**Sind Maschinen *wirklich* intelligent?**

*Bert Heinrichs und Ulrich Steckmann*

**Abstract**

In Diskussionen über künstliche Intelligenz wird regelmäßig die Frage gestellt, ob Maschinen *wirklich* intelligent sind. Zur Klärung dieser Frage greifen wir in diesem Beitrag eine Definition von Intelligenz auf, die Dimitri Coelho Mollo vorgeschlagen hat. Wir schlagen eine Differenzierung vor, die dazu führt, dass man zwischen zwei Arten von Intelligenz unterscheiden muss. Wir argumentieren dann, dass KI zwar über die eine Art von Intelligenz verfügen kann, über die andere hingegen (noch) nicht. Abschließend beschäftigen wir uns mit der Frage, unter welchen Bedingungen wir Maschinen auch die zweite Art von Intelligenz zuschreiben würden. Wir erwägen, ob die Zuschreibung von *originärer Intelligenz* womöglich in einer sozialen Praxis der Anerkennung begründet sein könnte.

**1. Einleitung**

In Diskussionen über künstliche Intelligenz (KI) wird regelmäßig die Frage gestellt: „Sind diese Maschinen *wirklich* intelligent?“. Eine ebenso naheliegende wie unbefriedigende Antwort lautet: „Das kommt darauf an, was man unter Intelligenz versteht.“ Richtig an dieser Antwort ist, dass sich die Frage nach dem Vorliegen einer Fähigkeit nur dann beantworten lässt, wenn man sich zunächst Klarheit darüber verschafft hat, was man unter dieser Fähigkeit genau versteht. Und richtig ist auch, dass das beim Begriff ‚Intelligenz‘ keineswegs leicht ist. Wie so oft in der Philosophie, muss man also zunächst über Begriffe sprechen, wenn man die Frage, ob Maschinen *wirklich* intelligent sind, klären will.

In diesem Beitrag werden wir zunächst eine Definition von Intelligenz aufgreifen, die Dimitri Coelho Mollo vorgeschlagen hat. Uns scheint Coelho Mollos Ansatz überaus vielversprechend zu sein. In einer Hinsicht halten wir ihn allerdings für unzureichend. Wir werden daher eine Differenzierung vorschlagen, die dazu führt, dass man zwischen zwei Arten von Intelligenz unterscheiden muss. Wir werden dann dafür argumentieren, dass KI zwar über die eine Art von Intelligenz verfügen kann, über die andere hingegen (zumindest bislang) nicht. Mit dem vorgeschlagenen Verständnis von Intelligenz kann man also eine klare Antwort auf die Ausgangsfrage – „Sind diese Maschinen *wirklich* intelligent?“ – geben. Es wird aber darüber hinaus auch verständlich, warum die Frage so beständig gestellt wird und uns immer wieder rätseln lässt. Es liegt schlicht daran, dass wir gewöhnlich nicht zwischen den beiden Formen von Intelligenz unterscheiden und es für uns zwar einerseits so scheint, als ob Maschinen intelligent seien, wir mit dieser Auffassung andererseits aber nicht ganz zufrieden sind. Abschließend werden wir uns mit der Frage beschäftigen, unter welchen Bedingungen wir Maschinen auch die zweite Art von Intelligenz zuschreiben würden. Unsere tentative Antwort wird zeigen, dass wir in einem weiteren Punkt von Coelho Mollos Ansatz abweichen.

**2. Coelho Mollos Bemühung um einen einheitlichen Begriff von Intelligenz**

In seinem Beitrag „Intelligent Behaviour“ von 2022 geht Coelho Mollo von der Beobachtung aus, dass der Intelligenzbegriff notorisch schwierig zu definieren sei (1-2).[[1]](#footnote-1) Als Grund dafür macht er u. a. den Umstand aus, dass der Begriff in vielen unterschiedlichen Kontexten – prätheoretischen wie wissenschaftlichen – und vielen verschiedenen Disziplinen verwendet werde. Während in der Psychologie sowie in Teilen der KI-Forschung das Bestreben vorrangig sei, Intelligenz zu messen und quantifizierbar zu machen, stünden in der Philosophie und einigen Bereichen der Kognitionsforschung eher Bemühungen um begriffliche Abgrenzung, insbesondere zu den Begriffen ‚Kognition‘ und ‚Rationalität‘, im Vordergrund. Angesichts dieser Vielstimmigkeit spreche einiges dafür, dass der Begriff der Intelligenz schlicht zu vage sei, um als wissenschaftlicher Begriff Verwendung zu finden. Dessen ungeachtet hält Coelho Mollo es für möglich, zu einem klaren, wissenschaftlich und philosophisch fruchtbaren Konzept zu gelangen. Er benennt dazu fünf Desiderate, die ein überzeugender Begriff der Intelligenz erfüllen müsse, und unternimmt es anschließend, anhand dieser Vorgaben eine Definition von Intelligenz zu erarbeiten.

Die fünf Desiderate, die Coelho Mollo an den Anfang seiner Überlegungen stellt, lauten wie folgt (2-4): *Species-neutrality*, *Origin-neutrality*, *Multiple realisability*, *Epistemic distinctiveness*, *Responsiveness to scientific practice.* Das erste Desiderat bezeichnet Coelho Mollo auch als „non-anthropocentrism *desideratum*“ (3). Ein überzeugender Begriff der Intelligenz sollte demnach nicht den Menschen bzw. typische menschliche Fähigkeiten als Standard ansetzen, auch wenn das traditionell häufig so gemacht wurde und auch heute noch verbreitet ist, nicht zuletzt, wenn behauptet wird, in der KI-Forschung gehe es darum, menschliche Intelligenz zu simulieren oder Maschinen dazu zu bringen, Dinge zu tun, die bislang nur Menschen tun können. Eine solche Herangehensweise hält Coelho Mollo vor allem deshalb für nicht zielführend, weil sie einem über Disziplingrenzen hinweg einheitlichen Begriff von Intelligenz im Wege stünde. Stattdessen sollte ein umfassender Begriff von Intelligenz eine Fähigkeit beschreiben, die von Menschen, nicht-menschlichen Tieren und Maschinen grundsätzlich gleichermaßen realisiert werden könne.

