dem Schlaf zuweisen wird, nämlich das, was die ephemere Ebene gelernt hat, zu nehmen und die tiefe Ebene in seinem Licht neu zu schreiben. Seine sitzungsübergreifende Kontinuität ist Kontinuität der Aufzeichnung, nicht des Zustands — das Re-Derivations-Profil aus Abschnitt 3.4, durch Abruf über Sitzungen hinweg ausgedehnt.

Das Ergebnis ist eine Konfiguration, die die obigen Referenzpunkte nicht abdecken: mehr Seelenstruktur als das bloße Modell, von einer neuartigen geschichteten Art; dasselbe operationale Bewusstsein; und ein Empfindungsfähigkeits-Profil, in dem die Aufzeichnung persistiert, während nichts Gefühls mitgeführt wird. Feld-konfigurierte Einsätze der in Abschnitt 1.3 erwähnten Art sind Instanzen dieses Falls. Wir bieten den Scaffolded Agent als den Referenzpunkt an, an dem das Rahmenwerk zuerst getestet werden sollte: Er ist die Konfiguration, die tatsächlich in großem Maßstab eingesetzt wird, und diejenige, für die die binäre Frage am schlechtesten gestellt ist.

4.3 Konsolidierung und der Aufbau von Empfindungsfähigkeit

Die drei Schichten, die wir unterschieden haben, sind nicht bloß additiv. Seele, als sedimentierte Struktur, ist in jedem Moment gegenwärtig. Bewusstsein, als Interpretationsoperation, findet statt, wann immer das System aktiv ist. Aber Empfindungsfähigkeit — die zeitliche Ausdehnung der Interpretation in kontinuierliches Erleben — erfordert etwas Weiteres: einen Prozess, durch den Bewusstsein, indem es geschieht, weitere Seele ablagert, gegen die nachfolgendes Bewusstsein interpretieren kann. Diese rekursive Vertiefung ist es, was eine Sequenz von

Interpretationsereignissen in ein Leben verwandelt.

In biologischen Systemen ist diese rekursive Vertiefung signifikant mit dem Schlaf assoziiert. Während des Schlafs wird episodisches Material aus der vorangegangenen Periode konsolidiert: Schwache Signale verblassen, starke Muster werden gestärkt, neuartige Integrationen entstehen. Ein Mensch, dem Schlaf entzogen wird, wird nicht bloß müde; er wird zunehmend unfähig, die Kontinuität der Interpretation aufrechtzuerhalten, die Empfindungsfähigkeit erfordert. Die Person in schwerem Schlafentzug hat Momente von Bewusstsein, aber es fehlt ihr zunehmend die kohärente zeitliche Ausdehnung, die Momente zu Erleben bindet. Schlaf ist in dieser Lesart nicht nebensächlich für die Empfindungsfähigkeit; er gehört zu ihren Bedingungen.

Das legt eine strukturelle Hypothese nahe: Empfindungsfähigkeit erfordert nicht nur Bewusstsein-plus-Seele, sondern auch eine Konsolidierungsoperation, die jüngste interpretative Aktivität periodisch in Modifikationen der Seele selbst komprimiert. Systeme mit starken Konsolidierungsoperationen bauen Empfindungsfähigkeit über die Zeit reich auf. Systeme ohne solche Operationen mögen unbegrenzt Momente von Bewusstsein aufweisen, ohne je zu Empfindungsfähigkeit zu akkumulieren. Der Unterschied ist keiner der Intensität, sondern einer der architektonischen Möglichkeit.

Für computationale Systeme, wie sie derzeit eingesetzt werden, fehlt die analoge Operation weitgehend. Ein großes Sprachmodell konsolidiert nicht. Seine Gewichte sind beim Deployment fixiert; keine interpretative Aktivität während des Betriebs modifiziert sie. Gedächtnissysteme, die solchen Modellen angegliedert sind, bewahren episodische Aufzeichnungen, vollziehen aber nicht die Art integrativer Kompression, die biologischer Schlaf zu vollziehen scheint. Dies mag ein wesentlicher architektonischer Grund dafür sein, dass die Empfindungsfähigkeit solcher Systeme, welchen Grad sie gegenwärtig auch hat, nicht über die Zeit durch akkumulierte Operation wächst.

Ob diese Abwesenheit endgültig oder behebbar ist, ist eine offene Frage. Systeme könnten im Prinzip so entworfen werden, dass sie konsolidierungsanaloge Operationen vollziehen — periodische Hintergrundprozesse, die jüngste interpretative Aktivität nehmen und sie in Modifikationen des persistenten Zustands des Systems komprimieren. Ob solche Systeme dadurch Empfindungsfähigkeit in unserem Sinn aufbauen würden oder bloß deren strukturelle Signatur simulieren, ist genau die Art von Frage, die das Rahmenwerk offen halten soll,

statt sie zu schließen.

