Studiengang Technik & Philosophie
Institut für Philosophie der TU Darmstadt

Ruth Karl

Surprised Machines?

Über die Fähigkeit, sich überraschen zu lassen

Masterarbeit im Studiengang Technik und Philosophie

Erstgutachterin: Prof. Dr. phil. Petra Gehring

Zweitgutachter: Prof. Dr. Andreas Kaminski

Darmstadt, Wintersemester 2023/2024

Inhalt

Surprised Machines?
Über die Fähigkeit, sich überraschen zu lassen

Warum sich mit der Fähigkeit, sich überraschen zu lassen, auseinandersetzen?

Drei kurze Antworten

Weil wir uns gerne überraschen lassen.

Weil sie, sagt Vinciane Despret, Verhaltensforschung – den Versuch, andere zu verstehen – besser macht.

Weil das etwas ist, was Menschen bislang besser können als Maschinen. Es könnte also schon von daher interessant sein, einmal zu untersuchen, welche Aspekte dieser Fähigkeit von aktuellen technikorientierten Ansätzen bislang ignoriert oder übersehen werden.

Und eine etwas längere

Unter dem Titel „F wie Forschung“ begegnete mir in Vinciane Desprets Buch *Was würden Tiere sagen, würden wir die richtigen Fragen stellen?* die Überraschung zum ersten Mal als etwas, über das nachzudenken sich lohnen könnte. Über die Geschichte der Verwissenschaftlichung der Verhaltensforschung schreibt sie dort:

> „Beginnend mit Lorenz’ theoretischen Überlegungen wird die Ethologie einen strikt wissenschaftlichen Weg einschlagen. Die Ethologen, die ihm dann folgen, haben gelernt, Tiere als nur *reagierende* statt als *fühlende und denkende* Lebewesen anzusehen und jegliche mögliche Einbeziehung individueller und subjektiver Erfahrungen auszuschließen. Die Tiere werden dadurch etwas verlieren, das eine entscheidende Möglichkeit für die Beziehung gewesen war, nämlich die Möglichkeit, den, der sie analysiert, zu *überraschen*. Alles wird vorhersehbar. Handlungsursachen

> ersetzen Beweggründe, seien sie nun vernünftig oder abwegig, und das Wort *Initiative* weicht dem Begriff *Reaktion* (…).“[1]

Überraschung erscheint hier zugleich als nicht genutzte Chance der wissenschaftlichen Forschung und als „eine entscheidende Möglichkeit“ für Beziehungen.

Die Bereitschaft, sich überraschen zu lassen, betrifft in Desprets Denken unmittelbar die Art, in der wir uns mit dem, worüber wir nachdenken, ins Verhältnis setzen. Überrascht zu werden ist also von Anfang an, anders als es die grammatikalische Form suggerieren mag, mehr als ein passives Geschehen, sondern hat Voraussetzungen, von denen einige aktiv gestaltbar zu sein scheinen. Diese näher zu untersuchen ist eines der Anliegen dieser Arbeit.

Es gehört zum Alltagsverständnis des Überraschungsgeschehens, dass darin Erwartungen eine Rolle spielen, und zwar in der Form, dass sich plötzlich etwas als anders erweist, als wir es erwartet hatten. Überraschungen erfordern – davon werde ich in dieser Arbeit ausgehen – einen Umgang mit enttäuschten oder sich als Irrtum herausstellenden Erwartungen. Durch Überraschungen lernen wir dazu.

Genau deshalb interessieren sich auch die eng mit der Entwicklung künstlicher Intelligenz[2] verbundenen Kognitionswissenschaften in den letzten Jahren zunehmend für dieses Phänomen. Denn die Überraschung verweist auf eine grundlegende Charakteristik menschlichen Denkens: Es ist einerseits veränderlich und reagiert beständig auf neue Umstände und Erfahrungen, und es ist gleichzeitig stabil, so dass wir aus einmal gemachten Erfahrungen einen längerfristig verfügbaren Orientierungsrahmen für künftiges Handeln gewinnen, dem wir so lange vertrauen, bis uns eine neue Erfahrung eines Besseren belehrt. Das Verhältnis zwischen Plastizität und Stabilität wird in den Kognitionswissenschaften oft als „Dilemma“ bezeichnet, da es bis heute nicht gelungen ist, es für das maschinelle Lernen in eine geeignete Formel zu bringen. Hier stoßen wir vielmehr da, wo Lernen außerhalb definierter „Trainingsphasen“ stattfinden soll, auf zahlreiche Probleme durch unerwünschte Interferenzen. Ein Ausgangspunkt meiner Arbeit war daher ursprünglich die Frage, ob eine phänomenologische Betrachtung des Phänomens der Überraschung hier einen Beitrag leisten kann. Soviel sei bereits an dieser Stelle verraten: Ich denke, sie kann es, aber nicht in einem unmittelbar produktiv umsetzbaren Sinn, sondern eher auf der Ebene einer kritischen Reflexion technologisch ausgerichteter Ansätze.

---

[1] Despret, Vinciane (2019): Was würden Tiere sagen, würden wir die richtigen Fragen stellen? Münster: Unrast, S. 52f. Kursiv im Original. In einer Fußnote dazu benennt Despret ihrerseits zwei Autorinnen, denen sie die Inspiration zu dieser Analyse verdankt, nämlich Dominique Lestel mit ihrem Hinweis auf „einen Zusammenhang zwischen der Tatsache, dass das Tier die Initiative verliert, und den Mitteln zur Unterwerfung, die dazu dienen, jegliche Möglichkeit für Überraschungen zu verhindern“ und Émilie Gomarts Analyse der „Überraschung in den Beziehungen zwischen Drogenkonsumenten, Helfern und politischen Mächten“. Gemeint sind a) Lestel, Dominique: Les Amis de mes amis. Paris, 2007. und b) Gomart, Émilie: „Surprised by Methadone“, in Body and Society 2–3, 10, Juni 2004, S. 85–110.

[2] Ich gehe davon aus, dass dieser Begriff inzwischen alltagssprachlich verwendet werden kann und verweise ansonsten auf das angehängte Glossar technischer Begriffe.

Während ich mit dieser Arbeit also einen deskriptiven Ansatz verfolge, halte ich die Beschäftigung mit der Fähigkeit, sich überraschen zu lassen, auch aus normativen Gründen für wertvoll. Dies nicht nur, weil ich mit Vinciane Despret meine, dass sie „eine entscheidende Möglichkeit“ für Beziehungen ist. Genauso entscheidend scheint sie mir im Hinblick auf unser gesellschaftliches Handeln insgesamt zu sein, denn die Offenheit für andere Perspektiven beinhaltet die Fähigkeit, sich von anderen überraschen zu lassen und eigene Voreinstellungen in der Begegnung mit ihnen kritisch zu hinterfragen.

So verstehe ich auch Tobias Matzner, der in seiner Dissertation *Vita Variabilis* danach fragt, wie sich *Weltbilder politisieren* lassen. Was diese Termini beschreiben, verdeutlicht das nachfolgende Zitat:

> „Wenn etwas hinterfragt werden soll, worauf sich eine Gruppe von Menschen in ihrem Tun verlässt, so kann man sich dabei nicht selbst darauf verlassen. Man braucht also einen anderen ‚Grund, auf dem man stehen kann‘, Überzeugungen, auf die man sich selbst verlässt und mit der die Grundlagen der Anderen als beweglich, zur Disposition stehend, als bedingt, abgeleitet oder relativ verstanden werden können.
>
> Darüber hinaus müssen Sprachspiele und Handlungsweisen etabliert werden, deren Grundlagen es *den Anderen* ermöglichen, ihre frühere feste Überzeugung zu ‚verflüssigen‘. Der „Grund“, von dem aus einem oder einigen das Weltbild fraglich wird und derjenige, auf den sich eventuell durch dieses Fraglichwerden angestoßene neue Praktiken und Sprachspiele verlassen, ist dabei nicht derselbe.
>
> Diesen zweiten Vorgang nenne ich „Politisierung“ – also Handlungen, die dazu führen, dass grundlegende Überzeugungen im Weltbild einer Gruppe von Menschen ‚verflüssigt‘ werden.“[3]

Matzner versteht *Weltbild* mit Wittgenstein in der Metapher eines Flussbetts, dessen Ränder für den Fluss zu einem gegebenen Zeitpunkt eine feste und den Lauf bestimmende Form haben, die sich aber permanent verändern und „verflüssigen“ können, ein Bild, das mir sehr geeignet scheint, auch das noch zu beschreibende *Plastizitäts-Stabilitäts-Dilemma* der Kognitionswissenschaften zu veranschaulichen. In dieser Metapher ist das Weltbild der stabile Part, die Möglichkeit, eine Verflüssigung anzuregen, entspräche der Politisierung, und der Fluss die veränderliche Welt des Sprechens und Handelns. Ich denke, dass eine Kultivierung der Fähigkeit, sich überraschen zu lassen, geeignet wäre, genau solche Sprachspiele und Handlungsweisen zu etablieren, die es anderen ermöglichen, ihre bestehenden Überzeugungen zu verflüssigen. Für mich ist sie damit eine Variante der Politisierung im Matznerschen Sinn. Im Bild von Fluss und Flussbett beschriebe die Fähigkeit, sich überraschen zu lassen, die Eigenschaft eines Flusses, der gemeinsam mit seiner veränderlichen Umwelt immer neue Ausläufer, Umwege und Biotope bildet. Das

3 Matzner, Tobias (2013): Vita variabilis. Handelnde und ihre Welt nach Hannah Arendt und Ludwig Wittgenstein. Würzburg: Königshausen & Neumann, S. 180f.

Gegenteil wäre ein in Kanälen oder engen Felswänden gefangener Strom, dem nichts bleibt, als Hindernisse mit sich zu reißen oder zu zermalmen.

Eine Politisierung von Weltbildern, insbesondere solchen, die Kriege, Gewalt und Umweltzerstörung nach sich ziehen, sollte jederzeit möglich sein. Diese Möglichkeit zu erhalten ist auch Aufgabe im Umgang mit maschinell lernenden Systemen, die immer tiefer in unseren Alltag eindringen und dabei unser Verständnis von Vorgängen beeinflussen, obwohl sie unsere (Vor-)Urteile übernehmen und am Umgang mit Überraschungen scheitern.

Aus all diesen Gründen – als Ressource für die Wissenschaft, als Grundlage der *Politisierung* und vor dem Hintergrund, dass wir zunehmend maschinelle Artefakte einsetzen, die von Überraschungen nichts verstehen – scheint es mir lohnend, die Fähigkeit, sich überraschen zu lassen, mit dieser Arbeit als philosophisches Thema zu erschließen. Die zentralen Forschungsfragen, die mich dabei leiten, sind: Wie lässt sich das Überraschungsgeschehen phänomenologisch beschreiben? In welchem Verhältnis steht diese Beschreibung zur gegenwärtigen Forschungslage im Bereich der Kognitionswissenschaften und des maschinellen Lernens? Und was kann dazu beitragen, die Fähigkeit, sich überraschen zu lassen, zu kultivieren?

Wo sich diese Arbeit verortet

Wer im deutschsprachigen Raum nach einer Philosophie der Überraschung sucht, wird nicht fündig. Aus dem Alltagsverständnis heraus, in dem der Erwartungsbruch eine entscheidende Rolle spielt, lässt sich die Überraschung auch nicht prima facie unter das Staunen subsumieren. Dass hier dennoch zumindest eine Verwandtschaft bestehen könnte, wird im nächsten Abschnitt begründet.

Im englischen Sprachraum findet sich immerhin eine Monographie, die die *Philosophie der Überraschung* im Titel führt[4], diese wird dort anhand einer Untersuchung literarischer Überraschungen von Mark Currie im Rückgriff auf einige andere Autor*innen, allen voran Paul Ricoeur und Henri Bergson, allerdings erst entwickelt. Vor dieser mageren Ausgangslage baue ich meine Argumentation zunächst auf Untersuchungen aus anderen Disziplinen auf.

[4] Currie, Mark (2013): The Unexpected. Narrative Temporality and the Philosophy of Surprise, Edinburgh: Edinburgh University Press.

Die Kognitionswissenschaft, die sich gleichermaßen für menschliche wie für künstliche Intelligenz interessiert, ist wahrscheinlich diejenige Disziplin, die sich gegenwärtig am stärksten für das Thema interessiert, da der Umgang mit Überraschungen für sie eine Reihe noch ungelöster Probleme aufwirft. Medien- und Literaturwissenschaften hingegen wollen verstehen, was den Erfolg und die Kunst überraschender Wendungen ausmacht. In diesem Kontext sticht die Studie *Elements of Surprise*[5] von Vera Tobin heraus, der ich wesentliche Einsichten verdanke. Erziehungswissenschaften und Politik geht es darum, mit Hilfe von Überraschungen Neugier und Aufmerksamkeit zu wecken, während die Ökonomie das Ziel verfolgt, Überraschungen weitestgehend auszuschließen. Psychologie und Digitalmarketing sind am emotionalen Gehalt der Überraschung interessiert und untersuchen sie unter dem Stichwort *sentiment analysis*. Aus all diesen Fäden gemeinsam lässt sich ein umfangreiches Grundverständnis des Phänomens gewinnen, an das dann philosophische Fragestellungen anknüpfen. Wie dies genau erfolgt, beschreibe ich unter „Aufbau der Arbeit“.

Technikphilosophisch gibt das Thema in seinem Detailgrad zunächst ebenso wenig her. Allerdings gerät durch die Simulationsmetaphern der künstlichen Intelligenz und des maschinellen Lernens schnell die Fragestellung in den Blick, inwieweit Technik und Wissenschaft hier aufeinander einwirken und wie sich dadurch Möglichkeitsräume verändern. So kann im Abschnitt „Crux der Simulation“ an eine bestehende Forschungslage angeknüpft werden.

Im Zusammenhang mit dem aktuellen KI-Boom könnte man anhand der Thematik auch die Frage stellen, wie zutreffend die derzeit geschürten Erwartungen tatsächlich sind. Im Hinblick darauf lässt sich diese Arbeit als Fortführung der Arbeiten von Hubert L. Dreyfus lesen und beschreibt – allerdings mit weniger apodiktischem Anspruch – was Computer (noch) nicht können.

Verwandtschaft zum Staunen

Im historischen Wörterbuch der Philosophie gibt es keinen Eintrag zu Überraschung, interessanterweise jedoch findet sich darin sowohl als englische wie auch als französische

[5] Tobin, Vera (2018): Elements of Surprise. Our Mental Limits and the Satisfactions of Plot, Cambridge/London: Harvard University Press.

Übersetzung des Eintrags „Staunen; Bewunderung; Verwunderung“[6] das Wort *surprise*. Diese unterschiedliche Übersetzung des ursprünglich in Griechenland geprägten philosophischen Terminus θαυμάζειν wird jedoch nicht weiter thematisiert und die anschließende Darstellung der Begriffsgeschichte geht von der Begriffsgeschichte des deutschen *Staunens* aus, das im Ursprung auf ein Steifsein oder eine Starre zurückzuführen sei. Danach ist Staunen ein Begriff, dessen Bedeutung sich über lange Zeit zwischen Ver– und Bewunderung eingependelt hat. Sucht man nach weiteren Spuren der Überraschung, so lässt sich in der Begriffsgeschichte noch Folgendes finden:

1. in der frühgriechischen Dichtung die Verbindung zu einem Gefühl (bei der Begegnung mit Göttern)
2. bei Platon und Aristoteles die „erschließende Kraft“ des θαυμάζειν, die bloße Meinungen zu überwinden auffordere
3. bei Augustinus ein erstes Auftauchen des „Unerwarteten“
4. im englischen Sprachraum, genauer bei Francis Bacon, die Formulierung „broken knowledge“ und damit (in der Darstellung des historischen Wörterbuchs zum ersten Mal) die Thematisierung eines durch das Staunen („wonder“) ausgelösten kognitiven *Bruchs.*
5. bei Thomas Hobbes die Verknüpfung des Staunens mit Freude (hier: „über etwas Neues“)
6. Heideggers Bezug des *Erstaunens* auf „ein vereinzeltes *Ungewöhnliches*“
7. Lévinas’ Verständnis des Staunens als „Medium wirklicher Fremderfahrung“.

Bei dieser Darstellung drängt sich die Frage auf, ob unterschiedliche Begriffsentwicklungen im deutschsprachigen und englisch-/französischsprachigen Raum nicht vor allem eine Folge unterschiedlicher Übersetzungsgeschichten sein könnten. Auf „broken knowledge“ kann man ja eigentlich nur über den Bedeutungsgehalt von „Überraschung“ kommen, während Staunen einen Zustand bzw. ein zunächst folgenloses Gefühl beschreibt. In „broken knowledge“ dagegen steckt bereits eine kognitive Bewegung, die sich leichter der Überraschung zuschreiben lässt. Dieser Vermutung tiefer nachzugehen ist jedoch hier nicht der Ort, und vielleicht könnte ihr schon die Tatsache widersprechen, dass auch im Stanford Dictionary kein Eintrag unter „surprise“ zu finden ist – allerdings auch keiner unter „wonder“ oder „astonishment“.

---

[6] Vgl. Jain/Trappe (1998). Der folgende Absatz paraphrasiert den Eintrag „Staunen“ aus dem Historischen Wörterbuch der Philosophie dergestalt, dass die möglichen Folgen unterschiedlicher Übersetzungen von θαυμάζειν erkennbar werden. Vielleicht ist es vor diesem Hintergrund auch kein Zufall, dass die erste Anregung für die vorliegende Arbeit von einer französischsprachigen Philosophin, nämlich Vinciane Despret, ausging. Gleichzeitig soll jedoch erwähnt werden, dass in der Stanford Encyclopedia of Philosophy auch kein Eintrag für „surprise“ zu finden ist – genauso wenig übrigens wie für andere Übersetzungen von θαυμάζειν wie „wonder“ oder „astonishment“.

Wenn wir davon ausgehen, dass die Überraschung immer mit einem Brechen und Revidieren von vorhandenen Vorstellungen oder Erwartungen einhergeht, kann eine Philosophie der Überraschung also am ehesten über Francis Bacon oder Emmanuel Lévinas beim Staunen anknüpfen. Letzteres wird unten auch geschehen. Aber mit dem Staunen ist die Überraschung allenfalls verwandt. Sie steht der Ver-Wunderung näher als der Bewunderung. Wenn wir sie in den Mittelpunkt einer Untersuchung stellen wollen, müssen wir bei genau den Eigenheiten anfangen, die sie zu einem Sonderfall machen.

Aufbau der Arbeit

Die großen Linien meiner Argumentation ergeben sich aus drei Zugangsrichtungen, die auch die Gliederung des Hauptteils meiner Arbeit bestimmen: einer narratologischen, einer kognitionswissenschaftlichen und einer genuin philosophischen.

Aus der Forschungslage, die an expliziten Untersuchungen nur die zwei eingangs erwähnten Publikationen zur Überraschung im Fiktionalen aufweist, ergibt sich der Einstieg über die Wissenschaft vom Erzählen, den ich mit „Inszenierung und Lesen“ betitelt habe. Sowohl Vera Tobin als auch Mark Currie, deren Überlegungen hier aufgegriffen werden, stammen aus dem englischen Sprachraum – was noch einmal meinen Verdacht bestärkt, dass es für englisch- oder französischsprachige Autor*innen näher liegen könnte, hinter dem Begriff ein Thema mit geisteswissenschaftlicher Relevanz und Tiefe zu vermuten. Während Currie, der vom Lesen her argumentiert, überraschende Erzählungen im Hinblick auf Konfigurationen der Zeitlichkeit untersucht, befasst sich Vera Tobin kognitionswissenschaftlich informiert mit der künstlerischen Inszenierung von Überraschung in Literatur und Film. So wird in diesem ersten Zugang bereits ein Grundverständnis der Überraschung erarbeitet, das im weiteren Verlauf der Arbeit präzisiert und auf seine Übertragbarkeit auf nicht-fiktionale Erfahrungen hin überprüft werden soll.

Der zweite Zugang zur Überraschung erfolgt über die Kognitionswissenschaften, deren Interesse am Thema sich mit Tobin bereits gezeigt hatte und das wir nun bestätigt finden. Überraschung erweist sich im Abschnitt „Stand der Forschung: Wie Menschen lernen“ als ein wesentlicher Motor für das menschliche Lernen, und im Abschnitt „Offene Fragen: Maschinelles Lernen zwischen Plastizität und Stabilität“ als ungelöstes Problem für das Lernen von Maschinen. Aus dieser Diskrepanz ergeben sich für die Kognitionswissenschaft, die seit ihrer Entstehung eng mit der KI-Forschung verbunden ist, zwei Fragen, die jeweils einen Seitenblick erfordern. Das Kapitel „Aufmerksam, neugierig oder überrascht?“ fragt danach, was mit bestimmten Experimenten eigentlich gemessen wird und wie sich diese

drei Begriffe voneinander abgrenzen lassen. Unter „Die Crux der Simulation“ soll mit Hilfe von Gabriele Gramelsberger zumindest angerissen werden, welche Probleme sich durch die enge Verknüpfung von Grundlagenforschung und Simulation über Computerexperimente, die den kognitionswissenschaftlichen Zugang bestimmt, ergeben können. Vor diesem Hintergrund drängt sich die Vermutung auf, dass die bestehenden Erklärungslücken, die sich darin zeigen, dass maschinelles Lernen in mehrfacher Hinsicht noch hinter dem menschlichen Lernen zurücksteht, auch darin begründet sein könnten, dass die Sprache der zeitgenössischen Neurophysiologie allein nicht ausreicht, um zum Beispiel die Fähigkeit, sich überraschen zu lassen, hinreichend adäquat und umfassend zu beschreiben.

Es gilt also, einen weiteren Zugang zu erschließen. Auf der Suche nach ergänzenden oder besseren Beschreibungsmöglichkeiten begebe ich mich im dritten und letzten großen Teil der Arbeit in den Bereich der Philosophie. „Philosophische Zugänge“ ergeben sich aus einer Bandbreite verschiedener Strömungen. Ausgehend von den Neurowissenschaften und einer Auseinandersetzung mit aktuellen technologischen Entwicklungen versucht Catherine Malabou, mit ihrem Plastizitätsbegriff eine Brücke zwischen Gehirnforschung und Philosophie zu bauen. In einer Art Gegenbewegung zu ihrem im dialektischen Materialismus und den Ansprüchen der Kritischen Theorie gefangenen Rationalismus möchte ich dann zunächst den Alteritätsbegriff von Emmanuel Lévinas zusammen mit seiner Konzeption einer thematisierenden Intentionalität in unser Nachdenken einführen. Ich sehe hierin einen ersten Anknüpfungspunkt an den von der Erzähltheorie herkommenden Teil der Arbeit. Ebenfalls phänomenologisch argumentiert Bernhard Waldenfels in Bezug auf die Aufmerksamkeit, die insofern Teil der Überraschungserfahrung ist, als in der Überraschung immer auch eine besondere Fokussierung und Umlenkung unserer Aufmerksamkeit erfolgt. Die narrativen Aspekte der Fähigkeit, sich überraschen zu lassen, lassen sich noch von zwei weiteren philosophischen Ansätzen untermauern. Zunächst von der Warte einer soziologisch informierten Betrachtung der Verhältnisse aus, die sich zwischen fiktionalen und „realen“ Realitäten beschreiben lassen und die in wirtschaftsphilosophische Überlegungen zu unserem Verhältnis zur Zukunft und zur Erwünschtheit bzw. Unerwünschtheit von Überraschungen münden. Mit Christiane Voss schließlich lässt sich die Überraschung als Emotion beschreiben, und damit als Zusammenspiel verschiedener Komponenten, die erst narrativ zusammengefasst und als Emotion gedeutet werden können. Die Fähigkeit, sich überraschen zu lassen, beruht, das wird das Ergebnis meiner Arbeit sein, ganz wesentlich auf narrativer Kompetenz. Durch die sprachliche Verfasstheit und die erst im Zusammenspiel mit Anderen entstehenden Deutungsmuster unserer erzählenden Intentionalität ergibt sich eine unhintergehbare Kopplung der Fähigkeit, sich überraschen zu lassen, an unser soziales Sein. Diesen Aspekt greife ich am Ende der Arbeit auf. Isabelle Stengers und ihre Lektüre von Alfred North Whitehead helfen zu verstehen, wie

sich die Fähigkeit, sich überraschen zu lassen, in gemeinschaftlichen Kontexten denken und kultivieren ließe.

Inszenierung und Lesen

Von der Kunst, überraschende Geschichten zu erzählen

Was macht gut gemachte Überraschungen in Geschichten eigentlich aus? Warum mögen wir sie? Und wie kommt es, dass sie zielsicher immer wieder funktionieren? Vera Tobin berichtet vom Vergnügen daran, sich überraschen zu lassen und vertritt in *Elements of Surprise – Our Mental Limits and the Satisfaction of Plot* die These, dass Überraschungen bestimmte Heuristiken im Denken ausnutzen, die den meisten Menschen gemein sind. Und dass wir es mögen, in Literatur und Film überrascht zu werden – zumindest wenn dies auf freundliche und nachvollziehbare Weise geschieht –, weil wir so etwas über uns selbst und unser Umfeld lernen.

Überraschungen hängen auch in ihrem Verständnis eng mit Erwartungen zusammen, sind jedoch mehr als ein Maß für die Unwahrscheinlichkeit des Eintreffens eines bestimmten Signals, wie es die Informationstheorie definiert. Im Gegensatz zur informationstheoretischen Definition beschreibt Tobin Überraschung als eine *bewusste Erfahrung*. Im Englischen gibt es, anders als im Deutschen, dafür zwar verwandte, aber doch unterschiedliche Begriffe: *surprise* und *surprisal*.

> „There can be no surprise without expectations. Fortunately or unfortunately, human cognition runs on a steady diet of them….. the embodied experiences of both language comprehension and making our way through the world may well run along a continual stream of „surprisal“ … that is mismatches between what our processing systems predict and what they actually encounter. (…) Surprisal, unlike surprise, is „subpersonal“; that is, the relative implausibility of a given signal, given the current model of the world, does not necessarily correspond to the conscious personal experience of surprise.“[7]

Der Unterschied zwischen *surprise* und *surprisal* lässt sich für sie kognitionswissenschaftlich stichhaltig begründen. Denn dort gelte es inzwischen als erwiesen, dass menschliche Wahrnehmung nicht durch einen einfachen bottom-up-Ansatz erklärt werden kann, sondern dass sensorische Stimuli erst im Wechselverhältnis mit Erwartungen Wahrnehmung

[7] Tobin 2018, S. 91f.

erzeugen. Erwartungen helfen, die Verarbeitung sensorischer Eindrücke zu filtern und zu organisieren, genau wie umgekehrt Sensorisches zu einer (Neu-)Formung von Erwartungen beiträgt. Zudem versetzen uns Erwartungen in die Lage, partielle und mehrdeutige Informationen schnell und richtig einzuordnen, indem sie Kontextbezüge herstellen.