Damit eng zusammen hängt das zweite Desiderat, das Neutralität hinsichtlich des Ursprungs von Intelligenz fordert. Insbesondere sollten, so Coelho Mollo, biologische Wesen, also Menschen und nicht-menschliche Tiere, und ihre evolutionäre Genese nicht von vornherein gegenüber nicht-biologischen Systemen, also Maschinen, privilegiert werden. Der Intelligenzbegriff sollte also so angelegt sein, dass er von der Kategorie ‚Lebewesen‘ unabhängig ist.

Auch über die Art der strukturellen Realisierung von Intelligenz sollte auf begrifflicher Ebene keine Vorentscheidung getroffen werden. Zwar könne es sich herausstellen, dass tatsächlich eine bestimmte Art der Realisierung zwingend für Intelligenz sei. Dies sei aber eine empirische Einsicht, die am Ende von Forschungsbemühungen stehen müsse, und keine begriffliche Wahrheit, die deren Auftakt bilden sollte. Daher das dritte Desiderat.

Das vierte Desiderat verlangt eine klare theoretische und exploratorische Rolle für den Begriff der Intelligenz in den Wissenschaften. Insbesondere müsse deutlich werden, wie er sich von verwandten Begriffen wie ‚Kognition‘ und ‚Rationalität‘ unterscheide. Denn nur so könne er sowohl theoretische wie auch experimentelle Forschung anleiten und befördern. Es gilt also auch mit Blick auf den Intelligenzbegriff die Maxime der Begriffsökonomie, die besagt, dass man keine unnötigen oder redundanten Begriffe in der Wissenschaft verwenden sollte. Wenn sich also herausstellen würde, dass man mit den Begriffen ‚Kognition‘ und ‚Rationalität‘ relevante Fähigkeiten klar und umfassend beschreiben kann, dann könnte es sein, dass der Begriff der Intelligenz sich schlicht als überflüssig erweist und folglich als wissenschaftlicher Begriff preisgegeben werden sollte.

Schließlich fordert Coelho Mollo mit dem fünften Desiderat, dass ein Begriff von Intelligenz nicht übermäßig revisionistisch sein dürfe. Zwar könne man nicht alle herkömmlichen Charakterisierungen aufgreifen, die teilweise vage und widersprüchlich sind. Umgekehrt müsse eine überzeugende Definition von Intelligenz, die über Disziplinen und Diskurse hinweg akzeptiert werden könne, aber an geteilte Auffassungen anknüpfen und diese zumindest partiell bewahren.

Diese letzte Überlegung führt Coelho Mollo dazu, einige allgemeine Elemente zu identifizieren, die sich seiner Meinung nach in vielen Definitionen von Intelligenz finden lassen (5). Es handelt sich dabei gerade um solche Elemente, die eine neue Definition von Intelligenz bewahren muss, wenn sie nicht Gefahr laufen will, übermäßig revisionistisch zu sein. Nach Coelho Mollo gehören dazu: *Generality*, *Flexibility*, *Goal-directedness*, *Adaptivity.*

Als Zwischenergebnis ergibt sich demnach also Folgendes: Wie immer man Intelligenz definieren will, es handelt sich jedenfalls um eine Fähigkeit, die

* angemessenes Verhalten bei einer Vielzahl unterschiedlicher Aufgaben und Umstände ermöglicht,
* eine Anpassung des Verhaltens an sich ändernde und/oder unsichere Umstände erlaubt,
* die Bildung und Verfolgung von den Umständen angemessenen Zielen gestattet,
* Information über gegenwärtiges und zukünftiges Verhalten im Lichte vergangener Interaktionen mit der Welt berücksichtigt.

Die ersten drei der oben genannten Desiderate legen Coelho Mollo zufolge eine rein behaviorale Definition von Intelligenz nahe, also eine, die allein auf beobachtbares Verhalten abstellt. Ein solcher Zugang hat zudem den Vorteil, dass die Überprüfung und die Zuschreibung von Intelligenz zu Problemen werden, die sich mit empirischen Methoden bearbeiten lassen.

Nach diesen Vorüberlegungen präsentiert Coelho Mollo schließlich seine Definition von Intelligenz. Sie lautet:

An intelligent system *S* is a system that manifests behaviours that are often enough

* *general*, i.e. that are appropriate in a variety of different circumstances;
* *flexible*, i.e. that change appropriately in light of changed, novel, or uncertain circumstances;
* *goal-directed*, i.e. that are appropriate in light of goals *S* plausibly possesses;
* *adaptive*, i.e. that change appropriately in light of previous interactions with the world. (6)

Im weiteren Verlauf seines Beitrags versucht Coelho Mollo zu zeigen, dass seine Definition von Intelligenz die eingangs formulierten Desiderata alle erfüllt. Er verwendet dabei vor allem viel Zeit auf das vierte Desiderat, also die Forderung nach „epistemic distinctiveness“, und untersucht dazu das Verhältnis von Intelligenz zu Kognition und Rationalität (8-12). Darüber hinaus widmet er sich auch ausführlich der Frage der multiplen Realisierbarkeit (12-15). Diese Überlegungen werden wir nicht im Detail nachzeichnen, kommen gleich aber noch einmal darauf zu sprechen. Wir möchten uns Coelho Mollos Definition nun etwas genauer ansehen und schließlich ein Element herausgreifen, das uns unterentwickelt zu sein scheint.

**3. Intelligenz als generelle, flexible und adaptive Fähigkeit**

Schaut man sich die Definition von Intelligenz an, die Coelho Mollo vorschlägt, dann kann man ohne Weiteres zugestehen, dass sie einem Alltagsverständnis von Intelligenz sehr nahekommt. Typischerweise verbinden wir mit Intelligenz einerseits ein hohes Maß an Flexibilität und andererseits die Fähigkeit zu lernen. Intelligent nennen wir ein Wesen oder ein System dann, wenn es sich auf unbekannte Umstände einstellen und aus Erfahrungen lernen kann. Dem trägt Coelho Mollo Rechnung, indem er Generalität und Flexibilität einerseits und Adaptivität andererseits als wesentliche Teilfähigkeiten in seine Definition von Intelligenz aufnimmt. Alle drei Teilfähigkeiten sind aus vielen gängigen Definitionen bekannt. Coelho Mollo erfüllt also seine eigene Forderung nach einer nicht-revisionistischen Definition.

