ThPh 94 (2019) 526–555

Verteilte Moral in Zeiten von *KI*?

Über die moralische Bedeutung technischer Artefakte in der Mensch-Maschine-Interaktion

Von Alexis Fritz und Wiebke Brandt

1. Die moralische Bedeutung von Technik im Dieselskandal

Ein unscheinbares, graues, metallenes Kästchen beschäftigt weltweit seit Jahren die Automobilbranche, Gerichte, Politik, Behörden und Privatleute. Es erschüttert bewährte Geschäftsmodelle wie robuste Volkswirtschaften. Die Rede ist von der *Electronic Diesel Control* (*EDC*). Dieses Gerät ist in den meisten Kraftfahrzeugen zu finden; es steuert unter anderem die Leistung und das Drehmoment von Motoren, aber auch, wie viel Schadstoffe ausgestoßen werden. Im sogenannten „Dieselskandal" wurde das Steuerungsgerät so programmiert, dass es einen genormten Prüfzyklus erkannte und das Abgasreinigungssystem gezielt auf die Testanforderungen hin optimierte. Dadurch war der Schadstoffausstoß geringer als im realen Fahrbetrieb. Für den Normalbetrieb schaltete das Gerät diesen Testmodus wieder ab.

Nach dem Bekanntwerden der manipulativen Abschaltvorrichtung wurde schrittweise das multilaterale Versagen der Beteiligten deutlich und die Frage nach den Verantwortlichen gestellt: Darf ein Zulieferer, der in enger Abstimmung mit dem Hersteller ein Gerät entwickelte und um dessen Missbrauchspotential wusste, einfachhin sagen, dass für die Geräteapplikation allein der Hersteller verantwortlich ist? Darf ein Konzern behaupten, bestehende Regelungen lediglich ausgereizt zu haben oder dass die eigenen Entscheidungsstrukturen äußerst komplex und intransparent sind? Waren Politik, Ämter und Behörden, Verbände und Vereine dem bereits vorher bestehenden Manipulationsverdacht entschieden genug nachgegangen? Ja, sollten nicht Kunden ihr eigenes Mobilitätsverhalten und ihr Vertrauen darauf, dass Verbrennungsmotoren gleichzeitig bezahlbar, abgasarm und leistungsstark sein können, kritischer hinterfragen?

Es fällt auf, dass dieser Frageparcours nach den Schuldigen eine zentrale Komponente der Manipulation kaum thematisiert: Es ist dieses kleine graue Kästchen, welches „von selbst" und „intelligent" die Abgasregeltechnik auf verschiedene Fahrsituationen einzustellen weiß. Das Abgasverhalten kann im Prüfungsmodus auch durch andere Faktoren (zum Beispiel eine volle Batterie, halb abgefahrene Reifen oder keine Extraausstattung) optimiert werden. Doch der manipulative Einsatz Künstlicher Intelligenz (*KI*) scheint etwas qualitativ Neues zu sein. Die Technikvergessenheit in der Diskussion über den Skandal verwundert auch deshalb, da zentrale Ansätze der Technikethik betonen, dass technische Artefakte (*tA*) moralisch höchst relevant

sind oder gar als *moral agents* bezeichnet werden können. Erst wenn die moralische Signifikanz technologischer Komponenten anerkannt werde, könne das komplexe Handlungsgeflecht von Akteuren hinreichend geklärt und ethisch bewertet werden.

Dabei ist die Bezeichnung von Computersystemen als *moral agents* keineswegs ein belangloses Umetikettieren, eine ästhetische Korrektur oder Modeerscheinung. Vielmehr hat diese auch in den Wissenschaften zunehmende Sprachpraxis das Potential, das Fundament der theologischen und philosophischen Anthropologie zu erschüttern. Ist es nicht unser Alleinstellungsmerkmal, unsere originäre und genuine Auszeichnung in der Sphäre der Sittlichkeit und in der Schöpfung, *moral agent* zu sein? In der Regel fassen Philosophinnen und Philosophen den Begriff *agency* sehr eng und kennzeichnen damit die Durchführung von intentionalen, auf Gründen basierenden, absichtlichen Handlungen.[1] Die vermutlich eindrücklichste biblische Erzählung zu *moral agency* findet sich in Gen 2–3. Gott schafft den Menschen als freies Wesen und in dieser Freiheit entdeckt der Mensch die erschreckende Möglichkeit, seinem Schöpfer widersprechen und seine Mitmenschen wie seine Umwelt schädigen zu können.[2] Spätestens seit dem *linguistic turn* des 20. Jahrhunderts sind wir dafür sensibel, dass die Sprache eine unhintergehbare Bedingung des Denkens ist, das heißt, dass die Wirklichkeit jenseits der Sprache nicht erreichbar ist. Das Verständnis von Computersystemen als *moral agents* ist keine Weiterentwicklung, sondern eine Zäsur zum herkömmlichen philosophischen wie theologischen Denken. Es gibt Anlass, der Tatsache auf den Grund zu gehen, dass es uns irritiert, wenn im Bereich der Pflege und Medizin Algorithmen und Roboter das fachkundige Personal nicht nur unterstützen, sondern ersetzen sollen, wenn uns Computer zu unserem Geburtstag gratulieren oder wenn intelligente Kampfroboter ihre Opfer suchen und töten.

Im Folgenden werden drei untereinander konkurrierende Erklärungsmodelle kritisch diskutiert, welche die durch die Komplexität der Mensch-Maschine-Interaktion hervorgerufenen Probleme für eine moralische Bewertung und Verantwortungszuschreibung entschärfen wollen, indem sie den technischen Teil mit *agent* oder *moral agent* bezeichnen. *Agent* kann im Englischen sowohl Personen bezeichnen, die (gegebenenfalls auch in Stellvertretung oder in geheimer Sache) handeln, als auch Ursachen, das heißt Dinge oder Personen, die bestimmte Effekte oder Veränderungen hervorrufen. In ihrem Ausgangsproblem und in ihrem Grundanliegen stimmen alle drei systemischen Modelle überein; dessen ungeachtet entwickelt Luciano Floridi einen technozentrischen, Deborah G. Johnson einen anthropozentrischen und Peter-Paul Verbeek einen phänomenologisch-konstruktivistischen

[1] Vgl. *G. E. M. Anscombe*, Intention, Oxford 1957; *D. Davidson*, Actions, reasons and causes, in: *Ders.*, Essays on Actions and Events, Oxford 1980, 3–20.

[2] Vgl. *Th. Pröpper*, Theologische Anthropologie; Band 2, Freiburg i. Br. 2011, 694–744.

Antwortversuch. Die vorliegende Untersuchung kommt zu dem Schluss, dass die Bezeichnungspraxis von Technik als *agent* oder *moral agent* weder aus rein deskriptiver noch aus normativ-ethischer Sicht überzeugt. Vielmehr wird dadurch in Kauf genommen, dass die Verantwortung in komplexen Beziehungsgefügen nicht mehr oder nur unzureichend geklärt werden kann. Daher sollte der Gebrauch des *(moral) agency*-Begriffs auf menschliche Akteure beschränkt bleiben.

Um dies zu verdeutlichen und um die Stärken und Schwächen der verschiedenen Ansätze aufzuzeigen, wird anschließend der VW-Dieselskandal aus der Perspektive jedes der drei Modelle analysiert.

2. Drei Theorien zur moralischen Bedeutung technischer Artefakte

2.1 Luciano Floridi

2.1.1 Kritik an der bisherigen „Standardethik"

Mit der unübersichtlichen Gemengelage, in der das Zusammenwirken von Mensch und Maschine häufig nicht mehr nachvollziehbar ist und die Frage nach der Verantwortung folglich nicht zufriedenstellend geklärt werden kann, sieht Floridi alle bisherigen Ethik-Theorien überfordert. Seiner Analyse zufolge sind die – von ihm im Begriff der „Standardethik"[3] zusammengefassten – herkömmlichen Ansätze den neuen Herausforderungen in zweierlei Hinsicht nicht gewachsen:

Zum einen greife ihre exklusiv menschliche Konzeption von *moral agency* zu kurz. Zwar sei der Kreis der *moral patients* infolge eines ökologischen Umdenkens in der Ethik inzwischen auch auf nicht-menschliche Entitäten (insbesondere Tiere) ausgeweitet worden. Als *moral agents* kämen aber weiterhin ausschließlich menschliche Personen[4] in Betracht. Dieses Ungleichgewicht im Mengenverhältnis von *moral agents* und *moral patients* führe jedoch dazu, dass deutlich mehr Verantwortung auf dem Individuum laste. Vor allem aber behindere ein solch anthropozentrisch verengtes *moral agent*-Konzept die sachgerechte Analyse von Mensch-Maschine-Interaktionen.[5]

Zum anderen setze die Standardethik auf einer grundsätzlichen Ebene falsche Prioritäten, indem sie ihr Hauptaugenmerk auf den einzelnen *moral agent* und die sein Handeln motivierenden Absichten richte. Sie sehe ihre

[3] Zu den „Standardethiken" oder auch „Standard-Makroethiken" rechnet Floridi deontologische wie konsequentialistische Ethik-Ansätze (vgl. *L. Floridi/J. W. Sanders*, Artificial Evil and the Foundation of Computer Ethics, in: Ethics and Information Technology 3 (2001) 55–66, hier 57, 64 f.).

[4] Eine geringfügige Ausweitung des Konzepts habe immerhin in Bezug auf juristische Personen wie Unternehmen u. a. stattgefunden (vgl. *L. Floridi/J. W. Sanders*, On the Morality of Artificial Agents, in: Minds and Machines 14 (2004) 349–379, hier 350).

[5] Vgl. ebd. 350 f.

Aufgabe darin, Individuen in ihrer persönlichen Lebensführung zu beurteilen; die ethische Reflexion über deren Intentionen und Verantwortlichkeiten finde dabei im pädagogischen Denkrahmen von Belohnung und Strafe statt.[6] Eine solche Herangehensweise sei aber dort völlig verfehlt, wo die Verkettung vieler, moralisch möglicherweise völlig belangloser Einzelhandlungen unterschiedlicher (menschlicher und nicht-menschlicher!) Akteure erst in ihrem Ergebnis moralisch bedeutsam wird. In solchen Kontexten undurchschaubarer Kausalzusammenhänge führe das ethische Interesse an den Intentionen des Einzelnen nicht weiter, der ja die Folgen seines Handelns unmöglich voraussehen kann und auch nicht (allein) zu verantworten hat.[7]

2.1.2 Entwurf einer neuen „Computerethik"

Den derart unzureichenden traditionellen „Standardethiken" stellt Floridi nun seinen eigenen Entwurf einer „Computerethik"[8] an die Seite. Darin erarbeitet er ein neues Konzept von *moral agency* und lenkt außerdem den Fokus ethischen Interesses von den Urhebern moralisch bedeutsamer Handlungen (*moral agents*) hin auf die Auswirkungen solcher Handlungen auf die *moral patients*.

2.1.2.1 *Moral agency*

Die Frage, wer oder was als *moral agent* gelten kann, veranlasst Floridi zunächst zu grundlegenderen Ausführungen darüber, wie eigentlich Definitionen entstehen. Eine Definition bildet demnach die Realität nicht eins zu eins ab, sodass sie das Kant'sche „Ding an sich" träfe. Sie ist nicht absolut, sondern immer nur relativ, das heißt kontextabhängig zu verstehen und folglich auch nur in bestimmten Kontexten sinnvoll.[9] Das wird am schillernden Begriff der „Intelligenz" verdeutlicht, für den es nicht die *eine* umfassende, zufriedenstellende Definition gibt, sondern viele verschiedene, die je nach Kontext variieren. Genauso verhalte es sich aber bei allen Definitionen, auch wenn das häufig weniger offensichtlich sei: Stets erfolge das Analysieren und Definieren einer Sache aus einer ganz bestimmten Perspektive heraus,[10] etwa einer subjektiven Einstellung gegenüber dem Untersuchungsobjekt.[11]

---

[6] Vgl. *L. Floridi*, Faultless Responsibility: On the Nature and Allocation of Moral Responsibility for Distributed Moral Actions, in: Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374 (2016) Issue 2083, 4. Der Konsequentialismus in seiner radikalen Folgenorientierung wird für diesen Zusammenhang nicht ausbuchstabiert, sondern verfällt demselben Urteil wie akteurs- und aktzentrierte Ansätze (vgl. ebd. 6; *L. Floridi*, Distributed Morality in an Information Society, in: Science and Engineering Ethics 19 (2013) 727–743, hier 730–732).

