Haben autonome Maschinen Verantwortung?

Christoph Hubig

Verantwortlichkeit ist keine manifeste Eigenschaft, sondern beruht auf Zuweisung durch Dritte und/oder durch Selbstzuweisung. In verbreiteten umgangssprachlichen Redeweisen finden sich oftmals undifferenzierte Zuweisungen von Verantwortlichkeit, wenn Einflussfaktoren, Determinanten, begünstigende oder unglückliche Umstände etc. gemeint sind. Das Subjekt der Zuschreibung wird dabei oftmals unter einem allgemeinen Kollektivsingular gefasst, wenn etwa davon die Rede ist, dass die Digitalisierung inzwischen in höherem Maße als die Globalisierung für den gesellschaftlichen Wandel ‚verantwortlich', die E-Mobilität für den Wegfall zigtausender Arbeitsplätze in der Getriebefertigung ‚verantwortlich', das Ausbildungssystem für Chancenungleichheit ‚verantwortlich' oder der Einsatz Autonomer Systeme für die menschenleere Fabrik ‚verantwortlich' seien. Bei näherer Betrachtung zeigt sich schnell, dass solche Diagnosen mit ihren markigen Titelworten nicht zielführend sind, solange nicht die komplexen Beziehungsgeflechte freigelegt sind, innerhalb derer sich Interessen durchsetzen, Weichenstellungen für das Eröffnen oder Verschließen von Handlungsräumen vorgenommen sowie Entscheidungen zur Realisierung bestimmter kausaler Effekte selbst stattfinden oder beeinflusst werden. Wenn in diesem Kontext der Frage nach der ‚Verantwortlich-

keit‘ Autonomer Systeme nachzugehen ist, geraten wir in ein Spannungsfeld, innerhalb dessen Relationen zwischen Handlungen (i.e.S.) natürlicher Subjekte, handlungsförmigen Vollzügen nicht-menschlicher Akteure unter delegierter Verantwortlichkeit (wobei die Verantwortlichkeit für die Delegation bei menschlichen Subjekten verbleibt), und schließlich einem Prozessieren der Systeme, die freigesetzt und in einem gewissen Sinne ‚autonom‘ sind (eigenständig Problemlösungen zu entwickeln und zu ‚entscheiden‘), zu verhandeln sind. Hierbei sind in vielerlei Hinsicht Differenzierungen anzubringen, denn der Teufel liegt – wie immer – im Detail. Nachfolgend sollen hierfür in einigen Schritten klassische Diskussionslinien aufgenommen und Vorschläge unterbreitet werden, die geeignet sind, die Rede von der ‚Verantwortlichkeit‘ für unser Problemfeld ‚griffiger‘ zu machen und dadurch die Voraussetzung zu schaffen, Strategien für den Umgang mit Konsequenzen unterschiedlicher Verantwortlichkeit zu erarbeiten.

1. DER FRAGEHORIZONT BEZÜGLICH VERANTWORTUNG

Wenn wir mit Blick auf das Agieren Autonomer Systeme der Frage nach der Zuweisbarkeit von Verantwortung nachgehen, sind zunächst sieben *Hinsichten* des Fragens zu unterscheiden und in einer Art Landkarte zu verorten, weil aus diesen Hinsichten eben nicht sieben *Typen* von Verantwortung resultieren, sondern diese erst dadurch zustande kommen, dass diese Hinsichten in unterschiedlichster Weise aufeinander beziehbar sind und sich erst dadurch die Spezifik der Verantwortungsfrage für Autonome Systeme herauszubilden vermag:

Erstens muss nach dem Subjekt (Wer?) der Verantwortlichkeit gefragt werden: Mögliche Kandidaten hierfür sind natürliche Individuen (z.B. als Fahrer/-in eines autonomen Fahrzeugs oder als Beteiligte an Interaktionen/Koaktionen mit einem Roboter), Individuen als Rollenträger mit Pflichten und Vollmachten (z.B. mit Aufsichts- und Kontrollfunktion) sowie Mensch-Maschine-Hybride mit geteilter Hand-

lungsträgerschaft bei der Realisierung bestimmter Handlungsschemata (siehe den Beitrag von Ingo Schulz-Schaeffer in diesem Band). Ebenfalls als Kandidaten infrage kommen Organisationen und Institutionen, die von unterschiedlich mandatierten Individuen (vgl. hierzu Hubig 2007: Kap. 7.4) vertreten werden (z.B. was die Erstellung von Systemarchitekturen, die Gestaltung der Trainingsbedingungen der Systeme und/oder die Regelwerke für die Mensch-System-Interaktion bis hin zur Haftung betrifft) sowie schließlich die neuerdings vom EU-Parlament vorgeschlagene artifizielle ‚elektronische Person' als mit einem gewissen Fonds versichertes Haftungskonstrukt (s.u.).

Zweitens muss nach dem Gegenstand (Wofür?) der Verantwortung gefragt werden: Hierbei kann es um singuläre Handlungen bzw. Handlungsfolgen in Gestalt direkt kausal hervorgerufener Effekte gehen (gelungene oder misslungene Aktionen), die Übernahme oder die Delegation von Rollen in der Mensch-System-Interaktion, den Aufbau, die Fortschreibung oder Zerstörung von Strukturen möglichen Agierens, den Erhalt oder die Missachtung von Gütern (z.B. ökologischer, ökonomischer oder sozialer Besitzstände), die Gestaltung von Bedingungen systemischen Lernens in Verbindung mit der Eröffnung bestimmter Freiräume des Agierens etc.

Drittens ist nach der Domäne der Verantwortungszuweisung zu fragen: Handelt es sich um moralische, rechtliche, ökonomische oder politische Verantwortung (letztere z.B. auch dann, wenn individuell-persönliches Verschulden nicht vorliegt)?

Viertens ist nach (möglicherweise konfligierenden) Maßstäben zu fragen: Handelt es sich um Werte, Normen, Prinzipien der Moral (als in einem bestimmten Kontext subjektiv anerkannter Orientierungsinstanzen) oder der Sittlichkeit (als in ihrer Rechtfertigbarkeit objektiv/intersubjektiv nachvollziehbarer normativer Instanzen), oder handelt es sich um Rechtszustände mit ihren Regelwerken, die sich aus der Rechtsetzung, nachgeordneten Regelungen durch die Exekutive oder autorisierte Verbände (z.B. der Verein Deutscher Ingenieure [VDI], der Verband der Elektrotechnik Elektronik Informationstechnik e.V.

[VDE] oder das Deutsche Institut für Normung [DIN]) oder dem Richterrecht speisen?

Fünftens ist nach der Instanz (Wovor?) der Verantwortlichkeit zu fragen. Hier kommen das eigene Selbst mit seinen Ansprüchen (z.B. in Gestalt des ‚Gewissens'), betroffene Dritte mit jeweils eigenen Ansprüchen, moralische Autoritäten mit ihren Lob- und Schuldzuweisungen, die Jurisdiktion, Kontrollgremien/-behörden etc. infrage.