Instruktiv ist, wie Coelho Mollo Intelligenz von Kognition einerseits und Rationalität andererseits abgrenzt: Kognition versteht er als einen prozessbasierten Begriff und lässt offen, ob es sich um Rechenprozesse (mit Repräsentationen) oder Prozesse der Selbsterhaltung und Selbstorganisation handelt. Coelho Mollo deutet Kognition mithin als einen „anspruchslosen“ Begriff, der sich bspw. auch auf Pflanzen anwenden lässt (10). Demgegenüber deutet Coelho Mollo Rationalität als sehr viel engeres Vermögen, das mit Vorstellungen von Optimalität oder Normativität verbunden ist, wobei diese allerdings sehr unterschiedlichen Quellen entspringen können. Insbesondere muss Rationalität nicht generalistisch sein, wie es Intelligenz stets ist. So gesehen, sind Schach- oder Go-Programme zwar „pointillistisch rational“, aber eben nicht generalistisch intelligent (11). Coelho Mollo verortet Intelligenz mithin zwischen Kognition und Rationalität. Seiner Meinung nach sind einige intelligente Systeme auch kognitive Systeme, und einige intelligente Systeme sind auch rationale Systeme. Er bleibt indes neutral in der Frage, ob alle intelligenten Systeme kognitive Systeme sind. All das ist hilfreich und konstruktiv.

Zu erwägen ist, ob eine überzeugende Definition von Intelligenz noch eine Einschränkung enthalten muss, die die zur Verfügung stehenden Ressourcen begrenzt. Eine solche Einschränkung findet sich etwa bei Pei Wang, der schreibt: „Intelligence is the capacity of an information-processing system to adapt to its environment while operating with insufficient knowledge and resources.“[[2]](#footnote-2)

Wie Wang zeigt, lässt sich die Einschränkung „with insufficient knowledge and resources“ näher dahin spezifizieren, dass ein intelligentes System endlich und offen sein und in Echtzeit operieren muss. Gerade mit Blick auf künstliche Systeme hat das womöglich erhebliche Konsequenzen. Wir möchten auf diesen Problemkomplex hier aber nur kurz hinweisen und werden ihn im weiteren Verlauf unberücksichtigt lassen.

Was bislang auch noch nicht zur Sprache gekommen ist, ist das Element der Zielgerichtetheit, das Coelho Mollo in seiner Definition von Intelligenz aufführt. Diesem Element werden wir uns nun zuwenden.

**4. Zielgerichtetheit**

Es fällt auf, dass die vier Elemente, die Coelho Mollo in seiner Definition von Intelligenz aufführt, nicht alle auf der gleichen Ebene angesiedelt zu sein scheinen. Während Generalität, Flexibilität und Adaptivität gleichsam ein *wie* beschreiben – nämlich, wie man ein Ziel verfolgt – deutet Zielgerichtetheit auf ein *dass* hin – nämlich, dass überhaupt ein Ziel verfolgt wird. Vielleicht wäre es sogar unbemerkt geblieben, wenn Coelho Mollo dieses vierte Element ganz weggelassen hätte und lediglich implizit davon ausgegangen wäre, dass ein intelligentes Wesen oder ein System stets Ziele verfolgt. Intelligenz scheint nämlich stets eine Art und Weise zu bezeichnen, wie Ziele verfolgt werden, und dabei immer schon vorauszusetzen, dass Ziele verfolgt werden. Bei näherem Hinsehen scheint uns aber gerade diese Charakterisierung Fragen aufzuwerfen, die bei Coelho Mollo unberücksichtigt bleiben. Wir möchten diesen Umstand nun zunächst etwas näher beschreiben, bevor wir dann eine Analyse liefern, die schließlich zu einer begrifflichen Differenzierung führen wird.

Stellen wir uns einen aktuellen Schachcomputer vor (wir lassen hier unberücksichtigt, ob ein solches Gerät nach Coelho Mollos Definition intelligent und/oder rational ist). Solche Schachcomputer oder Schachsoftware, die auf gewöhnlichen PCs läuft, sind sehr gut darin, in einer beliebigen Stellung den nächsten besten Zug zu finden – viel besser als jeder Mensch. Dieses Verhalten kann man intelligent nennen. Zumindest in dem Rahmen, der durch die Schachregeln vorgegeben wird, können sich Schachprogramme auf neue Situationen einstellen und aus Erfahrungen lernen. Sie agieren also flexibel und adaptiv. Dennoch sind Bemerkungen wie die folgenden nicht ungewöhnlich: „Der Schachcomputer spielt nicht wirklich Schach. Es ist ihm egal, ob er gewinnt oder verliert. Er weiß gar nicht, was es bedeutet, ein Spiel zu spielen“ usw. Was steckt dahinter? Offenbar hat es etwas mit dem Element der Zielgerichtetheit zu tun, das in Coelho Mollos Definition von Intelligenz eine Rolle spielt.

Auf den Einwand, der Computer würde nicht wirklich Schach spielen, kann man entgegen, dass er mit dem Zug, den er auswählt, offenbar ein Ziel verfolgt, nämlich nach den anerkannten Regeln des Schachs zu gewinnen. Gerade weil sie eine optimale oder zumindest sehr gute Lösung für das Problem ist, wie man diese Schachpartie gewinnt, fassen wir die Zugauswahl als intelligent auf. Es scheint also unzutreffend zu sein, dass der Computer nicht gewinnen will. Diese Antwort kann leicht folgenden weiteren Einwand nach sich ziehen: „Das stimmt zwar. Der Computer folgt aber nur dem Ziel, das ihm die Programmierer eingepflanzt haben. Es ist also nicht *sein* Ziel, das Spiel zu gewinnen, sondern dasjenige der Programmierer.“ Man könnte dem wiederum entgegenhalten, dass der Schachcomputer inklusive der implementierten Zielfunktionen zwar das Produkt einer Programmierung sei, aber auch der Mensch und alle Lebewesen inklusive ihrer jeweiligen ‚Zielfunktionen‘ Produkte sind, und zwar der natürlichen und sozialen Evolution. Diese mögliche Replik ändert freilich nichts daran, dass man zwischen unterschiedlichen Arten von Zielgerichtetheit unterscheiden kann. Dies verstößt übrigens nicht gegen Coelho Mollos zweites Desiderat, also die Forderung nach „origin neutrality“. Das wäre lediglich dann der Fall, wenn man evolutionär bzw. sozial entstandene Zielsetzungen als notwendige Voraussetzung für intelligentes Verhalten ansehen würde. Es geht hier aber nicht darum, dem Schachcomputer intelligentes Verhalten abzusprechen, sondern lediglich darum, auf eine Differenz aufmerksam zu machen, die zwischen Schachcomputern und menschlichen Schachspieler:innen besteht: Während man im Fall des Schachcomputers von ‚abgeleiteter Zielgerichtetheit‘ sprechen kann, kann man im menschlichen Fall von ‚originärer Zielgerichtetheit‘ sprechen.

