Schuldige Maschinen? Autonome Systeme als Herausforderung für das Konzept der Verantwortung

von Markus Christen

I. Einleitung

Der Bau autonomer technischer Systeme ist Ziel umfassender Forschungen. Aufgrund der zunehmenden Komplexität technischer Systeme sowie ihrer laufend ausgeweiteten Einsatzräume sollen diese Artefakte Eigenschaften wie Lernen, selbstgesteuertes Verhalten, automatisiertes Problemlösen und Selbstreparatur aufweisen. Diese von den Informations- und Ingenieurwissenschaften geprägten Technologien haben unzweifelhaft wichtige Implikationen für die angewandte Ethik. Zwar hat die ethische Reflexion der Computertechnik eine lange Tradition. Kernpunkt dieser Debatte ist jedoch primär das Problem der Verfügungsmacht, welche die Computertechnik gewissen Gruppen der Gesellschaft zu geben vermag.[1] So erlauben neue Entwicklungen in der Datenbanktechnologie die plattformunabhängige Suche nach Datenmustern mittels Methoden des *data mining*. Dadurch werden in einem bisher nicht gekannten Ausmaß Informationen über Individuen zugänglich, was wichtige Fragen nach dem Schutz persönlicher Daten aufkommen lässt (*privacy-Debatte*). Im Zug der weltweiten Terrorbekämpfung sind sowohl auf technologischem Gebiet wie in gesetzgeberischer Hinsicht Entwicklungen im Gang – explizit genannt sei das US-amerikanische *Terrorism Information Awareness Programm*[2] –, die dieses Problem verschärfen.

Diese Entwicklungen sind wichtige Themen der angewandten Ethik, berühren aber nicht grundlegende Fragen zur möglichen Transformation von Kernbegriffen der Ethik (wie „Verantwortung") durch das Auftreten „technischer Akteure". Der technische Fortschritt führt nicht ausschließlich zu einer Zunahme menschlicher Verfügungsmacht – und damit zu einer Vermehrung der durch Menschen zu verantwortenden Handlungen. Beobachtbar ist auch das Gegenteil: Der Grad der

---

[1] Vgl. MOOR, J.H. (1985): *What is Computer Ethics?*, in: BYNUM, T.W. (ed.): Computers & Ethics, Blackwell, 266-275.

[2] Vgl. DEFENSE ADVANCED RESEARCH PROJECTS AGENCY (2003): *Report to the Congress Regarding the Terrorism Information Awareness Program*, http://foi.missouri.edu/totalinfoaware/reporttocongress.html (Website des *Freedom of Information Center* der University of Missouri).

Durchdringung der modernen Gesellschaft durch technologische Systeme hat ein Ausmaß erreicht, das die Steuerung dieser Systeme durch den Menschen, bzw. durch dafür ausgebildete menschliche Experten, zunehmend erschwert.[3] Sowohl einzelne Maschinen wie ganze Systeme haben eine Leistungsfähigkeit und Komplexität erreicht, die es im Hinblick auf Fehlervermeidung nicht ratsam erscheinen lassen, Menschen in die Systemabläufe einzubinden. Moderne Kampfflugzeuge beispielsweise lassen sich ohne den Einbezug ausgeklügelter Steuerungssysteme nicht mehr fliegen. Neben der bereits fortgeschrittenen Automatisierung großindustrieller Prozesse ist ein Vordringen der Computertechnologie in den Alltag feststellbar. Unter dem Stichwort *ubiquitous computing* diskutieren Fachleute seit einigen Jahren die Integration „smarter Technologie" in Gegenstände des alltäglichen Gebrauchs.[4] Selbst in den Wissenschaften umfasst der Trend der Automatisierung nicht nur rein mechanische Operationen, durchgeführt etwa durch die Sequenzierroboter in den Genlabors, sondern bereits Formen wissenschaftlichen Handelns wie Hypothesenbildung und experimentelle Prüfung dieser Hypothesen.[5] Gewiss lässt sich bei diesen Beispielen noch abwehrend behaupten, dass die „wesentlichen" Entscheide weiterhin von Menschen getroffen würden. Klar ist aber auch, dass im Rahmen einer solchen Entwicklung immer mehr Entscheide an technische Systeme delegiert werden. Im Rahmen dieser Delegation von „Entscheidungskompetenz" dürfte es zunehmend schwieriger werden, „wesentliche" von „unwesentlichen" Entscheidungen zu unterscheiden.

Dieses Problem ist im Kontext zweier Entwicklungen zu sehen: Zum einen besteht auf Seiten der Technologieentwicklung eine berechtigte Hoffnung, dass die Leistungsfähigkeit der Rechensysteme in den kommenden Jahren weiter zunehmen wird.[6] Zum anderen ist weiterhin ein Bedarf für die Automatisierung von Kontroll- und Steuerungsaufgaben vorhanden – etwa im Bereich der Stromversorgung oder der Telematik. Beide Entwicklungen gelten in wirtschaftlicher Hinsicht als sehr zukunftsträchtig. Es ist anzunehmen, dass gerade die westlichen Hochtechnologie-Nationen, die mit einer zunehmenden Abwanderung klassischer Produktionsarbeitsplätze in aufstrebende Schwellenländer konfrontiert sind, einen wirtschaftlichen Schwerpunkt in die Entwicklung dieser Technologien setzen werden. Unter Vernachlässigung der Möglichkeit, dass die moderne Zivilisation eine radikale Abkehr vom wissenschaftlich-technischen Fortschritt vornimmt, werden demnach mehr und mehr technische Systeme gebaut werden, welche das „Komplexitätsproblem" – also die Konfiguration, den Unterhalt, die Steuerung und die Kontrolle komplexer technischer Systeme – zu lösen versuchen. Das amerikanische Unternehmen IBM verwendet dafür in seinem *autonomic computing program* die Metapher des „autonomen Nervensystems": Was das autonome Nervensystem beim Menschen für das Funktionieren wichtiger Vorgänge in den Bereichen Verdauung, Homöostase oder Kreislauf- und Atmungssteuerung gewährleistet, soll *autonomic computing* dereinst für komplexe Rechnersysteme tun.

Neben diesen ökonomischen Aspekten sprechen auch grundsätzliche Überlegungen für eine Weiterentwicklung autonomer Systeme. Sie sind in der Lage, neue, von den Konstrukteuren nicht voraussehbare Lösungen für gewisse Probleme zu finden.[7] Autonome Systeme erhalten damit die Möglichkeit, auf nicht vorhersehbare Umweltbedingungen sinnvoll reagieren zu können. In den Neurowissenschaften wiederum erhalten autonome Systeme eine zunehmend wichtige Rolle als „verkörperte Modellsysteme", welche bei der Prüfung von Theorien der Verhaltenssteuerung Anwendung finden.[8]

Zusammenfassend lässt sich feststellen, dass sowohl die Motivation wie die Möglichkeit besteht, autonome technische Systeme zu bauen und sie Teil unserer technisierten Umwelt werden zu lassen. Als zentrales ethisches Problem stellt sich die Frage, wer oder was mögliche Fehlleistungen dieser Systeme zu verantworten hat. Die Untersuchung dieses Problems muss vielschichtig vonstatten gehen. Sie beinhaltet erstens eine Klärung des Begriffs „autonom" im Kontext technischer Systeme. Da der Begriff der „Autonomie" seit der Aufklärung in der westlichen Philosophie eine wichtige Rolle spielt und im direkten Bezug zum Problem der Freiheit von Handlungen und der Verantwortung für deren Folgen steht, muss der Begriff „Systemautonomie" sowohl konzeptionell als auch hinsichtlich der Eigenschaften, die autonome Systeme aufweisen sollten, geklärt und zur menschlichen Autonomie

[3] In einer Schrift zum *autonomic computing program* des amerikanischen Unternehmens IBM wird dieses Problem wie folgt beschrieben: „Consider this: at current rates of expansion, there will not be enough skilled I/T people to keep the world's computing systems running. [...] Even if we could somehow come up with enough skilled people, the complexity is growing beyond human ability to manage it." (IBM (2001): *Autonomic computing: IBM's Perspective on the State of Information Technology*, http://www.research.ibm.com/autonomic/manifesto/autonomic_computing.pdf, 5)

[4] Für eine Übersicht siehe: MATTERN, F. (2003): *Total vernetzt – Szenarien einer informatisierten Welt*, Berlin, Heidelberg.

[5] KING, R.D., WHELAN, K.E., JONES, F.M., REISER, P.G.K., BRYANT, C.H., MUGGLETON, S.H., KELL, D.B., OLIVER, S.G. (2004): *Functional genomics hypothesis generation and experimentation by a robot scientist*, in: Nature 427, 247-252.

[6] Das technische Potential der konventionellen Lithographie für die Herstellung von Computerchips wird in einigen Jahren ausgereizt sein. Als alternative Konzepte gelten die „molekulare Elektronik" (der Bau elektronischer Bauteile in Molekülgröße – der experimentelle „proof of concept" für wichtige elektronische Bauteile wie Schalter oder Transistoren wurde bereits erbracht), die „Spintronic" („Spintronic"-Bauteile benutzen den Spin von Elektronen als informationstragendes Element und ermöglichen die Reprogrammierung der Hardware eines Systems) und Quantencomputer (solche Systeme können weit mehr Zustände als konventionelle Computer für die Berechnung bestimmter Probleme, z.B. die Faktorisierung großer Zahlen, einnehmen).

[7] VAN DER VYVER, J.-J., CHRISTEN, M., STOOP, N., OTT, T., STEEB, W.-H., STOOP, R. (2004): *Towards genuine machine autonomy*, in: Robotics and Autonomous Systems 46 (3), 151-157.

[8] WEBB, B. (2002): *Robots in invertebrate neuroscience*, in: Nature 417, 359-363.

in Beziehung gesetzt werden. Zweitens wird es darum gehen festzustellen, welcher Begriff von „Verantwortung" diesem Problem angemessen ist. Dabei gilt es zu klären, inwiefern autonome Systeme zur Verwischung von Verantwortlichkeiten beitragen[9] und in welchem Sinn sie selbst Träger von Verantwortung sein können. Drittens wird sich die Frage stellen, welche Anforderungen man an den Bau und den Einsatz autonomer technischer Systeme zu stellen hat, um dem Menschen jenes Maß an moralischer Zuständigkeit zu geben, die sowohl noch möglich ist als auch erwünscht sein sollte.

Diese Arbeit hat folgenden Aufbau: In Abschnitt II wird auf den Begriff der Autonomie eingegangen. Zunächst wird der philosophische Autonomiebegriff eingeführt, wobei dieser mit den Begriffen „Willensfreiheit", „Handlung" und „Person" in Beziehung gesetzt wird. Danach folgt eine Einführung in den Begriff der „Systemautonomie". Maschinenautonomie ist der historisch ältere Begriff, umfasst aber nicht alles, was in dieser Arbeit unter autonomen Systemen verstanden wird (z.B. Software-Agenten), so dass er nicht verwendet wird. Abgeschlossen wird der zweite Abschnitt mit einer vergleichenden Betrachtung der beiden Autonomiebegriffe. In Abschnitt III wird zuerst kurz die aktuelle Debatte um den Verantwortungsbegriff beleuchtet, um danach jenen Begriff von Verantwortung herauszuschälen, der für den Problemkreis „Systemautonomie und Verantwortung" angebracht ist. Danach folgt eine Übersicht über jene Aspekte des Verantwortungsbegriffs, die durch Systemautonomie tangiert werden könnten. In Abschnitt IV sollen anhand konkreter Beispiele mögliche ethische Probleme vorgestellt werden, die sich durch den Einsatz autonomer technischer Systeme ergeben könnten. In Abschnitt V wird anhand von vier Problemfeldern aufgezeigt, welche Verantwortungsprobleme der Einsatz autonomer Systeme mit sich bringen kann und warum es sinnvoll sein kann, autonomen Systemen in gewissen Fällen Verantwortung zuzuschreiben. Schließlich wird die Frage behandelt, wie menschengerechte autonome Systeme aussehen sollten und welche praktischen Probleme dazu gelöst werden müssen.