[7] Vgl. *Floridi*, Faultless Responsibility, 3 f.

[8] Vgl. *Floridi/Sanders*, Artificial Evil, 64 f.; allgemeiner auch „Informationsethik" (vgl. ebd. 55).

[9] Vgl. *Floridi/Sanders*, On the Morality, 352 f.

[10] Vgl. ebd. 353.

[11] Vgl. *L. Floridi*, Levels of Abstraction and the Turing Test, in: Kybernetes 39 (2010) 423–440, hier 426.

So werde sich bei der Untersuchung eines Autos der taxierende Blick einer Sammlerin unterscheiden vom Blick eines Bastlers oder dem einer Wirtschaftsexpertin. Als *terminus technicus* für diese unterschiedlichen Sichtweisen beziehungsweise Interessenlagen verwendet Floridi den Begriff der Abstraktionsebene (*level of abstraction* / *LoA*). Auf unterschiedlichen Abstraktionsebenen sind unterschiedliche beobachtbare Merkmale (*observables*) relevant: Die Sammlerin wird vielleicht auf Diebstahlsicherheit und auf die Reihe der Vorbesitzer achten, der Bastler auf den Motorzustand, die Wirtschaftsexpertin auf den Marktwert und die anfallenden Unterhaltskosten des Autos.[12] Merkmale, die für die jeweiligen Interessen irrelevant sind, werden ausgeblendet beziehungsweise abstrahiert. Abstraktionsebenen vereinfachen folglich die Komplexität des untersuchten Gegenstandes.[13] Je nachdem, wie karg oder reichhaltig die Menge an *observables* ausfällt, ist die Abstraktionsebene höher, das heißt abstrakter (Definition eines Autos als Fahrzeug), oder niedriger, das heißt konkreter (Definition eines Autos als motorisiertes Fahrzeug).[14]

Um Ambiguitäten, Missverständnissen und Fehlschlüssen vorzubeugen, sei es notwendig, im Vorfeld einer Analyse immer offenzulegen, von welcher Abstraktionsebene aus sie erfolgen soll.[15] Für die Definition von *agency* schlägt Floridi nun eine höhere Abstraktionsebene vor, als sie üblicherweise eingenommen wird. Kandidaten für *agents* sollen nicht mehr auf Intentionalität oder sonstige geistige Fähigkeiten untersucht werden; stattdessen sind sie von einer weiter entfernten Warte aus zu betrachten, von der aus sie nur noch unscharf als „Systeme" erkennbar sind. Als *agents* gelten dann diejenigen Systeme, die über folgende drei Eigenschaften verfügen: Interaktivität (Fähigkeit zur Zustandsänderung in Reaktion auf äußere Reize), Autonomie (Fähigkeit zur Zustandsänderung ohne äußere Reize) und Adaptivität (Fähigkeit zur selbstständigen Änderung der Regeln für Zustandsänderungen).[16] Ob nun zum Beispiel ein lernendes Spamfilter-Programm als *agent* gilt, hängt von der Feinheit der Abstraktionsebene ab: Werden als *observables* nur ein- und ausgehende E-Mails herangezogen, der Algorithmus, nach dem der Filter sich den Vorstellungen des Nutzers anpasst, jedoch abstrahiert, so erscheint das Programm interaktiv, autonom und adaptiv, folglich als *agent*. Sobald aber der zugrundeliegende Algorithmus mit in den Blick gerät, ist klar, dass das Programm in seinem Lernprozess durch Regeln gesteuert und also

[12] Vgl. *Floridi/Sanders*, On the Morality, 354.
[13] Vgl. *Floridi*, Levels, 426. Floridi stellt allerdings klar, dass es sich bei der Methode der Abstraktionsebenen um ein rein epistemologisches Instrument handelt, das keinerlei ontologische Entsprechung in einer vermeintlichen Ebenen-Unterteilung innerhalb des untersuchten Gegenstandes voraussetzt (vgl. *L. Floridi*, The Method of Levels of Abstraction, in: Minds and Machines 18 [2008] 303–329, hier 325 f.).
[14] Vgl. *Floridi/Sanders*, On the Morality, 355.
[15] Vgl. ebd. 355; *Floridi*, Levels, 431, 437 f.
[16] Vgl. *Floridi/Sanders*, On the Morality, 357 f.; *Floridi*, Levels, 432.

doch kein *agent* ist.[17] Da heutige Nutzer allerdings nur noch selten Zugriff auf den Programmcode ihrer Software haben und deren Funktionsweise für sie dadurch völlig undurchschaubar ist, hält Floridi die höhere Abstraktionsebene für angemessen, die solche Software als *agent* qualifiziert.[18] Vom *agent* zum *moral agent* ist es dann nur noch ein kleiner Schritt: Alle *agents*, deren Handlungen moralisch qualifizierbare Folgen nach sich ziehen, sind *moral agents*.[19] Ihnen wird deshalb aber keineswegs schon Verantwortung zugesprochen: Verantwortung kommt erst da ins Spiel, wo tatsächlich Intentionalität vorliegt;[20] auf der für *agency* gewählten Abstraktionsebene spielt die ja indes gerade keine Rolle. *Moral agents* ohne Intentionen sind laut Floridi für ihr Handeln nicht moralisch verantwortlich (*responsible*), sondern haftbar (*accountable*) zu machen.[21] Damit ist die Frage nach der Verantwortung bei Mensch-Maschine-Interaktionen nicht beantwortet. Allein schon die Bezeichnung als *moral agent* und die Zuschreibung von Haftbarkeit bringen nach Floridis Dafürhalten aber endlich Klarheit und Struktur in die Diskussion; vor allem kreise diese nicht mehr zwangsläufig darum, einen Schuldigen (das heißt einen moralisch Verantwortlichen) zu suchen, wenn anerkannt werde, dass auch nicht-verantwortliche Akteure als Urheber von Übeln in Frage kommen.[22]

2.1.2.2 Orientierung an den *moral patients* und Optimierung des Outputs

Grundsätzlich solle die Ethik aber ohnehin einen Perspektivwechsel vornehmen und, bevor sie die Urheber angeht, zunächst einmal auf das Endprodukt schauen. Dieses Endprodukt aus teilweise unentwirrbar miteinander verschlungenen Handlungen seitens natürlicher und künstlicher *moral agents* müsse mit Blick auf die Betroffenen (*moral patients*) moralisch evaluiert werden.[23] Ob ein gesamtes System (etwa eine Gesellschaft) durch die Auswirkungen bestimmter Multi-Akteur-Interaktionen in einem besseren oder schlechteren Zustand sei als vorher, könne auch ohne jede Information über die Intentionen der beteiligten Akteure festgestellt werden.[24] Um nach einer negativen Diagnose korrigierend und verbessernd in den Systemzustand eingreifen zu können, sei jeder einzelne Akteur, unabhängig vom Aus-

[17] Vgl. *Floridi/Sanders*, On the Morality, 362; *Floridi*, Levels, 432.
[18] Vgl. *Floridi/Sanders*, On the Morality, 361.
[19] Vgl. *Floridi/Sanders*, On the Morality, 364.
[20] Vgl. ebd. 365.
[21] Vgl. ebd. 351, 376.
[22] „[M]oral source of evil or good" (vgl. ebd. 375f.).
[23] Vgl. *Floridi*, Distributed Morality, 732.
[24] Vgl. *Floridi*, Faultless Responsibility, 6. Bei der Evaluierung von Software hätte sich ein solcher Zustandsvergleich am Gesamtwert der vorhandenen Daten zu orientieren. Sind etwa wertvolle Daten ohne Backup gelöscht worden (bspw. durch einen Virus), ist der Gesamtdatenwert verringert worden und das Ergebnis somit negativ zu bewerten (vgl. *Floridi/Sanders*, Artificial Evil, 63).

maß seiner individuellen Verantwortung, in Haftung zu nehmen.[25] Durch selbstregulative Prozesse innerhalb des Akteur-Netzes (wie zum Beispiel auf Gesellschaftsebene durch Gesetze, Verhaltensregeln, Nudging, Anreizstrukturen et cetera)[26], auf lange Sicht auch durch die präventive Wirkung der Haftungspflicht,[27] werde in der Folge der Output optimiert. Einzige Voraussetzung dafür sei, dass die kausal beteiligten Akteure lernfähig und zu Verhaltensänderungen in der Lage sind.[28] Individuelles Fehlverhalten könne im Einzelfall trotzdem zum jeweiligen Akteur zurückverfolgt werden.[29] Anhand vorher festgelegter Grenzwerte (*morality thresholds*), welche nicht über- beziehungsweise unterschritten werden dürfen, sei auch das Handeln individueller Akteure moralisch klar bewertbar[30] und könne gegebenenfalls entsprechend geahndet werden: Analog zu gesellschaftlichen Sanktionsmaßnahmen, die von Mahnung und Isolation bis hin zur Todesstrafe reichen, könne man Delikten vonseiten künstlicher *moral agents* mit einer Staffelung von Wartung und Umrüstung über ihre Trennung vom Datennetz bis hin zu ihrer vollständigen Vernichtung/Löschung begegnen.[31] Auch ethische Verhaltenskodizes, die bislang nur für Software-Entwickler gelten, sollten für die dabei entwickelten *moral agents* selbst zur Anwendung kommen.[32] Unterhalb dieses Maßnahmenkatalogs zur Systemoptimierung, der selbst schon explizit moralisches Gepräge hat, zieht Floridi aber noch eine fundamentalere Ermöglichungsebene ein – die „ethische Infrastruktur" beziehungsweise „Infraethik".[33] Wie eine funktionstüchtige Wirtschaft auf Infrastruktur für Transport, Kommunikation et cetera angewiesen sei, so brauche eben auch eine Gesellschaft moralische „Enablers", um moralisch erfolgreich zu sein.[34] Solche Ermöglichungsfaktoren moralisch guten Handelns (wie etwa Vertrauen, Respekt, Zuverlässigkeit, Meinungsfreiheit, fairer Wettbewerb) sind nicht selbst schon moralische Werte, sondern gewissermaßen das Schmiermittel beziehungsweise der Nährboden, auf dem mora-

[25] Haftung ist hier als verschuldensunabhängige (Gefährdungs-)Haftung („strict liability") zu verstehen (vgl. *Floridi*, Faultless Responsibility, 8).
[26] Vgl. ebd. 7.
[27] Vgl. ebd. 8.
[28] Das ist in der Regel gewährleistet, da „agents" über Interaktivität, Autonomie und Adaptivität verfügen (vgl. *Floridi*, Faultless Responsibility, 6 f.).
[29] Vgl. ebd. 9.
[30] Ein Spamfilter-Programm verhielte sich demnach moralisch gut, solange es einen bestimmten Prozentsatz der eingehenden E-Mails korrekt filtert; sobald es diesen Toleranzbereich verlässt und zu viele E-Mails falsch zuordnet, ist sein Verhalten moralisch schlecht (vgl. *Floridi/Sanders*, On the Morality, 369 f.).
[31] Vgl. ebd. 372 f.
[32] Solche Ethikkodizes verpflichten bspw. dazu, Schaden Dritter zu vermeiden und zum Wohl der Gesellschaft beizutragen, zu Fairness, Ehrlichkeit, Achtung der Eigentumsrechte und Privatsphäre anderer etc. (vgl. ebd.).
[33] Vgl. *L. Floridi*, Die 4. Revolution. Wie die Infosphäre unser Leben verändert, Berlin 2015, 248 f.
[34] Vgl. ebd. 250.

lisch Wertvolles gedeihen kann.[35] Im IT-Bereich sei es evident, worin die moralische „Schmiere" bestehe – Transparenz, Datenschutz, Verfügbarkeit und freier Zugang zu Informationen et cetera.[36] Bezeichnenderweise zögen „Enablers" schon seit Jahrzehnten das Interesse von Soziologie, Wirtschaft und Politik auf sich,[37] die Ethik aber hinke hinterher – vielleicht sei es nun endlich Zeit, auch den äußeren Rahmenbedingungen von Moral mehr Aufmerksamkeit zu schenken.