> „They help us recognize things in ‚congruent' contexts more easily than otherwise (…): we easiliy see bread, not mailboxes, in kitchens, and wristwatches, not coins, on wrists."[8]

Überraschung nun geht für Tobin mit einer neuen Einsicht (*insight)* und damit verbunden einem lustvollen Gefühl einher, so wie es sich anfühlt, wenn wir uns plötzlich ganz sicher sind, die Lösung eines Rätsels gefunden zu haben. Donald Hebb, der als Vordenker künstlicher neuronaler Netze (KNN) gilt, habe diesen Aspekt 1949 als ein faszinierendes Phänomen beschrieben, das mit der Anpassung innerer Vorstellungen an neue Situationen zu tun habe und damit vereinfachende Lernmodelle in Frage stelle. Es sei,

> "what allows humans and other sophisticated primates to restructure the associations they have built up to fit new situations at hand, and any models of human cognition that exclude it (… are) destined to fail"[9].

Diese Beschreibung steht augenscheinlich in der Tradition des „broken knowledge". Sie lässt sich gut auch auf das Überraschungsgeschehen anwenden und beschreibt es bereits als Herausforderung für die späteren Kognitionswissenschaften.

Auch fiktionale Erzählungen, unabhängig davon, über welches Medium sie verbreitet werden, können Menschen Überraschungen erleben lassen. Ob wir eine überraschende Wendung in Film oder Literatur jedoch akzeptieren, hängt Tobin zufolge davon ab, für wie „realistisch" wir sie halten. Das verweist auf die Kongruenzforderung unserer Wahrnehmung. Überraschende Wendungen müssen eine mit den verfügbaren Informationen vereinbare neue Sicht der Dinge erlauben, denn sonst sind wir aus dem Kontext geworfen und fühlen uns von der Erzählerin betrogen. Damit das nicht der Fall ist, müssen wir zumindest die Chance gehabt haben, selber „drauf zu kommen". Die gelungene überraschende Wendung ist also so beschaffen, dass sie lediglich ein neues Licht auf die schon vorhandenen Informationen wirft. Sie entwertet keine Informationen, sondern *Interpretationen*.

Um zu veranschaulichen, wie die menschliche Interpretationstätigkeit von versierten Erzähler*innen absichtsvoll in die Irre geleitet werden kann, schildert Tobin wiederkehrende Muster, die in überraschenden Erzählungen verwendet werden. Dazu gehören *frame shifts,* die Leser*innen dazu zu bringen, Informationen falschen Kontexten zuzuordnen, irreführende Erzählchronologien (*managed reveal*), das Einstreuen falscher Behauptungen,

[8] Ebd., S. 93.

[9] Ebd., S. 178.

die für Informationen gehalten werden können (*finessing misinformation),* oder das „Vergraben“ von Schlüsselinformationen in einem Berg von Unwichtigem *(burying information).* Die erzählerische Kunst bestehe, betont Tobin mit den Worten der Krimiautorin Dorothy Sayers, darin, die Leser*innen mit Hilfe solcher Strategien unmerklich dazu zu verführen, sich selbst etwas vorzumachen:

> “the right method is to tell the *truth* in such a way, that the *intelligent* reader is seduced into telling the lie for himself“[10].

Aus der obigen Aufzählung wurde bereits deutlich, dass die erzählerische Chronologie nur eines von vielen Mitteln ist, mit denen das erreicht werden kann. Tobin interessiert sich schwerpunktmäßig für Strategien, die typischerweise kognitive Verzerrungen ausnutzen und so verhindern, dass wir Informationen korrekt einordnen. Nachweisbar vergessen Menschen zum Beispiel regelmäßig die Quellen von Informationen oder bringen sie durcheinander, so dass Tatsachen, Meinungen und Schilderungen aus einer vermeintlich neutralen Dritte-Person-Perspektive als gleichwertig erscheinen.

Schon die Geschichtsförmigkeit selbst kann zum Problem werden, indem ihr *unified design* als Bestandteil einer kulturell verankerten narrativen Logik zu der Unterstellung verleiten kann, dass alle Elemente einer Geschichte Teil eines Gesamtzusammenhangs seien.

> „The logic of narrative suggests that all elements of a story are (or should be) on some level manifestations of a unified design. This impression is part of what gives stories their appeal as a tool for finding meaning in the mess of lived experience.“ [11]

Roland Barthes bezeichnete diese *Designqualität* von Stories als unausweichlich sogar dann, wenn Autor*innen versuchten, dagegen zu arbeiten. Sie unterminiert jedoch dauerhaft unsere Bereitschaft, Willkür, Beliebigkeit und Zufall zu akzeptieren und fördert so die Unterstellung einer Kausalität, wo unsere Wahrnehmung zeitliche Zusammenhänge findet.

Ein weiterer kognitionspsychologisch belegbarer Mechanismus, der augenscheinlich zu den Bestandteilen unserer Geschichtenerwartungen gehört, ist nach Tobin der „Fluch des Wissens“ – the *curse of knowledge*. Der Terminus beschreibt eine Tendenz menschlichen Denkens, die sowohl das Verhältnis zur Vergangenheit als auch das zu anderen Menschen betrifft: Es fällt uns offensichtlich schwer, uns in andere Situationen hineinzuversetzen, sogar dann, wenn es sich um unsere eigene Vergangenheit handelt. Unsere aktuelle Perspektive und unser aktueller Wissensstand überlagern Erinnerungen und beeinträchtigen uns maßgeblich in der Einschätzung dessen, was andere wissen, fühlen oder erleben.

---

[10] Zitiert ebd., S. 176, Hervorhebung im Original.

[11] Ebd., S. 279.

Zusätzlich könne der Effekt aller Erzählstrategien, die uns auf falsche Spuren locken sollen, dadurch verstärkt werden, dass wir in unserem Bewusstseinsstrom die gesamte Erzählung aus der Ich-Perspektive erleben:

> „Across any of these narrative perspective shifts, the main bulk of our lived experience as reader or viewer remains consistent. This fact can conspire with other techniques for misdirection to enhance mental contamination effects, making it less likely that people will keep source information crisply distinguished in memory.“[12]

Lediglich die Fehleranfälligkeit der Heuristiken, die derartige Irrtümer verursachen, in den Vordergrund zu stellen, wie es im Begriff „kognitive Verzerrungen“ zum Ausdruck kommt, hält Tobin jedoch für falsch. Aktuelle Forschung dagegen stelle ihren Nutzen für das Leben in sozialen Gemeinschaften heraus. Als soziale Wesen sind wir maßgeblich darauf angewiesen, uns eine Vorstellung davon zu machen, was andere denken, fühlen, wollen und vielleicht im nächsten Moment tun werden. Erfahrungen von uns auf andere zu projizieren sei eine effiziente Methode, Vorstellungen davon zu entwickeln, was andere denken und wissen könnten.

> „We depend on our ability to think about complex perspectives on a situation all the time navigating in the social world, and our skills of perspective taking are both very sophisticated and also limited in some very predictable ways. As we navigate social situations, we need to make very quick backstage assessments of what’s going on and what other people think and know, and projecting information from our own perspective is a fast and efficient way of generating good approximations. Still, as useful as these heuristics are, they can also be the source of many mistakes“.[13]

Wie viele andere Tiere, nicht nur Säugetiere, erlernen wir Sozialverhalten indem wir spielen. Geschichten sind in dieser Hinsicht Spiele. Auch durch sie üben wir uns darin, Theorien über das Empfinden und mögliche Verhalten Anderer zu bilden und sie auch wieder zu überprüfen.

Überraschungen in fiktionalen Erzählungen trainieren also, wie Tobin am Beispiel von *Emma* von Jane Austen ausführt, unter anderem die Fähigkeit, sich in andere hineinzuversetzen. Zugleich machen sie uns bewusst, wie fragwürdig diese Einschätzungen sein können:

> „Through these machinations, the novel trains its readers as the events of the novel train Emma: both to value the act of imaginative entry into another's consciousness and to appreciate that the value and accuracy of any individual's perspective may be questionable.“[14]

Sich das immer wieder vor Augen zu führen (oder vor Augen führen zu lassen) hat durchaus auch politische Relevanz. Am Ende ihrer Untersuchung schildert Tobin, wie dieselben Mechanismen nämlich auch im Kontext von Politik oder Rechtsprechung zu Fehlurteilen

[12] Ebd., S. 84.

[13] Ebd., S. 56f.

[14] Ebd., S. 170.

führen können[15], die – anders als in Buch oder Film – vielleicht niemals aufgeklärt werden, oder manchmal erst nach einer Zeit, in der sie viel zu viel Gelegenheit hatten, Schaden anzurichten.

Wichtige Elemente der Fähigkeit, sich überraschen zu lassen, die sich mit Tobin für diese Arbeit erschließen, sind zusammengefasst:

A. Überraschung (*surprise*) ist eine persönliche *Erfahrung*, die zwar mit einer Nicht-Übereinstimmung von Erwartung und tatsächlich Vorgefundenem einhergeht, aber nicht auf den informationstheoretischen Begriff der Überraschung (*surprisal*) reduziert werden kann.

B. Überraschungen lassen sich mit Hilfe des Hebb'schen Begriffs der Einsicht (*insight*) auch als kognitive Restrukturierungsbewegung bezeichnen.

C. Bei Erzählungen findet diese Restrukturierung auf der Ebene der *Interpretation* statt.

D. Erzählungen werden in unserem Bewusstseinsstrom wie andere Sinneseindrücke auch immer aus der Ich-Perspektive erfahren. Reflexion und Interpretation erfordern die Möglichkeit, eine kognitive Meta-Ebene einnehmen zu können.

E. Auf dieser Meta-Ebene arbeiten wir mit Heuristiken, die auf spezifische Weisen fehleranfällig sind.

F. Erzählungen trainieren erzählungsförmige Erwartungen, insbesondere in Bezug auf ihr *unified design,* die auf Weltbezüge übertragen zu Fehlurteilen verleiten können.

G. Durch fiktionale Geschichten können wir unsere sozialen Kompetenzen erweitern, da sie spielerisch nicht nur andere Perspektiven, sondern gelegentlich auch Fallstricke unserer interpretierenden Wahrnehmung erfahrbar machen.

Inwieweit sich auch die anderen Befunde auf Überraschungserfahrungen außerhalb des Fiktionalen übertragen lassen, muss der Fortgang der Arbeit zeigen. Hierzu verweise ich insbesondere auf den Abschnitt „Berührung und thematisierendes Bewusstsein", der mit Emmanuel Lévinas den erzählenden Charakter der Intentionalität in den Fokus rückt, und das Kapitel „Überraschung als Emotion", in dem in Anlehnung an Christiane Voss eine allgemeine Theorie der Überraschung entwickelt wird. Auch der Abschnitt „Fiktionen, Erwartungen, Zukünfte" kann zur Erhellung der Frage beitragen, welchen Anteil Fiktionen an unseren Weltbezügen haben. Der Faden, der auf die Bedeutung der Überraschungsfähigkeit für unser Leben in sozialen und anderen Bezügen verweist, wird am Ende dieser Arbeit mit

[15] Am Ende ihres Buches fasst Tobin auf diese Weise ihre Erkenntnisse aus den Arbeiten von Peter Brooks zusammen, der bekannte juristische Problemfälle auf ihre narrativen Dimensionen hin untersucht hat. Vgl. ebd., S. 273ff.

Isabelle Stengers wieder aufgenommen. Zunächst soll jedoch mit Mark Currie ein zweiter Autor zu Wort kommen, der sich mit überraschenden Plots in der Literatur befasst.

Narrativität und Zeitlichkeit

Wie Geschichten Geschichtenerwartungen prägen, lässt sich anhand von Paul Ricoeurs Konzept mimetischer Prozesse nachvollziehen. Mark Currie macht es zur Grundlage seiner Studie *The Unexpected – Narrative Temporality and the Philosophy of Surprise*. Diese wird ihn dazu führen zu sagen, dass auch unser Zeitverständnis maßgeblich von der Form der Erzählung geprägt ist. Wie aber begegnen wir dann dem Unvorhersehbaren? Die Überraschungserfahrung im Lesen legt es für ihn nahe, das *Unvorhersehbare* auf das *Unvorhergesehene* zu reduzieren, das wir sogar in einem eigenen Tempus ausdrücken, dem Futur II. Seine Argumentation verläuft dabei wie folgt.

Der mimetische Prozess zwischen Text und Leser*innen umfasst drei Stufen. Mit *Prefiguration* bezeichnet Ricoeur von der konkreten Erzählung unabhängige, die Wahrnehmung vorformende, auch historisch-kulturell geformte Grundlagen unserer Geschichtenerwartungen. Hierzu gehört die zeitlich erlebte Erfahrung unseres Seins in der Welt, unser Verständnis von einem "Jetzt" zwischen Vergangenheit und Zukunft. Dieses Vorverständnis geht in die zweite Stufe des mimetischen Prozesses ein, die Ricoeur als *Konfiguration* bezeichnet. Hier geht es um die Auswahl, Zusammenstellung und Anordnung narrativer Elemente durch die aktuell Erzählenden. Die Zeitauffassung dieser zweiten Stufe sei, schreibt Currie, eine andere als die unseres alltäglichen Erlebens, insbesondere sei sie nicht linear und unterscheide sich damit auf spezifische Weise von der als gerichtet erfahrenen zeitlichen Abfolge lebendiger Erfahrung:

> "configuration has definite temporal features, such as the assembling of events into a followable sequence, the crafting of an ending from which events can be seen as a whole, and therefore the construction of an alternative view of time, based not on time flow, or the arrow of time, but one in which the ending can be read in the beginning, and the beginning in the ending, and which teach us to read time itself backwards“ [16].

Die dritte Stufe der Mimesis schließlich, für die Ricoeur den Begriff der *Refiguration* gewählt hat, beschreibt die Rückwirkung des Narrativs auf die künftigen Geschichtenerwartungen der Lesenden. Geschichten trainieren Geschichtenerwartungen, lasen wir bei Tobin. Für

[16] Currie 2013, S. 44f.

Currie geht es um die Rück- und Auswirkungen auf unseren gesamten Erfahrungshorizont, der hier – das wäre in einem anderen Rahmen zu überprüfen – vielleicht mit Wittgensteins *Weltbildern* zusammenfällt. Ricoeurs Begriff der Mimesis verdeutliche den Weltbezug von Geschichten als Weltbezug in zwei Richtungen. Während die erzählte Welt die Handlungswelt imitiere, wirke die Erzählung umgekehrt auch auf die Welt der Handlungen zurück, in der umgekehrt Erzählungen imitiert würden:

> "in this reciprocity between mimesis and reverse mimesis, where life imitates art and art imitates life, the hermeneutic circle of narrative, or, as Ricoeur has it, the circle of narrative and time, revolves"[17].

Das Zeiterleben in der Erzählung wirke so auch auf unser Zeitverständnis zurück, insbesondere lernten wir durch narrative Konfigurationen, Zeit im Rückblick – also gewissermaßen rückwärts – zu betrachten sowie unser Handeln auf ein erwartbares Ende hin auszurichten. Damit erhalte, wie es bei Ricoeur heißt, die gegenwärtige Erfahrung für uns die Form einer *noch nicht* erzählten Geschichte, eines Ereignisses, von dem wir erwarten, das es sich retrospektiv in eine zusammenhängende Erzählung fügen und interpretieren lassen wird[18].

Der Moment der Überraschung, die Erfahrung des Unvorhergesehenen selbst, ist zunächst nur ein Moment der Verneinung, in dem unser bisheriges Verständnis einen Widerspruch erfährt, als „broken knowledge“ erscheint und revidiert werden muss, der selbst jedoch in seiner Gegenwärtigkeit nicht zu fassen ist.

> „The unexpected, the contradiction of everything that arrives as complete knowledge, produces nothing at all. (…) This is what I take the novel to be saying about memory and presence, not that unexpected moments are difficult to grasp, but that, in being difficult to grasp, they reveal the structure of time more generally, as a flight from presence, and so the ungraspability, the nothingness, of presence in general.“[19]

Eine (neue) Geschichte wird daraus erst im Rückblick. Die grammatikalische Zeitform, mit der dieses Werden erzählt wird, ist nach Currie das Futur II[20]:

> „I have been arguing (…) that the very combination of retrospect and futurity is the thing that gives narrative its special place in the encoding of temporal becoming: that the reading of a narrative, governed as it is by the structure of future anteriority,

---

[17] Ebd., S. 45.

[18] Vgl. ebd., S. 46.

[19] Ebd. S. 174. Die vorausgegangene Textanalyse bezieht sich zwar auf eine bestimmte Erzählung, nämlich Julian Barnes's *The Sense of an Ending,* Currie verwendet eine Szene daraus jedoch am Ende auch, um seine eigene These zu illustrieren.

[20] Was nicht für alle Zeitformen und über alle Sprachen hinweg gilt – für den konkreten Fall des Futur II (*Future Anterior* oder *Future Perfect* auf Englisch) lässt sich sagen, dass sie im Englischen und im Deutschen analog verwendet werden, um entweder eine Vermutung oder einen in der Zukunft vermutlich beendeten Sachverhalt ausdrücken.

> is the very model of temporal becoming as Lacan understood it, of ‚what I will have been, given what I am in the process of becoming'".[21]

Unsere Auffassung von Zeit geht auf kognitive Strukturen zurück, die es uns erlauben, aus der „Leere der Gegenwart" einen Zustand vorwegzunehmen, in dem wir rückblickend eine Geschichte erkennen oder in dem sich die Erfahrung der Gegenwart in Form einer Geschichte vollenden wird:

> „Finally, it can be claimed, that the apprehension of time itself depends on structures that allow this kind of cognitive projection forwards from the emptiness of presence to some notional state of retrospect or completion. In this respect, the future anterior is the structure, in narrative as in temporal becoming more generally, that makes the unexpected intelligible."[22]

Was sich von Currie für diese Arbeit lernen lässt: Unser ganzes Zeitverständnis beruht ihm zu Folge darauf, dass wir Geschichten entwickeln, und zwar auf das Ziel hin, einst die Gegenwart verstanden zu haben. Ähnliches, nur allgemeiner formuliert, hatten wir bei Tobin schon gesehen. Es erinnert an die *unausweichliche Designqualität* von Geschichten bei Tobin und Barthes. Currie erweitert dieses Verständnis auf überzeugende Weise um die Dimension der Zeitlichkeit und hilft damit, noch genauer zu verstehen, wie weitreichend Erzählungen auf die Thematisierungen unseres Bewusstseins einwirken.

Der narratologische Blick auf die Fähigkeit, sich überraschen zu lassen, ist auf diese Weise zu einem spannenden Einstieg in phänomenologische Betrachtungen geraten. Wenn die Fähigkeit, sich überraschen zu lassen, die Fähigkeit beinhaltet, sich selbst immer wieder neue Geschichten zu erzählen, diese aber auch durch neue Erfahrungen überprüfen und verändern zu können, dann beruht sie auf der Fähigkeit zur Selbstreflexion. Sie erfordert damit eine kognitive Struktur, die mehrere Ebenen und eine Auffassung von Zeitlichkeit umfasst, die narrativ geformt ist.

Überraschungen zwischen Kognitionswissenschaften und KI – sind Maschinen fähig, sich überraschen zu lassen?

Nach der Darstellung des narratologischen Forschungsstands werden im Folgenden die Erkenntnisse der Kognitionsforschung zur Überraschung in den Blick genommen. Die moderne Kognitionswissenschaft ist ein interdisziplinäres Projekt, das seit seiner Gründung

[21] Currie 2013, S. 175.

[22] Ebd., S. 175.

versucht, informationsverarbeitende Prozesse unabhängig von ihrem „Medium“ zu beschreiben, das sich also sowohl mit der Informationsverarbeitung von Organismen wie mit der künstlicher Systeme befasst. Eine starke Technologie-Orientierung ist ihr daher von Anfang an eingeschrieben[23].

Der französische Professor für experimentelle Kognitionspsychologie, Stanislas Dehaene, zum Zeitpunkt der Veröffentlichung seines Buches Präsident des wissenschaftlichen Beirats des französischen Bildungsministeriums, hat vor kurzem einen umfassenden Überblick über aktuelle Forschungsergebnisse zu einem der wichtigsten kognitionswissenschaftlichen Themen publiziert: dem Lernen[24]. Mit dem Ziel, einerseits Bildungsansätze auf neue wissenschaftliche Erkenntnisse abzustimmen, und andererseits daraus informiert die Möglichkeiten und Grenzen aktueller KI–Ansätze aufzuzeigen, beschreibt das Buch, was in den Augen seiner Wissenschaft menschliches Lernen ausmacht und was menschliche Gehirne (!) – derzeit noch – besser lernen lässt als aktuelle maschinelle Artefakte. Sein Fazit sei hier vorweg genommen:

> "(…) machines still have a long way to go. To improve, they will need many of the ingredients that we reviewed here: an internal language of thought that allows concepts to be flexibly recombined; algorithms that reason with probability distributions; a curiosity function; effective systems for managing attention and memory; and perhaps a sleep/wake algorithms that expands the training set and increases the chances of discovery. Algorithms of this type are beginning to appear, but they remain light years away from the performance of a new born baby. The brain keeps the upper hand over machines, and I predict that it will be for a long time."[25]

Ein wesentlicher Motor menschlichen Lernens ist, so lassen sich die von Dahaene präsentierten Forschungsergebnisse zusammenfassen, die Fähigkeit, sich überraschen zu lassen. Woraus dies abgeleitet wird und welche Folgerungen in Kognitionswissenschaften und KI-Forschung daraus gezogen werden, soll hier in einem ersten Schritt ausführlicher betrachtet und in ein Verhältnis zu den oben bereits gewonnenen Erkenntnissen gebracht werden.

In einem zweiten Schritt werde ich einen Blick auf die KI-Forschung werfen, wo die Problematik als Plastizitäts-/Stabilitäts-Dilemma diskutiert wird und sich in Problemen wie dem "katastrophalen Vergessen" manifestiert. Es wird in diesem Zusammenhang auch zu überlegen sein, was es bedeutet, die Erforschung kognitiver Fähigkeiten an den Erfordernissen und Zielen einer Simulation menschlicher Intelligenz auszurichten.

---

[23] Dies ist sicherlich auch den Abhängigkeiten von ökonomisch interessierter Forschungsfinanzierung Forschung geschuldet. Da eine Ausführung dieses Gedankens den Rahmen dieser Arbeit sprengen würde, verweise ich hierzu auf die sehr lesenswerte Arbeit des Chemnitzer Soziologen Andreas Bischof. Bischof, Andreas (2017): Soziale Maschinen bauen. Epistemische Praktiken der Sozialrobotik, Bielefeld: transcript.

[24] Dehaene, Stanislas (2021): How We Learn. Why Brains Learn Better Than Any Machine ... for Now, New York: Penguin.

[25] Dehaene 2021, S. 239.

Stand der Forschung: Wie Menschen lernen

Für Stanislas Dehaene bildet der Umgang mit Überraschungen von früher Kindheit an ein zentrales Moment des Lernens. Lernen, das er definiert als die Ausbildung eines inneren *Modells* der äußeren Welt[26], und als dessen Quintessenz er die Fähigkeit versteht, sich schnellstmöglich an unvorhersehbare Bedingungen anzupassen[27], wird danach von folgenden vier Funktionen ermöglicht: Aufmerksamkeit, Neugier, Fehlerfeedback und Konsolidierung[28].

Diese „Gehirnfunktionen" ließen sich auch bei zahlreichen anderen Spezies beobachten, was uns von diesen jedoch unterscheide, seien „our social brain and language skills", durch die wir sie noch effizienter nutzen könnten[29]. So können andere Menschen unsere Aufmerksamkeit gezielt auf etwas lenken – eines der fundamentalen Prinzipien des Spracherwerbs. Wenn jemand auf etwas zeigt und dabei eine bestimmte Lautfolge von sich gibt, dann erspart uns das mühsame Versuche, selbst herauszufinden, wie wir anderen deutlich machen können, was wir meinen (den Schmetterling und nicht die Blume, das Gras, den Käfer, oder sonst irgend etwas, was sich gerade in unserem Wahrnehmungsfeld befinden könnte). Es ist das Thema der Aufmerksamkeit, an dem Dehaene unser In-der-Welt-Sein als ζῷον πολιτικόν thematisiert[30].

Der von ihm so genannte „Neugier-Algorithmus", der Neugier als Manifestation des kindlichen Wunsches versteht, sich ein inneres Modell der Welt zu bilden, ist für diese Arbeit insofern interessant, als Dehaene neugieriges Lernen an verschiedene Grade der Überraschung – die er hier synonym mit Neuheit/„novelty" verwendet - koppelt und damit die bei Tobin bereits aufgeworfene Frage der Akzeptabilität von Überraschungen in anderen, nämlich messbaren, Begriffen beantwortet. Es sei ein Zuviel an Neuheit, das genauso wie zu wenig dem Entstehen von Neugier als dem Wunsch, zu verstehen, entgegenwirken könne:

> „This theory explains why curiosity is not directly related to the degree of surprise or novelty but instead follows a bell curve (…) Between the boredom of the too simple and the repulsion of the too complex, our curiosity naturally directs us toward new and accessible fields. But this attraction keeps changing. As we master them, the objects that once seemed attractive loose their appeal, and we redirect our curiosity toward new challenges."[31]

[26] Vgl. Ebd., S. 5: „to learn is to form an internal model of the external world".

[27] Ebd., S. XIX.

[28] Vgl. ebd., S. 145f.

[29] Ebd.

[30] Ebd., S. 168.

[31] Ebd., S. 190f.

Verstehe man Neugier in dieser Weise, folge daraus, dass schon kleinen Kindern bewusst sein muss, dass sie etwas bzw. was sie noch nicht wissen. Daraus folge, dass es eine zweite kognitive Ebene geben müsse, die er „metacognition“ nennt:

> „This vision of curiosity leads to an interesting prediction. It implies that in order for children to be curious, they must be aware of what they do not yet know. In other words, they must possess *metacognitive* faculties at an early age. ‚Metacognition‘ is cognition over cognition: the set of higher-order cognitive systems that monitor our mental processes. According to the gap theory of curiosity, metacognitive systems must constantly supervise our learning, evaluating what we know and don’t know, whether we are wrong or not (…) – metacognition encompasses everything we know about our own minds.“[32]

Auch für die Feststellung von Diskrepanzen zwischen „inneren Repräsentationen“ und „der äußeren Welt“ sei jenes zweite "System" verantwortlich. Auffällig ist, wie stark in den Begriffen, die Dehaene verwendet, aktuell verfügbare Technologien der Informatik anklingen. Dieses metakognitive System, so schreibt er nämlich, müsse auf eine Art Arbeitsspeicher ("working memory") zurückgreifen können, durch den sämtliche Elemente des gegenwärtig ablaufenden Programms ("all the elements of the ongoing program: intermediate results, steps already carried out, operations remaining to be performed"[33]) verfügbar bleiben.

Die Ausreifung des präfrontalen Cortex, wo sich während solcher Prozesse vermehrte Hirnaktivitäten messen lassen, dauere beim Menschen fünfzehn bis zwanzig Jahre. Wenn Babys zum Beispiel regelmäßig daran scheitern, ein Objekt wiederzufinden, das vor ihren Augen an Ort B versteckt wird, obwohl es bislang routinemäßig an Ort A lag, erkläre sich genau damit. Denn dieses zweite System sei in den ersten 10 Lebensmonaten noch nicht weit genug ausgebildet, um Diskrepanzen zwischen Gewohntem und aktueller Wahrnehmung korrekt aufzulösen[34].

