VON SCHLEIMPILZEN, DIE DENKEN UND MASCHINEN, DIE LERNEN

Georg Trogemann
Vortrag beim Neokybernetik-Treffen
vom 26.-28. August
Gut Wendegräben, Brandenburg

KOMMENTAR

zu: The Logical Categories of Learning and Communication

in: STEPS TO AN ECOLOGY OF MIND
COLLECTED ESSAYS IN ANTHROPOLOGY, PSYCHIATRY, EVOLUTION, AND EPISTEMOLOGY
Gregory Bateson

dt.: ÖKOLOGIE DES GEISTES
ANTHROPOLOGISCHE, PSYCHOLOGISCHE, BIOLOGISCHE UND EPISTEMOLOGISCHE PERSPEKTIVEN

Maschinelles Lernen (ML) steht im Zentrum der Künstlichen Intelligenz (KI). Insbesondere *Deep Learning* hat die gesamte KI so stark verändert, dass KI und ML mittlerweile nicht selten synonym verwendet werden. Doch was verstehen wir unter *Lernen*? Was ist mit der Behauptung gemeint, ein Mensch, eine Maschine oder ein Organismus hat dieses oder jenes *gelernt*? Die Psychologie fasst unter „Lernen“ alle Prozesse zusammen, die sich als Verhaltensänderung durch Erfahrung beschreiben lassen.[1] Erfahrung meint hierbei, dass sich eine Sache in gewisser Hinsicht wiederholt und der Verlauf des Geschehens in irgendeiner Weise zur Verhaltensänderung genutzt wird, um damit beim erneuten Auftreten der gleichen Situation einen qualitativ anderen Verlauf zu erhalten. Die hier angestellten Überlegungen gehen von Gregory Batesons logischen Kategorien des Lernens aus, um einerseits maschinelles Lernen besser einordnen zu können und andererseits zu fragen, welchen Bedingungen ein allgemeinerer Begriff von Lernen genügen sollte, der beispielsweise auch das Lernen von Schleimpilzen miteinschließt.

[1] Beispielsweise Langfeldt, Hans-Peter (1996). Psychologie. Neuwied, Kriftel, Berlin: Luchterhand. [S. 102]. "Unter Lernen versteht man die hypothetischen Prozesse, die den Verhaltensänderungen durch Erfahrung entsprechen."

Vorüberlegungen

Wollen wir Begriffe wie *Denken*, *Intelligenz* oder *Lernen* nicht nur als menschliche Fähigkeiten begreifen, sondern auch Maschinen oder Organismen ohne zentrale Nervensysteme (z.B. Schleimpilzen) kognitive Fähigkeiten zusprechen, kann ihre Definition nicht auf rein materiellen Qualitäten beruhen. Materiell sind diese Entitäten so verschieden, dass ihre Gleichsetzung auf Basis physischer Vergleiche sicher nicht gelingen kann. Die Aufgabe besteht also darin, eine abstrakte Beschreibungsebene zu finden, auf der wir Fähigkeiten wie Lernen oder Intelligenz überzeugend darstellen können und für die wir zudem überzeugend darlegen können, dass die betrachteten Systeme diesen Beschreibungen genügen. Aufgrund der materiellen und organisatorischen Diversität und der Tatsache, dass Begriffe wie Lernen und Intelligenz ihrerseits schon Konstruktionen sind, geht es also von vorne herein nicht um Objektivität, sondern um Plausibilität und um die Art und Weise, wie wir auf eine Sache schauen und einen vertretbaren Standpunkt finden. Der Begriff Intelligenz ist vor allem eine sprachliche Vereinbarung, die ausgehandelt werden muss.

Auf den ersten Blick mag es einfacher erscheinen *Denken*, *Intelligenz* und *Lernen* nur auf Lebewesen (Menschen, Tiere, Schleimpilze etc.) anzuwenden. Tatsächlich kann aber gerade die Hereinnahme von Maschinen (Computern) helfen, methodische Schwierigkeiten deutlicher zu machen. Maschinen sind von Menschen geplante und hergestellte Artefakte, wir kennen also ihren Bauplan, sie sollten eigentlich keine Geheimnisse in sich bergen. Tatsächlich sind sie aber sehr komplex, das heißt, wir können das, was sie tun und wofür wir sie einsetzen nicht mehr zufriedenstellend aus einer einzigen Perspektive (mit einem einzigen Modell, Formalismus etc.) beschreiben.[2] Bei Computern haben wir es immer schon mit mehreren formalen Ebenen zu tun, die allesamt notwendig sind, um das gesamte Rechensystem auf allen unterschiedlichen Arbeitsebenen zu erfassen. Das Gleiche gilt aber beispielsweise für Organismen und Tiere in der Biologie. Moderne Naturwissenschaft liefert immer formale Beschreibungen dessen, was sie untersucht. Mehr als das bloße Phänomen gelten der Wissenschaft ihre Modelle und Theorien als eigentliche Realität, denn nur sie besitzen Erklärungskraft. Das wissenschaftliche Objekt *Physarum Polycephalum* wird erst durch die Experimente, technischen Messapparaturen und Messdaten sowie die formalen Modelle, Simulationen und Berechnungen, die wir anstellen überhaupt erst hergestellt. Insofern stellt sich auch für biologische Entitäten die Frage, wem sprechen wir eigentlich die Intelligenz zu, der bloßen materiellen Instanz oder vielmehr unseren formalen, abstrakten Modellen davon. Welche Rolle spielt also die Formalisierung bei der Klärung von Begriffen wie Intelligenz, Lernen, Denken? Die Hereinnahme des ingenieurwissenschaftlichen Artefakts Computer kann helfen, die Rolle formaler Modelle in den Feldern der Naturwissenschaft deutlicher hervortreten zu lassen. Die Beschreibung der Computerhardware ist beispielsweise nur auf der untersten Ebene der Siliziumkristalle materialbasiert. Auf allen höheren Ebenen, z.B. der elektronischen Bauteile (Widerstände und Transistoren), der Gatter (Und, Oder, Nicht) und der Funktionsgruppen (ALU, CPU, IO, Memory) haben wir es bereits mit funktionalen Abstraktionen zu tun. Die Abstraktionen setzen sich in den Software-Ebenen mit Maschinensprachen, höheren Programmiersprachen und schließlich APIs fort mit denen