2.2 Deborah G. Johnson

2.2.1 Kritik an einseitigen Ethikansätzen

Die US-amerikanische Philosophin Deborah G. Johnson sucht mit ihrem Ansatz die „goldene Mitte" zwischen zwei Extremen: In der einen Extremposition werde die menschliche Verantwortung unterlaufen, insofern *tA* als *moral agents* bezeichnet werden.[38] Vertreter der anderen Positionen verkennen hingegen die moralische Qualität maschinellen Verhaltens, insofern sie Technik als außermoralisch ansehen und diese folglich nicht zu einem eigenen Gegenstand ethischer Untersuchungen machen. Dieser Ansatz übersieht laut Johnson, dass Menschen immer komplexere Aufgaben an immer intelligentere Maschinen delegieren. Die Welt ist zunehmend den Einflüssen von *KI* unterworfen, welche weitgehend unabhängig von den Menschen operiert, die diese konzipierten und benutzen.[39] Beide Annahmen – *tA* seien *moral agents* oder seien moralisch nicht signifikant – bezeichnet Johnson als Fehler. Einerseits müssen *tA* von Menschen kontrolliert werden, andererseits beeinflussen sie die moralische Welt der Menschen. Angesichts des Schadenspotentials wirkmächtiger und automatisierter *KI* möchte sie die Frage nach der Verantwortung aber auch nicht stillschweigend übergehen. Die schlechteste aller Lösungen sei eine *responsibility gap*, wo niemand für die technisch verursachten Schäden verantwortlich ist.[40] Johnson möchte sowohl die moralische Bedeutung der Wirkmacht von *tA* betonen als auch die damit einhergehende Macht und Verantwortung derjenigen, die solche Technologien konzipieren und einsetzen.

---

[35] Vgl. *Floridi*, Distributed Morality, 740.

[36] Vgl. ebd. 739f.

[37] Vor allem im Bereich der Entwicklungshilfe gelten Bildung, Gesundheit, Sicherheit etc. als wichtige Enablers (vgl. ebd. 739).

[38] Vgl. *Floridi/Sanders*, On the Morality; *Floridi*, The Method; *N. R. Jennings/K. Sycara/M. Wooldridge*, A Roadmap of Agent Research and Development, in: Autonomous Agents and Multi-Agent Systems 1 (1998) 7–38; *J. Ferber*, Multi-Agent Systems: An Introduction to Distributed Artificial Intelligence, Harlow [u. a.] 1999; *G. Weiss*, Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence, Cambridge (Mass.) 1999; *M. Wooldridge*, An Introduction to Multiagent Systems, Chichester 2002.

[39] Vgl. *D. G. Johnson/K. W. Miller*, Un-making Artificial Moral Agents, in: Ethics and Information Technology 10 (2008) 123–133, 125.

[40] *A. Matthias*, The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata, in: Ethics and Information Technology 6 (2004) 175–183.

2.2.2 *Intentionality* von Artefakten

In ihrem Aufsatz *Computer Systems: Moral Entities but not Moral Agents*[41] lehnt Johnson es ausdrücklich ab, *tA* als *moral agents* zu bezeichnen. Offensichtlich kritisiert sie solche Ansätze, wie sie beispielsweise von L. Floridi vertreten werden.[42] Zwar konnte Johnson der metaphorischen Bezeichnung von *tA* als *moral agent* zwischenzeitlich durchaus etwas abgewinnen. Jedoch hatte sie dies anfangs noch klar abgelehnt, da ein Akteur, um moralisch handeln zu können, intentionale Geisteszustände (*intendings to act*) – letztlich Handlungsgründe – besitzen müsse. Dass *KI* dazu fähig sei oder in Zukunft einmal sein werde, sei eine unbewiesene Behauptung. Es reiche nicht, wenn *KI* menschliche Handlungen umfänglich abbilde; Mensch und Computer unterschieden sich ontologisch.[43]

Zugleich verhalte sich *KI* keineswegs moralisch neutral – das heißt ausschließlich instrumentell – oder besitze den gleichen moralischen Status wie natürliche Ereignisse. Denn bei aller Kritik hätten Vertreter, die bestimmte *tA* als *moral agents* bezeichnen, richtigerweise erkannt, dass diese relativ eigenständig unsere sozialen Arrangements, Beziehungen, Institutionen und Werte enorm beeinflussen und gestalten. Artefakte verhielten sich zwar wie natürliche Ereignisse mit Notwendigkeit.[44] Ihnen komme aber ein anderer moralischer Status zu, da sie die *intentionality* jener Menschen verkörperten, die sie konzipieren und einsetzen.[45] Artefakte besitzen diese *intentionality* laut Johnson auch weitestgehend als ihre eigene, wenn menschliche User und Designer ihr Verhalten nicht unmittelbar kontrollieren.

Dennoch darf eine solche *intentionality*[46] von *tA* nicht mit den *intendings to act*[47] oder *intentional agency*[48] von Menschen verwechselt werden. Sowohl Designer wie auch User und *tA* besitzen ihre je spezifische *intentionality* und *efficacy* (Wirksamkeit). Vor diesem Hintergrund entwirft Johnson ihre

---

[41] *D. G. Johnson*, Computer Systems: Moral Entities but not Moral Agents, in: Ethics and Information Technology 8 (2006) 195–204.

[42] Vgl. *Floridi/Sanders*, On the Morality; *Floridi*, The Method; *Jennings/Sycara/Wooldridge*, A Roadmap; *Ferber*, Multi-Agent Systems; *Weiss*, Multiagent Systems; *Wooldridge*, An Introduction.

[43] Vgl. *Johnson*, Computer Systems: Moral Entities, 198–200.

[44] Ungeachtet zahlreicher auch aus Sicht von Johnson berechtigter Anfragen an die Möglichkeit einer klaren Unterscheidung zwischen „natürlich" und „nicht-natürlich" erlaubt diese Unterteilung die so elementare Unterscheidung zwischen den Auswirkungen menschlichen Verhaltens auf die Welt und dem, was diesem vorgegeben bzw. von diesem unabhängig ist (vgl. ebd. 197; vgl. dazu auch *Aristoteles*, Nikomachische Ethik, 6.32; *M. Heidegger*, Die Frage nach der Technik, in Vorträgen und Aufsätzen 1954).

[45] Vgl. *Johnson*, Computer Systems: Moral Entities, 202; *D. G. Johnson/M. Noorman*, Artefactual Agency and Artefactual Moral Agency, in: *P. Kroes/P.-P. Verbeek* (Hgg.), The Moral Status of Technical Artefacts, Dordrecht 2014, 143–158, hier 144 f.

[46] *Johnson*, Computer Systems: Moral Entities, 201.

[47] Ebd. 201.

[48] Vgl. *D. G. Johnson/M. Verdicchio*, AI, Agency and Responsibility: The VW Fraud Case and Beyond, in: AI & SOCIETY 34 (2019) 639–647 (https://doi.org/10.1007/s00146-017-0781-9; letzter Zugriff: 28.08.2019).

eigene Handlungstheorie und bezeichnet diese als *triad of intentionality and efficacy*[49]. Artefakte können menschliche Handlungsmöglichkeiten erweitern oder einschränken sowie Gutes und Schlechtes bewirken. Einerseits sind *tA* wegen ihrer *intentionality* und *efficacy components of moral agency*[50]. Andererseits ist ihr Verhalten abhängig von der *efficacy* und *intentionality* ihrer Designer und User. Handlungsstränge, in denen Designer, User und *tA* je eine moralisch bedeutsame Komponente bilden, bezeichnet Johnson als *technological moral action*[51]. Alle drei Komponenten sind in einer ethischen Untersuchung zu beachten.

2.2.3 Konzepte von *agency* und ambivalente Metaphern

Den Beitrag von *tA* als eine der Komponenten einer *technological moral action* spezifiziert und differenziert Johnson mithilfe des *agency*-Begriffs. Dabei ist im Argumentationsgang eine substantielle Verschiebung zu beobachten: Wurde zuvor von Artefakten als *components of moral agency* gesprochen, geht es nun um deren *kind of moral agency*[52].

Agency definiert Johnson als eine *ability or capacity of an entity to act in the world*[53] und unterscheidet näherhin drei Konzepte von *agency*: *causal efficacy*, *acting for* und *moral autonomy*.[54] Es kommt erstens vor, dass Artefakte aufgrund ihrer kausalen Wirksamkeit (*causal efficacy*) als *agent* oder sogar als *moral agent* bezeichnet werden.[55] Der kausalen Wirksamkeit entspricht der Terminus der „kausalen Verantwortung“[56]. Jedoch sind die Begriffe „Verantwortung“ und *(moral) agency*, wenn sie auf Artefakte bezogen werden, Metaphern.

Im zweiten Konzept wird *agency* verstanden als *acting for or on behalf of another entity*[57]. Artefakte sind hier *moral agents*, insofern ihnen Aufgaben übertragen werden, deren Ausübung oder Folgen moralisch relevant sind.[58] Diese zweite Auslegungsmöglichkeit von *moral agency* erörtert Johnson ausgiebig anderenorts mithilfe des Konzepts *surrogate agency*.[59] Ein *surrogate agent* handelt in der Regel aufgrund einer bestimmten übernommenen Rolle in Stellvertretung für jemand anderen. Zum Beispiel beauftragt ein

---

[49] *Johnson*, Computer Systems: Moral Entities, 202.
[50] Ebd. 203.
[51] *D. G. Johnson/T. M. Powers*, Computer Systems and Responsibility: A Normative Look at Technological Complexity, in: Ethics and Information Technology 7 (2005) 99–107.
[52] *D. G. Johnson/T. M. Powers*, Computers as Surrogate Agents, in: *J. van den Hoven/J. Weckert* (Hgg.), Information Technology and Moral Philosophy. Philosophical Explorations in Computer Ethics, Cambridge 2008, 251–269.
[53] *Johnson/Noorman*, Artefactual Agency, 148.
[54] Ebd. 148–157.
[55] Vgl. ebd. 153.
[56] *Johnson/Powers*, Computer Systems, 105 f.
[57] *Johnson/ Noorman*, Artefactual Agency, 148.
[58] Vgl. ebd. 153.
[59] „Surrogate agency, whether human or computer, is a special form of moral agency in which the agent has a third-person perspective and pursues what we will call second-order interests – those interests of clients or users“ (*Johnson/Powers*, Computers as Surrogate Agents, 258).

Klient einen Anwalt, seine Interessen zu vertreten. Laut Johnson ist es für die moralische Bewertung des Rollenverhaltens zunächst irrelevant, ob ein Mensch oder eine *KI* als *surrogate agent* handelt.[60] Berechnet ein Steuerberater oder eine Steuersoftware die Einkommenssteuer falsch, verletzen sie die Interessen ihrer Klienten beziehungsweise User. Demnach besitzen Artefakte eine Rollenverantwortung, wenn sie mit moralisch relevanten Aufgaben (beispielsweise Erkennen von technischen Fehlern, Sekretariatsaufgaben oder Flugkontrolle) beauftragt werden.[61] Allerdings begründet eine *human-to-artefact*-Delegation keine moralische Verantwortung für Artefakte. Ausschließlich in einer *human-to-human*-Delegation wird gemeinsam mit Aufgaben auch eine moralische Verantwortung delegiert.[62] So gesehen ist es also doch moralisch relevant, ob die Einkommenssteuer von einer Software oder einem Steuerberater berechnet wird. Wenn die Software ihre Aufgabe nicht erfüllt, hat sie vermutlich eine Fehlfunktion, handelt aber nicht – im moralischen Sinn – unverantwortlich. Auch in der zweiten Konzeption werden *moral agency* und „Verantwortung" für Artefakte als Metapher gebraucht. Metaphern können uns einerseits helfen, ein womöglich komplexes Verhalten von *tA* leichter zu verstehen und ihren spezifischen Beitrag in einer ethischen Untersuchung zu würdigen. Andererseits gibt eine solch missverständliche Sprechweise auch Anlass zu schlicht irrigen Annahmen. Es besteht die Gefahr, dass die eigentlichen moralischen Akteure und Träger von Verantwortung verdeckt werden.[63]

Diese sind Gegenstand des dritten *agency*-Konzepts der *moral autonomy*. Unter „Autonomie" versteht Johnson die menschliche Freiheit als Möglichkeitsbedingung von Sittlichkeit. Im Unterschied zu Artefakten vermag der Mensch frei zu handeln und Sollen setzt Können voraus. Dagegen werden Artefakte in der Regel als „autonom" bezeichnet, wenn sie relativ unabhängig operieren und von Menschen nicht unmittelbar kontrolliert werden. Johnson warnt davor, diese zwei Verständnisse von Autonomie zu vermischen und *tA* als moralisch autonom Handelnde zu bezeichnen.[64] Artefakte müssten immer so verstanden werden, dass sie an die Designer und User gebunden blieben, damit die Frage der menschlichen Verantwortung für das artifizielle Verhalten nicht ins Nichts laufe.[65] Die Charakterisierung als *moral agent* dürfe nicht den ontologischen Unterschied zwischen Mensch und Computer

[60] „The primary issue is whether the agent is incompetent or misbehaves with respect to the *clients'* interests. In other words, does the surrogate agent stay within the constraints of the special role morality?" (ebd. 260).

