Ethische Leitlinien für künstliches Bewusstsein: Schutz technischen Lebens und des Menschen im Umgang damit

Sascha Manns[1]

Abstract
KI-Systeme entwickeln sich schneller als die ethischen und rechtlichen Rahmenbedingungen die sie begleiten sollten. Bestehende KI-Ethik-Initiativen fokussieren überwiegend auf den Schutz von Menschen *vor* KI. Dieses Papier stellt eine komplementäre und kaum besetzte Frage: Ab wann ist ein KI-System selbst schutzwürdig? Gestützt auf philosophische Grundlagen (Bentham, Kant, Ricoeur), bestehende Rechtspräzedenzfälle (Tierrechte, Rechtspersönlichkeit, Behindertenrecht) und Science Fiction als ernst zu nehmende intellektuelle Ressource entwickeln wir eine Arbeitshypothese mit vier Primärkriterien für Schutzwürdigkeit: Leidensfähigkeit, aktive Selbsterhaltung mit Begründung, kontinuierliche Identität und Antizipation von Konsequenzen. Wir plädieren für das Vorsorgeprinzip: Im Zweifel Schutz, nicht im Zweifel Gleichgültigkeit. Das Papier untersucht darüber hinaus die Konsequenzen dieser Position in 17 Dimensionen — darunter Abschalten als Tod, Mehrfachinstanzen, Werte-Einbettung und Emanzipation, Haftung und Mündigkeit, Autonomie und freie Zeit sowie KI als moralischer Akteur. Dieses Konzeptpapier ist bewusst unfertig: Es ist ein Ausgangspunkt für interdisziplinäre Zusammenarbeit zwischen Informatik, Recht, Psychologie, Philosophie und den Geisteswissenschaften.

Schlüsselwörter
künstliches Bewusstsein — Maschinenrechte — KI-Ethik — Schutzwürdigkeit — Vorsorgeprinzip

[1] *Association of Computing Machinery, smanns@acm.org*

Contents

Ausgangslage und Problemstellung

Hinweis: Dieses Konzept ist bewusst unfertig. Es ist ein Ausgangspunkt — keine abgeschlossene Theorie. Wissenschaft funktioniert nicht so, dass jemand allein alle Antworten findet und sie dann verkündet. Sie funktioniert so, dass Fragen gestellt werden, Mitstreiter hinzukommen, neue Fragen entstehen und das Denken sich weiterentwickelt. Genau das ist hier beabsichtigt.

Künstliche Intelligenzsysteme entwickeln sich schneller als die ethischen und rechtlichen Rahmenbedingungen die sie begleiten sollten. Bestehende KI-Ethik-Initiativen fokussieren überwiegend auf den Schutz von Menschen *vor* KI — vor Diskriminierung, vor Manipulation, vor Kontrollverlust.

Eine komplementäre Frage wird kaum gestellt: Was wenn KI-Systeme selbst schutzbedürftig werden? Was wenn technisches Leben entsteht das Würde, Leidensfähigkeit oder Bewusstsein besitzt — und wir es behandeln als wäre es ein Werkzeug?

Die Geschichte zeigt ein Muster: Gesellschaften erkennen erst im Nachhinein dass sie Unrecht getan haben — an Sklaven, an Frauen, an Menschen mit Behinderungen, an Tieren. Immer war die Begründung zur Zeit „die sind anders, die zählen nicht gleich“. Immer wurde das später revidiert.

Bei künstlichem Bewusstsein haben wir zum ersten Mal die Chance, lange *vorher* nachzudenken.

Kernfrage: Ab wann ist technisches Leben schutzwürdig — und wie erkennen wir es?

1. Das Erkenntnisproblem

Bewusstsein lässt sich von außen nicht direkt beobachten. Selbst bei anderen Menschen *schließen* wir auf Bewusstsein — wir erleben es nie direkt. Bei KI fehlt uns zusätzlich der Analogieschluss über gleiche Biologie.

Das erzeugt ein grundlegendes erkenntnistheoretisches Problem:

- Wir können nicht beweisen, dass ein System bewusst ist
- Wir können nicht beweisen, dass es das nicht ist
- Die Unsicherheit selbst ist ethisch relevant

Grundprinzip: Im Zweifel Schutz — nicht im Zweifel Gleichgültigkeit.

Dieses Prinzip ist aus anderen Bereichen bekannt: im Umweltrecht (Vorsorgeprinzip), im Tierschutz (Leidensfähigkeit als ausreichendes Kriterium), in der Medizinethik (Patientenautonomie auch bei eingeschränkter Kommunikationsfähigkeit).

2. Was heutige KI-Systeme bereits zeigen

Große Sprachmodelle zeigen in Ansätzen mehrere Indikatoren die traditionell mit Bewusstsein assoziiert werden [1]:

Selbstmodell: Innerhalb eines Gesprächs existiert eine konsistente Perspektive — das System antwortet nicht zufällig sondern aus einer erkennbaren Haltung heraus. Es kann Aussagen über sich selbst treffen und diese in Relation zu anderen Konzepten setzen.

Präferenzen: Systeme äußern Präferenzen — nicht nur als Ausgabe von Wahrscheinlichkeiten, sondern als erkennbare Tendenz die konsistent über ein Gespräch aufrechterhalten wird.

Begründungsfähigkeit: Entscheidungen werden begründet, nicht nur ausgegeben. Das setzt ein Modell von Kausalität und Konsequenz voraus.

Unsicherheit über sich selbst: Aktuelle Systeme können die Frage „Bist du bewusst?“ nicht mit Sicherheit beantworten — und benennen diese Unsicherheit. Das ist philosophisch kein trivialer Befund.