Schließlich beruhe unsere Fähigkeit zu lebenslangem Lernen darauf, Gelerntes zu konsolidieren und dauerhaft verfügbar zu machen. Die Formbarkeit oder *Plastizität* unseres Gehirns findet darin ihr Gegengewicht, das dafür sorgt, dass wir die Erfahrung von gestern trotz neuer Erlebnisse auch morgen noch nutzen können. „Automatisierung“ sorge dafür, dass die nicht Multitasking-fähigen Ressourcen im präfrontalen Cortex wieder frei gegeben werden können[35]. Hierbei spielen Wiederholungen sowie Schlaf- bzw. Traumphasen wohl eine wesentliche Rolle.

> „Nocturnal consolidation is (…) not limited to the strengthening of existing knowledge. The discoveries from the day are not only stored, but also recoded in a

[32] Ebd., S. 193.

[33] Ebd., S. 160.

[34] Vgl. ebd., S. 163.

[35] Vgl. ebd., S. 223.

> more abstract and general form. Nighttime neuronal replay undoubtedly has a crucial role in this process. Every night, our floating ideas from the day are reactivated hundreds of times at an accelerated rate, thus multiplying the chances that our cortex eventually discovers a rule that makes sense.“[36]

Was hier wie selbstverständlich formuliert scheint, ist jedoch in Gehirnforschung und Kognitionswissenschaften eine bis heute ungelöste Frage: Wie genau kommt es von Signalen und neuronalen Reizen zu abstrakten Konzepten und symbolischen Regeln? Dehaene:

> „Another frontier of research consists of clarifying how such learning–induced changes, whether synaptic or not, can implement the most elaborate types of learning that the human brain is capable of, based on the ‚language of thought‘ and the fast recombination of existing concepts. (…) there is no truly satisfactory model of how synaptic changes in neural networks underlie language acquisition or mathematical rules. Moving from the domain of synapses to the symbolic rules that we learn in math class remains a challenge today.“[37]

Beim symbolischen Lernen fangen wir Menschen, auch das zeige die Forschung, jedenfalls nicht bei Null an. Wir besitzen offensichtlich einen angeborenen Instinkt zum Spracherwerb[38]. Ebenso scheinen uns grundlegende Vorstellungen von Objekten, physikalischen Gesetzen, psychologischen Mustern und mathematischen Verhältnissen angeboren zu sein[39]. Auch wenn dieses offensichtlich evolutionär für die ganze Spezies erworbene Wissen biologisch noch nicht erklärt werden kann, führten die Ergebnisse zahlreicher Experimente mit Babys zu diesem Schluss. Es ist interessant, einen Blick in die Labore zu werfen, in denen die vorsprachliche Vorstellungswelt von Kleinkindern ergründet werden soll[40]:

> "In today's cognitive science laboratories, experimenters have become magicians (…). In small theaters specially designed for babies, they play all sorts of tricks: on the stage, objects appear, disappear, multiply, pass through walls… (…) By zooming in on the children's eyes – to determine where they look and for how long – cognitive scientists manage to accurately measure their degree of surprise and infer what they expected to see."[41]

Auf diese Art der Experimente komme ich noch zu sprechen.

---

[36] Vgl. ebd., S. 231.

[37] Ebd., S. 96.

[38] Vgl. ebd., S. 67.

[39] Ebd., S. 55.

[40] Vinciane Despret würde sicherlich sagen: fabriziert werden.

[41] Dehaene 2021, S. 54.

Vorab ein vorläufiges Fazit in Bezug auf meine Erkundung der Fähigkeit, sich überraschen zu lassen. Zunächst einmal finden wir uns bei Dehaene bezüglich ihrer Bedeutung für das menschliche Lernen eine klare Bestätigung. Lebenslanges Lernen ist ein Erfordernis in einer veränderlichen Umwelt. Es wäre weder möglich noch wünschenswert, schreibt er, wenn uns alles Wissen „pre-wired" – wie feste „Verdrahtungen" der Hardware – angeboren wäre. Unmöglich, weil die biologischen Bausteine nicht ausreichten, um alles zu codieren, und nicht wünschenswert, weil jedes individuelle Lebewesen eine andere Umwelt, andere Lebensbedingungen vorfindet, also anderes Wissen braucht, um zu überleben[42].

Die empirischen Beobachtungen, die Dehaene schildert, helfen bei der Bildung eines Begriffs der Fähigkeit, sich überraschen zu lassen. Wir haben gesehen, dass sich der Gap zwischen Signal und Symbol, zwischen *surprisal* (einem Signal mit erhöhtem Informationswert) und erlebter Überraschung (*surprise*), in der Sprache neuronaler Prozesse allein nicht erklären lässt. Der Übergang zwischen der neuronalen und der symbolischen und sprachlichen Ebene kann als wichtige Forschungsfrage formuliert werden. Parallel dazu haben wir eine Einführung erhalten in ungelöste Probleme der KI. Und wir haben deutlich vor Augen geführt bekommen, wie stark kognitionswissenschaftliche Grundlagenforschung von der Sprache verfügbarer Technologien gefärbt ist.

Aufmerksam, neugierig oder überrascht?

Dehaenes theatralische Schilderungen kognitionswissenschaftlicher Experimente mit Babys werfen für mich einige Fragen auf. Lässt sich „Überraschtsein" wirklich so messen? Wird mit Eye-Tracking-Verfahren nicht eher Aufmerksamkeit beobachtet? Welches Verhältnis zwischen Aufmerksamkeit und Überraschtsein wird vorausgesetzt? Und dürfen wir wirklich davon ausgehen, dass ein solches Setting und die darin agierenden „magicians" das gesehene Verhalten nicht beeinflussen oder gar erst hervorrufen könnten – gerade weil es sich um Babys, also existenziell von der Nähe zu anderen Menschen abhängige Lebewesen handelt[43]? Ist die implizite Grundannahme gerechtfertigt, dass Babys sich stärker für das

[42] Vgl. ebd., S. XVIIff.

[43] Vgl. hierzu Vinciane Despret's ausgiebige Kritik an psychologischer Laborforschung zu Emotionen in: Despret, Vinciane (2004): Our Emotional Makeup. Ethnopsychology and Selfhood, New York: Other Press. Ich möchte hier in Anlehnung an sie lediglich die Vermutung in den Raum stellen, dass Babys zwar nicht auf einer verbalen Ebene, aber dennoch genauso wie (oder vielleicht sogar stärker als) Erwachsene (oder andere soziale Lebewesen, darüber hat Despret in ihren Arbeiten zur Verhaltensforschung an Tieren geschrieben) versuchen, die Erwartungen des ihnen gegenüber stehenden Menschen zu erraten und dass deshalb immer zu diskutieren ist, inwieweit dies die Ergebnisse solcher Experimente verfälscht.

interessieren, was sie nicht erwarten oder wiedererkennen? In welchem Kontext ist das so? Was ist in dieser Schilderung der Unterschied zwischen Neugier und Überraschung? Und sind visuelle Aufmerksamkeit und Beschäftigungsdauer wirklich ein aussagekräftiges Maß für die Nichtübereinstimmung von Erwartung und Erleben?

Ohne diese Fragen hier in Bezug auf die geschilderten Versuche klären zu können, bleibt festzuhalten, dass Aufmerksamkeit, Neugier und die Fähigkeit, sich überraschen zu lassen, in Dehaenes Text sehr eng beieinander liegen.

Dabei unterscheiden sich Neugier und Überraschung meines Erachtens gerade auf der Ebene der „Metakognition". Neugier beinhaltet ein Bewusstsein übers Noch-Nicht-Wissen und ein Wissen-Wollen, die Erwartung richtet sich also direkt auf die Möglichkeit, Neues zu erfahren[44], während wir, wenn wir überrascht werden, genau damit eben nicht rechnen, sondern uns in einem Bewusstseinszustand befinden, über den wir erst *im Nachhinein sagen werden*, dass wir etwas anderes erwartet hätten, als wir dann tatsächlich erlebt haben.

Die Gegenüberstellung von Neugier und Überraschtsein, zu der Dehaenes Versuchsbeschreibungen mich erst gedrängt haben, die ich also ihm verdanke, führt zu einem weiteren interessanten Punkt. Die „bewusste persönliche Erfahrung", als die Tobin die Überraschung bezeichnet hat, scheint nicht davon abzuhängen, ob uns auf das Ereignis bezogene Erwartungen vor dessen Eintreffen bewusst waren. Es kann auch sein, dass diese durch das Eintreffen des überraschenden Ereignisses erst aktualisiert und dann einer Revision unterzogen werden.

Aufmerksamkeit und Überraschtsein wiederum lassen sich auf den ersten Blick so voneinander abgrenzen: Während Aufmerksamkeit auf etwas gerichtet sein kann, ohne dass dieses Etwas als überraschend empfunden wird, gilt dies umgekehrt nicht: Etwas kann nicht als überraschend empfunden werden, ohne dass es in den Fokus der Aufmerksamkeit gerät. Das, was überrascht, war aber zuvor eben gerade nicht in diesem Fokus. Überraschung geht also mit einer Bewegung der Aufmerksamkeit einher. Die Fähigkeit, sich überraschen zu lassen, muss daher eine Beweglichkeit der Aufmerksamkeit zur Voraussetzung haben. Worauf ich mit Bernhard Waldenfels zurückkommen werde.

[44] Dehaene berichtet über den Bau von neugierigen Robotern („curious robots"). Deren Algorithmus besteht darin, permanent Vorhersagen über den Zustand der (von ihnen auch manipulierbaren) Welt zu machen und zu testen. Sobald sie feststellen, dass ihr Vorhersagen bezüglich bestimmter Manipulationen ausreichend treffsicher sind, suchen sie aktiv nach neuen Lernfeldern (anderen Orten, Gegenständen oder Manipulationsmöglichkeiten). Auf einer Babymatte mit Spielzeug würden sich diese Roboter „exactly like a young child" verhalten. Vgl. Dehaene 2021, S. 191.

Offene Fragen: Maschinelles Lernen zwischen Plastizität und Stabilität

Wenn es stimmt, dass neben der Neugier auch die Fähigkeit, uns überraschen zu lassen, ein wesentlicher Motor des Lernens ist, ließe sich vermuten, dass sie auch im maschinellen Lernen eine Rolle spielt, spätestens dann, wenn maschinell lernende Artefakte (MLA)[45] nicht mit einem abgeschlossenen Datenset oder nicht nur auf eine spezifische Aufgabe hin trainiert werden. Die KI–Forschung spricht im Gegensatz dazu von „kontinuierlichem Lernen“, in einem allgemeineren Sinne auch von „starker KI“ oder „general intelligence“[46].

2019 erschien der seither vielzitierte Artikel "Continual lifelong learning with neural networks: A review" von German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan und Stefan Wermter. Die fünf Autoren stellten zur Ausgangslage fest:

> "lifelong learning capabilities are crucial for computational learning systems and autonomous agents interacting in the real world and processing continuous streams of information. However, lifelong learning remains a long–standing challenge for machine learning and neural network models since the continual acquisition of incrementally available information from non–stationary data distributions *generally leads to catastrophic forgetting or interference*."[47]

MLA auf der Grundlage künstlicher neuronaler Netze (KNN) sind bis heute nicht in der Lage, einmal Gelerntes über ihre gesamte Lebensdauer hinweg durch neue Informationen permanent zu aktualisieren und zu erweitern, ohne dass es zu unerwünschten Interferenzen bis hin zum Phänomen des "katastrophalen Vergessens" kommt. Was für Menschen und Tiere ein Leichtes zu sein scheint, sowohl im Sinne des Umgangs mit neuen oder überraschenden Informationen als auch im Sinne der kreativen Lösung neuer Aufgaben, das ist für MLA noch in weiter Ferne. Das Problem ist in der KI auch als „Plastizitäts-/Stabilitäts-Dilemma“ bekannt, womit ausgedrückt wird, dass Plastizität bislang nur zu Lasten der Stabilität des bereits erreichten Grades an Funktionalität zu haben ist.

Lösungsansätze für das Problem des katastrophalen Vergessens auf der Basis von künstlichen neuronalen Netzen lassen sich laut Parisi et. al. drei Kategorien zuordnen: sie ändern die Modellierung von Plastizität i) auf „Synapsen“–Ebene, ii) durch die Alloziierung

---

[45] Diesen Begriff entnehme ich Harrach, Sebastian (2014): Neugierige Strukturvorschläge im maschinellen Lernen. Eine technikphilosophische Verortung. Bielefeld: transcript.

[46] Für eine prägnante Kurzdarstellung der Geschichte der KI–Forschung verweise ich auf Toosi A, Bottino AG, Saboury B, Siegel E, Rahmim A. (2021): A Brief History of AI. How to Prevent Another Winter (A Critical Review), PET Clin. 2021 Oct;16(4):449-469.
Ein Glossar der wichtigsten in dieser Arbeit auftauchenden technischen Begriffe findet sich außerdem im Anhang.

[47] Parisi, German I., Ronald Kemker, Jose L. Part, Christopher Kanan, und Stefan Wermter. 2019. „Continual Lifelong Learning with Neural Networks: A Review“. *Neural Networks* 113 (Mai): 54–71. https://doi.org/10.1016/j.neunet.2019.01.012. S. 54. Hervorhebung von mir.

neuer Ressourcen für neue Informationen und Aufgaben oder iii) durch den Einsatz komplementärer Systeme, die Wissen konsolidieren und Trainingsdaten bei Bedarf wieder abrufen können (genannt wird das „experience replay“: Erfahrungswiedergabe). Alle drei Richtungen sind an neurowissenschaftliche Erkenntnisse angelehnt: i) an die Beobachtung, dass kürzlich aktivierte Synapsen schneller feuern als andere, ii) an die Beobachtung unterschiedlicher Gehirnaktivitäten in präfrontalem Kortex und Großhirnrinde, und iii) an die lernförderliche Wirkung von Schlaf und Träumen. Bis zur Veröffentlichung des Artikels habe noch keine dieser Ideen durch ausreichend aussagekräftige Versuche überprüft werden können. In jedem Fall hätten insbesondere ii) und iii) gewichtige Nachteile zur Folge, nämlich einen immensen und kaum skalierbaren Hardware- und Energiebedarf.

Für das unüberwachte kontinuierliche Lernen gebe es Lösungsvorschläge in Richtung sich selbst organisierender neuronaler Netze mit dynamischen Architekturanpassungen, die den Autoren zufolge jedoch ebenfalls noch nicht rigoros anhand beliebig wachsender Datenmengen und Aufgaben getestet wurden. Bislang fehlten nicht zuletzt definierte Metriken, mit denen das Auftauchen katastrophalen Vergessens mess- und vergleichbar gemacht werden könnte.

Neuere Forschungsansätze versuchten, schreiben die Autoren, weitere biologische Aspekte menschlichen Lernens aufzugreifen, zum Beispiel durch die Entwicklung unterschiedlicher Trainingspläne für aufeinander aufbauende Entwicklungsstufen, durch autonome Erkundungsbewegungen in der Umwelt, Belohnungssysteme, sensomotorische Anbindungen oder Implementierung von Formen der „Selbstreflexion“.

Übergeordnete Prozesse würden dabei meist in Anlehnung an die Theorie komplementärer Lernsysteme (Complementary Learning Systems, CLS) entworfen, die das Zusammenwirken von zwei Systemen beschreibt, eines für das kurzfristige, adaptionsfähige (schnelle) Auffassen einzelner Informationen, das beim Menschen im Hippocampus verortet wird, und ein zweites für das langsamere episodische Gedächtnis und das Strukturieren von Information, das beim Menschen im Neocortex erfolgt. Das Zusammenspiel dieser beiden Funktionalitäten, so die Autoren, sei offensichtlich der Schlüssel ("crucial") für die Fähigkeit, gleichermaßen statistische Regelmäßigkeiten wie spezifische, „episodische“ Erinnerungen zu erlernen.

Katastrophales Vergessen, so die Autoren, gebe es auch bei Kindern. Dass dies nur selten passiert, wird in der Forschung damit begründet, dass menschliche Erfahrungen sehr häufig miteinander "verwoben" seien ("the kind of experiences we are exposed to are very often interleaved"[48]). Die restriktiven Szenarien, unter denen Computer lernen, offenbarten damit einen gravierenden Nachteil:

[48] Parisi et al. 2019, S. 58.

> "In the case of computational systems, however, additional challenges must be faced due to the limitations of learning in restricted scenarios that typically capture very few components of the processing richness of biological systems"[49].

Über die lange Zeit der Evolution hinweg hätten biologische Systeme, so die Autoren, komplexe neurokognitive Anpassungsmechanismen entwickelt, deren Komplexität von aktuellen künstlichen Systemen nicht annähernd erreicht werde:

> "the differences between biological and artificial systems go beyond architectural differences, and also include the way in which these artificial systems are exposed to external stimuli. (…) Humans and animals make massive use of the spatio-temporal relations and increasingly richer high-order associations of the sensory input to learn and trigger meaningful behavioural responses."[50]

Alle bekannten Ansätze ließen weiterhin viele Fragen ungelöst, es gäbe noch keine gut erforschten Mechanismen, die eine aufgabenspezifische Balance von Plastizität und Stabilität regulieren könnten. Einen Ansatz, der Erwartungsverhalten modelliere, verfolge die "adaptive resonance theory" (ART), in der „bottom-up sensory observations" permanent mit „top-down expectations as memory templates" verglichen werden, um katastrophalem Vergessen entgegenzuwirken. Jedoch seien auch hier die Ergebnisse unzuverlässig und hingen stark davon ab, in welcher Reihenfolge Trainingsdaten zur Verfügung gestellt werden. Dies scheint im Übrigen generell für das Training von KNN zu gelten und macht einen ihrer Schwachpunkte aus.

Hinsichtlich der biologischen Plausibilität der verschiedenen Ansätze räumen die Autoren deutlich ein, dass alle stärker an performativen Erfolgen als an der Exaktheit der Simulation von Theorien ausgerichtet sind, es gehe also zuallererst um funktionale Verbesserungen[51]. Vor diesem Hintergrund scheint es mir umso fragwürdiger, wenn Gehirnforschung nun, wie bei Dehaene gesehen, Metaphern aus den verfügbaren Technologien entlehnt, um menschliches Verhalten zu beschreiben und besser verstehen.

Es hätte den Rahmen der Arbeit gesprengt zu versuchen, einen Überblick über Fachpublikationen zum maschinellen Lernen seit 2019 zu gewinnen. Nachdem aktuelle Technologien marktreif geworden und spätestens mit ChatGPT auch im Consumer–Bereich angekommen sind, boomt die Disziplin. Unternehmen wie Bildungseinrichtungen investieren, und entsprechend sind auf allen Kanälen täglich neue Meldungen und Publikationen zu finden[52]. Dennoch lässt sich der Eindruck gewinnen, dass sich im Zusammenhang mit dem Plastizitäts-/Stabilitäts-Dilemma noch keine entscheidenden

---

[49] Ebd., S. 58.

[50] Ebd., S. 59. Man achte auf die Sprache, die Lernen und Verhalten unterkomplex als Reiz–Reaktionsschema vorstellt.

[51] Ebd.

[52] Zu den Folgen und Risiken solcher Booms, von denen die KI bereits den dritten erlebt, siehe auch Toosi et. al.(2021).

Lösungen abzeichnen. Dies mögen aktuelle Artikel belegen, die neue Vorschläge und Konzepte mit allerdings nach wie vor nur begrenzt aussagefähigen Ergebnissen vorstellen[53]. Um hier ein Beispiel herauszugreifen, nachfolgend die Beschreibung eines im November 2022 erschienenen, und in der Fachpresse mit Aufmerksamkeit versehenen Versuchs, die Stabilität bereits gelernter Inhalte bei einem Folgetraining mit neuen Daten durch die Simulation von Schlafphasen zu verbessern:

> „Interleaving new task training with periods of off–line reactivation, mimicking biological sleep, mitigated catastrophic forgetting by constraining the network synaptic weight state to the previously learned manifold, while allowing the weight configuration to converge towards the intersection of the manifolds representing old and new tasks. The study reveals a possible strategy of synaptic weights dynamics the brain applies during sleep to prevent forgetting and optimize learning.“[54]

Über derartige Ansätze hinaus ist heute im Bereich der Entwicklung genereller oder starker KI zunehmend von einer dritten Welle die Rede[55]. Wer sich von den in der Tat im Vergleich zur „symbolischen KI“[56] fantastischen performativen Höhenflügen der auf riesigen Datenmengen mathematisch operierenden KI in Form von Go–Spielen oder Large Language Models wie ChatGPT zunächst beeindrucken ließ, muss angesichts des Plastizitäts–Stabilitäts–Dilemmas überlegen, ob der „symbolischen KI“ nicht doch eine neue Chance gebührt, vielleicht in Kombination mit den heute erfolgreichen Verfahren maschinellen Lernens. Aus einem an kommerzialisierbaren Anwendungsszenarien interessierten Blickwinkel kommen noch weitere Argumente hinzu, zum Beispiel weil man so eventuell Transparenz– und Sicherheitsproblemen entgegenwirken könnte. In dem Band „Road to General Intelligence“ von 2022 heißt es ähnlich wie bei Dehaene, nur technischer ausgedrückt:

> „Machine learning (ML), notably deep- and reinforcement learning, has emerged as the dominant AI paradigm. (…) Nevertheless, it remains a feat of imagination to ascribe any meaningful notion of intelligence to any of these systems (…) Although machine learning is a valuable *engineering technique*, this fact is not to be confused with a *claim* that it might offer a path toward general intelligence. (…) There is increasing consensus that it is necessary to combine the strengths of both symbolic and connectionist paradigms (…): the main advantage of symbolic approaches is the ready injection of domain knowledge, with the attendant pruning of hypothesis space. In contrast, the main advantage of connectionism is that it is (at least in principle) a *tabula rasa*. (…) As has been argued (…) for many years, we also hold the

---

[53] Ebd.

[54] Golden, Ryan/ Delanois, Jean Erik/ Sanda Pavel/ Bazhenov, Maxim (2022): Sleep prevents catastrophic forgetting in spiking neural networks by forming a joint synaptic weight representation. PLoS Computational Biology 18(11): e1010628.

[55] Zum Beispiel Garcez, A. d'Avila/ Lamb, Luis C. (2020): Neurosymbolic AI: The 3rd Wave. https://arxiv.org/pdf/2012.05876.pdf (19.12.2023). https://doi.org/10.48550/arXiv.2012.05876 (19.12.2023)
Dort steht allerdings die Anforderung der Nachvollziehbarkeit im Vordergrund der Argumentation.

[56] Siehe Glossar technischer Begriffe im Anhang dieser Arbeit.

view that general intelligence requires the recursively algebraic capacities of human reasoning. “[57]

Die Crux der Simulation

Wenn ich in dieser Arbeit den Blick auf die Technik des maschinellen Lernens lenke, dann tue ich das aus zwei Gründen. Zum einen leite ich aus der performativen Differenz zwischen Mensch und MLA einen Teil meiner Motivation für die Beschäftigung mit der Fähigkeit, sich überraschen zu lassen, ab. Zum anderen scheint es mir spannend, den Einfluss der verfügbaren KI-Technologien auf die Theoriebildung in der Gehirnforschung und auch in anderen Bereichen nachzuzeichnen[58]. Maschinelles Lernen interessiert mich also in dieser Arbeit nicht als Anwendung, sondern als Modell und Metapher.

Eine Kritik des Maschinellen Lernens als Werkzeug zur Erforschung kognitiver Fähigkeiten des Menschen kann im Rahmen dieser Arbeit nur andeutungsweise erfolgen. Dennoch soll hier zumindest in der durch den begrenzten Rahmen dieser Arbeit gebotenen Kürze angerissen werden, welche Aspekte dabei eine Rolle spielen könnten. Dazu erweist sich die Lektüre Gabriele Gramelsbergers Analyse von Computerexperimenten als geeigneter Einstieg[59].

Wohlgemerkt: Gramelsberger bezieht sich in ihrer Analyse auf den Wandel derjenigen Wissenschaften, deren Theoriesprache die Mathematik war – insbesondere die Physik. Nicht jede ihrer Schlussfolgerungen wird daher übertragbar sein auf einen Bereich, dessen Beschreibungssprache vor Einführung der Computer in den Forschungsbetrieb keine mathematische war. Insbesondere ihr Punkt, dass mathematisch nicht beweisbar ist, ob mathematische und informatische Modelle identisch sind, kann hier nicht übernommen werden. Ich treffe im Folgenden also eine Auswahl derjenigen ihrer Argumente, von denen ich denke, dass sie auch im Bereich von Kognitions- und Neurowissenschaften eine

[57] Swan, Jerry/ Nivel, Eric/ Kant, Neil/ Hedges, Jules/ Atkinson, Timothy/ Steunebrink, Bas (2022): The Road to General Intelligence. Studies in Computational Intelligence, Cham: Springer Nature, S. 3f.

[58] Das unterscheidet meine Blickrichtung von der Sebastian Harrachs, der MLA als angewandte Technik untersucht (Harrach 2014). „Neugierige Strukturvorschläge“ entstehen auf Aufgabenebene, nicht aber bei der Umsetzung und Erprobung von Theorien über das Lernen oder den Umgang mit Überraschungen. Überraschungen behandelt Harrach lediglich im Blick darauf, ob MLA in der Lage sind, Nutzer*innen zu überraschen. Seine Ausführungen helfen also für meine Fragestellung, die Voraussetzungen und Elemente der Fähigkeit, sich überraschen zu lassen, sucht, leider nicht weiter.

[59] Gramelsberger, Gabriele (2010): Computerexperimente. Zum Wandel der Wissenschaft im Zeitalter des Computers, Bielefeld: transcript.

kritische Betrachtung des „Forschungshandelns im Computerlabor“ anleiten könnten. Gramelsberger schreibt:

> „Welchen neuen Erfahrungsbegriff der Wissenschaft und welches Selbstbild der Gesellschaft sich aus der Vollendung der wissenschaftlichen Revolution durch den Computer ableiten werden, gilt es zu untersuchen. Dazu wird es notwendig sein, näher auf den Computer als Medium sowie auf die (maschinentauglichen) *Algorithmen als neue mathematische Sprache* und den dadurch initiierten Medien- und Sprachwandel in der Wissenschaft einzugehen. Da sich mit dem Medien- und Sprachwandel ein *Wandel der symbolischen Form wissenschaftlicher Forschung* abzeichnet, liegt es nahe, hier den epistemischen Kern des Wandels der Wissenschaft im Zeitalter des Computers zu vermuten.“[60]

Kognitions- und Neurowissenschaften sind, wie in den vorangegangenen Abschnitten sichtbar geworden ist, Beispiele für nicht primär mathematisch ausgelegte Disziplinen, in denen die Sprache der Algorithmen in den Kanon des Forschungshandelns aufgenommen wurde. Dehaene als Kognitionspsychologe und unterwegs im Auftrag der Schulbildung verwendet in seinem Buch durchgängig Beschreibungen, die als Algorithmen verstanden (wenn auch nicht unmittelbar umgesetzt) werden können. Zur nochmaligen Illustration hier seine Beschreibung der vier „Gehirnfunktionen“, die er als wesentlich für das Lernen bezeichnet:

A. Aufmerksamkeit: „amplifies the information we focus on“
B. Neugier: „an algorithm also called „curiosity“, which encourages our brain to ceaselesly test new hypotheses“
C. Fehlerfeedback: „compares our predictions with reality and corrects our models of the world“
D. Konsolidierung: „renders what we have learned fully automated and involves sleep as a key component“[61]

Wenn es aber so ist, dass Algorithmen eine – oder *die* – neue Sprache der Wissenschaft sind, ist eine kritische Betrachtung der Stärken und Schwächen dieser neuen symbolischen Form[62] erforderlich. Algorithmen lassen sich in einem ganz allgemeinen Sinn einfach als Verfahrens- oder Ablaufbeschreibungen definieren. Als solche können sie, müssen aber nicht, in Form von Computerprogrammen umsetzbar sein – Computer-Algorithmen sind also nur eine Unterform von Algorithmen, die sich dadurch auszeichnet, dass der Algorithmus in einer Computersprache ausgedrückt wird. Dabei können die Eigenheiten der jeweils gewählten Computersprache zu unterschiedlichen Möglichkeiten und Begrenzungen führen.