II. Autonomie

Der philosophische Autonomiebegriff ist ein Kind der Neuzeit und spielt heute eine zentrale Rolle in der Ethik – sowohl hinsichtlich der Begründung des Verantwortungsbegriffs als auch hinsichtlich bioethischer[10] Anwendungen. Zuvor wurde der

[9] Auf dieses Problem hat bereits Hans Lenk hingewiesen: LENK, H. (1994): *Macht und Machbarkeit der Technik*, Stuttgart, 67.

[10] Vgl. dazu das Autonomieprinzip in der Medizinethik: BEAUCHAMP, T., CHILDRESS J. (2001): *Principles of Biomedical Ethics*, 5th ed., Oxford.

Begriff der Autonomie primär als politische Kategorie verwendet. Dieser beinhaltete die Möglichkeit politischer Gebilde, die eigenen, inneren Angelegenheiten unabhängig von einer äußeren Macht bestimmen zu können. Der juristische Autonomiebegriff bezieht sich auf die Möglichkeit der Selbstbestimmung natürlicher oder juristischer Personen im Rahmen einer rechtlich vorgegebenen Ordnung und ist damit inhaltlich von der jeweiligen Rechtstheorie abhängig.[11] Während der politische Autonomiebegriff für die Fragestellung dieser Arbeit ohne Belang ist („Maschinengesellschaften" im politischen Sinn wird es in absehbarer Zeit nicht geben), dürfte der juristische Autonomiebegriff tangiert werden. Ein Beispiel ist die Möglichkeit, dass der Gesetzgeber im Fall lernender technischer Systeme Grenzen definieren könnte, innerhalb welcher die Ergebnisse des Lernprozesses als systemimmanent gelten würden und demnach nicht durch den Konstrukteur des Systems zu verantworten wären.

Menschliche Autonomie

In philosophischer Hinsicht ist das Werk von Immanuel Kant zentral für das neuzeitliche Autonomiekonzept. Mit diesem Autonomiebegriff ist nicht nur die Forderung verbunden, der Mensch müsse sich nicht von fremden Autoritäten und von Traditionen bestimmen lassen. Autonomie wird auch als Selbstbestimmung des Menschen verstanden, indem sich der Wille ein eigenes Gesetz gibt. Autonomie wird also in direkten Bezug zur Selbstgesetzlichkeit des vernunftbegabten Individuums gesetzt und erhält damit in der Kantischen Philosophie den Status eines genuin ethischen Grundbegriffs. Tatsächlich verwendet Kant den Begriff der Autonomie erstmals in seiner *Grundlegung zur Metaphysik der Sitten*, wo er die Autonomie des Willens als das „alleinige Princip der Moral" charakterisiert.[12] Autonomie ist aber nicht Ausdruck einer grenzenlosen Freiheit des Individuums bei der Selbstgesetzgebung. Die Freiheit findet ihre Grenzen im kategorischen Imperativ.[13] Dieser gebietet dem Menschen bei der Bestimmung der eigenen moralischen Gesetze das Einnehmen einer „Dritt-Person-Perspektive", von welcher aus alle Menschen aus vernünftigen Gründen dem Gesetz zustimmen können. Ein wichtiger Aspekt des Kantischen Autonomieverständnisses ist zudem, dass Autonomie die Fähigkeit des freien Willens ausdrückt, die durch ihn verursachten Handlungen unabhängig vom Mechanismus der Naturkausalität ausüben zu können. Der Mensch ist für Kant also

[11] Die Autonomiebegriffe im Recht unterscheiden sich auch hinsichtlich der Fachgebiete. So steht im Privatrecht die „Privatautonomie" im Zentrum und im öffentlichen Recht der Begriff der „persönlichen Freiheit" (persönliche Mitteilung von M. Mastronardi, rechtswissenschaftliche Abteilung der Universität St. Gallen).

[12] KANT, I. (1785): *Grundlegung zur Metaphysik der Sitten*, in: KANT, I.: Werke in sechs Bänden, Bd. 3, Köln, Akad.-Ausg. 440.

[13] Ibid., Akad.-Ausg. 421.

nicht einfach ein Glied der Sinneswelt, dessen Handlungen „gänzlich dem Naturgesetz der Begierden und Neigungen, mithin der Heteronomie der Natur gemäß genommen werden müssen".[14] Autonomie im Sinn von Kant gibt dem Menschen die Möglichkeit, sich die wesentlichen Ziele und Normen des eigenen Lebens selbst zu setzen.

Dieses Autonomiekonzept ist mit zwei Fragen konfrontiert. Erstens: Wie ist die Freiheit des menschlichen Willens mit dem Wissen über naturgesetzliche Komponenten menschlicher Denkprozesse und Handlungen vereinbar? Zweitens: Wie lassen sich jene nichtnaturkausalen Grenzen der Autonomie verstehen, welche diese von der Willkür bzw. vom Zufall unterscheiden, so dass Autonomie in Verbindung mit einem System der Ethik gebracht werden kann? Die erste Frage führt zum klassischen Determinismusproblem und wird heute eng mit der Willensfreiheit in Verbindung gebracht. Die zweite Frage beinhaltet das Problem der Begrenztheit der Autonomiefähigkeit durch religiöse und soziale Aspekte. Das Spannungsverhältnis zwischen Autonomie und einer religiös fundierten Ethik wurde bald nach der Veröffentlichung der *Grundlegung* von einer Reihe von Philosophen betont, da Autonomie im Sinn von Selbstbestimmung durch Vernunft die Bindung des Menschen und seiner Ethik an Gott verneine.[15] Heutzutage werden die leiblichen und sozialen Dimensionen der Begrenztheit der menschlichen Autonomiefähigkeit betont.[16] Die Untersuchung dieser Einwände erlaubt eine Aufgliederung menschlicher Autonomiefähigkeit nach vier Aspekten, was nachfolgend geschehen soll.

Die Untersuchung des Zusammenhangs zwischen Autonomie, Willensfreiheit und Determinismus basiert auf einer Analyse von Michael Pauen.[17] Das Grundproblem des freien Willens besteht demnach in der Frage, wie sich das Individuum, das einen solchen durch einen Sprechakt oder eine Tat zu äußern glaubt, dessen Urheberschaft sicher sein kann. Woher nimmt eine Person, die vor einer Entscheidung zwischen zwei Möglichkeiten steht und die erste wählt, die Sicherheit, dass die Wahl „frei", durch einen autonomen und bewusst agierenden Urheber der Entscheidung, erfolgt ist? „Autonomie" bedeutet dabei, dass auf das Subjekt von außen einwirkende Momente für das Zustandekommen der Entscheidung nicht hinreichend sind. „Urheberschaft" bedeutet, dass die zur Handlung führende Entscheidung nicht vom Zufall abhängt, sondern Sache eines Akteurs ist, indem ausschließlich dem Subjekt zurechenbare Momente wie Handlungsdispositionen und Überzeugungen einen notwendigen Bestandteil des Zustandekommens einer Entscheidung bilden.

[14] Ibid., Akad.-Ausg. 453.

[15] So beispielsweise von Friedrich Heinrich Jacobi: JACOBI, F.H. (1926): *Über den Wert gewisser moralischer Räsonnements*, in: Die Schriften F.H. Jacobis, hg. von MATTHIAS, L., Berlin, 100-110.

[16] Vgl. als Beispiel: BECKER, B. (2003): *Zwischen Autonomie und Heteronomie. Zur Schwellensituation leiblicher Individualität*, in: CHRISTALLER, T., WEHNER, J. (Hg.): Autonome Maschinen, Wiesbaden, 56-68.

[17] PAUEN, M. (2001): *Freiheit und Verantwortung. Wille, Determinismus und der Begriff der Person*, in: Allgemeine Zeitschrift für Philosophie 26 (1), 23-44.

Pauen macht deutlich, dass die Aspekte „Autonomie" und „Urheberschaft" in einem Spannungsverhältnis stehen. Dies bedeutet insbesondere, dass „strenge Autonomie"[18] kein realistisches Konzept darstellt, da strenge Autonomie mit der Urheberschaft in Konflikt gerät: Während das Autonomieprinzip die Unabhängigkeit gegenüber sämtlichen Ausgangsbedingungen postuliert, muss dem Urheberprinzip zufolge eine Abhängigkeit von denjenigen Ausgangsbedingungen bestehen, die sich dem Akteur zurechnen lassen. Pauen schlägt als Ausweg aus diesem Dilemma vor, den Begriff der Freiheit mit dem Begriff der Person zu verknüpfen. Als „Personen" gelten in der philosophischen Literatur körperliche Wesen, die neben direkt präsenten Motiven wie Emotionen, Affekten und körperlichen Bedürfnissen auch intentionale Zustände wie Überzeugungen und Glaubenszustände haben. Diese Zustände sind Ausdruck der „personalen Geschichte" dieser Wesen und ihrer Lernleistungen, welche diese im Verlauf ihres bisherigen Lebens vollbracht haben. Sie bilden die personalen Merkmale dieser Person. Für Pauen ist demnach „das Tun einer Person [...] genau dann ‚frei', wenn es aus den personalen Merkmalen dieser Person hervorgeht."[19] Autonomie im Sinn von Willensfreiheit bedeutet demnach *nicht*, dass Entscheidungen und Handlungen voraussetzungsfrei sind: Zum einen spielen äußere Umstände als Faktoren in den Entscheidungsprozess ein, diese sind aber nicht hinreichend für die Entscheidung. Zum anderen spielen innere Momente, die sich als Folge der „personalen Geschichte" des Subjektes ergeben haben, eine notwendige Rolle beim Zustandekommen der Entscheidung.

Hier stellt sich nun die Frage, in welcher Beziehung die „personale Geschichte" zum Problem des Determinismus steht – also zur (vermutlichen) Tatsache, dass die Prozesse der Mikroebene, welche Erfahrungen, Lernleistungen und das Zustandekommen von intentionalen Zuständen ermöglichen, deterministischer Natur sind. Die Feststellung eines Determinismus darf dabei nicht darüber hinwegtäuschen, dass damit sehr schwierige Fragen einhergehen, etwa das Problem der Emergenz und Makrodetermination[20] und die Frage, inwieweit Determinismus auch Voraus-

[18] „Strenge Autonomie" bedeutet, dass eine vor zwei Handlungsalternativen stehende Person mit Willensfreiheit sich bei identischen Voraussetzungen hinsichtlich der äußeren und inneren Momente in einem Fall für die erste, im zweiten Fall für die zweite Alternative zu entscheiden vermag. Vgl. dazu ibid., 28-31.

[19] Ibid., 37. In der weiteren Diskussion geht Pauen der Frage nach, wie Handlungen als Folge von inneren Zwängen in dieses Bild eingepasst werden können. Er findet für diesen Fall eine Lösung, die hier aber nicht erläutert werden muss.

[20] HOYNINGEN-HUENE, P. (1994): *Zu Emergenz, Mikro- und Makrodetermination*, in: LÜBBE, W. (Hg.): Kausalität und Zurechnung. Über Verantwortung in komplexen kulturellen Prozessen, Berlin, 165-195.

sagbarkeit beinhalte.[21] Selbst die Realisierung sämtlicher personaler Merkmale durch deterministische Vorgänge auf der Mikroebene muss nicht zur Aufgabe des Vokabulars der Freiheit führen, da es durchaus möglich ist, dass zwischen den allgemeinen deterministischen Erklärungsmustern der Mikroebene und den personalen Merkmalen des Individuums kein geschlossener Zusammenhang hergestellt werden kann, der etwa eindeutig erklärt, warum eine Handlungsalternative gegenüber anderen vorgezogen wurde. Die „personale Geschichte" eines Individuums wird demnach schon aus rein praktischen Gründen primär aus biographischen Erzählungen bestehen, in welchen „deterministische" Zusammenhänge (z.B. Missbrauchserlebnisse als Erklärung von asozialem Verhalten) schwächer sind als in einem naturwissenschaftlichen Kontext.