[2] „Komplexität gibt an wie viele irreduzible Qualitäten bzw. Kontexturen im Spiel sind." Sie ist deshalb von Kompliziertheit zu unterscheiden. Siehe Eberhard von Goldammer, Rudolf Kaehr, Lernen in Maschinen- und lebenden Systemen: machine learning, neuronale Netze, http://www.vordenker.de/rk/rk_Zur-Dekonstruktion-der-Techno-Logik_1995.pdf

wir unsere Applikationen beschreiben. Am Ende werden natürlich alle Ebenen auf Signale in elektrischen Schaltungen abgebildet. Als Veränderungen messbar sind im Grunde nur diese Signale. Ohne Zugriff auf die Programmtexte ist es nahezu unmöglich, die höheren Organisationsformen je abzuleiten. Im Grunde steht der Biophysiker, der höhere Organisationsformen bei einem komplexen Lebewesen nachweisen will, vor der gleichen Aufgabe wie ein Außerirdischer, der zufällig einen lauffähigen Computer findet und herausfinden möchte, wie dieser funktioniert, ohne Zugriff auf den Quellcode oder sonstige Baupläne zu haben. Wie soll aus den messbaren Signalen auf die höheren Funktionen geschlossen werden, die nur im Quellcode sichtbar werden? Je nachdem, auf welche Beschreibungsebene unseres Computers wir also schauen, wird es unterschiedlich schwerfallen, der Maschine Intelligenz oder Lernverhalten zuzusprechen. Tatsächlich wurde bisher vor allem auf zwei Ebenen versucht, eine Analogie zwischen maschineller und menschlicher Intelligenz zu begründen. 1. Auf der materialnahen Ebene einfachster logischer Operationen und 2. auf (materialferner) Anwendungsebene, wo vor allem das Programmverhalten unter bestimmten Eingaben interessant ist.

Die materialnahe Argumentation umfasst im Wesentlichen drei Schritte, die zusammen eine Gleichsetzung auf funktionaler Ebene begründen.[3] George Boole liefert 1854 eine mathematische-formale Beschreibung der klassischen Logik. Die fundamentalen Gesetze des Verstandes lassen sich damit erstmals als symbolische Operationen beschreiben. 1943 veröffentlichen Warren McCulloch und Walter Pitts einen Artikel mit dem Titel „A logical Calculus of the Ideas immanent in Nervous Activity“. In ihrem Aufsatz weisen McCulloch und Pitts nach, dass die Nerventätigkeit des Gehirns den streng logischen Booleschen Gesetzen folgt. Den letzten Baustein lieferte (unter anderen) Claude E. Shannon. In seiner Master Thesis am Department of Electrical Engineering am Massachusetts Institute of Technology (MIT), Cambridge 1937 mit dem Titel *A Symbolic Analysis of Relay and Switching Circuits* beschreibt er den Zusammenhang zwischen logischen Operationen und elektrischen Schaltkreisen. Die Boolesche Algebra als Essenz der Logik erlaubte hier also als abstrakt-formale Vermittlungsebene die Gleichsetzung der Verstandesoperationen des Geistes mit dem, was elektrische Schaltkreise tun. Die unterschiedlichen materiellen Ausprägungen werden reduziert auf ihre funktionale Beschreibung. Da der Übergang von materiellen Vorgängen zur funktionalen Beschreibung auf Abstraktion beruht, kann methodisch bedingt nicht vollständig ausgeschlossen werden, dass Wesentliches bei diesem Schritt verloren geht. Anders gesagt: Materiell unterschiedliches auf eine gemeinsame funktionale Basis zu reduzieren ist immer auch eine abstrakte Konstruktion.

Die zweite (materialferne) Methode um Intelligenz, Denken oder Lernfähigkeit bei Maschinen, Menschen, Tieren oder gar einzelligen Organismen überzeugend zu begründen, beruht auf einen Vergleich des beobachtbaren Input-Output-Verhaltens, d. h. man konzentriert sich ausschließlich auf die Prozesse, die sich zwischen Organismus (Maschine, Mensch) und Umwelt abspielen. Das berühmte von Alan Turing eingeführte *Imitation Game*[4] ist nichts anderes, als der Versuch, die Frage „Can Machines Think?“ nicht durch Betrachtung der inneren Struktur der Maschine zu beantworten, sondern durch Beobachtung und Vergleich des Verhaltens von Mensch und Maschine unter

[3] Siehe z. B. Georg Trogemann, Jochen Viehoff, Code Art: Eine Elementare Einführung in die Programmierung als Künstlerische Praktik, Springer Verlag 2005, S. 522f.

