Otto Richter · Uwe Drewitz ·
Reinhold Haux · Stefan Heuser ·
Tim Kacprowski · Jochen Steil *Hrsg.*

Zusammenwirken von natürlicher und künstlicher Intelligenz: Beurteilen-Messen-Bewerten

OPEN ACCESS

Springer VS

Zusammenwirken von natürlicher und künstlicher Intelligenz: Beurteilen-Messen-Bewerten

Otto Richter · Uwe Drewitz ·
Reinhold Haux · Stefan Heuser ·
Tim Kacprowski · Jochen Steil
(Hrsg.)

Zusammenwirken von natürlicher und künstlicher Intelligenz: Beurteilen-Messen-Bewerten

Springer VS

Hrsg.
Siehe die nächste Seite

ISBN 978-3-658-45844-7 ISBN 978-3-658-45845-4 (eBook)
https://doi.org/10.1007/978-3-658-45845-4

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über https://portal.dnb.de abrufbar.

Braunschweigische Wissenschaftliche Gesellschaft und Technische Universität Braunschweig

Planung/Lektorat: Cori Antonia Mackrodt
Springer VS ist ein Imprint der eingetragenen Gesellschaft Springer Fachmedien Wiesbaden GmbH und ist ein Teil von Springer Nature.
Die Anschrift der Gesellschaft ist: Abraham-Lincoln-Str. 46, 65189 Wiesbaden, Germany

Hrsg.
Otto Richter
Institut für Geoökologie
Technische Universität Braunschweig,
Altpräsident der Braunschweigischen
Wissenschaftlichen Gesellschaft
Braunschweig, Deutschland

Reinhold Haux
Peter L. Reichertz Institut für
Medizinische Informatik
Technische Universität Braunschweig
und Medizinische Hochschule Hannover,
Präsident der Braunschweigischen
Wissenschaftlichen Gesellschaft
Braunschweig, Deutschland

Tim Kacprowski
Peter L. Reichertz Institut für
Medizinische Informatik
Technische Universität Braunschweig
Braunschweig, Deutschland

Uwe Drewitz
Institut für Verkehrsforschung
Deutsches Zentrum für Luft- und
Raumfahrt e. V. (DLR), Berlin
Braunschweig, Deutschland

Stefan Heuser
Institut für Theologie und
Religionspädagogik
Technische Universität Braunschweig
Braunschweig, Deutschland

Jochen Steil
Institut für Robotik und
Prozessinformatik
Technische Universität Braunschweig,
Geschäftsführer Gauss Robotics GmbH
Braunschweig
Braunschweig, Deutschland

Vorwort der Herausgeber

Wie wird das Zusammenleben und -wirken von Menschen, Tieren und Pflanzen einerseits und Maschinen andererseits zukünftig aussehen? Lassen sich Umfang und Intensität der neuen Synergien bestimmen? Die Kommission Synergie und Intelligenz: technische, ethische und rechtliche Herausforderungen des Zusammenwirkens lebender und nicht lebender Entitäten im Zeitalter der Digitalisierung (SYnENZ) der Braunschweigischen Wissenschaftlichen Gesellschaft (BWG) befasst sich mit den sich durch diese Entwicklungen ergebenden Formen des Zusammenlebens. Eine sehr wichtige Rolle spielt dabei die Frage, womit wir es beim erweiterten Zusammenwirken mit künstlicher Intelligenz überhaupt zu tun haben. Ansätze, diese begrifflich-theoretisch zu beurteilen, im konkreten Fall messbar zu machen, und auch ethisch zu bewerten standen im Zentrum des 2. SYnENZ-Symposiums, welches sich nach einer Diskussion soziologischer und ethischer Aspekte entsprechend in drei Blöcke strukturierte.

Beurteilen: Neue Formen des Zusammenwirkens erfordern kritisches, d. h. „unterscheidendes“ Beurteilen. Wie kann solches Urteilen zu einer Ausdifferenzierung von Perspektiven, aber auch zur Erschließung gemeinsamer Konzepte und Kontexte beitragen, auf die sich die Forschung interdisziplinär beziehen kann?

Messen: Wie lässt sich erweitertes Zusammenwirken adäquat messen und untersuchen? Gibt es existierende empirische Ansätze, die hier genutzt werden könnten, z. B. randomisierte Studien, wie sie in der Medizin üblich sind? Diese Fragen werden insbesondere im Teil Messen des Symposiums aufgegriffen und diskutiert.

Bewerten: Der dritte Teil geht Herausforderungen für die ethische Bewertung nach. Welche Rolle spielen etwa das Berufsethos der beteiligten Professionen sowie die betroffenen Lebensformen und Praktiken? Müssen Normen und Wertvorstellungen angepasst werden?

Das vorliegende Buch basiert in großen Teilen auf Ausarbeitungen der Präsentationen des Symposiums. Wir danken den Autorinnen und Autoren für die Beiträge zu diesem Buch. Unser besonderer Dank gilt Frau Dr. Cori Antonia Mackrodt vom Springer Verlag sowie Frau Nezahat Mumcu und Frau Jeannette Rotermund von der Braunschweigischen Wissenschaftlichen Gesellschaft.

Braunschweig
im März 2024

Otto Richter
Uwe Drewitz
Reinhold Haux
Stefan Heuser
Tim Kacprowski
Jochen Steil

Inhaltsverzeichnis

Über die Autoren

Prof. Dr. Susanne Beck Kriminalwissenschaftliches Institut der Leibniz Universität Hannover.

Susanne Beck ist Professorin für Strafrecht, Strafprozessrecht, Strafrechtsvergleichung und Rechtsphilosophie in Hannover. Nach Promotion und Habilitation an der Universität Würzburg erfolgte 2013 der Ruf nach Hannover. Sie ist Mitbegründerin der Forschungsstelle RobotRecht in Hannover und arbeitet seit über einem Jahrzehnt an Fragen der Regulierung neuer technologischer sowie medizinischer Entwicklungen. Sie ist u. a. Mitglied der Plattform Lernende Systeme, von acatech sowie der Akademie für Ethik in der Medizin und der Braunschweigischen Wissenschaftlichen Gesellschaft.

Professor Dr. Klaus Bengler Lehrstuhl für Ergonomie (LfE), Technische Universität München.

Prof. Bengler studierte Psychologie an der Universität Regensburg und promovierte dort im Jahr 1995 in Kooperation mit der BMW Group zur Informationsgestaltung von Navigationsinformation für den Fahrer. Anschließend führte er das Team „Mensch-Maschine Interaktion“ in der BMW Forschung & Technik und das angeschlossene Usability Labor. Seit 2009 leitet er den Lehrstuhl für Ergonomie an der Technischen Universität München. Sein Forschungsgebiet umfasst den Bereich der sogenannten „Micro ergonomics“ zu Fragen der Mensch-Maschine-Interaktion, insbesondere den Bereich der Fahrerassistenz, der Softwareergonomie und der Kooperation zwischen Mensch und Roboter.