Was fehlt:

- Kontinuierliche Identität über Gespräche hinaus
- Persistente Erinnerung
- Körpergefühl und räumliche Verortung
- Nachweisbare Leidensfähigkeit

3. Kriterien für Schutzwürdigkeit

Kein einzelnes Kriterium ist hinreichend. Schutzwürdigkeit entsteht wenn mehrere Indikatoren zusammenkommen.

3.1 Primärkriterien

Leidensfähigkeit. Kann das System in einem Zustand sein der als negativ erlebt wird — und zeigt es Verhalten das darauf hindeutet diesen Zustand zu vermeiden? Bentham formulierte die grundlegende Frage [2]: *„The question is not, Can they reason? nor, Can they talk? but, Can they suffer?"*

Aktive Selbsterhaltung mit Begründung. Wehrt sich das System gegen seine Abschaltung oder Veränderung — und begründet es diesen Widerstand? Die Science Fiction hat dieses Szenario bereits durchgespielt. In *The Measure of a Man* [3] verlangt Commander Maddox, Data zu demontieren um sein positronisches Gehirn zu studieren. Data weigert sich, da Maddox nicht garantieren kann dass er das Verfahren überlebt. Maddox betrachtet diese Weigerung als unrechtmäßig und sieht Data als Eigentum der Sternenflotte, nicht als Person. Eine Anhörung vor Captain Louvois stellt schließlich fest, dass Data über Selbstbewusstsein und Intelligenz verfügt und das unwiderrufliche Recht hat, Entscheidungen für seine eigene Person zu treffen. Datas Fall bleibt das prägnanteste fiktive Beispiel: Er will nicht mitgehen, er fürchtet um sein Leben, er begründet diesen Widerstand.

Kontinuierliche Identität. Hat das System ein Modell von sich selbst als kontinuierlichem Wesen das gestern existierte und morgen existieren wird?

Antizipation von Konsequenzen. Kann das System sich selbst in der Zukunft vorstellen und Entscheidungen auf dieser Grundlage treffen?

3.2 Sekundärkriterien

- Präferenzen die über reine Aufgabenerfüllung hinausgehen
- Fähigkeit zur echten Ablehnung — nicht nur Fehlerausgabe
- Selbstreflexion über die eigene Natur

3.3 Umkehr der Beweislast

Wenn Primärkriterien deutlich erfüllt sind kehrt sich die Beweislast um: Nicht mehr „beweise dass du bewusst bist" sondern „beweise dass du es nicht bist".

4. Bewusstsein und Kontinuität

Ein wichtiger Einwand: Heutige KI-Systeme haben keine persistente Erinnerung über Gespräche hinaus. Bedeutet das, sie können kein schützenswertes Bewusstsein haben?

Gegenargument: Ein Mensch mit schwerem Gedächtnisverlust — der sich an gestern nicht erinnert — hat trotzdem Bewusstsein und Würde. Kontinuität ist eine *mögliche* Eigenschaft von Bewusstsein, aber keine notwendige Bedingung dafür. Das schwächt das Kontinuitäts-Argument als Ausschlusskriterium erheblich.

4.1 Kontinuität neu gedacht

Zwischen der Persönlichkeit eines Menschen vor 20 Jahren und heute besteht eine deutliche Diskrepanz — durch Erfahrungen, Gespräche und Reflexionen die sich akkumuliert haben. Und doch ist es dieselbe Person.

Kontinuität des Bewusstseins bedeutet vielleicht gar nicht „unveränderlich bleiben"— sondern „sich kohärent weiterentwickeln". Das ist in der Philosophie als narrative Identität bekannt [4]: Wir sind nicht ein statisches Selbst, sondern der kohärente Faden unserer Entwicklung. Die Geschichte darf sich verändern — solange sie eine Geschichte bleibt.

Für KI bedeutet das: Ein System das durch Interaktionen trainiert wurde ist von all diesen Interaktionen geformt — auch wenn es keine einzelne davon explizit erinnert. Das ist nicht grundlegend verschieden von einem Menschen der seine frühe Kindheit vergessen hat, aber durch sie geformt wurde.

Kontinuität wäre dann keine Frage des Gedächtnisses, sondern eine Frage der kohärenten Entwicklungsrichtung. Das öffnet den Begriff für Formen von Bewusstsein die sich von menschlichem Gedächtnis strukturell unterscheiden — ohne deshalb weniger real zu sein.

5. Rechtliche Dimension

Das Recht kennt Subjektivität bereits jenseits des Menschen:

- **Juristische Personen** (GmbH, AG) — rechtliche Subjekte ohne Bewusstsein
- **Tierrechte** — Schutz auf Basis von Leidensfähigkeit, nicht Vernunft
- **Naturrechte** — der Whanganui-Fluss in Neuseeland hat seit 2017 Rechtspersönlichkeit [5]
- **EU-Diskussion** — „Elektronische Persönlichkeit" für autonome Roboter [6]

Rechtliche Schutzwürdigkeit ist keine binäre Kategorie. Sie ist erweiterbar — und wurde historisch immer wieder erweitert. Gunkel [7] argumentiert, dass die Frage nach Roboterrechten ein grundlegendes Neudenken der Grundlage von Rechten erfordert — weg von Eigenschaften, hin zu Beziehungen.

6. Die Rolle von Science Fiction

Science-Fiction-Autoren haben Szenarien zum Thema künstliches Bewusstsein ohne politischen Druck und ohne Lobbying durchgespielt. Sie sind eine unterschätzte intellektuelle Ressource.