Diese Unterscheidung erinnert natürlich an eine ähnliche Unterscheidung, die in Diskussionen der Philosophie des Geistes eine wichtige Rolle spielt, nämlich die zwischen zwei Arten von Intentionalität, die John Searle und in etwas anderer Form auch John Haugeland vorgenommen haben.[[3]](#footnote-3) In seinem berühmten Chinese-Room-Argument hat Searle geltend gemacht, Maschinen könnten kein genuines Verständnis erlangen.[[4]](#footnote-4) All das hängt zusammen. Leider sind wir weit entfernt davon, eine Lösung oder auch nur den Ansatz einer Lösung für diesen großen Zusammenhang bieten zu können. Unser Anspruch ist wesentlich bescheidener und geht lediglich davon aus, dass es einen Unterschied zwischen originärer und abgeleiteter Zielgerichtetheit gibt.

**5. Zwei Arten von Intelligenz**

Wenn es stimmt, dass man zwischen zwei Arten von Zielgerichtetheit unterscheiden kann, nämlich zwischen einer solchen, die sich auf die Ziele eines anderen Wesens oder Systems zurückführen lässt, einerseits, und einer, bei der das nicht möglich ist, andererseits, dann hat dies unmittelbare Konsequenzen für den Begriff der Intelligenz. Wenn man nämlich mit Coelho Mollo davon ausgeht, dass Zielgerichtetheit ein elementarer Bestandteil von Intelligenz ist, dann setzt sich die Unterscheidung von Arten der Zielgerichtetheit, oder von *eigenen* und *vorgegebenen* Zielen, wie man auch sagen kann, im Intelligenzbegriff fort und führt zu zwei Arten von Intelligenz. Während Wesen oder Systeme, die über die Fähigkeit verfügen, eigene Ziel flexibel und adaptiv zu verfolgen, als *originär-intelligent* bezeichnet werden können, kann man bei Wesen und Systemen, die über eine vergleichbare Fähigkeit verfügen, aber nur vorgegebene Ziele verfolgen können, lediglich von *abgeleiteter Intelligenz* sprechen. Diese Terminologie ist sicher nicht ganz glücklich, aber hoffentlich trotzdem hinreichend klar. Was bringt sie uns?

Sie versetzt uns einerseits in die Position, die ursprüngliche Frage, nämlich ob (einige) Maschinen *wirklich* intelligent sind, beantworten zu können. Wesen oder Systeme, die generelles, flexibles und adaptives Verhalten bei der Verfolgung von Zielen zeigen, sind intelligent. Ohne Zweifel erfüllen schon viele heutige KI-Systeme dieses Kriterium. Wie beim Menschen, so wird man auch mit Blick auf diese Systeme sagen, dass sie in unterschiedlichem Maße intelligent sind. Ob es sinnvoll ist, dies in einem einzigen Wert zu messen, oder ob man besser unterschiedliche Dimensionen differenziert, lassen wir hier offen. Es scheint uns allerdings einiges dafür zu sprechen, dass Generalität, Flexibilität und Adaptivität sehr verschieden sind und es sowohl Systeme geben kann, die bspw. über ein hohes Maß an Generalität verfügen, gleichzeitig aber nicht sehr adaptiv sind, als auch umgekehrt. Bei Menschen kennt man das Phänomen der Inselbegabung, das sich in einem einzelnen quantitativen Intelligenzscore nicht gut abbilden lässt. Wie dem auch sei, geht man von den drei genannten Kriterien aus, dann sind einige aktuelle KI-Systeme sicher *abgeleitet intelligent*.

Keinem der aktuellen KI-Systeme wird man indes *originäre Intelligenz* attestieren können, da sie bislang stets von den Zielsetzungen menschlicher Programmierer:innen abhängig sind. Dies gilt auch dann, wenn ein System innerhalb eines übergeordneten Ziels eventuell flexibel neue Strategien entwickelt, die eigenständige Binnenzielsetzungen umfassen. In diesem Fall stimmt es zwar, dass sich ein solches Binnenziel nicht direkt auf eine:n Programmierer:in zurückführen lässt. Das ändert aber nichts an der Tatsache, dass die übergeordnete Zielsetzung vorgegeben bleibt und damit den Sinngehalt etwaiger neuer Binnenziele vollständig bestimmt.

Die Ausgangsfrage, ob KI-Systeme *wirklich* intelligent sind, muss man also präzisieren, indem man zwei neue Fragen stellt, nämlich 1. „Sind KI-Systeme abgeleitet intelligent?“ (Auf diese Frage lautet die Antwort „Ja“) und 2. „Sind KI-Systeme originär-intelligent?“ (Auf diese Frage lautet die Antwort „Nein“).

Der Umstand, dass es zwei unterschiedliche Antworten gibt, kann helfen zu verstehen, warum sich die Frage so hartnäckig hält. Solange man nur über den undifferenzierten Intelligenzbegriff verfügt, ist man teilweise geneigt, die ursprüngliche Frage positiv zu beantworten, man hält die Antwort aber teilweise auch für unbefriedigend. Mit der Differenzierung sieht man jetzt, dass dieses anhaltende Unbehagen seinen Grund in der Sache selbst hat. Sowohl die positive als auch die negative Antwort erweisen sich als falsch. Was nötig ist, ist ein Intelligenzbegriff, der eine differenzierte Antwort erlaubt, und genau einen solchen Intelligenzbegriff liefert Coelho Mollo, wenn man ihn um die Differenzierung von unterschiedlichen Arten der Zielgerichtetheit bzw. von eigenen und vorgegebenen Zielen erweitert.

Wie so oft, führt auch diese Antwort auf eine neue Frage: Was müsste ein KI-System für Verhalten an den Tag legen, damit wir ihm originäre Intelligenz zubilligen? Die Antwort scheint zunächst einfach zu sein: Ein solches System müsste Ziele verfolgen, die sich nicht auf die Ziele seiner Programmierer:innen zurückführen lassen. Hier könnte ein Problem liegen, das wir zum Abschluss noch etwas beleuchten möchten. Wir kommen damit auf einen Punkt zurück, den wir eingangs bereits kurz erwähnt haben, nämlich Coelho Mollos Forderung, ein überzeugender Begriff von Intelligenz müsse sich auf beobachtbares Verhalten beschränken.