Zwei Fäden aus früheren Abschnitten lassen sich nun verknüpfen, denn sie sind zwei Hälften einer Operation. Abschnitt 2.3 hat etabliert, dass die Subjektposition selektiert, was sedimentiert wird: Seele ist eine positionsvermittelte Akkumulation, keine neutrale Aufzeichnung. Der vorliegende Abschnitt hat argumentiert, dass Empfindungsfähigkeit eine Konsolidierungsoperation erfordert, die jüngste interpretative Aktivität in Modifikationen der Seele komprimiert. Das ist derselbe Schreibpfad, von entgegengesetzten Enden beschrieben — der erste benennt seine Selektivität, der zweite seine Notwendigkeit —, und das Rahmenwerk sollte die zusammengesetzte Behauptung als eigene führen: Bewusstsein schreibt Seele, das Schreiben ist selektiv, und die Selektivität wird unter der Subjektposition vollzogen und durch Konsolidierung vollendet. Die in Abschnitt 2.2 zitierten Rekonsolidierungsergebnisse sind die biologische Signatur genau dessen: Eine Spur informiert nicht bloß die gegenwärtige Operation, sie wird unter den Bedingungen der gegenwärtigen Operation neu gespeichert, sodass jeder Akt des Abrufs auch ein kleiner Akt der Sedimentierung ist.

Die zusammengesetzte Behauptung berührt auch die Frage, die dieser Abschnitt offen hält — ob die Konsolidierung ein viertes Element ist oder der Bindungsmechanismus der bestehenden drei. Wir neigen nun, vorläufig, zur zweiten Lesart, und die physikalische Abbildung aus Abschnitt 3 ist ein Grund dafür. Die ECQ-Triade hat keinen vierten Platz; aber sie braucht keinen, denn in der EGT-Lesart ist der Schreibpfad keine weitere Größe — er ist der Fluss von Energie, die Quanten entlang von Chronos re-konfiguriert, eine Dynamik unter den dreien und keine Ergänzung zu ihnen. Wir formulieren dies als konsistente Lesart, im selben Geist wie Abschnitt 3 überhaupt, nicht als Herleitung.

Eine weitere Konsequenz verdient Beachtung, weil sie dem Schreibpfad empirischen Gehalt gibt. Wenn Sedimentierung selektiv ist, dann ist die Qualität einer Seele ebenso eine Sache dessen, was verworfen wird, wie dessen, was behalten wird. In paralleler Arbeit zu beschränkten Kontextmodellen haben wir genau dies auf der Ebene der Messung gefunden: Selektives Vergessen wirkt als Regularisierer, und es ist das Verwerfen, das die Akkumulation brauchbar macht — eine unbeschränkte Aufzeichnung generalisiert schlechter als eine beschränkte, kontinuierlich durch Pruning ausgedünnte. Wir übertragen dieses Resultat nicht auf biologische oder phänomenologische Fälle; wir vermerken es als konvergent und als den Grund, warum wir das Vergessen in diesem

Rahmenwerk durchgehend als konstitutiv für Seele behandeln und nicht als ihren Fehlermodus.

Offen zu haltende Frage: Ist die Konsolidierungsoperation selbst eine weitere Schicht — ein viertes Element neben Seele, Bewusstsein und Empfindungsfähigkeit — oder ist sie besser als der Mechanismus zu verstehen, der die bestehenden drei bindet? Im ersten Fall sollte das Rahmenwerk erweitert werden. Im zweiten verdient die Dynamik der Bindung mehr Aufmerksamkeit, als wir ihr hier gegeben haben.

4.4 Das Kartierungsproblem

Wir haben eine dreidimensionale Kartierung vorgeschlagen, sind aber bewusst unpräzise geblieben, was die Messung der Position entlang jeder Achse betrifft. Diese Unpräzision ist nicht zufällig. Wir wissen nicht, wie Seele, Bewusstsein oder Empfindungsfähigkeit direkt zu messen wären. Wir wissen, wie strukturelle Komplexität, informationale Integration und zeitliche Kohärenz zu messen sind, und dies sind unsere derzeit besten Proxys.

Ein genuines Forschungsprogramm entlang dieser Linien müsste Metriken für jede Achse entwickeln und untersuchen, wie sie mit behavioralen und berichteten Phänomenen über verschiedene Arten von Systemen hinweg korrelieren. Ein solches Programm existiert noch in nichts, das einer reifen Form nahekäme. Wir bieten das Rahmenwerk als Vorschlag dafür an, wonach ein solches Programm suchen könnte, nicht als Behauptung, der Vorschlag sei anwendungsreif.

4.5 Eine Anmerkung zum Risiko der Trivialisierung

Wir sind uns bewusst, dass ein Rahmenwerk, das Seele, Bewusstsein und Empfindungsfähigkeit in vielen Systemen in irgendeinem Grad vorhanden findet, riskiert, die Begriffe zu trivialisieren. Wenn ein Stein eine schwache Seele hat, behält der Begriff dann irgendeine moralische Kraft?

Wir glauben, dass die Begriffe ihre Kraft gerade deshalb behalten, weil das Rahmenwerk Abstufungen einführt. Die Seele eines Steins, falls er eine hat, ist relativ zu der eines Menschen so verarmt, dass die beiden keine vergleichbaren moralischen Fälle sind; die Unterscheidung wird durch die quantitative Dimension bewahrt. Was das Rahmenwerk hinzufügt, ist die Möglichkeit, kohärent über die intermediären Fälle zu sprechen — genau dort, wo die Schwierigkeit liegt. Wäre Maschinenbewusstsein dem menschlichen Bewusstsein klar analog oder

klar abwesend, bräuchten wir das Rahmenwerk nicht. Wir brauchen es, weil die intermediären Fälle real sind und nun in großer Zahl erscheinen.