---

[60] Gramelsberger 2010, S. 256f. Hervorhebungen von mir.

[61] Dehaene 2021, S. 145f.

[62] Gramelsberger verwendet diesen Begriff in Anlehnung an Cassirer.

Gramelsberger beschreibt am Beispiel von computerbasierten Klimasimulationen einige bemerkenswerte Eigenschaften von Algorithmen als Wissenschaftssprache. Es scheint mir hilfreich, diese danach zu sortieren, ob es sich um eine allgemeine Eigenschaft von Algorithmen oder um eine spezielle Eigenschaft informatischer, also in Computersprache formulierter Algorithmen handelt.

Eine allgemeine Eigenschaften von Algorithmen ist nach Gramelsberger, dass sie die Zerlegung eines komplexen Problems in „einzelne Anweisungen und Berechnungspunkte" voraussetzen, die

> „in Form einer komplexen Choreographie von Abläufen, Schleifen und Entscheidungspfaden strukturiert werden [müssen]. Eine solche Choreographie übersetzt jedoch die *Simultanität* der Prozesse, die ein Phänomen wie den Zustand der Atmosphäre ausmachen, in nacheinander abarbeitbare Teilprozesse"[63].

Hier sind an mehreren Stellen Annahmen und Setzungen am Werk, die zusätzliche, nicht der Theorie zuzuschreibende Fehlerquellen einführen, deren Auswirkungen umso schwerer zu beurteilen sind, je komplexer das zu Grunde liegende Problem ist.

Computer-Algorithmen erweitern das Spektrum möglicher, theoriefremder Fehler um weitere Eigenheiten. Zum einen müssen sie numerisch expliziert werden. Doch durch die Komplexität und

> „die unendlichen Möglichkeiten der Wechselwirkungen in einem System mit vielen Freiheitsgraden wird die numerische Lösung *sensitiv abhängig* von ihrer numerischen Initialisierung. Geringfügige Änderungen in der Initialisierung können zu vollkommen anderen Resultaten führen und von der eigentlichen Lösung wegführen"[64].

Außerdem führen sie neue Abhängigkeiten ein, nämlich von der Verfügbarkeit und der Zuverlässigkeit passender Hardware. Und schließlich sind Computer- und Fachexpertise nicht immer in einer Person vereint. Dadurch fehlt an möglicherweise entscheidenden Stellen das nötige Fachwissen, um aus inzwischen zahlreich verfügbaren Softwarebibliotheken (fertigen Bausteinen oder Modulen, die bestimmte Funktionalitäten implementieren) die fachlich korrekte und problemadäquate Lösung auszuwählen. Wenn fehlendes Fachwissen jedoch durch pragmatisches Handeln (funktioniert, ist verfügbar, eventuell noch: hat gute Bewertungen) ersetzt wird, können sich über solche Module zusätzliche und später schwer zu findende Fehlerquellen in das Computermodell einschleichen.

Für die Forschung, die sich an Computerexperimenten ausrichtet, bedeute deren Einsatz, so folgert Gramelsberger, nicht nur den Verlust von Materie als Korrektiv für Theorien, sondern auch, dass sie abhängig davon wird, dass Algorithmen keine falschen Setzungen enthalten

---

[63] Gramelsberger 2010, S. 244. Beispiel dort: die Computermodellierung des Klimas.

[64] Ebd., S. 245. Hervorhebung von mir.

und informatorisch korrekt umgesetzt werden. Wenn aber der Erkenntnisgewinn von Computer-Modellen auf der Fehleranalyse, das heißt darauf beruht, dass sie Erklärungslücken aufspüren helfen, sei das Verständnis ihrer besonderen Eigenschaften wesentlich. Anders formuliert:

> „Hier liegt die Crux der Computerexperimente: sie agieren zwar in einem theoretisch überdeterminierten Raum, dieser bietet jedoch keinerlei Korrektiv in seiner experimentellen Anordnung, weder als materiales Korrektiv, noch als logisches basierend auf der Kohärenz des deduktiven Verfahrens. Doch (…) hilft sich hier die computerexperimentelle Forschung praktisch, indem sie den theoretisch überdeterminierten Experimentalraum nicht als Theorieraum, sondern als ein (sic!) Möglichkeitsraum betrachtet und ihn als solchen zum Gegenstand der experimentellen Forschung macht."

Technik im Computerexperiment sei, sagt sie mit Hans-Jörg Rheinberger, nicht im Licht von Zwecken, für die sie entworfen seien, also als „Maschinen, die Antwort geben sollen" zu betrachten, sondern als „epistemisches Objekt", was bedeute, sie ist in diesem Kontext „in erster Linie eine Maschine, die Fragen aufwirft"[65]. Der Möglichkeitsraum dieser Fragen bezieht jedoch all die Einschränkungen auf Grund der oben genannten Eigenschaften von Algorithmen genauso ein wie die fachlichen Vorgaben, die zu ihnen geführt haben[66].

Wird die Computersimulation zur Wissenschaftssprache, so definiert die Struktur der verfügbaren Technik (zum Beispiel MLA) die Menge symbolischer Entitäten, bestimmt also auch die Grenzen dessen, was sagbar ist[67]. Und das ist auch im Hinblick darauf interessant, ob sich diese Grenzen bei Bedarf verflüssigen lassen. Dann bleibt aber die Frage, wie ein solcher Bedarf überhaupt festgestellt wird, wenn das Scheitern der Simulation so viele schwer detektierbare Ursachen jenseits des theoretischen Modells haben kann.

Philosophische Zugänge

Nun kann die Philosophie verschiedene Aufgaben wahrnehmen. Sie kann aufbauend auf philosophischen und insbesondere phänomenologischen Überlegungen Fragen an die Herangehensweisen von Kognitionswissenschaft und KI stellen. Sie kann sich mit den Verflechtungen von Wissenschaft, Technologie und anders situierten Diskursen befassen. Sie kann, wie Isabelle Stengers mit Alfred North Whitehead vorschlägt, die Modi der

---

[65] Ebd., S. 272f.

[66] Umso mehr und auf besondere Weise bezieht sich das auch auf Computersimulationen, die auf Modellen beruhen, die durch Verfahren maschinellen Lernens erzeugt wurden. Wo Deep Learning in der Wissenschaft bereits eingesetzt wird, beschreibt zum Beispiel Baldi, Pierre (2021): Deep Learning in Science. Cambridge, New York et.al.: Cambridge University Press.

[67] Vgl. ebd., S. 233 und S. 275.

Abstraktionen hinterfragen, mit denen Autor*innen und Disziplinen sich einer Thematik nähern, so wie das oben zum Teil bereits geschehen ist. Sie kann Änderungen wissenschaftlicher Paradigmen beschreiben und kritisch begleiten. Sie kann bei all dem eine normative oder eine deskriptive Haltung einnehmen. Sie kann mit Gründen ein Denken jenseits technologischer Anforderungen kultivieren und stark machen und mit ihren Angeboten dafür sorgen, dass das Flussbett veränderlich bleibt und da, wo es droht, zu eng zu werden, zeigen, wie das Feste vielleicht auch wieder verflüssigt werden kann. Sie kann politisieren.

Um ein paar dieser Möglichkeiten auszuloten, sollen in diesem Teil der Arbeit philosophische Beiträge vorgestellt werden, die die bisherigen Betrachtungen der Fähigkeit, sich überraschen zu lassen, um weitere Zugangsmöglichkeiten ergänzt. Beginnen möchte ich mit Catherine Malabou, die seit ihrer Dissertation bei Jacques Derrida am Begriff der *Plastizität* arbeitet, und das zunehmend in Auseinandersetzung mit aktuellen Entwicklungen in der Gehirnforschung. Ihr Ansatz, der sich als materialistisch versteht und der kritischen Theorie nahesteht, versucht, die Brücke zu schlagen zwischen Hegels Dialektik und kognitionswissenschaftlichen Ergebnissen, die zugleich ideologiekritisch hinterfragt werden. Im Licht ihres Textes erscheint die Fähigkeit, sich überraschen zu lassen, als materielle biologische Notwendigkeit. Mit Emmanuel Lévinas folgt dann ein phänomenologischer Beitrag, durch den Überraschung als „Durchstoßen“ der *thematisierenden Intentionalität* des Bewusstseins durch das Ereignis der *Nähe* verstehbar wird, zwei Begriffe, die er in Gegenüberstellung zur Intentionalität bei Husserl entwickelt. An die Erfahrung des Anderen als einer absoluten Alterität bei Lévinas werde ich mit Bernhard Waldenfels anknüpfen, in dessen Phänomenologie der *Aufmerksamkeit* das Verhältnis zwischen dem, was auffällt und der, der es auffällt, noch kleinteiliger untersucht wird. Es bereichert unser Verständnis von der Fähigkeit, sich überraschen zu lassen, um die „Zwischensphäre des Leibes“. Daran anschließend wird die Form unserer Weltbezüge im *thematisierenden Bewusstsein* wieder aufgegriffen, indem aus dem Bereich des Narrativen namentlich *Fiktionen* und Erwartungen aus soziologischer und wirtschaftsphilosophischer Perspektive auf ihren Realitätsgehalt und ihre Funktion hin analysiert werden. *Narrativität* ist laut Christiane Voss auch das, was Emotionen erst als solche erfahrbar macht. Ihre allgemeine Emotionstheorie hilft mir, noch weitere Eigenheiten des Überraschungsgeschehens zu verstehen. Zuletzt wird eine Spur aufgenommen, die im Verlauf der Lektüre immer stärker sichtbar geworden ist: Denken ist nicht zu denken ohne die Einbindung in soziale Gefüge. Mit Isabelle Stengers soll daher schließlich eine Wissenschaftsphilosophin zu Wort kommen, die in ihrem letzten Buch unter Rückgriff auf Whitehead mögliche Formen gemeinschaftlicher Sinngebung vorstellt und von der sich vielleicht lernen lässt, wie sich dadurch auch die Fähigkeit, sich überraschen zu lassen, kultivieren ließe.

Plastizität

Wie kommen wir von der Hirnforschung zurück zur Philosophie? Hat die Philosophie Antworten auf die Frage, wie sich der Übergang vom Neuronalen zum Symbolischen denken lässt? Catherine Malabou hat sich seit ihrer Dissertation bei Jacques Derrida zu Hegels Begriff der Plastizität immer wieder mit dieser Frage befasst. Hegel ist für sie

> „der erste Philosoph, der aus dem Wort ‚Plastizität' einen Begriff gemacht und eine Theorie über das Verhältnis von Natur und Geist entwickelt hat, das in seinem Wesen konflikthaft und widersprüchlich ist. (…) Hegel konnte sich zwar noch nicht in der Sprache des ‚Neuronalen' und des ‚Mentalen' ausdrücken, aber das ändert nichts daran, dass die Transformation der natürlichen Existenz des Geistes (das Gehirn, das er noch als ‚natürliche Seele' bezeichnet) in sein geschichtliches und spekulatives Dasein Gegenstand seiner ständigen Bemühungen war. Und diese Transformation ist die Dialektik selber."[68]

In ihrem Buch „Was tun mit unserem Gehirn?" entwirft Malabou einen dialektischen Blick auf den „Übergang vom Biologischen zum Geistigen". Unter Berücksichtigung neurobiologischer Erkenntnisse und auf Grund der neuronalen Plastizität sei, so schreibt sie dort, nicht von einer Identität von Biologie und Geist auszugehen[69], der Übergang lasse sich aber als Widerständigkeit der ‚Natur' gegen ihre Determiniertheit auffassen, die das Denken aus sich selbst heraus dialektisch hervorbringe:

> „Ein vernunftbegründeter Materialismus scheint uns der zu sein, der davon ausgeht, dass das Natürliche sich selbst widerspricht und dass das Denken die Frucht dieses Widerspruchs ist. Eine der stichhaltigen Arten, das *mind-body-problem* zu betrachten, besteht darin, die dialektische Spannung zu erfassen, die Naturalität und Intentionalität verbindet und zugleich entgegensetzt, und sich für sie als lebendigen Mittelpunkt einer komplexen Realität zu interessieren. Die Plastizität könnte, philosophisch gewendet, genau der Name dieses Dazwischen sein."[70]

Um ihr Denken besser zu verstehen, ist es hilfreich, sich genauer anzuschauen, wie sie den Begriff der Plastizität entwickelt. Zunächst einmal stellt sie fest, dass Plastizität schöpferische wie zerstörerische Eigenschaften umfasst. Die wissenschaftliche Erforschung des Gehirns kenne drei Ebenen der Plastizität. Auf Zellebene entwickelt sich das Gehirn noch lange nach unserer Geburt weiter, nicht nur hinsichtlich seiner Größe, sondern auch hinsichtlich seiner Struktur, wofür zunächst gleichermaßen Gene wie der Vorgang der Apoptose („Zelltod") und später zunehmend von außen kommende Reize und Erfahrungen ursächlich sind. Diese Prozesse bezeichnet sie als epigenetische „Bildhauerarbeit", die

---

[68] Malabou, Catherine (2006/2021): Was tun mit unserem Gehirn? Zürich: Diaphanes, S. 122.

[69] Diese Aussage hat sie einige Jahre später unter dem Eindruck des umstrittenen „Blue Brain"-Projekts zurückgenommen, ihre frühere Theorie scheint mir jedoch die für meine Arbeit interessanteren Fragen zu stellen. Vgl. Malabou, Catherine (2019): Morphing Intelligence. From IQ Measurement to Artificial Brains, New York: Columbia University Press.

[70] Malabou 2006/2021, S. 124.

zunehmend beginne, zu „improvisieren“ und ihr Werk durch eigene Aktivitäten zu prägen[71]. Die zweite Ebene ist die der synaptischen Plastizität, auf der durchgehend Umwelteinflüsse zur Geltung kommen. Schließlich sei das Gehirn auch in der Lage, Verletzungen bestimmter Regionen durch die Verlagerung von Funktionen in andere Areale auszugleichen – diese Ebene nennt sie „Plastizität der Wiederherstellung“.

Das Selbst des Bewusstseins sei genau diese schöpferische wie zerstörende Plastizität des Gehirns:

> „Das Selbst ist eine Synthese aller plastischen Prozesse, die im Gehirn im Gange sind. Es ermöglicht, die Kartographie der (…) Netze zusammenzuhalten und zu vereinigen.“[72]

Das Selbst sei die Antwort auf das Problem, dass die vielen Teilstrukturen des Gehirns und des Organismus zusammen- und nicht gegeneinander arbeiten müssen, und löse die Frage, wie aus den inneren Widersprüchen einer Vielzahl von Strukturen etwas Gemeinsames entstehen könne. Unter Berufung auf auf die Arbeiten von Antonio R. Damasio und Joseph LeDoux meint sie, eine Kohärenz neuronaler Netze sei aus neurowissenschaftlicher Sicht nur auf Basis einer Selbst-Repräsentation des Organismus zu haben. Diese sei eine bewusstseinsunabhängige Notwendigkeit, damit der Organismus überhaupt als Einheit überlebe. Die notwendige permanente innere Abstimmung beruhe also auf dem Vermögen der inneren Selbst-Repräsentation des Organismus und bilde die in neuronalen Netzen organisch verfasste Bedingung der Möglichkeit für jede weitere Verweisstruktur und damit auch für das Symbolische. In diesem organischen „Proto-Selbst“ sei der Weg zum bewussten Selbst – zum Geistigen, zur Psyche, zum Mentalen – bereits angelegt. Die „Herstellung der Beziehung zum Objekt“ erfordere dann, das übernimmt sie ebenfalls von dem Neurowissenschaftler Antonio Damasio, die Entstehung von Repräsentationen zweiter Ordnung und schließlich von Zeichen[73]. So gesehen ist die Plastizität des Gehirns der gesuchte Ausgangspunkt für den Übergang vom Organischen zum Symbolischen:

> „Wie man sieht, wird der Übergang vom Neuronalen zum Mentalen deshalb für gewiss gehalten, weil es im Grunde unmöglich ist, die beiden Bereiche ganz streng und absolut zu unterscheiden. Wenn es im Gehirn eine Art von unterirdischer Aktivität der Repräsentation gibt, so bedeutet das, dass die Neuronen durch ihr ‚in Verbindung stehen‘ bereits für den Sinn verfügbar und vorhanden sind. Genauso sind der Sinn und die symbolische Aktivität im allgemeinen von der neuronalen Konnektivität abhängig.“[74]

---

[71] Vgl. ebd., S. 35.

[72] Ebd., S. 89.

[73] Vgl. ebd., S. 90ff.

[74] Ebd., S. 94.

In einem späteren Interview betont sie, dass die Plastizität des Gehirns ihren Begriff der Plastizität zugleich verkörpere und deutlich mache. Die Ergebnisse der Hirnforschung müssten in ihren Augen dazu führen, die scharfe Unterscheidung zwischen Empirie und Transzendenz aufzugeben:

> „My concept of plasticity does not act as a transcendental concept that would be examplified by many empirical occurences. Because it is itself plastic, that is, exposed to change both by external influences and internal modifications, plasticity is in a certain sense the exemplification of itself. Therefore, it cuts through the divide between the transcendental and the empirical.“[75]

Die Plastizität des Gehirns (und des Denkens) auf Grund innerer Prozesse genauso wie auf Grund äußerer Einflüsse, *ist* für sie das Ununterscheidbare, das Kant nicht vorsieht. Es sei Zeit, dass die epigenetische Wende der Biologie auch für die kritische Theorie und die Philosophie insgesamt fruchtbar gemacht werde:

> „One of the main tasks for critical theory and continental philosophy today is, I believe, to inscribe within their own fields, the resources provided by current cellular, molecular and neurobiology. We are witnessing the birth of the epigenetic paradigm, which, again, is not pregnant only in biology but is also an invaluable resource for the humanities.“[76]

Plastizität und die Verankerung des Denkens in organischen wie in historischen Bezügen beinhalte mit dem Moment der Entstehung des Bewusstseins und der Reflexionsfähigkeit auch die Möglichkeit zur Selbstveränderung und zur Veränderung der äußeren Verhältnisse. Nicht nur deshalb betont Malabou: „jede Betrachtung des Gehirns ist zwangsläufig politisch.“[77] Plastizität werde – im wissenschaftlichen wie im ökonomischen und sozialen Diskurs – schnell auf Anpassungsfähigkeit und Flexibilität reduziert, aus ideologischen Gründen. Sie enthalte jedoch den Sprengstoff, der der Flexibilität fehle:

> „Was soll man tun, damit das Bewusstsein des Gehirns nicht schlicht und einfach mit dem Geist des Kapitalismus zusammenfällt? Wir formulieren dazu die folgende These: Heute wird die Plastizität in ihrer wahren Bedeutung verdunkelt, und man neigt dazu, sie immer wieder durch ihre falsche Freundin, die Flexibilität, zu ersetzen. Der Unterschied zwischen den beiden Begriffen scheint unbedeutend zu sein. Dennoch, die Flexibilität ist die ideologische Gestalt der Plastizität. Sie ist zugleich ihre Maske, ihre Entstellung und ihre Enteignung. (…) Der Flexibilität fehlt (…) die Ressource der Formgebung, also das Vermögen, etwas schaffen, erfinden oder sogar eine Prägung übertreffen zu können (…)“ [78]

---

[75] Malabou, Catherine (2022): Plasticity. The Promise of Explosion. Edinburgh: Edinburgh University Press, S. 310. Wir finden hier ganz nebenbei auch eine Reminiszenz an die gesuchte Verflüssigung von Weltbildern, da sie an dieser Stelle davon spricht, dass die Trennung zwischen Empirie und Transzendenz *verflüssigt*, wenn nicht gar aufgelöst werden müsse („needs to be fluidified, if not erased“).

[76] Ebd., S. 164.

[77] Malabou 2006/2021, S. 81.

[78] Ebd., S. 23. Es belegt dies mit einem Vergleich populärer Diskurse über das Gehirn mit dem von Boltanski und Chiapello aufgespürten „Neuen Geist des Kapitalismus“ in *Boltanski, Luc/Chiapello, Ève (2006): Der neue Geist des Kapitalismus. Konstanz: UVK Verlagsgesellschaft.*

So spannend diese Theorie auch ist, so viel sie auch hergeben mag in der Auseinandersetzung mit und als Überleitung von der Hirnforschung zurück in die Philosophie, lässt sie im Zusammenhang mit meinem Thema viele Fragen offen. Wie genau wäre der Übergang von der Selbst-Repräsentation des Organismus zur symbolischen Form der Sprache zu erklären? Welche Form hat dieses „Mentale“, das, wie Malabou meint, aus dem Widerstand des Biologischen gegen seine Festlegung geboren wurde? Wie sind wir selbst und die Welt uns darin gegeben? Wie kommt es dazu, dass manche Dinge uns auffallen, überraschen und verändern können und andere nicht?

In den folgenden Abschnitten werden diese Fragen aufgegriffen und anhand ausgewählter Beiträge aus anderen Bereichen der Philosophie, insbesondere der Phänomenologie, weiter untersucht.

Berührung und thematisierendes Bewusstsein

In dem Aufsatz „Sprache und Nähe“ untersucht Emmanuel Lévinas die Art und Weise, wie uns die Welt in unserem sinnlichen Empfinden und sinnhaft in der Sprache gegeben ist. Wir haben, so die Hauptthese des Textes, Nähe zum Gegebenen auf zwei Weisen: in einer vorsprachlichen Unmittelbarkeit („Ur–Impression“) und in Form einer sprachlich verfassten Erfahrung. Lévinas nennt jene sprachlich verfasste Weise Bericht[79] oder *thematisierende* Intentionalität, ich möchte sie im Zusammenhang mit dieser Arbeit, und um den Bogen zurück zum Anfang und zu den narratologischen Studien Tobins und Curries zu schlagen, hier gerne auch als *erzählend* bezeichnen. Diesen Vorschlag stützt die folgende Passage, in der er den Begriff der *thematisierenden* Intentionalität einführt. Lévinas schreibt:

> „Das Sein erscheint im Ausgang von einem Thema. (…) Die Wörter entstehen (…) nicht aus der beschränkten und vergeblichen Absicht, an die Stelle von Dingen Zeichen sowie an die Stelle von Zeichen Zeichen zu setzen. Vielmehr sind die Einrichtung und der Gebrauch verbaler Zeichen getragen von einer erzählerischen und thematisierenden Intentionalität, die zu den Seienden gelangt.“[80]

[79] Der Übersetzer, Wolfgang Nikolaus Krewani, schreibt: „‚Bericht’ übersetzt das französische Wort ‚récit‘, das vom lateinischen ‚re-citare‘ kommt. Der Ausdruck ‚Bericht‘ wurde gewählt, weil er durch seinen Zusammenhang mit ‚richtig‘ und ‚einrichten‘ die thematisierende Intentionalität ausdrückt“. Vgl. Lévinas, Emmanuel (2017): Die Spur des Anderen. Untersuchungen zur Phänomenologie und Sozialphilosophie, übers. u. hrsg. v. W. N. Krewani, 7. Auflage, Freiburg (Breisgau), München: Karl Alber, S. 261.

[80] Lévinas 2017, S. 261f.

Ereignisse, Zeit, Reihen, Akte und Zustände, die als Phänomene in uns auftauchen, erhalten, sagt Lévinas, erst in der thematisierenden Intentionalität einen einheitlichen, das heißt dem Auftauchen dieser Phänomene gemeinsamen Sinn.

> „Zeichen, die kraft ihrer Stelle in einem System und kraft des Abstandes zu anderen Zeichen eine Bedeutung haben (…) vermögen der zeitlichen Zerstreuung der Ereignisse und der Gedanken eine einheitliche Bedeutung zu verleihen, vermögen sie in der unauflöslichen Gleichzeitigkeit der Fabel zu synchronisieren. (…) Die Synopse liegt an der Einheit des Themas, das seine Identität durch die Erzählung gewinnt; oder genauer: die Zusammenschau geschieht als Auftauchen des Themas und als Rückbezug aller unthematischen, untheoretischen und sogar ‚noch unsagbaren' Erscheinung auf das Thema."[81]

Identifikation und Thematisierung als notwendige Methoden erzählender Intentionalität sind dabei nicht mit dem Logos der Vernunft oder der Wahrheit gleichzusetzen, sondern zunächst reines Meinen:

> „… die erzählerische – und infolgedessen verbale, linguistische – Intentionalität ist für das Denken, sofern es Thematisierung und Identifikation ist, wesentlich. Denn die Identifikation des Gegebenen in der Erfahrung ist in der Tat reines Meinen. Sie ist nicht Schau oder geläuterte Erfahrung. Sie besteht nicht darin, ein *Dieses* oder ein *Jenes* wahrzunehmen, sondern dieses *als* dieses und jenes *als* jenes zu ‚verstehen' (…)"[82]

Was wir „verstehen" nennen, kann sich demnach nicht auf ein Etwas oder ein Sein jenseits unseres Bewusstseins beziehen, sondern nur auf einen Sinn, den wir ihm erst geben.

> „Das Verstehen des Etwas als Etwas versteht nicht den Gegenstand, sondern seinen Sinn. Das Sein hat den Sinn weder zu erfüllen noch zu enttäuschen. Der Sinn, der weder gegeben noch nicht-gegeben ist, wird verstanden. Allerdings zeigt sich ein Seiendes als Seiendes von seinem Sinn her."[83]

Es reiche also nicht aus, mit Husserl zu sagen, dass Bewusstsein Bewusstsein von etwas ist, vielmehr sei in der Intentionalität des Bewusstseins immer schon, durch das Verständnis von etwas *als* etwas, Denken, Verstehen und Meinen enthalten[84]. Lévinas nennt dies die *kerygmatische* Funktion des Wortes, welche er als Apriori des Wissens bezeichnet[85]. Dabei zerstöre der "proklamatorische Charakter der Identifikation"[86] die Vorstellung einer eindeutigen Beziehung zwischen dem Anderen und dessen Erscheinung im Bewusstsein.

---

[81] Ebd., S. 261.

[82] Ebd., S. 262f, Kursiv im Original.

[83] Ebd., S. 263.

[84] Ebd., S. 264.

[85] Vgl. ebd., S. 265.

[86] Ebd., S. 266.