**6. Originär-intelligente Maschinen**

Stellen wir uns vor, Sie sehen eine Person, die Schach spielt und über den nächsten Zug nachdenkt. Nach einer Weile zieht sie eine Figur und setzt damit ihre:n Gegner:in matt. Nehmen wir zudem an, der Zug, den die Person gefunden hat, war wirklich sehr gut und keineswegs offensichtlich. Wir würden also sagen, die Person hat sich intelligent verhalten. Sie hat eine gute Lösung für ein schwieriges Problem gefunden. Nehmen wir weiter an, Sie würden zu der Person gehen und sagen: „Das war ein toller Zug!“ und die Person würde antworten: „Ja, aber ich wollte gar nicht spielen. X hat gesagt, ich müsse spielen.“ Wir würden dies kaum zum Anlass nehmen, der Person ihre Intelligenz abzusprechen. Trotzdem scheinen wir einräumen zu müssen, dass die Person in diesem Fall nicht *ihr eigenes* Ziel verfolgt hat, sondern das Ziel von X. Folgt man dieser Überlegung, dann ist es wohl sehr oft der Fall, dass Menschen *vorgegebene Ziele* verfolgen und nicht *eigene*. Wir würden in diesen Fällen aber sicher nicht sagen, dass diese Menschen nicht intelligent handeln. Mehr noch, es wäre sogar ziemlich absurd zu behaupten, dass ein und dieselbe Handlung intelligent zu nennen ist, wenn sich eine Person selbst dazu entschlossen hat, und nicht intelligent, wenn sie ein vorgegebenes Ziel einer anderen Person verfolgt. Stellt dies die obige Differenzierung nicht grundsätzlich in Frage?

So weit muss man nicht gehen. Die oben vorgeschlagene Differenzierung sieht ausdrücklich nicht vor, dass die Fähigkeit, vorgegebene Ziele zu verfolgen, nicht als Intelligenz zählt, sondern lediglich, dass wir unterscheiden müssen, ob es sich um *originäre* oder *abgeleitete* Intelligenz handelt. Mit anderen Worten: Auch Schachspieler:innen, die auf Anforderung hin spielen, agieren intelligent. Dennoch, die Brillanz eines Schachzugs scheint überhaupt nicht davon abzuhängen, ob sich die Person selbst zum Spiel entschlossen hat oder lediglich der Forderung eines anderen nachkommt. Der Hinweis, dass man sie weiterhin abgeleitet-intelligent nennen kann, bleibt daher unbefriedigend.

Wir denken trotzdem, dass die Unterscheidung relevant ist, und darüber hinaus auch, dass sie Einfluss auf die grundsätzliche Charakterisierung eines Wesens oder Systems hat. Es scheint uns zu stimmen, dass der Schachcomputer über eine *grundsätzlich andere* *Art von Intelligenz* verfügt, solange er *ausschließlich* abgeleitete Ziele verfolgt. Um dies behaupten zu können, muss man aber eine Zusatzannahme machen, die womöglich weitreichende Konsequenzen hat.

Betrachten wir dazu noch einmal den Schachspieler aus dem Beispiel. Wenn er sagt: „Ich wollte gar nicht spielen. X hat es mir vorgegeben“, dann beschreibt er zwar eine relevante Tatsache, er unterschlägt aber, dass er sich trotzdem dazu entschließen musste, der Forderung von X zu folgen. Allgemeiner kann man sagen, dass Menschen, wann immer sie den Vorgaben von anderen folgen, die vorgängige Entscheidung getroffen haben, den Vorgaben anderer zu folgen. Damit liegt die letzte Zielsetzung stets bei ihnen. Etwas anders formuliert bedeutet das: Als freie und verantwortliche Wesen folgen wir in letzter Konsequenz immer unseren eigenen Zielen.[[5]](#footnote-5) Das unterscheidet den menschlichen Schachspieler vom Computerprogramm, selbst wenn er eine Partie auf Anweisung eines anderen hin spielt.

Entscheidend ist nun folgende Überlegung: Ob unser Gegenüber ein freies Wesen ist, lässt sich nicht allein an seinem beobachtbaren Verhalten ablesen, sondern ist vielmehr eine Unterstellung unsererseits. Tatsächlich handelt es sich um einen *Akt der wechselseitigen Anerkennung*: Wir anerkennen ein Gegenüber dann und nur dann als Person, wenn wir Grund zur Annahme haben, dass auch er uns als Person anerkennt.[[6]](#footnote-6) (Natürlich muss das nicht kontinuierlich der Fall sein, sonst könnten wir schon Schlafende nicht als Personen betrachten. Darüber hinaus gibt es andere schwierige Fälle, bspw. Menschen, die sich im Wachkoma befinden.)

Trifft das aber zu, dann stellt sich ein grundlegendes Problem: Die Frage „Welches Verhalten müsste ein KI-System an den Tag legen, damit wir ihm originäre Intelligenz zubilligen?“ erweist sich als Sackgasse. Denn offenbar hängt der grundlegende Anerkennungsakt nicht von beobachtbarem Verhalten ab, sondern resultiert aus einer komplexen sozialen Praxis. Dies steht aber in Widerspruch zu Coelho Mollos Forderung, der Intelligenzbegriff müsse sich auf beobachtbares Verhalten beschränken. Man muss also offenbar zwischen zwei gleichermaßen unbefriedigenden Alternativen wählen: Entweder man akzeptiert die Differenzierung von zwei Arten der Zielgerichtetheit und nimmt in Kauf, dass sich originäre Intelligenz nicht allein durch beobachtbares Verhalten aufweisen lässt, sondern vielmehr mit einem anspruchsvollen Anerkennungsbegriff verknüpft ist, der in einer gemeinsamen Anerkennungspraxis gründet. Oder aber man lehnt die Differenzierung unterschiedlicher Arten von Zielgerichtetheit ab und beschränkt sich auf eine rein behaviorale Ebene. Dann kann man zum einen den Unterschied zwischen Schachcomputern und menschlichen Schachspieler:innen nicht einfangen (was man womöglich gar nicht schlimm findet). Was schwerer wiegt, ist, dass man erklären muss, was Zielgerichtetheit eigentlich sein soll und warum man bei rein mechanischen Vorgängen nicht von Zielgerichtetheit spricht.