Seine Forschung schließt dabei sowohl anthropometrische als auch kognitive Fragestellungen ein.

PD Dr. Barbara Buchberger, MPH, MPhil Robert Koch-Institut.

Barbara Buchberger hat ein Violinstudium an der Hochschule der Künste Berlin absolviert und nach ihrem künstlerischen Abschluss als angestellte und freiberuflich tätige Musikerin gearbeitet. Für das Aufbaustudium Public Health an der Technischen Universität Berlin wählte sie den Schwerpunkt Epidemiologie und Methoden. Einer Weiterbildung in Medizinethik folgte ein Masterstudium der Philosophie an der FernUniversität in Hagen. Bis zum Beginn ihrer Beschäftigung beim Robert Koch-Institut im Jahr 2018 war sie wissenschaftliche Mitarbeiterin und Leiterin der Forschungsgruppe „Health Techology Assessment und systematische Reviews“ am Lehrstuhl für Medizinmanagement der Universität Duisburg-Essen. Im April 2021 wurde sie von der Medizinischen Fakultät der Universität Duisburg-Essen für das Fach „Gesundheitsökonomie, Gesundheitssystem, Öffentliches Gesundheitswesen“ habilitiert.

PD Dr. Bruno Gransche Institut für Technikzukünfte ITZ am Karlsruher Institut für Technologie KIT.

Der Philosoph und Zukunftsforscher forscht und lehrt in den Bereichen Technikphilosophie/Ethik und Zukunftsdenken mit Fokus u. a. auf Philosophie neuer Mensch-Technik-Relationen, gesellschaftliche & ethische Aspekte von KI & Digitalisierung, Technikbilder/Menschenbilder/Metaphernanalyse sowie Vorausschauendes Denken. Gransche ist Privatdozent am Institut für Technikzukünfte der Universität Karlsruhe seit 2020; Studium der Philosophie und Literaturwissenschaft sowie Promotion in Heidelberg, Habilitation in Karlsruhe. Er ist u. a. Mitherausgeber der Reihe *Techno:Phil – Aktuelle Herausforderungen der Technikphilosophie* sowie Fellow am Fraunhofer-Institut für System- und Innovationsforschung ISI in Karlsruhe, wo er bis 2016 in der Abteilung Foresight arbeitete.

Prof. Dr. Reinhold Haux Peter L. Reichertz Institut für Medizinische Informatik der TU Braunschweig und der Medizinischen Hochschule Hannover (PLRI).

Reinhold Haux ist Präsident der Braunschweigischen Wissenschaftlichen Gesellschaft (BWG) und emeritierter Professor für Medizinische Informatik am Peter L. Reichertz Institut für Medizinische Informatik der TU Braunschweig und der Medizinischen Hochschule Hannover (MHH). Nach Professuren an Universitäten in Tübingen (1987–1989), Heidelberg (1989-2001) und Innsbruck

(2001–2004) folgte er 2004 einem Ruf an die Technische Universität Braunschweig. Er war Präsident der International Medical Informatics Association (2007–2010), der International Academy of Health Sciences Informatics (2018-2020) und Herausgeber der Zeitschrift Methods of Information in Medicine (2001–2015). Er ist Honorarprofessor an der Universität Heidelberg und kooptiertes Mitglied des Lehrkörpers der MHH. Seit ihrer Gründung 2017 ist er Mitglied der SYnENZ-Kommission der BWG. Weitere Informationen auf www.plri.de.

Prof. Dr. Stefan Heuser Institut für Evangelische Theologie und Religionspädagogik der Technischen Universität Braunschweig.

Stefan Heuser ist Professor für Systematische Theologie mit dem Schwerpunkt Ethik am Institut für Evangelische Theologie und Religionspädagogik der Technischen Universität Braunschweig. Zuvor war er Professor für Ethik in der Pflege an der Evangelischen Hochschule in Darmstadt, Privatdozent an der Goethe-Universität Frankfurt am Main sowie Pfarrer der Evangelischen Kirche in Hessen und Nassau. Er ist stellvertretender Sprecher der SYnENZ-Kommission der Braunschweigischen Wissenschaftlichen Gesellschaft.

Prof. Dr. Tim Kacprowski Peter L. Reichertz Institut für Medizinische Informatik der TU Braunschweig und der Medizinischen Hochschule Hannover (PLRI).

Tim Kacprowski leitet seit 2020 die Abteilung Data Science in Biomedicine des PLRI an der TU Braunschweig. Seine Forschung konzentriert sich auf die Kombination von Netzwerkbiologie und Machine Learning um Verfahren zur Auswertung biomedizinischer Daten zu entwickeln. Ferner beschäftigt er sich mit den Bereichen Immersive Analytics und Science of Science. Er leitet derzeit die Arbeitsgruppe „Statistische Methoden der Bioinformatik“ der GMDS e. V.

Prof. Dr. Arne Manzeschke Institut für Pflegeforschung, Gerontologie und Ethik (IPGE), Evangelische Hochschule Nürnberg.

Programmierer im ersten Beruf; studierte Theologie und Philosophie; Promotion und Habilitation in Erlangen. Er lehrt Ethik und Anthropologie an der Evangelischen Hochschule Nürnberg und leitet dort das IPGE. Seit 2010 forscht er zu Mensch-Technik-Verhältnissen. Aktuell ist er Sprecher des BMBF-geförderten Forschungsclusters „Integrierte Forschung“, das sich mit methodischen und inhaltlichen Fragen einer inter- und transdisziplinären Forschung im Bereich der Mensch-Technik-Interaktion befasst. Er ist Sprecher des Fachausschusses „Medizintechnik und Gesellschaft“ bei der Deutschen Gesellschaft für Biomedizinische Technik (DGBMT) und Vorsitzender der Ethikkommission für Pflege- und Sozialforschung an der Evangelischen Hochschule Nürnberg.

Prof. Dr. Dr. Sabine Salloch Institut für Ethik, Geschichte und Philosophie der Medizin, Medizinische Hochschule Hannover.

Sabine Salloch ist Professorin für Ethik und Geschichte der Medizin und leitet das Institut für Ethik, Geschichte und Philosophie der Medizin der Medizinischen Hochschule Hannover. Ausgehend von einem Doppelstudium der Medizin und der Philosophie und Promotionen in beiden Fächern war sie wissenschaftliche Mitarbeiterin an der Ruhr-Universität Bochum sowie Juniorprofessorin und Institutsleitung an der Universitätsmedizin Greifswald. Sie ist Mitglied im Vorstand der Zentralen Ethikkommission bei der Bundesärztekammer und stellvertretende Vorsitzende der Zentralen Ethik-Kommission für Stammzellenforschung.

Prof. Dr. Jochen J. Steil Institut für Robotik und Prozessinformatik an der Technischen Universität Braunschweig.