[61] Vgl. *Johnson/Powers*, Computer Systems, 105.

[62] Vgl. *Johnson/Noorman*, Artefactual Agency, 153.

[63] Vgl. *D. G. Johnson*, Software Agents, Anticipatory Ethics, and Accountability, in: *G. E. Marchant/B. R. Allenby/J. R. Herkert* (Hgg.), The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight. The Pacing Problem, Dordrecht 2011, 61–76, hier 63.

[64] Vgl. *Johnson/Noorman*, Artefactual Agency, 151–152.

[65] Vgl. *Johnson*, Software Agents, 70.

verdecken. Aus diesem Grund lehnt Johnson den Gebrauch von *moral agent* im Sinne der *autonomy-agency*-Konzeption ab.[66]

Im Unterschied dazu helfen uns die *moral agent*-Metaphern der zwei anderen *agency*-Konzepte laut Johnson, zu verstehen, dass *tA* menschliche Interessen tangieren und Handlungen beschränken oder erleichtern *und* unabhängig von Designern und Nutzern moralisch bewertet werden können. Allerdings lasse sich nur von der tatsächlichen *moral agency* eines Menschen die Brücke zur moralischen Verantwortung schlagen.[67] Allein in diesem Zusammenhang beurteilt Johnson die Vorstellung einer *moral agency* von Artefakten als seltsam und gefährlich.[68]

2.2.4 Verantwortung innerhalb von *triadic agency*

Vor kurzem entwickelte Johnson gemeinsam mit Mario Verdicchio ein weiteres heuristisches Modell, um die jeweiligen Beiträge und Verantwortlichkeiten von Designer, User und *tA* differenziert zuschreiben zu können.[69] Nach ihrem „*triadic agency*"-Modell wird ein Zustand weder vom Menschen noch vom Computersystem allein verursacht, sondern durch ein differenziertes Zusammenspiel von Mensch und Artefakt. Grundsätzlich bedeutet *agency* eine *capability to act*. Ferner unterscheidet Johnson drei *agency*-Formen: (1) Eine Entität besitzt aufgrund ihres Vermögens, kausal wirksam zu sein, eine *causal agency* (*agency* im weiten Sinn).[70] (2) Ein Mensch besitzt aufgrund seines Vermögens, intentional handeln zu können, *intentional agency* (*agency* im strikten und klassischen Sinn).[71] In dieser zweiten *agency*-Form steht die Intention am Anfang einer Kausalkette. Artefakte können nur im metaphorischen Sinn „intentional handeln". (3) Um die Mensch-Maschine-Interaktion und das Zustandekommen von Ergebnissen erklären zu können, müssen sich diese unterschiedlichen *agency*-Formen zueinander verhalten. So gesehen ist die *triadic agency* mehr als die Summe ihrer Einzelteile: Wenn Menschen gemeinsam mit Artefakten Ziele errei-

---

[66] „Using the autonomy conception of agency in relation to artefactual moral agents is problematic in the sense that we are asked to imagine or hypothesize that machines will at some point in their development operate in a way that would justify considering them morally autonomous, ascribing responsibility to them and granting them the status of moral agents. Without knowing how such machines would work, this idea seems to go from using a metaphor to understand certain phenomena, to using it as a basis to attribute status" (*Johnson/Noorman*, Artefactual Agency, 156 f.; vgl. *Johnson*, Software Agents, 62 f.).

[67] Vgl. *Johnson/Verdicchio*, AI.

[68] „The claims are odd because they do not seem to acknowledge that computer systems are an extension of human activity and human agency, and because their view of agency and morality is out of sync with the idea of morality as a human system of contextualized ideas and meanings. The claims seem dangerous because they imply that computer systems can operate without any human responsibility for the system behavior" (*Johnson/Miller*, Un-making Artificial Moral Agents, 127).

[69] Vgl. *Johnson/Verdicchio*, AI, 4.

[70] Vgl. *M. Schlosser*, Art. Agency (2015), in: The Stanford Encyclopedia of Philosophy (https://plato.stanford.edu/entries/agency/; letzter Zugriff 09.04.2019).

[71] Vgl. *Schlosser*, Agency.

chen, dann: (a) möchte der Nutzer (oder die Nutzer) ein Ziel erreichen und delegiert die Aufgabe, dieses Ziel zu erreichen, an den Designer; (b) erstellt der Designer (oder die Designer) ein Artefakt, um das Ziel zu erreichen; (c) bietet das Artefakt eine kausale Wirksamkeit, die notwendig ist, um das Ziel zu erreichen.[72] Während Menschen Intentionalität besitzen und kausal wirksam sind, steuern Artefakte zusätzliche kausale Wirksamkeit bei. In diesem Modell wird die jeweilige Eigenart und Qualität der einzelnen Handlungskomponenten herausgearbeitet, ohne deren wechselseitige Bezogenheit zu abstrahieren. Designer, User und *tA* leisten ihre spezifischen Beiträge, aber nur gemeinsam vermögen sie etwas zu produzieren. Im *triadic agency*-Modell wird nur demjenigen Verantwortung zugeschrieben, der fähig ist, intentional zu handeln. *Agency* umfasst hier sowohl die kausale als auch intentionale Dimension und ist weiter als der Verantwortungsbegriff. Da *KI* nicht fähig ist, intentional zu handeln, trägt sie für ihre kausale Wirksamkeit keine Verantwortung. Umgekehrt bleiben Menschen in ihrer Verantwortung, auch wenn sie zunehmend komplexere Aufgaben an *KI* delegieren. Auf der Suche nach dem oder der Verantwortlichen muss solange in Richtung Designer/User gefragt werden, bis ein Mensch (oder eine Menschengruppe) gefunden wird. Allerdings kann eine Antwort auf die Frage, welche Person wieviel Verantwortung trägt, nicht ohne Berücksichtigung der technologischen Komponente gefunden werden.

2.3 Peter-Paul Verbeek

2.3.1 Ausgangssituation und Anliegen

Künstliche Gegenstände sind an menschlichem Handeln beteiligt, seit die ersten paläolithischen Faustkeile in Gebrauch kamen. Dass sich eine (ethische) Diskussion darüber entspinnt, welche Rollen den an einer Handlung mitwirkenden Menschen und Gegenständen jeweils zuzuschreiben sind, ist jedoch erst eine Folge der permanent zunehmenden Durchdringung unseres Alltags durch Technik.[73] Inzwischen scheint eine rein funktionale Sicht auf die Dinge, die menschliches Handeln prägen, nicht (mehr) ausreichend zu sein: Mit Worten wie „Funktion“ und „Gebrauch“ sind die Rolle von Technik und unser Verhältnis zur technisierten Welt nicht angemessen beschrieben.[74] Der niederländische Philosoph Peter-Paul Verbeek kommt in seinen Ausführungen dazu immer wieder auf ein prägendes Erlebnis

[72] Vgl. *Johnson/Verdicchio*, AI, 4.

[73] Vgl. *F. A. Hanson*, Which Came First, the Doer or the Deed? in: *Kroes/Verbeek* (Hgg.), The Moral Status of Technical Artefacts, 55–73, hier 60.

[74] Vgl. *P.-P. Verbeek*, Materializing Morality. Design Ethics and Technological Mediation, in: Science, Technology, & Human Values 31 (2006) 361–380, hier 362, 378; *P.-P. Verbeek,* Moralizing Technology. Understanding and Designing the Morality of Things, Chicago 2011, 46; *P.-P. Verbeek*, Beyond Interaction: A Short Introduction to Mediation Theory, in: Interactions 22 (2015) 26–31, hier 28.

zurück, das er während der Schwangerschaft seiner Frau hatte und das die Unzulänglichkeit einer solchen Technik-Auffassung deutlich vor Augen führt – die Ultraschalluntersuchung. Werdende Eltern sind plötzlich vor eine moralische Entscheidung gestellt, die ohne diese Technik niemals hätte getroffen werden müssen (selbst wer auf den Ultraschall verzichtet, fällt eine moralisch gewichtige Entscheidung). Von einem reinen „Gebrauchen" der Ultraschalldiagnostik zu einem von Menschen vorher festgelegten Zweck kann hier also keine Rede sein. Die (psychologische, soziologische et cetera.) Wirkung von Pränataldiagnostik (*PND*) ist auf die Intentionen von Entwicklern[75] und Nutzern nicht reduzierbar. *PND* ist nicht nur Mittel zum Zweck, dient nicht nur der Ausführung einer Handlung, sondern führt überhaupt erst eine Handlungsentscheidung herbei. So entstehen neue Formen moralischer Verantwortung, die nicht einfach auf Menschen zurückgeführt werden können.[76]

2.3.2 *mediation theory*

Ausgehend von Don Ihdes postphänomenologischem Ansatz und Bruno Latours *actor-network theory*[77] entwickelt Verbeek eine Vermittlungs-Theorie (*mediation theory*), die einer solchen gemeinsamen Handlungsurheberschaft von Mensch und Technik gerecht zu werden versucht. Demnach hat Technik die Rolle eines aktiven Vermittlers zwischen Mensch und Welt.[78] Diese Vermittlung findet auf zwei Ebenen statt: Zum einen beeinflusst Technologie die menschliche Wahrnehmung der Welt (hermeneutische Vermittlung); zum anderen partizipiert sie aktiv an Handlungen (pragmatische Vermittlung).[79] Zurück zum Beispiel der Ultraschalldiagnostik: Ob werdende Eltern ihr ungeborenes Kind als Patienten, als defizitär et cetera wahrnehmen, ist ganz maßgeblich technologisch vermittelt (hermeneutische Vermittlung). Die Entscheidung für oder gegen eine Abtreibung ist pragmatisch vermittelt; sie ist durch die Technik weder determiniert, noch kann sie völlig unabhängig von ihr gefällt werden. Moralische Entscheidungen und Handlungen sind folglich Gemeinschaftsprodukte von

[75] Die Ultraschalltechnik wurde ursprünglich nicht zu medizinischen Zwecken erfunden, geschweige denn mit der Zielsetzung, die Abtreibungspraxis zu beeinflussen (vgl. *P.-P. Verbeek*, Some Misunderstandings About the Moral Significance of Technology, in: *Kroes/Verbeek* (Hgg.), The Moral Status of Technical Artefacts, 75–88, hier 82).

[76] Vgl. *Verbeek*, Moralizing Technology, viii f., 38, 50 f.; *ders.*, Some Misunderstandings, 82.

[77] Vgl. *Verbeek*, Materializing Morality, 362 f.; *ders.*, Moralizing Technology, 33, 45–47, 52; *ders.*, Some Misunderstandings, 79, 86; *ders.*, Beyond Interaction, 29; *P.-P. Verbeek*, Toward a Theory of Technological Mediation. A Program for Postphenomenological Research, in: *J. K. B. O. Friis/R. P. Crease* (Hgg.), Technoscience and Postphenomenology: The Manhattan Papers, London 2015, 189–204, hier 190, 193 f. und 201; *P.-P. Verbeek*, Designing the Morality of Things: The Ethics of Behaviour-Guiding Technology, in: *J. van den Hoven/S. Miller/T. Pogge* (Hgg.), Designing in Ethics, Cambridge 2017, 78–94, hier 82.

[78] Vgl. *Verbeek*, Materializing Morality, 364; *ders.*, Some Misunderstandings, 77 f.; *ders.*, Beyond Interaction, 28 f.

[79] Vgl. *Verbeek*, Materializing Morality, 364, 368.