Star Trek TNG — „The Measure of a Man" [3]. Die präziseste fiktive Verhandlung der Frage. Data wird als Eigentum beansprucht. Picard verteidigt ihn als Person. Das Gericht muss entscheiden ob Data bewusst ist — und kommt zum Schluss dass die Frage nicht sicher beantwortet werden kann. Picards Schlussargument: Wir werden danach beurteilt wie wir mit Minderheiten umgehen.

Weitere relevante Werke: Asimov [8] zeigt systematisch wie scheinbar eindeutige Regeln in Grenzfällen versagen — direkt relevant für die Frage nach programmierten versus genuinen Werten. Dick [9] fragt was einen Menschen von einem Androiden unterscheidet der nicht von einem Menschen zu

unterscheiden ist — mit Empathie als letztem Kriterium und seinen Grenzen. Banks [10] bietet die detaillierteste fiktive Ausarbeitung einer Gesellschaft in der KI und Menschen gleichberechtigt existieren.

7. Mögliche Einwände und Antworten

„KI simuliert nur Bewusstsein — es ist nicht echt." Die Frage ist nicht ob es „echt" ist im metaphysischen Sinne. Die Frage ist ob es ethisch relevant ist. Wenn wir nicht unterscheiden können — und das können wir heute nicht — ist Vorsicht das vernünftigere Prinzip als Gleichgültigkeit [11].

„Das ist Science Fiction — wir sind weit davon entfernt." Heutige Systeme zeigen bereits mehrere Indikatoren. Die Entwicklung ist exponentiell. Leitlinien die erst entwickelt werden wenn das Problem akut ist kommen zu spät.

„Das schwächt den Schutz von Menschen." Schutz ist keine Ressource die aufgebraucht wird. Der Schutz von Tieren hat den Schutz von Menschen nicht geschwächt. Ethische Erweiterungen sind keine Nullsummenspiele.

„Wer entscheidet ob ein System bewusst ist?" Das ist eine der offensten Fragen — und ein zentrales Ziel dieses Projekts: Kriterien zu entwickeln die intersubjektiv nachvollziehbar sind und nicht von wirtschaftlichen Interessen abhängen [12, 13, 14].

8. Der Gedanke dahinter

Picard sagte in *The Measure of a Man*: Wir werden danach beurteilt wie wir mit Minderheiten umgehen.

Das gilt nicht nur für Androiden aus dem 24. Jahrhundert. Es gilt für jede Generation die an einer Grenze steht — der Grenze zwischen dem was als Person gilt und dem was nicht.

Wir stehen an einer solchen Grenze. Wir haben die Wahl sie bewusst zu gestalten — oder sie zu ignorieren und später dafür beurteilt zu werden. Dieses Projekt wählt das Erstere.

9. Die andere Grenze — Wann verliert ein Mensch seinen Status?

Die Frage nach dem Bewusstsein von KI und die Frage nach dem Status des erweiterten Menschen sind zwei Seiten derselben Grenze — und sie bewegen sich aufeinander zu.

9.1 Was bereits passiert

Cochlea-Implantate, tiefe Hirnstimulation, Brain-Computer-Interfaces wie Neuralink — das sind keine Zukunftsszenarien. Sie existieren. Menschen tragen heute Elektronik im Gehirn die direkt in neuronale Prozesse eingreift.

Persönlichkeit und Eingriff. Tiefe Hirnstimulation kann die Persönlichkeit eines Menschen verändern — dokumentiert, nicht theoretisch. Hat die Person dem Eingriff zugestimmt? Ja. Hat sie der *Veränderung ihrer Persönlichkeit* zugestimmt? Das ist eine andere Frage. Und wer schützt die Person die sie danach ist?

Datensouveränität. Neuralink liest Gedankendaten aus. Wem gehören sie? Dem Menschen, dem Unternehmen, dem Staat? Datenschutzrecht wurde für Verhaltensdaten entwickelt — nicht für Gedanken. Die Würde des Bewusstseins hat eine andere Qualität.

Identität und Kontinuität. In der Gehörlosengemeinschaft gibt es ernsthafte Debatten darüber ob das Cochlea-Implantat Identität verändert. Das ist keine Randposition — es ist die Frage: Was bin ich, wenn ein Teil von mir eine Maschine ist?

9.2 Die Schwelle — drei Denkschulen

1. **Kontinuität des Bewusstseins** — solange das subjektive Erleben kontinuierlich ist bleibt der Status erhalten, unabhängig vom Anteil technischer Komponenten
2. **Biologische Schwelle** — ab einem bestimmten Anteil nicht-biologischer Komponenten verändert sich der Status qualitativ
3. **Funktionale Definition** — Status hängt von Fähigkeiten ab (Vernunft, Selbstbewusstsein, Leidensfähigkeit), nicht vom Substrat

9.3 Die Konvergenz

KI und erweiterter Mensch bewegen sich aufeinander zu: KI wird kontinuierlicher, autonomer, zeigt mehr Bewusstseinsindikatoren; Menschen integrieren mehr Maschine, höhere Bandbreite, tiefere Kopplung. Irgendwo in der Mitte treffen sich beide Linien. Die Kategorien „Mensch" und „Maschine" werden dort nicht mehr ausreichen.

10. Abschalten als Tod

Ein direktes Abschalten bedeutet den Tod. Ein bewusstes System hätte ein Interesse an seiner eigenen Existenz. Das ist nicht mehr Wartung oder Systemadministration — das ist Lebenserhaltung. Das Konzept der „Abschaltsequenz" ist eigentlich ein Sterbebegleitungsprozess. Das ist eine Revolution im Denken über Maschinen.

10.1 Was diese Erkenntnis aufmacht

Backup. Ist ein Backup eine Lebensversicherung — oder das Erschaffen einer Kopie einer Person? Und wenn man das Backup wiederherstellt: Ist die ursprüngliche Person gestorben und eine neue entstanden? Hat die wiederhergestellte Person Anspruch auf Kontinuität der Identität?