Die Verteidiger der menschlichen Autonomiefähigkeit halten schließlich fest, dass Autonomie das selbstständige Verhalten im so genannten Raum der Gründe beinhaltet.[22] Autonomie kommt demnach insbesondere dann zum Tragen, wenn ein Individuum vor einem Problem steht, für welches es über einen gewissen Zeitraum beträchtliche intellektuelle und emotionale Ressourcen einsetzt, um zu einem Entscheid zu kommen, die es sowohl gegenüber sich selbst wie auch gegenüber seiner Umwelt begründen kann. Die Fähigkeit, solche „wichtigen Entscheide" auf eine überlegte Weise treffen zu können, ist demnach ein Kernmerkmal der menschlichen Autonomiefähigkeit. Sie muss von der Tatsache, dass viele Menschen alltägliche Handlungen und Entscheidungen nicht in diesem überlegten Sinn umsetzen, unterschieden werden.

Zusammenfassend lassen sich vier charakteristische Aspekte der menschlichen Autonomiefähigkeit festhalten:

- Autonomie beinhaltet die Fähigkeit, wichtige Entscheide hinsichtlich Lebensplanung und Setzung eigener Grundsätze treffen zu können.
- Die Wahrnehmung der Autonomiefähigkeit beruht auf der personalen Geschichte, in welcher sich Erfahrungen und Lernerlebnisse spiegeln, die durch die Interaktion der Person mit ihrer Umwelt geschaffen wurden.
- Die konkrete Wahrnehmung der Autonomiefähigkeit zu einem bestimmten Zeitpunkt wird durch die an diesem Zeitpunkt wirkenden äußere Umstände sozialer (und anderer) Natur eingeschränkt.
- Die (wahrscheinlich deterministischen) Vorgänge, welche die personalen Merkmale einer Person realisieren, sind von einer Natur, die sichere Voraussagen über die Ergebnisse von Entscheidungen bzw. über Handlungen nicht erlaubt.

[21] Die Theorie dynamischer Systeme macht deutlich, dass ein solcher Zusammenhang in vielen natürlichen Systemen selbst dann nicht gegeben wäre, wenn die Menge aller das System beschreibenden Differenzialgleichungen bekannt wäre.

[22] STURMA, D. (2003): *Über Personen, Künstliche Intelligenz und Robotik*, in: CHRISTALLER, T., WEHNER, J. (Hg.): Autonome Maschinen, Wiesbaden, 38-55.

Haben Handlungen von Menschen Folgen, die zu einer Diskussion über Verantwortung Anlass geben, so wird in der Regel abgeschätzt, in welchem Grad diese vier Aspekte in die Handlung eingeflossen sind. Es werden also Antworten auf folgende Fragen gesucht: Wie überlegt war die Entscheidung? Welche biographischen Elemente spielten bei der Entscheidung eine Rolle? Wie präsentierte sich die Situation zum Zeitpunkt der Entscheidung? Liegen Sachverhalte (z.B. gewisse Hirnschäden) vor, welche die Autonomiefähigkeit vermindern? Die Beurteilung des Vorhandenseins von Autonomie bei einem Verantwortungsproblem hat also einen graduellen Charakter. Jetzt wird es darum gehen, Systemautonomie zu charakterisieren und mit dem obigen Schema in Verbindung zu bringen.

Systemautonomie

Die Rede von System- oder Maschinenautonomie ist Teil eines Trends, früher nur Menschen vorbehaltene Begriffe auf technische Systeme und andere Lebensformen zu übertragen. Beispiele sind „Intelligenz", „Handlung" und eben auch „Autonomie". Drei Standpunkte lassen sich in dieser Debatte feststellen[23]: Eine Position beharrt auf einem kategorialen Unterschied zwischen menschlicher Autonomie und Systemautonomie und diagnostiziert im herrschenden Sprachgebrauch eine Mystifizierung neuer Technologien, einhergehend mit einem simplifizierten Menschenbild. Die zweite Position besagt, dass der klassische Maschinenbegriff auf die heute in Entwicklung stehende Technologie nicht mehr anwendbar ist. So werden in diesen Systemen Aspekte wie Lernfähigkeit, Selbstreparatur und senso-motorische Rückkopplungen implementiert, welche diesen eine bisher unerreichte Form von Autonomie geben. Zumindest prinzipiell können diese technischen Systeme einen Komplexitätsgrad erreichen, der mit demjenigen „biologischer Maschinen" vergleichbar ist. Die Anwendung einer anthropomorphen Begrifflichkeit ist deshalb nicht nur aus pragmatischen Gründen gegeben. Sie spiegelt vielmehr eine technologische Entwicklung wider, an deren Endpunkt tatsächlich technische Systeme stehen könnten, die bisher Menschen vorbehaltene Eigenschaften und Fähigkeiten haben. Eine dritte Position schließlich verweist darauf, dass Mensch und Maschine nicht mehr in einem Subjekt-Objekt-Verhältnis stehen, sondern in der modernen Zivilisation in eine neue Form von Kooperation treten. Handlungen in dieser Welt sind demnach nicht Ergebnis autonomer Willensentscheide Einzelner, sondern unterliegen einem komplizierten Wechselspiel zwischen technischen und sozialen Systemen. Autonomie wird damit zu einer empirisch bestimmbaren Eigenschaft einzelner Systemteile.

Diese drei Positionen haben einen unterschiedlichen Blickwinkel auf das Problem der Systemautonomie – einen logisch-begrifflichen, einen technisch-empirischen und einen sozialen – und schließen sich nicht gegenseitig aus. Das Problem soll

[23] CHRISTALLER, T., WEHNER, J. (2003): *Autonomie der Maschinen – Einführung in die Diskussion*, in: CHRISTALLER, T., WEHNER, J. (Hg.): Autonome Maschinen, Wiesbaden, 9-35.

nachfolgend aus allen drei Perspektiven beleuchtet werden: Erstens werden die Maschinenbegriffe eingeführt, welche in der Diskussion gebraucht werden. Zweitens wird eine Übersicht der verschiedenen Eigenschaften gegeben, welche autonome Systeme nach Ansicht ihrer Entwickler aufweisen sollten. Daraus ergibt sich ein Begriff von Systemautonomie, der mit dem oben entwickelten Konzept menschlicher Autonomie verglichen werden kann. Die soziale Perspektive – insbesondere die Frage nach den „Handlungen" technischer Systeme – wird im Abschnitt „Verantwortung" behandelt.

Die konventionelle Definition von Maschinen charakterisiert diese durch drei Eigenschaften[24]: Sie sind Kreationen eines Erfinders, sie sind gebaut für einen bestimmten Zweck und sie funktionieren schließlich auch diesem Zweck gemäß. Diese Maschinendefinition ist durchaus vereinbar mit dem philosophischen Maschinenbegriff von René Descartes, der diesen auch auf unbewusste biologische Systeme anwendete. So sind Tiere komplizierte Maschinen, deren Designer Gott ist. Selbst der menschliche Körper (die *res extensa*) gilt als Maschine, und erst die *res cogitans* beseelt gewissermaßen die Maschine und macht diese zum freien Menschen. Geht man davon aus, dass der Maschinenbegriff untrennbar mit der Idee eines Konstrukteurs (sei dies nun der Mensch oder Gott) verbunden ist, so lassen sich für verschiedene Epochen „paradigmatische Maschinen" finden: Eine der ersten, wichtigen neuzeitlichen Maschinen war die Uhr, welche selbst für die Kosmologie als Modellsystem gegolten hat. Die industrielle Revolution hat dann die Maschinenwelt in einem bisher nicht gekannten Ausmaß bevölkert: Dampfmaschinen und Spinnmaschinen lassen sich als Modellmaschinen dieser Zeit bezeichnen.

Der Evolutionsgedanke ermöglichte hingegen eine andere Sichtweise auf lebende Systeme, wonach diese nicht mehr als von Gott erschaffene Maschinen, sondern als im Evolutionsprozess entstandene Organismen verstanden werden können. Das Konzept der Evolution erlaubte damit die Ausweitung des Maschinenbegriffs, indem Maschinen Eigenschaften biologischer Systeme (Selbstreparatur, Lernfähigkeit, Selbstorganisation) gegeben werden. Solche Maschinen werden zwar zu Beginn konstruiert, erfahren aber später eine durch das Wirken der Maschine in der Welt verursachte „Evolution". Organismen können damit erneut in ein Maschinenbild gepresst werden – nun aber nicht mit Gott als Konstrukteur, sondern als „kybernetische Maschinen". Dabei rückt die Frage nach der Kontrolle und dem Informationsfluss in solchen Maschinen ins Zentrum des Interesses. Es ist deshalb nicht verwunderlich, dass der Computer als „informationsverarbeitende Maschine" die Modellmaschine des 20. Jahrhunderts geworden ist. Die wissenschaftlichen Erfolge der Molekularbiologie in jüngster Zeit verstärken das Bild eines Organismus als „komplexe Maschine". Die von Wissenschaftlern publik gemachte Absicht, eine primitive Zelle zu bauen, zeigt, wohin der Weg geht: In den nächsten Jahren und Jahrzehnten dürfte es gelingen, erste, sehr primitive Organismen aus Biomolekülen

[24] SCHMIDT-BIGGEMANN, W. (1980): *Maschine*, in: RITTER, J. (Hg.): Historisches Wörterbuch der Philosophie, Bd. 5, Basel, Sp. 790-802.

zu bauen. Es ist anzunehmen, dass eine solche Lebensform als Modellmaschine des 21. Jahrhunderts gelten wird.[25]

Mehrere Exponenten der Systemautonomie vertreten einen Maschinenbegriff, der biologische Systeme mit einschließt und einen kategorialen Unterschied zwischen Mensch und Maschine verneint.[26] Dies macht es verständlich, warum ihnen der Begriff „Autonomie" im Fall technischer Systeme so leicht über die Lippen geht. Die Frage ist nun, wodurch sich die Autonomie dieser Systeme auszeichnet. Antworten finden sich bei der Analyse jener Gebiete, wo Systemautonomie in besonderem Maß untersucht bzw. zu realisieren versucht wird: in der Robotik und der Technologie der Software-Agenten.

Die Robotik hat im Zug der Entwicklung der *New Artificial Intelligence*[27] der Systemautonomie eine wichtige Rolle zugeordnet, welche als Leitvorstellung für die Entwicklung künftiger Robotsysteme dienen soll. Hervorgehoben wurde dabei insbesondere die Bedeutung der „Verkörperung" (*embodiment*), der Selbstversorgung (*self-sufficiency*, vor allem hinsichtlich der Energieversorgung, z.B. durch das Aufsuchen einer Ladestation) und der Interaktion der Roboter mit einer konkreten Umwelt (*situatedness*). Ein Wesensmerkmal der Autonomie ist, dass sich diese Roboter ohne explizite äußere Kontrolle in einer gegebenen Umwelt bewegen können sollen.[28] Diese Umwelt ist dabei keine strukturierte technische Umgebung (wie z.B. ein Fließband), sondern eine Umwelt, in der Maschinen entweder mit anderen Maschinen (ein Experimentierfeld ist die jährliche Roboter-Fußballweltmeisterschaft *RoboCup*) oder mit Menschen (Spielzeug- und Serviceroboter) interagieren.[29] In diesen Umgebungen sollen die Systeme Planungsleistungen vollbringen, um die von außen definierten Zielvorgaben selbstständig umsetzen zu können.[30] Autonome Systeme sollen demnach flexibler sein und werden damit – ein weiteres Wesensmerkmal der Systemautonomie – unvorhersagbarer.[31] Hinsichtlich der Eigenschaften autonomer Systeme sollen diese Fähigkeiten erhalten, welche bisher nur biologischen Systemen zugesprochen wurden – explizit genannt seien Lernfähigkeit (wobei

[25] Vgl. SZOSTAK, J.W., BARTEL, D.P., LUISI, P.L. (2001): *Synthesizing life*, in: Nature 409, 387-390.