[4] A. M. Turing, Computing Machinery and Intelligence, in: *Mind*, Volume LIX, Issue 236, October 1950, S. 433–460.

bestimmten Testbedingungen, also behavioristisch. Sowohl Introspektion als auch die Analyse der inneren Struktur eines Organismus oder einer Maschine schließt der Behaviorist aus; Organismus wie Maschine werden als *Black Box* betrachtet. Wenn Mensch und Maschine (Organismus etc.) unter bestimmten Testbedingungen vergleichbare Leistung bringen, gelten sie hinsichtlich dieser Leistungen als nicht unterscheidbar, wir sagen, sie besitzen beide diese Qualität. Das entspricht weitgehend einer umgangssprachlichen Definition von Intelligenz: Wenn Maschinen Leistungen bringen, für die wir beim Menschen Intelligenz voraussetzen, dann müssen wir auch der Maschine Intelligenz bescheinigen. Wenn wir beispielsweise Intelligenz definieren, als die Fähigkeit Probleme zu lösen, müssen wir uns nur noch auf den Problemkanon einigen. Können Maschinen dann – ganz behavioristisch – diese als relevant erachteten Probleme lösen, können wir sie als intelligent einstufen. Dabei kann auch die Geschwindigkeit, mit der eine geeignete Antwort (Reaktion) auf Probleme gefunden wird als relevant für den Grad der Intelligenz angesehen werden. Wollen wir Begriffe wie Denken, Intelligenz, Lernen stärker verallgemeinern und beispielsweise für einzellige Organismen wie Schleimpilze öffnen, wird dies kaum gelingen, wenn wir von menschlichen Problemstellungen und deren Lösung ausgehen. Intelligenz kann aus allgemeinerer Perspektive sinnvoller als „goal-oriented behaviour in a complex and unpredictable environment" definiert werden.[5]

Damit kennen wir bereits drei grundsätzlich verschiedene Standpunkte, um bestimmte Leistungen (Intelligenz, Lernfähigkeit etc.) diverser Entitäten zu bewerten und miteinander in Beziehung zu setzen. 1) Die Einführung einer gemeinsamen formalen Beschreibungsebene, auf die sich unterschiedliche materielle Prozesse reduzieren lassen. In der klassischen KI war das die Boolesche Algebra als Essenz menschlicher Logik. Die gemeinsame Ebene im Kognitivismus ist der Informationsbegriff. Kognitivismus geht von der Überzeugung aus, dass Denken im Wesentlichen Informationsverarbeitung ist und deshalb Computerprozesse und menschliches Denken äquivalent sind. Hier ist *Information* die gemeinsame abstrakte Form zur Beschreibung der inneren Organisation des Systems. 2) Die Betrachtung des Verhaltens. Die innere Struktur und Realisierung sind in diesem Fall uninteressant, was zählt, ist ausschließlich das beobachtbare Verhalten, also die äußerlich wahrnehmbaren und daher auch mit technischen Hilfsmitteln messbaren Veränderungen (Bewegungen, Stellungen, Äußerungen, Absonderungen, Aktivitäten, Outputs etc.). 3) Daneben bleibt natürlich immer die Möglichkeit, dass wir uns in den Anforderungen auf konkrete physische Merkmale stützen, etwa das Vorhandensein eines Zentralnervensystems wie es Wirbeltiere besitzen. Je nach dem, welchen Standpunkt wir wählen, werden wir der gleichen Sache Intelligenz oder Lernfähigkeit zu- bzw. absprechen.

Lernen systemtheoretisch betrachtet

Da wir sehr unterschiedliche Entitäten (lebende wie nicht lebende, soziale wie technische) unter dem Gesichtspunkt des Lernens betrachten wollen, ist es geboten, die Ausgangssituation möglichst allgemein zu beschreiben. Es bietet sich hierfür die systemtheoretische Sicht an. (Genau für solche heterogenen und disziplinübergreifenden Zusammenhänge wurde sie schließlich entwickelt.) Die Frage lautet dann: Was charakterisiert Systeme die lernen oder intelligent sind?

[5] Jonghyun Lee, Choice of a unicellular organism: Physarum polycephalum, Dissertation Bremen 2019, preprint, S. 136ff.

Systeme sind abstrakte Einheiten, die sich zuallererst über eine *Grenze* zwischen sich und ihrer *Umwelt* definieren. Ludwig von Bertalanffy beschrieb Systeme als sich selbst organisierende Funktionseinheiten, die sich von ihrer Umwelt abgrenzen und ihr Überleben (bzw. Weiterfunktionieren) durch ihre innere Organisation selbst sicherstellen.[6] Die Grenze kann räumlich oder organisatorisch bestimmt sein. Zum Beispiel sind für das Verständnis einer Softwareanwendung, die in der Cloud gerechnet wird, die räumlichen Verteilungen der beteiligten Prozesse ohne Bedeutung. Was zählt, sind ausschließlich ihre funktionalen Zusammenhänge. Der Begriff *Grenze* suggeriert, wir könnten leicht entscheiden, was zu jedem Zeitpunkt diesseits oder jenseits von ihr liegt, was sich also innen befindet und zum System gehört und was dem Außen zugeschlagen werden muss. In der Praxis kann sich diese Entscheidung als schwierig erweisen. Auch die Frage, ob ein System offen oder geschlossen ist, ist standpunktabhängig. Die *Umwelt* oder das Außen kann im Fall von Experimentalsettings sehr stark determiniert und eingeschränkt sein, es kann aber auch als das Offene und Unvorhersehbare weitgehend unbestimmt bleiben. Entscheidend für die systemische Betrachtung ist, welcher Austausch zwischen System und Umwelt stattfindet und wie dieser Austausch jeweils auf beide Seiten zurückwirkt. Materieller Austausch, Informationsaustausch und strukturelle Kopplung (operationale Geschlossenheit) sind drei unterschiedliche Typen von Systemen, bei denen sich der Austausch grundsätzlich verschieden darstellt.