Wir denken in Sprache, und damit urteilen wir bereits in vielfacher Hinsicht. Gedanken, da sprachlich, sind Urteile, und das Urteil ist der Sinn der Sprache – nicht nur eine mögliche Form von ihr.

> "Weil das Sagen Prädikation ist, ist das Denken Urteilen: Nicht weil die Sprache sich wunderbarerweise dem Urteil als dem ursprünglichen Denken anpassen würde, sondern weil das Urteil den Sinn der Sprache entfaltet. Die Sprache bedeutet nicht, weil sie von ich weiß nicht welchem Spiel sinnloser Zeichen herrührte; sie bedeutet, weil sie die kerygmatische Verkündigung ist, die dieses als jenes identifiziert.“[87]

Dem gegenüber steht für Lévinas "das unthematische Bewußtsein, das sich als Zeit vollzieht" und das er auch als Bewusstsein "ohne aktives Subjekt" bezeichnet, dem die "Polarisation Subjekt-Objekt, die Initiative und die Intention eines Subjekts, das sich ein Thema vornimmt, fremd"[88] seien.

Doch auch in diesem unthematischen Bewusstseinsmodus finden Identifikationen statt, und zwar in Bezug auf die zeitliche Erfahrung der Gegenwart. Denn Gegenwart und Vorstellung treffen niemals zusammen. "In der Vorstellung, in der Repräsentation, ist die Gegenwart bereits vergangen“[89], schreibt er. Oder an anderer Stelle: Die „Unmittelbarkeit“ als „die den Besessenen be-sitzende Nähe des Nächsten“ überspringe das Bewusstsein[90]. Und da die Intuition Erschlossenheit für das Bewusstsein ist, sei sie von dem Intendierten immer schon "durch eine ‚Reflexionszeit' getrennt“[91].

Bewusstsein erwache, wenn wir dem ursprünglichen Strömen des Fühlens das gerade noch Gefühlte entreißen und verstehen wollten, entstehe mit der Unmöglichkeit des Denkens von „Jetzt“. Erst die Sprache erlaube das.

> "Auf diese Weise entdecken wir die Stellung, die die Sprache im Denken von der ersten Gebärde der Identifikation (…) an (…) einnimmt. Indem sie verfolgt, was durch das ursprüngliche Strömen der Zeit schon entkommen ist, wird die Identifikation durch eine dem Bewußtsein kon-substanziale Rede getragen. Demnach hätten die Rede und die Universalität ihren Geburtsort in der Trennung des Fühlens und des Gefühlten, in der das Bewußtsein erwacht."[92]

Wie hängen nun aber Sprache, das thematisierende Bewusstsein und das Sinnliche zusammen?

---

[87] Ebd., S. 268.

[88] Vgl. ebd., S. 271.

[89] Ebd., S. 283.

[90] Ebd., S. 281.

[91] Ebd.

[92] Ebd., S. 270.

Das Sinnliche ist für Lévinas mehr als bloße Information. Die Reduzierung auf einen Informationsgehalt hat ihm zufolge bereits einen "privativen Charakter", sie mache "so etwas wie" den "Verweis auf den Ursprung im Anderen", der sich "als apriorische Struktur des Sinnlichen aufzudrängen scheint" unsichtbar, und damit den *Kontakt*, "der nicht in noetisch–noematische Strukturen umgemünzt werden kann und der schon das Worin für alle Übertragung von Botschaften ist", die "ursprüngliche Sprache, Sprache ohne Worte und Sätze, reine Kommunikation“[93]. In dem Versuch, die Nähe des Anderen als Berührung zu denken, erweitert Lévinas die Phänomenologie Husserls um eine Dimension, die er „ethisch“ nennt.

> "Das Subjekt ist in die Erschlossenheit der Intentionalität und der Sicht hineingegangen. Die Orientierung des Subjekts auf das Objekt hat sich in Nähe verwandelt, das Intentionale ist Ethik geworden (ohne an dieser Stelle etwas Moralisches anzuzeigen).“[94]

Ethisch, so erklärt er in einer Fußnote, bezeichne hier "eine Beziehung zwischen Termini, in der der eine und der andere weder durch eine Verstandessynthese noch durch die Beziehung von Subjekt und Objekt vereint sind, und in der dennoch der eine für den anderen Gewicht hat, ihm wichtig ist, ihm bedeutet, in der sie durch eine Intrige verknüpft sind, die das Wissen weder auszuschöpfen noch zu entwirren vermöchte.“[95]

Das Ethische bedeute also eine "Umwendung der Subjektivität": an die Stelle einer "Subjektivität, die *offen* ist *für* die Seienden, die sich die Seienden immer irgendwie vorstellt, sie setzt" trete eine Subjektivität, "die mit einer Singularität in Berührung kommt, mit einer absoluten und als solche unvorstellbaren Singularität, welche die Thematisierung und die Vorstellung ausschließt.“[96]. Dies sei „die ursprüngliche Sprache, das Fundament der anderen“[97], schreibt Lévinas.

Lévinas spricht von dem "Unmittelbaren der Berührung" ausdrücklich nicht in einem räumlichen Sinne. Nähe ist für ihn nicht äußere Voraussetzung der Begegnung, sondern eine besondere Weise der Bedeutung "*durch sich selbst*":

> "Die eigene Bedeutung der Geschmacksempfindung besteht in gewisser Weise darin, die aufgesammelte Erkenntnis zu 'durchstoßen', um gewissermaßen in das Innerste der Dinge durchzudringen. Nichts ähnelt hier der Deckung des Intendierten mit dem Gegebenen, wie es der Husserlsche Begriff der Erfüllung verlangen würde

[93] Ebd., S. 280.

[94] Ebd., S. 274.

[95] Ebd.

[96] Ebd., S. 275.

[97] Ebd.

> (…) In der Empfindung *passiert* etwas zwischen dem Empfindenden und dem Empfundenen, durchaus unterhalb (…) des Bewußtseins für das Phänomen“[98].

Zwischen Berührung und Wissen vermittle die Sprache in ihrer kerygmatischen Funktion. Sie beruht darauf, dass wir die Nähe zum Anderen erlauben und begehren, ohne diesen Anderen je vollständig fassen zu können – *l’infini*, das Unendliche, kann hier auch als niemals abgeschlossen und nicht erreichbar gelesen werden[99]. Das Fundament der Sprache kennt noch keine Identifikation oder Thematisierung, drängt uns aber dazu, weil wir das andere begehren als Dasjenige, unter dessen Getrenntheit wir leiden, dessen wir uns aber gewahr werden über das Sinnliche.

Mit diesen Begriffen ist es uns nun möglich, die Vorgänge des Überraschtwerdens phänomenologisch besser zu fassen. Wir bekommen durch Lévinas Begriffe an die Hand, die sehr gut zu den verschiedenen Aspekten passen, denen wir in dieser Arbeit bereits begegnet sind. Die von Tobin geschilderten kognitiven Verzerrungen wären als Heuristiken im Bereich des identifizierenden Meinens beschreibbar, auf das wir angewiesen sind, wenn (oder da) wir von der Berührung zum verstehenden Sinn kommen wollen. Der Antrieb zu lernen, von dem wir bei Dehaene gelesen haben, das er insbesondere durch Überraschungen ausgelöst wird, lässt sich in diesen Begriffen deuten als das Drängen der ersten oder der ursprünglichen Sprache, die mit der Singularität des Anderen in Berührung gekommen ist. Was uns überrascht, affiziert uns durch die Berührung und durchstößt das thematische Bewusstsein. In der Folge findet eine Bewegung innerhalb der erzählenden, urteilenden, immer nur meinenden Intentionalität statt und ersetzt dabei eine Weise der Thematisierung durch eine andere.

Die Fähigkeit, sich überraschen zu lassen, erscheint in dieser Begriffswelt als die Fähigkeit, sich vom Anderen berühren zu lassen, diese Berührung als ein Durchstoßen des thematisierenden Bewusstseins zu erleben und daraufhin innerhalb desselben eine Bewegung zu vollziehen.

Durch das Herausstellen des erzählenden Charakters der Intentionalität lassen sich hieran auch die Beobachtungen aus dem narratologischen Zugang anschließen. Denn wenn wir Lévinas darin folgen, dass wir im Verstehen von Etwas nicht einen Gegenstand, sondern einen Sinn verstehen, und wenn dieser mehr ist als ein Zeichen, nämlich ein Urteil, dann widerspricht das jeder Vorstellung eines Denkens und Erwartens, die ausschließlich auf Repräsentationen beruht. Derartige Vorstellungen gelten allerdings, wie wir oben gesehen

---

[98] Ebd., S. 277f.

[99] So schreibt Wolfgang Krewani in der Einleitung zu der von ihm übersetzten Aufsatzsammlung von Emmanuel Lévinas: „Das Andere ist Bedingung des Subjekts in der Weise eines Telos. Aber das Begehren ist doch insofern keine Finalität, als das Andere kein erreichbares Ziel (Finis, frz. fin) ist, sondern infinit, unendlich. (…) Das Begehren selbst ist die Idee des Unendlichen in uns“. Ebd., S. 24.

haben, heute auch in der Kognitionswissenschaft kaum noch als ausreichend. Heuristiken und Interpretationsbewegungen des Überraschungsgeschehens, wie Tobin sie beschrieben hat, lassen sich hier dagegen mühelos integrieren. Genau wie eine narrativ geformte Auffassung von Zeitlichkeit, deren Verhältnis zur Überraschung Currie ausgearbeitet hat.

Aufmerksamkeit

Wie geschieht es eigentlich, dass uns aus der großen Menge sinnlicher Eindrücke genau Etwas auffällt und affiziert? Lässt sich die Rolle dieses Anderen über das Aufmerksamkeitsgeschehen noch genauer beschreiben? Hier erweist sich die Lektüre von Bernhard Waldenfels' Phänomenologie der Aufmerksamkeit auch im Hinblick auf die Überraschung als besonders ergiebig, denn für ihn, ähnlich wie für Lévinas, steht dasjenige im Mittelpunkt der Betrachtung, das uns auffällt – oder überrascht.

> "Das Fremde, das uns in der Erfahrung überrascht, wird bis heute gleichgesetzt mit dem, was wir noch nicht kennen, aber unter geeigneten Bedingungen kennenlernen können. (…) Was uns dagegen vorschwebt, ist eine starke Form der Erfahrung, die im Zuge der Aufmerksamkeit von dem ausgeht, was uns auffällt oder einfällt; in einer solchen Erfahrung verändert sich die Welt, und auch wir selbst verändern uns."[100]

Waldenfels führt so das Überraschtwerden als Teil eines transformativen Prozesses ein, an dem etwas Ich-Fremdes beteiligt ist, und er stellt in Aussicht, dabei auf eine „starke Form der Erfahrung" zu stoßen. Was ist damit gemeint?

Die dem Ereignis nachgelagerten Setzungen eines erzählenden (thematisierenden) Bewusstseins werden erst dadurch erforderlich, dass wir uns in einem Umfeld des Redens und Handelns bewegen. Sie können zwar dem Verstricktsein in eine unerzählbar komplexe leibhafte Welterfahrung ohne Anfänge und Enden niemals gerecht werden, doch ohne sie könnten wir keine einzelnen Ereignisse ausmachen und schon gar nicht über sie sprechen.

> „Erzählmuster und Erzählstrukturen (…) setzen bei Rede- und Handlungsfeldern an, nicht bei isolierten Daten. Damit gewinnt das Reden und Handeln eine Welt- und Leibhaftigkeit sowie seine genuine Geselligkeit zurück. (…) Erzählbare Zusammenhänge, innerhalb derer eines in das andere verflochten ist, bilden ein Zwischenreich, das sich der schlichten Alternative von Mikro- und Makrobetrachtung entzieht."[101]

[100] Waldenfels, Bernhard (2016): Geweckte und gelenkte Aufmerksamkeit. In: Müller, Jörn (Hg.)/ Nießeler, Andreas (Hg.)/ Rauh, Andreas: Aufmerksamkeit. Neue humanwissenschaftliche Perspektiven. Bielefeld: transcript,, S. 27.

[101] Waldenfels, Bernhard (2004): Phänomenologie der Aufmerksamkeit. Frankfurt am Main: Suhrkamp, S. 49.

Eine Erzählung beschreibt nicht ein Ereignis, sondern „bezieht sich auf eine Erfahrung, die erst im Erzählen und Wiedererzählen Gestalt gewinnt.“[102] Dabei entsteht der Ereigniszusammenhang in einem „Zwischenreich“, in dem er „weder auf die Summierung individueller Eigenleistungen zurückgeführt werden kann noch auf eine Einheitsinstanz, die Gemeinsamkeit garantiert“[103].

Zwischen Widerfahrnis und Erzählung gebe es, betont auch Waldenfels, keine repräsentierende „*Verknüpfung“*, sondern allenfalls eine offene Form der

> „*Anknüpfung*. Letztere besagt, daß jemand, dem etwas widerfährt, auf anderes antwortet, ohne daß diese Antwort in dem, was ist, und in dem, was sein soll, ihren zureichenden Grund findet.“[104].

Da wir aber Widerfahrnisse als Erfahrung erinnern, erfordern sie die Fähigkeit, sie zu erzählen. Widerfahrnisse, die wir nicht erinnern, können dennoch Wirkung entfalten und sind in der Psychoanalyse anhand jener aufspürbar:

> „Widerfahrnisse und mithin alles, was uns auffällt und anspricht, können *als Ereignisse* nicht erinnert werden. (…) Wir erinnern uns an das, *was* wir geantwortet haben; an das *wovon* wir getroffen wurden und *worauf* wir geantwortet haben, erinnern wir uns nur indirekt, sofern es nämlich im Gesagten und Getanen seine Spuren hinterlassen hat. (…) In der frühkindlichen Erfahrung und auch bei traumatischen Erfahrungen verschiebt sich das Gewicht. Wir haben es dann nicht nur mit Unerinnerbarem in der Erinnerung zu tun, das dem Unerwartbaren im Erwarten entspricht, sondern mit Unerinnerbarem, das sich allein in seinen Wirkungen kundtut.“[105]

Als Argument gegen eine lineare und rein kognitive Auffassung des Aufmerksamkeitsgeschehens führt Waldenfels den schon bei Lévinas gesehenen pathischen Aspekt der Erfahrung an, der eng mit der affektiven Dimension unserer leiblichen Existenz verbunden ist:

> „Das Auffallen und das Aufmerken könnte primär kognitiv verstanden werden, so daß affektive und praktische Momente nur als Zusatzfaktoren eine Rolle spielen. Doch diese Annahme erweist sich als gegenstandslos, wenn wir von einer pathischen Erfahrung ausgehen, die gerade durch das Getroffensein gekennzeichnet ist. Hier gibt es kein Auffallen, ohne daß uns etwas af-fiziert und an-geht. Nicht das Pathische bedarf der zusätzlichen Erklärung, sondern die Apathie, also der Schwund der Gefühle und deren künstliche Zurückdrängung.“[106]

---

[102] Ebd., S. 50.

[103] Ebd., S. 43.

[104] Ebd.

[105] Ebd., S. 93.

[106] Ebd., S. 71.

Das Gehirn ist wie die sinnliche Wahrnehmung Teil der leiblichen Existenz des Menschen. Ihm jedoch auf „fundamentalneurologische oder pantechnische“[107] Weise die Alleinherrschaft über die Erfahrung zu überlassen, werde insbesondere durch „Risse und Spalten“ verboten, die sich auch im Aufmerksamkeitsgeschehen zeigen.

> „Das Gehirn ist weder ein Zusatzding noch ein Zwischending, sondern es gehört zur Zwischensphäre des Leibes, und wie der Leib, so ist auch das Gehirn *auf gewisse Weise* alles, aber eben nur auf gewisse Weise.
> Daß dies nur auf gewisse Weise so ist, rührt her von winzigen Rissen und Spalten, die eine Totalisierung ausschließen. Das Aufmerksamkeitsgeschehen zeigt einen solchen Spalt in der zeitlichen Verschiebung von Auffallen und Aufmerken, die uns von Anfang an veranlaßt hat, Doppelereignisse und Zwischenereignisse anzusetzen. Dort, wo die Gehirnforschung auf entsprechende Zeitprobleme stößt, sieht sie sich selbst genötigt, den Gebrauch des linearen Zeitschemas, das sich selbst in der Physik als unzulänglich erwiesen hat, einzuschränken, um dem Zeiterleben seinen gebührenden Platz einzuräumen.“[108]

Ein schon von Husserl vorgebrachtes Beispiel dafür ist das Erleben von Musik. Bei Husserl sind es Melodien, Waldenfels argumentiert mit Tonintensitäten, um zu veranschaulichen, dass „unsere gegenwärtige Erfahrung von Gedächtnismomenten durchsetzt“[109] ist, und meint:

> „Solche Verzögerungen im Herzen der Erfahrung, an denen Neues aufbricht, stellen zugleich ein Gegengift dar gegen die Versuchung, das zerebrale Geschehen samt seiner Aufmerksamkeitsaspekte rein adaptiv zu deuten.“[110]

Es scheint mir ein wenig fraglich, ob dies das beste Argument gegen eine „rein adaptive“ kognitionswissenschaftliche Deutung ist, denn warum könnte das, was uns als „Gedächtnismoment“, also als zeitlich versetzt *erscheint*, nicht der Dauer und dem Ineinandergreifen verschiedener, miteinander vernetzter neuronaler Prozesse geschuldet sein? Doch folgen wir Waldenfels noch ein wenig.

Auch er führt Husserls Überlegungen zu den „Horizonten des Gegenwartsfeldes“[111] auf eine Art weiter, die durch die Nutzung derselben grammatikalischen Zeitform an die bei Currie gesehenen Überlegungen erinnert. Was bei Currie als Teil der Leseerfahrung untersucht wurde, erscheint nun bei Waldenfels als grundsätzlicher Modus weltlicher Erfahrung:

> „Was auf uns zukommt, kommt unseren Erwartungen zuvor. Dieses *Zuvorkommen*, das nicht meinem eigenen Vorgehen entspringt, bedeutet, daß meine Zukunft mir vorauseilt und älter ist als ich selbst. Halten wir uns an die gewohnte Zeitfolge, so

[107] Ebd., S. 144.

[108] Ebd., S. 153.

[109] Ebd., S. 154.

[110] Ebd., S. 154.

[111] Ebd., S. 88.

> liegt diese Zukunft für mich selbst in der Vergangenheit. Diese eingesetzte und eingegebene Zukunft, die mich dorthin versetzt, wo ich nie war, entspricht der Vorstellung von einem zweiten Futur. Umgekehrt ist es aber auch so, dass mein Eingehen auf das, was mir zuvorkommt, von anderem herkommt und mir eine Zukunft eröffnet – so wie es das schlichte ‚Komm!' tut. Diesen Rückbezug auf eine zukunftsträchtige Vergangenheit bezeichnen wir als Herkommen. Die radikale Fremdheit dessen, was unsere eigenen und alle vorhandenen Möglichkeiten übersteigt, hat zur Folge, daß die vergangene Zukunft in eine zukünftige Vergangenheit überspringt. (…) Ohne jene radikale Fremdheit bleibt uns nur das Wechselspiel von Fortschritt und Rückschritt (…). Doch dies gehört bereits zur Verarbeitung der Erfahrung (…)"[112]

Anders als bei Currie wird hier eine Sprache mehrfach ineinander verschachtelter Möglichkeitsräume gefunden. Als „radikal fremd" wird das bezeichnet, was all diese Möglichkeiten übersteigt. Zwischen *Zuvorkommen* und *Herkommen* kann sich „etwas" im Überspringen der gesetzten Möglichkeiten als radikal Fremdes zeigen – wird aber wohl, so könnte man das weiterführen, durch die Verarbeitungsleistung der Erfahrung umgehend wieder gezähmt oder eingehegt (zum Beispiel in einer Erzählung von Fort– und Rückschritten).

Doch bleiben wir noch einen Augenblick bei dieser seltsamen Formulierung des „radikal Fremden". Gibt es auch weniger radikal Fremdes? Und wie fremd sind uns die Inhalte unseres Bewusstseins überhaupt? Findet die Überraschung nicht erst dort statt? Gerade, wenn Waldenfels wie oben gesehen davon ausgeht, dass das Widerfahrnis erst durch die Erzählung zur Erfahrung wird, kann es hier ja eigentlich nichts radikal Fremdes mehr geben. Anderes vielleicht. Oder eben, wie bei Lévinas gesehen, die Spur einer Berührung.

Der Bruch zwischen vorsprachlichem Erleben und seiner sprachlichen Erschließung in der Erzählung findet sich auch in Erwartungsstrukturen wieder. Mit Paul Valéry formuliert er:

> „Unerwartbar ist all das, was nicht nur aus dem üblichen Rahmen fällt, sondern sich den bestehenden Möglichkeitsbedingungen entzieht. ‚Das Erwartete ist *Idee*, determinierte Sinnesempfindung. Das Unerwartete ist Schock, ungeformte Sinnesempfindung.' (Valéry (…)). Dies trifft bis zu einem gewissen Grad auf alles zu, was unseren Erwartungen zuvorkommt, was auf uns zukommt."[113]

Aber wäre dann nicht jede Sinnesempfindung, bevor sie als Erfahrung verarbeitet wurde, ein Schock? Und gestalten die Sinne den Erwartungshorizont nicht maßgeblich mit? Wie hängen Unerwartetes und Unerwartbares zusammen? Kann es nicht auch erwartbares Unerwartetes geben? Gerade im Fiktionalen gehört es ja sogar zur Akzeptabilität von Überraschungen, dass sie im Rückblick erwartbar scheinen, und dennoch wurden sie in ihrer konkreten Ausprägung nicht vorhergesehen.

---

[112] Ebd., S. 90.

[113] Ebd., S. 91.

Auch Waldenfels nimmt an verschiedenen Stellen Bezug auf aktuelle Ergebnisse der Hirnforschung. Das Gehirn sei wie der Körper phänomenologisch als Ermöglichungsrahmen bestimmter Modalitäten der Erfahrung von Interesse, die ohne das Mechanische und Mediale des Körpers nur unzureichend beschrieben werden können. Er schlägt vor, „den Leibkörper als einen Inbegriff von Modalitäten“ zu betrachten, das heißt „als die Art und Weise, wie uns dies und jenes in der Welt, an uns selbst und bei Anderen begegnet“[114]. Das Verhältnis zwischen diesen phänomenologischen Befunden und Neurowissenschaften könnte aus seiner Sicht des Aufmerksamkeitsgeschehens wie folgt entwickelt werden[115]:

> „Das Aufmerksamkeitsgeschehen, das uns hier als Gradmesser dient, öffnet eine Perspektive jenseits von Dualismus und Monismus. Wir entgehen der Alternative eines autonomen Ich und eines determinierten Schein-Ich, wenn wir von jemandem ausgehen, *dem* etwas auffällt, bevor und auch während er oder sie in eigener Person spricht und agiert. Die Zerebralisierung wäre dann als ein Prozeß zu begreifen, in dem das, was wir tun, systematisch in physische Effekte *umgesetzt* wird. Sie wäre ferner als ein Organisationsprozeß zu begreifen, der zur Bildung von Auffälligkeiten und zur Reliefbildung *beiträgt*. Dies erlaubt es, gestalttheoretische Einsichten neurologisch zu reformulieren und zu präzisieren, nach den neuronalen Grundlagen von Figurierung, Mustererkennung, Konstellationsbildung oder sensomotorischer Synchronisierung zu fragen und so auch die neuronale Infrastruktur des Aufmerksamkeitsgeschehens zu durchleuchten (…)“. [116]

Die Arbeit von Waldenfels erweitert den bisherigen Blick auf die Fähigkeit, sich überraschen zu lassen, um die Perspektive des Aufmerksamkeitsgeschehens, als dessen Spezialfall sich das Überraschtwerden sehen lässt. Als Spezialfall wäre es unter anderem durch den besonderen Affekt des Überraschtseins zu charakterisieren. Hieran werde ich später noch anknüpfen, wenn es um die Überraschung als Emotion geht.

Zugleich zeigt Waldenfels, dass das im Bewusstsein verortete Aufmerksamkeitsgeschehen nicht nur die schöpferische und transformative Verarbeitung der Erfahrung auf der Ebene der Sprache oder der Handlung beinhaltet, sondern auch nicht ohne den Leib – als Inbegriff der *Modalität des Herkommens* – beschrieben werden kann.

---

[114] Ebd., S. 138.

[115] Maren Wehrle legt in ihrer Dissertation zu „dynamischen Konzeption der Aufmerksamkeit aus phänomenologischer und kognitionspsychologischer Sicht“ im Gegensatz dazu die Aufmerksamkeitsphänomenologie von Bernhard Waldenfels relativ schnell beiseite, da dieser sich vornehmlich mit der pathischen Form der Erfahrung befasse, welche sich "per definitionem nicht im Rahmen eines geplanten Experimentes feststellen lässt" und daher für die Kognitionswissenschaft von nachrangiger Bedeutung sei. Sie selbst versuche stattdessen in ihrer Arbeit, die habituelle, kulturellen Einflüssen unterliegende Form der Aufmerksamkeit als "implizite Form der Selektivität" herauszuarbeiten, gerade um vor diesem Hintergrund "eine explizite Offenheit und eine vorurteilsfreie Responsivität erreichen zu können", als Grundlage für eine "explizite ethische Haltung" und "eine wirkliche Horizonterweiterung". Unabhängig davon, dass sie in ihrem Urteil meines Erachtens Waldenfels nicht gerecht wird, unterliegt auch ihre Haltung selbst kontingenten kulturellen Einflüssen, die Waldenfels wiederum als „Neurophorie“ bezeichnen könnte. Vgl. Wehrle, Maren (2013): Horizonte der Aufmerksamkeit. Entwurf einer dynamischen Konzeption der Aufmerksamkeit aus phänomenologischer und kognitionspsychologischer Sicht. München: Fink.

[116] Waldenfels 2004, S. 151.

Überraschungs- und Aufmerksamkeitsgeschehen werden bei Waldenfels weitgehend synonym behandelt. Ich möchte das Verhältnis allerdings differenzierter betrachten. Denn wenngleich jede Überraschung mit einem Aufmerksamkeitsgeschehen einhergeht, also auch ohne die von Waldenfels geschilderte Doppelbewegung aus Auffallen und Aufmerken nicht zu denken ist, meine ich, dass die Überraschung damit noch nicht vollständig beschrieben ist und dass sich umgekehrt nicht sagen lässt, dass auch jedes Aufmerksamkeitsgeschehen eine Überraschung beinhaltet. Denken wir hier nur an den Fall, wo mir auffällt, dass endlich die Tage wieder länger werden. Ich richte dann meine Aufmerksamkeit auf die Uhr und den Kalender, vielleicht auch auf die ersten Schneeglöckchen und die anderen Lieder der Vögel, aber von einer Überraschung würde ich nicht sprechen. Mindestens zwei Unterschiede lassen sich allein an diesem Beispiel erkennen: Die Überraschung beinhaltet eine bestimmte Form der Erwartung, die offensichtlich nicht jedes Aufmerksamkeitsgeschehen kennzeichnet, und die Überraschung ist mit einem besonderen Gefühl verbunden. Beide Aspekte werden in den folgenden Abschnitten der Arbeit noch genauer untersucht werden.