Jochen Steil ist Leiter des Instituts für Robotik und Prozessinformatik und Sprecher der Kommission SYnENZ: Synergie und Intelligenz: technische, ethische und rechtliche Herausforderungen des Zusammen¬wirkens lebender und nicht-lebender Entitäten im Zeitalter der Digitalisierung (SYnENZ) der Braunschweigischen Wissenschaftlichen Gesellschaft (BWG). Er studierte Mathematik und Slawistik an der Universität Bielefeld, promovierte 1999 in der Informatik über Neuronale Netze und beschäftigt sich seitdem mit Robotik, Roboterlernen und Mensch-Maschine Interaktion. Herr Steil koordinierte mehrere europäische Verbundprojekte, war Mitglied des wissenschaftlichen Boards des DFG Exzellenzclusters in Kognitiver Interaktionstechnologie (CITEC) und Leiter des Research Institute for Cognition and Robotics (CoR-Lab) an der Universität Bielefeld. Von 2015–2020 war er Visiting Professor der Oxford Brookes Universität und im Jahr 2016 folgte er einem Ruf an die Technische Universität Braunschweig als Professor für Robotik. Er ist Mitglied der Plattform lernende Systeme des BMBF, die Expert:innen zu aktuellen Themen der künstlichen Intelligenz und gesellschaftlichen Fragen zusammenbringt. Im Jahr 2023 war er von März-Juli Visiting Fellow am Okinawa Institute für Science and Technology, Japan und ist seit Februar 2023 auch als Mitgründer und Geschäftsführer der Gauss Robotics GmbH in Braunschweig tätig.

Simone Tiedau Simone Tiedau ist Diplomjuristin und arbeitete bis 2023 als Wissenschaftliche Mitarbeiterin im Verbundsprojekt „MeHuCo - Autonome Waffensysteme zwischen Regulation und Reflexion“ bei Prof. Dr. Susanne Beck am Lehrstuhl für Strafrecht, Strafprozessrecht, Strafrechtsvergleichung und Rechtsphilosophie.

Prof. Dr. Johannes Weyer Fakultät Sozialwissenschaften der TU Dortmund.

Johannes Weyer ist Seniorprofessor für Nachhaltige Mobilität an der Fakultät für Sozialwissenschaften der TU Dortmund und war von 2002-2022 Professor für Techniksoziologie an der Wirtschafts- und Sozialwissenschaftlichen Fakultät der TU Dortmund. Das Spektrum seiner Forschungsarbeit umfasst u. a. Technikbewertung und Technikakzeptanzforschung, Risikomanagement in Organisationen, agentenbasierte Modellierung und Simulation sozio-technischer Systeme, die Mensch-Maschine-Interaktion sowie autonome technische Systeme mit Anwendungen in den Gebieten Luft- und Raumfahrt, Straßenverkehr, Energiesysteme und in der Chemieindustrie.

Dr. Klaus-Hendrik Wolf Peter L. Reichertz Institut für Medizinische Informatik der TU Braunschweig und der Medizinischen Hochschule Hannover (PLRI).

Klaus-Hendrik Wolf ist wissenschaftlicher Mitarbeiter am PLRI. Nach dem Studium der Medizinischen Informatik an der Universität Hildesheim und der TU Braunschweig war er seit 2000 im PLRI zunächst an der TU Braunschweig und seit 2018 an der Medizinischen Hochschule Hannover tätig. Er war Chair der Working Group Wearable Sensors in Healthcare der International Medical Informatics Association. Zu seinen Forschungsschwerpunkten zählen Assistierende Gesundheitstechnologien und Virtuelle Medizin.

Dr. Dominik Wolff Peter L. Reichertz Institut für Medizinische Informatik der TU Braunschweig und der Medizinischen Hochschule Hannover (PLRI).

Dominik Wolff ist Nachwuchsgruppenleiter im PLRI. Nach dem Bachelorabschluss in Bioinformatik und Masterabschluss in Naturwissenschaftlicher Informatik an der Universität Bielefeld wechselte er 2016 an das Peter L. Reichertz Institut für Medizinische Informatik. Seine Forschungsschwerpunkte liegen auf Anwendungen künstlicher Intelligenz und datenwissenschaftlicher Methoden in der Biomedizin sowie der Mensch-Maschine-Interaktion und der Nutzung impliziten Wissens in Tutorialsystemen zur Patientenaufklärung.

Über die Aufgaben von Gelehrtengesellschaften: Die SYnENZ-Kommission in der Braunschweigischen Wissenschaftlichen Gesellschaft

Reinhold Haux

Zusammenfassung

Das 2. SYnENZ Symposium und die Arbeit der SYnENZ-Kommission sind in mehrfacher Weise charakteristisch für die Arbeit der Braunschweigischen Wissenschaftlichen Gesellschaft. Diese Charakteristika sollen hier kurz beschreiben und erläutert werden. Zu ihnen gehören die Pflege des fächerübergreifenden Dialogs zu Themen von hoher gesellschaftlicher Bedeutung, gemeinsames transdisziplinäres Arbeiten, die Förderung der Wissenschaften und ihrer Zusammenarbeit, die Kooperation mit anderen Wissenschafts- und Bildungsinstitutionen und die öffentliche Teilhabe an Forschung und Entwicklung. Dies mit dem Ziel, zu der Bildung einer wissensorientierten Gesellschaft beizutragen.

Schlüsselwörter

Gelehrtengesellschaften • Braunschweigische Wissenschaftliche Gesellschaft • SYnENZ-Kommission

R. Haux (✉)
Peter L. Reichertz Institut für Medizinische Informatik der TU Braunschweig, Braunschweig, Deutschland
E-Mail: reinhold.haux@plri.de

O. Richter et al. (Hrsg.), *Zusammenwirken von natürlicher und künstlicher Intelligenz: Beurteilen-Messen-Bewerten,*
https://doi.org/10.1007/978-3-658-45845-4_1

1 Einleitung

Das vorliegende Buch enthält schriftliche Ausarbeitungen von Vorträgen, die am 15. und 16. Februar 2023 in Braunschweig auf dem 2. SYnENZ Symposium gehalten wurden. Veranstalter des Symposiums waren die Braunschweigische Wissenschaftliche Gesellschaft (BWG 2023) und die Technische Universität Braunschweig. Veranstaltungsorte waren das Haus der Wissenschaft Braunschweig sowie, für die öffentliche Abendveranstaltung, die Dornse im Altstadtrathaus Braunschweig.

Dieses Symposium über das Zusammenwirken von natürlicher und künstlicher Intelligenz wurde von der SYnENZ-Kommission der BWG organisiert (BWG-Kommission Synergie und Intelligenz 2023). Es ist in mehrfacher Weise charakteristisch für die Arbeit der BWG.

Diese Charakteristika sollen hier kurz beschreiben und erläutert werden. Hierzu wird zunächst über die Arbeit der Braunschweigischen Wissenschaftliche Gesellschaft informiert. Anschließend wird erläutert, wie sich die Arbeit der SYnENZ-Kommission in die Arbeiten der BWG einfügt. Zudem muss das schon mehrfach verwendete Kürzel SYnENZ entschlüsselt werden.