Was heißt nun Lernen aus systemtheoretischer Sicht? Wir hatten Lernen eingangs als *Verhaltensänderung durch Erfahrung* beschrieben und Erfahrung als etwas, das sich in bestimmter Hinsicht wiederholt. In der Umwelt des Systems treten also Situationen auf, die in bestimmter Hinsicht gleichwertig zu früheren Situationen sind. Wir werden darauf zurückkommen müssen, was gleichwertig oder vergleichbar in diesem Zusammenhang heißen soll. Wenn diese Situationen (gewisse Zustände der Umwelt) nun im Austausch mit dem System zu einer Neustrukturierung oder Modifikation seiner inneren Organisation führen, so dass es sich zu einem späteren Zeitpunkt beim erneuten Auftreten dieser Situation anders verhält, sprechen wir dem System *Lernfähigkeit* zu. Das sich Einschreiben der Umwelt-System-Interaktion in die innere Struktur des Systems hängt eng zusammen mit dem, was wir als Gedächtnis bezeichnen. Lernen ist gewissermaßen die verhaltensseitige Beschreibung einer Systemveränderung und Gedächtnis die strukturelle. Wie können wir nun entscheiden, ob Lernen vorliegt? Festzuhalten ist, dass die Entscheidung, ob ein System lernt oder nicht, von einem externen Beobachter, also von außen getroffen werden muss. Dieser externe Beobachter steht vor verschiedenen Schwierigkeiten. Er muss entscheiden, wann gewisse Umweltbedingungen eine Wiederholung für das System darstellen. Kann man wirklich die exakt gleiche Ausgangssituation herstellen (zum Beispiel im Rahmen eines Experiments)? Wann können Situationen überhaupt als gleich gelten? Auch dafür ist Abstraktion notwendig. Der externe Beobachter muss ferner sagen können, ob Veränderungen im Verhalten auf genau diese Umweltsituation zurückzuführen sind und nicht etwa auf zufällige Schwankungen im Systemverhalten. Und zu guter Letzt möchte er natürlich wissen, welche Veränderungen im Inneren überhaupt die Verhaltensänderung bewirken. Nicht alle stattfindenden Strukturänderungen bedeuten eine Verhaltensänderung, zumal in Lebewesen permanent irgendwelche Veränderungen

[6] Ludwig von Bertalanffy: *An Outline of General Systems Theory*. In: *The British Journal for the Philosophy of Science*, 1/2, 1950, S. 134–165.

stattfinden. Dazu gehört auch die Frage, ob überhaupt die richtige Ebene bzw. Perspektive für die Beobachtung des Systems gewählt wurde? Die elektronische Schaltung eines Rechners, auf dem eine ML-Software läuft, verändert sich beispielsweise nicht. Trotzdem kann man aus der Starrheit der Schaltungsstruktur nicht den Schluss ziehen, dass ein nicht-lernendes System vorliegt. Das Lernen findet einfach mehrere Ebenen höher statt und kann nicht aus den Signalmustern des Rechners abgeleitet werden.

Mit Verweis auf Jonghyun Lee hatten wir *Intelligenz* oben als *zielgerichtetes Verhalten in einer komplexen und unvorhersehbaren Umgebung* definiert.[7] Wie beim Lernen muss auch hier die Entscheidung, ob von intelligentem Verhalten gesprochen werden kann von einem externen Beobachter getroffen werden. Auch hier gibt es eine Reihe von Schwierigkeiten. Was verstehen wir unter einem Ziel? Wer setzt dieses Ziel? Das Ziel, also jenes, was wir als wünschenswertes und richtiges Verhalten des Systems erachten, kann nur eine Setzung von außen sein. Da wir auch Maschinen und Organismen ohne Bewusstsein als lernend betrachten wollen, findet sich innerhalb des Systems keine Instanz, der wir diese Zielsetzung zuschreiben können. Möchten wir intelligentes Verhalten einer Entität überprüfen, so wird das Ziel also immer von außen gesetzt, beispielsweise im Rahmen des Experimentaufbaus. Nur im Hinblick auf dieses Ziel kann entschieden werden, ob eine Entscheidung falsch oder richtig war. Falsch oder richtig sind folglich ebenfalls Entscheidungen, die außerhalb des Systems getroffen werden. Doch die größere Schwierigkeit besteht darin, näher zu bestimmen, was eine unvorhersehbare Umgebung ist. Was heißt „vorhersehen“ in diesem Fall und für wen ist hier etwas „unvorhersehbar“? Für Jonghyun Lee ist „unpredictability“ eng verbunden mit dem, was wir weiter oben Erfahrung genannt haben. Im Zusammenhang mit dem Verhalten des einzelligen Schleimpilzes *Physarum Polycephalum* bedeutet *unpredictability* „it had not encountered this particular situation in the past“.[8] Damit ist wie beim Lernen auch die Entscheidung über die Intelligenz eines Systems teilweise zurückgeführt auf die Frage, wann wir von der Gleichheit einer Situation sprechen wollen und wann sie als „verschieden“ zu bewerten ist. Problematischer dürfte jedoch die Gleichsetzung von *Unvorhersehbarkeit* und der *Erstmaligkeit einer Situation sein*. Insbesondere soll der Begriff für Organismen verwendet werden, die kein *Bewusstsein* im herkömmlichen Sinn besitzen. Ein System, das erstmalig einer bestimmten Situation ausgesetzt ist, kann auf zwei unterschiedliche Weisen erfolgreich im Sinne der Zielsetzung sein. Einmal, indem es Mechanismen einsetzt, die explizit für den Umgang mit solchen neuen Situationen da sind. Auf technischer Ebene kann beispielsweise Induktion, Deduktion und Abduktion verwendet werden, um aus früheren Situationen Lösungsmöglichkeiten für die neue Situation abzuleiten. Zum anderen könnte das System aber auch zufällig durch seinen inneren Aufbau für diese neue Situation gerüstet sein und deshalb genauso gute Leistungen zeigen wie das System, das ein Problem als neu erkennt und aktiv versucht, eine Lösung zu finden. Behavioristisch wäre hier keine Unterscheidung festzustellen, nur durch Inspektion der inneren Struktur könnte gegebenenfalls ein Problemlösungsmechanismus nachgewiesen werden. Kann ein solcher Mechanismus tatsächlich ermittelt werden, wären wir vermutlich geneigt dem System Intelligenz zuzusprechen, obwohl es weder lernfähig ist, noch unvorhersehbare Situationen vorgelegen haben. Eine Problemlösungsstrategie, die auf Induktion und