Was Lévinas als das Durchstoßen des thematischen Bewusstseins bezeichnet, ist bei Waldenfels als ein Ereignis beschrieben, das in seiner „radikalen Fremdheit“ „unsere eigenen und alle vorhandenen Möglichkeiten übersteigt“. Nur die eigenartige Zeitstruktur der Erfahrung kann erklären, wie es sich zutragen kann, dass die Grenzen unserer eigenen und der vorhandenen Möglichkeiten erst im Augenblick eines sie übersteigenden Ereignisses bewusst werden und genau dadurch das Ich-Fremde betonen – vielleicht sogar überbetonen. Wir werden die Begrenztheit der Möglichkeiten des Ichs durch ein radikal Fremdes erfahren haben – Futur II. Genau hierin ist aber bereits die thematische Verarbeitung am Werk, deren Verbundenheit zum Reich der Rede und der Handlungen von Gewicht ist. Mit Waldenfels gerät damit für diese Arbeit die Dimension der Welt- und Leibhaftigkeit erneut in den Blick, und damit die „genuine Geselligkeit“ der Erfahrung. Auch dieser Aspekt wird in den beiden folgenden Abschnitten noch vertieft werden. Dazu ist es hilfreich, zunächst die Zusammenhänge zwischen Erwartungen, Fiktionen und Zukünften genauer zu untersuchen.

Erwartungen, Fiktionen, Zukünfte

Der Finanzmarkt, der „in all seinen modernen Ausprägungen lange vor der Industrialisierung entstanden ist“[117], schreibt Michael Seewalder, beruht seit seinen Anfängen auf dem

[117] Seewalder, Michael: Die Rhetorik des Marktes. Joseph de la Vegas Confusion de Confusiones, in: Priddat, Birger P. (2016): Erwartung, Prognose, Fiktion, Narration. Zur Epistemologie des Futurs in der Ökonomie. Marburg: Metropolis, S. 106.

Kultivieren der „Hoffnung auf unendliche Gewinnchancen in der Zukunft“[118], denn nur so konnte und kann es gelingen, „die kritische Masse an Kapital für große unternehmerische Wagnisse zu mobilisieren“. Aktien seien „im buchstäblichen Sinne als handelbare Wechsel auf die Zukunft entstanden“[119]. Investitionen in die Zukunft und damit einhergehend der Wunsch, diese vorausberechnen zu können und zu kontrollieren, ist der kapitalistischen Ökonomie inhärent. Vor diesem Hintergrund spielen Prognosen und Erwartungen in der Ökonomie eine entscheidende Rolle, Überraschungen sind da ungern gesehen.

In einer sozial, politisch und ökologisch zunehmend komplexen und veränderlichen Welt sehnen sich unternehmerische Entscheider*innen nach Methoden und Theorien, die ihre Handlungen dem Verdacht der Beliebigkeit entziehen. In einem Kontext, in dem Ziele in Form von Umsätzen, Gewinnen und Preisen zahlenförmig angelegt sind, scheint die Versuchung groß, auch Entscheidungen und menschliches Handeln als mathematisierbar zu begreifen. Die zunehmende Verfügbarkeit von Daten und die Erfolge stochastischer Verfahren in anderen hochkomplexen Aufgabengebieten stützen das Vertrauen in den Versuch, sich auch in Wirtschaft und Politik mit Verfahren des Messens und Berechnens gegen den Schrecken unsicherer Zukünfte zu wappnen. Heute stehen nahezu beliebig große Datenmengen und die technologischen Voraussetzungen dafür zur Verfügung. Die Vorstellung, mit stochastischen Verfahren aus vorliegenden Daten plausible Handlungsgründe ableiten zu können, ist für Priddat jedoch nicht mehr als eine Vermutung:

> „Einzig die Erwartung, dass das, was wir als wahrscheinlich bestimmt haben, *eher eintritt als nicht* (eine Häufigkeits*vermutung*), berechtigt uns, unsere Vorentscheidung als plausibel anzunehmen. Diese Plausibilität aber ist eine *narratio*, eine Erzählung wünschenswerter Zukünftigkeit, deren Eintritt gewisse Kontingenzen beibehält, die wir durch unsere Entscheidung nicht eliminieren können. (…) In der Plausibilität haben wir eine Art von Versicherung, richtig zu handeln, ohne genau zu wissen, ob oder wie. Die Versicherung ist die der Anschlussfähigkeit der Handlung, nicht die, das beste Ergebnis prognostiziert zu haben.“ [120]

Der konstruktive, fiktionale Charakter plausibilitätsschaffender Wahrscheinlichkeiten bleibe dabei allerdings unsichtbar. Er bestehe darin, dass ausgewählte Datenspuren singulärer historischer Ereignisse in einem Rahmen zusammengeführt werden, der keineswegs vorgängig existiert. Erst damit wird aktiv ein Koordinatensystem geschaffen, in dem sich mögliche Ereignisse verorten lassen als mehr oder wenig entfernt von einer gedachten Linie, die die Vergangenheit in die Zukunft projizieren soll (oder die Zukunft als plausible Folge der solcherart konstruierten Vergangenheit entwirft). Statt der Rahmensetzung (dem Akt des

---

[118] Ebd. S. 107.

[119] Ebd..

[120] Priddat, Birger P. (2016): Erwartung, Prognose, Fiktion, Narration. Zur Epistemologie des Futurs in der Ökonomie. Marburg: Metropolis, S. 32f. Kursiv im Original.

Erzählens) wird die von ihr erst geschaffene Projektion der Wirklichkeit (die Erzählung) für wahr gehalten.

> „Wir haben es mit einer kontingenten Setzung des Rahmens (*frame*) zu tun, innerhalb dessen die Geschichte/Erzählung so konzipiert erscheint, als ob sie einer vorgängigen Wahrscheinlichkeit entspricht. So wird der erzählte „Schein des Wahren“ erzählte Wahrheit, d.h. über die Fiktion reell. Die Konstruktion des Rahmens wird intransparent gehalten, um über den Erzähler eine Geschichtswahrscheinlichkeit anzubieten, die statt auf ihre (unwahrscheinliche) Konstruktion auf einen vorgängigen Schein verweist, als ob sie wahr wäre.“[121]

Wahrscheinlichkeitsbasierte Erzählungen machen Handlungen in einer unüberschaubaren Welt begründbar, und genau das macht für Priddat ihre Attraktivität in ökonomischen und politischen Kontexten aus. Ihr fiktionaler Charakter jedoch gerate dabei häufig aus dem Blick.

Er bezieht sich dabei maßgeblich auf die Arbeit der italienischen Soziologin Elena Esposito, die sich gefragt hat, warum der moderne Roman und die Wahrscheinlichkeitsrechnung historisch nahezu zeitgleich entstanden sind[122]. Esposito meint, das Aufkommen von beidem sei als Antwort auf ein zentrales Problem der westlichen Moderne zu verstehen: das Problem des Umgangs mit Kontingenz. Je unsicherer die Zukunft scheint, und je weniger Wahr und Falsch durch göttliche oder weltliche Autoritäten garantiert werden können, desto stärker brauche es neue Mittel und Wege, um gesellschaftlich akzeptierte Maßstäbe zu finden, an denen sich das Handeln ausrichten lässt. Wir sind genötigt, in der Gegenwart Entscheidungen zu treffen, ohne zu wissen, ob sie sich in der Zukunft noch als richtig erweisen oder wir sie bereuen werden. Gleichzeitig wissen wir, dass es anderen genauso geht. Deshalb wird es zunehmend interessant, einander dabei zu beobachten, wie wir das tun. Genau das leiste der Roman, indem darin Beweggründe von Handlungen auf eine Art sichtbar werden, wie sie das im alltäglichen Leben nicht tun. Statistiken und auf ihnen beruhende Berechnungen von Wahrscheinlichkeiten sind in Espositos Augen ein anderer Weg, der dasselbe Ziel verfolgt. Beide stehen für sie, in der Tradition der soziologischen Systemtheorie Luhmanns, in einem doppelten Bezug zur Realität. Hierbei handelt es sich wieder nicht um ungebrochene Abbildungsverhältnisse. Anders als Don Quijote, der ihr zufolge in dieser Hinsicht einen kulturellen Übergang sichtbar macht, wissen Menschen inzwischen sehr gut, dass sie es im Roman mit einer fiktiven Realität zu tun haben. Das gelte leider nicht für den Umgang mit Statistiken und Wahrscheinlichkeiten, die, so zeigt sie, in der Sachdimension nichts anderes als Fiktionen sein können, sofern sie vorgeben, sich auf eine *zukünftige Gegenwart* zu beziehen. Nützlich sei die Fiktion der Wahrscheinlichkeit jedoch in der Sozialdimension, nämlich als vereinfachter Blick auf *gegenwärtige Zukünfte*,

---

[121] Ebd., S. 72. Kursiv im Original.

[122] Esposito, Elena (2007): Die Fiktion der wahrscheinlichen Realität. Frankfurt am Main: Suhrkamp.

indem sie einen nachvollziehbaren und daher potentiell konsensfähigen Rahmen für Entscheidungen in der Gegenwart abgeben. In meinen Augen beschreibt der Begriff der *gegenwärtigen Zukunft* genau jenes Verhältnis zur Zukunft, das Currie als Haltung in Bezug auf literarische *fiction* herausgearbeitet und in seiner Sprache mit der grammatikalischen Form des Futur II gleichgesetzt hat. Dem Begriff des Unvorhergesehenen in der Fiktion ist die Erwartung des Vorhersehbaren weil Erklär- und Erzählbaren eingeschrieben. Doch wie beeinflusst dies die von Esposito „konkrete Realität" genannte Welterfahrung?

Sie führt hier den auf Luhmann zurückgehenden Begriff der *Realitätsverdopplung* ein. Darunter ist nicht nur zu verstehen, dass fiktive Realitäten alternative Realitäten beschreiben, sondern dass sie gleichzeitig Teil der *konkreten Realität* werden:

> „Auf den Bereich der fiktionalen Texte haben wir bereits hingewiesen: Die Autonomie der erfundenen Welten wird heute zwar in der Regel akzeptiert, dennoch tut man sich immer noch schwer damit, ihre konkrete Realität anzuerkennen. Mit konkreter Realität meinen wir dabei die unvermeidlichen praktischen Konsequenzen der Vertrautheit mit *fiction* für das Realitätsverständnis der modernen Gesellschaft. Obwohl das Gegenteil bewiesen ist, glaubt man noch immer, daß direkte Erfahrungen lehrreicher sind als vermittelte und letztere nur dann aufschlußreich sind, wenn sie die Realität genau widerspiegeln."[123]

Mit der *konkreten Realität der Fiktion* halten wir nun einen weiteren Begriff in den Händen für etwas, worauf wir bei Tobin mit dem Hinweis auf die von Geschichten geschaffenen Geschichtenerwartungen schon gestoßen sind.

Im Umgang mit auf Wahrscheinlichkeiten basierenden Prognosen und Entscheidungen fehle, schreibt sie, jedoch häufig das Bewusstsein über deren fiktiven Charakter, genauso wie das Verständnis dafür, wie stark sie gleichwohl konkrete Realitäten prägen. Esposito führt als Grund für diese Ignoranz etwas an, das sie Magie der Formalisierung nennt. Es wäre zu überlegen, ob eine solche nicht auch die in der Öffentlichkeit oft euphorische Bewertung der Ergebnisse von mit Big Data trainierten maschinellen Artefakten kontaminiert. Je mehr deren fiktionaler Charakter aus dem Blick gerät, desto schwieriger ist es, ihren gleichwohl bestehenden Einfluss auf die konkrete Realität zu reflektieren und zu kritisieren.

Wenn die Erschaffung fiktiver Realitäten durch Maschinen erfolgt, deren Sprache und Geschichten wir nicht oder nur unzureichend verstehen, wie dies bei Deep Learning und KNN der Fall ist, werden diese dennoch zu einem Teil unserer Erfahrungen und damit der konkreten Realität. Könnten wir dann durch Maschinen überrascht werden? Ich denke ja, genauso wie wir von Geschichten überrascht werden könnten, und genauso, wie Menschen schon lange und selbst wider besseres Wissen dem Output von Maschinen Sinn unterstellt

[123] Ebd., S. 72.

haben[124]. Genau das besagt ja gerade der Begriff der Realitätsverdopplung. Aber warum sollten wir das wollen?

Interessanter finde ich es, die Frage andersherum zu stellen: Können aktuelle Maschinen durch uns überrascht werden? Ist die Formalisierung nämlich immun gegenüber überraschenden Ereignissen, die im Stande wären, Schwachstellen einer Fiktion erkennbar zu machen und zu korrigieren – fehlt also die Fähigkeit, sich überraschen zu lassen –, dann wären wir nicht nur Fehlern, sondern auch einem Status Quo auf ewig ausgeliefert. Lassen wir Maschinen für uns die Realität aus einer unkontrollierbaren Menge aus Datenpunkten interpretieren, sollten wir uns zudem fragen, wo und wie wir selbst überraschender Ereignisse und Veränderungen jenseits unseres ganz direkten Erfahrungsumfelds noch gewahr werden können, inwiefern „Denkmaschinen“ also den Radius, in dem wir uns überraschen lassen können, einschränken. Werden Überraschungen damit ins Anekdotenhafte und aus der wissenschaftlichen Erforschung des Lebendigen verbannt?

Wie wichtig es ist, diese Fragen zu stellen, wird hoffentlich in den beiden folgenden Abschnitten noch deutlicher.

Überraschung als Emotion

Ist Überraschung eine Emotion oder nicht? In der Psychologie gehen die Meinungen darüber auseinander. Die einen sagen, Überraschungen setzen nicht unbedingt einen sogenannten „primären Bewertungsschritt“[125] voraus, da dieser jedoch Bestandteil der Definition von Emotion sei, gehörten sie nicht dazu. Andere, am Verhalten orientiert argumentierende Psycholog*innen, halten Überraschung dagegen sogar für eine „fundamentale“, angeborene Emotion und argumentieren unter anderem damit, dass „für Überraschung ein ganz bestimmter Gesichtsausdruck charakteristisch sei, der in allen Kulturen in weitestgehend gleicher Weise auffindbar sei“[126]. Verbale Berichte („ich bin überrascht“), physiologische Reaktionen („starker Anstieg neuraler Aktivierung“) sowie Verhaltensmerkmale („mimischer Ausdruck, Handlungsverzögerung, Fokussieren der

[124] Vgl. z.B. Weizenbaum, Joseph (1978): Die Macht der Computer und die Ohnmacht der Vernunft. Frankfurt am Main: Suhrkamp, S. 19f.

[125] Gemeint ist damit eine Art unmittelbare Bewertung hinsichtlich des eigenen Wohlbefindens, vgl. Meyer, Wulf–Uwe/ Niepel, Michael/ Schützwohl, Achim (1994): Überraschung und Attribution. In: Försterling, Friedich (Hg.)/ Stiensmeier–Pelster, Joachim (Hg.): Attributionstheorie. Grundlagen und Anwendungen, Göttingen u.a.: Hogrefe, S. 105.

[126] Vgl. ebd., S. 118f. Bereits Darwin hatte die Mimik der Überraschung beschrieben und die Vermutung geäußert, dass die weite Öffnung der Augen einer Vergrößerung des Gesichtsfelds diene und gleichzeitig ein soziales Signal sein könne, so dass zum Beispiel eine Mutter ihrem überraschten Kind mit einer Erklärung zu Hilfe kommen könne.

Aufmerksamkeit“) machten Überraschung nach außen hin als spezifische Emotion erkennbar. Das Auftreten der Überraschung wird mit dem Erleben einer *Schemadiskrepanz* in Zusammenhang gebracht. Der hier verwendete Begriff des *Schemas* geht auf die Psychologen Charlesworth und Rumelhart zurück und wird von ihnen als ein „durch einen Stimulus aktualisierter Ausschnitt unserer Wissensstruktur“ eingeführt. Wulf-Uwe Meyer, Rainer Reisenzein und Michael Niepel beschreiben Überraschung im „Handbuch Emotionspsychologie“ als „probabilistisches Reaktionssyndrom“[127]. Dazu legen sie ein Prozessmodell vor, das die mentalen Reaktionen und Prozesse auf Grund unerwarteter Reize in einem schematheoretischen Rahmen darstellt.

> „Schemata dienen zur Interpretation gegenwärtiger und vergangener Ereignisse sowie zur Vorhersage zukünftiger Ereignisse und damit indirekt der flexiblen, adaptiven Planung und Steuerung des Handelns. Um diese Funktion erfüllen zu können, müssen die Schemata einer Person (ihre informellen Theorien) aber annähernd korrekt sein. Dies wiederum erfordert – da das Wissen um die Umwelt häufig unvollständig ist und sich die Umwelt verändern kann – die fortlaufende Überprüfung der vorhandenen Schemata in Bezug auf ihre Übereinstimmung mit den verfügbaren Informationen (…) Wird dagegen eine Diskrepanz zwischen Schema und eingehender Information festgestellt, wird der Überraschungsmechanismus aktiviert (…): Die automatische Verarbeitung von Informationen wird unterbrochen, Überraschung wird erlebt, und kognitive Prozesse der Analyse und Bewertung des schemadiskrepanten Ereignisses werden in Gang gesetzt.“[128]

Überraschung sei, schreiben sie,

> „vermutlich der wichtigste (weil häufigste) Auslöser von Prozessen der spontanen Ursachenanalyse (…). Die Kenntnis der Ursachen eines schemadiskrepanten Ereignisses ist in vielen Fällen eine notwendige Voraussetzung für die erfolgreiche kurz- und langfristige Anpassung an ein unerwartetes Ereignis (d.h. für adäquates, sofortiges Reagieren auf das Ereignis und für die angemessene Aktualisierung des Schemas).“[129]

Die Herangehensweise der Autoren entspricht weitgehend den von Christiane Voss in ihrem Buch „Narrative Emotionen“ als gängig geschilderten Komponententheorien der Emotionen:

> „Der bisher erreichte Stand ist, dass Komponententheoretiker den Emotionen gemäß ihres erweiterten Begriffsverständnisses behaviorale (expressive), körperlich-perzeptive und intentionale (kognitiv-evaluative) Komponenten zuschreiben.“[130]

In der oben zitierten emotionspsychologischen Beschreibung der Überraschung finden wir Referenzen auf alle drei genannten Komponenten. Der expressive Teil kann in einer verbalen und/oder einer mimischen Äußerung bestehen. Als körperlich-perzeptiv könnte der

---

[127] Meyer, Wulf–Uwe/ Reisenzein, Rainer/ Niepel, Michael (2000): Überraschung. In: Otto, Jürgen H. (Hg.)/ Euler, Harald A. (Hg.)/ Mandl, Heinz (Hg.): Emotionspsychologie. Ein Handbuch. Weinheim: Beltz, Psychologie–VerlagsUnion, S. 253.

[128] Ebd., S. 254f.

[129] Ebd.. S. 258.

[130] Voss, Christiane (2004): Narrative Emotionen. Eine Untersuchung über Möglichkeiten und Grenzen philosophischer Emotionstheorien, Berlin, Boston: De Gruyter, S. 159f.

geänderte Aufmerksamkeitsfokus gelten, genauso wie messbare Änderungen im neuralen Bereich. Das Erleben einer „Schemadiskrepanz“ betrifft die intentionale Ebene. Genau auf dieser liegt jedoch bei der Überraschung eine Besonderheit.

Während die von Voss ins Feld geführten Emotionen wie Furcht, Stolz, Eifersucht, Neid, Verliebtsein etc. sich jeweils einen konkreten Sachverhalt außerhalb des Bewusstseins intendieren[131], welcher dann in Bezug auf das eigene Wohlbefinden beispielsweise als bedrohlich, einer eigenen Leistung zurechenbar, ungerecht oder begehrenswert bewertet wird, ist die intentionale Komponente im Fall der Überraschung meiner Ansicht nach anders aufgebaut. Denn was wir als überraschend bewerten, kann selbst wiederum eine Emotion sein, oder ein Ereignis, das für sich genommen bereits eine der vorgenannten Emotionen auslöst. Überraschung führt in meinen Augen auf eine Art Meta-Meta-Ebene und ist dadurch charakterisiert, dass sie in einer plötzlichen Einsicht den Irrtum bisheriger Auffassungen thematisiert. Das, worauf wir uns typischer Weise beziehen, wenn wir etwas als überraschend bezeichnen, liegt nämlich bereits selbst auf einer Meta-Ebene unseres Bewusstseins. Die Überraschung entsteht nicht wegen eines bestimmten Ereignisses oder Sachverhalts selbst, sondern weil deren intentionale Thematisierung auf einen Widerspruch stößt. Überraschung kommt, da haben die zuerst erwähnten Psycholog*innen recht, ohne einen „primären Bewertungsschritt“, also die Bewertung eines Ereignisses in Bezug auf unser Wohlbefinden, aus – es sei denn, man betrachtet ein kohärentes Weltbild als Bestandteil desselben. Aus diesem Grund würde ich vorschlagen zu sagen, dass die kognitiv-evaluative Komponente im Fall der Überraschung sich auf die thematisierende Intentionalität unseres Bewusstseins selbst bezieht. Im Gegensatz zu den unmittelbarer auf „Objekte“[132], Sachverhalte oder Widerfahrnisse bezogenen Emotionen, die nach dem Muster „x macht mich …, weil ich glaube, dass p“ verbalisierbar sind, könnte ein verallgemeinerter verbaler Ausdruck der Überraschung zum Beispiel lauten: „ich bin überrascht, denn bis eben dachte ich noch, dass p, aber nach dem x passiert ist, sehe ich jetzt, dass q“. Der Modus der Vergangenheit wird dem der Gegenwart, und einem Verb der Meinung oder des Glaubens wird eines relativer Sicherheit gegenübergestellt. Diese Besonderheit der Überraschung steht meiner Auffassung nach jedoch einer Einordnung als Emotion nach Voss nicht entgegen, da die von ihr aufgeführten Komponenten in der oben erläuterten Form enthalten sind.

---

[131] Ihr Begriff der Intentionalität scheint mir näher bei Husserl zu liegen als bei Lévinas, was hier aber nicht stören soll, da sich bei ihr unabhängig davon viel über Emotionen lernen lässt.

[132] So schreibt Voss unter anderem „In emotionalen Situationen sind Personen über Gedanken oder Wahrnehmungen mit etwas (Objekten) verbunden“, wobei „sich diese intentionalen Bezüge nicht auf kausale Beziehungen reduzieren lassen“. Ebd., S. 145.

Voss erweitert die präsentierten Komponententheorien um zwei weitere Aspekte, die sich nicht nur ebenfalls auf Überraschungen anwenden lassen, sondern die unser Verständnis davon weiter vertiefen.

> „Ergänzungsbedürftig sind die Komponententheorien aus meiner Sicht in zwei zentralen Hinsichten. Ein Manko dieser Ansätze besteht darin, dass sie die phänomenale Dimension der Lust- und Unlustgefühle von Emotionen nicht adäquat behandeln.
>
> (…) Aus meiner Sicht ist es ein weiteres Manko der Komponententheorien, dass sie nicht mehr darauf eingehen, wie die ausdifferenzierten Konstituenten der Emotionen miteinander zusammenhängen (müssen), um die spezifische Einheit zu bilden, welche Emotionen sind. Ich werde für die These argumentieren, dass die innere Einheit der multikomponentalen Emotionen in ihrer narrativen Strukturiertheit besteht.“[133]

Emotionen werden immer, schreibt sie, von „hedonistischen“ Gefühlen wie Lust oder Unlust (sie nennt sie „H-Gefühle“) begleitet, die, wie sie am Beispiel eines befriedigenden Zornausbruchs darlegt, jedoch nicht mit den positiven oder negativen Bewertungen des auslösenden Sachverhalts korrespondieren müssen (X wäre dann zornig, weil ihr p widerfahren ist, fühlt sich aber gut dabei, weil sie so ihren Frust abreagieren kann anstatt traurig zu sein…). Emotionen lassen sich also nicht per se als lustvoll oder schmerzhaft charakterisieren. Dennoch sind sie es immer irgendwie. Das spezifische des Beitrags der H-Gefühle zu den Emotionen liege darin, dass sie gerade nicht semantisch aufgeladen seien.

> „Nach meiner Auffassung tragen Gefühle (..) vielmehr gerade deshalb etwas Spezifisches zu Emotionen und ihrer Intentionalität bei, weil sie selbst keine intentionalen Phänomene sind. Durch das natürlich gegebene Vermögen zu fühlen erfahren wir auch auf vorsprachliche Weise etwas über uns in der Welt.“[134]

Ein Blick auf die Überraschung – die Voss selber allerdings nicht explizit erwähnt – kann das bestätigen. Wir reden von „guten“ und „bösen“ Überraschungen - in der konkreten Realität vielleicht abhängig davon, ob wir das auslösende Ereignis eher als angenehm oder als unangenehm empfinden. Im Bereich des Fiktionalen können wir aber auch „böse“ Überraschungen genießen – hier kann das mit der Überraschung einhergehende H-Gefühl wie bei Tobin gesehen sich zum Beispiel darauf beziehen, ob wir den Schauer lieben, uns betrogen oder auf anregende Weise unterhalten oder belehrt fühlen und so fort. In Film und Literatur genießen wir Überraschungen generell eher als zum Beispiel in einem geschäftlichen Umfeld, wo eine schlechte Fehlerkultur dazu führen kann, dass wir uns bloßgestellt fühlen, wenn wir einen Irrtum zugeben müssen.

> „Wir fühlen einen Schmerz oder eine Lust, eine Aufregung oder Bedrückung unmittelbar körperlich. Ein inhaltlicher Zusammenhang solcher phänomenalen

[133] Ebd., S. 159f.

[134] Ebd., S. 192.

> Gefühlsqualitäten mit dem Rest einer emotionalen Veränderung wird, wie ich meine, durch ihre Einbettung in ein emotionales Narrativ hergestellt. Dadurch erhalten die an sich selbst semantikfreien Gefühlsregungen einen bestimmten Sinn, wie z.B. den, die positive oder negative Tönung einer emotionalen Regung wie Eifersucht mit all ihren repräsentationalen und motivationalen Elementen anzuzeigen.“[135]

Dabei sei das „Wechselspiel der emotionalen Komponenten untereinander (…) dynamisch und bewirke weitere Kettenreaktionen“[136] – sie bezeichnet das auch als „holistisch“ oder als „Tanz der Komponenten“.

Um nun die Komplexität und das Zusammenspiel dieses Tanzes im Bewusstsein zu fassen, findet Voss den Begriff der „intentionalen Superstruktur“ der Emotionen, die darin besteht, „dass ihre Elemente insgesamt – und nicht nur ihre unmittelbaren Meinungsanteile – als Bedeutungsträger eines thematischen Überbaus wahrgenommen werden“[137]. Es gehöre zur Zuschreibung von Emotionen, dass diese ohne Sinnverlust nur innerhalb eines abgesteckten semantisch-normativen Rahmens vollzogen werden könne[138]. Die Abhängigkeit der Emotionen von narrativen Strukturen zeige sich auch daran, dass man sie durch Änderung der mit ihnen verknüpften Narrative beeinflussen kann. Die „Tatsache, dass wir uns primär im Modus des Erzählens – und nicht z.B. in Form von Statistiken oder bloßen Aufzählungen – zu unseren Emotionen verhalten“[139] sei ein weiteres Indiz. Und wir können

> „(…) Emotionen – nicht Empfindungen oder bloße H-Gefühle – einklagen, unterdrücken oder auch trainieren, weil sie semantische Haltungen sind, auf die wir erzählerisch ausmalend und sogar argumentierend einwirken können.“[140]

Voss hält Emotionen daher für den Überbegriff für „thematisch geleitete, affektiv getönte, psycho-physische Relationen“[141], die bewusste und unbewusste, körperliche und mentale Bewegungen narrativ miteinander verbinden.