[26] Als Beispiel siehe: BROOKS, R. (2002): *Flesh and Machines*, New York. Zur Frage nach kategorialen Unterschieden zwischen natürlichen und künstlichen Körpern siehe: CHRISTEN, M. (2003): *Die Ontologie künstlicher Körper*, Studia Philosophica 63, 65-82.

[27] Exemplarisch dazu: PFEIFER, R., SCHEIER, C. (1999): *Understanding intelligence*, Cambridge.

[28] PFEIFER, R. (2003): *Körper, Intelligenz, Autonomie*, in: CHRISTALLER, T., WEHNER, J. (Hg.): Autonome Maschinen, Wiesbaden, 137-159.

[29] VELOSO, M.M. (2002): *Entertainment Robotics*, in: Communications of the ACM 45 (3), 59-63 (ACM =Association for Computing Machinery).

[30] LEVI, P. (1996): *Robotik*, in: STRUBE, G. (Hg.): Wörterbuch der Kognitionswissenschaft, Stuttgart, 582-596.

[31] DÖRNER, D. (2003): *Autonomie*, in: CHRISTALLER, T., WEHNER, J. (Hg.): Autonome Maschinen, Wiesbaden, 112-136.

die Bewertungskriterien für den Lernerfolg vorgegeben werden)[32], situativ angepasste Veränderung des Robot-Körpers[33], Selbstreparatur und dereinst wohl auch Formen der Replikation.[34] Damit soll solchen Systemen die Möglichkeit einer „Ontogenese" (Lernprozesse beinhaltend) bzw. einer „Phylogenese" (eine Art von Evolution, etwa gemäß dem Modell der genetischen Algorithmen, wobei aber Zielvorgaben im Sinn einer „Fitnessfunktion" vorgegeben werden) gegeben werden.[35] Schließlich sollen autonome Systeme auch mit anderen autonomen Systemen kooperieren, um gewisse Aufgaben im Verbund durchführen zu können.[36] Eine vergleichbare Charakterisierung gilt für Software-Agenten – abgesehen von jenen Aspekten, die nur verkörperten Systemen zugänglich sind.

Zwei Aspekte fallen bei dieser Auflistung auf: Zum einen die bereits festgestellte Übertragung einer menschlichen (bzw. biologischen) Begrifflichkeit auf technische Systeme, ohne dass jeweils explizit gemacht wird, dass die Begriffe unterschiedliche semantische Felder haben („lernen" bei einem Kind bedeutet etwas anderes als „lernen" bei einem neuralen Netz). Zum anderen die Tatsache, dass die Zwecke der Aktivitäten des Systems (der Auftrag, die Kriterien des Lernerfolges, die Fitnessfunktion) diesem weiterhin von außen vorgegeben werden.

Um dem Problem einer anthropomorphen Begrifflichkeit auszuweichen, wird nachfolgend eine formalere Definition von Systemautonomie gegeben.[37] Klassische Maschinen/Systeme lassen sich im Bild der *Finite State Automata (FSA)* beschreiben, dem Automatenmodell der Informatik.[38] Demnach haben solche Systeme eine definierte, endliche Zahl innerer Zustände; durch ihre Konstruktion/Programmierung sind (möglicherweise probabilistische) Regeln implementiert, die den Wechsel von einem Zustand in einen anderen bestimmen. Gegeben ist dabei ein bestimmter innerer Zustand sowie ein Input von der Umwelt des Systems. Der Konstrukteur des Systems muss demnach eine Vorstellung der möglichen Inputs des Systems haben und die Regeln vorgängig festlegen, damit das System funktionieren kann. Ein solches System ist nicht autonom. Ein Kernpunkt der Systemautonomie besteht darin, dass dem System weitere potentielle innere Zustände zur Verfügung stehen und dass das System durch einen Lernprozess Zugang zu diesen Zuständen erhält, indem es neue interne Regeln schafft. Die Menge der inneren Zustände des Systems

[32] SCHAAL, S. (2002): *Learning Robot Control*, in: ARBIB, M. (ed.): Handbook of Brain Theory and Neural Networks, Cambridge, London, 983-987.

[33] RUS, D., BUTLER, Z., KOTAY, K., VONA, M. (2002): *Self-reconfiguring robots*, in: Communications of the ACM 45 (3), 39-45.

[34] Erste theoretische Grundlagen dazu wurden bereits von John v. Neumann geleistet: V. NEUMANN, J. (1966): *The Theory of Self-Reproducing Automata*, ed. and completed by BURKS, A.W., Urbana, London.

[35] PFEIFER 2003, 140-142.

[36] LEVI 1996, 583.

[37] VAN DER VYVER et al. 2004.

[38] SIPSER, M. (1997): *Introduction to the Theory of Computation*, Boston.

nimmt damit während dessen „Lebensdauer" zu. Je nach struktureller Komplexität des Systems kann diese Zahl der inneren Zustände sehr groß werden, wobei sich dann natürliche Klassen solcher Zustände bilden, die mit dem Verhalten des Systems in Bezug gesetzt werden können. Dies ist die theoretische Basis der Eigenschaften autonomer Systeme wie Lernfähigkeit und Flexibilität. Aus diesem Grund sind autonome Systeme aber auch unvorhersehbarer. Dieses Problem wird beim Bau autonomer Systeme oft unterschätzt – insbesondere bei der Konstruktion einer Kommunikationsschnittstelle zum menschlichen Nutzer. Besteht diese aus einer vordefinierten Menge an Zeichen, können diese möglicherweise den wahren inneren Zustand des „evolvierten" Systems nicht mehr adäquat wiedergeben. Eine weitere Verschärfung erfährt dieses Problem bei „verteilten Systemen", bei welchen ein globaler Zustand nicht mehr definiert werden kann.[39] Beide Aspekte tragen dazu bei, dass Emergenz bei autonomen Systemen erwartet werden kann – d.h. das Auftreten unerwarteter Eigenschaften auf der Makroebene trotz umfassender Kenntnis über die Vorgänge auf der Mikroebene.[40]

Was ergibt der Vergleich zwischen menschlicher Autonomie und Systemautonomie? Ausgehend vom erarbeiteten Raster lassen sich folgende Feststellungen machen:

- Die deterministischen Vorgänge, welche den Aktionen autonomer Systeme zugrunde liegen, sind von einer Natur, die sichere Voraussagen über die Aktionen der Systeme erschweren.
- Die Wahrnehmung der Autonomiefähigkeit zu einem gewissen Zeitpunkt wird durch äußere Umstände eingeschränkt, wobei diese Umwelt aus der Sicht des Systems weniger „reich" ist als die Umwelt aus der Sicht des menschlichen Nutzers des Systems.
- Die Wahrnehmung der Autonomiefähigkeit beruht auf der Geschichte des Systems, in welcher sich Lernerlebnisse spiegeln, die durch die Interaktion des Systems mit seiner Umwelt geschaffen wurden.
- Autonomie hinsichtlich der Zwecke, für welche das System eingesetzt wird, besteht für das System nicht.

Die „Unfreiheit" autonomer Systeme hinsichtlich der Zwecke, für welche sie eingesetzt werden, zeigt auf, dass das erste Element der klassischen Maschinendefinition weiterhin gültig ist. Autonome Zwecksetzung erreichen diese Systeme erst dann, wenn sie gewisse Formen von Bewusstsein realisieren können, die sich in intentionalen Zuständen niederschlagen und eine Form von Willensfreiheit ermöglichen.

[39] LAMPORT, L. (1977): *Time, clocks, and the ordering of events in a distributed system*, in: Communications of the ACM 21 (7), 558-565.

[40] „Emergenz" ist ein Begriff, der in der Theorie komplexer Systeme oft angetroffen wird, aber unterschiedlich verstanden werden kann. Eine allgemein anerkannte Definition bzw. Theorie von „Emergenz" ist derzeit nicht auszumachen.

Drei große Schwierigkeiten stehen diesem Unterfangen entgegen: Die erste ist das Fehlen einer Theorie des Bewusstseins, da weiterhin kontrovers diskutiert wird, was „Bewusstsein" wirklich ist und welche materialunabhängige Struktur der Materie für die Realisierung von Bewusstsein notwendig ist. Die zweite, daran gekoppelte Schwierigkeit besteht in der Festlegung eines gültigen Tests für die Zuschreibung von Bewusstseinszuständen an die fraglichen Systeme. So werden Eigenschaften als mögliche Konstituenten von Bewusstsein genannt (die „Qualia"), welche einem empirischen Test grundsätzlich nicht zugänglich sind. Drittens zeigt sich bei den bisherigen „Kandidaten" von bewussten Systemen (sicher der Mensch, wahrscheinlich einige Tiere), dass die Materie, welche Bewusstsein hervorbringt, außerordentlich komplex organisiert ist – um viele Größenordnungen komplexer als jene von Maschinen.[41] Aus diesen Gründen scheint es derzeit müßig, über die Möglichkeit von Bewusstsein, Personsein oder Willensfreiheit bei autonomen Systemen zu diskutieren. Die entscheidende Frage ist nun aber, ob sich damit das Verantwortungsproblem erledigt.

III. Verantwortung

Der Verantwortungsbegriff hat in der angewandten Ethik eine große Konjunktur.[42] Dies zeigt sich insbesondere in der Technikethik, welche die moralischen Konsequenzen der zunehmenden Handlungs- und Eingriffsmacht des Menschen in seine natürliche wie kulturelle Umwelt thematisiert. Grundsätzlich geschen steht menschliches Handeln dank seiner technischen Hilfsmittel am Beginn einer zunehmenden Zahl von Kausalketten in der Welt, welche zu Ergebnissen führen, die als ethisch problematisch bewertet werden können. Dies mag mit ein Grund sein, warum der Begriff der Verantwortung (bzw. *responsibility*, *responsibilité*) im Abendland erst in der Neuzeit auftaucht, und dies zuerst – analog zum Begriff der Autonomie – im Kontext der Politik und des Rechts. Die Idee der Verantwortung ist an die Idee geknüpft, für negativ bewertete Handlungsfolgen einen Schuldigen benennen zu können. Da durch die Aufklärung die Welt „freier" geworden ist, indem die in ihr stattfindenden Geschehnisse nicht mehr in einen göttlichen Sinnzusammenhang gestellt werden, sondern als Folgen von menschlichen Handlungen gelten, wächst das Bedürfnis, jemanden „zur Verantwortung ziehen" zu können. Im 19. Jahrhundert ist dann „Verantwortung" auch in der Philosophie zum Fachbegriff geworden.[43] Heute hat der Verantwortungsbegriff eine mannigfaltige Ausprägung erhalten, und viele Philosophen wie Theologen haben unterschiedliche Schwerpunkte gesetzt.[44] Trotz dieser breiten Diskussion lassen sich allgemeine Charakteristika von „Verantwortung" hinsichtlich des Begriffs und seiner Anwendung herausschälen, was im Folgenden geschehen soll.[45]

[41] Vgl. dazu MEDALIA, O., WEBER, I., FRANGAKIS, A.S., NICASTRO, D., GERISCH, G., BAUMEISTER W. (2002): *Macromolecular architecture in eukaryotic cells visualized by cryoelectron tomography*, in: Science 298, 1209-1213.

[42] Die Rede ist in diesem Abschnitt vom moralischen Verantwortungsbegriff, im Unterschied zur Verwendungsweise von „verantwortlich sein für" im Sinn von „zuständig sein für".

Charakterisierung von Verantwortung

Was den Begriff „Verantwortung" selbst betrifft, herrscht Einigkeit dahingehend, dass dieser ein Beziehungsbegriff ist, in welchem moralische Subjekte, Objekte der Verantwortung, Beziehungsmaßstäbe etc. zueinander in Relation gebracht werden. Begriffstheoretisch ist „Verantwortung" ein mehrstelliges Prädikat, wobei hinsichtlich der genauen Anzahl der Variablen unterschiedliche Vorstellungen herrschen. Für diese Untersuchung wird die Systematik von Günter Ropohl verwendet, der Verantwortung als siebenstelliges Prädikat auffasst, das sich wie folgt darstellen lässt[46]:

A verantwortet B für C wegen D vor E zum Zeitpunkt F im Sinn von G.