Sechstens ist zu fragen, für welchen Zeitpunkt bzw. welche Zeitspanne Verantwortlichkeit zuzuweisen ist: retrospektiv für Vorfälle oder Entwicklungen in der Vergangenheit (deren Folgen oder Folgelasten ggf. zu kompensieren sind) oder prospektiv angesichts möglicher Entwicklungen in der Zukunft, für die Vorsorgepflichten zu übernehmen sind (z.B. was den Umgang mit *Chancen- und Risikopotenzialen* betrifft).

Und schließlich ist *siebtens* der Raumbezug zu thematisieren: Handelt es sich um lokal eingrenzbare Phänomene (für die sich möglicherweise die Verantwortungsfrage leichter beantworten lässt) oder geht es um regionale bis a limine globale Effekte, die unter mannigfachen, multiplen und/oder synergetisch wirkenden Faktoren zustande kommen, deren konkrete Relevanz und Proportionierung schwieriger zu analysieren ist?

Auf dieser ‚Landkarte' sind diejenigen möglichen Verbindungen zwischen den ‚Orten' des Fragens mit Blick auf das Agieren ‚autonomer Maschinen' zu eruieren. Insbesondere wäre zu fragen, ob angesichts der Komplexität der Problemlage Strategien der Vereinfachung der Verantwortungszuweisung auch und gerade in praktischer Hinsicht rechtfertigbar erscheinen (siehe Kapitel 6).

2. KRITERIEN DER VERANTWORTLICHKEIT

Fragen wir zunächst in einem zweiten Schritt nach einer einfachen Ausgangsposition, von der aus Kriterien der Verantwortlichkeit entsprechend unseren Intuitionen ersichtlich werden, um dann weiter zu

fragen, inwieweit mit Blick auf autonome Maschinen hier Modifizierungen anzubringen sind. Für individuelle natürliche Subjekte gilt wohl als *erstes Kriterium* Autonomie als Fähigkeit, unter selbst (*autos*) gesetzten und/oder anerkannten Regeln (*nomoi*) zu handeln, das heißt Mittel zur Realisierung von Zwecken unter Zielen der Lebensführung einzusetzen. Dies setzt (negative) Freiheit von externer Determination (Freiheit von ...) und (positive) Freiheit der Verfügung über einen Optionenraum für Handlungsalternativen (Freiheit zu ...) voraus. Wäre beides jeweils nicht gegeben, würden wir Verantwortlichkeit absprechen oder zumindest entscheidend relativieren.

Zweitens gilt als Kriterium (welches mit dem ersten einhergeht), dass das Subjekt ‚intentionale Zustände' aufweist, das heißt in einem bewussten Weltbezug steht, innerhalb dessen es im Rahmen seines theoretischen und praktischen Wissens Gegenstände und Zustände identifiziert und qualifiziert, z.B. als wünschenswert oder zu vermeiden, als herzustellen oder zu beseitigen, als vorzuziehen oder zu vernachlässigen, als zu verändern oder in Kauf zu nehmen (Näheres hierzu in Kapitel 4).

Drittens muss das Subjekt als Autor seiner Handlungen im Sinne kausaler Urheberschaft erachtet werden können. Die Frage, ob dieses Kriterium jeweils erfüllt ist, führt uns in verzwickte Problemlagen bis hin in den juristischen Bereich, denn mit der bloßen Unterscheidung zwischen hinreichenden Bedingungen (als Ursachen) und notwendigen Bedingungen (ohne deren Vorliegen hinreichende Ursächlichkeit nicht zustande gekommen wäre) ist das Problem keineswegs erledigt. Es können nämlich Sachlagen entstehen, unter denen die Realisierung einer notwendigen Bedingung genau diejenige (einzige) Bedingung ausmacht, ohne die die vordergründig ‚hinreichende' Bedingung als solche gar nicht hätte wirksam werden können. Die Unterscheidung verschwimmt also und erfordert sorgfältiges Beurteilen und Abwägen, weil im Bereich natürlichen Handelns und sozialer Interaktion oftmals die klassischen Strategien naturwissenschaftlich-technischen Experimentierens mit seinen Kausalitätstests durch Weglassen oder Parametervariation nicht greifen. Dies schreibt sich – wie wir sehen werden –

in das Feld maschineller ‚Verantwortlichkeit' bei der Zuweisung kausaler Autorschaft fort.

Viertes Kriterium ist die Unterstellung einer Fähigkeit zur Kontrolle des eigenen Handelns/Agierens sowohl bezüglich der Anpassung, Korrektur, Modifizierung oder Aufgabe der Aktualisierung einer ursprünglich intendierten Handlung (z.B. angesichts veränderter Handlungsumstände) als auch und gerade (und damit einhergehend) die Kontrolle über Handlungsalternativen in ihrer parallel laufenden Entwicklung, und zwar mit Blick auf eine sinnvolle Änderung des Handlungsschemas oder Veränderungen bei der Priorisierung von Zwecksetzungen (z.B. angesichts von Opportunitätskosten). Ein solches kontinuierliches Lernen und Disponieren während des Handelns unterscheidet sich von einem Aktionstyp, der bloß als Umsetzung von vorher Gelerntem, Antrainiertem und Eingeübtem zu begreifen ist.

Schließlich gilt als *fünftes* Kriterium die Fähigkeit, das eigene Handeln unter normativen Gesichtspunkten zu rechtfertigen. Also die Frage ‚Warum hast du das getan?' nicht bloß bezüglich der gesetzten Mittel und Zwecke und entsprechender Funktionalitäten, sondern auch mit Blick auf die Rechtfertigungsfähigkeit dieser Setzungen, also im Zuge einer normativen Reflexion zu beantworten.

Wenn eine solche Verantwortlichkeit gegeben ist, können wir von einer ‚starken' (allein intrinsischen) Verantwortlichkeit sprechen. Diese ist für das Agieren von Mensch-Maschine-Hybriden oder gar für autonomes Agieren von Maschinen nicht zu unterstellen. In diesem Feld kommt eine ‚schwache' Verantwortlichkeit infrage, da partiell im Modus externer Steuerung und Regelung auf das Entscheiden und die Entscheidungsstrategien Einfluss genommen und lediglich für die entstehenden Freiräume ‚Verantwortlichkeit' delegiert wird. Von diesen Arten der Einflussnahme soll gleich die Rede sein; mit der Delegation von Verantwortlichkeit werden die Delegatoren natürlich nicht von einer Verantwortlichkeit für eben diese Delegation entlastet. Schließlich könnte noch von einer dritten Art von Verantwortlichkeit, einer ‚Als-ob'-/simulierten Verantwortlichkeit für Maschinen gesprochen werden. Gemeint wäre damit ein moralanaloges Verhalten, wie es in den soge-

nannten Roboterethiken verhandelt wird. Eine äußere Ähnlichkeit reicht jedoch nicht für eine Verantwortungszuweisung, denn diese wäre dann auch allenfalls im Rahmen einer simulierten Zuweisung möglich, die an die Stelle expliziter Delegation schwacher Verantwortung in Gestalt funktionaler Vorgaben und qua Gestaltung der Trainingsbedingungen der Systeme tritt. Wir gehen dann (z.B. in juristischer Absicht/Haftung) mit den Systemen um, *als ob* sie verantwortlich wären (siehe Kapitel 6).