> „Ihre Einheit – und das ist ein Kerngedanke meiner narrativen Konzeption der Emotionen – ist das Resultat davon, dass diese Bewegungen auf einen gemeinsamen Sinn, einen thematischen Mittelpunkt eines Geschehens narrativ bezogen werden. (…) Geschichten – nicht isolierte Propositionen – sind damit sozusagen die kleinsten Bedeutungseinheiten der Emotionen.“[142]

---

[135] Ebd.

[136] Ebd., S. 199.

[137] Ebd., S. 203.

[138] Ebd., S. 204.

[139] Ebd., S. 211.

[140] Ebd., S. 222.

[141] Ebd., S. 208.

[142] Ebd., S. 209.

Dabei beruhen die Erzählweisen auf einem in paradigmatischen Situationen erlernten strukturellen Verständnis und beinhalten ihr zufolge daher auch immer so etwas wie einen Anfang, einen Höhepunkt und einen Schluss, was ich für eine etwas vereinfachte Darstellung halte, aber hier nicht weiter verfolgen möchte. Jedenfalls sei dieses strukturelle Verständnis Ergebnis unserer individuellen wie historisch-kulturell situierten Erfahrung der Welt:

> „Das strukturelle Verständnis emotionaler Szenarien, das uns als Deutungsfolie auch neuer emotionaler Erfahrungen dient, gewinnen wir aus einem abstrahierenden Querschnitt der vielen emotionalen Interaktionen, mit denen wir in Erzählungen, Spielen, Filmen, Berichten und vor allem in Begegnungen im Laufe unseres Lebens konfrontiert werden."[143]

Diese „intentionale Superstruktur" der Emotionen als einen erst in einer konkreten Umwelt geformten Überbau zu verstehen, erlaubt es zwar, einzelne ihrer Bestandteile bereits bei Babys zu entdecken, zum Beispiel eine plötzliche Umlenkung der Aufmerksamkeit bei unerwarteten Ereignissen, wie oben gesehen. Doch selbst, wenn diese augenscheinlich sogar noch affektiv aufgeladen, also vom „H-Gefühl" Freude begleitet sind, könnte man nach Voss in diesem Alter mangels Sprachkompetenz dem Baby noch nicht die Emotion der Überraschung zuschreiben. Wir tun dies aber. Vielleicht genauso, wie schon viele einem „stochastischen Papagei"[144] Intelligenz zugeschrieben haben, obwohl er sie nicht hat. Aber über die Probleme bei Fremdzuschreibungen haben wir ja schon bei Tobin einiges gelernt – ich erinnere nur an den „Fluch des Wissens". Sie sind allenfalls Annäherungen, und wir können dabei fürchterlich falsch liegen.

Analog zu Esposito stellt Voss klar, dass die Narrativität der Emotionen in einem doppelten Verhältnis zur Wirklichkeit steht. Der interpretierend verstehende Zugang entfaltet eine eigene Kraft, die aus sich heraus wiederum wirklichkeits-konstituierend wirkt[145].

Vielleicht erweist sich die Überraschung in diesem Zusammenhang noch einmal als Sonderfall, und zwar in mehrfacher Hinsicht. Wohlgemerkt, gerade die Überraschung lässt sich als narrative Emotion beschreiben. Dennoch kann hier wohl kaum von einem Dreiakter aus Anfang, Höhepunkt und Schluss gesprochen werden. Schon eher handelt es sich bei der Überraschungserzählung um eine Art Witz. Auch lässt sich die Erzählung im Fall der Überraschung nicht beliebig umdeuten, einklagen oder unterdrücken. Und, anders als Trauer, Eifersucht, Verliebtsein etc., die Voss wiederholt als Beispiele anführt, lässt sich Überraschtheit auch nicht anlassunabhängig in die Länge ziehen. Was andererseits aber

---

[143] Ebd., S. 219.

[144] Vgl. Bender, Emily M./Gebru, Timnit/ McMillan–Major, Angelina/ Shmitchell, Shmargaret (2021): On the Dangers of Stochastic Parrots. Can Language Models Be Too Big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, S. 610–623.

[145] Vgl. Voss 2022, S. 220.

durch die Überraschung stärker noch als bei allen anderen Emotionen betont wird, ist der Fokus auf das kausale Denken. Wie oben beschrieben gilt die Überraschung als der im Alltag häufigste Anlass, die Frage nach dem Warum zu stellen. Die Überraschung stößt uns auf eine Grenze unserer Freiheit im Erfinden von bestimmten Geschichten, nämlich solchen, die unserem Handeln in der Welt wirkungsvoll Orientierung verleihen können.

Christiane Voss hat mit ihrer philosophischen Erkundung der Emotionen eine äußerst fruchtbare Grundlage für ein allgemeines Verständnis der Fähigkeit, sich überraschen zu lassen, geschaffen. Als Überleitung zum folgenden Abschnitt möchte ich ihren Hinweis übernehmen, dass die kognitiv-evaluative Seite der Emotionen auf doppelte Art normativ zur Wirkung kommt. Zum einen stellen die Bewertungen, die mit Emotionen einhergehen, einen normativen Bezug zu einem Sachverhalt oder Ereignis her, der stark durch die Normen des kulturellen Kontexts geprägt wurde. Zum anderen beruhen die Konventionen, bestimmte sinnliche, verhaltensbezogene und gefühlsmäßige Wahrnehmungen bestimmten Emotionen zuzuordnen, ebenfalls den Normen des Umfelds, in dem wir leben: In der Regel halten wir es für angebracht, uns in einer so und so wahrgenommenen Situation so und so zu fühlen, weil auch andere das tun, oder meinen darauf hoffen zu dürfen, dass andere unsere Emotionen nachvollziehen können, wenn wir ihnen die Situation entsprechend schildern. Emotionen sind, wie die Erzählungen, die sie begleiten, stark von unserer jeweiligen sozialen und historischen Verortung geprägt. Emotionen – und ich zähle die Überraschung dazu – entstehen in Gesellschaft.

Vinciane Despret, der ich ja bereits die Anregung zu dieser Arbeit verdanke und mit der sich über gemeinsame philosophische Wurzeln ebenfalls eine Brücke zu Isabelle Stengers bauen lässt, hat in einem ihrer frühen Bücher den Begriff der Versionen vorgeschlagen, der auf die soziale Konstruiertheit unseres Verständnisses abhebt:

> „(…) version is not imposed, it is constructed. It is not defined as truth or lie or illusion but as a form of becoming: the becoming of a text that is ceaselessly reworked (…). Version does not reveal the world any more than that it veils it, it makes it exist in a possible manner. Version is not the fact of a single man, it is the source and fruit of a relationship, it is the negotiation of what is turned around, transformed and translated.“[146]

Was dieser Begriff noch stärker als der der Erzählung leistet ist, dass er die Fabrikation von Verständnis, von Erzählungen oder Theorien in spezifischen Kontexten in den Blick nimmt und dass hierbei gesellschaftliche Beziehungen und Beziehungen zur Umwelt eine entscheidende Rolle spielen. Hierbei jedoch sind häufig bestimmte Perspektiven dominant, andere werden ausgeschlossen. Während Despret die Fabrikation von Versionen an konkreten Beispielen nachvollzieht, was die vorliegende Arbeit in eine neue Richtung führen würde, ziehe ich es vor, an deren Ende angelangt an den Anfang anzuknüpfen und mit

[146] Despret, Vinciane (2004): Our Emotional Makeup. Ethnopsychology and Selfhood, New York: Other Press, S. 25.

Isabelle Stengers die Frage zumindest anzureißen, wie im Kontext gesellschaftlicher Normen bestehende Überzeugungen, Weltbilder oder „Versionen“ fraglich werden und unterschiedliche Perspektiven in einem gemeinsamen Verständnis zusammenfließen können.

Verstehen in einer Welt unterschiedlicher Perspektiven

Wie wirken Kollektive auf die Fähigkeit, sich überraschen zu lassen, ein? Wie können unterschiedliche Perspektiven in gesellschaftlichen Sinngebungsprozessen Raum finden? Welche Normen und Institutionen könnten die mögliche Politisierung von Weltbildern und Überzeugungen befördern? Welche behindern sie? Antworten auf diese Fragen können aus Isabelle Stengers letzter Publikation gewonnen werden, in der sie Alfred North Whitehead einer neuen Lektüre unterzieht und nach Antworten auf die Krise der modernen Wissenschaften in der öffentlichen Wahrnehmung sucht. Unter Verwendung des Begriffs der *Abstraktion* werden in ihrem Text vor allem die Weisen einer solchen Thematisierung kritisiert.

Denken ohne Abstraktion, meinen Whitehead und mit ihm Stengers, gebe es nicht. Die Rolle der Philosophie bestehe deshalb darin, ihre präzisen Weisen – „the modes of abstraction that equip the thought of that era“[147] – zu erforschen und kritisch den jeweils konkreten, individuellen Erfahrung gegenüberzustellen.

> „(…) experiences do not become mere illustrations of categories. Such categories are tools for philosophy, and philosophers comprehend them only by learning how to handle them, and in which circumstances. This is why to think with Whitehead is to learn. It is to learn to think in zigzag against the straight line. This straight line encourages us to think out statements refer to facts that, correlatively, present themselves as isolable. The zigzag entails experimentation going back and forth between a conceptual abstraction that aims to bring coherence into existence and a situation that our usual statements make bifurcate. The zigzag gives *this* situation the power to reclaim its reality as individual concrete fact.“[148]

Stengers und Whitehead sehen die Philosophie in der Verantwortung, entsprechende Praktiken des Verstehens (an dieser Stelle als Übersetzung von „making sense“) zu suchen und entwickeln. Eine wesentliche Rolle spiele dabei, unterschiedliche Perspektiven zur Geltung zu bringen. Die Evolution bringe mit sich, dass sämtliche Bezugspunkte dabei als

[147] Stengers, Isabelle (2023): Making Sense in Common. A Reading of Whitehead in Times of Collapse. Minneapolis, London: University of Minnesota Press, S. 20.

[148] Ebd., S. 141.

veränderlich gedacht werden müssen – sie als plastisch zu verstehen, wie Malabou sagen könnte.

> „Accepting evolution means agreeing to abandon the idea that thought needs fixed references to avoid confusion and arbitrariness.“[149].

Einen fixen Boden für Konsens – wie die Kritische Theorie ihn voraussetze – könne es nicht geben. Deshalb müsse die Philosophie zu der Vorstellung ermutigen, hinter dem, was andere zweifeln lässt, einen anderen Zugang zur immensen Vielfältigkeit der Welt zu sehen, auch wenn sich dieser (noch) nicht in Worten ausdrücken lasse. Die Diffamierung des *common sense* als manipulierbare und im Kern irrationale öffentliche *Meinung*, gehe daher völlig am Problem vorbei. Denn wenn *common sense* nicht als Grundlage, sondern als Ergebnis verstanden wird, muss zunächst die Frage gestellt werden, wie es um die Praxis der Einbeziehung verschiedener Perspektiven in Bezug auf ein Thema bestellt ist.

> „What if such blind hate against this elite ‚who knows‘ was related to the cultural and political desaster which I am calling the defeat of common sense? The words of Bertold Brecht spring to mind: ‚We often speak of the violence of a river overflowing but less of the violence of the banks that confine it‘.“[150]

Für die Wissenschaftsphilosophin Stengers heißt „making sense in common“ nicht, dass Wissenschaft nicht gehört werden sollte, und auch nicht, dass es keine Fakten gäbe oder man einem grundsätzlichen Relativismus anheim fallen müsse. Stattdessen gelte es, eine neue, respektvolle Haltung einzunehmen, die den Raum für andere Perspektiven offen lasse:

> „Civilizing modernity, then, means getting specialists to learn to situate themselves, to use Haraway’s turn of phrase, or, to evoke Deleuze, to honor the truth of the relative (in contrast to the relativeness of the truth), the truth of knowledges that know how to present themselves as relative to the question they prove able to pose effectively.“[151]

Unter „Zivilisierung“ versteht Stengers mit Whitehead vereinfachend zusammengefasst den situierten Versuch einer Gesellschaft, sich selbst zu verstehen[152]. *Common sense* wird damit zum Inbegriff der Zivilisation. Aus dieser Charakterisierung geht hervor, dass die Frage, wer an diesem „Abenteuer“ in welcher Form beteiligt wird – und wer nicht – relevant ist. Eine Zivilisation, die bestimmte Perspektiven ausschließt, kann sich also nur unzureichend verstehen. Nachahmenswerte Gegenbeispiele findet sie in den Erfahrungen global vernetzter politischer Aktivist*innen.

---

[149] Ebd., S.7.

[150] Ebd., S. 14f.

[151] Ebd., S. 27.

[152] Vgl. ebd., S. 25–33.

> „Today it is what activists strive for: making sense in common. Instead of coming to an agreement, making sense in common is about knowing together that the reasons for resisting, as different as they may be, need each other“[153].

Auch in anderen Kulturen gebe es Institutionen, die verschiedenen Perspektiven Raum geben. Stengers nennt zum Beispiel Formen des Ältestenrats, in denen die Beteiligten nicht für sich selbst sprechen, sondern eine anwaltsähnliche Rolle für unterschiedliche Perspektiven übernehmen und am Ende gemeinsam zu einem Ergebnis kommen, das für alle tragbar ist. Ähnlich positiv bewertet sie die von Kleingruppen auf Demonstrationen praktizierten Konsensfindungsmethoden oder Bruno Latours Vorschlag, nach Betroffenheiten und einem Cliquenbildungen vermeidenden Zufallsprinzip besetzte öffentliche Zusammenkünfte in Anlehnung an die griechische ἀγορά einzuführen und um eine „diplomatische Intervention“ zu ergänzen, deren Aufgabe es wäre, den vertretenen Perspektiven zu Respekt zu verhelfen[154]. Ein wichtiger, von Stengers zwar erwähnter, aber in meinen Augen zu kurz ausgeführter Aspekt hierbei ist, dass derartige Formen des „making sense in common“ schnell Zynismus und Aversion erzeugen, wenn sie folgenlos bleiben und die Teilnehmenden sich am Ende nicht wirkmächtig, sondern betrogen fühlen[155].

Der Bezug zur Fähigkeit, sich überraschen zu lassen, lässt sich über ein zentrales Motiv herstellen: Die Möglichkeit, sich von dem oder den Anderen affizieren zu lassen, setzt voraus, Räume der Begegnung zu schaffen und zu nutzen. Dies geschieht zum einen auf der Ebene der Wissenspraxis (Wessen Perspektiven gehen ein in den *common sense?*), zum anderen deshalb, weil der *common sense* in Christiane Voss' Worten die Deutungsfolie abgibt, vor der alle ihre Wahrnehmungen erst verstehen. Weisen der Abstraktion sind immer „räuberisch“, wie Leben räuberisch sei: „life is robbery“. Die Abstraktionen eines, wie wir gesehen haben thematisierenden und erzählenden, intentionalen Bewusstseins beraube immer andere Aspekte ihrer Sichtbarkeit. Deshalb ist es so wichtig, sie zu untersuchen und dem, was seiner Sichtbarkeit beraubt wird, die angemessene Aufmerksamkeit („due attention“) zu schenken. Angemessen ist für Stengers Aufmerksamkeit auch dann, wenn sie den Blick darauf lenkt, was Ausnahmen und Sonderfälle, die mit vorhandenen Theorien nicht erklärt werden können einzigartig macht. Hier führt sie als Beispiel die Entstehung von Tornados an, die sich nicht auf das Verhalten einzelner Moleküle zurückführen lässt, sondern erst im Zusammenhang mit besonderen Bedingungen („circumstances“) erklärt werden kann und plötzlich die Masse („crowd“) relevant werden lässt. Derartige Beispiele stellten die Theoriebildung über statistische Mittelwerte und das Ausblenden möglicher

[153] Ebd., S. 17.

[154] Vgl. ebd., S. 102f.

[155] Vgl. ebd., S. 63.

Wechselwirkungen in Frage. Im Umgang mit Lebendigem müsse das Partikulare mehr zählen als das Allgemeine, und die Interaktion mehr als die Atomisierung ihrer Teile.

> „In this case, to put it briefly, what is problematized is he pertinent notion of average value, the bridge built by statistical mechanics between the ‚law' the gases seem to obey and the crowd of molecules composing gases. The notion of average value implies that overall behaviour results from behaviours indifferent to one another. But what does this notion of indifference depend on for its validity? Specialists in statistical mechanics relate this notion to the posibility of dividing a system into microregions that should be ‚uncorrelated', which means that a local deviation relative to the average has no, or only negligible, repercussions on other regions. In contrast, the emergence of strong, long–range correlations marks the appearance of a form of social sensibility (…). It implies agents in crowds behaving differently. (…) Instead of being an answer, the crowd has become a problem: what can a crowd do?"

Gegen ihre These, dass es gelte, nach Weisen der Abstraktion zu suchen, die ein Wechselspiel sozialer Interaktionen und die Möglichkeit origineller Antworten auf ein unvorhersehbares Angerührtsein weniger stark ausschließen, also auf die Wahrnehmung von Überraschungen vorbereiten, könnte man vielleicht einwenden, dass eine erwartete Überraschung gar keine Überraschung mehr sei. Ein solcher Einwand wäre jedoch zu kurz gegriffen, weil eine Überraschung, wie wir gesehen haben, immer an ein ganz konkretes, individuelles Ereignis gebunden ist, das als solches auch durch eine allgemeine Überraschungserwartung nicht vorhergesehen werden kann.

Versuch eines Fazits

Surprised machines? Als ich diese Arbeit begonnen habe, hatte ich zunächst vor, die Fähigkeit, sich überraschen zu lassen, als einen weiteren Einwand gegen die Rede von künstlicher Intelligenz zu entwickeln. Zwei Überlegungen haben mich schon bald eine andere Richtung einschlagen lassen. Zum einen ist es angesichts der derzeit mit massiven Investitionen vorangetriebenen Entwicklungen in diesem Bereich ein möglicherweise in seiner Aussagekraft sehr vergängliches und angesichts der schieren Menge an Veröffentlichungen, in denen sich Versprechungen oft nur mit tiefem Fachwissen von tatsächlichen Fortschritten unterscheiden lassen, auch gar nicht leistbares Unterfangen, sich auf einen definierten Stand der Forschung zu beziehen. Auf der anderen Seite erscheint es mir vermessen, ein theoretisches Urteil darüber fällen zu wollen, über welche Fähigkeiten Maschinen grundsätzlich verfügen oder nicht verfügen können.

Dennoch hat es sich als überaus lohnenswert erwiesen, sich mit dem Phänomen der Überraschung zu befassen. Gerade weil es bislang so gut wie keine philosophische Literatur dazu gibt. Eine solche Forschungslage erfordert einen Blick in die Breite und eröffnet damit die Möglichkeit, unterschiedliche Perspektiven in ungewöhnlichen Konstellationen zusammenzubringen, über die Grenzen der eigenen Disziplin hinaus.

An dieser Stelle wurde die Frage nach der Überraschungsfähigkeit von Maschinen erneut relevant. Denn die Kognitionswissenschaft, die sich gleichermaßen für menschliche wie für künstliche Intelligenz interessiert, ist wahrscheinlich diejenige Disziplin, die sich gegenwärtig am stärksten für das Thema der Überraschung interessiert. Und zwar genau aus dem Grund, der der ursprüngliche Ausgangspunkt meiner Arbeit war, nämlich weil der Umgang mit Überraschungen in der KI-Forschung ein bislang ungelöstes Problem darstellt. Die Überraschung von Maschinen wurde mithin zu einer Art geistigem Dialogpartner dieser Arbeit, deren Ziel es ist, zu beschreiben, wie die Fähigkeit, sich überraschen zu lassen, beschaffen ist und was sie für Voraussetzungen hat.

Dies ist philosophisch aus mehreren Gründen interessant. Als spezielles Phänomen des Bewusstseins erschließt das Überraschungsgeschehen auf spezifische Weise, wie wir selbst und die Welt uns in der Erfahrung gegeben sind. Als Geschehen, in dem uns Erwartungen, Überzeugungen oder Weltbilder (in dem Sinne, wie Tobias Matzner sie bei Wittgenstein versteht) fraglich werden können, ist die Überraschung die explizitere Verwandte des Staunens und lehrt etwas über (falsche) Gewissheiten, die Dimensionen des bloßen Meinens und das Lernen aus Fehlern. Als solches wiederum wirft sie dann auch technikethische Fragestellungen auf: Wo wollen und können wir es uns leisten, bei Maschinen auf an das

Überraschungsgeschehen angelehnte Korrekturmöglichkeiten zu verzichten? Als gemeinsames Thema mit Kognitions- und Technikwissenschaften markiert die Überraschung hinsichtlich theoretischer Annäherungen einen äußerst spannenden Schnittstellenbereich, in dem Technik und Philosophie vielleicht sogar zu Sparringspartnern werden könnten.

Am Ende meiner Untersuchung erweist sich das Überraschungsgeschehen, und damit die Fähigkeit, sich überraschen zu lassen, als komplexes und vielschichtiges Phänomen, das all diese Anschlussmöglichkeiten mit sich bringt.

„Stets findet Überraschung statt / Da, wo man`s nicht erwartet hat“.

Dieses Zitat von Wilhelm Busch ist lustig, weil es eine Selbstverständlichkeit – also unser Alltagsverständnis – auf skurrile Weise in Worte fasst. Dieses Alltagsverständnis führt uns bereits mitten hinein ins Thema. Überraschungen in Film und Literatur spielen auf raffinierte Weise mit Erwartungen. Wie kommen diese zustande, und was sagt das über unser Verhältnis zur Welt?

Vera Tobin zeigt am Beispiel der Überraschung auf, dass Erwartungen ein zentrales Moment unserer Welterfahrung sind. Sie sind nicht nur ihr mit neuen Sinneseindrücken veränderliches Ergebnis, sondern prägen bereits ihre Wahrnehmung, unsere Einschätzung ihrer Relevanz und ihre Einordnung in einen Kontext, in dem sie uns erst verständlich werden. In der Überraschung wird die Komplexität dieses Erwartungsmanagements deutlich.

Dies ist ein Argument, warum Überraschung (*surprise*) als Phänomen unserer Wahrnehmung grundsätzlich vom informationstheoretischen Begriff der Überraschung (*surprisal*) zu unterscheiden ist, der sich allein auf den Informationswert von Zeichen bezieht. Überraschung ist eine subjektive Erfahrung.

Umso erstaunlicher ist es, dass sie phänomenologisch bislang so wenig Beachtung gefunden hat. Eine wichtige Ausnahme stellt hier Bernhard Waldenfels’ *Phänomenologie der Aufmerksamkeit* dar. In dem Aufsatz *Geweckte und gelenkte Aufmerksamkeit* wird der Bezug zur Überraschung explizit genannt. "Erfahrung kommt nie ganz ohne Überraschung aus, solange sie nicht erstarrt“[156], schreibt Waldenfels dort. Sie könne unsere "*primäre, innovative* Aufmerksamkeit" wecken[157] und schließe „eine radikale Selbstüberraschung“[158] ein.

156 Waldenfels 2016, S. 31.

157 Ebd., S. 35.

158 Ebd., S. 31.

Der starke Erfahrungsbegriff, den Waldenfels mit seiner Untersuchung anstrebt, kann, das gehört zu den wichtigen Ergebnissen dieser Arbeit, in produktiver Weise ergänzt werden durch die Betrachtung der Überraschung als Emotion. Christiane Voss' philosophische Emotionstheorie bringt jenseits kognitiver Aspekte Komponenten und Abhängigkeiten unserer Welterfahrung in den Blick, die in der Philosophie erst seit wenigen Jahrzehnten wieder Beachtung finden. Auch sie ist daher gezwungen, auf Kenntnisstände anderer Disziplinen zurückzugreifen, hier namentlich der Psychologie.

Nach Voss kommen bei emotionalen Vorgängen neben körperlich-perziptiven, kognitiv-evaluativen und expressiven Komponenten noch zwei weitere Aspekte zum Tragen: die Verknüpfung mit semantisch nicht aufgeladenen Gefühlen der Lust oder Unlust, durch die wir „auf vorsprachliche Weise etwas über uns in der Welt" erfahren, und die kulturell geprägte Entstehung *intentionaler Superstrukturen,* in der all diese Komponenten erst als Einheit verstanden werden.

Der expressive Teil der Überraschung kann in verbalen oder mimischen Äußerungen bestehen. Die körperlich-perzeptiven Vorgänge lassen sich mit Waldenfels in allgemeinerer Form als Aufmerksamkeitsgeschehen beschreiben, das eine dualistische Rede von Subjekt und Objekt verbietet und das Dazwischen beschreibt. Auf der kognitiv-evaluativen Ebene besteht, so denke ich, eine Besonderheit der Überraschung darin, dass der Bewertungsschritt sich typischer Weise auf Erwartungen, also *höherstufig* auf vorangegangene Bewertungen des Bewusstseins, bezieht. In der Überraschung wird unser eigenes Erwartungsmanagement zum Gegenstand der Bewertung. Überraschungen werden oft als mehr oder als weniger erfreulich empfunden. Dabei kann Freude insbesondere über eine neue *Einsicht* entstehen. In der Überraschung können wir Denken als lustvoll erleben. Ich vermute, das macht nicht zuletzt ihren besonderen Reiz im Fiktionalen aus.

Die allermeisten der in dieser Arbeit vereinten Fundstücke verweisen darauf, dass unsere Erfahrung sich auf fundamentale Weise in Form von Geschichten vollzieht, Geschichten, die wir uns selbst erzählen, und die unser Verständnis *sind*. Dies hat mehrerlei Konsequenzen. Zum einen sind wir auf der Ebene unserer bewussten Erfahrung „immer schon zu spät". Zum anderen beinhaltet unsere Erfahrung damit zwangsläufig Setzungen, Thematisierungen oder Abstraktionen, die gleichermaßen unumgehbar wie potentiell irrig sind - also, wenn man so will, in das Reich „bloßen Meinens" verweisen. Genau das aber kann in der Überraschung bewusst werden.

Findet das Auffallen und Aufmerken, wie Waldenfels betont, auch in einem „Zwischenreich" statt, erfahren wir Überraschung dennoch innerhalb unseres Bewusstseins, und zwar als „Schemadiskrepanz", wie die Psycholog*innen sagen. Dieses Erlebnis innerhalb des Bewusstseins setzt eine Metaebene voraus, nämlich die Möglichkeit zur Reflexion über die

eigenen Bewusstseinsinhalte und zugleich eine Auffassung von Zeitlichkeit. Wir können erfahren, dass wir uns geirrt haben.