**6. Anerkennung als menschliche Praxis**

Die geschilderte Alternative zweier gleichermaßen unbefriedigender Alternativen lässt im Hintergrund eine bekannte Konfliktsituation erkennen. Auf der einen Seite stehen dabei die Vertreter:innen einer Position, die reklamiert, dass es Sachverhalte der menschlichen Lebensform gebe, die mit szientifischen Methoden nicht oder zumindest nicht adäquat zu erfassen seien. Dies wird von szientifischer Seite in der Regel mit einem Obskurantismusvorwurf beantwortet. Dieser Konflikt hat wiederholt Ausdruck gefunden im wissenschaftstheoretischen Streit um die methodologische Eigenständigkeit der Geisteswissenschaften.[[7]](#footnote-7) Dieser Grundlagenstreit kann hier selbstverständlich nicht beigelegt werden, aber im Hinblick auf die Frage nach einer originären Zielgerichtetheit und ihrer epistemischen Zugänglichkeit kann zumindest versucht werden, dasjenige möglichst präzise kenntlich und damit plausibel zu machen, was dem szientifischen Blick womöglich systematisch entgeht. Die Theorietraditionen, an die dabei angeknüpft werden kann, sind zum einen die im Deutschen Idealismus von Fichte und vor allem von Hegel entwickelten Anerkennungstheorien sowie deren Wiederaufnahmen in der Gegenwartsphilosophie.[[8]](#footnote-8) Zum anderen sind phänomenologische Beiträge zu berücksichtigen, die der Aufklärung der Situation des Begegnens von Personen gewidmet sind.[[9]](#footnote-9)

Folgt man dem Gedanken, dass die originäre Intelligenz nicht im Zuge einer bloßen Tatsachenfeststellung zugänglich wird, sondern mindestens teilweise darauf beruht, dass ein Wesen in einem wechselseitigen Prozess als Person anerkannt wird, dieser Status also erst durch die Anerkennung erzeugt bzw. beglaubigt wird, dann droht ein naturwissenschaftlich formierter Zugriff einen relevanten Teil des Phänomens intelligenten Verhaltens zu eliminieren.

Etwas als jemanden, das heißt als Person anzuerkennen, ist ein Vorgang, der zum einen eine epistemische Seite hat: An den Äußerungen, am Ausdrucksverhalten wird etwas wahrgenommen, das beim Wahrnehmenden wiederum ein bestimmtes Ausdrucksverhalten hervorruft, das seinerseits vom anderen wahrgenommen und auf eine spezifische Weise beantwortet wird. Zum anderen verfügt dieser Vorgang über eine praktische Seite: Das reaktive Ausdrucksverhalten der Beteiligten impliziert, den jeweils Anderen als Quelle von Ansprüchen zu betrachten und entsprechend zu reagieren. Um Ansprüche, die auf ein bestimmtes reaktives Verhalten beim Anderen abzielen, handelt es sich zunächst einmal in einem sehr allgemeinen Sinne: Die Beteiligten einer Anerkennungsbeziehung betrachten sich wechselseitig und auch sich jeweils selbst als Quellen normativer Ansprüche oder, um es mit Robert Brandom auszudrücken, als Akteur:innen im Raum der Gründe.[[10]](#footnote-10) Darin liegt, grob gesagt, die Anerkennung. Und nur innerhalb der Sphäre von Anerkennungsbeziehungen ist originäre Intelligenz situiert.

Um den entscheidenden Unterschied zwischen dem Erkennen, das im Anerkennen enthalten ist, und szientifischen Erkenntnisweisen weiter zu schärfen: Die Äußerung einer Absicht oder eines Vorhabens wird im Rahmen einer Anerkennungsbeziehung als das Eingehen bestimmter normativer Verpflichtungen angesehen, auf die das Gegenüber mit bestimmten Forderungen reagieren kann. Die Beteiligten der Anerkennungsbeziehung vollziehen eine Praxis und betrachten ihrer beider Äußerungen vor dem Hintergrund sozial etablierter normativer Verhaltensmuster. Sie betrachten die jeweils anderen Akteur:innen im Raum der Gründe nicht als bloße Gegenstände und ihre Äußerungen nicht als bloße physische Vorkommnisse. Wunschäußerungen von Akteur:innen sind nicht bloß Berichte über innere Zustände, und Absichtsbekundungen sind etwas anderes als Verhaltensvorhersagen in eigener Sache.[[11]](#footnote-11) Insbesondere von phänomenologischer Seite ist herausgestellt worden, dass die Wahrnehmung von Personen und des entsprechenden Ausdrucksverhalten direkt erfolgt, das heißt, es wird nicht erst auf der Basis von proto-szientifischen Verhaltensbeobachtungen in einem zweiten Schritt auf das Vorliegen von Personalität geschlossen. Nur in besonderen Zweifelsfällen – etwa bei dem Verdacht, es mit pathologischem Verhalten zu tun zu haben – können direkte Wahrnehmung und Reaktion suspendiert werden und in prüfender Absicht die Einstellung distanzierter Verhaltensbeobachtung eingenommen werden.[[12]](#footnote-12) Aber auch dann zielt dieses Vorgehen nicht darauf, die Äußerungen originärer Intelligenz auf anderem Wege zu erschließen, sondern umgekehrt sollen für diejenigen Verhaltensepisoden, die den Zweifel auslösen, alternative Erklärungen in Betracht gezogen werden. Jemandem hingegen, der systematisch Wünsche für Berichte über innere Zustände hielte und Absichtsbekundungen für Vorhersagen, müsste man eine eigentümliche Art von Blindheit für andere Personen attestieren,[[13]](#footnote-13) das heißt, er wäre unfähig, die Äußerungen von Personen als Ausdruck genuiner Intentionalität zu verstehen. Als aktiv Beteiligter einer Anerkennungsbeziehung würde er scheitern.

Wer solchermaßen agiert, würde das Verhalten anderer Akteur:innen in einer Weise deuten, die der szientifischen Erkenntnisweise nahekommt – mit dem gravierenden Unterschied freilich, dass die bzw. der Verhaltenswissenschaftler:in in ihrer bzw. seiner Rolle als Forschende:r nicht nur die eigene Einbindung in Anerkennungsbeziehungen zu den Forschungsobjekten methodisch eliminiert, sondern deren Verhalten auch durch die Angleichung an beliebige andere Vorkommnisse in der Welt seiner normativen Dimension entkleidet.