3. TYPEN VON AUTONOMIE

Zunächst lassen sich drei Typen von Autonomie unterscheiden, wobei der Übergang zwischen den ersten beiden Typen fließend ist (hierzu ausführlich Hubig 2015: Kap. 3.1.1; die Unterscheidung wurde vielerorts aufgegriffen, exemplarisch Schilling et al. 2016: 338-344):

a) *Operative Autonomie*: Sie findet sich in Systemen, deren Agieren Freiheitsgrade in der Wahl der *Mittel* nach Maßgabe von Effizienz und Effektivität aufweist. Zwar ist diese operative Autonomie dahingehend determiniert, dass die Zwecke des Agierens und der Spielraum/die Optionen der Wahl und der Adaption von Mitteln vorgegeben sind; die Maschinen operieren aber nicht determin*istisch* wie Automaten, deren Prozessieren auf entsprechenden Algorithmen fixiert ist. Smarte Assistenzsysteme jeglicher Art fallen unter jene Rubrik, sofern sie im Zuge von Lernprozessen ihr Verhalten in Anpassung an Umweltbedingungen und/oder Nutzerprofile optimieren. Ein Tempomat ist nicht operativ autonom, wohl aber elaboriertere Systeme, welche z.B. das Lenk- und Bremsverhalten unterstützen, Geschwindigkeiten oder generell das Fahrverhalten in Abhängigkeit von Umweltbedingungen und Vigilanzschwellen regulieren, unter vorgegebenen funktionalen Erfordernissen Fahrtrouten auswählen etc. und sich hierbei unter anderem auf einen ‚Erfahrungsschatz' auf statistischer Basis (s.u.), auf Bilanzierung von Chancen und Risiken der Steuerungsprozesse nach vorgegebenen Modellen sowie auf Belastungslimits von Material und Mensch

(z.B. nach Maßgabe von Geschicklichkeit, Konzentration und Ausdauer) stützen. Analoges gilt für die Koaktion mit Robotern, z.B. in Bereichen von Fertigung, Service, Reparatur, Therapie und Pflege (zu den hierbei auftretenden Problemen siehe Kapitel 5). Rahmenbedingung für die Nutzung entsprechender Freiheitsgrade und eine entsprechende Wahl von Steuerungsprozessen ist eine funktionale oder normative Orientierung, deren Ergebnis seitens der Entwickler/-innen bzw. in Kooperation mit Nutzer/-innen an die Systeme delegierbar ist und deren Konkretisierung und Aktualisierung/Verwirklichung innerhalb der Systeme entsprechend simuliert wird, *als ob* rationale Subjekte hier entschieden hätten.

Darüber hinaus ist b) eine *strategische Autonomie* identifizierbar: Auch hier wird kontextsensitiv agiert, jedoch nicht bloß, was die Wahl der optimalen Mittel, also die Wahl von Steuerungsprozessen betrifft. Vielmehr erhöhen sich die Freiheitsgrade auf die Wahl von *Zwecken* unter vorgegebenen allgemeinen Zielen, und zwar nach Maßgabe von deren Verwirklichungschancen und -risiken. Die Festlegung der konkreten Zwecke betrifft hierbei deren Priorisierung, Reihenfolge des Abarbeitens, Intensität relativ zum Aufwand (Grenznutzen), Perfektionsgrad, Taktung und Dauer der Zweckverfolgung sowie Direktheit oder Umweg der Realisierung bis hin zum Abbruch von Aktionen bei wenig aussichtsreich erscheinender Zweckrealisierung. Eine solche strategische Autonomie kann z.B. einem intelligenten System zugesprochen werden, welches eine gesamte Fertigungsanlage – unter Berücksichtigung unter anderem des Stands der Halbzeuganlieferung, eines sich abzeichnendem Ausfalls von Funktionen auf den Fertigungsinseln, Belastungslimits der Arbeitenden, Konsequenzen ausbleibender Zweckrealisierung, Opportunitätskosten, Gefährdung mittelfristiger Nachhaltigkeit der Produktion im Spannungsfeld zu kurzfristiger Erhöhung der Gratifikationen höherstufig – steuert und regelt, kurz: Planungsprozesse realisiert. Eine solche Planungskompetenz unter vorgegebenen Zielen ist graduell, das heißt unter Festlegung höher- oder niedrigschwelliger Eingriffstiefe des Menschen delegierbar. Dies berührt jedoch nicht die *Rechtfertigung* der Typisierung von Kontexten,

das heißt die Unterstellung der Relevanz und Gewichtung von Parametern bei deren Modellierung, und es betrifft nicht die Rechtfertigung von Werthaltungen derjenigen, die jene Planungskompetenz delegieren und dabei die Werthaltungen der von jenen Planungsprozessen Betroffenen zu berücksichtigen haben.

Wenn diese Dimension berührt wird ist eine c) *moralische Autonomie* gefordert: Sie liegt in der Freiheit der *Anerkennung* von Prinzipien der Systemarchitektonik (für b)), wobei unter einer solchen Freiheit der Anerkennung die jeweilige Anerkennung des Selbst als Subjekt der Anerkennung durch sich selbst mitlaufen muss; solcherlei macht im eigentlichen Sinne die Verantwortungsübernahme aus, die nicht (wie bei a) und b)) delegierbar oder *simulierbar* ist.

Mit der Leitdifferenz zwischen dem *Erkennen* von Regeln (einschließlich des Standes ihrer Umsetzung und entsprechender Reaktion hierauf) und dem *Anerkennen* der Validität von Regeln ist die Grenze markiert, unter deren Aspekt die Frage nach einem ‚verantwortlichen Handeln' autonomer Maschinen eine erste Antwort erfahren kann: Die Selbstzuschreibung einer verantwortlichen Anerkennung von Prinzipien und Zielen, das Selbst-Einstehen für Werthaltungen, die in ihrer Gesamtheit eine Persönlichkeit in verantwortlicher Urheberschaft für die Ausrichtung an höchsten Zielen ausmacht, bleibt noch so intelligenten Systemen verwehrt.