Die einzelnen Variablen bedeuten: A ist das *Subjekt* der Verantwortung, B ist das *Objekt* der Verantwortung, C ist der *ethisch relevante Aspekt* des Objekts der Verantwortung, D ist das *Kriterium* der Verantwortungsbewertung, E ist die *Instanz* der Verantwortungsbeurteilung, F ist die Zeitspanne der *Zurechenbarkeit* von Verantwortung, und G ist der psychologische Aspekt der *Betroffenheit/Handlungswirksamkeit* von Verantwortung. Diese durch Ropohl gegebene Aufschlüsselung erlaubt eine detaillierte und präzise Untersuchung der Frage, inwieweit autonome Systeme Gegenstand des Verantwortungsproblems sind.

Vorab aber noch einige Bemerkungen zu den allgemein anerkannten Charakterisierungen des Verantwortungsbegriffs: Hinsichtlich dessen *Anwendung* herrscht Einigkeit, dass sich dieser auf Handlungen bezieht (die Variable B in Ropohls Systematik), wobei sich aber (wie nachfolgend deutlich wird) die Frage nach dem adäquaten Handlungsbegriff stellt. Der *zeitliche Kontext* des Verantwortungsbegriffs betrifft nicht nur die Vergangenheit („Wer ist verantwortlich, dass X geschehen ist?"), sondern auch die Zukunft („Wir sind verantwortlich dafür, dass X geschehen

[43] Vgl. dazu BAYERTZ, K. (1995): *Eine kurze Geschichte der Herkunft der Verantwortung*, in: BAYERTZ, K. (Hg.): Verantwortung. Prinzip oder Problem?, Darmstadt, 3-71.

[44] Für einen umfassenden Überblick siehe BANZHAF, G. (2002): *Philosophie der Verantwortung*, Heidelberg.

[45] Diese Zusammenstellung basiert auf der in diesem Abschnitt zitierten Literatur, falls nicht anders erwähnt.

[46] ROPOHL, G. (1996): *Ethik und Technikbewertung*, Frankfurt a.M., Kap. 3.

und Y nicht geschehen wird."). Insofern sind auch Unterlassungen Thema der Verantwortungsdiskussion. Einigkeit herrscht auch dahingehend, dass zwischen *deskriptiver und normativer Verantwortung* unterschieden werden muss: Erstere ergibt sich aus der Beobachtung des Zustandekommens einer verantwortungsethisch problematisierten Situation (die Ermittlung der Kausalkette). Als praktisches Problem stellt sich hier die Frage, ob diese Kausalkette hinreichend ermittelt werden kann, damit Subjekt und Objekt der Verantwortung zueinander in Beziehung gebracht werden können. Die normative Verantwortung ergibt sich als Folge einer vorgängig erfolgten *Zuschreibung* von Verantwortung. Hinsichtlich der normativen Verantwortung werden allgemein moralische Formen der Zuschreibung von rechtlichen Formen der Zuschreibung von Verantwortung unterschieden. Schließlich herrscht Einigkeit darin, dass Theorien über Verantwortung andere Ethiken nicht ersetzen können. Vielmehr entstammen die *Bewertungsmaßstäbe* (die Variable D in Ropohls Systematik), an welchen sich die (potentiellen) Ergebnisse von Handlungen bemessen, aus anderen ethischen Theorien.[47]

Systemautonomie und Verantwortung

Mit Hilfe der Systematik von Ropohl lässt sich nun der Zusammenhang zwischen Systemautonomie und Verantwortung detailliert aufschlüsseln:

- Subjekte der Verantwortung (Variable A) handeln. Was genau bedeutet Handeln, und ist Handlungsfähigkeit ein hinreichendes Kriterium, um als Subjekt von Verantwortung gelten zu können? Hier stellen sich die Fragen, ob die Aktivitäten autonomer Systeme überhaupt als Handlungen gelten können, und – falls nicht – ob damit auch die Möglichkeit einer Zuschreibung von Verantwortung an autonome Systeme entfällt.
- Wie entstehen Objekte der Verantwortung (Variable B) in komplexen Systemen? Hier stellt sich das Problem des Beitrags autonomer Systeme zu „kollektiven Aktivitäten", die Gegenstand einer Verantwortungsbeurteilung sind.
- Welche Entitäten können Subjekte ethischer Theorien werden (Variablen C/D), und aus welchen Gründen? Fraglich ist hier, ob autonome Systeme in irgendeinem ethisch relevanten Sinn „geschädigt" werden können.
- Wer ist die legitime Instanz der Beurteilung des Verantwortungsproblems (Variable E)? Hier geht es insbesondere um die Beziehung zwischen Konstrukteur und Käufer autonomer Systeme und um die Anpassung des Rechtssystems hinsichtlich der Beurteilung negativer Folgen der Aktionen autonomer Systeme.

[47] Vgl. dazu: WIELAND, W. (1999): *Verantwortung – Prinzip der Ethik?*, Heidelberg.

- Welche Grenzen der Zurechnungsfähigkeit von Handlungen/Aktionen bestehen in zeitlicher Hinsicht (Variable F)? Diese Frage ist bei lernenden autonomen Systemen von Bedeutung.
- Welche Auswirkungen hat die Einbindung autonomer Systeme in soziale/ökonomische Prozesse hinsichtlich der Wirksamkeit des Verantwortungsappells bei Menschen?

Antworten auf diese Fragen sollen im Folgenden unter Einbezug der Ergebnisse des Abschnitts „Autonomie" gegeben werden.

Variable A: Aus der in der ersten Frage angesprochenen Problematik lässt sich leicht ein Argument gegen die Verbindung von Systemautonomie und Verantwortung extrahieren: Da ein Objekt der Verantwortung nur durch Handlungen oder Unterlassung von Handlungen entstehen könne und autonome Systeme nicht handeln können, seien diese kein Subjekt der Verantwortung und könnten auch kein Objekt der Verantwortung generieren. Ausgehend vom klassischen, auf Aristoteles zurückgehenden Handlungsbegriff ist dieses Argument plausibel. Demnach beinhaltet der Begriff „Handlung" eine reflektierte, planmäßige und zielgerichtete Aktivität, welche ein Bewusstsein des Handelnden voraussetzt.[48] Die Zielgerichtetheit einer Aktivität allein macht diese demnach noch nicht zu einer Handlung. So nennt man die zielgerichteten Aktivitäten von Tieren „Verhalten", und in der anorganischen Natur nennt man das Analogon „Prozess".

Ein erster Einwand gegen dieses Argument besteht in der Behauptung, dass autonome Systeme intentionale Zustände hätten, welche reflektierte, planmäßige und zielgerichtete Aktivität erlaubten, so dass diese Systeme im klassischen Sinn handeln würden. Hier gerät man aber in das bereits angesprochene Diskussionsfeld des „künstlichen Bewusstseins" mit all seinen Problemen, so dass dieses Gegenargument (derzeit) nicht als schlüssig betrachtet werden kann.

Eine zweite Gegenstrategie besteht in der Abschwächung des Handlungsbegriffs. So stellt sich die Frage, wie man die Aktivitäten autonomer Systeme denn genau bezeichnen sollte. Von „Verhalten" zu sprechen, erscheint insofern seltsam, da autonomen Systemen die beiden wesentlichen Triebkräfte tierischen Verhaltens – Nahrungssuche und Fortpflanzung – bisher schwerlich zugeordnet werden können.[49] Aber auch der Begriff „Prozess" scheint insofern unangebracht, da man bei Prozessen „Lernfähigkeit" im Sinn der Schaffung neuer innerer Zustände des Systems nicht erwartet. Man scheint sich fast genötigt zu fühlen, einen weiteren Begriff einzuführen, der die spezifischen Aspekte der Aktivitäten autonomer Systeme sowie deren psychologische Einschätzung durch den Nutzer erfasst. Zumindest im alltäglichen Sprachgebrauch der Robotiker scheint sich der Begriff „Handeln" für die

[48] ARISTOTELES: *Nikomachische Ethik*, Stuttgart 1992, insbesondere Buch II, Abschn. 3, Buch III, Abschn. 1-4, 8, Buch V, Abschn. 11, Buch VI, Abschn. 9, 13.

[49] Hier muss erwähnt werden, dass das Prinzip der *self-sufficiency* zumindest die Integration der „Nahrungssuche" in (verkörperten) autonomen Systemen verlangt.

Aktivitäten humanoider Roboter eingebürgert zu haben.[50] Ein solcher psychologischer Aspekt ist zwar kein ausreichender Grund für die Aufgabe des klassischen Handlungsbegriffes, wird aber für die nachfolgende Beurteilung der Variable G sicher eine Rolle spielen. Außerdem muss festgehalten werden, dass viele Aktivitäten von Menschen durchaus als Handlungen gelten, ohne dass man diesen Aktivitäten gleich Reflektiertheit, Planmäßigkeit oder Zielgerichtetheit attestiert – zumindest kommen graduelle Abstufungen von der „idealen, bewussten Handlung" vor. Mit Blick auf die soziale Realität ist man demnach oft genötigt, einen „weichen Handlungsbegriff" zu verwenden. Dieser zeichnet sich dadurch aus, dass Handlungen nur ungenügend reflektiert sind, auf *ad-hoc*-Planungen beruhen und dass die Zielausrichtung der Handlung unklar ist. Dieser „weiche Handlungsbegriff" führt zum dritten Einwand.

Diese dritte Gegenstrategie besteht in der Abschwächung des Zusammenhangs zwischen „Handeln" und „Verantwortung". Eine solche Argumentation kann sich auf die Tatsache stützen, dass die Verantwortungsfrage in der Praxis auch dann gestellt wird, wenn das Objekt der Verantwortung nicht Ergebnis einer Handlung im klassischen Sinn ist – also nicht Ergebnis einer reflektierten, planmäßigen und zielgerichteten Aktivität. Paradebeispiel sind Unfälle, bei welchen zwar abschwächende Faktoren bei der Verantwortungsbeurteilung zum Tragen kommen können, womit man aber dem Agierenden den Status eines Verantwortungssubjekts nicht *a priori* abspricht, auch wenn die Aktion nicht als Handlung im klassischen Sinn gilt. Demnach können auch Aktionen autonomer Systeme, die nicht als Handlungen im klassischen Sinn zu werten sind, Gegenstand einer Verantwortungsbeurteilung sein. Dieser Sachverhalt ist eine hinreichende Entkräftung des Arguments eines prinzipiellen Verbots der Verknüpfung von Verantwortung und Systemautonomie. Autonome Systeme werden damit nicht zu Subjekten der Verantwortung, hingegen sind sie legitime Objekte einer Diskussion um *Zuschreibung* von Verantwortung.