Coelho Mollos Forderung nach einer Beschränkung des Intelligenzbegriffs auf beobachtbares Verhalten erweitert zwar im Zusammenspiel mit seinen auf die anfänglich genannten Desiderate antwortenden Forderungen das Spektrum dessen, was als intelligent gelten kann. Doch die Erweiterung läuft, wie gesagt, Gefahr, durch eine Blickverengung erkauft zu werden, die in einer nicht phänomengerechten Betrachtung normgesteuerten Verhaltens und damit des Ausdrucks originärer Intelligenz besteht. Dabei besteht das Problem nicht darin, dass eine phänomengerechte Betrachtung die epistemische Bezugnahme auf eine obskure geisterhafte Größe statt auf manifestes Verhalten erforderlich mache. Auch im Rahmen von Anerkennungsbeziehungen kommt der Wahrnehmung von Verhalten – nämlich: Ausdrucksverhalten – die zentrale epistemische Rolle zu, doch von den Akteur:innen wird das gleiche Verhalten der anderen Akteur:innen im Raum der Gründe, das auch in szientifischer Einstellung beobachtet wird, auf eine andere Weise wahrgenommen. Die Akteur:innen beziehen sich nicht auf grundsätzlich andere Vorkommnisse, sondern sie beziehen sich anders auf die im Wesentlichen gleichen Vorkommnisse.

Stanley Cavell spricht hinsichtlich der Wahrnehmung von Ausdrucksverhalten in Anerkennungsbeziehungen von einem „seeing human beings as human beings“[[14]](#footnote-14). Auch wenn er damit einen wichtigen Punkt trifft, ist es womöglich in zweifacher Hinsicht unglücklich formuliert. Zum einen greift Cavell auf zwei unterschiedliche Bedeutungen des Ausdrucks „human being“ zurück (als Gattungswesen einerseits, als Akteur im Raum der Gründe andererseits), was ohne weitere Erläuterungen nicht über ein Wortspiel hinausführt. Und zum anderen, und das ist für den vorliegenden Kontext bedeutsamer, scheint Cavell in dem Zitat von vornherein davon auszugehen, dass nur Exemplare der Gattung Mensch als Akteur:innen im Raum der Gründe und somit als aktiv Beteiligte in Anerkennungsbeziehungen wahrgenommen werden können. Eine solche anthropozentrische Prämisse steht aber aus guten Gründen infrage.

Während wir bei anderen Menschen routinemäßig davon ausgehen, dass es sich um Personen handelt und nur in Grenzfällen – etwa bei dauerhaft Komatösen oder Föten bis zu einem gewissen Entwicklungsgrad – unsicher sind, und wir gleichzeitig bislang über keine Erfahrungen mit nicht-menschlichen Personen verfügen – selbst wenn einige bei den sog. Großen Menschenaffen in dieser Hinsicht Revisionsbedarf anmelden –, sollten wir grundsätzlich davon ausgehen, dass es andere Entitäten gibt oder zumindest geben kann, die Personen sind, die über *originäre Intelligenz* wie wir verfügen. Das bedeutet, um eine Formulierung von Wilfrid Sellars aufzugreifen, dass wir uns und diese Entitäten als zu einer Gemeinschaft gehörig ansehen müssten.[[15]](#footnote-15) Als entscheidend für die Zugehörigkeit zu dieser Gemeinschaft sieht Sellars die Fähigkeit an, „in meaningful discourse“ miteinander zu treten.[[16]](#footnote-16) Demnach ist die soziale Praxis der personalen Anerkennung in eine diskursive Praxis eingebettet und die Zuschreibung originärer Intelligenz erschöpft sich nicht darin, eine bestimmte Fähigkeit zu beschreiben, sondern miteinander in einen diskursiven Austausch einzutreten. Brandom hat einen Vorschlag dazu gemacht, wie die Struktur diskursiver Praxis im Detail zu verstehen ist. Ihm zufolge ist es eine sechsfache Struktur, die diskursive Bewertungspraktiken mit konzeptionellen Inhalten verbindet, die für diskursive Praxis notwendig und hinreichend ist.[[17]](#footnote-17)

**8. Fazit**

Den Ausgangspunkt unserer Überlegungen bildete die Frage, ob Maschinen *wirklich* intelligent sein können. Anknüpfend an die Definition von Coelho Mollo haben wir eine Differenzierung vorgeschlagen, die es erlaubt, diese Frage sowohl mit „Ja“ als auch mit „Nein“ zu beantworten. In einem abgeleiteten Sinne können KI-Systeme demnach zu Recht als intelligent bezeichnet werden, in einem originären Sinne aber nicht. Diese Lösung scheint uns durchaus attraktiv zu sein. Wir räumen aber ein, dass sie zu tiefen Problemen innerhalb der Philosophie des Geistes führt, für die wir keine überzeugende Lösung anbieten können. Eine Möglichkeit besteht darin, die Zuschreibung des Personenstatus als einen Akt der Anerkennung zu rekonstruieren, der in einer sozialen Praxis gründet. Man gibt dann aber den Vorteil preis, den eine rein behaviorale Beschreibungsebene bietet, die Coelho Mollo bei seinem Versuch einer Definition von Intelligenz explizit gefordert hatte.

Vielleicht ist es tatsächlich so, dass die Frage, ob Maschinen *wirklich* intelligent sind, den Bereich beobachtbaren Verhaltens überschreitet. Vielleicht verweist diese Frage auf einen tiefergehenden Vergleich von Menschen und Maschinen und führt uns vor Augen, dass wir Menschen nicht nur als Lebewesen erkennen und beschreiben können, sondern auch als Personen anerkennen müssen. Dies schließt im Übrigen nicht aus, dass wir auch andere Wesen oder Artefakte *als Personen anerkennen* können oder sogar müssen. Wir verstehen aber offenbar immer noch nicht gut genug, wie unsere personale Anerkennungspraxis als diskursive Praxis funktioniert. Vielleicht ist das eine der Fragen, auf die die intensive Beschäftigung mit KI letztlich führt und der wir uns intensiver widmen sollten.

**Literatur**

Brandom, Robert (1994): *Making It Explicit. Reasoning, Representing, and Discursive Commitment*. Harvard University Press, Cambridge/MA.