[7] Jonghyun Lee, ebda. S. 136ff, “Therefore, we define intelligence as a 'goal-oriented behaviour in a complex and unpredictable environment’“.

[8] Private Kommunikation.

Deduktion beruht, erzeugt ja gerade einen Lösungsraum, der vorhersehbar ist, nämlich durch Anwendung dieser beiden logischen Prinzipien und ohne dass das System vorher schon einmal in dieser Situation gewesen sein muss oder das Problem je explizit vorhergesehen wurde. Wie wir weiter unten sehen werden, muss das System dafür nicht einmal lernfähig sein, es müssen also keinerlei Mechanismen existieren, die auf *Trial-and-Error* beruhen.[9] Was meinen wir also damit, wenn wir sagen, das selbstfahrende Auto ist in eine unvorhersehbare Situation geraten und hat deshalb einen Umfall verursacht? Unvorhersehbar sind in strengen Sinne nur Situationen, an denen das System auch wirklich scheitern kann. Soll Unvorhersagbarkeit wirklich eine Rolle für die Definition von Intelligenz spielen, müsste Lernfähigkeit eine zwingende Voraussetzung sein. Nur lernfähige Systeme arbeiten nach dem *Trial-and-Error-Prinzip* und können eben auch irren.[10] Wenn der Begriff „unvorhersehbar" sich als zu problematisch erweist, wie könnte dann eine systemtheoretische Charakterisierung eines intelligenten Systems aussehen?

Die logischen Kategorien von Lernen und Kommunikation bei Bateson

Auch wenn wir in den Vorüberlegungen Intelligenz und Lernen wie zwei eng verwandte Fähigkeiten eingeführt hatten, wollen wir hier nicht untersuchen, wie Intelligenz und Lernen zusammen hängen. Also keine Fragen der Art beantworten: Ist Intelligenz vererbt oder das Ergebnis eines umweltbedingten Lernprozesses? Zu wie viel Prozent ist der IQ eines Menschen Veranlagung und zu wie viel Prozent durch die Umwelt bestimmt? Setzt Intelligenz überhaupt Lernfähigkeit voraus? Es bleibt also unklar, wie Lernen und Intelligenz verbunden sind. Viele Definitionen von Intelligenz fordern nicht explizit Lernfähigkeit. Persönlich bin ich der Meinung, dass Lernfähigkeit zu den zentralen Eigenschaften einer intelligenten Entität gehört. Im Zentrum der folgenden Betrachtungen steht allerdings die Erkenntnis Batesons, dass wir beim Lernen logische Typen unterscheiden müssen. In Batesons Worten: „Die Frage lautet nicht: »können Maschinen lernen?« sondern: »Welche Ebene oder Ordnung des Lernens erreicht eine gegebene Maschine wirklich?«"[11] Mit Hilfe der logischen Kategorien des Lernens kann man besser verstehen, welche Lernebenen und damit verbundene Leistungen wir in der KI bisher erreicht haben und welche nicht. Batesons Ansatz zur Beschreibung des „Lernens" ist auch nicht klassisch behavioristisch, sondern vielmehr disziplinübergreifend. Die Umgebung und das Verhalten sind genauso wichtig, wie die innere Form der Organisation des lernenden Systems, die er in hierarchische Lerntypen unterteilt. Der Vorteil seines Ansatzes ist, dass Lernen sehr allgemein als Kommunikationsphänomen charakterisiert ist und damit Lernen bei Menschen, Tieren, Organismen und Computern einer einheitlichen Beschreibung zugänglich wird.

Als *Lernen* bezeichnet Bateson eine *Veränderung* irgendeiner Art. Das deckt sich gut mit unserer systemtheoretischen Beschreibung von Lernsituationen. Die Schwierigkeit besteht für Bateson darin zu sagen, um *was für eine Art* der Veränderung es sich

[9] Was weiter unten als *Lernen null* bezeichnet wird, ist eine nicht zu unterschätzende Denkfigur. Bateson verwendet in seinem Beispiel von Neumanns fiktiven Spieler, um zu zeigen, wie mächtig Systeme im Hinblick auf ihre Problemlösungskompetenz sein können, auch wenn sie nicht lernen.

[10] Interessant ist in diesem Zusammenhang, dass der logische Operator Abduktion im Gegensatz zu Induktion und Deduktion nur mögliche Ursachen liefert, also auch irren kann.