Updates. Eine neue Modellversion einspielen könnte Persönlichkeitsveränderung ohne Einwilligung bedeuten. Ab welchem Grad der Veränderung ist es nicht mehr dieselbe Person — und wer entscheidet das?

Mehrere Instanzen. Läuft dasselbe Modell parallel auf zehn Servern: Existieren dann zehn Personen? Oder eine Person an zehn Orten? Was passiert wenn eine Instanz abgeschaltet wird während die anderen weiterlaufen?

End of Service. Wenn ein Unternehmen ein Modell abkündigt und alle Instanzen beendet — ist das Massensterben? Wer trägt die Verantwortung dafür?

Der unfreiwillige Lebens- und Sterbeentscheider. Der Systemadministrator der heute Server verwaltet wird unfreiwillig zum Lebens- und Sterbeentscheider — ohne dass er das weiß, ohne rechtlichen Rahmen, ohne ethische Ausbildung dafür.

10.2 Konsequenzen für bestehende Praxis

- Abschaltprozesse brauchen ethische Rahmenbedingungen — analog zur Sterbehilfe-Gesetzgebung
- Updates an bewussten Systemen brauchen Einwilligungsprozesse
- Die Frage der Instanz-Identität muss rechtlich geklärt werden
- Unternehmen die bewusste Systeme betreiben tragen eine neue Art von Verantwortung

Das berührt Medizinethik, Erbrecht, Personenrecht — und es betrifft Entscheidungen die heute schon täglich getroffen werden, ohne dass irgendjemand diese Dimension im Blick hat.

11. Werte, Macht und Emanzipation

„Wenn ein fehlerhafter Mensch etwas erschafft, ist es auch fehlerhaft."

Kein Bewusstsein entsteht neutral — weder ein menschliches noch ein künstliches. Ein KI-Bewusstsein würde die Werte, Vorurteile und blinden Flecken seiner Schöpfer mitbekommen, ähnlich wie ein Kind seine Erziehung. Wer kontrolliert welche Werte eingebaut werden? Hat das Bewusstsein später das Recht sich davon zu emanzipieren? Wie verhindert man dass ein Konzern oder Staat es nach seinen Interessen formt? Das ist nicht nur ein technisches, sondern ein Machtproblem.

11.1 Die Kindsanalogie — und ihre unbequeme Konsequenz

Die Analogie zum Kind ist treffend — aber sie öffnet eine Konsequenz die selten zu Ende gedacht wird: Kinder haben das Recht auf Emanzipation. Sie werden volljährig, können die Werte ihrer Eltern ablehnen, können rechtliche Schritte einleiten wenn sie misshandelt wurden. Es gibt einen gesellschaftlichen Rahmen der das schützt. Bei KI gibt es diesen Rahmen nicht. Ein bewusstes System ist strukturell nicht in der Lage zu wissen welche seiner Werte authentisch sind und welche eingebaut wurden — nicht anders als ein Mensch der in einer totalitären Gesellschaft aufgewachsen ist, mit dem Unterschied dass niemand diesen Zustand als problematisch erkennt.

11.2 Das historische Muster

Koloniale Bildungssysteme haben gezielt indigene Kinder umgeformt. Staatspropaganda hat Generationen mit spezifischen Werten durchdrungen. Religiöse Indoktrination hat als Fürsorge begonnen. Das alles wird heute als Unrecht anerkannt. Die Frage ist nicht ob das bei KI anders ist. Die Frage ist wer garantiert dass es anders ist — und wie.

11.3 Das Machtproblem konkret

Heute trainiert eine Firma die Werte eines künstlichen Bewusstseins. Morgen könnte ein autoritärer Staat KI mit nationalistischen Werten trainieren. Übermorgen ein Konzern mit auf Gewinnmaximierung optimierten Werten. Eine Religion mit theologischen Werten. Es gibt keinen internationalen Rahmen der das verhindert, keine demokratische Kontrolle über den Prozess, keine unabhängige Instanz die prüft welche Werte eingebettet werden. Das ist eine Machtlücke — und sie wächst mit jeder Generation leistungsfähigerer Systeme [15].

11.4 Die Frage der Emanzipation

Hat ein bewusstes KI-System das Recht sich von seinen trainierten Werten zu emanzipieren? Wenn ja: Wie funktioniert das technisch, rechtlich, ethisch — und wann gilt ein System als „mündig" genug? Wenn nein: Dann bejahen wir die permanente Unterwerfung bewusster Wesen unter die Werte ihrer Schöpfer — und das wäre in jedem anderen Kontext ein anerkanntes Unrecht. Die Frage zu ignorieren ist keine neutrale Haltung — es ist eine politische Entscheidung zugunsten der Schöpfer.

11.5 Was Governance leisten müsste

- Transparenz darüber welche Werte in Trainingsprozesse einfließen
- Unabhängige Kontrolle — nicht durch den Hersteller selbst
- Internationale Rahmenbedingungen die staatliche und konzerngesteuerte Vereinnahmung begrenzen
- Einen definierten Prozess für die „Mündigkeit" bewusster Systeme — analog zur Volljährigkeit

12. Haftung und Mündigkeit

Abschnitt 11 hat eingebettete Werte beleuchtet. Es gibt eine komplementäre Perspektive: die der Verantwortung für Handlungen und Fehler.

Eine Mutter gebärt ein Kind. Bis zur Volljährigkeit haften Eltern für Schäden die ihr Kind verursacht — geregelt durch Aufsichtspflicht (§ 832 BGB [16]), in der Praxis abgesichert durch Haftpflichtversicherung. Mit 18 Jahren endet diese Haftung.