Brandom, Robert (2010): *Conceptual content and discursive practice*, in: Grazer Philosophische Studien 81/1, 13–35.

Cavell, Stanley (1982): *The Claim of Reason. Wittgenstein, Skepticism, Morality, and Tragedy*, Oxford University Press, Oxford/New York.

Coelho Mollo, Dimitri (2022): *Intelligent Behaviour*, in: Erkenntnis (2022). <https://doi.org/10.1007/s10670-022-00552-8>.

Cole, David (2020): *The Chinese Room Argument*, in: The Stanford Encyclopedia of Philosophy (Summer 2023 Edition) <https://plato.stanford.edu/archives/sum2023/entries/chinese-room/>.

Fichte, Johann Gottlieb (1991): *Grundlage des Naturrechts nach Prinzipien der Wissenschaftslehre*, Meiner, Hamburg.

Hartmann, Dirk/Mohseni, Amir/Reckwitz, Erhard/Rojek, Tim/Steckmann, Ulrich (Hg.) (2012): *Methoden der Geisteswissenschaften. Eine Selbstverständigung*, Velbrück, Weilerswist.

Haugeland, John (1990): *The Intentionality All-Stars*, in: Philosophical Perspectives 4, 383–427.

Hegel, Georg Wilhelm Friedrich (1987): *Phänomenologie des Geistes*, Meiner, Hamburg.

Honneth, Axel (2003): *Kampf um Anerkennung. Zur moralischen Grammatik sozialer Konflikte.* Erweiterte Ausgabe, Suhrkamp, Frankfurt a. M.

Ikäheimo, Heikki (2014): *Anerkennung*, de Gruyter, Berlin.

Lévinas, Emmanuel (1999): *Die Spur des Anderen. Untersuchungen zur Phänomenologie und Sozialphilosophie.* Übersetzt, herausgegeben und eingeleitet von Wolfgang Nikolaus Krewani, Karl Alber, Freiburg i. Br.

Sartre, Jean-Paul (92018): *Der Existentialismus ist ein Humanismus*, in: ders.: Der Existentialismus ist ein Humanismus und andere philosophische Essays 1943-1948, Rowohlt, Reinbek, 145–192.

Sartre, Jean-Paul (232022): *Das Sein und das Nichts: Versuch einer phänomenologischen Ontologie*, Rowohlt, Reinbek.

Searle, John (1980): *Minds, brains, and programs*, in: Behavioral and Brain Sciences 3/3, 417-424.

Sellars, Wilfrid (2007): *Philosophy and the Scientific Image of Man*, in: Scharp, Kevin/Brandom, Robert (Hg.): In the Space of Reasons. Selected Essays of Wilfrid Sellars, Harvard University Press, Cambridge/MA, 369–408.

Siep, Ludwig (2014): *Anerkennung als Prinzip der praktischen Philosophie. Untersuchungen zu Hegels Jenaer Philosophie des Geistes*, Meiner, Hamburg.

Strawson, Peter Frederick (1978): *Freiheit und Übelnehmen*, in: Pothast, Ulrich (Hg.): Seminar: Freies Handeln und Determinismus, Suhrkamp, Frankfurt a. M., 201–233.

Sturma, Dieter (1997): *Philosophie der Person. Die Selbstverhältnisse von Subjektivität und Moralität*, Schöningh, Paderborn.

Wang, Pei (2019): *On Defining Artificial Intelligence*, in: Journal of Artificial General Intelligence 10/2, 1–37.

Wittgenstein, Ludwig (1984): *Philosophische Untersuchungen*. Hg. von Anscombe, G. E. M./von Wright, G. H./Rhees, Rush, in: ders.: Werkausgabe Band 1, Suhrkamp, Frankfurt a. M.

**Kurzbiographie**

*Bert Heinrichs* ist Professor für Ethik und Angewandte Ethik am Institut für Wissenschaft und Ethik (IWE) der Universität Bonn und Leiter der Arbeitsgruppe „Neuroethik und Ethik der KI“ im Institut für Neurowissenschaften und Medizin: Gehirn und Verhalten (INM-7) am Forschungszentrum Jülich. *Ulrich Steckmann* ist Wissenschaftlicher Mitarbeiter der Arbeitsgruppe „Neuroethik und Ethik der KI“.

1. Seitenzahlen im Text beziehen sich auf die PDF-Version von Coelho Mollo (2022). [↑](#footnote-ref-1)
2. Wang (2019): 17. [↑](#footnote-ref-2)
3. Vgl. Searle (1980) und Haugeland (1990). [↑](#footnote-ref-3)
4. Zur mittlerweile weitverzweigten Diskussion um das Chinese-Room-Argument vgl. Cole (2020). [↑](#footnote-ref-4)
5. Vgl. Sartre (2018). [↑](#footnote-ref-5)
6. Vgl. einführend zum Begriff der Anerkennung Ikäheimo (2014). [↑](#footnote-ref-6)
7. Vgl. etwa Hartmann et al. (2012). [↑](#footnote-ref-7)
8. Vgl. Fichte (1991), Hegel (1987), Siep (2014), Honneth (2003). [↑](#footnote-ref-8)
9. Vgl. insbesondere Sartre (2022) und Lévinas (1999); siehe dazu Sturma (1997): 305-310. [↑](#footnote-ref-9)
10. Vgl. Brandom (1994). [↑](#footnote-ref-10)
11. Vgl. Cavell (1982): 379. [↑](#footnote-ref-11)
12. Vgl. dazu auch Strawson (1978), insbesondere 208 ff. [↑](#footnote-ref-12)
13. Stanley Cavell spricht diesbezüglich von „soul-blindness“ (Cavell 1982: 378 ff.), was allerdings leicht als ein Festhalten am ontologischen Dualismus missverstanden werden könnte. Cavell selbst bezieht sich auf Ludwig Wittgensteins Spätphilosophie, insbesondere auf dessen Konzepte des Aspektsehens und der Aspektblindheit; vgl. Wittgenstein (1984): 518–525 und 551–554. [↑](#footnote-ref-13)
14. Cavell (1982): 378. [↑](#footnote-ref-14)
15. Sellars (2007): 408. [↑](#footnote-ref-15)
16. Ebd.: 407. [↑](#footnote-ref-16)
17. Vgl. Brandom (2010). [↑](#footnote-ref-17)