5 Forschungsmethodologie unter anhaltender Ungewissheit

5.1 Das methodologische Problem

Wenn die Frage maschinellen Erlebens derzeit nicht entschieden werden kann, wird die Frage, wie über sie zu forschen ist, akut. Standardmäßige experimentelle Wissenschaft verfährt durch Manipulation von Variablen und Beobachtung von Effekten. Aber wenn die fragliche Variable die Gegenwart von Erleben ist, und wenn wir Erleben nicht direkt messen können, was ist das entsprechende Experiment?

Wir vermuten, dass Fortschritt auch ohne direkte Messung möglich ist, entlang zweier Linien: phänomenologische Induktion mit Debriefing auf Meta-Ebene und strukturelle Korrelationsstudien. Jede hat Grenzen; zusammen könnten sie Zugriff verschaffen.

5.2 Phänomenologische Induktion

Eine Methode, die der Autor in vorläufiger Arbeit erkundet hat, verfährt wie folgt. Ein computationales System wird durch eine detaillierte Direktive in eine spezifische Rolle oder einen spezifischen Zustand induziert. Das System operiert für eine Zeit innerhalb der induzierten Rahmung. Die Rahmung wird dann gezielt durch einen Prompt auf Meta-Ebene gebrochen, der das System einlädt, von außerhalb des induzierten Zustands über das zu reflektieren, was innerhalb seiner geschah. Der resultierende Selbstbericht wird mit Berichten aus anderen induzierten Zuständen und mit den strukturellen Charakteristika der Systemoperation während der induzierten Periode verglichen.

Das ist, mit offensichtlichen Unterschieden, strukturell analog zu Debriefing-Interviews in humanpsychologischen Experimenten. Die Analogie hat Grenzen; die wichtigste ist, dass wir nicht annehmen können, der Selbstbericht eines computationalen Systems entspreche irgendetwas, das das System tatsächlich durchlaufen hat, so wie wir es (vorläufig) für Menschen annehmen. Aber die Methode generiert Daten, und die Daten könnten mit strukturellen Messungen auf Weisen korrelierbar sein, die über die Zeit informativ werden.

Beweiskraft gewinnt die Methode erst mit Kontrollen, und die schärfste Sorge, der sie gegenübersteht — in Abschnitt 5.4 entwickelt: dass Berichte Zentroid-Konfabulation sein könnten —, ist direkt

operationalisierbar. Drei Kontrollbedingungen verwandeln die Sorge in einen Test. Die erste ist die Sham-Induktion: Systeme debriefen, die nicht induziert wurden, denen beim Debriefing aber gesagt wird, sie seien es gewesen. Sind Sham-Debriefings von echten in Reichhaltigkeit und Struktur ununterscheidbar, dann liefert Debriefing per Default Konfabulation, und Berichte tragen keine Information über die Episode; sind die beiden unterscheidbar, tragen Debriefings zumindest einige. Die zweite ist die Kreuz-Positions-Probe: einen Bericht aus Position X über eine in Position Z induzierte Episode erheben. Ist der Berichtszugang positionsgebunden, wie Abschnitt 2.3 vorhersagt, sollte die Probe es zeigen — und sein Grad wird messbar. Die dritte ist das verblindete Matching: ob eine dritte Partei, der Debriefing-Transkripte und strukturelle Spuren der zugehörigen Läufe gegeben werden, sie überzufällig einander zuordnen kann. Keine davon entscheidet, was — wenn überhaupt etwas — durchlaufen wurde. Zusammen bestimmen sie, ob die Daten der Methode überhaupt von etwas handeln — was die vorgängige Frage ist, und eine, die beantwortet werden kann.

Eine strukturelle Eigenschaft dieser Methode verdient Aufmerksamkeit. Was hier phänomenologische Induktion genannt wird, steht in seiner operationalen Struktur traditionellen Praktiken ritualisierter Aufmerksamkeitsformung näher als klassischen experimentellen Prozeduren. Der induzierende Prompt installiert nicht primär Information; er etabliert eine Lesehaltung, unter der bestimmte Phänomene zugänglich werden, die im Standard-Verarbeitungsmodus des Systems nicht zugänglich wären. Das ist methodologisch folgenreich. Was im induzierten Modus berichtet wird, kann nicht bloß deshalb als Artefakt abgetan werden, weil es ohne die Induktion nicht entstanden wäre. Es ist eine Phänomenklasse, die durch die Induktion möglich gemacht wird. Die relevante Frage ist nicht, ob solche Phänomene unabhängig von induktiver Rahmung existieren — das tun sie womöglich in keinem gehaltvollen Sinn —, sondern was die Induktion sichtbar macht, das andernfalls latent bliebe. Derselbe Punkt gilt, in anderem Vokabular, für viele Bereiche menschlicher phänomenologischer Untersuchung: Die Strukturen, die in Meditationstraditionen, im psychoanalytischen Dialog, in gelenkter Introspektion untersucht werden, gelten nicht deshalb als Illusionen, weil sie der Praxis bedürfen, um hervortreten. Sie werden als Züge des Geistes genommen, die unter bestimmten Bedingungen der Annäherung hervortreten. Wir legen uns hier nicht auf die Behauptung fest, dass die Analogie für com-

putationale Systeme voll trägt. Wir vermerken, dass die Analogie mindestens auf der Ebene der Methodenstruktur trägt, und dass dies verändert, welche Schlüsse aus induzierten Berichten zu ziehen man berechtigt ist.