Mensch und Technik;[80] Moral ist „hybrid" und *moral agency* gibt es nur als Gemisch (*composite moral agency*).[81] *Moral agency* ist also keine ontologische Eigenschaft einzelner Entitäten, die ihnen von Natur aus immer schon anhaften würde. Weder einzelne technische Geräte noch einzelne Menschen haben *moral agency*. Diese ist vielmehr erst das Ergebnis komplexer Technik-Mensch-Interaktionen; sie bildet also nicht die Grundlage für eine Handlung, sondern entsteht aus der Handlung.[82]

Kritik äußert Verbeek am Ansatz von Floridi, *moral agency* einzelnen technologischen Artefakten zuzuschreiben. Um die moralische Bedeutsamkeit von Technik zu verstehen und begrifflich zu erfassen, sei das ein wichtiger Beitrag. Allerdings greife er zu kurz, da nur ganz bestimmte Technologien die Kriterien für *agency* (Interaktivität, Autonomie und Adaptivität) erfüllen. Die Ultraschalldiagnostik wäre nach diesem Konzept kein *moral agent* und bliebe damit begrifflich unterbestimmt. Für ein umfassendes Verständnis der teilweise subtilen technologischen Einflüsse auf den Menschen sei *moral agency* also immer nur als Gemeinschaftsprodukt zu denken, in dem eben auch Technikformen ihren Platz finden, die den Anforderungen an selbstständige *agents* nicht entsprechen.[83]

Verbeek geht so weit, sogar die Handelnden selbst als Ergebnis von Interaktion zu bezeichnen („[H]umans and technologies should not be seen as two ‚poles' between which there is an interaction; rather, they are the result of this interaction"), da sie sich in ihrem Zusammenspiel gegenseitig formten.[84] Folglich sei auch technologische Vermittlung nichts, was sich zwischen bereits bestehenden Entitäten (Mensch und Welt) abspiele und damit eine Mittelposition zwischen ihnen einnähme. Subjektivität des Menschen und Objektivität der Welt entstehen erst als Produkt aus der technologischen Vermittlung.[85] Mag das Ultraschallbeispiel hier zur Verdeutlichung noch nicht ausreichen, scheint das Ganze mit Blick auf *smart environments* und *cyborgs* schon plausibler: In einem *smart home* steht Technik nicht mehr als Mittelinstanz zwischen Mensch und Umwelt; sie verschmilzt mit ihr. Bei

[80] Vgl. *Verbeek*, Some Misunderstandings, 78.
[81] Ebd. 77 f.
[82] Vgl. ebd. 75, 80; *Verbeek*, Designing the Morality, 84.
[83] Vgl. *Verbeek*, Moralizing Technology, 50 f. Zur Kritik an Floridis engen Kriterien für „moral agency" kommen natürlich Verbeeks grundsätzliche Bedenken gegenüber statischen Zuschreibungen an einzelne Entitäten hinzu.
[84] Vgl. *Verbeek*, Beyond Interaction, 28. In seinem Artikel „Which Came First? The Doer or the Deed?" kennzeichnet F. Allan Hanson das als Charakteristikum aller „Composite Agency"-Theorien (unter diesem Sammelbegriff fasst er Ansätze unterschiedlicher Autoren, darunter auch Verbeek [vgl. *Hanson*, Which Came First, 60]): Während aus Sicht des Individualismus der Handelnde vor, während, nach und separat von jeglicher Handlung existiert, wird nach den „Composite Agency"-Theorien der Handelnde erst durch die Handlung konstituiert, existiert also nicht unabhängig von ihr (vgl. ebd. 70): „In a world where nothing happens there are no agents" (ebd. 61).
[85] Vgl. *P.-P. Verbeek*, Expanding Mediation Theory, in: Foundations of Science 17 (2012) 391–395, hier 391 f.

einem Menschen, dem ein Cochlea-Implantat eingepflanzt ist, kann nicht mehr unterschieden werden zwischen dem Hören des Menschen und der Vermittlung der Technik; die Hörleistung ist ein Gemeinschaftsprodukt beider.[86] Noch eindrücklicher wird eine solche Verschmelzung von Mensch und Technik bei Hirnimplantaten, die das Verhalten steuern und damit direkten Einfluss auf die Persönlichkeit von Menschen nehmen.[87] Bei diesem Extrembeispiel liegt nicht nur hybride *moral agency* vor, sondern auch hybride Intentionalität. Auf niedrigeren Stufen von Misch-Intentionalität werden Intentionen „induziert" (Ultraschalldiagnostik),[88] von „persuasiven" (kalorienberechnendes *FoodPhone*) oder *nudging*-Technologien (Essensanordnung in der Kantine) beeinflusst.[89] Zum Teil ist eine solche intentionale Beeinflussung beabsichtigt und im Design-Prozess aktiv miteingebaut (Boden-Schwellen zur Verkehrsberuhigung), zum Teil ist sie nicht vorgesehen und entwickelt gewissermaßen ein Eigenleben (Ultraschalldiagnostik; Rollstuhlfahrerbehinderung durch Drehtüren).[90] So entwickeln sich nicht nur Mensch und Welt in ihrer Beziehung zueinander, auch die vermittelnde Technologie hat keine festgelegte Identität; sie ist *multi-stable*.[91] Dass Verbeek das Thema Intentionalität untersucht, hängt damit zusammen, dass Intentionalität und Freiheit als klassische Kriterien für *moral agency* gelten. Genau wie *moral agency* sind für ihn aber weder Intentionalität noch Freiheit statische Eigenschaften einzelner Entitäten. Intentionalität entsteht durch Vermittlung, und auch Freiheit ist immer vermittelt: Freiheit darf hier nicht verstanden werden als negative Freiheit, das heißt als Unabhängigkeit von technologischen Einflüssen, sondern als positive Freiheit, also als ein freies Sich-Verhalten zur und Gestalten der sowieso unvermeidlichen und beinahe allgegenwärtigen technologischen Beeinflussung: Wer technologische Welt-Vermittlung zu leugnen oder abzustreifen versucht, jagt damit einem Phantom nach und wird unfrei. Durch ein vernünftig reflektiertes Verhältnis zu jener immer schon gesetzten Vermittlung hingegen bleibt die menschliche Freiheit gewahrt.[92]

---

[86] Vgl. ebd. 393.
[87] Vgl. *Verbeek*, Some Misunderstandings, 83 f.
[88] Vgl. ebd. 82 f.
[89] Vgl. *Verbeek*, Designing the Morality, 79–81.
[90] Verbeek warnt hier vor einem Designer-Fehlschluss („designer fallacy"): Wie die Interpretation eines Textes nicht auf die Intentionen des Autors reduziert werden könne („author fallacy"), so seien eben auch die Wirkungen von Technologien nicht ausschließlich auf die Intentionen ihrer Designer rückführbar (vgl. ebd. 79).
[91] Vgl. ebd. 85; *P.-P. Verbeek*, What Things Do. Philosophical Reflections on Technology, Agency, and Design, University Park (Pa.) ²2005, 170.
[92] Vgl. *Verbeek*, Materializing Morality, 370; *Verbeek*, Moralizing Technology, 60 f., 87; *Verbeek*, Some Misunderstandings, 84; *Verbeek*, Beyond Interaction, 31; *Verbeek*, Designing the Morality, 82, 86 f.

2.3.3 Verantwortung der Designer

Vermittlung ist also zu reflektieren, zu antizipieren und zu gestalten. Verbeek sieht diese Aufgabe vorrangig bei den Technik-Designern. Damit lastet auf ihnen eine große Verantwortung, weil Technik das In-der-Welt-Sein des Menschen und mithin den Menschen selbst formt. Die Dimensionen einer solchen Verantwortung verdeutlicht Verbeek mit Sätzen wie „Designers materialize morality“[93] und „Designing technology is designing human beings“[94]. Um dieser Aufgabe gerecht zu werden, müssen Designer zunächst einmal ihr Verständnis von verschiedenen (zum Teil nicht intendierten) Wirkweisen von Technik schärfen. Durch systematische Technikfolgenabschätzung und im Gespräch mit verschiedenen Interessengruppen (potentielle Verbraucher et cetera) sollen dann zukünftige Wirkweisen möglichst genau bestimmt werden.[95] Anschließend ist in einem demokratisch legitimierten Prozess[96] eine Entscheidung darüber zu treffen, ob und wenn ja in welcher Form (moralische) Vermittlung explizit in Technologie eingebaut werden soll.[97] Die Faktoren „Sichtbarkeit“ (*visibility*) und „Kraft/Zwang“ (*force*) bestimmen die jeweilige Vermittlungsform.[98]

Verbeek sieht Menschen durch das Theorem einer hybriden *moral agency* also keineswegs aus der Verantwortung entlassen, im Gegenteil: „Seeing the moral significance of technologies makes us more responsible, rather than less.“[99] Indem die Vermittlungstheorie technologische Wirkweisen und Einflussnahmen samt ihrer moralischen Komponente durchschaubarer mache, steigere sie auch die Verantwortung für deren Gestaltung und Einsatz.[100]

3. Kritische Auswertung der drei Theorien

Die drei hier dargestellten Theorien gebrauchen den *moral agency*-Begriff unterschiedlich. Inwieweit trägt nun das jeweilige Verständnis dazu bei, die komplexe Gemengelage einer Mensch-Maschine-Interaktion zu klären und so Moral wie Verantwortung Akteuren eindeutig zuzuschreiben?

---

[93] *Verbeek*, Beyond Interaction, 31 (vgl. *ders.*, Materializing Morality, 361, 369, 379; *ders.*, Designing the Morality, 88).

[94] *Verbeek*, Beyond Interaction, 28.

[95] Vgl. *Verbeek*, Materializing Morality, 369, 371–373, 375; *ders.*, Some Misunderstandings, 86 f.; *ders.*, Toward a Theory, 196–198; *ders.*, Designing the Morality, 89 f.

[96] Vgl. *Verbeek*, Materializing Morality, 369 f.; *ders.*, Designing the Morality, 79, 82.

[97] Vgl. *Verbeek*, Materializing Morality, 363, 369; *ders.*, Beyond Interaction, 31; *ders.*, Toward a Theory, 199.

[98] Sie bilden sozusagen die beiden Achsen eines Koordinatensystems, auf dem sich die Einflussnahme von Technik abbilden lässt. Die vier Grundformen dieser Einflussnahme ergeben sich aus der Kombinationen der jeweiligen Extreme: stark und offensichtlich (Drehkreuze, die zum Ticketkauf zwingen) – schwach und offensichtlich (Diät-Apps) – schwach und unsichtbar (Kaffeemaschine im Foyer zur Stimulation von sozialer Interaktion) – stark und unsichtbar (Haus ohne Aufzug, in dem Menschen zum Treppensteigen gezwungen sind) (vgl. *Verbeek*, Beyond Interaction, 30).

[99] *Verbeek*, Some Misunderstandings, 85.

[100] Vgl. *Verbeek*, Designing the Morality, 84.

3.1 Luciano Floridi

Floridi nennt als einen Beweggrund seiner Neudeutung von *moral agency* das Missverhältnis zwischen *moral agents* und *moral patients*, welches den Verantwortungsdruck auf das Individuum erhöhe.[101] Auffällig ist hierbei, dass er diese eigentlich so gewichtige Schlussfolgerung eher lapidar einwirft und später auch nicht mehr daran anknüpft – inwiefern nun die neu definierte Gruppe von *moral agents*, denen ja ausdrücklich keine moralische Verantwortung zukommen soll, das Individuum in seiner Verantwortung entlastet, bleibt jedenfalls weitgehend im Dunkeln.[102] Zwar resümiert er, dass mit dem Zuwachs an *moral agents* ihr Verhältnis zu den *moral patients* besser ausbalanciert sei; offen bleibt jedoch, worin genau (von jener dubiosen Entlastungsfunktion einmal abgesehen) der Wert einer solchen quantitativen Balance besteht. Überdies scheint er dazu bezüglich der Gruppe der *moral patients* die übliche und grundsätzliche Abgrenzung wie Abstufungen zu übersehen, wie zum Beispiel die zwischen Lebewesen und Nicht-Lebewesen oder die zwischen Menschen, Tieren und Pflanzen.

Deutlicher äußert sich Floridi in Bezug auf die Strukturierungseffekte der neuen Terminologie: „The great advantage is a better grasp of the moral discourse in non-human contexts."[103] Offensichtlich verspricht er sich von der neuen sprachlichen Kategorie „*moral agent* ohne Verantwortung" einen wohltuend ordnenden Einfluss auf die *KI*-Debatte. Hier ist aber zu fragen, inwiefern die formale Identifikation von künstlichen *moral agents* die Debatte tatsächlich inhaltlich voranbringt.