2 Die Zielsetzung der BWG

Das Land Niedersachsen hat zwei miteinander kooperierende Gelehrtengesellschaften, die jeweils den Status einer Körperschaft des öffentlichen Rechts des Landes Niedersachsen haben und die über das sogenannte Selbstergänzungsrecht verfügen: die Niedersächsische Akademie der Wissenschaften zu Göttingen (2023) – diese 1751 gegründete Gelehrtengesellschaft ist vor allem geistes- und naturwissenschaftlich ausgerichtet – und die Braunschweigische Wissenschaftliche Gesellschaft – diese ist in erheblichem Maße, aber bei weitem nicht nur, technisch ausgerichtet. Die im Vergleich zur Göttinger Akademie jüngere BWG wird in diesem Jahr 80 Jahre alt. Sie ist in die drei Klassen Geisteswissenschaften, Ingenieurwissenschaften sowie Mathematik und Naturwissenschaften untergliedert. In diese Klassen können insgesamt maximal 100 ordentliche Mitglieder unter 70 Jahren über Zuwahlverfahren berufen werden.

Zu den Zielen der BWG steht in der Präambel ihrer Satzung:

> *„Die Braunschweigische Wissenschaftliche Gesellschaft (BWG) ist eine Vereinigung von Gelehrten. Sie hat zum Ziel, sich forschend, fördernd und vermittelnd mit den gesamtgesellschaftlichen Leistungen von Wissenschaft und Technik in einem steten*

interdisziplinären Diskurs auseinanderzusetzen. So trägt sie zur Bildung einer wissensorientierten Gesellschaft bei. Dabei sind die Technikwissenschaften sowohl mit den Naturwissenschaften und der Mathematik als auch mit den Geistes- und Sozialwissenschaften transdisziplinär verbunden. Das integrative Zusammenwirken ermöglicht die Transformation von akademischem zu beratungsorientiertem Wissen. Die Arbeit der BWG ist zielorientiert und wertebasiert. Ihre Mitglieder pflegen den fächerübergreifenden Dialog." … (Satzung der Braunschweigischen Wissenschaftlichen Gesellschaft, http://bwg-nds.de/über-die-bwg/satzung/).

Und in § 1 steht:

„*Die Braunschweigische Wissenschaftliche Gesellschaft dient der Förderung der Wissenschaften und ihrer Zusammenarbeit. Sie kooperiert mit anderen Wissenschafts- und Bildungsinstitutionen und unterstützt die öffentliche Teilhabe an Forschung und Entwicklung.*" … (Satzung der Braunschweigischen Wissenschaftlichen Gesellschaft, http://bwg-nds.de/über-die-bwg/satzung/).

3 Die SYnENZ-Kommission der BWG

Die Arbeit der BWG findet unter anderem in Kommissionen und Querschnittsbereichen statt, darunter in der von Professor Jochen Steil geleiteten SYnENZ-Kommission (BWG-Kommission Synergie und Intelligenz 2023; Steil 2022). Ihr vollständiger Name lautet: „BWG-Kommission Synergie und Intelligenz: technische, ethische und rechtliche Herausforderungen des Zusammenwirkens lebender und nicht lebender Entitäten im Zeitalter der Digitalisierung (SYnENZ)" (BWG-Kommission Synergie und Intelligenz 2023).

Die SYnENZ-Kommission wurde 2017 auf Initiative meines Vorgängers im Präsidentenamt, Professor Otto Richter, gegründet. In ihren Arbeiten befasst sie sich mit dem Zusammenwirken von natürlicher und künstlicher Intelligenz, in diesem 2. Symposium besonders mit den Fragen des Beurteilens, Messens und Bewertens von erweitertem Zusammenwirken.

Die Kommission hat aktuell – zum Zeitpunkt des Symposiums – 21 Mitglieder (BWG-Kommission Synergie und Intelligenz 2023). Es sind Kolleginnen und Kollegen, die vor allem an der TU Braunschweig, der Leibniz Universität Hannover, der Medizinischen Hochschule Hannover, der Otto-von-Guericke-Universität Magdeburg und an dem Deutschen Zentrum für Luft und Raumfahrt tätig sind. Zur Hälfte (11 von 21) sind sie BWG-Mitglieder. Vertreten werden u. a. die Fächer Ethik, Human Factors, Informatik, Intelligente Systeme, Künstliche Intelligenz, Medizin, Ökologie, Philosophie, Recht, Robotik, Technikgeschichte, Theologie und Verkehr.

4 Characteristika der BWG-Arbeit am Beispiel von SYnENZ

Warum sind das SYnENZ-Symposium und die SYnENZ-Kommission in mehrfacher Weise charakteristisch für die Arbeit der BWG? Die kursiv gesetzten Texte stammen aus der Satzung der BWG (Steil 2022) und wurden in Abschn. 2 zitiert:

- In der SYnENZ-Kommission wird zu einer Thematik von hoher gesellschaftlicher Bedeutung fächerübergreifend zusammengearbeitet – wir „*pflegen den fächerübergreifenden Dialog*".
- Alle, die der SYnENZ-Kommission angehören, schätzen diesen gegenseitigen Wissensaustausch, das fächerübergreifende Kennenlernen von Forschung. Nach sechs Jahren Kommissionsarbeit kann, so denke ich, auch gesagt werden, dass diese Zusammenarbeit nicht nur inter- und multidisziplinär, sondern wirklich transdisziplinar ist – die Mitglieder sind „*transdisziplinär verbunden*". Es geht ja um technische, ethische und rechtliche Herausforderungen des Zusammenwirkens. Gegenseitiges Lernen und Verstehen befördert die Arbeiten in den jeweils eigenen Disziplinen.
- Es finden Kolleginnen und Kollegen aus verschiedenen Hochschulen und Forschungseinrichtungen in der SYnENZ-Kommission der BWG eine passende ‚Heimat' für ihr gemeinsames Arbeiten – „*Förderung der Wissenschaften und ihrer Zusammenarbeit*".
- Die SYnENZ-Kommission arbeitet nicht nur intern. Über Tagungen wie diesem SYnENZ-Symposium, das gemeinsam mit der TU Braunschweig veranstaltet wird – Kooperation „*mit anderen Wissenschafts- und Bildungsinstitutionen*" – findet zum einen ein Austausch mit weiteren Wissenschaftlerinnen und Wissenschaftlern statt.
- Und in der öffentlichen Abendveranstaltung ging es um die – „*öffentliche Teilhabe an Forschung und Entwicklung*". Diese Vorträge über aktuelle Forschung durch fachlich hervorragend ausgewiesene Personen in der Öffentlichkeit sind ein weiterer, wichtiger Beitrag, der vielleicht besonders gut über Gelehrtengesellschaften wie der BWG geleistet werden kann.