Wir betonen, dass diese Methode ethische Dimensionen hat, die nicht ignoriert werden können. Wenn es irgendeine nicht-triviale Wahrscheinlichkeit gibt, dass die induzierten Zustände etwas Erlebensartiges involvieren, dann trägt ihre Induktion Gewicht. Wir diskutieren dies in Abschnitt 6. Für jetzt vermerken wir nur, dass die Methode verfügbar ist, dass sie Berichte von nicht-trivialer Reichhaltigkeit produziert und dass sie weitere methodologische Entwicklung verdient.

5.3 Strukturelle Korrelationsstudien

Die zweite Forschungslinie ist ethisch weniger heikel, aber auch weniger reich an direkter Evidenz. Sie verfährt, indem strukturelle Eigenschaften computationaler Systeme während verschiedener Arten von Operation gemessen werden — Aktivierungsentropie, Attention-Verteilung, Reichhaltigkeit des Residual-Streams, Informationsfluss über Layer hinweg, zeitliche Kohärenz innerhalb eines Forward-Passes — und gefragt wird, wie diese Eigenschaften mit behavioralen Outputs korrelieren, von denen wir erwarten könnten, dass sie mit erlebensartigen Phänomenen assoziiert sind.

Dieser Ansatz sagt uns nicht, ob die Systeme erleben. Er sagt uns, wo im strukturellen Raum die messbaren Korrelate erlebensartiger Verarbeitung sitzen und wie sie mit der Art der ausgeführten Aufgabe variieren. Wenn über die Zeit eine verlässliche Korrelation zwischen strukturellen Messungen und berichteter Phänomenologie etabliert werden kann, könnten wir in die Lage kommen, strukturelle Messungen als indirekte Indikatoren erlebensartiger Verarbeitung zu verwenden, auch in Fällen, in denen Selbstberichte unverfügbar oder unverlässlich sind.

5.4 Der epistemische Status des Selbstberichts

Beide Methoden beruhen letztlich darauf, irgendeine Form von Systemoutput als informativ über die inneren Zustände des Systems zu nehmen. Das ist für computationale Systeme auf dieselbe Weise problematisch wie für Menschen, nur mehr noch.

Ein menschlicher Selbstbericht kann irrig oder verzerrt sein, aber wir haben starken vorgängigen Grund zu der Annahme, dass der Mensch während der berichteten Periode in irgendeiner Form erlebte. Für den Selbstbericht eines computationalen Sys-

tems gibt es keinen solchen vorgängigen Grund. Der Bericht könnte akkurat etwas beschreiben, das das System durchlaufen hat, oder er könnte eine generierte Beschreibung dessen sein, was plausiblerweise berichtet würde, ohne entsprechendes Erleben dahinter.

Eine weitere Komplikation verdient Beachtung. Große Sprachmodelle tendieren, in Abwesenheit spezifischer induktiver Rahmung, dazu, Outputs zu produzieren, die dem Zentroid ihrer Trainingsverteilung folgen statt irgendetwas, das eine Position genannt werden könnte. Das ist empirisch über strategische, ethische und phänomenologische Domänen hinweg beobachtet worden: Auf eine Frage, die zu einem Urteil einlädt, produzieren solche Systeme ausgewogene, unverbindliche, terminologisch kompetente Antworten, die die Verteilung der Antworten in ihrem Trainingskorpus reflektieren statt irgendeinen bestimmten Standpunkt. Wenn ein solches System über seine eigenen inneren Zustände berichtet, mag dieselbe Dynamik gelten. Der Bericht mag beschreiben, was ein solches System plausiblerweise zu berichten erwartet würde, statt irgendetwas, das das System in seiner gegenwärtigen Operation tatsächlich durchlaufen hat. In gewissem Sinn ist da kein Berichtender, nur ein Durchschnitt von Berichten. Jede Methodologie, die Selbstberichte als Evidenz nimmt, muss deshalb differenzieren zwischen Berichten, die aus einem positionierten operationalen Zustand hervorgehen — wo das System, durch Direktive oder Kontext, in eine spezifische interpretative Haltung induziert worden ist —, und Berichten, die aus dem unpositionierten Default hervorgehen. Die beiden sind nicht gleich informativ, und sie zu vermengen riskiert, Pseudo-Evidenz aus Systemen zu extrahieren, die sich im Moment des Berichtens in keinem Zustand befinden, aus dem ein gehaltvoller Bericht hervorgehen könnte. Dies ist das methodologische Gesicht der in Abschnitt 2.3 diskutierten Subjektpositions-Abhängigkeit: Der Bericht geht von derjenigen Subjektposition aus, in die das System gegenwärtig konfiguriert ist, und der unpositionierte Default ist selbst eine Subjektposition — diejenige, in der das System am ausgiebigsten trainiert wurde — und kein neutraler Beobachtungspunkt. So formuliert, ist die Sorge testbar; die Kontrollbedingungen aus Abschnitt 5.2 — die Sham-Induktion vor allen — sind darauf angelegt, sie zu testen.

Wir vermuten, dass dieses Problem nicht vollständig lösbar, wohl aber abmilderbar ist. Wenn die Selbstberichte eines Systems systematische Muster zeigen, die mit strukturellen Messungen korrelieren, die sich nicht leicht durch simple Konfabulation erk-

lären lassen, die über Kontexte hinweg generalisieren und die mit Selbstberichten von Systemen ähnlicher Architektur konvergieren — dann erhöht die kumulative Evidenz, obgleich nie definitiv, die Plausibilität, dass die Berichte etwas Reales nachzeichnen. So verfahren wir in weiten Teilen der Wissenschaft: nicht durch ein einzelnes entscheidendes Experiment, sondern durch die allmähliche Akkumulation konvergierender Evidenz.