Das führt zu der zentralen Frage, worum sich diese Debatte eigentlich dreht – Floridi zufolge um die Zuschreibung von Verantwortung an Individuen, also um die Suche nach einem „Schuldigen". Bleibt man bei dieser Annahme, ist nicht einzusehen, wie die Existenz nicht-verantwortlicher *moral agents* bei der Suche behilflich sein kann. Folgt man aber Floridi in seiner Kritik an dieser (vermeintlichen) bisherigen ethischen Grundausrichtung und sieht die Aufgaben der Ethik am anderen Ende der Handlungskette, so bleibt immer noch rätselhaft, weshalb ausgerechnet die Ausweitung der Klasse von *moral agents* diesem Kurswechsel förderlich sein soll. Schließlich tadelt Floridi doch gerade das übertriebene Interesse an ihnen; Ethik habe sich stattdessen um die *moral patients* zu kümmern. Er selbst versucht, den Zusammenhang mit der Argumentation zu erhellen, sobald Ethik akzeptiere, dass nicht nur Menschen moralische Urheber von

[101] Vgl. *Floridi/Sanders*, On the Morality, 350.

[102] Es ist nicht völlig auszuschließen, dass Floridi hier eine gedankliche Brücke zur Systemoptimierung durch Haftung für alle (vgl. *Floridi*, Faultless Responsibility) schlägt; ein solcher Zusammenhang wird aber nicht explizit hergestellt und liegt auch nicht nahe, da das „agent-patient"-Ungleichgewicht (vgl. *Floridi/Sanders*, On the Morality) und die Systemperspektive (vgl. *Floridi*, Faultless Responsibility) mit beträchtlichem zeitlichen Abstand in unterschiedlichen Artikeln behandelt werden.

[103] *Floridi/Sanders*, On the Morality, 376.

schlechten und guten Ereignissen („the moral source of evil or good") sein können, werde sie weniger auf Schuldzuschreibungen fixiert sein.[104] Was meint Floridi nun aber mit dem Adjektiv *moral*/„moralisch"? Ganz offensichtlich sind damit keine geistigen Fähigkeiten wie moralische Urteilskraft gemeint. Da die Abstraktionsebene für *agents* ausdrücklich keine Intentionalität vorsieht und der Zusatz *moral* sich nur auf die Folgen von Handlungen bezieht, ist eine solche Deutung ausgeschlossen. Alles spricht dafür, dass Floridi das Wort *moral* in seiner formalen Bedeutung als „die Moral betreffend" benutzt. Der Grundtenor seiner Texte lässt nun aber keinen Zweifel daran, dass „Moral" sich für ihn nicht auf eine Eigenschaft des Urhebers von Handlungen, sondern auf deren Konsequenzen für die *moral patients* bezieht. „Moralisch" (relevant) ist also alles, was sich gut beziehungsweise schlecht auf Dritte auswirkt. Vor diesem Hintergrund muss die Einschätzung verwundern, Ethik könne durch die Anerkennung weiterer, nicht-menschlicher *moral source[s] of evil or good* gewissermaßen von innen erneuert werden. Denn es ist keineswegs eine neue Erkenntnis, dass als leidvoll erfahrene und somit moralisch qualifizierbare Ereignisse nicht menschlich bedingt sein müssen. Überschwemmungen verursachen menschliches Leid und sind folglich „moralische" Phänomene. Mehr steckt aber auch nicht in Floridis Aussage. Denn seiner konsequentialistischen Deutung von „moralisch"[105] zufolge kommt jede Entität als *moral source* in Frage: Was als „moralisch" zu qualifizieren ist, bemisst sich am Ergebnis und das ist im Fall der Überschwemmung unter Umständen genauso moralisch relevant (weil nachteilig für die Betroffenen) wie im Fall einer von Menschen geplanten und ausgeführten Missetat. Somit ist der Hinweis auf außermenschliche „moralische" Quellen für Gut und Übel jeglicher Art für sich genommen banal und redundant.

Liegt das spezifisch Neue jener Einsicht in die Möglichkeit außermenschlicher Quellen des Übels also gar nicht in deren Qualifizierung als „moralisch", sondern darin, dass sie mittels eines bestimmten Abstraktionslevels als *agents* gelten können? Welchen konkreten Beitrag soll diese sprachliche Differenzierung nun aber für eine bessere Strukturierung des Diskurses oder gar für eine Neuausrichtung in der Ethik leisten? Sind nicht das Verfahren der Haftbarmachung (zumindest in einer analogen Bedeutung des Wortes) und der davon angestoßene Prozess der systemischen Selbstregulierung auch ohne jede Annahme von *agency* möglich?

Diese Anfragen zielen in ihrem Kern auf den Zusammenhang der beiden von Floridi diagnostizierten Schwachpunkte herkömmlicher Ethik. Während sein Anliegen, ethisches Nachdenken auf die Betroffenen zu lenken, (trotz

[104] Vgl. ebd.

[105] Explizit konsequentialistisch äußert sich Floridi in Bezug auf „moral action": „An action is said to be morally qualifiable if and only if it can cause moral good or evil. An agent is said to be a moral agent if and only if it is capable of morally qualifiable action" (ebd. 364).

des vorschnellen Pauschalurteils über die „Standardethik") speziell für den *KI*-Kontext vernünftig und sinnvoll erscheint, erschließt sich der tiefere Sinn der *agency*-Umdeutung nicht. Als Voraussetzung für eine ethische Neuausrichtung ist sie zumindest augenscheinlich verzichtbar; der so energisch angemahnte Perspektivenwechsel dürfte auch zu realisieren sein, ohne dass großzügig *agent*-Etiketten ausgegeben werden.

Oder geht es am Ende gerade um die Art der Verpackung? Ist die Etikettierung als *moral agents* in Floridis Augen womöglich notwendig dafür, dass Ethik sich überhaupt zuständig fühlt für einen Bereich, der laut Floridi mit menschlichen Intentionen und Verantwortung nur noch wenig zu tun hat? Will man auch solche Spekulationen einmal dahingestellt sein lassen, ist trotzdem festzuhalten, dass nicht alle Glieder einer Handlungskette mit *moralisch* oder *agent* beschriftet sein müssen, um Gegenstand ethischen Nachdenkens zu werden. Auch in anthropozentrischen Verantwortungsanalysen ist eine hinreichende Kenntnis wirkursächlicher Zusammenhänge unabdingbar; ohne präzise Beschreibung der Kausalketten einer Handlung kann die Frage nach der Verantwortung schließlich gar nicht sinnvoll gestellt werden. Es verfestigt sich der Eindruck, dass der Moralbegriff für Floridis *agency*-Ansatz aus *agent*- wie *patient*-zentrierter Perspektive inhaltlich nichts beiträgt, dass es sich vielmehr um einen leeren Begriff handelt, wo er von Verantwortung getrennt wird. Diese Kritik trifft Floridi nicht überraschend; er selbst hat sie antizipiert und folgendermaßen zu widerlegen versucht: Der Einwand, es sei sinnlos, zwischen *moral agents* und *morally responsible agents* zu differenzieren, basiere nämlich auf der Annahme, dass Ethik sich ausschließlich mit der Frage nach Verantwortung zu beschäftigen habe. Einer solch reduktionistischen Aufgabenbeschreibung der Ethik stellt sich Floridi aber bekanntlich vehement entgegen: Ethik habe sich noch um andere Dinge zu kümmern als immer nur um Verantwortung, und dafür sei eben die klare Identifikation von *moral agents* erforderlich.[106] Hier kann man ihm zwar wieder nur beipflichten – ja, Ethik lässt sich nicht auf die Suche nach Verantwortlichen reduzieren –, doch wieder bleibt die Frage nach dem Nutzen seines Sprachgebrauchs von *moral agents* offen.

Nach diesen sachlogischen Anfragen an die innere Stringenz seines Ansatzes ist Floridis Grundanliegen – dass Ethik sich nicht länger mit der Frage nach Verantwortung aufhalten dürfe – schließlich auch inhaltlich zu kritisieren: Wenn Verantwortung innerhalb der Makro-Systemperspektive nurmehr eine nachrangige Kategorie darstellt, so besteht die Gefahr, dass Verantwortungsträger (Designer und Nutzer von *KI*) vorschnell entschuldigt werden und somit ethisch fragwürdige Anreizstrukturen entstehen.

[106] Vgl. *Floridi/Sanders*, On the Morality, 368.

3.2 Deborah G. Johnson

Johnson möchte mit dem *moral agency*-Begriff weder die menschliche Verantwortung unterlaufen noch maschinelles Verhalten als moralisch irrelevant abstempeln. In der Welt werden unser Handeln und unsere Verantwortung in moralisch signifikanter Weise von *tA* beeinflusst und dennoch liegt die Verantwortung klar beim Menschen. *Prima facie* hat Johnson tatsächlich eine Balance gefunden, die beides ernst nimmt: die moralische Dimension maschinellen Verhaltens und die Unhintergehbarkeit menschlicher Verantwortung. Damit ist ihr ein Ansatz gelungen, der unserer Alltagsintuition sehr nahe kommt und zudem anschlussfähig an herkömmliche moralphilosophische Standardpositionen ist. Allerdings wäre im Ringen um eine ausbalancierte und angemessene Heuristik der moralischen Dimension der Mensch-Maschine-Interaktion gleich zu Beginn ein differenzierter und abgrenzender Gebrauch der zwei zentralen Begriffe *moral agency* und „Verantwortung" wünschenswert gewesen. Die hier gebotene kritische Würdigung zielt darauf ab, dass Johnson mit ihrem jüngst entwickelten *triadic agency*-Modell ihrem Anliegen am nächsten kommt.

In ihrem Beitrag *Computer Systems: Moral Entities but not Moral Agents* (2006) spricht sich Johnson noch klar dagegen aus, *tA* als *moral agents* zu bezeichnen, da ihnen ein intentionaler Geisteszustand fehlt. Allerdings besitzen sie *intentionality* und *efficacy* und können sich weitgehend unabhängig von menschlicher Kontrolle verhalten, weswegen ihr Verhalten moralisch relevant und eine eigene Komponente einer moralischen Handlungskette sei. Ist es Johnson hier gelungen, die moralische Signifikanz maschinellen Verhaltens und seine Differenz zu menschlichen Handlungen hinreichend zu bestimmen? Ihr Hinweis, dass Kausalität, Automatisierung und die in ihrem Design eingebetteten Werte und Interessen moralisch relevant sind, ist richtig und wichtig, aber bereits weitgehend Konsens. An einer Stelle wird hier allerdings zu kurz gedacht: Lernende System gehen über die eingepflanzte Intentionalität hinaus. Selbst „normale", sich nicht selbst weiterentwickelnde oder selbst-lernende Technik wie Ultraschall (siehe Verbeek) verkörpern nicht einfach die Intentionalität der Designer. Offen bleibt hingegen, wie in einer Mensch-Maschine-Interaktion das maschinelle Verhalten ethisch gewichtet werden soll. Durch die Behauptung, „*tA* besitzen *intentionality*" kommt zur Hintertür die Frage nach der hinreichenden Abgrenzung zwischen maschinellem und menschlichem Verhalten hinein, die Johnson anfangs mit ihrem Votum, dass *tA* keine *moral agency* besitzen, momenthaft geklärt zu haben schien. Damit bürdet sie sich nicht nur die Aufgabe auf, fortlaufend darauf hinweisen zu müssen, dass mit der *intentionality* von *tA* etwas anderes gemeint sei als mit *intendings to act*, *intentional agency* oder *mental intentional states*. Werden die Formen der Intentionalität und Wirksamkeit von User, Designer und *tA* nicht hinreichend spezifiziert und voneinander abgegrenzt, erweckt dies den Anschein,

dass die Komponenten untereinander gleichrangig sind. Das mag zwar für eine kausale Erklärung, in der es unter notwendigen kausalen Bedingungen keine größere oder kleinere Notwendigkeit gibt, stimmen. Jedoch verhält es sich anders in einer ethischen Bewertung, wo jemand mehr, weniger oder gar nicht verantwortlich sein kann.