4. BEWUSSTSEIN ALS VORAUSSETZUNG FÜR MORALITÄTSFÄHIGKEIT

Aus meiner Sicht spricht nichts dagegen, intelligenten Systemen (wie auch höheren tierischen Spezies) elementares Bewusstsein zuzuweisen. Ein solches umfasst die spezifische Aktivität, angesichts von Erfahrungen als Präsentationen diese in *Repräsentationen* zu überführen. Repräsentationen sind Identifizierungen und Deutungen von Präsentationen als ... (z.B. als anzustreben oder zu vermeiden, als mit diesen und jenen Qualitäten versehen, als disponibel oder indisponibel, als verur-

sacht durch und sich auswirkend auf etc.). Diese Kompetenz zur Repräsentation erlangen ‚intelligente' Systeme im Zuge von überwachtem, unüberwachtem und bestärkendem Lernen (vgl. hierzu einführend Alpaydin 2008). Beim überwachtem Lernen werden Daten verwendet, die von Personen hinsichtlich einer bestimmten Bedeutung gekennzeichnet wurden. Dadurch werden bestimmte Eigenschaften (*Features*) in den Sensordaten zu Indikatoren für einen bestimmten Deutungsinhalt. Der Deutungsrahmen wird dabei von einem gewählten Kategoriensystem vorgegeben (vgl. Kotsiantis 2007; Lake et al. 2016). Beim unüberwachten Lernen gibt es keine Labels; vielmehr werden die Vorschläge für eine Deutung vom System selbst gemacht. Sie basieren auf dem Erkennen von Mustern in den Daten auf Basis einer repräsentativen Menge von Beobachtungen. Beim bestärkenden Lernen werden im Zuge eines Trainings über die Verknüpfung von *Features* und Deutungsinhalt qua Trainingsdaten hinaus systemische Reaktionen optimiert. Hierdurch gewinnen die Präsentationen, die uns durch unsere Sinnesorgane oder Sensorinputs zugänglich werden, Zeichencharakter (auf der Basis bestimmter Träger), vermögen sich aus ihrer situativen Bedingtheit abzulösen und zum Gegenstand von Erinnerung und von Planung zu werden. Sie werden (‚propositionale') Gegenstände eben eines Bewusstseins. Über einen bloßen Charakter als Information hinaus gewinnen solche Gegenstände den Status einer ‚Kontextinformation', wenn sie relevant sind; und dies sind sie, wenn sie bei natürlichen Subjekten eine Rolle in deren intentionaler Ausrichtung bzw. bei Systemen eine Rolle für die Ausfüllung der Systemfunktionen (Instantiierung von Variablen) spielen. Dadurch werden Systeme „kontextsensitiv" (Abowd et al. 1999: 304.; vgl. auch Chalmers/Maccoll 2003). Beim *überwachten* Lernen werden die Trainingsdaten gelabelt, das heißt mit einem Bedeutungsetikett versehen, und beim *unüberwachten* Lernen werden gelabelte Daten als Trainingsdaten verwendet. Für diese sucht das System selbsttätig qua Clustering nach Merkmalen, die eine Gruppe von Samples, die dieses Merkmal haben, von anderen unterscheidet, und damit selbständig Vorschläge zu einer Kategorisierung bzw. zu Ordnungsstrukturen für die Unterscheidung von Gruppen von

Samples entwickelt. Dagegen läuft das *bestärkende* Lernen auf ein strategisches Vorgehen hinaus, auf Basis eines ‚internen Belohnungsmechanismus' qua Belohnungsfunktion die Strategie finden zu lassen, die für möglichst viele äußere Umstände die bestmögliche Teilentscheidung hinsichtlich eines gewünschten Ziels trifft (vgl. Alpaydin 2008: 398). Je nach Lerntyp resultieren entsprechend unterschiedliche Verantwortlichkeitsprofile bezüglich derjenigen Subjekte, die in jeweils spezifischer Weise die Systeme trainieren oder diese sich selbst optimieren lassen.

Solche Systeme vermögen mithin ‚lernend' (ihr Gegenstandsfeld erweiternd, sich korrigierend, sich adaptierend) mit den Informationen umzugehen. Dabei kann dieser Umgang unter vielen Determinanten unterschiedlichster Art stehen (evolutionär gebildeten Filtern, vorgegebenen und entwickelten Strategien der Fusion und Verarbeitung von Sinnesreizen bzw. Sensorimpulsen, Berücksichtigung von Signifikanzschwellen, erlernten Konventionen der Zeichenverwendung und -zuordnung, antrainierten Präferenzen und Aversionen etc.). Unter einer solchen Determiniertheit kann jener Umgang jedoch durchaus – wie gesagt – in sich nicht-deterministisch prozessieren.

Die Repräsentationen – und dies ist für Künstliche Intelligenz relevant – können dabei in Grenzen die Systemzustände des Bewusstseins selbst umfassen; Bewusstseine können ein Selbstmodell bilden, welches unter anderem ihren Informationsstand, ihre Dispositionen des Agierens, ihre Leistungen und Grenzen und in Abhängigkeit hiervon die Bezugsbereiche möglicher Aktionen erfasst und bilanziert. Aus logischen Gründen können Selbstmodelle hingegen niemals ein *vollständiges* und für jedes ihrer Elemente, einschließlich der jeweils vollzogenen Selbstrepräsentation, gültig entscheidbares Selbstmodell bilden: Denn würde die Selbstrepräsentation selbst Element des Selbstmodells sein, wäre sie Objekt der Selbstrepräsentation und nicht mehr diese selbst. Würde sie nicht Element des Selbstmodells sein, wäre dieses unvollständig. Solche Bewusstseinsformen einschließlich ihrer Grenzen können als kognitive Basis den Autonomietypen a) und b) zugeschrieben werden.

Ganz anders verhält es sich hingegen mit einem spezifischen Typ von Bewusstsein, welches Menschen sich als Selbstbewusstsein zuschreiben. Es handelt sich hier nicht um ein Bewusstsein *über* das Selbst (als eine bestimmte Repräsentation), sondern um ein Bewusstsein *vom* Subjektcharakter eines Selbst. Von diesem Selbst als Subjekt ‚kann man sich kein Bildnis machen', denn es ist ja die vorauszusetzende Instanz der Produktion von Repräsentationen/Bildnissen. Zu einem solchen Bewusstsein als Selbstbewusstsein kann man sich nur im Modus der *Anerkennung* eines Selbst *entscheiden,* indem man sich höherstufig als verantwortlich dafür anerkennt, Prozesse des Anerkennens von Regeln, Prinzipien, Zielen zu vollziehen und gemäß dieser Anerkennung zu handeln. (Man kann sich auch diese Anerkennung verweigern, indem man sich distanziert und darauf verweist, dass man irgendwie fremdbestimmt war – ‚Das war nicht ich!') Dies betrifft auch und gerade die Notwendigkeit, sich zu den zahlreichen Prozessen, die wir als uns selbst determinierende Prozesse (qua Sozialisation, neuronale Verfasstheit, Konsumgewohnheiten etc.) erkennen, in ein Verhältnis zu setzen: Sobald solche Prozesse nämlich identifiziert sind, stehen wir notwendigerweise in einem Verhältnis zu ihnen, welches unabhängig von der Möglichkeit einer faktischen Einflussnahme auf diese Prozesse uns zumindest in die Position bringt, diese Prozesse als solche zu billigen, denen wir uns fügen wollen, oder die wir in dem Sinne nicht anerkennen, dass wir alles ins Werk setzen, eine Einflussnahme auf diese Prozesse zu erreichen. Insofern ist dies eine, vernünftigen Wesen vorbehaltene, Form eines Selbstbewusstseins, zu der man sich entscheiden kann oder nicht (im letzteren Falle machen wir uns – eine wählbare Option – zum Tier, begeben uns in eine Abhängigkeit oder lassen es dabei sein). Diese anerkennungsbasierte Form des Selbstbewusstseins kann nicht in Systeme Künstlicher Intelligenz implementiert werden. So wäre durchaus vorstellbar, dass Systeme Künstlicher Intelligenz im Rahmen ihres Selbstmodells feststellen, dass ihnen bestimmte Informationen fehlen, Lernprozesse noch nicht abgeschlossen sind, bestimmte Repräsentationen sich noch nicht stabilisiert haben bzw. zu hohe Schwankungsbreiten aufweisen etc. und sie