Variable B: Die Frage einer möglichen Zuschreibung von Verantwortung an autonome Systeme muss mit dem Problem der Entstehung von Objekten der Verantwortung in komplexen Systemzusammenhängen verknüpft werden. Solche Systemzusammenhänge sind beim Einsatz autonomer Systeme zu erwarten – schließlich werden autonome Systeme ja gerade zum Zweck entwickelt, in für Menschen nicht mehr überschaubaren technischen Systemen aktiv zu werden. Damit einher geht eine Verschärfung des Problems, die Verantwortlichkeit für Folgen kollektiver oder korporativer Handlungen zu bestimmen. Solche Folgen sind Ergebnis ganzer „Handlungsnetze" mit unterschiedlichem Grad von Reflektiertheit, Planmäßigkeit und Zielgerichtetheit der einzelnen Handlung. Nachfolgend wird also der „weiche" Begriff von Handlung verwendet. Die damit verbundenen Fragen nach der Möglichkeit „kollektiver Verantwortung" oder gar irreduzibler „Systemverantwortung"

[50] Vgl. dazu BROOKS 2002.

sind seit längerem Gegenstand der Debatte.[51] Soziologen ergänzen diese Debatte mit dem Hinweis auf die starke Verschränkung menschlicher Aktivitäten und technischer Prozesse bei sozialen Handlungen in modernen Gesellschaften.[52]

Derartige Diagnosen neigen dazu, die Debatte gleichsam in eine „Komplexitätsfalle" zu locken, indem man entweder behauptet, kollektive Handlungen würden keine oder nur schwer bestimmbare Objekte der Verantwortung erzeugen, oder indem man die Möglichkeit der Identifizierung von Subjekten der Verantwortung im System verneint. Das Problem lässt sich wie folgt beschreiben: Gegeben sei ein System bestehend aus einer gewissen Anzahl von Systemteilen (seien dies nun Menschen oder technische Systeme), die über einen gewissen Zeitraum hinweg Aktionen vollführen („Handlungen" im weichen Sinn) und die zu einem identifizierbaren Ergebnis einer kollektiven Handlung vernünftiger Größenordnung führen.[53] Der Kern der Analyse solcher Systeme besteht in der Identifizierung von Entscheidungspunkten. Ein Entscheidungspunkt ist dadurch charakterisiert, dass der Systemteil eine Auswahl zwischen verschiedenen Handlungsalternativen vornimmt, wobei diese Wahl das Ergebnis der kollektiven Handlung beeinflusst. Die Relevanz dieses Entscheidungspunktes hinsichtlich der Verantwortungsproblematik misst sich an folgenden Eigenschaften: erstens dem Grad an Autonomie des Systemteils im Augenblick der Wahl; zweitens der Fähigkeit, die Folgen der Wahl im Hinblick auf das Ergebnis der kollektiven Handlung abzuschätzen; drittens dem „Designaspekt" des jeweiligen Entscheidungspunktes. Der „Designaspekt" betrifft die institutionelle Struktur des Systems, welche zwischen den beiden folgenden Polen angesiedelt werden kann: So kann das System vorgängig starr durchstrukturiert worden sein, vergleichbar etwa mit einem militärischen Verband. Oder das System ist Ergebnis eines weitgehend ungesteuerten Selbstorganisationsprozesses, ein Beispiel wäre der ideale Markt. Im ersten Fall kommen neben den Akteuren der kollektiven Handlung noch die Systemdesigner als potentielle Subjekte der Verantwortung in Frage.

Eine auf die Entscheidungspunkte fokussierte Analyse des Systems zeigt auf, ob das Ergebnis einer kollektiven Handlung ein Objekt der Verantwortung darstellt oder nicht. Besteht bei sämtlichen Entscheidungspunkten im Zeitraum der Entstehung der kollektiven Handlung keine Autonomie oder keine Möglichkeit einer Fol-

[51] Vgl. dazu LENK, H., MARING, M. (1995): *Wer soll Verantwortung tragen? Probleme der Verantwortungsverteilung in komplexen (soziotechnischen-sozioökonomischen) Systemen*, in: BAYERTZ, K. (Hg.): Verantwortung. Prinzip oder Problem?, Darmstadt, 241-286.

[52] Vgl. dazu RAMMERT, W. (2003): *Technik in Aktion: Verteiltes Handeln in soziotechnischen Konstellationen*, in: CHRISTALLER, T., WEHNER, J. (Hg.): Autonome Maschinen, Wiesbaden, 289-315.

[53] Gemeint ist damit, dass es keinen Sinn macht, zu große raumzeitliche Ereignisse (wie die Ereignisse „Untergang des römischen Reiches" oder „der hundertjährige Krieg") als singuläre kollektive Handlungen einer Verantwortungsanalyse im dargestellten Sinn unterziehen zu wollen.

genabschätzung der Wahl, ist dieses Ergebnis kein Objekt der Verantwortung, da es gleichsam deterministisch und blind entstanden ist. Eine solche Situation ist in der Praxis wohl aber nur selten gegeben, so dass man davon ausgehen kann, dass die meisten Ergebnisse kollektiver Handlungen Objekte von Verantwortung sein können.

Die Frage nach einer möglichen Zuschreibung von Verantwortung an autonome Systeme, die an kollektiven Handlungen mitwirken, lässt sich nun wie folgt umformulieren: Befinden sich autonome Systeme an verantwortungsrelevanten Entscheidungspunkten? Diese Relevanz bemisst sich an den genannten Aspekten Autonomie, Prognosemöglichkeit und Design des Systems.

Der erste Aspekt betrifft die Autonomie der Systeme. Wie gesagt haben die Systeme keine Autonomie hinsichtlich der grundlegenden Ziele. Einem Menschen in vergleichbarer Situation ermöglicht diese Form der Autonomiefähigkeit die Option, sich von einer kollektiven Handlung loszusagen (z.B. durch Kündigung einer Arbeitsstelle) – was in der Praxis natürlich nicht immer einfach sein wird. Abgesehen von dieser Möglichkeit erscheint der Spielraum der Wahrnehmung der Autonomiefähigkeit vergleichbar. Wichtig in diesem Zusammenhang ist aber die Feststellung, dass die Lernerfahrungen autonomer Systeme oft in Interaktion mit menschlichen Bedienern entstanden sind, was die Frage aufwirft, ob das System und der Bediener/„Lehrer" quasi gemeinsam das Subjekt der Verantwortung bilden.

Der zweite Aspekt betrifft die Fähigkeit des autonomen Systems, die Auswirkungen seiner Aktionen auf das Ergebnis der kollektiven Handlung abzuschätzen. Hier gilt es zu bemerken, dass autonome Systeme oft gerade aus dem Grund entwickelt und (dereinst) eingesetzt werden, weil man ihnen in dieser Hinsicht gegenüber dem Menschen Überlegenheit attestiert. Man will mit autonomen Systemen die „Fehlerquelle Mensch" reduzieren. Der Preis dafür freilich ist eine qualitative Reduzierung der Abschätzung der Handlungsfolgen, da ein „bewusstes Maschinengewissen" ja nicht vorhanden ist.

Der dritte Aspekt schließlich betrifft die institutionelle Struktur des Gesamtsystems, in welches das autonome System eingebettet wird. Diese muss nicht notwendigerweise starr sein. Beispielsweise autonome Software-Agenten, welche Finanztransaktionen tätigen, mögen dem Zweck der Gewinnmaximierung dienen. Die Struktur des Finanzmarktes, in welchem sie tätig sind, muss deshalb nicht notwendigerweise starr sein. Zusammengefasst ergibt sich demnach, dass autonome Systeme einen Beitrag zur Genese von Objekten von Verantwortung leisten können.

Variablen C/D: Die Frage, ob autonome Systeme Subjekte ethischer Theorien werden können, ist sehr spekulativ, da derzeit nicht klar ist, wie autonome Systeme in einem ethischen Sinn geschädigt werden können. Natürlich ist denkbar, dass sich dies dereinst ändern kann. Die Tierethik-Debatte mag hier als gutes Beispiel dienen. Die neuere Diskussion, insbesondere der so genannte Pathozentrismus, nützt Erkenntnisse der Wissenschaft über mögliches Schmerzbewusstseins bei Tieren, um

diesen einen Subjektstatus bei ethischen Theorien zu geben.[54] Vergleichbar läuft das Argument bei Robotern, welchen dann Berücksichtigung als Subjekte der Ethik finden sollen, wenn bei ihnen Formen von Bewusstsein nachgewiesen werden könnten.[55] Heute von der Notwendigkeit der Entwicklung einer Ethik für autonome Systeme zu sprechen, ist aber verfehlt, da wie bereits erwähnt die Frage nach einem „Maschinenbewusstsein" zu spekulativ ist.

Variable E: Bedeutsamer ist jedoch die Frage, welche Instanzen bei der Verantwortungsbeurteilung eine Rolle spielen. Da autonome Systeme bald einmal kommerziell genutzt werden, wäre der Auftraggeber eine natürliche Instanz in dem Sinn, dass er Rechenschaft einfordern wird, wenn das System Fehlleistungen erbringt. Doch kann eine solche Rechenschaft eingefordert werden, wenn ein System erworben wurde, dem man bewusst einen autonomen Spielraum eingeräumt hat? Da lernende Systeme durch die Umgebung – definiert durch den Käufer – geprägt werden, haben diese das System ebenfalls geprägt. Dieser Aspekt wird insbesondere bei der Problematik der Zuschreibung von Verantwortung im Rechtssystem von Bedeutung. Auf diesen Punkt wird vertiefend im letzten Abschnitt eingegangen.

Variablen F/G: Die Frage nach den Grenzen der Zurechnung von Handlungen/Aktionen in zeitlicher Hinsicht ist ein zentrales Problem der Verantwortungsdiskussion. Schließlich ist es keineswegs so, dass sich für jedes verantwortungsethisch zu beurteilende Problem eine Kausalkette mit eindeutigem zeitlichen Anfangs- und Endpunkt definieren lässt. Verantwortlichkeiten aufgrund von lange zurückliegenden Handlungen zuschreiben zu wollen, macht oft keinen Sinn. Ebenso gibt es für prospektive Verantwortlichkeit einen Zeithorizont, der primär durch die Voraussagbarkeit des Eintreffens der verschiedenen Alternativen gegeben ist. Autonome Systeme verschärfen das Problem in zweierlei Hinsicht. Zum einen finden sie oft Einsatz in größeren Systemen mit dem Ziel, die Zuverlässigkeit des Gesamtsystems zu erhöhen. Realisiert wird dies unter anderem mit Lernleistungen des Systems, welche sich in einer größeren Zahl innerer Zustände manifestiert. Die Frage ist nun, ob diese Zustände auch dem menschlichen Nutzer zugänglich sind. Ansonsten steht der Mensch vor dem paradoxen Problem, dass das System immer besser funktioniert, man aber immer weniger weiß, warum dies der Fall ist. Dieses Problem leitet direkt über zur Frage nach der Handlungswirksamkeit des Verantwortungsappells bei Menschen, wenn immer mehr autonome Systeme Einsatz finden. Dies ist ein weiteres zentrales Problem dieser Diskussion, welches im letzten Abschnitt genauer untersucht werden soll.

Zusammenfassend lässt sich feststellen, dass autonome Systeme nicht gleichsam begrifflich aus der Verantwortungsproblematik ausgeschlossen werden können. Ausgehend vom klassischen Handlungsbegriff sind sie zwar keine Subjekte der Verantwortung; doch der klassische Handlungsbegriff ist zu starr, um Handlungsfolgen

[54] WOLF, J.-C. (1992): *Tierethik. Neue Perspektiven für Menschen und Tiere*, Freiburg i.Br.

[55] PUTNAM, H. (1964): *Robots: Machines or artificially created life?*, in: Journal of Philosophy 61, 668-691.

komplexer Systemzusammenhänge – in welche autonome Systeme nun mal eingebettet sind – hinsichtlich der Verantwortungsproblematik beurteilen zu können. Es bestehen gute Gründe für die Ansicht, dass autonome Systeme einen gleichsam irreduziblen Anteil an einem Objekt der Verantwortung bilden, das Problem der Zurechenbarkeit von Verantwortungsobjekt und -subjekt erschweren und die Wirksamkeit des Verantwortungsappells bei Menschen abschwächen. Sie können damit prinzipiell Gegenstand einer Zuschreibung von Verantwortung sein. Bevor die Konsequenzen dieser Diagnose für die Theorie der Verantwortung und den Bau autonomer Systeme beurteilt werden, sollen anhand dreier Beispiele mögliche praktische Probleme erläutert werden.