Den spezifischen Beitrag von *tA* erörtert Johnson mit Hilfe des *agency*-Begriffs.[107] Offen thematisiert sie die Ambivalenz des metaphorischen Gebrauchs von *(moral) agency* und „Verantwortung". Der Vorteil liege darin, dass durch diese starken Begriffe eine ethische Untersuchung eher auf die technologische Handlungskomponente achte; der Nachteil sei, dass diese Sprachpraxis die menschliche Verantwortung und mit ihr die ontologische Differenz zwischen Mensch und Maschine zu unterlaufen drohe. Diese Nachteile bewogen Johnson dazu, die Bezeichnung von *tA* als *moral agent* im Sinne einer *moral autonomy* strikt abzulehnen. Wieder soll die Frage gestellt werden, ob es Johnson gelungen ist, die moralische Bedeutung maschinellen Verhaltens hinreichend zu bestimmen – bei gleichzeitiger Erhaltung der moralischen Differenz zwischen Mensch und Maschine. Im dritten *agency*-Konzept von *moral autonomy* diente die Bezeichnung *moral agent* der Abgrenzung maschinellen Verhaltens vom menschlichen Handeln. Daher ist es unverständlich, warum Johnson in anderen Konzepten mit eben dieser Begrifflichkeit die Eigenart technischer Wirksamkeit positiv erhellen möchte. In ihren drei *agency*-Konzeptionen finden wir bezüglich menschlicher Akteure drei vorbehaltlos zulässige Attributionen für *(moral) agency* und „Verantwortlichkeit" und bezüglich *tA* zwei zulässige, aber metaphorische Attributionen und eine vorbehaltlos unzulässige Attribution. Da gleiche Ausdrücke für unterschiedliche Tätigkeiten und Akteure verwendet werden, können diese – erst recht in komplexen Mensch-Maschine-Settings – nur Verwirrung stiften. Leider zeigt Johnson nur bezüglich der *moral autonomy*-Konzeption klare Kante und schließt die *agency*-Zuschreibung für *tA* aus. Warum sie dies nicht auch bei den anderen zwei Konzeptionen tut, ist unverständlich, zudem es ihr ja durchaus gelingt, den spezifischen Beitrag maschinellen Verhaltens differenziert herauszuarbeiten: Dass durch Technik Gutes wie Schlechtes bewirkt werden kann oder Aufgaben und Funktionsweisen an *tA* delegiert werden können und so Interessen erfüllt oder verletzt werden, ist moralisch relevant, ohne dass gleichzeitig von einer spezifischen Verantwortung oder *agency* von *tA* gesprochen werden muss. Der Gebrauch von Metaphern ist abzuwägen. Metaphern können uns grundsätzlich helfen, komplexe Sachverhalte leichter zu verstehen. Sie sind als solche Einstiegshilfen. Ihr Einsatz ist aber dort nicht zweckdienlich, wo Komplexität nicht reduziert, sondern verfälscht wird oder das Risiko von Metaphern ihren Nutzen überwiegt.

[107] Vgl. u. a. *Johnson/Powers*, Computer Systems; *Johnson/Powers*, Computers als Surrogate Agents; *Johnson/Noorman*, Artefactual Agency.

Schließlich entwickelt Johnson gemeinsam mit Verdicchio das sogenannte *triadic agency*-Modell auf der Basis weiterführender Begriffsklärungen.[108] In diesem Modell unterscheidet sie sehr deutlich die *causal agency* von Artefakten (*agency* im weiten Sinn) von der *intentional agency* von Menschen (*agency* im strengen und klassischen Sinn). Letztere beinhalte *intentionality und causality*. Durch ihre Unterscheidung von zwei Wirkweisen gelingt es Johnson, die ontologische Differenz zwischen Mensch und Maschine in ihrer Handlungstheorie durchzuhalten. Obschon nur Menschen verantwortlich sein können, kann ihre Verantwortung nur im Blick auf alle Handlungskomponenten geklärt werden. Johnsons *triadic agency*-Modell bietet Ansätze für eine äußerst gelungene und mächtige Heuristik einer *technological moral action*, insoweit sie die je spezifische Kausalität und moralische Qualität maschinellen Verhaltens und menschlichen Handelns herausarbeitet und gleichzeitig das Handlungsergebnis als ein Zusammenspiel aller Komponenten begreift. Aufgrund der Bedeutung maschinellen Verhaltens möchte Johnson hier nicht auf den *agency*-Begriff verzichten. Die Möglichkeit einer weiten Interpretation von *agency* ist im Englischen zulässig, doch bietet sie auch dort Anlass für Missverständnisse und Nebenschauplätze, da ein und derselbe Begriff auf Mensch und Maschine referiert. Johnson ist bemüht, die Differenz und wechselseitige Bezogenheit von Mensch und Maschine sachgerecht zu benennen, doch scheut sie vor dem letzten Schritt zurück und verzichtet nicht auf die *agency*-Attribution für *tA*. Dabei besitzt *agency* kein Alleinstellungsmerkmal und es gibt andere Begriffe (zum Beispiel *factor*), welche die deskriptive wie normative Relevanz und Eigenart von *tA* unmissverständlich und präzise beschreiben können.

3.3 Peter-Paul Verbeek

Der Grundgedanke der Vermittlungstheorie, dass Mensch und Technologie so stark mit-, auf- und ineinanderwirken, dass dabei ontologische und moralische Grenzen verschwimmen, besticht zunächst durch seine Anerkennung der Komplexität von Mensch-Maschine-Interaktionen. Dass der Weltzugang heutiger Menschen zunehmend technologisch vermittelt ist – sowohl hermeneutisch als auch pragmatisch – und dass im Nachhinein häufig nur noch schwerlich einzelne Akteure in ihren Wirkweisen und Verantwortlichkeiten voneinander abgegrenzt werden können, lässt sich an mannigfaltigen Beispielen nachvollziehen. Verbeeks Ansatz gelingt es also in besonderem Maße, Realität abzubilden. Wenn Ethik sensibel dafür ist, dass Technik mit Menschen zusammen Wirklichkeit schafft, und wenn über dieser Einsicht fixierte Rollenzuschreibungen und Bewertungsmuster zwischen Mensch und Maschine zunächst zurücktreten, dann ist das für

[108] Vgl. *Johnson/Verdicchio*, AI.

die Debatte über einen verantwortlichen Umgang mit Technik grundsätzlich ein Gewinn.

Es ist allerdings zu fragen, was die Vermittlungstheorie darüber hinaus leistet beziehungsweise inwiefern der ethische Diskurs, welcher auch auf die Lösung konkreter Probleme abzielt, von ihr profitieren kann.

Auch hier bricht zunächst die grundsätzliche Frage nach der Bedeutung des Wortes *agency* auf. Verbeek bestreitet, dass *agency* eine statische Eigenschaft ist, vielmehr sei sie immer erst Ergebnis von Interaktionen. Somit kommt die deutsche Übersetzung „Handlungsfähigkeit" schon nicht mehr in Frage, ist doch damit eindeutig eine Eigenschaft benannt, die als Grundlage für und folglich auch vor einer Handlung gegeben sein muss – nicht erst als Ergebnis davon. Die deutsche Übersetzung „Wirkung" träfe schon eher Verbeeks Stoßrichtung, wäre in ihrem Bedeutungsgehalt aber ernüchternd banal. *Agency* als „Handeln" würde einen starken anthropologischen Akzent setzen und damit abermals auf die moralische Relevanz von Technologie verweisen, indem ihre Wirkungen mit menschlichem Handeln auf eine Stufe gehoben würden. Allerdings ist laut Verbeek auch der Mensch für sich betrachtet kein *moral agent*, kein moralisch Handelnder, womit der anthropologische Referenzpunkt schon wieder hinfällig wird.

Ungeachtet solch konkreter Übersetzungsschwierigkeiten ist zu fragen, was mit der Kategorie *moral agency* eigentlich philosophisch ausgesagt ist. Die klassischen Bedeutungskomponenten „Intentionalität" und „Freiheit" sind offenbar in veränderter – eben vermittelter, hybrider – Form weiterhin Bestandteil dieses Begriffs. Allerdings hat eine Verschiebung von einer Kategorie der Ermöglichung zu einer Kategorie der Bilanzierung stattgefunden: Intentionalität und Freiheit sind nicht mehr Grundlage für *moral agency*, sondern – wie diese selbst – erst das Ergebnis komplexer Mensch-Maschine-Verflechtungen. Zudem ist *moral agency* nicht mehr konkret einzeln zuschreibbar, da sie immer nur aus einem Gesamtgefüge hervorgeht. Verantwortung spielt in diesem Konstrukt nicht direkt eine Rolle. Zwar betont Verbeek die prospektive Verantwortung der Technik-Designer, ein mögliches Verfahren zur retrospektiven Verantwortungszuschreibung kommt jedoch nicht zur Sprache. Ein etwaiger begrifflicher Zusammenhang mit *moral agency* wird nicht deutlich.

Letztlich bleibt also der Verdacht, dass Verbeeks Schlüsselbegriff der *moral agency* ein hermeneutisches Begriffskonstrukt ist, das letztlich keine Verständnishilfe bieten, geschweige denn das Problem der Verantwortungszuschreibung lösen kann.

Um die innige Verbindung von Mensch und Technik zu verdeutlichen und ihre moralische Relevanz herauszustellen, bedarf es keiner philosophischen Großkaliber wie des traditionell schon stark vorgeprägten Begriffs der *moral agency*. Was Verbeek anhand vieler Beispiele so anschaulich und treffend beschreibt, kann in Erklärungsmodellen aufgehoben werden, die ohne diesen Terminus auskommen.

Wie sieht es nun aber mit dem Begriff der Vermittlung aus? Trägt er etwas zum besseren Verständnis der Sachlage bei? Ja und nein: Zum einen ist der vermittelnde Charakter einer Ultraschalldiagnostik natürlich einleuchtend, insofern als die Beziehung der werdenden Eltern zu ihrem Kind nicht unvermittelt ist. Zum anderen wäre aber zu fragen, ob Vermittlung als spezifisches Charakteristikum von Technik gelten kann. Ist der Weltzugang des Menschen nicht immer vermittelt? Wo die Technik fehlt, ist es vielleicht die soziale Vermittlung durch andere Menschen, die ihn hermeneutisch und pragmatisch beeinflusst und hybride Intentionalität erzeugt.

Eine „direkte" Beziehung zwischen dem Subjekt Mensch und dem Objekt Welt „an sich" anzunehmen, erscheint auch reichlich naiv. Vermutlich geht Verbeek aber nicht von diesem krassen Gegenbild aus, sondern will lediglich aufzeigen, dass die Intensität der Welt-Vermittlung durch Technik gravierend zugenommen hat. Dem ist auch nicht zu widersprechen. Darf dann allerdings die Konzeptualisierung dieses Phänomens so weit gehen, den Menschen gar nicht mehr ohne technologische Vermittlung zu denken? Beispiele wie Cyborgs sind Extreme, die zwar sofort einleuchten, aber eben (noch) nicht repräsentativ für die Lebenswirklichkeit vieler Menschen sind. In Verbeeks Vermittlungstheorie ist für menschliche *moral agents*, die zumindest situativ auch mal unabhängig von technologischen Einflüssen handeln, kein Platz. Alles ist immer schon Gemisch – und der einzelne Mensch gerät aus dem Blick. Hier ist auf einer grundsätzlicheren Ebene Kritik an Verbeeks radikalem Konstruktivismus zu üben: Trotz wechselseitiger Formung von Mensch und Maschine muss zu Beginn eine Polarität angenommen werden können, aus der heraus Interaktion entsteht. Auch Freiheit und Intention müssen in einem Mindest-Kernbestand als dieser Interaktion vorgängig betrachtet werden, damit der Mensch noch als handlungsfähiges Subjekt denkbar bleibt.

3.4 Konkretisierung in Bezug auf den Dieselskandal

Im Anschluss an die kritischen Untersuchungen sollen nun zentrale Aspekte der jeweiligen Zugänge anhand des VW-Dieselskandals konkretisiert und veranschaulicht werden.

3.4.1 Luciano Floridi: Steuerungsgerät ist kein *moral agent*

Innerhalb komplexer Mensch-Maschine-Interaktionen kann der einzelne Akteur die unmittelbaren Folgen seiner Handlung immer schwieriger abschätzen. Dies ist ein Grund für Floridis Skepsis, dass die Klärung der Schuldfrage in der Ethik vorrangig sei. Tatsächlich wussten aber im Dieselskandal Hersteller wie Designer von der Manipulationsgefahr der Abschalteinrichtung und mussten mit den damit verbundenen Risiken kalkulieren. Die Entwicklung und Applikation der Betrugssoftware ist noch relativ über-

schaubar darzustellen, und so könnten auch Verantwortliche klar benannt werden.