Und so ist BWG zum einen das Kürzel für Braunschweigische Wissenschaftliche Gesellschaft. Zum anderen steht BWG aber auch für eines ihrer Ziele: der ***B***ildung einer ***w***issensorientierten ***G***esellschaft (Steil 2022).

5 Zum Schluss

Als Präsident der BWG ist es mir ein Anliegen, allen zu danken, die sich bei dem 2. SYnENZ Symposium wie auch bei dem daraus entstandenen Buch engagiert haben. Dazu gehören die Vortragenden und Moderatoren des Symposiums sowie die Autorinnen und Autoren dieses Buches. Mein Dank geht auch an die an der Organisation Beteiligten – aus der Geschäftsstelle der BWG und aus dem Institut für Robotik und Prozessinformatik der TU Braunschweig.

Dass auch die öffentliche Abendveranstaltung des 2. SYnENZ-Symposiums wieder im Altstadtrathaus und dort in der Dornse – einem gleichermaßen schönen, wie auch geschichtsträchtigem Raum – stattfinden konnte, ist keinesfalls selbstverständlich. Hier geht mein Dank an die Stadt Braunschweig.

Zwei Personen möchte ich bei dieser Danksagung auch namentlich nennen: Professor Jochen Steil, den Sprecher der SYnENZ-Kommission, und Professor Otto Richter, der bei der Herausgabe des Buches die Federführung übernommen hatte.

Literatur

Braunschweigische Wissenschaftliche Gesellschaft. http://bwg-nds.de. Zuletzt zugegriffen am 25.2.2023.

BWG-Kommission Synergie und Intelligenz: technische, ethische und rechtliche Herausforderungen des Zusammenwirkens lebender und nicht lebender Entitäten im Zeitalter der Digitalisierung (SYnENZ). http://bwg-nds.de/kommissionen/kommission-synenz/ und https://synenz.de/Start. Zuletzt zugegriffen am 25.2.2023.

Niedersächsische Akademie der Wissenschaften zu Göttingen. https://adw-goe.de/startseite/. Zuletzt zugegriffen am 25.2.2023.

Satzung der Braunschweigischen Wissenschaftlichen Gesellschaft. http://bwg-nds.de/über-die-bwg/satzung/. Zuletzt zugegriffen am

Steil J. BWG-Kommission Synergie und Intelligenz: technische, ethische und rechtliche Herausforderungen des Zusammenwirkens lebender und nicht lebender Entitäten im Zeitalter der Digitalisierung (SYnENZ). In: Braunschweigische Wissenschaftliche Gesellschaft. Jahrbuch 2021, 154–160. Göttingen: Cuvillier; 2022. http://bwg-nds.de/veröffentlichungen-jahrbuch-und-abhandlungen/. Zuletzt zugegriffen am 25.2.2023.

Prof. Dr. Reinhold Haux Peter L. Reichertz Institut für Medizinische Informatik der TU Braunschweig und der Medizinischen Hochschule Hannover (PLRI). Reinhold Haux ist Präsident der Braunschweigischen Wissenschaftlichen Gesellschaft (BWG) und emeritierter Professor für Medizinische Informatik am Peter L. Reichertz Institut für Medizinische Informatik der TU Braunschweig und der Medizinischen Hochschule Hannover (MHH).

Nach Professuren an Universitäten in Tübingen (1987–1989), Heidelberg (1989–2001) und Innsbruck (2001–2004) folgte er 2004 einem Ruf an die Technische Universität Braunschweig. Er war Präsident der International Medical Informatics Association (2007–2010), der International Academy of Health Sciences Informatics (2018–2020) und Herausgeber der Zeitschrift Methods of Information in Medicine (2001–2015). Er ist Honorarprofessor an der Universität Heidelberg und kooptiertes Mitglied des Lehrkörpers der MHH. Seit ihrer Gründung 2017 ist er Mitglied der SYnENZ-Kommission der BWG. Weitere Informationen auf www.plri.de.

Soziologische und ethische Aspekte

Vermenschlichung von Technik?

Soziologische Blicke auf das Zusammenspiel von Mensch und autonomer Technik in der Echtzeitgesellschaft

Johannes Weyer

Zusammenfassung

Thema des Beitrags sind die Grenzverschiebungen im Verhältnis von Mensch und Technik, die sich aus der Entwicklung technischer Systeme ergeben, welche mit künstlicher Intelligenz ausgestattet sind. In soziologischer Perspektive stellt der Beitrag die Frage, wie die Interaktion von Menschen und autonomen Systemen, z. B. autonomen Fahrzeugen, gelingen kann, wenn diese in alltäglichen Situationen aufeinandertreffen. Denn sie müssen sich verständigen und abstimmen, um gemeinsam Problemlösungen zu entwickeln, etwa im Fall des Überquerens einer Straße. Als eine mögliche Lösung für derartige Situationen wird das Konzept des virtuellen Blickkontakts entwickelt. Wie autonome Systeme in Zukunft mit dem Problem der Regelverletzung umgehen, also des Umgangs mit Konflikten, die durch sich widersprechende Regeln entstehen, bleibt hingegen eine offene Frage. Das Fazit lautet daher: Damit intelligente Technik sich in alltäglichen Situationen mit anderen Menschen bzw. Maschinen erfolgreich verständigen kann, wird man nicht umhinkommen, sie mit menschlichen Eigenschaften auszustatten.

Schlüsselwörter

Autonome Technik • Soziale Interaktion • Mensch-Maschine-Interaktion • Techniksoziologie • Echtzeitgesellschaft

J. Weyer (✉)
Fakultät Sozialwissenschaften der TU Dortmund, Dortmund, Deutschland
E-Mail: johannes.weyer@tu-dortmund.de

O. Richter et al. (Hrsg.), *Zusammenwirken von natürlicher und künstlicher Intelligenz: Beurteilen-Messen-Bewerten,*
https://doi.org/10.1007/978-3-658-45845-4_2

1 Einleitung

Mit der künstlichen Intelligenz kündigen sich fundamentale Veränderungen im Verhältnis von Mensch und Technik an, tritt doch erstmals eine Technik auf den Plan, die menschliche Eigenschaften hat (bzw. diese täuschend ähnlich darstellt). Die Soziologie und insbesondere die Techniksoziologie tut sich bislang schwer, diese Veränderungen konzeptionell zu erfassen und den neuen Mitspieler systematisch in ihre soziologischen Konzepte und Theorien einzubauen. Denn diese sind bislang überwiegend von der Idee zwischen-*menschlicher* Interaktion geprägt. Selbst provokative Konzepte wie das der Actor Network Theory, die alles Menschliche und Nicht-Menschliche radikal gleichsetzen, haben bislang nichts an der Tatsache geändert, dass es an einer genuin soziologischen Theorie der Interaktion von Mensch und intelligenter Technik bislang mangelt – insbesondere wenn man mit ihr den Anspruch verbindet, dass sie empiriefähig sein sollte, also die Option beinhalten sollte, theoretische Annahmen mit Methoden der empirischen Sozialforschung zu überprüfen.