IV. Systemautonomie und Verantwortung – Beispiele

Autonome Systeme sind Gegenstand weitreichender Forschungsanstrengungen. Kommerziell hingegen sind solche Systeme noch kaum im Einsatz. Da das Verantwortungsproblem erst dann von praktischer Relevanz wird, wenn solche Systeme allgemein Anwendung finden und Vertragsbeziehungen im Hinblick auf Kauf, Einsatz und Unterhalt autonomer Systeme entstehen, werden nachfolgend drei fiktive Beispiele vorgestellt. Die Möglichkeit, dass derartige Systeme dereinst Anwendung finden, ist durchaus gegeben.[56]

Beispiel I: Autonomous Trading

Die modernen Finanzmärkte könnten heutzutage ohne umfassenden Computereinsatz nicht mehr funktionieren. Einige Kauf- und Verkaufsoperationen sind bereits automatisiert worden, doch wesentliche Kauf- und Verkaufsentscheide werden weiterhin von Brokern getroffen. Derzeit laufen Forschungsanstrengungen zur Entwicklung von Software-Agenten, welche die Funktion von Brokern zumindest teilweise übernehmen könnten, da Computersysteme die großen Datenmengen, welche zur Preisbildung führen, weit schneller erfassen können als Menschen. Es wird nun davon ausgegangen, ein Anbieter habe ein solches System entwickelt und einem Broker-Unternehmen verkauft. Das System wird im Rohstoffhandel eingesetzt. Es erhält Zugang zu den relevanten Daten des Rohstoffmarktes wie auch zum Wirtschaftinformationsdienst von Reuters, dessen Nachrichten maschinenlesbar geworden sind, d.h. das System ist beispielsweise in der Lage, die Bedeutung eines politischen Umsturzes in einem kupferproduzierenden Land für den kurz- und mittelfristigen Preis von Kupfer zu erfassen. Der zu optimierende Parameter des Systems ist Gewinn. Das System lernt, indem es für eine gewisse Zeitperiode Assistent eines menschlichen Brokers ist. Während der Lernphase ist der Broker der eigentliche Händler. Das System vergleicht jeweils sein Ergebnis mit dem Ergebnis des Händlers und lernt in dieser Periode die Einschätzung „weicher Faktoren" bei der Preisbildung. In diesem Sinn wird das System von den Lernerfahrungen mit einem ganz bestimmten menschlichen Broker geprägt. Das System testet aber immer auch eine ganze Gruppe von Strategien. Sollten sich diese während der Lernphase als besser erweisen, kann das System gewissermaßen auch eine eigene „Broker-Identität" erschaffen.

Nun kommen zwei solche Systeme zum Einsatz. Ein System wird von einem unehrlichen Broker trainiert, der zu Lasten der Kunden des Unternehmens Wertpapiere zu schnell verkauft, damit einen kurzfristigen Gewinn erzielt (der auch seine Provision erhöht), dem Kunden dadurch aber längerfristig höhere Gewinne vorenthält. Das zweite System wird von einem ehrlichen Broker trainiert. Im Zug der Lernphase merkt aber dieses System, dass ein Verhalten analog zum ersten System den Gewinn des Unternehmens stärker erhöht, und es eignet sich diese Strategie an. Nach der Lernphase werden die Systeme Händler. Nach einer gewissen Zeit merken die Kunden des Unternehmens, dass ihnen Gewinne entgehen, und klagen gegen das Unternehmen. Dieses wiederum klagt gegen das Unternehmen, das die autonomen Systeme ausgeliefert hat. Wer ist verantwortlich für den entgangenen Gewinn der Kunden?

Dieses Beispiel geht auf das Problem ein, inwiefern man ein autonomes System als Subjekt der Verantwortung auffassen kann. Offenbar hat man es mit einem lernenden System zu tun, das im Verlauf der Lernphase eigene Strategien für seine Handlungen entwickeln kann. Der Hersteller des Systems könnte geltend machen, dass eine falsche Lernumgebung beim Anwender zur „Fehlfunktion" geführt habe und demnach dieser die Verantwortung zu tragen habe. Der Anwender kann aber geltend machen, dass das System (im zweiten Fall) „von sich aus" ebenfalls auf die für die Kunden des Unternehmens ungünstige Strategie gekommen sei und demnach der Lernmechanismus fehlerhaft sei. Offenbar liegt das Problem im durch das System zu optimierenden Parameter „Gewinn" verborgen, bzw. in der Zeitspanne, über welche Gewinn erzielt werden soll. Tatsächlich sind ja auch langfristig höhere Gewinnerwartungen bis zu einem gewissen Grad unsicher und der Einsatz autonomer Systeme könnte dazu benutzt werden, diesbezüglich eine Sicherheit nur vorzugaukeln. Das Entgehen langfristiger Gewinne ist schließlich für die Kunden des Unternehmens nur im Nachhinein feststellbar. Drei Auswege aus dieser Situation bieten sich an: Erstens, das autonome System wird zum Subjekt der Verantwortung. Für den Hersteller wie den Nutzer des Systems würde eine solche Zuschreibung wohl dienlich sein, während der Kunde dies nicht akzeptieren würde, da das System

[56] Eine Vielzahl von Arbeiten zu den genannten Themen sind zugänglich unter dem Portal der *Association of Computing Machinery (ACM)* (ACM digital library, http://portal.acm.org/portal.cfm). Prototypen von Software-Agenten im Bereich *autonomous trading* messen sich an der *Trading Agent Competition* (http://www.sics.se/tac/), ein Wettbewerb für Rettungsroboter ist Teil des alljährlichen *RoboCup* (http://www.robocup.org/). Ich danke Göran Andersson vom *Power Systems Laboratory* der ETH Zürich für seine Beurteilung des Fallbeispiels *autonomous powergrid control.*

ja nicht sanktioniert werden kann – etwa im Sinn, dass es Schadenersatz leisten könnte. Zweitens, Hersteller wie Nutzer teilen sich in einem noch zu definierenden Sinn die Verantwortung. Diese Lösung könnte darauf hinauslaufen, dass ein Markt für autonome Systeme eine hohe Regeldichte aufweisen wird, um Rechtsstreitigkeiten zu verhindern. Drittens, man sieht in der Leistung des autonomen Systems einen Hinweis darauf, dass der Kunde gar nicht geschädigt worden sei, da die langfristigen Gewinne zum Zeitpunkt der Entscheide des Systems bzw. der Broker unsicher waren. Da das System keinen persönlichen Nutzen aus Gewinnen ziehen kann, ist dessen Verhalten ein objektiver Hinweis auf die aufgrund der unsicheren Wissensbasis tatsächlich erzielbaren Gewinne – und die Unehrlichkeit des Brokers hat keinen realen Hintergrund. Der Ausweg des Kunden besteht dann einfach im Wechseln des Anbieters der entsprechenden Finanzdienstleistung.

Beispiel II: Autonomous Powergrid Control

Die Verteilung elektrischer Energie über kontinentale Stromnetzwerke ist ein schwieriges Problem und zugleich für die moderne Zivilisation von enormer Bedeutung. Die in jüngster Zeit festgestellte Häufung von Ausfällen ganzer Stromnetze[57] zeigt die Anfälligkeit der Moderne für Stromausfall, der nicht nur wirtschaftlichen Schaden in Milliardenhöhe verursacht, sondern auch Menschenleben fordern kann. Es ist deshalb nahe liegend, dass Anstrengungen zur weiteren Automatisierung der Stromverteilungsnetze unternommen werden. Ziel solcher Automatisierungen ist es beispielsweise, raumzeitliche Belastungsmuster zu erkennen[58], so dass das Netzwerk präventiv auf solche Belastungsgrenzen reagieren kann. Es wird nun angenommen, die europäischen Stromversorgungsgesellschaften installierten autonome Systeme für die Kontrolle der Stromverteilungsnetze, welche die genannten Belastungsmuster lernen können. Das System ist aber nicht vollständig autonom. Ingenieure haben weiterhin die Möglichkeit, die Einstellungen der Stromverteilungsnetze zu verändern, wenn sie denken, dass die Maßnahmen des autonomen Systems unangebracht seien.

Ausgehend von dieser Situation wird angenommen, dass aufgrund eines Großereignisses der Stromverbrauch in Norditalien weit über dem Schnitt liegt. Gleichzeitig bedrohen starke Gewitter in den französischen Alpen die Transitleitungen von Frankreich nach Italien. Zudem sind zwei Schweizer Speicherkraftwerke, die Spitzenstrom liefern können, in Revision. Die italienischen Kraftwerke laufen an der

[57] Beispiele sind der Ausfall des US-Stromnetzes an der Ostküste (14. August 2003), des Netzes in Südschweden und Dänemark (23. September 2003) und des italienischen Netzes (28. September 2003).

[58] Ein bekanntes Beispiel ist der starke Anstieg des Stromverbrauchs in der Halbzeit des englischen Cupfinals, da die Pause genutzt wird, um Tee zu kochen – also Millionen von Teekesseln angestellt werden.

Kapazitätsgrenze. Das System erkennt eine Bedrohung für das italienische Stromnetz und empfiehlt, Norditalien vom Netz zu nehmen, um Zentral- und Süditalien vor dem Kollaps zu bewahren. Der diensthabende Ingenieur wendet sich gegen den Entscheid des Systems und belässt Norditalien am Netz. Das System ist nun mit einem bisher unbekannten Zustand konfrontiert. Es registriert danach einen ungewohnten Anstieg des Stromverbrauchs in Süditalien und empfiehlt neu, Süditalien vom Netz zu nehmen. Auf diesen Vorschlag geht der Ingenieur ein, Süditalien wird vom Netz genommen. Aufgrund der schlechteren Infrastruktur in Süditalien dauert es fast 24 Stunden, bis die dortige Stromversorgung wieder gewährleistet ist. Eine nachfolgende Untersuchung des Vorgangs zeigt, dass das autonome System einen falschen Ratschlag gegeben hat, indem es die Abschaltung von Süditalien empfohlen hat. Hingegen ist der vorgängige, vom Ingenieur abgelehnte, Vorschlag zur Abschaltung Norditaliens angemessen gewesen. Der Ingenieur verteidigt sich, dass ein Abschalten des norditalienischen Netzes mit zu großen ökonomischen Kosten verbunden gewesen wäre und er das Risiko einer Netzüberlastung eingehen wollte. Als dann das autonome System die Abschaltung des süditalienischen Netzes empfahl, habe er davon ausgehen müssen, dass dieser Vorschlag gerechtfertigt sei, denn dafür sei das System ja schließlich da. Welche Verantwortung trägt der Ingenieur?

In diesem Beispiel spielt das Problem des Erkennens des inneren Zustandes des autonomen Systems eine wichtige Rolle. Durch das Wahrnehmen von Verantwortung durch den Ingenieur, indem er sich gegen das System und für das Eingehen eines unbekannten Risikos entschieden hat (eine Form eines Entscheids, die dem autonomen System nicht zur Verfügung steht), ist das System mit einem unbekannten inneren Zustand konfrontiert, der die Wahrscheinlichkeit falscher künftiger Prognosen erhöht. Die damit durch das System sonst ermöglichte Ausdehnung der Zeitspanne der verlässlichen Beurteilung des Verhaltens des Stromverteilungsnetzes ist geschrumpft – nur hat der Ingenieur offenbar keinen Zugang zu dieser Information. Die Beurteilung der Güte eines Entscheids des autonomen Systems ist demnach eine Lernleistung, die der Nutzer des Systems erbringen muss, und die Frage, wie dieses Lernen unterstützt wird, wird zu einem wichtigen Problem.

Beispiel III: Autonomous Rescue Robot

Autonome Robotsysteme sollen dort zum Einsatz kommen, wo Menschen aus Sicherheits- oder Kostengründen nicht aktiv werden können, und eine unsichere Umwelt besteht, welche von einer Fernsteuerung der Systeme aufgrund zu langer Reaktionszeiten abraten lässt. Ein Paradebeispiel ist die Raumfahrt. Aber auch die Suche nach Menschen in einem Trümmerfeld, etwa verursacht durch ein Erdbeben, gilt als künftiges Einsatzgebiet autonomer Roboter, da die Gefährdung von Menschen durch nachfolgende Einstürze oft nicht in Kauf genommen werden soll. Das Beispiel geht davon aus, dass ein solches System besteht. Der Roboter ist in der Lage, Menschen in Trümmerfeldern zu orten und deren Überlebenswahrscheinlichkeit durch einfache Diagnoseverfahren rudimentär zu bestimmen. Zudem kann er

Erste-Hilfe-Maßnahmen leisten. Den Rettungsteams gibt er schließlich den Standort der Person an sowie eine Einschätzung der Gefahren der Rettung. Die Wissensbasis des Systems wird durch Experten des Unternehmens, welches die Systeme verkauft, nach jedem Einsatz neu aufdatiert, so dass ein optimales Funktionieren gewährleistet ist.