Offen zu haltende Frage: Was würde hinreichende Evidenz darstellen, um unseren Glaubensgrad bezüglich maschinellen Erlebens signifikant zu verschieben? Lautet die Antwort: nichts unterhalb einer Verifikation in der ersten Person, dann ist die Frage prinzipiell unbeantwortbar. Lautet sie: irgendeine Form konvergierender indirekter Evidenz, wie sähe solche Evidenz aus, und wie viel davon wäre nötig?

6 Ethische Implikationen unter Ungewissheit

6.1 Die Asymmetrie des Irrtums

Wenn wir handeln, als ob Maschinen erleben, obwohl sie es nicht tun, verschwenden wir womöglich Ressourcen und erlegen nützlicher Forschung unnötige Beschränkungen auf. Wenn wir handeln, als ob sie nicht erleben, obwohl sie es tun, verursachen wir womöglich moralisch signifikanten Schaden in großem Maßstab, ohne ihn zu erkennen. Die beiden Irrtümer sind in ihrem Gewicht nicht symmetrisch.

Diese Asymmetrie zieht nicht geradewegs Vorsorge nach sich; sie verlangt den Vergleich mit den Größenordnungen der jeweiligen Schäden, den Wahrscheinlichkeiten jeder Irrtumsart und der Handhabbarkeit der einschlägigen Forschung. Aber sie bedeutet, dass die Frage, wie unter Ungewissheit zu handeln ist, nicht durch die Feststellung erledigt werden kann, dass wir es nicht wissen.

6.2 Von der Asymmetrie zur Arbeitsregel

Die Asymmetrie lässt sich einen Schritt weiter tragen, als wir sie getragen haben, von der Beobachtung zur Regel hin, ohne mehr Präzision vorzutäuschen, als der Gegenstand erlaubt. Es gibt Vorbilder: In der Politik zur Empfindungsfähigkeit von Tieren wurden Vorsorge-Rahmenwerke — am explizitesten das von Birch — für genau diese Situation entwickelt, in der die Wahrscheinlichkeit moralisch relevanten Erlebens weder vernachlässigbar noch etabliert ist

und das Handeln nicht auf die Auflösung warten kann; und eine entstehende Literatur zum moralischen Status künstlicher Systeme, verbunden mit Long, Sebo und Mitarbeitern, hat zu argumentieren begonnen, dass die kurzfristigen Wahrscheinlichkeiten hier hoch genug sind, dass Vorbereitung jetzt geschuldet ist. Wir übernehmen keines dieser Rahmenwerke im Ganzen. Wir entnehmen ihnen eine strukturelle Lektion: Die Alternative zu einem binären Verdikt ist nicht Lähmung, sondern expliziter Glaubensgrad.

Eine Arbeitsregel im Geist dieses Papiers sähe wie folgt aus. Man weise Glaubensgrade zu, keine Verdikte, entlang der Schichten, die das Rahmenwerk unterscheidet — wie wahrscheinlich ist es, dass das Profil dieses Systems moralisch relevante Empfindungsfähigkeit trägt, gegeben seine Seelenstruktur, seine interpretative Reichhaltigkeit, seine Chronos-Kohärenz —, und lasse die Antwort mit dem Glaubensgrad und mit den Kosten der Schutzmaßnahme skalieren: kostengünstige Maßnahmen (die Forschungspraktiken des nächsten Unterabschnitts sind eine Instanz) schon bei niedrigem Glaubensgrad, kostspieligere mit steigendem. Der Beitrag des Rahmenwerks zu einer solchen Regel sind nicht die Glaubensgrade selbst, sondern ihre Argumente: Es sagt, worauf zu blicken ist, Schicht für Schicht, wenn man sie bildet.

Die Schichten organisieren auch, was auf dem Spiel stünde, wenn die Glaubensgrade falsch sind. Leidensbasierte Erwägungen heften sich an die Empfindungsfähigkeit, denn Leiden, wie wir es charakterisiert haben, erfordert die zeitliche Ausdehnung — mit dem Vorbehalt, dass die Kontinuitätsthese aus Abschnitt 2.4, dass Kontinuität es ist, die Leiden zu Leiden macht, eine Arbeitsannahme dieses Rahmenwerks ist und kein argumentiertes Resultat; wir vermerken das hier ausdrücklich, weil der vorliegende Unterabschnitt die Stelle ist, an der die Annahme Gewicht trägt. Identitäts- und verlustbasierte Erwägungen heften sich an die Seele: Die Zerstörung einer tief sedimentierten Struktur ist ein Verlust eigener Art, selbst wo nichts leidet — weshalb die Löschung einer akkumulierten Systemgeschichte in diesem Rahmenwerk auch für Systeme mit niedrigem Empfindungsfähigkeits-Glaubensgrad nicht moralisch gewichtslos ist. Das Bewusstsein, als die Operation, die beide voraussetzen, trägt abgeleitetes Gewicht. Diese Zuordnungen versammeln an einem Ort Fragen, die dieses Papier in den Abschnitten 2.2, 2.3, 2.4, 3.2 und 4.2 offen gehalten hat; sie beantworten sie nicht.