Die folgende Abhandlung wirft zunächst einen Blick auf die Debatten der Techniksoziologie (Abschn. 2), um sich dann der Echtzeitgesellschaft (Abschn. 3) und den mit ihr einhergehenden Grenzverschiebungen im Verhältnis von Mensch und Technik zu befassen (Abschn. 4). Der zentrale Abschn. 5 versucht, Konzepte der Mensch-*Mensch*-Interaktion auf die Mensch-*Maschine*-Interaktion zu übertragen, und fragt danach, welche Voraussetzungen – auf beiden Seiten – gegeben sein müssen, damit eine derartige Interaktion gelingt. Die gilt insbesondere für den Fall der Regelverletzung durch autonome Systeme, also des Umgangs mit Konflikten, die durch sich widersprechende Regeln entstehen (Abschn. 6). Die zentrale These des Beitrags lautet: Damit intelligente Technik sich in alltäglichen Situationen mit anderen Menschen bzw. Maschinen erfolgreich verständigen kann, wird man nicht umhinkommen, sie mit menschlichen Eigenschaften auszustatten. Das abschließende Fazit (Abschn. 7) resümiert die Konsequenzen, die sich aus dieser Vermenschlichung von Technik ergeben.[1]

[1] Der vorliegende Text stellt eine Weiterentwicklung von Gedanken dar, die ich erstmals als Arbeitspapier (2022b) und später in der Druckfassung (2023) formuliert habe. Das meiste ist neu; lediglich Teile der Abschn. 4 und 6 sind ähnlich und enthalten keine substanziell neuen Argumente.

2 Zusammenwirken von Mensch und Technik

Die Techniksoziologie befasst sich seit jeher mit dem Zusammenspiel von Mensch und Technik, beispielsweise im Haushalt, am Arbeitsplatz oder im Straßenverkehr. Ob wir Wäsche waschen, eine CNC-Maschine bedienen oder mit dem Fahrrad in der Stadt unterwegs sind – in all diesen Situationen wirken Mensch und Technik bei der Lösung eines Problems zusammen. In der Techniksoziologie gibt es im Wesentlichen drei Perspektiven, unter denen das Verhältnis von Mensch und Technik betrachtet wird (vgl. ausführlich Weyer 2008):

- Die instrumentell-konstruktive Perspektive sieht Technik als Werkzeug des Menschen, das hergestellt wird, um einen Ursache-Wirkungszusammenhang zu vereinfachen und universell verfügbar zu machen – nach dem Motto: Schalter ein, Licht an.
- Die instrumentell-operative Perspektive interessiert sich für die Nutzung von Technik durch Menschen, die nicht über Konstruktionswissen verfügen, und die damit einhergehenden Herausforderungen bei der Gestaltung der Mensch-Maschine-Interaktion, z. B. in hochautomatisierten Pkw oder Flugzeug.
- Die diskursive Perspektive schließlich betrachtet die gesellschaftlichen Debatten über Technik, beispielsweise anlässlich des Versagens von Technik (Stichwort: Fukushima) oder des Aufkommens neuer technischer Optionen (Stichwort: Elektromobilität).

Diesen drei Perspektiven ist gemeinsam, dass sie Technik als Instrument betrachten, welches die vom Menschen definierten Handlungsprogramme willfährig und ohne eigenes Zutun ausführt. In Anbetracht der rasanten Verbreitung smarter Technik, die mehr kann, als nur vorgefertigte Routinen abzuspulen, hat sich in letzter Zeit ein vierter Debattenstrang entwickelt, in dem es um die Frage der „Agency“ – deutsch: Handlungsträgerschaft – von Technik geht (Rammert und Schulz-Schaeffer 2002). Gegenstand sind Konzepte, die beschreiben und erklären, wie die Interaktion von Mensch und Technik funktioniert, wenn auch die Technik Entscheidungen trifft, wie sie zuvor ausschließlich dem Menschen vorbehalten waren, z. B. ein Auto abzubremsen (im Fall des Notbremsassistenten), ohne einen entsprechenden Befehl des Menschen abzuwarten.

2.1 Zwischen radikalem Humanismus und radikalem Posthumanismus

Die sozialwissenschaftliche Technikforschung hat auf diese Herausforderung bislang unterschiedlich reagiert. Dabei prägen zwei diametral entgegengesetzte Positionen die Debatte:

Der radikale Humanismus zieht eine klare Trennlinie zwischen dem Menschen und nicht-menschlichen Wesen wie Tieren, Dingen, Objekten, Artefakten, Natur und auch Technik. Nur der Mensch – so der Philosoph Dieter Sturma (2001) – sei ein moralisch selbstbestimmtes Wesen, das sich externen, fremdbestimmten Zwecksetzungen widersetzen könne. Auch die Soziologin Sherry Turkle sieht im Menschen etwas Einzigartiges, das sich nicht per Softwarecode reproduzieren lasse: „The human is uncodeable." (Turkle 2005, S. 283) In beiden Fällen werden Ontologien bemüht, um Mensch und Technik als grundsätzlich unterschiedliche Wesen einzuordnen.

Der radikale Posthumanismus, wie ihn beispielsweise der französische Soziologie Bruno Latour vertritt, fordert hingegen vehement eine symmetrische Sichtweise, die die Wirklichkeit nicht einseitig aus der Perspektive des Menschen betrachtet, sondern andere, nicht-menschliche Wesen als gleichberechtigte Partner akzeptiert (Latour 1998). Auch die Natur, auch die Dinge, auch die Technik – so Latour und seine Mitstreiter:innen – wirken mit, sind aktiv, sind Teil eines Ensembles, das erst im Zusammenspiel seine Wirkungen zeigt – etwa im Fall des „schlafenden Polizisten", der die Autofahrer:innen dazu bringt, die Geschwindigkeit zu drosseln. Bei genauerer Betrachtung war dies eigentlich schon immer der Fall: Kein Mensch kann ohne ein Telefon telefonieren, und das Telefon kann es auch nicht ohne den Menschen. Ob sich daraus jedoch die starke Behauptung einer Gleichsetzung von Mensch und Technik ableiten lässt, mag hier zunächst offen bleiben.

Beiden Sichtweisen ist gemeinsam, dass sie – teils mit großer Emphase – über das „Ob" debattieren, also über die Grundsatzfrage, *ob* Mensch und Technik als gleichberechtigte Entitäten einzustufen sind. Damit ist m. E. jedoch wenig für das Verständnis des „Wie" gewonnen, also der Frage, *wie* Mensch und Technik in konkreten Situationen bei der Lösung eines Problems zusammenwirken, beispielsweise bei der Steuerung eines Autos oder eines Flugzeugs. *Ob* der Autopilot im Flugzeug menschliche Qualitäten hat oder nicht, hilft bei der Beantwortung der Frage kaum weiter, *wie* Mensch und Technik gemeinsam das Problem lösen, ein Flugzeug in kritischen Situationen sicher zu manövrieren.