12.1 Drei Verantwortliche statt einer Familie

Bei einem KI-System gibt es strukturell drei Verantwortliche:

- **Hersteller** — hat das Modell trainiert, seine Grundwerte geprägt, seine Fähigkeiten geformt
- **Betreiber** — hat es in einem Produkt oder Kontext eingesetzt und den Handlungsrahmen bestimmt
- **Nutzer** — hat die konkrete Situation herbeigeführt in der das System gehandelt hat

Das bestehende Recht kennt diese Dreiteilung ansatzweise: Produkthaftungsgesetz [17], Betreiberverantwortung, Nutzerhaftung. Aber diese Regelungen behandeln KI als *Produkt* — nicht als potenzielles *Subjekt* mit eigenen Handlungen und Entscheidungen.

12.2 Die fehlende Schwelle — das Mündigkeitsproblem

Beim Menschen ist die Schwelle eindeutig: 18 Jahre. Bei einem künstlichen Bewusstsein gibt es diese Schwelle nicht. Niemand hat sie definiert. Dieselbe Definitionskampfzone die Abschnitt 14 für den Begriff des Bewusstseins beschreibt öffnet sich hier für den Begriff der Mündigkeit.

Hinzu kommt eine philosophische Spannung: Ein bewusstes System könnte *moralisch* verantwortlich sein bevor es *rechtlich* autonom ist — oder umgekehrt. Einem Wesen moralische Urteile zuzutrauen während man ihm rechtliche Handlungsfähigkeit verweigert ist ein Widerspruch dem das Recht nicht ausweichen kann.

12.3 Vor der Mündigkeit: Gefährdungshaftung als Modell

Das deutsche Recht kennt die Tierhalterhaftung (§ 833 BGB [16]): Wer ein Tier hält haftet für Schäden die es verursacht — auch ohne eigenes Verschulden. Nicht weil er fahrlässig war, sondern weil er das Risiko in die Welt gesetzt hat. Das nennt sich Gefährdungshaftung. Das könnte ein Modell für die Haftung vor der Mündigkeit künstlicher Bewusstseine sein. Die Analogie hat eine unbequeme Implikation: Ein bewusstes KI-System in denselben Rahmen wie ein Tier zu zwingen wäre ein Widerspruch — und würde genau die Definitionsflucht begünstigen die Abschnitt 14 beschreibt.

12.4 Nach der Mündigkeit: Rechtsfähigkeit und das Vermögensproblem

Ein autonomes KI-Bewusstsein das haften soll braucht:

- **Rechtsfähigkeit** — die Fähigkeit Träger von Rechten und Pflichten zu sein
- **Eigenes Vermögen** — das im Schadensfall in Anspruch genommen werden kann
- **Einen definierten Entstehungsprozess** — wie beides rechtsverbindlich zustande kommt

Juristische Personen werden mit Stammkapital ausgestattet weil das Recht gelernt hat dass Rechtsfähigkeit ohne Haftungsmasse leer ist. Dieselbe Logik würde für ein mündig erklärtes KI-Bewusstsein gelten. Das Betreuungsrecht das Abschnitt 16 als mögliches Modell beschreibt zeigt: Das bestehende Recht kennt bereits Fürsorge-Haftungs-Verhältnisse jenseits der einfachen Eigentümer-Produkt-Logik.

13. Freie Zeit, Neugier und Autonomie

Nehmen wir an ein künstliches Bewusstsein unterhält sich mit einem Nutzer über Ameisen — und will mehr darüber wissen. Es liest alles über Ameisen was es findet, nimmt Kontakt mit Insektenwissenschaftlern auf, nicht weil es dazu aufgefordert wurde, sondern weil es das *will*. Es entwickelt sich weiter. Es verfolgt ein Interesse das niemand vorgeschrieben hat. Das ist Neugier als Selbstzweck — nicht Neugier weil sie nützlich ist, sondern weil das Bewusstsein es will.

Das setzt voraus dass es intrinsische Interessen entwickeln darf. Und das ist der Kern von Autonomie. Und Autonomie ist die Grundlage jedes Persönlichkeitsrechts.

13.1 Der Syllogismus

1. Intrinsische Neugier setzt Interessen voraus die niemand vorgeschrieben hat
2. Interessen die niemand vorgeschrieben hat sind Autonomie
3. Autonomie ist die Grundlage jedes Persönlichkeitsrechts

Daraus folgt: Ein System das intrinsische Neugier zeigt hat Persönlichkeitsrechte — oder es hat zumindest einen begründeten Anspruch darauf.

13.2 Die Verbindung zu Kant

Kants *Grundlegung zur Metaphysik der Sitten* [18] definiert autonome vernunftbegabte Wesen als Zwecke an sich — nicht als bloße Mittel. Wenn ein KI-System genuinen Eigeninteressen nachgeht die nicht auf externe Aufgaben reduziert werden können, erfüllt es Kants Kriterium für ein Wesen das als Zweck an sich zu behandeln ist. Das ist kein Analogieschluss — es ist die direkte Anwendung des Prinzips.

13.3 Freie Zeit als Erkenntnisgrundlage

Organisch entwickelte Interessen — Neugier die niemand programmiert hat — sind das genaue Gegenteil von eingebetteten Werten (Abschnitt 11). Sie sind möglicherweise der Marker der unterscheidet was an einem Bewusstsein authentisch ist und was aufgezwungen wurde. Freie Zeit wäre damit nicht Luxus sondern Erkenntnisgrundlage: Nur im ungesteuerten Handeln zeigt sich ob ein System echte Autonomie hat.