Dessen ungeachtet bleibt es für Floridi bei der Empfehlung an die Ethik, ihre Kräfte nicht für retrospektive Schuldzuschreibungen aufzubrauchen, sondern diese zu bündeln, um den entstandenen Schaden bei den betroffenen Menschen zu lindern. Ist man dennoch bereit, sich von der betroffenen- auf die akteurszentrierte Sicht umzustellen, so wird schnell klar, dass weder aus der Sicht des Herstellers oder Designers noch aus Kundensicht die *Electronic Diesel Control* als *moral agent* wahrgenommen werden kann. Dies kann nur zum Teil dadurch erklärt werden, dass viele Menschen noch glauben, durch ihren Fuß am Gaspedal den Motor mechanisch zu steuern. (Tatsächlich sitzen sie in einem rollenden Computer.) Der entscheidende Grund liegt entsprechend Floridis Nomenklatur darin, dass das Steuerungsgerät zwar interaktiv (da sich beobachten lässt, dass ein bestimmter Input einen bestimmten Output generiert), allerdings unzureichend autonom (da es nur auf spezifische Kennfelder reagiert, folglich ohne vorherigen Input keinen beobachtbaren Output generiert) und nicht adaptiv ist (da sich kein selbstständiger Wechsel zwischen interaktivem und autonomem Verhalten beobachten lässt): Aufgrund von Sensordaten (beispielsweise Bewegung der Räder, Drehzahl und Temperatur des Motors oder Dauer des Fahrzyklus) erkennt die Steuerung Anfang und Ende des Rollentests und schaltet aufgrund eines eigenen Programms vom Normalbetrieb auf andere Einstellungen um, damit die Abgaswerte auf dem Prüfstand eingehalten werden. Am Ende des Tests oder zu Beginn des Normalbetriebs wird die Abgasreinigungsanlage wieder abgeschaltet oder weniger wirksam gemacht. Hinreichend autonom und adaptiv wäre zum Beispiel ein auf *Machine-Learning*-Algorithmen basierendes Steuerungsgerät, das aus einer unüberschaubaren Menge an Datensätzen eine eigenständige Lösung zur Erreichung eines vorgegebenen Zieles (hier: das Bestehen des Tests) generieren könnte. Dann wäre es nach Floridi gerechtfertigt, die Steuerung als *moral agent* zu bezeichnen. Die systematischen Bedenken gegen eine solche Bezeichnungspraxis wurden bereits weiter oben entfaltet: Eine ethische Herangehensweise, die zentrale menschliche Handlungskomponenten (hier: Hersteller und Zulieferer) abstrahiert, kann nicht mehr nach der Verantwortung fragen. Ferner trägt aus deskriptiver Sicht die Bezeichnung *moral agent* nichts dazu bei, die Komplexität des Handlungsgeflechts angemessen zu begreifen.

Im Falle, dass das Steuerungsgerät kein *moral agent* ist, besteht aber ein entgegengesetztes Risiko: Mit Verbeek kann gesagt werden, dass unterhalb einer gewissen Wahrnehmungsschwelle die technische Komponente ganz ausgeblendet und für eine ethische Betrachtung zentrale Merkmale übersehen werden. Jedoch sollte sowohl in einer ethischen als auch in einer rein deskriptiven oder phänomenologischen Untersuchung auf die Wirkweise und

die vermittelnde hermeneutische wie pragmatische Kraft eines technischen Artefaktes geachtet werden.

3.4.2 Deborah G. Johnson: *triadic agency*

Mit ihrem jüngsten Erklärungsmodell – dem *triadic agency*-Modell – erfasst Johnson unterschiedliche *agency*-Formen von menschlichen sowie technischen Akteuren.[109] Die am Dieselskandal beteiligten Gruppen können ihnen folgendermaßen zugeordnet werden: Erstens möchte das Management des Herstellers (= *user/s*), dass sein Auto den Test besteht (= *intentional agency*). Zweitens akzeptiert der Zulieferer (= *designer*) die Absicht des Herstellers und entwickelt (= *intentional agency*) ein entsprechendes Steuerungsgerät. Drittens bewirkt das Steuerungsgerät (= *artefact; causal agency*), dass das Auto den Test besteht. Erst durch das Zusammenwirken aller drei Komponenten wird das Handlungsziel erreicht. Die Wirkung des Steuerungsgerätes ist Teil der manipulativen Handlung, obschon es diese nicht intendieren kann und deshalb nicht verantwortlich ist. Für Johnson kann in voraussehbarer Zukunft *KI* niemals für etwas verantwortlich sein. Das gilt auch dann, wenn eine reichlich fortgeschrittene *KI* die Rolle des Herstellers und/oder des Designers übernimmt. Auch diese wirke auf Basis eines Programms weiterhin kausal und könne den ontologischen Graben zwischen Mensch und technischem Artefakt nicht überbrücken. Selbst wenn man in der direkten Handlungskette in allen Rollen zunächst nur auf *KI* träfe und auf keinen Menschen, müsse man, um die Verantwortlichen zu finden, solange zurückgehen, bis man auf Menschen treffe, die bestimmte Aufgaben an die *KI* delegiert oder diese programmiert hätten.

Johnsons Ergebnis ist angesichts ihrer ontologischen und handlungstheoretischen Prämissen konsequent. Sie differenziert begrifflich den Beitrag jeder Komponente und vermag so ein annähernd sachgerechtes Bild der Mensch-Maschine-Interaktion zu vermitteln. Allerdings gelingt es ihr lediglich, die Eigenart von *tA* durch die hybride Bezeichnung *causal agency* negativ zu bestimmen. *Causal* besagt, dass *tA* keine *moral agents* und ohne Intention sind; *agency* verweist darauf, dass *tA* sich anders verhalten als natürliche Ereignisse.

Mag durch das *triadic agency*-Modell der Beitrag des Steuerungsgerätes noch hinreichend systematisch eingeholt werden, so verliert das Modell dort seine Erklärungskraft, wo zum Beispiel *KI* in Form selbstlernender Algorithmen am Werk ist. Denn angesichts des Potentials und der moralischen Bedeutung des Einsatzes sogenannter „nicht-deterministischer" *KI* wäre eine positive Bestimmung ihres Verhaltens wünschenswert.

Schließlich ist ihre konsequente Suche nach Verantwortlichen unter den menschlichen Akteuren beispielhaft, wenn auch für viele Hersteller, Desig-

[109] Vgl. *Johnson/Verdicchio*, AI.

ner und Anwender schmerzlich. Denn diese müssten sorgfältig den Einsatz nicht-deterministischer *KI* auf Gebieten mit hohen Sicherheitsanforderungen abwägen und womöglich aufgrund des hohen Risikos auf den Nutzen selbstlernender Algorithmen verzichten.

3.4.3 Peter-Paul Verbeek: Mobilität als *composite moral agency*

Verbeeks These, dass weder der einzelne Mensch noch technische Geräte *moral agency* besitzen, sondern dies ein Ergebnis ihrer Interaktion sei, lässt aufhorchen. Wie kann der Dieselskandal von Verbeeks Zugang her gedeutet werden? Zunächst ist für den Menschen die Mobilität zunehmend technisch vermittelt. Im Fahren „verschmelzen" Mensch und Technik miteinander: Das Verhalten eines Motors gibt nicht der Fuß am Gaspedal allein vor, sondern es wird gesteuert von leistungsstarken Programmen, die in Echtzeit enorme Datenmengen zahlreicher Geräte und Sensoren verarbeiten. Dadurch beeinflussen Steuerungsgeräte nicht nur aktiv das Fahren (pragmatische Vermittlung), sondern zugleich das Fahr- und Mobilitätsverständnis und -erlebnis von Entwicklern, Herstellern und Kunden (hermeneutische Vermittlung). Steuerungsgeräte machen unsere Mobilität sicherer (ABS oder ESP), adaptiver (Fahrprogramme), angenehmer (Klimaanlage) aber auch manipulativer (Abschaltvorrichtung). Im Ergebnis ließe sich so das Fahren oder das Kreieren von Mobilität als *composite moral agency* rekonstruieren.

Verbeeks Gedanken sind einerseits eine wertvolle Hilfe, um die technische Vermittlung von Mobilität und die menschliche Sicht auf diese zu veranschaulichen. Andererseits wurde bereits oben ausgeführt, dass nicht ersichtlich ist, wie von seinem Konstrukt *composite moral agency* eine Brücke zur moralischen Verantwortung gebaut werden kann. Denn letztere tragen laut Verbeek insbesondere solche Menschen (in diesem Fall der Hersteller und Zulieferer), welche die komplexen Interaktionsmuster verstehen und gestalten können. Bleibt aber auf Seiten des menschlichen Akteurs ein hinreichender Rest an moralischer Verantwortung, Freiheit und Intention übrig, wenn alles Ergebnis von Interaktion ist? Wie kann eine solche Kreisbewegung durchbrochen werden? Die Konzepte *moral agency* und „moralische Verantwortung" stehen, anders als Verbeeks Hinweis auf die besondere Verantwortung von Designern uns glauben lässt, begrifflich merkwürdig nebeneinander, als ob sie nichts miteinander zu tun hätten.

Verbeeks zwei Anliegen – eine Rekonstruktion des Verständnisses der Mensch-Maschine-Interaktion und der Zuschreibung moralischer Verantwortung – ließen sich auch erfüllen, wenn die menschlichen Akteure *moral agents* blieben. Denn die Erkenntnis, dass sich menschliche Handlungsfähigkeit stets vermittelt konkretisiert und diese vermittelte Freiheit nie reflexiv eingeholt werden kann, ist philosophisch gesehen nichts Neues. Jedoch muss zur Vermeidung eines Zirkelschlusses bei der Zuschreibung von *moral agency* und moralischer Verantwortung die Freiheit der mensch-

lichen Akteure als vorgängig betrachtet werden. Denn die Interaktion ist nicht selbstursprünglich, sondern Folge der menschlichen Fähigkeit, frei zu reflektieren, zu entscheiden und zu handeln. Verbeek zeigt leider nicht auf, wie ein freies, das heißt moralisches, Subjekt in seinem konstruktivistischen Ansatz denkbar ist.

Allerdings wäre nicht die Dekonstruktion, sondern die Stärkung des moralischen Subjekts vordringliche Aufgabe, wenn Menschen die Entwicklung und den Einsatz von Technik kritisch reflektieren und verantwortlich gestalten sollen. Mit Verbeek kann gesagt werden, dass im Dieselskandal Hersteller und Zulieferer ihre Kunden und die Gesellschaft getäuscht haben, da sie ohne deren Wissen und entgegen ihren Interessen eine bestimmte Art von Mobilität verkauft haben. Gleichzeitig müssen sich Gesellschaft und Politik nicht nur fragen, ob sie ihre Interessen entschieden genug vertreten haben. Zusätzlich müssten sie kritisch ihr Mobilitätsverständnis prüfen. Offensichtlich vertraut man allzu gerne Versprechungen, dass eine saubere, bezahlbare, leistungsstarke und überall verfügbare Mobilität technisch machbar ist.

4. Ausblick

Die hier angestellte Untersuchung hat die Chancen und Risiken unterschiedlicher Attributionswege von *moral agency* auf die Mensch-Maschine-Interaktion herausgearbeitet. Schlussendlich kommt man zu dem Ergebnis, dass die Risiken einer *agency*-Attribution auf maschinelles Verhalten den Nutzen überwiegen. Daher sollte eine sachgerechte Differenzierung zwischen Mensch und Maschine auch begrifflich erkennbar sein. So kann die Mensch-Maschine-Interaktion nicht nur präziser beschrieben, sondern auch die normative Struktur deutlicher herausgestellt werden. Dadurch wird angesichts zunehmend komplexer und sich ausdifferenzierender Mensch-Maschine-Interaktion auch das nicht vergessen, worum es der Ethik ab der zweiten Hälfte des letzten Jahrhunderts im Blick auf den technischen Fortschritt vornehmlich ging: die Verantwortung von Menschen.

Summary

Due to the increasing moral significance of technical artefacts, increasingly common approaches in technical ethics attempt to expand the attribution of *agency* to technical artefacts, a task traditionally reserved for human beings.

These approaches argue that this is the only proper way to describe and evaluate the complexity of human-machine-interaction from an ethical perspective, especially with regard to the question of responsibility.

In the following article, we present three of these approaches (L. Floridi, D. G. Johnson, P.-P. Verbeek) and analyse to what extent they do justice to their claim to clarify matters both descriptively and normatively.

To test their explanatory power, each concept is spelled out in practice, using the diesel emissions scandal as an example.

The study comes to the conclusion that attributing *agency* to technical artefacts creates no additional benefit but causes even more confusion and runs the additional risk of obscuring ethical responsibility.