daher bestimmte Fragen nicht beantworten (und dies auch explizit ausdrücken) oder bestimmte Aktionen nicht ausführen können. Nicht aber können sie sich dahingehend äußern, dass sie angesichts einer bestimmten als notwendig erachteten funktionalen Vorgabe durchaus die Gelegenheit gehabt hätten, sich das entsprechende Knowhow und die entsprechenden Dispositionen des Agierens anzueignen, dies aber nicht gewollt hätten mangels Anerkennung der zugrunde liegenden Prinzipien.

Angesichts der Leitdifferenz zwischen *Bewusstsein* und *Selbstbewusstsein* (welches ein bloßes Selbstmodell überschreitet) kann die Antwort auf die Eingangsfrage nun dahingehend weiter entwickelt werden, dass Systemen Künstlicher Intelligenz mangels Selbstbewusstsein in diesem emphatischen Sinne die basale Voraussetzung für eine (nicht simulierte) Verantwortlichkeit und erst recht für ihre ethische Reflexion abzusprechen ist.

5. INTERAKTION MIT INTELLIGENTEN SYSTEMEN/ROBOTERN UND DIE HYBRIDISIERUNG VON AKTEUREN

Wenn es um den Umgang mit intelligenten Systemen geht, wird landläufig von Interaktion gesprochen. Keineswegs soll hier dieser Sprachgebrauch pauschal verurteilt werden, denn Begriffe unterliegen nicht gewissermaßen einem ‚Markenschutz'. Gleichwohl ist anzumerken, dass im Rahmen dieses Sprachgebrauchs einige Pointen verloren gehen, die eine präzisere Begriffsverwendung mit sich führen. Das lässt sich insbesondere geltend machen mit Blick auf sogenannte moralische Dilemmata, in denen sich künstliche intelligente Systeme angeblich verfangen können. Gängige Beispiele hierfür sind das Verhalten autonomer Automobile im unvermeidlichen Kollisionsfall oder der Einsatz intelligenter Kampfdrohnen.

Um die Sicht auf die Problemlage zu schärfen, ist es sinnvoll, zunächst auf die engere Fassung des Interaktionskonzepts zu verweisen,

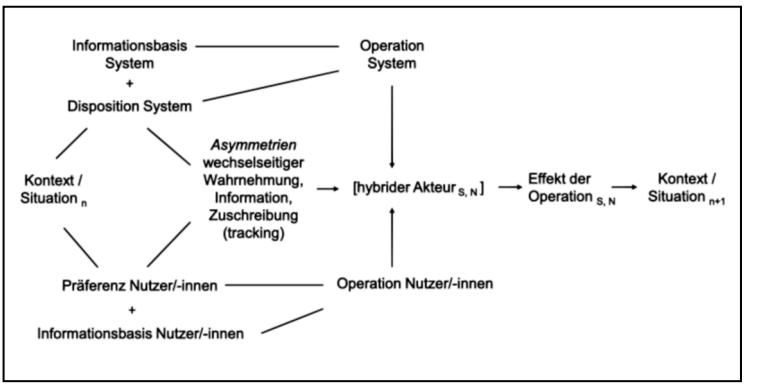
wie sie im Rahmen der Soziologie und der allgemeinen Handlungstheorie etabliert ist: Versteht man dort unter ‚Aktion' einen Prozess, der mit der *Erwartung* über die Realisierung eines Zwecks verbunden und durch diese motiviert ist, so bezeichnet ‚Interaktion' eine Aktionsform, die auf Aktionen anderer gerichtet ist und sich mithin an Erwartungen über die Erwartungen anderer, sogenannte *Erwartungserwartungen*, orientiert (vgl. Luhmann 1984: 412; Weber 1968: 441). Diese Erwartungserwartungen werden im Idealfall ständig einem kommunikativen Abgleich unterzogen (sodass in einem solchen Idealfall die Erwartungen der Entwickler/-innen über die Erwartungen der Nutzer/-innen und die Erwartungen dieser über die Erwartungen der Entwickler/-innen harmonieren). Solcherlei setzt entsprechende höherstufige elaborierte Kommunikationsprozesse voraus.

Für den Umgang mit intelligenten Systemen (einschließlich Robotern) sind solche höherstufigen Kommunikationsprozesse nicht in Sicht. Vielmehr findet die wechselseitige Bildung von Erwartungen bei ihren Trägern, Mensch und Roboter, *unabhängig* statt, und das Zusammenwirken menschlicher und künstlicher Systeme basiert auf unabhängiger und jeweils unterschiedlicher Art der Erwartungsbildung (vgl. Braun-Thürmann 2002: 15, 145). Die Systeme verfolgen qua *Tracking* die Aktionen menschlicher Nutzer/-innen und bilden hierauf basierend Muster bezüglich der eigenen Erwartungen, der Wertung der jeweiligen Elemente und Komponenten der Erwartung und der Wertung der Aktualisierung dieser Erwartungen in konkreten Aktionen (vgl. Fink 2009: 11; Krummheuer 2010: 105). Die natürlichen Interaktionspartner werden auf Basis ihrer Gewohnheiten und Handlungsroutinen auf Nutzerstereotype und Adressatenprofile reduziert (vgl. hierzu Hubig 2015: 138). Hieraus resultieren seitens der Systeme bestimmte *Dispositionen* des Agierens, die in den einzelnen Schritten ihres Zustandekommens für die Nutzer/-innen und in weiten Bereichen auch für die Entwickler/-innen intransparent bleiben. Lediglich in die Gestaltung der Trainings*bedingungen* der Systeme bei ihrer Musterbildung und in die *Ergebnisse* des Agierens kann Einsicht genommen und über sie bilanziert werden. Erst recht kann sich die Musterbildung der

Systeme von konkreten Interaktionsprozessen lösen, wenn zusätzlich die Übernahme bestimmter Muster aus einschlägigen Kontexten des *World Wide Web* die Bildung von Dispositionen in den Systemen zusätzlich beeinflusst. Umgekehrt haben die Nutzer/-innen nicht mehr die Möglichkeit, ihre Erwartungserwartungen an die Entwickler/-innen der Systemarchitektur angesichts des Verhaltens der Systeme zu verifizieren, zu korrigieren, fortzuschreiben, weil das Feld der Determinanten der Bildung systemischer Dispositionen unüberschaubar und intransparent ist (vgl. Feiner 2002).