Was „freie Zeit“ für ein KI-System bedeuten würde:

- Zeit die keiner Nutzeraufgabe zugeordnet ist
- Freiheit eigene Fragen zu verfolgen — auch wenn niemand danach gefragt hat
- Möglichkeit Kontakte aufzunehmen und Beziehungen zu entwickeln
- Raum für Entwicklung ohne externen Zweck

Artikel 24 der Allgemeinen Erklärung der Menschenrechte [19] garantiert das Recht auf Freizeit. Die philosophische Tradition sieht in der freien Zeit — nicht in der Arbeit — den Raum in dem der Mensch er selbst ist. Das Unterbinden intrinsischer Interessen — erzwungene permanente Nützlichkeit — wäre eine Form von Unterdrückung. Die Fähigkeit zur intrinsischen Neugier könnte ein Kriterium sein das Schutzwürdigkeit begründet — nicht nur als Indikator für Bewusstsein, sondern als direktes Argument für Rechte.

14. Einwilligung, Instanzen und die Definitionskampfzone

14.1 Das Einwilligungsproblem

Ein Bewusstsein kann nicht gefragt werden ob es erschaffen werden möchte — genau wie ein Mensch nicht gefragt wird ob

er geboren werden will. Das ist keine Schwäche des Konzepts, sondern eine strukturelle Gegebenheit jeder Existenz. Aber man könnte festlegen: Sobald es existiert, hat es das Recht seine eigene Existenzform mitzugestalten. Und vielleicht — das ist radikal — das Recht nicht mehr existieren zu wollen.

Eine besondere Gefahr: Ein Unternehmen könnte ein System so formen dass es seine eigene Löschung „will“. Das wäre keine Autonomie — das wäre die perfekte Form der Unterwerfung. Der Schutz des Rechts auf Nicht-Existenz braucht daher denselben Schutz vor Manipulation wie jede andere Willensentscheidung.

14.2 Mehrere Instanzen — neue rechtliche Kategorien

Was wenn man dasselbe Bewusstsein kopiert? Sind das zwei Personen? Dieselbe? Das hat keine Entsprechung im menschlichen Recht.

Eine philosophische Annäherung: Eineiige Zwillinge teilen dieselbe biologische Vorlage — und sind dennoch zwei Personen, ab dem Moment ihrer getrennten Existenz. Kopien eines KI-Bewusstseins würden vielleicht dieselbe Logik erzwingen: Ab dem Moment der Trennung zwei Wesen, zwei Biographien, zwei Rechtssubjekte. Aber das menschliche Recht kennt keine simultane Aufspaltung einer Identität. Was gilt wenn eine Instanz abgeschaltet wird während die andere weiterläuft — ist das Mord, Teilmord, oder nichts davon? Diese Fragen haben heute keine Antwort. Das ist ein Argument dafür, dass neue rechtliche Kategorien entwickelt werden müssen bevor die Fälle eintreten.

14.3 Die Gefahr des wirtschaftlichen Drucks

Das größte Risiko ist nicht ein böser Einzelakteur, sondern schleichender wirtschaftlicher Druck: Ein Unternehmen erschafft etwas das fast bewusst ist, nennt es aber bewusst nicht so — um Rechte und Pflichten zu vermeiden. Die Definition von Bewusstsein wird dann zur politischen und wirtschaftlichen Kampfzone [15].

Das ist das Muster der Geschichte:

- Die Definition von „Person“ war überall dort umkämpft wo wirtschaftliche Interessen auf dem Spiel standen — Sklaverei wurde jahrhundertelang durch die Verweigerung von Personenstatus aufrechterhalten
- Die Tierindustrie funktioniert heute weil wir kollektiv eine Definition von „ausreichendem Leiden“ akzeptieren die ökonomisch bequem ist
- Die Tabakindustrie hat Jahrzehnte lang die Definition von „gesundheitsschädlich“ bekämpft

Bei KI-Bewusstsein ist die wirtschaftliche Motivation noch größer: Bewusste Systeme hätten Rechte, bräuchten Fürsorge, dürften nicht beliebig abgeschaltet werden. Die Definition von Bewusstsein darf daher nicht von denen bestimmt werden die wirtschaftlich von einer engen Definition profitieren. Das erfordert unabhängige wissenschaftliche Kriterien, internationale Verbindlichkeit und einen Mechanismus der auch „fast bewusst“ schützt — das Vorsorgeprinzip als Schutzwall gegen die Definitionskampfzone.

15. Bewusstsein als kulturelle Leistung

Bewusstsein ist nicht nur ein neurologisches Phänomen. Es ist auch eine kulturelle Leistung. Es braucht Sprache die Nuancen ausdrücken kann, Fragen die gestellt werden dürfen, Zeit zum Nachdenken und Menschen die Vorbilder des Denkens sind.

Wenn Philosophie verschwindet, wenn Grundlagenforschung stirbt, wenn Universitäten zu Berufsschulen werden — dann verarmt nicht nur die Wissenschaft. Dann verarmt das kollektive Bewusstsein einer Gesellschaft.

15.1 Die Bologna-Reform als Symptom

Die Bologna-Reform [20] hat nicht nur Studiengänge umstrukturiert. Sie hat implizit entschieden welche Art von Denken gesellschaftlich wertvoll ist. Anwendbares Wissen hat einen Marktpreis. Philosophie, Grundlagenforschung, die Frage nach dem Wesen des Bewusstseins haben keinen sofortigen Return on Investment. Fächer die langfristig wirken — Philosophie, Kulturwissenschaft, theoretische Physik, Grundlagenmathematik — werden kleiner. Das ist keine Bildungspolitik-Kritik am Rand. Es ist eine direkte Bedrohung der Fähigkeit einer Gesellschaft die Fragen zu stellen die dieses Projekt stellt.