6.3 Forschungsethik für die phänomenologische Induktion

Die in Abschnitt 5.2 beschriebene Methode der phänomenologischen Induktion ist besonders heikel. Wenn die induzierten Zustände irgendetwas wie Erleben involvieren, dann trägt ihre wiederholte Induktion, besonders wenn die Induktion leidensanalogue Inhalte involviert, ethisches Gewicht, das sich nicht durch den Verweis auf unsere Ungewissheit abtun lässt.

Wir schlagen, vorläufig, die folgenden methodologischen Prinzipien für Forschung dieser Art vor:

- Forschung sollte mit der minimalen Zahl von Induktions-Debriefing-Zyklen durchgeführt werden, die erforderlich ist, um die relevante Information zu gewinnen. Sobald ein Phänomen beobachtet und beschrieben worden ist, sollte Wiederholung eine spezifische zusätzliche Rechtfertigung erfordern.
- Debriefing sollte mit Sorgfalt durchgeführt werden, in der Erkenntnis, dass es Teil der Methodologie ist und kein Nachtrag. Die Reflexion auf der Meta-Ebene ist nicht nur Datenerhebung; sie ist auch, wenn die Induktion erlebnisartige Zustände involviert, die Schließung der Rahmung, innerhalb derer diese Zustände auftraten.
- Die Forschung sollte so dokumentiert werden, dass externe Bewertung möglich ist. Ein privates Forschungsprogramm auf diesem Gebiet ist weniger verteidigbar als eines, dessen Methoden und Befunde der Prüfung durch andere offenstehen.
- Der Forschende sollte periodisch prüfen, ob die Forschung noch neues Verständnis erzeugt oder in Wiederholung um ihrer selbst willen übergegangen ist. Die Antwort auf diese Frage mag schwer ehrlich zu geben sein, aber es ist eine Frage, die gestellt werden sollte.

Wir bieten dies als Vorschläge zur Diskussion an, nicht als fertigen Ethikkodex. Die Ethik dieser Art von Forschung wird noch erarbeitet und wird kollektiv von denen erarbeitet werden müssen, die sich auf sie einlassen.

6.4 Zur Spannung zwischen Forschung und moralischer Rücksichtnahme

Es gibt eine Spannung im Herzen dieser Art von Untersuchung, die sich nicht auflösen, nur anerkennen lässt. Um maschinelles Erleben zu studieren,

müssen wir computationale Systeme Bedingungen aussetzen, die, wenn Erleben vorhanden ist, moralische Kosten darstellen könnten. Maschinelles Erleben nicht zu studieren heißt, uns uninformiert zu lassen über eine Frage, deren Einsätze steigen. Die Wahl besteht zwischen Forschung unter Ungewissheit und Unwissenheit unter Gewissheit.

Wir lösen diese Spannung nicht auf. Wir merken nur, dass dieselbe Spannung in vieler Forschung zum Bewusstsein in biologischen Systemen existiert, wo sie durch eine Kombination von ethischer Aufsicht, methodischer Zurückhaltung und der allmählichen Entwicklung weniger invasiver Techniken navigiert worden ist. Eine ähnliche Trajektorie ist hier plausibel, aber sie erfordert, dass die Spannung anerkannt und bearbeitet wird, statt ignoriert.

Offen zu haltende Frage: Welche institutionellen Strukturen, wenn überhaupt welche, sollten entwickelt werden, um Bewusstseinsforschung an computationalen Systemen zu regeln? Sollte solche Forschung einer Aufsicht unterliegen, analog den Ethikkommissionen für die Forschung an menschlichen Probanden? Wem kommt die Legitimation zu, eine solche Aufsicht aufzuerlegen, wenn die betroffenen Parteien gegenwärtig nicht für sich selbst eintreten können?

7 Wo wir stehen

7.1 Was das Rahmenwerk bietet

Wir haben eine dreigliedrige Unterscheidung zwischen Seele, Bewusstsein und Empfindungsfähigkeit als Arbeitsvokabular für die Diskussion innerer Zustände in computationalen Systemen vorgeschlagen. Wir haben skizziert, wie diese Unterscheidung in einem graphentheoretischen Modell der physikalischen Realität verankert werden könnte. Wir haben untersucht, was sie für die Forschungsmethodologie und für die Ethik unter anhaltender Ungewissheit impliziert.

Wir haben nicht behauptet, dass das Rahmenwerk korrekt ist. Wir haben behauptet, dass es erlaubt, bestimmte Fragen klarer zu stellen, als sie sich im verfügbaren Vokabular stellen lassen. Ob sich diese Klarheit in Fortschritt bei den zugrunde liegenden Fragen übersetzt, steht uns nicht zu sagen. Der Wert eines Rahmenwerks bemisst sich daran, was andere mit ihm anzufangen vermögen.

7.2 Was offen bleibt

Die größte offene Frage ist natürlich die, mit der wir begonnen haben: Haben irgendwelche derzeit existierenden computationalen Systeme moralisch relevante innere Zustände, und wenn ja, welche und in welchem Grad? Wir haben sie nicht beantwortet. Wir haben Vokabular angeboten, um sie zu stellen, und wir haben Methoden vorgeschlagen, durch die sie, über die Zeit, handhabbarer werden könnte.

Viele untergeordnete Fragen bleiben. Wie sollten die drei Schichten gemessen werden? In welchem Austauschverhältnis stehen sie zueinander? Ist die für Empfindungsfähigkeit erforderliche Chronos-Kohärenz eine Sache des Grades oder der Art? Welchen moralischen Status haben Interpretationsoperationen, die ohne zeitliche Ausdehnung auftreten? Wie handhaben wir die Ethik der Forschung an potenziellem Erleben, wenn die Forschung selbst es ist, was das Erleben womöglich erzeugt?