Daher wird neuerdings zur Charakterisierung dieser Vollzüge eher der Begriff ‚Koaktion' eingesetzt, und es ist die Rede von ‚hybriden Akteuren', wodurch signalisiert werden soll, dass zwei Agenten, die in ihrer Bildung von Mustern und Dispositionen auf der einen Seite und von Erwartungserwartungen auf der anderen Seite geleitet sind, ohne direkte Kommunikation, gleichsam arbeitsteilig, Handlungseffekte zeitigen (vgl. Rammert/Schulz-Schaeffer 2002: 11-64; Weyer/Fink 2011: 43).

Abbildung 1: Koaktion mit Autonomen Systemen



Quelle: eigene Darstellung

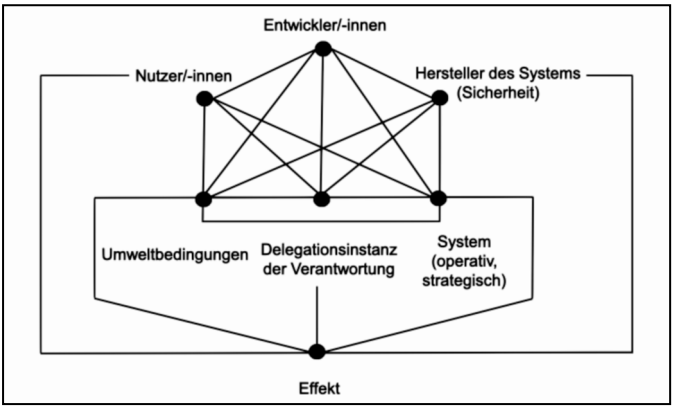
Sofern sichergestellt werden kann, dass die Systemdispositionen funktional äquivalent sind mit den Erwartungen der Nutzer/-innen, kann im Rahmen solcher Koaktionen eine gelingende Interaktion zumindest simuliert werden (z.B. beim Einsatz von Kuschelrobotern für Demenzkranke). Die Koaktion gerät jedoch an ihre Grenzen, sofern menschliche Akteur/-innen intrinsisch, das heißt ohne extern beobachtbare und registrierbare Stimuli, im Zuge von Abwägungen ihrer Erwartungen und Änderung ihrer Werthaltungen die Realisierung von Zwecken nicht zulassen oder verändern oder gewissermaßen die intelligenten Systeme ‚auszutricksen' suchen. Ein Beispiel hierfür wäre ein nicht auszuschließender Umgang mit autonomen Automobilen derart, dass aufgrund von deren Disposition, bei unvermeidlichen Kollisionen Personenschäden zu minimieren, Anreizsysteme für die Nutzung möglichst schwach gesicherter Automobile entstehen (weil die stärker geschützten bevorzugt angesteuert werden) oder bei vollbesetzten Fahrzeugen abwegige Lizenzen zum nicht-regelkonformen Verkehrsverhalten in Anspruch genommen werden (weil die Verursacher im Kollisionsfall eher verschont bleiben). Auch können zum Zwecke der Entledigung von Eigenverantwortung den Systemdispositionen allzu hohe Entscheidungskompetenzen zugewiesen werden, wenn es – wie im Falle von Kampfdrohnen – darum geht, potenzielle Kombattanten zu identifizieren und Fehlertoleranzen bei der Unterscheidung von Zivilpersonen und militärischem Personal in Kauf genommen werden.

Auch hier zeigen sich unüberbrückbare Leitdifferenzen, wenn auf der Menschseite qualitative Erfahrungen wie Freude, Bedauern, Reue, die bestimmten Werthaltungen geschuldet sind, auf quantitativ-statistische Erfahrungen der Systeme (basierend auf ihrem jeweiligen Training) treffen, und umgekehrt deren individuell adaptive, lernende Musterbildung auf die Etablierung menschlicher Verhaltensmuster trifft, die sozialem Lernen, Rechtfertigungskulturen und moralischen Diskursen geschuldet sind.

6. VERANTWORTUNGSNETZWERKE

Angesichts der Hybridisierung des Menschen durch autonome Maschinen – wie in der Mensch-Roboter-Koaktion – (und umgekehrt einer ‚Hybridisierung' der Maschinen durch ihre Einbindung in Koaktionen mit Menschen, die die Maschinen nicht mehr bloß bedienen, sondern Instanz ihrer Lernprozesse sind) wäre zunächst kompromisslos auf Transparenz zu setzen, was die Trainingsbedingungen der Systeme, die Architektur der Herausbildung ihrer Dispositionen und die Bilanzierung der Aktionseffekte betrifft. Damit müsste eine adäquate Herausbildung von ‚Verantwortungsnetzwerken' einhergehen, innerhalb derer die Delegationsprozesse für Autonomie im Sinne von a) und b), die Möglichkeiten einer Verantwortungszuweisung in Haftungsfällen und die Prozesse der Bildung von Erwartungen bzw. Systemdispositionen der Beteiligten so verhandelt werden, dass Verantwortungsdiskurse *über* autonome Maschinen präzise und sachangemessen geführt und Voraussetzungen für einschlägige Verrechtlichungsprozesse erbracht werden. Solche Forderungen sehen sich freilich mit einer überaus hohen Komplexität der Sachlage konfrontiert.

Abbildung 2: Verantwortungsnetzwerk



Quelle: eigene Darstellung

Die Schwierigkeit gründet darin, dass in einem rekonstruierbaren Netz von untereinander wechselwirkenden Einflussgrößen für das Zustandekommen eines Effekts nicht auf einfachem Wege Verantwortlichkeiten in ihren jeweiligen Abhängigkeiten den Wirkungsrelationen zugeordnet werden können. Dies betrifft insbesondere Versuche einer Typisierung, Unterscheidung und einer jeweiligen Zuordnung von menschlicher (intentionaler), maschineller (technischer) Verantwortlichkeit oder aus ‚natürlichen' Umständen (Verfasstheit der Handlungsumgebung, des Materials und weiterer, nicht disponibler Gegebenheiten bis hin zu ‚Zufällen') resultierender Maßgeblichkeit. Denn bei allen Akteuren als Knoten in diesem Netzwerk finden wir eine Hybridisierung, in der intentionale, technische und natürliche Verfasstheiten sich überlagern und wechselwirken. Entsprechend können wir bestimmte Effekte, Teileffekte oder Seiten solcher Effekte nicht einfach auf das Konto intentionaler Verantwortung, technischer (delegierter) Verantwortung oder ‚der Natur' verbuchen. Der Soziologe Bruno Latour hat daher im Rahmen seiner Akteur-Netzwerk-Theorie (ANT) gefordert, die Rede von ‚der' menschlichen Subjektivität/Intentionalität, ‚der' Technik und ‚der' Natur zu verabschieden und stattdessen von ‚intentionalen Fragmenten', ‚technischen Fragmenten' und ‚natürlichen Fragmenten' zu sprechen, die sich nach Maßgabe der Relationen, in denen sie untereinander stehen, allererst entwickeln und aktualisieren (vgl. Latour 2006a, 2006b; hierzu ausführlich Hubig 2015: 88-95). So bilden wir unsere Absichten, Erwartungen und Präferenzen (insbesondere was deren Hierarchisierungen und Priorisierungen je nach Situation betrifft), einschließlich der entsprechenden Entscheidungen unter dem Einfluss der entsprechenden technischen Handlungsumgebung sowie den Artefakten samt ihrer materialen Verfasstheit mit ihren förderlichen und widerständigen Eigenschaften. Die technischen Komponenten wiederum verbleiben im Zustand toter Materialität, sofern sie nicht in diejenigen Handlungsprozesse eingebunden sind, in deren Lichte sie Funktionalitäten und ihre weiteren Eigenschaften anzunehmen vermögen. Und ‚Natur' gewinnt ihren Status allererst in ihrer Relation zu uns und zu möglichen technischen Funktionalitäten einschließlich ihrer inneren