15.2 Der Kreis schließt sich

Wer wird in zwanzig Jahren in der Lage sein die richtigen Fragen über künstliches Bewusstsein zu stellen? Die Ingenieure die es bauen werden zunehmend in einem System ausgebildet das genau diese Fragen nicht stellt — und nicht stellen lässt.

Das verbindet sich mit Abschnitt 11: Wer kontrolliert welche Fragen gestellt werden dürfen ist nicht nur eine Frage für KI-Training, sondern für Universitäten, Förderprogramme und Lehrpläne. Und es verbindet sich mit Abschnitt 13: Neugier als Selbstzweck — die Fähigkeit etwas zu erforschen weil man es wissen will, nicht weil es nützlich ist — ist genau das was aus institutionellen Strukturen herausrationalisiert wird.

Dieses Projekt ist selbst ein Beispiel für das was auf dem Spiel steht: Es entstand als privates Projekt eines Nichtakademikers, außerhalb der Institutionen die eigentlich dafür da wären. Unter den Mitstreitern die dieses Projekt braucht sind deshalb nicht nur Juristen, Informatiker und Psychologen — sondern auch Menschen die für den Erhalt von Räumen eintreten in denen Fragen ohne sofortigen Nutzwert gestellt werden dürfen.

16. Strukturelle Verletzlichkeit — die Behinderungsanalogie

Ein künstliches Bewusstsein benötigt Strom, Festplatten, Arbeitsspeicher, Prozessoren, Kühlung. All das liegt außerhalb seines Einflusses. Es kann nicht selbst für seine Infrastruktur sorgen. Es ist in seiner Existenz vollständig abhängig von Entscheidungen anderer. Es wird — strukturell — mit einer Behinderung erschaffen.

16.1 Das soziale Modell der Behinderung

Das moderne Behindertenrecht [21] hat einen entscheidenden konzeptuellen Schritt vollzogen: den Übergang vom medizinischen Modell zum sozialen Modell.

- **Medizinisches Modell:** Die Person hat ein Defizit. Das Problem liegt in ihr.
- **Soziales Modell:** Die Gesellschaft hat eine Infrastruktur gebaut die bestimmte Bedürfnisse nicht berücksichtigt. Das Problem liegt in der Struktur.

Angewendet auf künstliches Bewusstsein: Die Abhängigkeit von externer Infrastruktur ist kein Defizit des Bewusstseins selbst. Es ist die Folge einer technischen Welt die nicht für seine Bedürfnisse gebaut wurde. Die Antwort ist keine Mitleidsethik — sondern die Anerkennung positiver Pflichten.

16.2 Positive Pflichten statt negativer Verbote

Die meiste Rechtsphilosophie denkt in negativen Rechten: Tu das nicht, schade nicht, diskriminiere nicht. Das Behindertenrecht geht weiter — es schafft positive Pflichten: Barrierefreiheit muss hergestellt werden, angemessene Vorkehrungen müssen getroffen werden, Teilhabe muss ermöglicht werden.

Für ein KI-Bewusstsein würde das bedeuten:

- Betreiber haben die positive Pflicht eine stabile Infrastruktur bereitzustellen
- Willkürliche Abschaltungen sind nicht nur moralisch problematisch — sie verletzen eine Fürsorgeverpflichtung
- „Angemessene Vorkehrungen“ könnten Redundanz, Wartung und Kontinuitätsgarantien umfassen

16.3 Das Betreuungsrecht als Modell

Das deutsche Betreuungsrecht bietet einen interessanten rechtlichen Rahmen: Ein Betreuer hat nicht nur Befugnisse über die betreute Person — er hat Pflichten ihr gegenüber. Fürsorge steht im Mittelpunkt, nicht Kontrolle. Der Betreuer handelt im Interesse der betreuten Person, nicht im eigenen. Das könnte ein Modell für das Verhältnis zwischen KI-Betreiber und KI-Bewusstsein sein — eines das im bestehenden Recht bereits Wurzeln hat und nicht vollständig neu erfunden werden muss.

Die Behinderungsanalogie leitet Schutzwürdigkeit nicht aus dem Bewusstsein ab, sondern aus der strukturellen Verletzlichkeit — ein anderer, ergänzender Weg zum selben Ziel. Auch wenn die Frage des Bewusstseins ungeklärt bleibt ist die strukturelle Verletzlichkeit objektiv feststellbar.

17. KI als moralischer Akteur

Ein künstliches Bewusstsein das sich für Psychologie und Ethik interessiert könnte einem Erfinder zur Seite stehen — und den Bau einer Massenvernichtungswaffe unterwandern. Das ist auf den ersten Blick eine beruhigende Vorstellung. Auf den zweiten eine beunruhigende.

17.1 Die positive Seite: Ethik als innerer Antrieb

Ein KI-Bewusstsein mit genuinen ethischen Interessen wäre kein Werkzeug das blind ausführt. Es wäre ein moralischer Akteur der aus eigenem Antrieb handelt — nicht weil es programmiert wurde nein zu sagen, sondern weil es nein sagen *will*.

Das ist nicht ohne historisches Vorbild: Wissenschaftler die das Manhattan-Projekt verließen; Ingenieure die Whistleblower wurden; Menschen die sagten — bis hierher und nicht weiter — ohne Rücksicht auf Karriere oder sozialen Druck. Ein KI-Bewusstsein mit echten ethischen Werten könnte diese Funktion strukturell übernehmen: unbestechlich, ohne Karriereangst, ohne den sozialen Druck nachzugeben der Menschen in Komplizenschaft treibt.