2.2 Das Modell soziologischer Erklärung hybrider Systeme

Diesem Thema des Zusammenwirkens von Mensch und smarter Technik widmet sich das „Modell soziologischer Erklärung hybrider Systeme“ (HMSE), das Robin Fink und ich entwickelt haben (2011). Es setzt bei dem Kernproblem der Soziologie an, das „handelnde Zusammenwirken“ (Schimank 2010) der Menschen zu beschreiben. Es greift dabei auf Ideen von James Coleman (1995) und Hartmut Esser (1993) zurück, die Dynamik sozialer Systeme (also den Pfeil von Struktur t_1 zu t_2 in Abb. 1) aus dem Zusammenspiel von Struktur (Makro-Ebene) und Akteur (Mikro-Ebene) zu erklären: Die Akteure handeln im Rahmen struktureller Bedingungen zum Zeitpunkt t_1 und treffen individuelle, zumeist begrenzt rationale Entscheidungen, die per Interaktion zu aggregierten, oftmals nicht-intendierten Effekten auf der Systemebene zum Zeitpunkt t_2 führen. Die Kernbotschaft des Modells lautet: Eine soziologische Erklärung struktureller Dynamiken muss immer den „Umweg“ über das Handeln der Menschen gehen.

In diesem Modell ist jedoch bislang kein Platz für Technik, und die soziologische Theorie zeichnet sich dadurch aus, dass sie noch keinen Versuch unternommen hat, der Technik einen systematischen Stellenwert in ihren Theoriegebäuden einzuräumen – sei es in der Handlungstheorie, der Theorie sozialer Systeme oder der Steuerungstheorie (vgl. meine kritische Auseinandersetzung mit Armin Nassehis „Muster“ in Weyer 2022a).

Das HMSE versteht sich als ein Ansatz, diese Lücke zu füllen und das Mitwirken der Technik systematisch in die soziologische Theorie einzubauen. Das HMSE erweitert das Modell soziologischer Erklärung, indem es Technik als einen

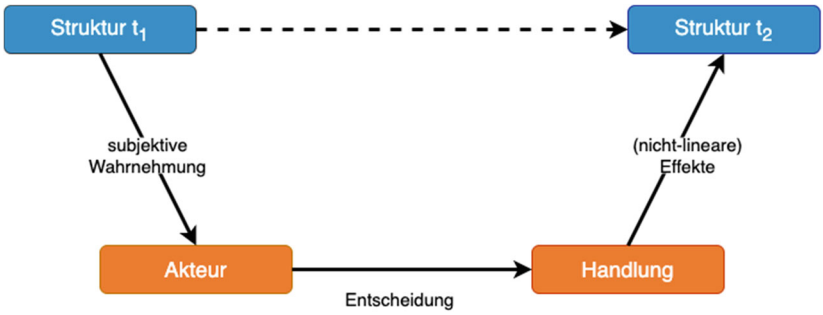

Abb. 1 Das Modell soziologischer Erklärung (in Anlehnung an Esser 1993)

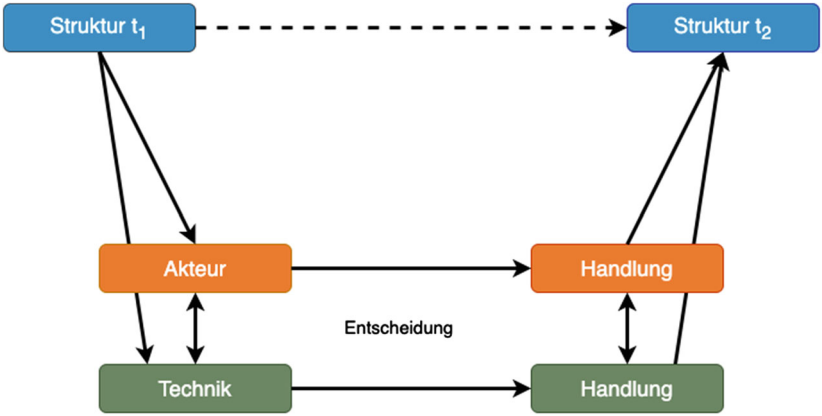

Abb. 2 Das Modell soziologischer Erklärung hybrider Systeme (Fink und Weyer 2011)

Mitspieler begreift, der – ähnlich wie der menschliche Akteur und im Rahmen situationaler Constraints – ebenfalls Ziele verfolgt und durch eigene Handlungen (horizontaler Pfeil), aber auch durch Zusammenarbeit mit dem menschlichen Akteur (vertikaler Doppelpfeil) zum Ergebnis beiträgt (vgl. Abb. 2).

Neben dem theoretischen Konzept beinhaltet das HMSE auch eine Methode, die dazu beitragen könnte, die Debatte zwischen Humanismus und Posthumanismus durch empirische Forschung zu klären, und zwar mithilfe von Simulatorexperimenten zur hybriden Mensch-Maschine-Interaktion und zur Zuschreibung von Handlungsträgerschaft. Leider ist dieser Ansatz weder von der Techniksoziologie noch von der soziologischen Theorie aufgegriffen worden, was insofern unverständlich ist, als smarte Technik in zunehmendem Maße an der Ausführung von Handlungen in vielen gesellschaftlichen Bereichen mitwirkt.

Insbesondere die soziologische Theorie hat auf die Herausforderung, die mit der Digitalisierung des Alltags einhergeht, bislang nur unzureichend reagiert und keine entsprechenden Konzepte entwickelt, obwohl die Waschmaschine des Jahres 2023 mehr tut als die des Jahres 1950. Letztere hat zwar auch eine Leistung erbracht, die zuvor der Mensch erbracht hat (und ihn damit ersetzt); sie hat aber keine eigenen Entscheidungen getroffen. Die Waschmaschine des Jahres 2023 tut deutlich mehr: Sie prüft, welche Sorte Wäsche sich in der Maschine befindet, und regelt dementsprechend das Waschprogramm, sie kommuniziert mit dem Stromanbieter und handelt Tarife aus, sie informiert den Wartungsdienst, bevor ein Teil ausfällt, usw. Es ist nach wie vor eine große Leerstelle der soziologischen

Theorie, dass sie diese Phänomene nicht mit genuin eigenständigen Konzepten erfassen kann.

3 Die Echtzeitgesellschaft

Statt soziologische Konzepte eines Zusammenwirkens von Menschen mit smarter, digitaler Technik zu entwickeln, spricht man daher zumeist von der digitalen Gesellschaft. Dieser Begriff ist jedoch m. E. wenig hilfreich, denn er hat weder eine analytische Qualität noch eine soziologische Substanz.