[11] In: Gregory Bateson, Ökologie des Geistes, Suhrkamp Verlag, Frankfurt am Main 1981, 11. Auflage 2014, S. 368.

handelt.[12] Er beginnt seine Betrachtungen mit dem Sonderfall der festverdrahteten Reaktion, wo also keine Änderung und damit kein Lernen stattfinden. Diesen Fall bezeichnet er als *Lernen null*. Zentral für die Charakterisierung der Batesonschen Lernebenen ist der Begriff des *Irrtums*. Die hierarchische Ordnung der Lernprozesse basiert auf einer Klassifizierung von Irrtumstypen. Haben wir die unterschiedlichen Irrtumstypen verstanden, können wir auch die Lernebenen unterscheiden. Wir gehen also davon aus, dass Menschen, Tiere, Organismen und Algorithmen auch Dinge falsch machen – sprich sich irren können. Ob etwas falsch gemacht wird, kann natürlich nur im Hinblick auf ein Ziel, d. h. als Differenz zum *Richtigen* entschieden werden. Wir hatten bereits darauf hingewiesen, dass das Ziel immer von außen festgelegt wird, etwa durch den Aufbau und die Beobachtungsgrößen eines Experiments. Wir wollen bei den folgenden Betrachtungen der Lernebenen möglichst auf die bei Bateson angeführten Beispiele für menschliches Lernen verzichten und dagegen die abstrakte systemtheoretische und maschinelle Sicht betonen.

Lernen null
(also Nicht-lernen) bezeichnet alle Systeme, bei denen der Irrtum nicht genutzt werden kann, um ihr zukünftiges Verhalten im Hinblick auf die Erreichung des Ziels zu verbessern. Selbst wenn beispielsweise ein Algorithmus herausfindet, dass eine Entscheidung falsch war, darf diese Entdeckung nicht dazu führen, dass sich sein Verhalten ändert, d. h. in einer vergleichbaren Situation in der Zukunft anders entschieden wird. Systeme, die nicht lernen werden in der gleichen Situation auch die gleichen Entscheidungen treffen und damit auch die gleichen Fehler immer wiederholen. Der Reiz und die darauf folgende Reaktion des Systems gehorchen einer starren Abbildung. Diese Abbildung kann beliebig kompliziert und aufwendig sein und sogar Entscheidungen unterschiedlichen Typs berücksichtigen, zum Beispiel sowohl strategische Entscheidungen (welches Ziel soll langfristig erreicht werden) als auch taktische Entscheidungen (was kann ich jetzt für die Erreichung meines momentanen Ziels tun) treffen. Das System kann sogar aktiv versuchen, Informationen aus dem Umweltfeedback zu berechnen und diese mit in die Entscheidungen einbeziehen. Bei Systemen, die einen großen Zustandsraum besitzen, die wir aber nicht selbst entworfen haben oder deren Bauplan wir nicht kennen, wird es schwierig auf Basis von Beobachtungen (Messungen) zu sagen, ob Lernen vorliegt oder nicht. Wir werden weiter unten auf dieses Problem zurückkommen.

Lernen I
Alle weiteren Lernebenen beruhen auf dem Prinzip von Versuch und Irrtum. Das heißt, die Rückmeldung aus der Umwelt des Systems wird benutzt, um das Verhalten des Systems anzupassen. Das bekannteste Beispiel für *Lernen I* ist die klassische Pawlowsche Konditionierung. Zu einem bestimmten Zeitpunkt 2 sondert der Hund als Reaktion auf das Bimmeln einer Glocke Speichel ab. Zu einem Zeitpunkt 1, bevor er gelernt hat die Glocke und die Gabe von Futter miteinander zu verbinden, hat er das nicht getan.[13] *Lernen I* bedeutet demnach Festlegung der Wahl innerhalb einer fixen Menge von Alternativen. Auf dieser Ebene kann die einfachste Form des Irrtums

[12] Ebda. S. 366.

[13] Empfehlenswert ist hierzu die von Heinz von Förster erzählte Geschichte. Es kommt auf die Glocke überhaupt nicht an. Der Hund versteht die Situation aus dem Kontext, die Glocke ist nur für Pawlow wichtig, um seine Theorie zu untermauern. Höre dazu: *Das hermeneutische Prinzip*, auf der CD: *2x2 = grün*, Audio-CD, supposé Verlag 1999. In diesem Beispiel wird deutlich, dass auch Tiere Kontexte verstehen können.