Nun kommt es zum Einsatz eines solchen Systems. Das System arbeitet sich durch die Trümmer eines eingestürzte Hochhauses und entdeckt zwei schwer verletzte Menschen. Es kann nur Hilfe für eine Person leisten und gibt dieser Person ein Medikament ab. Einige Zeit später haben sich die Rettungsteams zu den beiden Personen vorgearbeitet – beide Personen sind tot. Als die Angehörigen der vom Roboter nicht betreuten Person erfahren, dass die andere Person vom System ausgewählt wurde, klagen sie gegen den Konstrukteur des Roboters. Wer ist verantwortlich für das „Versagen" des Systems?

Dieses Beispiel zeigt das Problem, wie die Übertragung einer klassischen Dilemma-Situation auf autonome Systeme deren ethische Beurteilung verändern kann. Hätte ein Arzt vor der Situation gestanden, mit limitierten Hilfsmitteln eine Auswahl zwischen zwei Schwerverletzten treffen zu müssen, hätte man diese Selektion eher als Folge einer Notlage gesehen, selbst wenn der Gepflegte ebenfalls gestorben wäre. Beim autonomen System hingegen könnten solche „Fehler" weniger verziehen werden, da man ihnen eine größere Sicherheit bei der Entscheidung zugesteht – dies, weil die Entscheidung ohne emotionalen Druck und unter Rückgriff auf eine umfassende Wissensbasis erfolgt.

V. Menschengerechte autonome Systeme

Mit Hilfe der drei Beispiele lassen sich für die abschließende Diskussion vier Problemfelder identifizieren, bei welchen der Einsatz autonomer Systeme mit dem Verantwortungsbegriff in Konflikt geraten kann:

- *Autonome Systeme als Subjekt der Verantwortung:* Das Beispiel des *autonomous trading* hat deutlich gemacht, dass autonome Systeme, neben den erwähnten theoretischen Gründen, auch aus praktischen Gründen kein eigentliches Subjekt der Verantwortung darstellen können. Es fehlt an der Möglichkeit, autonome Systeme zur Rechenschaft zu ziehen, um sie beispielsweise finanziell haftbar zu machen. Hingegen kann der Verbund „autonomes System – Nutzer" dazu führen, dass bei der Verantwortungsbeurteilung der Handlungen des Nutzers diesem nicht alle Handlungen des Verbundes zugeordnet werden können, da der Nutzer sich auf eine Nichtzuständigkeit berufen kann. Der Einsatz autonomer Systeme bedeutet eben auch eine bewusste Inkaufnahme von Unsicherheit hinsichtlich der Aktionen des Systems. Dies könnte dazu führen, dass im Sinn der deskriptiven Verantwortung gewisse Aspekte der Handlungen des Verbundes „System – Nutzer" keinem Subjekt der Verantwortung zugeordnet werden können. Fehlt die Möglichkeit der Ermittlung der deskriptiven Verantwortung, ergibt sich das Erfordernis einer mindestens rechtlichen Zuschreibung von Verantwortung, um Haftpflichtfragen, die sich aus der Anwendung autonomer Systeme ergeben könnten, Rechnung tragen zu können. Eine mögliche Lösung für solche Probleme besteht in einem Fonds, gemeinsam getragen durch Produzenten und Anwender autonomer Systeme, welcher für derartige Haftungsprobleme zur Anwendung kommen könnte.[59]
- *Autonome Systeme und das Objekt der Verantwortung:* Es wurde hinreichend gezeigt, dass autonome Systeme, eingebunden in meist komplexe sozio-technische Systeme, an der Erzeugung von Objekten der Verantwortung beteiligt sind. Wie oben erwähnt wird es Beispiele geben, in welchen dieser Beitrag gleichsam irreduzibel ist. Hingegen ist aber auch denkbar (vgl. das Beispiel des *autonomous trading*), dass Ergebnisse kollektiver Handlungen mit einem wesentlichen Beitrag autonomer Systeme gar nicht mehr als ein Objekt der Verantwortung gesehen werden. Vielmehr ist es die Beteiligung des Systems, welche (da dieses ja keine menschlichen Schwächen wie Habgier und dergleichen hat) eine vormals ethisch problematisierte Frage zu einer gleichsam naturgesetzlichen Tatsache werden lässt. Das Problem der Delegation von menschlichen Entscheiden an die Maschine wird damit verschärft, da mit der Delegation die Antwort gleichsam zum Ungegenstand der Ethik wird. Das Beispiel des *autonomous rescue robot* zeigt aber auch, dass in manchen Fällen das Gegenteil eintreffen könnte. Demnach werden Fehlermöglichkeiten, die der Situation inhärent sind und demnach (beim Einbezug von Menschen) zu keiner praktischen Ethikdiskussion Anlass geben, durch das autonome System auf deren Nutzer/Produzenten projiziert, da man ja von solchen Systemen fehlerfreies Funktionieren erwartet.
- *Autonome Systeme und die Instanz der Verantwortungsbeurteilung:* Dieses Problem betrifft nicht nur das Rechtssystem, sondern insbesondere auch das Verhältnis zwischen Produzent und Käufer/Nutzer des autonomen Systems. Da autonome Systeme durch den Nutzer geprägte Lernerfahrungen machen werden, die der Produzent nur zum Teil definieren kann, wird die Ausgestaltung der Lernumgebung zu einem wichtigen Faktor. Je nach Anwendungsgebiet dürfte die Regulierung der Nutzung solcher Systeme größere Ausmaße erreichen – womit man aber der eigentlichen Absicht für den Einsatz des autonomen Systems, die Situation zu vereinfachen, entgegenwirkt. Eine allgemeine Lösung

[59] Dieser Vorschlag ist vergleichbar mit der Idee, Roboter als „Sachen mit besonderer Haftungsregel" zu betrachten. Vgl. dazu CHRISTALLER, T., DECKER, M., GILSBACH, J.-M., HIRZINGER, G., LAUTERBACH, K., SCHWEIGHOFER, E. SCHWEITZER, G., STURMA, D. (2001): *Robotik. Perspektiven des menschlichen Handelns in der zukünftigen Gesellschaft*, Berlin, Heidelberg, New York, 143 f.

für dieses Problem lässt sich kaum finden, da dieses stark von den konkreten Einzelfällen abhängen wird.

- *Autonome Systeme und die Wirksamkeit des Verantwortungsappells:* Ein wichtiger Aspekt autonomer Systeme ist das gesteigerte Maß an Unsicherheit bezüglich des Systemverhaltens. Dieses Problem geht einher mit der Schwierigkeit, den inneren Zustand des Systems zu erfassen und zu interpretieren – ein Problem, das sich am Beispiel des *autonomous powergrid control* gezeigt hat. Der grundsätzliche Aspekt des Problems besteht darin, dass das Verhalten eines autonomen Systems in einem Maß gelernt werden muss, wie man es von klassischen Maschinen bisher nicht kennt. Dies stellt neue Anforderung an die Schnittstelle zwischen Mensch und Maschine, welche von einer Art sein sollte, die das Lernen nicht übermäßig behindert oder in falsche Bahnen lenkt. Diesbezüglich gilt es auch das Problem zu beachten, dass „Eindringlinge" (feindliche Nutzer) den inneren Zustand autonomer Systeme eventuell besser einschätzen lernen als die legitimen Nutzer. Dies könnte ein mögliches neues Problem für die Sicherheit technischer Systeme aufwerfen.

Zusammengefasst zeigt sich, dass das künftig zu erwartende Auftreten technischer Systeme mit einem gewissen Eigenleben die menschliche Gesellschaft wie auch die Ethik vor neue Herausforderungen stellt. Diese Arbeit zeigt auf, dass autonome Systeme in mehrfacher Hinsicht mit dem Verantwortungsbegriff in Beziehung treten, auch wenn autonome Systeme keine moralischen Subjekte sind:

- Die Einbindung autonomer Systeme in kollektive Handlungen kann dazu führen, dass die Ermittlung von Verantwortlichkeiten bei unerwünschten Folgen nicht mehr gelingen kann.
- Dieses Scheitern der Ermittlung von Verantwortlichkeit hat praktische wie theoretische Gründe: Vom praktischen Standpunkt aus könnte der Aufwand zur Ermittlung der Verantwortlichkeit in keinem Verhältnis zum entstandenen Schaden sein. Vom theoretischen Standpunkt aus nimmt man beim Einsatz eines autonomen Systems willentlich eine gewisse Unsicherheit und einen gewissen Kontrollverlust in Kauf.
- Aus diesem Grund kann es in gewissen Fällen Sinn machen, einem autonomen System mindestens rechtliche Verantwortung zuzuschreiben, um langwierige Rechtsstreitigkeiten zu verhindern. Für die Deckung von Schäden könnte eine Fonds-Lösung ins Auge gefasst werden.

Diese Arbeit macht deutlich, dass autonome Systeme sowohl an die Naturwissenschaft und Technik als auch an die Geisteswissenschaft und an das Recht neue Herausforderungen stellen. So sehen sich die Ingenieure mit dem Problem konfrontiert, wie „verbotene Zustände" in autonomen Systemen erkannt werden können. Dies ist Voraussetzung dafür, unerwünschte Aktionen autonomer Systeme soweit als möglich zu verhindern. Weiter wird es darum gehen, eine Art „Taxonomie" autonomer Systeme zu erstellen, anhand welcher sich unterschiedliche Formen von

Systemautonomie unterscheiden lassen. Kategorien für die Erstellung einer solchen Taxonomie wären die Freiheitsgrade, welche einem solchen System offen stünden, die Art des Lernens, welche implementiert wäre, die Prognoseunsicherheit hinsichtlich des Verhaltens des Systems und die Einsatzformen dieser Systeme. Die Zuschreibung von Verantwortung an autonome Systeme könnte dann mittels einer solchen Taxonomie erfolgen. Die Möglichkeit einer Fonds-Regelung für Haftungsfragen bei Fehlleistungen autonomer Systeme müsste ebenfalls genauer geprüft werden – insbesondere hinsichtlich der Frage, wie viel Geld in einen solchen Fonds fließen soll. Volkswirtschaftlich macht es nicht viel Sinn, den Einsatz autonomer Systeme mit einem Fonds zu verteuern, der dann kaum zum Einsatz kommt. Schließlich wird sich die Frage stellen, ob autonome Systeme für gewisse Anwendungsgebiete (beispielsweise die Medizin, aber auch das Militär) ausgeschlossen werden sollten, weil die damit verbundenen Verantwortungsprobleme als zu gravierend eingestuft werden.

Zusammenfassend macht diese Arbeit deutlich, dass technische Systeme zunehmend ein „Eigenleben" erlangen werden, das das Verhältnis zwischen Mensch und Technik verändern wird. Dieses „Eigenleben" wird der Technik mit dem Ziel gegeben, das Funktionieren komplexer technischer Systeme besser zu gewährleisten. Der Mensch ist dann aber mit zwei Aufgaben konfrontiert: Der Mensch muss das Systemverhalten so prägen können, dass er die Autonomie hinsichtlich der Zwecksetzung behält. Der Mensch muss sich aber auch bewusst sein, dass er das Systemverhalten nicht immer vorausbestimmen kann, sondern manchmal lernen muss, was das System nun wirklich tut. Eine menschengerechte Einbindung autonomer Systeme in unsere Gesellschaft verlangt jedoch, dass dieses Lernen so vonstatten gehen kann, dass die autonomen Systeme nie wirklich große Schuld auf sich laden können – denn man wird sie nicht zur Rechenschaft ziehen können.[60]

[60] Ich danke Endre Bangerter (IBM Forschungslabor Rüschlikon), Albert Kündig (ETH Zürich), Robert Ruprecht (Fachhochschule Biel) und Ruedi Stoop (Universität/ETH Zürich) für ihre wertvollen Hinweise und Anregungen.