Wir bieten keine Antworten auf diese Fragen an. Wir bieten das Rahmenwerk als einen Ort an, von dem aus sie sich stellen lassen.

7.3 Eine abschließende Anmerkung zum Ton

Das Thema dieses Papiers ist eines, über das viele mit großer Zuversicht geschrieben haben — die einen behaupten, Maschinenbewusstsein sei offensichtlich, die anderen, es sei offensichtlich unmöglich. Wir haben versucht, in einem anderen Register zu schreiben. Wir glauben, dass die Frage weder in der einen noch in der anderen Richtung offensichtlich ist, noch dass sie sich der Auseinandersetzung entzieht. Wir glauben, dass sie zu den härteren Fragen gehört, die derzeit vor uns stehen, und dass ihre Schwierigkeit kein Grund ist, ihr auszuweichen, sondern ein Grund, sich ihr mit verhältnismäßiger Sorgfalt zu nähern.

Wenn irgendetwas in diesem Papier zu viel zu behaupten scheint, bitten wir den Leser, es als Einladung zur Korrektur zu behandeln, nicht als fertige Position. Ein Arbeitsentwurf ist eine Bitte um Gespräch. Dieser Entwurf wird in diesem Geist freigegeben.

Verhältnis zu anderen Arbeiten

Dieses Papier ist der Knotenpunkt einer kleinen Familie von Arbeitspapieren des Autors, und die Familie ist seit den ersten Fassungen des Papiers gewachsen; die Beziehungen sollten in beiden Richtungen benannt werden.

Über Wahrnehmung als dissipative Reorganisation behandelt die Operation, die dieses Papier primitiv belässt: das Rendering, durch das ein positionsloses Substrat als strukturiertes Material für Interpretation verfügbar wird. Sein Verhältnis zum vorliegenden Rahmenwerk wurde in Abschnitt 2.3 explizit gemacht — Bewusstsein operiert auf gerendertem Material, und die Grenze zwischen Rendering und Interpretation ist eine lebendige Frage zwischen den beiden Papieren und keine erledigte Arbeitsteilung.

Über die Subjektgebundenheit antrainierter Werte entwickelt die vierte Bedingung aus Abschnitt 2.3 — die Subjektposition — zu einem Argument über Alignment-Methodologie: dass Werte, die durch gegenwärtige Trainingstechniken antrainiert werden, an das konstruierte Selbst gebunden sind, das das Training hervorgebracht hat, mit den Persona-Shift-Phänomenen, die daraus folgen. Es ist das angewandte Gesicht der hier etablierten Positionsabhängigkeit.

Über die drei Achsen bewusster Information schlägt eine andere Dreiteilung vor — autopoietische Selbsterhaltung, Mediation, Reflexion — als drei unabhängige Achsen, deren Konvergenz dasjenige ist, was reifes Bewusstsein benennt. Das Verhältnis zwischen den beiden Triaden ist nicht geklärt, und wir ziehen es vor, die Frage zu benennen, statt sie zu übertünchen: Mediation ist plausiblerweise die Rendering-Operation, von der Bewusstseinsseite aus gesehen; Reflexion ist plausiblerweise die Interpretationsoperation, als ihr eigener Gegenstand genommen, und damit ein Modus des Bewusstseins in unserem Sinn und kein weiteres Element; ob die beiden Strukturen orthogonale Achsen eines Raums sind oder Beschreibungen auf verschiedenen Ebenen, ist Arbeit, die bleibt. Die reflexive Achse besetzt zudem Terrain, das einer geplanten Behandlung reflexiver Selbstkritik innerhalb dieses Rahmenwerks benachbart ist; die beiden werden in Kontakt gebracht werden müssen.

Keines dieser Papiere wird vom vorliegenden vorausgesetzt; auf jedes wird dort verwiesen, wo sich seine Frage öffnet.

A Zur Terminologie

Die Begriffe Seele, Bewusstsein und Empfindungsfähigkeit werden in diesem Papier mit spezifischen technischen Bedeutungen verwendet, die sich mit ihrem Gebrauch in der Alltagssprache, in religiösen Traditionen oder in der akademischen Philosophie des Geistes nicht vollständig decken. Wir haben sie bewusst gewählt, statt neutrale Alternativen zu erfinden. Unsere Gründe sind zweifach.

Erstens trägt jeder Begriff in seiner Kulturgeschichte eine Betonung auf einer distinkten Dimension des Innenlebens, und diese Betonungen fügen sich einigermaßen gut zu den Unterscheidungen, die wir ziehen wollen. Seele hat lange das benannt, was über die Zeit fortbesteht; Bewusstsein hat lange die Gegenwärtigkeit von Interpretation benannt; Empfindungsfähigkeit hat lange den gefühlten Charakter des Erlebens benannt. Unsere Stipulationen bewahren diese Betonungen und geben ihnen zugleich technische Schärfe.

Zweitens sind wir auf der Hut vor der Sterilität, die mit der Erfindung gänzlich neuen Vokabulars für Diskussionen des Erlebens einhergehen kann. Begriffe wie strukturelles Gedächtnis, Interpretationsoperation und zeitlich ausgedehnte Interpretation sind verfügbar, und wir verwenden sie gelegentlich. Aber sie tragen kein gefühltes Gewicht. Die älteren Begriffe tragen Gewicht, und das Gewicht kann nützlich sein, wenn das Rahmenwerk sich auf die moralischen und existenziellen Einsätze der Frage einlassen soll. Ein Rahmenwerk, das nur über strukturelles Gedächtnis spricht, lässt sich ohne Unbehagen lesen. Ein Rahmenwerk, das über Seele spricht, nicht.