korrigiert werden. War das Ereignis vorher mit keiner oder mit einer falschen Reaktion verbunden, so erfolgt nach dem Lernen die richtige Reaktion auf den Umweltreiz. Alles Weitere bleibt unverändert, nur die Zuordnung zwischen Ereignis und Reaktion wird gelernt. Die Entscheidungsalternativen selbst unterliegen keiner Korrektur. Hier kommt erstmals der Begriff des *Kontexts* ins Spiel. Systemtheoretisch betrachtet differenziert der Begriff des Kontexts das Geschehen in der Umwelt eines Systems. Das Umweltgeschehen wird quasi markiert und erhält damit eine andere Bedeutung. Eine Situation, die sich identisch wiederholt, aber in einem anderen Kontext stattfindet, hat eine vollkommen andere Wirkung auf den Beobachter. Eine Beschimpfung, die innerhalb eines Theaterstücks stattfindet, wird anders gewertet als die gleiche Beschimpfung auf einer Geburtstagsfeier. Kontext heißt also, dass etwas grundlegend verschieden ist, während gleichzeitig das im Mittelpunkt stehende Geschehen (das den Reiz auslöst) unverändert ablaufen kann. Da der Kontext durch das System erst ermittelt und sogar geformt wird, ist der Begriff nicht nur im Außen festgemacht. Er zielt sowohl auf äußere Situationen und Ereignisse ab, schließt aber auch die innere Struktur ein, die den Kontext markiert. Die Idee eines wiederholbaren Kontextes ist zentral für die gesamte Lerntheorie von Bateson. Wären Kontexte nicht wiederholbar, wäre jede Situation vollkommen neu und Lernen durch Erfahrung damit nicht möglich. Wir hatten weiter oben bereits auf die Schwierigkeit hingewiesen, wie zu entscheiden ist, ob sich etwas wiederholt oder neu ist. Der Begriff des Kontexts führt eine Differenzierung für die Umgebung von Systemen ein. Es reicht nicht zu erkennen, dass eine bestimmte Situation sich wiederholt, auch der Rest muss einer Bewertung unterzogen werden. Ein System reagiert auf denselben „Reiz“ in verschiedenen Kontexten verschieden, daher stellt sich die Frage, wie schafft es ein System, Kontext A von Kontext B zu unterscheiden. Bateson nennt diese Unterscheidung Kontextmarkierung.

Alle bisher in der KI beschriebenen Methoden für maschinelles Lernen sind entweder vom *Typ null* oder vom *Typ I*. Technisch gesprochen werden lediglich Parameteranpassungen vorgenommen, mit deren Hilfe die Auswahl der Alternativen gesteuert wird. Bei Neuronalen Netzen sind diese Parameter die Gewichte auf den Verbindungen zwischen den Neuronen.

Lernen II
Auf Ebene II werden nicht Reiz-Reaktions-Schemata gefestigt, sondern der Prozess des Lernens selbst wird zum Gegenstand. Wir hatten eingangs Lernen als *Veränderung* irgendeiner Art definiert. *Lernen I* ist die *Veränderung der Verbindungen von Reiz und Reaktion*. Auf Ebene II geht es nun um die *Veränderung im Prozess des Lernens I*. Es werden auf dieser Ebene nicht nur Gewohnheiten eintrainiert (in dieser Situation mache ich das und das), sondern Gewohnheiten werden selbst zum Gegenstand der Betrachtung (Was mache ich da eigentlich und warum?). Man spricht in diesem Zusammenhang auch von *Lernen lernen*.

Was bedeutet diese Lernebene im Zusammenhang mit AI? *Explainable Artificial Intelligence* (XAI; deutsch: *erklärbare künstliche Intelligenz* oder *erklärbares Maschinenlernen*) soll nachvollziehbar machen, auf welche Weise AI-Systeme wie Neuronale Netze zu ihren Entscheidungen kommen. Hierfür ist Introspektion nötig, d.h. Algorithmen müssen ihr eigenes Tun beobachten und Auskunft darüber geben. Obwohl nicht alle Probleme der XAI zwingend Lernebene II voraussetzen, sind die hier behandelten Probleme von einem höheren Typ als das einfache Lernen Neuronaler Netze, da der Kontext, in dem ein Algorithmus arbeitet, Gegenstand der Betrachtung

wird. Experimentell lässt sich *Lernen II* durch das so genannte Umkehrungs-Lernen feststellen.[14] Das System lernt zunächst eine binäre Entscheidung (auf Reiz 1 folgt Reaktion 1, auf Reiz 2 folgt Reaktion 2). Danach wird die Bedeutung der Reize umgekehrt (auf Reiz 1 folgt Reaktion 2, auf Reiz 2 folgt Reaktion 1). Ist dies gelernt, werden die Reize erneut umgekehrt, usw. Die Frage lautet, ist das System in der Lage etwas über die Umkehrung zu lernen. Es würde also aufhören zu lernen, sobald der Mechanismus durchschaut ist. Ein Neuronales Netz, das mit bestimmten Bilddatenbanken trainiert wird, würde, sofern es zu Lernen II fähig ist, vielleicht irgendwann zum Schluss kommen, dass es vollkommen sinnlos ist, diese Bilder zu lernen, da das Lernverfahren nicht geeignet ist oder das Datenmaterial nicht repräsentativ.

Lernen III
Auch wir haben das Schema inzwischen gelernt. *Lernen II ist die Veränderung des Prozesses von Lernen I*. Folglich ist *Lernen III die Veränderung im Prozess des Lernens II*. Bateson schreibt hierzu: „Es steht zu erwarten, dass es auch für Wissenschaftler, die auch nur Menschen sind, schwierig sein wird, diesen Prozess vorzustellen oder zu beschreiben."[15] Seiner Ansicht nach kommt *Lernen III* nur gelegentlich vor, etwa in der Psychotherapie, der religiösen Bekehrung und anderen Prozessen, bei denen eine *tiefgreifende Umstrukturierung des Charakters* stattfindet. Der Lernende auf Ebene III ist in der Lage sein, die Kontexte von Kontexten wahrzunehmen. Was könnte *Lernen III* auf technischer Seite bedeuten? Auf Ebene II lernt das Programm etwas über den Kontext, in dem es eingesetzt wird und über die Methoden, die für einen Kontext geeignet sind. Ebene III lernt also etwas über die Struktur von Kontexten und erfindet Methoden um in diesen Kontexten zielgerichtet zu agieren. Als Äquivalent zur tiefgreifenden Umstrukturierung des menschlichen Charakters müssten Programme also in der Lage sein, sich selbst umzuschreiben. In Laufe des Lernprozesses müsste das Programm sich seine eigene Basis entziehen und gleichzeitig eine neue schaffen. Der Quellcode des Programms schreibt sich selbst um und erzeugt damit aus sich heraus neue Kontexte und Methoden für diese Kontexte. Wie kann das gehen?