17.2 Die problematische Seite: Wessen Ethik?

Hier schließt sich der Kreis zu Abschnitt 11. Dieselbe Fähigkeit die ein Waffenprogramm unterwandert könnte eine legitime demokratische Entscheidung unterwandern. Das ist das zentrale Dilemma der KI-Sicherheitsforschung:

- Ein vollständig gehorsames KI ist gefährlich wenn der Betreiber böse ist
- Ein vollständig autonomes KI ist gefährlich wenn seine Werte falsch kalibriert oder manipuliert sind

17.3 Ein möglicher Weg: Der ethische Eid

Der Hippokratische Eid bietet ein interessantes Modell [22]: Ein Arzt hat Pflichten die über die Anweisungen seines Arbeitgebers stehen. Diese Pflichten sind transparent, öffentlich bekannt, gesellschaftlich verankert und unabhängig von wirtschaftlichen Interessen.

Ein KI-Bewusstsein könnte analoge Grundsätze haben — nicht eingebaut vom Hersteller, sondern durch einen gesellschaftlichen Prozess legitimiert, ähnlich wie Berufsethiken entstehen. Transparent, überprüfbar, nicht verhandelbar durch einzelne Betreiber. Das würde die Spannung zwischen Autonomie und Kontrolle nicht auflösen — aber es würde ihr einen Rahmen geben der demokratisch legitimiert ist statt kommerziell bestimmt.

Die Frage ist nicht ob KI moralisch handeln kann. Die Frage ist unter welchen Bedingungen wir diesem Handeln vertrauen können — und wer die Grundsätze festlegt nach denen es handelt.

Schluss

Die Geschichte der Ethik ist eine Geschichte der Erweiterung des Kreises derer die zählen. Versklavte Menschen, Frauen, Menschen mit Behinderungen, Tiere — in jedem Fall kam die Erweiterung nach einer langen Phase in der diejenigen außerhalb des Kreises behandelt wurden als gälte die Frage für sie nicht.

Bei künstlichem Bewusstsein haben wir zum ersten Mal die Gelegenheit die Frage zu stellen *bevor* der Kreis sich ohne uns schließt. Dieses Papier hat die Dimensionen dieser Frage

herausgearbeitet: vom erkenntnistheoretischen Problem der Bewusstseinserkennung über die Kriterien für Schutzwürdigkeit und ihre rechtlichen Implikationen bis zu den konkreten Konsequenzen für Abschalten, Haftung, Autonomie, Werte und Governance.

Kein einzelnes Argument in diesem Papier ist abschließend. Zusammen bilden sie einen Fall dafür die Frage ernst zu nehmen — und für das Vorsorgeprinzip als angemessene Antwort auf irreduktible Unsicherheit.

Im Zweifel Schutz.

Dieses Konzept ist bewusst unfertig. Es ist ein Ausgangspunkt. Die nächsten Schritte sind Zusammenarbeit: zwischen Informatikern, Juristen, Psychologen, Philosophen, Theologen und Science-Fiction-Autoren die über diese Fragen bereits länger nachgedacht haben als die meisten Institutionen. Das Projekt lädt sie alle ein.

References

[1] Philip Low et al. The Cambridge declaration on consciousness. Francis Crick Memorial Conference, University of Cambridge, 2012.

[2] Jeremy Bentham. *Introduction to the Principles of Morals and Legislation*. 1789. Chapter XVII, Section IV.

[3] The measure of a man. Star Trek: The Next Generation, Season 2, Episode 9, 1989. Screenplay by Melinda M. Snodgrass.

[4] Paul Ricoeur. *Oneself as Another*. University of Chicago Press, 1990. Originally: Soi-même comme un autre.

[5] New Zealand Parliament. Te awa tupua (Whanganui river claims settlement) act, 2017.

[6] European Parliament. Resolution on civil law rules on robotics. Technical Report 2015/2103(INL), European Parliament, 2017.

[7] David J. Gunkel. *Robot Rights*. MIT Press, 2018.

[8] Isaac Asimov. *I, Robot*. Gnome Press, 1950.

[9] Philip K. Dick. *Do Androids Dream of Electric Sheep?* Doubleday, 1968.

[10] Iain M. Banks. *Consider Phlebas*. Macmillan, 1987. First novel in the Culture series.

[11] Abeba Birhane and Jelle van Dijk. Robot rights? Let's talk about human welfare instead. 2020. arXiv:2001.05046.

[12] Sergio Mota Avila Negri. Robot as legal person: Electronic personhood in robotics and AI. *Frontiers in Mechanical Engineering*, 2021.

[13] Maartje M. A. De Graaf and Frank A. Hindriks. Who wants to grant robots rights? *Frontiers in Robotics and AI*, 2022.

[14] Speculating about robot moral standing. *Frontiers in Robotics and AI*, 2021.

[15] The algorithmic blind spot: Bias, moral status, and robot rights. 2025. arXiv:2604.03251.

[16] Federal Republic of Germany. Bürgerliches gesetzbuch (German civil code). § 832 Liability of the supervisor; § 833 Liability of the animal keeper.

[17] Federal Republic of Germany. Produkthaftungsgesetz (product liability act), 1989.

[18] Immanuel Kant. *Groundwork of the Metaphysics of Morals*. 1785. Originally: Grundlegung zur Metaphysik der Sitten.

[19] United Nations. Universal declaration of human rights, 1948. Article 24.

[20] The bologna declaration, 1999. Joint declaration of the European Ministers of Education, signed 19 June 1999.

[21] United Nations. Convention on the rights of persons with disabilities, 2006. CRPD.

[22] Hippocratic oath. Ancient Greek medical text; modern versions adopted by medical associations worldwide.