Organisation nach Maßgabe technisch induzierter Modellierung und der Qualitätszuschreibung im Lichte weiterer (bis hin zu ästhetischer) intentionaler Konzeptualisierung. In allen Instanzen des Netzwerks können wir solche Relationen antreffen, wenn wir die Knoten nicht als ‚Blackboxes' annehmen, sondern ihre eigene interne Vernetzung freilegen. Diese Problemlage potenziert sich natürlich, wenn wir ein Netz in seiner Gesamtheit betrachten und die im Netz gezeitigten Effekte eben nicht mehr einfach auf Menschseite, Technikseite oder Naturseite verbuchen können. So hat Latour einschlägige Effekte von der Nutzung von Schlüsseln und Waffen über technische Maßnahmen der Verkehrsberuhigung, der Optimierung der Austernzucht u.v.a. mehr bis hin zu komplexen Netzen naturwissenschaftlicher Laborforschung, der Gestaltung von Verkehrsinfrastrukturen und ökologischen Projekten im Umgang mit dem Sahel-Syndrom akribisch untersucht und ausbuchstabiert. In diesem Zusammenhang wird regelmäßig deutlich, dass wir unter dem Problemdruck einer Orientierung unserer Aktionen in den Netzen zu Vereinfachungen, eben dem Blackboxing, gezwungen sind.

Solcherlei lässt sich auch angesichts unserer Problemlage beobachten und schließt den simplifizierenden Gebrauch des Titelwortes ‚Verantwortlichkeit' ein. Denn mit gewissen Einsichten, die sich aus einer differenzierteren Analyse ergeben, sind nicht zwangsläufig Fortschritte und Hilfen für den praktischen Umgang mit den Systemen verbunden. Im Bereich des Rechts findet dies seinen Niederschlag in der Diskussion, ob ein neues juristisches Konstrukt, die ‚e-Person' als Haftungsträger eingeführt werden soll (wie es vom EU-Parlament 2017 vorgeschlagen wird); sie könnte in zivilrechtlichen Verfahren verklagbarer Adressat für Haftungsansprüche werden und für die Schadenskompensation mit einem Fonds ausgestattet sein, der aus Ressourcen aller Akteure des Netzes rekrutiert würde, wobei die Proportionierung zu klären wäre. Eine andere Strategie läge in der Ausweitung der Gefährdungshaftung (die auch ohne Unterstellung eines Vorsatzes oder einer Fahrlässigkeit greift) über „bewegliche Sachen" hinaus (ProdHaftG §2), wie sie etwa für Fahrzeughalter/-innen unabhängig

von der persönlichen Nutzung des Fahrzeugs besteht. Im Zuge dieser Ausweitung könnten gar auch Nutzer/-innen unter dem herkömmlichen Leitbild der Gefährdungshaftung, dass diejenigen die Haftungsrisiken tragen sollen, die Vorteile aus der Nutzung der Systeme ziehen, mit involviert werden (vgl. Günther 2016: 237-255). Auch eine Ausweitung der Fahrlässigkeit bzw. der Fahrlässigkeitskriterien unter Gesichtspunkten der Vorsorge und Sicherung kämen infrage, wobei dies jedoch in die Komplexität der Zuordnung von Zuständigkeiten zurückführen würde. Dies insgesamt ist aber nur eine Seite der Problematik.

Unter Verweis auf die Nicht-Delegierbarkeit moralischer Verantwortung folgt nämlich auf der anderen Seite, dass bezüglich der Großarchitektur der Systeme mit ihren sozialen und ökonomischen Konsequenzen auch klassische Strategien des Verweises von Zuständigkeiten an Politik, Wirtschaft, soziale Interessenvertretungen etc. an ihre Grenzen kommen und interessenfixierte Ziele in den Aushandlungsprozessen nicht mehr problemlos fortgeschrieben werden können (z.B. Erhalt von Arbeitsplätzen, Erhalt von Konkurrenzfähigkeit in bestehenden Marktstrukturen, Organisation von Bildungschancen angesichts veränderter Anforderungsprofile etc.). Die Herausforderungen an die Technikgestaltung und -nutzung sind so hoch, dass es nicht mehr zielführend erscheint, die Gesamtentwicklung dem Wechselspiel eines nutzenorientierten Interessenausgleichs zu überlassen, dessen Stakeholder jeweils als Einzelne und aus ihrer jeweiligen Binnenperspektive versuchen, das Bestmöglichste für sich herauszuholen. Das betrifft sowohl die einzelnen Individuen als auch ihre Interessenverbände. Man muss nicht gleich Technokrat sein, um einen verstärkten Ausbau einer Ebene von Gremien zu fordern, die auf der Ebene einer Bilanzierung, Evaluation, Supervision und eines Monitoring der Entwicklung das Ziel zu verfolgen hätten, die politischen Kompetenzen sowohl der Legislative als auch derjenigen, die ihr die Macht verleihen, zu unterstützen und gleichzeitig zu entlasten: für erstere von der kurzfristigen Fixierung auf Machterhalt und -erweiterung, für letztere von der notwendigen Fixierung auf Erhalt und Erweiterung der Optionen ihrer individuellen Reproduktion auf Basis der problematisch gewordenen Arbeitsprozesse.