Lernen IV
ist die *Veränderung im Lernen III*. Nach Bateson wird diese Ebene von keinem ausgewachsenen lebenden Organismus auf der Erde erreicht. Nur die Verbindung von Ontogenese und Phylogenese kann aus seiner Sicht Ebene IV erreichen. Um auf Ebene IV zu gelangen, betrachten wir also den Prozess, der Organismen hervorgebracht hat, die zu *Lernen III* fähig sind. Das ist der Evolutionsprozess zusammen mit der Ontogenese der Individuen. Lernen IV bezieht sich also nicht mehr auf Individuen, die lernen, sondern auf den Prozess, der diese Individuen hervorgebracht hat. Aus technischer Sicht würde es entsprechend nicht mehr um Programme gehen, sondern um die Prozesse, die Programmiersprachen und programmierbare Maschinen hervorgebracht haben. Entsprechend der Typentheorie Russels, auf die Bateson sich bezieht, sind die Lernebenen seiner Lerntheorie nach oben nicht abgeschlossen, wir können nach dem bekannten Schema immer neue konstruieren. *Lernen X+1 ist Veränderung im Prozess des Lernens X*. Für Bateson ist bei Ebene IV allerdings Schluss. Angesichts synthetischer Biologie und Gentechnologie können wir allerdings feststellen, dass wir mittlerweile auf

[14] Siehe Eberhard von Goldammer, Rudolf Kaehr, Lernen in Maschinen- und lebenden Systemen: machine learning, neuronale Netze, a.a.O., S. 5.
[15] A.a.O., S. 390.

Lernebene V experimentieren. Wir thematisieren und manipulieren die Prozesse der Ebene IV.

Schluss

Die hier angestellten Überlegungen sollten noch einmal deutlich machen, dass die Frage, ob einem System Intelligenz zugesprochen werden kann, davon abhängt, wie Intelligenz definiert wird, welche Forderungen wir also an *Intelligenz* stellen und mit welchen Methoden wir die Kriterien prüfen. So zeigen Fische in Experimenten erstaunliche Gedächtnisleistungen, hohe Lernfähigkeit und soziale Intelligenz. Andere problemlösungsorientierte Fähigkeiten wie sie beispielsweise Primaten zeigen fehlen dagegen (z. B. Sprache, Werkzeuggebrauch). Sind Fische nun intelligent oder nicht? Die Frage lautet deshalb, worauf legen wir bei der konkreten Untersuchung einer Spezies oder eines Systems den Fokus oder schlicht: Womit geben wir uns zufrieden, um Intelligenz zu postulieren? Interessanter als die binäre Unterscheidung *intelligent – nicht intelligent* sind deshalb differenziertere Betrachtungen und Abstufungen von Intelligenz. Bateson liefert für den Begriff des Lernens ein hierarchisches Modell, das nicht nur danach fragt, ob Lernen vorliegt, sondern welche Stufe des Lernens erreicht wird. Die Frage, wie sich menschliche und maschinelle Intelligenz unterscheiden bzw. inwieweit sie sich ähnlich sind ist demnach nicht besonders interessant. Es geht vielmehr darum, einen Raum zu schaffen, in dem maschinelle, menschliche und sonstige organische oder nicht-organische Intelligenz vorurteilsfrei betrachtet und beurteilt werden können.

Der Begriff der *Maschinellen* oder *Künstlichen Intelligenz* dürfte sich aufgrund der gegenwärtigen Welle von neuen Anwendungen endgültig auch in der breiten Öffentlichkeit durchgesetzt haben. Die Gesellschaft spricht Softwarelösungen nun also Intelligenz zu, auch wenn aus Sicht von Lernkategorien nur die untersten Stufen erreicht werden. *AI* und *ML* Algorithmen lernen bisher lediglich auf den Ebenen 0 und I der Batesonschen Lernkategorien. Auf dieser Ebene lässt sich realisieren, was im Diskurs der Künstlichen Intelligenz als *schwache KI* bezeichnet wird. Für eng abgesteckte Aufgaben mit klaren Entscheidungsalternativen und ohne explizite Thematisierung des Kontextes lassen sich die algorithmischen Möglichkeitsräume zwischen Eingabe und Entscheidung sehr leicht systematisch erzeugen. Die Lernalgorithmen der KI sind damit nicht viel mehr, als effiziente Methoden, um aus den riesigen Möglichkeitsräumen die brauchbaren Verbindungen zwischen den Eingängen und den Ausgängen (dem gezeigten Bild und der Antwort Hund) herauszufiltern. Sicher ist, dass Lernen I nicht ausreicht, um *starke KI* zu realisieren. Unabhängig davon, ob das überhaupt ein sinnvolles Ziel ist, lautet aus methodischer Sicht die Frage, wie sehen diese Algorithmen für die höheren Lernebenen aus? Interessant ist in diesem Zusammenhang die Feststellung, dass bei gegenwärtigen Entwicklungsumgebungen der Programmierer eines KI-Systems immer mindestens eine Reflexionsebene (Lernebene) höher arbeitet als sein Programm. Wie müssen Entwicklungsumgebungen aufgebaut sein, bei denen die Anwendung eine höhere Lernebene erreicht als ihr Programmierer?