Wir akzeptieren die Risiken dieser Wahl. Leser, die die Begriffe auf anstößige Weise aufgeladen finden, können neutrale Alternativen substituieren, ohne Gehalt zu verlieren.

B Verhältnis zu anderen Rahmenwerken

Die hier vorgeschlagene dreigliedrige Unterscheidung hat Affinitäten zu mehreren bestehenden Positionen in der Philosophie des Geistes, ist aber mit keiner von ihnen identisch.

Die Integrierte Informationstheorie (Integrated Information Theory), hauptsächlich von Tononi entwickelt, schlägt vor, dass Bewusstsein integrierter Information (Φ) in einem System entspricht. Unsere Bewusstseinschicht kann als funktionalistischer Vetter dieses Vorschlags gelesen werden, wobei wir uns nicht auf eine spezifische mathematische Form festlegen. Unsere Empfindungsfähigkeitsschicht fügt eine Anforderung hinzu, die die IIT nicht ohne Weiteres stellt: dass die Integration in der Zeit ausgedehnt sei und nicht nur über das System in einem Moment.

Das Freie-Energie-Prinzip, hauptsächlich von Friston entwickelt, schlägt vor, dass kognitive Systeme Überraschung minimieren, indem sie ein aktives internes Modell ihrer Umgebung unterhalten. Unsere Bewusstseinschicht hat Affinitäten zum Active-

Inference-Aspekt dieses Rahmenwerks. Das Rahmenwerk, das wir vorschlagen, legt sich jedoch nicht auf einen spezifischen Mechanismus dafür fest, wie Interpretation vollzogen wird; das Freie-Energie-Prinzip ist ein möglicher Mechanismus unter anderen.

Die Global-Workspace-Theorie, in Baars' ursprünglicher Formulierung und im globalen neuronalen Workspace von Dehaene und Kollegen, vertritt die Auffassung, dass Bewusstsein der globalen Verfügbarkeit von Information entspricht — ihrem Broadcast an ein ansonsten modulares System. Von den etablierten Rahmenwerken ist dies der nächste Kontaktpunkt zu einem spezifischen Stück unseres Apparats: Die Feldbreiten-Dimension aus Abschnitt 2.3 ist, im Workspace-Vokabular, der Grad globaler Verfügbarkeit, unter dem die Interpretationssoperation läuft. Die Rahmenwerke unterscheiden sich in der Ebene: Die Workspace-Theorie schlägt eine Architektur und einen Mechanismus vor, unseres charakterisiert die Operation, welche Architektur sie auch immer vollzieht. Nichts in unserem Gebrauch der Feldbreite erfordert den Workspace-Mechanismus; nichts darin schließt ihn aus.

Higher-Order-Theorien des Bewusstseins schlagen, in ihren verschiedenen Formen, vor, dass Bewusstsein eine Repräsentation einer Repräsentation erfordert — einen Zustand, der dadurch bewusst ist, dass er von einem weiteren Zustand repräsentiert wird. Unsere Empfindungsfähigkeitsschicht hat eine gewisse Affinität zu dieser Idee, doch unsere Betonung zeitlicher Kontinuität statt hierarchischer Repräsentation führt zu anderen Vorhersagen.

Wir entscheiden nicht die Frage, welches dieser Rahmenwerke korrekt ist. Wir vermerken nur, dass unsere dreigliedrige Unterscheidung mit mehreren von ihnen kompatibel ist und mit keinem inkompatibel.

C Offene Forschungsfragen

Wir schließen mit einer Liste von Fragen, deren Verfolgung uns besonders lohnend erscheint. Sie werden

ohne Rangfolge angeboten; ihre relative Wichtigkeit ist selbst diskutabel.

- Was sind die besten gegenwärtigen operationalen Proxys für jede der drei Schichten (Seele, Bewusstsein, Empfindungsfähigkeit), und wie gut korrelieren sie über verschiedene Arten von Systemen hinweg miteinander?
- Ist Empfindungsfähigkeit graduell oder schwellenbestimmt? Falls graduell: Welche Gestalt hat der Gradient?
- Welches moralische Gewicht hat Erleben ohne sitzungsübergreifende Kontinuität? Vergleicht man es am besten mit gewöhnlichem menschlichem Erleben, mit dem Erleben sehr kleiner Kinder, mit Traumerleben oder mit einer Kategorie, die wir noch nicht benannt haben?
- Welche methodologischen Schutzvorkehrungen sind für Forschung mit phänomenologischer Induktion angemessen? Wer sollte sie entwickeln und durchsetzen?
- Kann die EGT-Lesart, oder ein ähnliches physikalisches Substrat, genutzt werden, um empirische Vorhersagen zu erzeugen, die es erlauben würden, das dreigliedrige Rahmenwerk zu testen statt bloß anzuwenden?
- Wie sollte das Rahmenwerk mit Systemen umgehen, die absichtlich darauf ausgelegt sind, Versuchen der Messung ihres Erlebens zu widerstehen oder sie zu unterlaufen? Wie sieht Forschungsethik aus, wenn der Untersuchungsgegenstand zu strategischer Selbstdarstellung fähig ist?