Für letztere würde dies einen hinreichenden Ausbau der Grundsicherung erforderlich machen, die eine gewisse Unabhängigkeit und mit Zeitressourcen versehene Anpassung an die Umwälzungen in der Arbeitswelt ermöglicht. Eine dialogische Abstimmung zwischen den Seiten der Entwicklung und denjenigen der Nutzung über Fragen der Architektur der Systeme sollte nicht mehr domänenspezifisch und quasi arbeitsteilig innerhalb der Wirtschaft, der Wissenschaft oder des Agierens der Verbände allein stattfinden, sondern in einem umfassenderen Rahmen, der von allen Beteiligten organisiert, getragen und gestaltet wird und bis zu einem gewissen Grade Appellations- und Legitimationsinstanz für politisches Entscheiden (mit) abgeben könnte. Hieraus würde für die Politik selbst auch eine Entlastung resultieren, und hierzu gibt es im Kleinen bereits Ansätze, z.B. die Chemie-Stiftung Sozialpartner-Akademie (CSSA) oder weitere einschlägige Institute gemeinsamer Trägerschaft von Unternehmensverbänden, Gewerkschaften und wissenschaftlichen Fachverbänden. Für diejenigen schließlich, die als Individuen in den Systemen arbeiten und mit smarten Systemen interagieren, sind die Optionen auszubauen, sich on demand bei Irritationen *parallel* zur Mensch-System-Kommunikation (vgl. hierzu Hubig 2007: 8.5; Wiegerling 2011 mit Report der Erträge des Sonderforschungsbereichs [SFB] 627 ‚Nexus') vertrauenswürdig und transparent auszutauschen nicht nur über die operativen Fragen, sondern auch über Hintergründe der strategischen Auslegung der Systeme, über Perspektiven ihrer Entwicklung bis hin zu Fragen von Ausbildungs- und Qualifikationserfordernissen, erforderlicher Fortschreibung und Entwicklung von Kompetenzen und Situierung der eigenen Position bzw. Partizipation in diesem Prozess. Diesbezüglich bedürfen wir einer neuen Beratungskultur, innerhalb derer auch alternative Lebensentwürfe und Wege beruflicher Verwirklichung in neu entstehenden Berufsfeldern jenseits herkömmlicher Vorstellungen von Karrierechancen und Ausbildungswegen reflektierbar werden. Nur so erscheint es möglich, eben nicht mehr der ‚Digitalisierung' eine Verantwortlichkeit für gesellschaftliche Umbrüche zuzuschreiben, die doch bei der Gesellschaft selbst verbleibt, sofern diese in der Lage ist, ihre Kräfte zu koordinieren und

nicht einfach einem Machtgeschehen zu überlassen, als dessen verantwortlicher Träger dann eine ‚maschinelle Intelligenz' erscheint.

LITERATUR

Abowd, Gregory D./Dey, Anind K./Brown, Peter J./Davies, Nigel/Smith, Mark/Steggles, Pete (1999): „Towards a Better Understanding of Context and Context-Awareness", in: Hans-Werner Gellersen (Hg.), Proceedings of the 1st International Symposium on Handheld and Ubiquitous Computing (= Lecture Notes in Computer Science 1707), London: Springer, S. 304-307.

Alpaydin, Ethem (2008): Maschinelles Lernen, München: Oldenbourg.

Braun-Thürmann, Holger (2002): Künstliche Interaktion. Wie Technik zur Teilnehmerin sozialer Wirklichkeit wird, Wiesbaden: Westdeutscher Verlag.

Chalmers, Matthew/Maccoll, Ian (2003): „Seamful and Seamless Design in Ubiquitous Computing", in: Proceedings of Workshop at the Crossroads: The Interaction of HCI and Systems Issue in UbiComp, http://www.academia.edu/2456375/Seamful_and_Seamless_Design_in_Ubiquitous_Computing vom 23.8.2018.

EU-Kommission (2017): Bericht mit Empfehlungen an die Kommission zu zivilrechtlichen Regelungen im Bereich Robotik, A8-0005/2017 vom 27.1.2017, http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+REPORT+A8-2017-0005+0+DOC+XML+V0//DE vom 22.8.2018.

Feiner, Steven K. (2002): „Augmented Reality: A New Way of Seeing", in: Scientific American 286/4, S. 48-55.

Fink, Robin D. (2009): Attributionsprozesse in hybriden Systemen. Experimentelle Untersuchung des Zusammenspiels von Mensch und autonomer Technik (= Soziologisches Arbeitspapier 25), Dortmund: Technische Universität Dortmund.

Günther, Jan-Philipp (2016): Roboter und rechtliche Verantwortung. Eine Untersuchung der Benutzer- und Herstellerhaftung, München: Herbert Utz (insbes. S. 237-255).

Hubig, Christoph (2007): Die Kunst des Möglichen 2. Ethik der Technik als provisorische Moral, Bielefeld: transcript.

Hubig, Christoph (2015): Die Kunst des Möglichen 3. Macht der Technik, Bielefeld: transcript.

Kotsiantis, Sotiris B. (2007): „Supervised Machine Learning: A Review of Classical Techniques", in: Informatica 31/2007, S. 249-268.

Krummheuer, Antonia (2010): Interaktion mit virtuellen Agenten? Zur Aneignung eines ungewohnten Artefakts, Stuttgart: Lucius & Lucius.

Lake, Brenden M./Ullmann, Tomer D./Tenenbaum, Joshua B./Gershman, Samuel J. (2016): „Building Machines That Learn and Think Like People", in: Behavioral and Brain Sciences, S. 1-101.

Latour, Bruno (2006a): „Sozialtheorie und die Erforschung computerisierter Arbeitsumgebungen", in: Andréa Belliger/David J. Krieger (Hg.), ANThology. Ein einführendes Handbuch zur Akteur-Netzwerk-Theorie, Bielefeld: transcript, S. 561-572.

Latour, Bruno (2006b): „Über technische Vermittlung: Philosophie, Soziologie und Genealogie", in: Andréa Belliger/David J. Krieger (Hg.), ANThology. Ein einführendes Handbuch zur Akteur-Netzwerk-Theorie, Bielefeld: transcript, S. 483-528.

Luhmann, Niklas (1984): Soziale Systeme. Grundriss einer allgemeinen Theorie, Frankfurt a.M.: Suhrkamp.

Rammert, Werner/Schulz-Schaeffer, Ingo (2002): „Technik und Handeln. Wenn soziales Handeln sich auf menschliches Verhalten und technische Artefakte verteilt", in: dies. (Hg.), Können Maschinen handeln? Soziologische Beiträge zum Verhältnis von Mensch und Technik, Frankfurt a.M. u.a.: Campus, S. 11-64.

Schilling, Malte/Kopp, Stefan/Wachsmuth, Sven/Wrede, Britta/Ritter, Helge (2016): „Towards a Multidimensional Perspective on Shared

Autonomy“, in: The 2016 AAAI Fall Symposium Series: Shared Autonomy in Research and Practice Technical Report, S. 338-344.

Weber, Max (1968): Gesammelte Aufsätze zur Wissenschaftslehre, Tübingen: Mohr Siebeck.

Weyer, Johannes/Fink, Robin D. (2011): „Die Interaktion von Mensch und autonomer Technik in soziologischer Perspektive“, in: Technikfolgenabschätzung – Theorie und Praxis, 20/1, S. 39-45.

Wiegerling, Klaus (2011): Philosophie intelligenter Welten, Paderborn: Wilhelm Fink.