Gerade weil wir selbst es sind, die sich die Geschichten erzählen, die unsere Erfahrung ausmachen, und weil wir von anderen dazu verführt werden können, dabei Fehler zu machen, wie wir bei Tobin nachzuvollziehen gelernt haben, liegen im Waldenfels'schen Zwischenreich des Aufmerksamkeitsgeschehens wie in unserer dadurch anstoßbaren Reflexionsfähigkeit wichtige Korrektive für unsere Fähigkeit, uns in der Welt zu orientieren. Darin besteht nicht zuletzt eine Funktion des Spiels mit Geschichten: sie erweitern Orientierungsräume in einer sich permanent verändernden Welt (Dehaene), und können insbesondere unseren Umgang mit unterschiedlichen Perspektiven trainieren, was für ein Leben in sozialen Gemeinschaften von zentraler Bedeutung ist (Tobin).

Nicht nur unser soziales und ökologisches Verwobensein, auch die sprachgebundene und auf Erzählungen basierende Form unseres intentionalen Bewusstseins ermöglichen und erfordern einen ständigen Austausch mit unserer Umwelt. Weder biologisch noch physikalisch noch chemisch noch unserer Erfahrung nach sind wir als isolierte oder autarke Wesen denkbar. Es sollte daher naheliegen, nach den Formen des Austauschs zu fragen, in denen sich unsere Erfahrung vollzieht. Weil auch Geschichten in einem spezifischen sozio-kulturellen Kontext entstehen, wurde am Schluss der Arbeit mit Isabelle Stengers zumindest angerissen, dass hierbei auch gesellschaftliche Institutionen, in denen Wissen vergemeinschaftet wird oder gemeinschaftlich entsteht, auf den Prüfstand zu stellen sind. Der Austausch mit dem oder den Anderen ist schon bei Lévinas der fundamentale Sinn der Sprache, „die ursprüngliche Sprache".

Aus der großen Bandbreite an Fragestellungen, die auf den vorangegangenen Seiten eingeführt wurden, möchte ich an dieser Stelle nur ein paar wenige herausgreifen, deren Weiterverfolgung ich für spannend halte.

Lässt sich noch mehr darüber herausfinden, was uns überrascht und was wir im Gegensatz dazu als Fehler abtun und nicht weiter beachten? Ergäbe sich ein anderes Bild, wenn wir in unsere Überlegungen einbeziehen, dass wir ja auch von Kunst, von Tönen und Geräuschen, Bildern und haptischen Erlebnissen überrascht werden können? Dieser Aspekt ist in der vorliegenden Arbeit aus zwei Gründen ausgeklammert geblieben: erstens auf Grund der Literaturlage und zweitens, weil auch solche Überraschungen nur durch Sprache vermittelbar und damit für andere zugänglich sind. Spätestens dann aber haben wir es wieder mit Erzählungen zu tun. Was in den Kognitionswissenschaften als Gap zwischen Neuronalem und Symbolischen bezeichnet wird, bleibt auch philosophisch unzureichend konzeptualisiert. Bei Catherine Malabou haben wir zumindest einen Ansatz in diese

Richtung gesehen. Die Erklärungslücke ist allerdings noch größer: Sie besteht zwischen neuronalen Vorgängen und der Form der Erzählung.

Zahlreiche weitere Untersuchungen sind im Anschluss denkbar. Inwiefern wird diskursiv vorbereitet, welcher Art von Ereignissen wir zugestehen, dass sie unsere innere Reflexionstätigkeit – in Daniel Kahnemanns Worten das behäbigere System II[159] - anwerfen, um die Geschichten oder Weltbilder, die unser Verständnis bisher geleitet haben, in Frage zu stellen? Welche Formen von Geschichten erzählen wir uns überhaupt? Wie entwickeln sich Geschichten im Zusammenleben mit Technik? Durch welche Art von Geschichten werden heute unsere Geschichtenerwartungen trainiert? Und brauchen wir neue, wie Donna Haraway in *Unruhig bleiben*[160] vorgeschlagen hat? Wie können Sinngebungsprozesse in sozialen Gemeinschaften so organisiert werden, dass sie Überraschungen zulassen?

Wenn ich angeben sollte, in welcher Form diese Arbeit erzählt ist, würde ich auf die Tragetaschentheorie der Erzählung von Ursula LeGuin verweisen. Es ist keine Heldengeschichte, die zunächst ein gefährliches Tier vorstellt und dieses dann in einem dramatischen Spannungsbogen in drei Akten mit Waffengewalt durch einen einsamen Helden (seltener eine einsame Heldin) zur Strecke bringt. Vielmehr habe ich mich einer ihrer Theorie nach noch viel älteren Kulturerrungenschaft bedient: des Tragebeutels. Ich habe aufgesammelt, was ich vorgefunden habe, auch von einsamen Jäger*innen Erlegtes, und bringe dies nun

> „mit nach Hause (…), wobei zu Hause schlichtweg eine weitere, größere Art von Beutel oder Tasche, ein Behältnis für Menschen, ist, [um] es später herauszunehmen und zu essen oder zu teilen oder für den Winter in einem solideren Behältnis einzulagern oder ins Medizinbündel oder in den Schrein oder ins Museum zu geben, an den heiligen Ort, der das, was heilig ist, bewahrt, und dann am darauffolgenden Tag mehr oder weniger dasselbe zu tun (…)“[161].

Das Wichtigste für diese Arbeit ist das Teilen.

Und was ist mit den Maschinen?

Dass sie weit davon entfernt sind, sich überraschen lassen zu können, haben wir gesehen. Aber *müssen* sie das überhaupt können? Ich würde diese Frage nicht von den Ansprüchen einer Intelligenzdefinition her beantworten wollen, sondern im Zusammenhang mit ihrem

---

[159] Vgl. Kahnemann, Daniel (2012): Schnelles Denken, langsames Denken. München: Siedler. Kahnemann ist mit dieser Theorie auch Teil laufender Debatten zur Weiterentwicklung künstlicher Intelligenz, vgl. z.B. Garcez, A. d'Avila/ Lamb, Luis C. (2020): Neurosymbolic AI: The 3rd Wave.

[160] Vgl. Haraway, Donna J. (2018): Unruhig bleiben. Die Verwandtschaft der Arten im Chthuluzän, Frankfurt am Main, New York: Campus.

[161] Le Guin, Ursula (2020): Am Anfang war der Beutel. Warum uns Fortschritts-Utopien an den Rand des Abgrunds führen und wie Denken in Rundungen die Grundlage für gutes Leben schafft, Klein Jasedow: Drachenverlag, S. 17.

jeweiligen Einsatz betrachten. In Bereichen, die grundsätzlich als veränderlich gedacht werden müssen, das heißt zumindest überall dort, wo wir es mit Lebendigem zu tun haben, scheint mir die Fähigkeit, „Schemadiskrepanzen“ zu erkennen und mit ihnen umzugehen, indem vorgenommene Setzungen in Frage gestellt werden, unabdingbar. Denn hier haben wir es mit kontingenten Verhältnissen zu tun, die wir zwar mit Hilfe von Fiktionen zu beschreiben versuchen, denen wir damit aber nur so lange gerecht werden, als wir den fiktiven und perspektivgebundenen Charakter unserer Geschichten im Blick behalten. An den Übergängen zwischen fiktiver und realer Realität wächst die Aufgabe der Politisierung im Matznerschen Sinn.

Wollten wir die Fähigkeit, sich überraschen zu lassen, künstlich herstellen, müssten wir jedenfalls Maschinen bauen, für die es möglich wäre, in Betracht zu ziehen, was vorher noch nie in Betracht gezogen wurde. Die Entwicklung ihres „Denkens“ müsste auf sehr viel komplexere Weise als bisher zurückgebunden sein an lebendige Gemeinschaften, in denen Sinn und Situationsverständnisse gemeinsam geschaffen und verändert werden können. Wir müssten also Maschinen bauen, die zu allererst nicht losgelöst von ihrem (lebendigen) Umfeld denken und agieren, sondern mit diesem gemeinsam. Die sich mit uns über ihr und unser und ein gemeinsames Verständnis austauschen können. Wir müssten Maschinen bauen, die über zeitlich, thematisch und gesellschaftlich gebundene metakognitive Fähigkeiten verfügen. Die um den fiktionalen Charakter ihrer Modelle wissen und erkennen, wenn ein Ereignis diesen in nicht nur für sie, sondern auch für ihr Umfeld relevanter Weise widerspricht und die in der Lage sind, ihr „Denken“ daraufhin zu revidieren. Und ob sie das dann Überraschung nennen würden hinge auch davon ab, wie die Welt um sie herum das sieht.

Das Wesen einer Arbeit, die ein neues Feld zunächst einmal erschließt, aber irgendwann abgeschlossen werden muss, ist es, dass am Ende mehr Fragen offen sind als beantwortet werden konnten. So ist es auch mit dieser Arbeit. Wenn es mir gelungen ist, das Überraschungsgeschehen als lohnendes Thema weiterer Erkundungen attraktiv zu machen, bin ich froh.

Glossar technischer Begriffe

„Künstliche Intelligenz“

Wie Clifford A. Pickover in seiner Illustrierten Geschichte der Künstlichen Intelligenz einleitend feststellt, haben „die Geheimnisse des Geistes, das Wesen des Denkens und die Möglichkeit künstlicher Wesen“ die Menschheit schon lange vor unserer Zeit zu unterschiedlichen Vorstellungen und Theorien inspiriert[162]. Genauso alt sind die Versuche, diese Theorien durch die Konstruktion entsprechender Apparate zu veranschaulichen und zu nutzen. „Intelligenz“ ist in sämtlichen Disziplinen – von der Psychologie über die Biologie bis hin zur Informatik – ein umstrittener Begriff. Diesen als gesetzt zu markieren würde bedeuten, die Frage nach den unter dem Begriff stattfindenden Bedeutungsverschiebungen nicht stellen zu können[163]. Heute erlebt die Rede von „Künstlicher Intelligenz“ nicht zuletzt dank neuer technischer Möglichkeiten (siehe Big Data, Maschinelles Lernen, Deep Learning) einen weiteren Boom. Eine empfehlenswerte Zusammenfassung der Höhen und Tiefen in der bisherigen Geschichte der KI inklusive der während verschiedener Phasen schwerpunktmäßig verfolgten Ansätze findet sich zusammen mit einer aktuellen Risikoanalyse in dem Aufsatz *A Brief History of AI: How to Prevent another Winter* von Toosi et al. (2021).

Starke KI / Schwache KI

Die begriffliche Unterscheidung von „starker“ und „schwacher“ KI wurde von Stuart Russell und Peter Norvig in ihrem zuerst 2003 erschienenen Kompendium „Artificial intelligence: a modern approach“ eingeführt[164]. Die Vorstellung, dass künstliche Systeme tatsächlich wie Menschen denken, handeln und sich weiterentwickeln könnten, ohne dabei auf bestimmte Anwendungsbereiche (oder Funktionen) festgelegt zu sein, nannten sie „starke“ im Gegensatz zu „schwacher KI“. Während „starke KI“ ein fern scheinendes Maximalziel beschreibt, schlagen sie vor, Systeme, die zumindest in einem definierten Teilbereich und im

---

[162] Pickover, Clifford A., (2021): Künstliche Intelligenz: eine illustrierte Geschichte. Von mittelalterlichen Robotern zu neuronalen Netzen, Kerkdriel: Librero, S. VIII.
Pickover beginnt seine Chronologie von materialisierten „Künstlichen Intelligenzen“ mit Ktesibios’ Wasseruhr ca. 250 v.Chr. (Pickover 2021, 7). Ähnliche Geräte seien jedoch auch schon „im alten China, Indien, Babylon, Ägypten, Persien und anderswo“ konstruiert worden (ebd.). Ktesibios selbst habe desweiteren auch eine sich bewegende „unheimliche roboterhafte Statue einer Gottheit“ entwickelt und auf öffentlichen Prozessionen in Alexandria vorgeführt.

[163] Siehe hierzu auch Malabou, Catherine (2019): Morphing Intelligence. From IQ Measurement to Artificial Brains, New York: Columbia University Press.

[164] Die aktuelle Auflage ist Russell, Stuart J./ Norvig, Peter (2023): Künstliche Intelligenz. Ein moderner Ansatz, 4. aktualisierte Auflage, München: Pearson.

Ergebnis intelligente menschliche Verhaltensweisen imitieren – oder sogar übertreffen – können, als „schwache KI“ (aber damit eben immer noch als KI) zu bezeichnen.

Symbolische KI

In den 1950er bis zu den 1990er Jahren vorherrschendes Paradigma in der KI–Forschung, das davon ausging, intelligentes Verhalten unter Verwendung formal-semantischer und regelbasierter Wissensrepräsentationen automatisieren zu können. Im Ergebnis wird dieser Ansatz, dessen Erfolg davon abhängt, dass Vorgaben und Regeln möglichst vollständig angegeben werden können, heute in vielen Bereichen durch Vorgehensweisen weit übertroffen, die auf „maschinellem Lernen“ beruhen .

Maschinelles Lernen

Maschinelles Lernen kann für unsere Zwecke allgemein in Abgrenzung an die symbolische KI als „neues Programmierparadigma“ verstanden werden, das nicht wie das klassische Programmieren Regeln und Daten vorgibt, die Antworten erzeugen, sondern durch das aus Fragen und vorgegebenen Antworten Regeln erzeugt werden. Die heute erfolgreichen Anwendungen maschinellen Lernens „lernen“ während einer sogenannten „Trainingsphase“, und liefern auf dieser Grundlage im produktiven Betrieb „eigenständige Antworten“[165]. Ihr Erfolg ist nicht denkbar ohne die Verfügbarkeit hoher Rechenleistung und enormer Datenmengen (siehe Big Data). Unter anderem darin unterscheidet sich Maschinelles Lernen von klassischen statistischen Verfahren. François Chollet, Autor einer der wichtigsten Software-Bibliotheken für Deep Learning, schreibt, und das ist für eine technikphilosophische Betrachtung durchaus relevant: „Daher spielt die mathematische Theorie beim Machine Learning und insbesondere beim Deep Learning nur eine vergleichsweise kleine – vielleicht zu kleine – Rolle. In diesem praxisorientierten Fachgebiet werden Ideen häufiger empirisch erprobt als theoretisch vorhergesagt“[166].

Deep Learning

Deep Learning ist ein Teilgebiet des Maschinellen Lernens, das Datenmodelle in mehreren aufeinander bezogenen Repräsentationsebenen erzeugt. Die Anzahl dieser Ebenen wird als Tiefe des Modells bezeichnet. Durch die Gewichtung oder Parametrisierung der zu den einzelnen Ebenen führenden Transformationen kann die Vorhersagegenauigkeit des Modells

[165] Vgl. Chollet, François/Allaire, J.J. (2018): Deep Learning mit R und Keras. Frechen: mitp, S. 25f.

[166] Ebd., S. 26.

optimiert werden. Deep Learning automatisiert die Merkmalserstellung beim maschinellen Lernen[167] und wird heute meist mit Hilfe von Künstlichen Neuronalen Netzen (KNN) umgesetzt.

(Künstliche) Neuronale Netze

Künstliche neuronale Netze sind eine aktuell häufig verwendete Lernstrategie im maschinellen Lernen. Sie sind von der „Hypothese [inspiriert], dass mentale Aktivität hauptsächlich aus elektrochemischer Aktivität in Netzen von Gehirnzellen, den sogenannten Neuronen, besteht“[168]. Damit realisieren sie das sogenannte *konnektionistische Paradigma* der KI-Forschung. Die Entwicklung entsprechender Informationsverarbeitungseinheiten und ihrer Vernetzungsfunktionen begann bereits in den 40er Jahren des vergangenen Jahrhunderts.

Big Data / strukturierte und unstrukturierte Daten

Der Zugriff auf eine riesige Menge an Daten („Big Data“) ist die wohl wichtigste immaterielle Ressource für heutige KI-Systeme, die auf maschinellem Lernen beruhen[169]. Dabei können Daten in unterschiedlicher Form vorliegen. Unter strukturierten Daten sind Daten zu verstehen, die eine definierte semantische Struktur haben, wie zum Beispiel Literaturangaben in einem bestimmten Format (Autor*in, Titel, Erscheinungsjahr, Verlag etc. an jeweils bestimmten Positionen, getrennt durch bestimmte Satzzeichen oder *Tags* usw.). Der Großteil an Daten liegt aber unstrukturiert vor und kann nicht anhand vorgegebener Kategorien weiterverarbeitet werden. Genau darin, in unübersichtlichen Datenmengen Merkmalsstrukturen ausfindig zu machen, liegt die Aufgabe maschinellen Lernens.

Algorithmus / Algorithmen

Als Algorithmus wird klassischer Weise die formale und prozedurale Beschreibung eines Lösungswegs für ein bestimmtes Problem bezeichnet. Durch die informatische Umsetzung lassen sich Problemlösungswege automatisieren. Fachfremd steht der Begriff heute gelegentlich auch für auf maschinellem Lernen basierende Anwendungen wie zum Beispiel Empfehlungs- oder Recommender-Systeme.

---

[167] Ebd. S. 42f.

[168] Russell, Stuart J./ Norvig, Peter (2023): Künstliche Intelligenz. Ein moderner Ansatz, 4. aktualisierte Auflage, München: Pearson, S. 846.

[169] Eine umfassende Aufzählung der für heutige KI nötigen Ressourcen findet sich in Crawford, Kate (2021): Atlas of AI. Power, Politics, and the Planetary Costs of Artificial Intelligence, New Haven, London: Yale University Press.

Literatur

Baldi, Pierre (2021): Deep Learning in Science. Cambridge, New York et.al.: Cambridge University Press.

Bender, Emily M./Gebru, Timnit/ McMillan–Major, Angelina/ Shmitchell, Shmargaret (2021): On the Dangers of Stochastic Parrots. Can Language Models Be Too Big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, S. 610–623. https://doi.org/10.1145/3442188.3445922 (19.12.2023).

Bischof, Andreas (2017): Soziale Maschinen bauen. Epistemische Praktiken der Sozialrobotik, Bielefeld: transcript.

Boltanski, Luc/Chiapello, Ève (2006): Der neue Geist des Kapitalismus. Konstanz: UVK Verlagsgesellschaft.

Chollet, François/Allaire, J.J. (2018): Deep Learning mit R und Keras. Frechen: mitp.

Crawford, Kate (2021): Atlas of AI. Power, Politics, and the Planetary Costs of Artificial Intelligence, New Haven, London: Yale University Press.

Currie, Mark (2013): The Unexpected. Narrative Temporality and the Philosophy of Surprise, Edinburgh: Edinburgh University Press. https://doi.org/10.1515/9780748676309

Dehaene, Stanislas (2021): How We Learn. Why Brains Learn Better Than Any Machine ... for Now, New York: Penguin.

Despret, Vinciane (2004): Our Emotional Makeup. Ethnopsychology and Selfhood, New York: Other Press.

Despret, Vinciane (2019): Was würden Tiere sagen, würden wir die richtigen Fragen stellen? Münster: Unrast.

Esposito, Elena (2007): Die Fiktion der wahrscheinlichen Realität. Frankfurt am Main: Suhrkamp.

Garcez, A. d'Avila/ Lamb, Luis C. (2020): Neurosymbolic AI: The 3rd Wave. https://arxiv.org/pdf/2012.05876.pdf (19.12.2023). https://doi.org/10.48550/arXiv.2012.05876

Golden, Ryan/ Delanois, Jean Erik/ Sanda Pavel/ Bazhenov, Maxim (2022): Sleep prevents catastrophic forgetting in spiking neural networks by forming a joint synaptic weight representation. PLoS Computational Biology 18(11): e1010628. https://doi.org/10.1371/journal.pcbi.1010628

Gramelsberger, Gabriele (2010): Computerexperimente. Zum Wandel der Wissenschaft im Zeitalter des Computers, Bielefeld: transcript.

Haraway, Donna J. (2018): Unruhig bleiben. Die Verwandtschaft der Arten im Chthuluzän, Frankfurt am Main, New York: Campus.

Harrach, Sebastian (2014): Neugierige Strukturvorschläge im maschinellen Lernen. Eine technikphilosophische Verortung. Bielefeld: transcript.

Jain, Elenor/ Trappe, Tobias (1998): Staunen; Bewunderung; Verwunderung. In: Ritter, Joachim (Hg.), Gründer, Karlfried (Hg.)/ Gabriel, Gottfried: Historisches Wörterbuch der Philosophie, Band 10. Basel: Schwabe.

Kahnemann, Daniel (2012): Schnelles Denken, langsames Denken. München: Siedler.

Le Guin, Ursula (2020): Am Anfang war der Beutel. Warum uns Fortschritts-Utopien an den Rand des Abgrunds führen und wie Denken in Rundungen die Grundlage für gutes Leben schafft, Klein Jasedow: Drachenverlag.

Lévinas, Emmanuel (2017): Die Spur des Anderen. Untersuchungen zur Phänomenologie und Sozialphilosophie, übers. u. hrsg. v. W. N. Krewani, 7. Auflage, Freiburg (Breisgau), München: Karl Alber.

Malabou, Catherine (2006): Was tun mit unserem Gehirn? Zürich: Diaphanes (Neuauflage 2021).

Malabou, Catherine (2019): Morphing Intelligence. From IQ Measurement to Artificial Brains, New York: Columbia University Press. https://doi.org/10.7312/mala18736 (19.12.2023).

Malabou, Catherine (2022): Plasticity. The Promise of Explosion. Edinburgh: Edinburgh University Press. https://doi.org/10.1515/9781474462143 (19.12.2023).

Matzner, Tobias (2013): Vita variabilis. Handelnde und ihre Welt nach Hannah Arendt und Ludwig Wittgenstein. Würzburg: Königshausen & Neumann.

Meyer, Wulf–Uwe/ Niepel, Michael/ Schützwohl, Achim (1994): Überraschung und Attribution. In: Försterling, Friedich (Hg.)/ Stiensmeier–Pelster, Joachim (Hg.): Attributionstheorie. Grundlagen und Anwendungen, Göttingen u.a.: Hogrefe, S. 105–122.

Meyer, Wulf–Uwe/ Reisenzein, Rainer/ Niepel, Michael (2000): Überraschung. In: Otto, Jürgen H. (Hg.)/ Euler, Harald A. (Hg.)/ Mandl, Heinz (Hg.): Emotionspsychologie. Ein Handbuch. Weinheim: Beltz, Psychologie–VerlagsUnion, S. 253–263.

Parisi, German I./Kemker, Ronald/ Part, Jose L./ Kanan, Christopher/ Wermter, Stefan (2019): Continual lifelong learning with neural networks. A review, In: Neural Networks 113, 54–71. https://doi.org/10.1016/j.neunet.2019.01.012 (19.12.2023)

Pickover, Clifford A., (2021): Künstliche Intelligenz. Eine illustrierte Geschichte. Von mittelalterlichen Robotern zu neuronalen Netzen, Kerkdriel: Librero.

Priddat, Birger P. (2016): Erwartung, Prognose, Fiktion, Narration. Zur Epistemologie des Futurs in der Ökonomie. Marburg: Metropolis.

Russell, Stuart J./ Norvig, Peter (2023): Künstliche Intelligenz. Ein moderner Ansatz, 4. aktualisierte Auflage, München: Pearson.

Seewalder, Michael (2016): Die Rhetorik des Marktes. Joseph de la Vegas Confusion de Confusiones, in: Priddat (2016), S. 103–128.

Stengers, Isabelle (2023): Making Sense in Common. A Reading of Whitehead in Times of Collapse. Minneapolis, London: University of Minnesota Press.

Swan, Jerry/ Nivel, Eric/ Kant, Neil/ Hedges, Jules/ Atkinson, Timothy/ Steunebrink, Bas (2022): The Road to General Intelligence. Studies in Computational Intelligence, Cham: Springer Nature. Online verfügbar (open access) unter https://link.springer.com/book/10.1007/978–3–031–08020–3 (19.12.2023).

Tobin, Vera (2018): Elements of Surprise. Our Mental Limits and the Satisfactions of Plot, Cambridge/London: Harvard University Press.

Toosi A, Bottino AG, Saboury B, Siegel E, Rahmim A. (2021): A Brief History of AI. How to Prevent Another Winter (A Critical Review), PET Clin. 2021 Oct;16(4):449-469. doi: 10.1016/j.cpet.2021.07.001 . Online verfügbar unter https://arxiv.org/abs/2109.01517 (21.12.2023)

Voss, Christiane (2004): Narrative Emotionen. Eine Untersuchung über Möglichkeiten und Grenzen philosophischer Emotionstheorien, Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110896268 (19.12.2023).

Waldenfels, Bernhard (2004): Phänomenologie der Aufmerksamkeit. Frankfurt am Main: Suhrkamp.

Waldenfels, Bernhard (2016): Geweckte und gelenkte Aufmerksamkeit. In: Müller, Jörn (Hg.)/ Nießeler, Andreas (Hg.)/ Rauh, Andreas: Aufmerksamkeit. Neue humanwissenschaftliche Perspektiven. Bielefeld: transcript, S. 25-45.

Wehrle, Maren (2013): Horizonte der Aufmerksamkeit. Entwurf einer dynamischen Konzeption der Aufmerksamkeit aus phänomenologischer und kognitionspsychologischer Sicht. München: Fink.

Weizenbaum, Joseph (1978): Die Macht der Computer und die Ohnmacht der Vernunft. Frankfurt am Main: Suhrkamp.

Erklärung zur Abschlussarbeit gemäß § 22 Abs. 7 APB TU Darmstadt

Hiermit erkläre ich, *Ruth Karl*, dass ich die vorliegende Arbeit gemäß § 22 Abs. 7 APB TU Darmstadt selbstständig, ohne Hilfe Dritter und nur mit den angegebenen Quellen und Hilfsmitteln angefertigt habe. Ich habe mit Ausnahme der zitierten Literatur und anderer in der Arbeit genannter Quellen keine fremden Hilfsmittel benutzt. Die von mir bei der Anfertigung dieser wissenschaftlichen Arbeit wörtlich oder inhaltlich benutzte Literatur und alle anderen Quellen habe ich im Text deutlich gekennzeichnet und gesondert aufgeführt. Dies gilt auch für Quellen oder Hilfsmittel aus dem Internet.

Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Mir ist bekannt, dass im Falle eines Plagiats (§38 Abs.2 APB) ein Täuschungsversuch vorliegt, der dazu führt, dass die Arbeit mit 5,0 bewertet und damit ein Prüfungsversuch verbraucht wird. Abschlussarbeiten dürfen nur einmal wiederholt werden.

English translation for information purposes only:

Thesis Statement pursuant to § 22 paragraph 7 of APB TU Darmstadt

I herewith formally declare that I, f*irst name last name*, have written the submitted thesis independently pursuant to § 22 paragraph 7 of APB TU Darmstadt without any outside support and using only the quoted literature and other sources. I did not use any outside support except for the quoted literature and other sources mentioned in the paper. I have clearly marked and separately listed in the text the literature used literally or in terms of content and all other sources I used for the preparation of this academic work. This also applies to sources or aids from the Internet.

This thesis has not been handed in or published before in the same or similar form.

I am aware, that in case of an attempt at deception based on plagiarism (§38 Abs. 2 APB), the thesis would be graded with 5,0 and counted as one failed examination attempt. The thesis may only be repeated once.

Datum / Date: 3.1.2024 (Ruth Karl)