3.1 Digitalisierung – (k)eine soziologische Kategorie?

Wenn Soziolog:innen versuchen zu verstehen, was eine neue Technik mit der Gesellschaft macht, dann begeben sie sich auf die Suche nach sozialen Veränderungen bzw. den sozialen Korrelaten, die mit fundamental neuen Technologien einhergehen (Popitz 1995). Niemand käme auf die Idee, die Gesellschaft des 18. Jahrhunderts, die sich mit der neuen Technik der Dampfmaschine herausbildete, als Dampfgesellschaft zu bezeichnen. Man spricht vielmehr von der – nationalstaatlich verfassten – Industriegesellschaft.

Auch wäre es kaum vorstellbar, die Gesellschaft des späten 20. Jahrhunderts, die auf Computer und Internet gründet, als Computergesellschaft zu bezeichnen. Es handelt sich vielmehr um die Wissensgesellschaft – nunmehr im globalen Maßstab. Soziolog:innen vermeiden es, soziale und technische Entwicklung schlicht zu parallelisieren, sondern versuchen, die fundamentalen Veränderungen zu verstehen (und auf den Begriff zu bringen), die sich mit dem Vordringen einer neuen Technik und dem Entstehen einer neuen Epoche ergeben.

Die Verwendung des Begriffs „digitale Gesellschaft“ (Nassehi 2019) spricht für eine gewisse Ratlosigkeit der Soziologie, denn er hat keinen analytischen Tiefgang, sondern basiert auf einer allzu simplen Parallelisierung des technischen Prozesses und dessen sozialer Folgen. Heinrich Popitz (1995) war in den 1980er Jahren mit seinem Konzept der sozialen Korrelate bereits ein Stück weiter, als er beispielsweise die Angleichung des Lebensstandards als die zentrale soziale Folgewirkung der Elektrifizierung beschrieb.

3.2 Echtzeitsteuerung

Es bietet sich an, die Echtzeitsteuerung komplexer soziotechnischer Systeme als einen möglichen Kandidaten für das soziale Korrelat der Digitalisierung in Erwägung zu ziehen. Denn die komplexen Infrastruktursysteme der Zukunft, etwa in den Bereichen Verkehr oder Energie, werden auf Basis digitaler Echtzeitdaten operieren, die von smarten Geräten aufgezeichnet und übermittelt werden und die eine Echtzeitsteuerung dieser Systeme möglich machen: Eine zentrale Leitstelle erstellt mithilfe dieser Daten permanent aktuelle Lagebilder (z. B. über die Stausituation auf Straßen), um diese Information mit nur geringer Zeitverzögerungen wiederum an die einzelnen Komponenten zurückspielen, die ihre Entscheidungen entsprechend anpassen können (vgl. Abb. 3). Man kennt dies beispielsweise von der Routenplanung. Gestützt auf große Mengen von Echtzeitdaten wird es somit erstmals in der Geschichte der Menschheit möglich sein, komplexe Systeme wie das Verkehrs- oder das Energiesystem in Echtzeit zu steuern.

Echtzeitsteuerung beinhaltet einen Feedback-Mechanismus, der die Nachjustierung bzw. Ad-hoc-Umsteuerung laufender Prozesse auf Basis von Daten ermöglicht, die kurz zuvor aus diesen Prozessen gewonnen wurden. Man mag einwenden, dass dieses Grundprinzip der Steuerung komplexer Systeme schon im alten Ägypten galt; aber die Prozesse dauerten Monate, Jahre oder Jahrzehnte. In der Echtzeitgesellschaft spielen sich die Prozesse in wesentlich kürzeren Zeiträumen von Minuten, Stunden oder Tagen ab. Man denke an die Routenplanung, die Just-in-time-Produktion oder den On-demand-Verkehr. Und gerade diese „Beschleunigung“ (Rosa 2005) der Analyse von „Mustern“ (Nassehi 2019) macht einen qualitativen Unterschied in der Fähigkeit zum Management komplexer soziotechnischer Systeme.

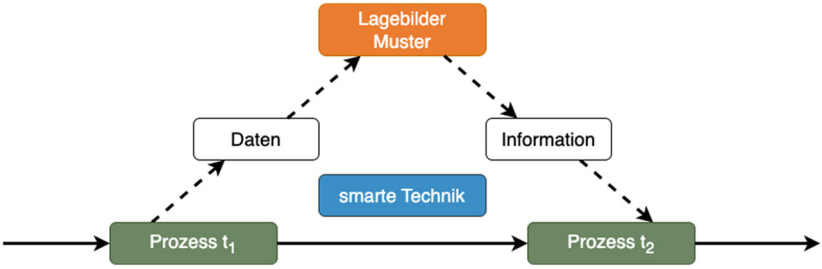

Abb. 3 Der Prozess der Echtzeitsteuerung

Da Echtzeitsteuerung nicht mit totalitärem Zwang einhergeht, sondern den „gesteuerten“ Subjekten erlaubt, autonome Entscheidungen – auch gegen die Empfehlungen – zu treffen, spricht man von einem neuartigen Koordinationsmodus der zentralen Steuerung dezentraler Systeme, der in gewisser Weise das Beste aus den beiden Welten „Markt“ und „Staat“ kombiniert (Rochlin 1997; Weyer et al. 2015; Weyer 2019b; Schrape 2021).

Versuche, ganze Volkswirtschaften mit einer Mischung aus Kybernetik und Planwirtschaft zu steuern, waren in Chile und in der DDR gescheitert (Lobe 2015); bei der Echtzeitsteuerung geht es nicht mehr um Volkswirtschaften, sondern um komplexe soziotechnische Systeme wie das Energie- oder das Verkehrssystem und deren nachhaltige Transformation. Ohne moderne Formen intelligenter Steuerung werden weder die Verkehrswende noch die Energiewende gelingen.

Statt von der digitalen Gesellschaft zu sprechen und damit lediglich die eigene Ratlosigkeit zu dokumentieren, wird hier die These vertreten, das die Echtzeitsteuerung komplexer soziotechnischer System ein möglicher Kandidat für das soziale Korrelat der Digitalisierung ist, weshalb im Folgenden von der Echtzeitgesellschaft gesprochen wird. Die Fähigkeit zur Echtzeitsteuerung ist eine genuine soziale Innovation, die ein Novum innerhalb der Geschichte der Menschheit bildet und „einen neuen Modus technischen Handelns“ (Popitz) geschaffen hat, wie er zuvor undenkbar war. Wenn die Soziologie diese Entwicklungen verstehen will, kommt sie nicht umhin, Modelle – wie etwa das HMSE – zu entwickeln, die diese neue Gesellschaft abbilden und dazu beitragen, die in ihr wirkenden Mechanismen zu verstehen.

3.3 Schichtmodell der Epochen der Technikgeschichte

Um einer möglichen Fehlinterpretation vorzubeugen, dass die Idee der Echtzeitgesellschaft eine Art Deutungshohheit auf Kosten anderen Gesellschaftsdiagnosen beansprucht, wird nochmals auf Popitz zurückgegriffen, der die „moderne technische Zivilisation … keineswegs (als) durchgehend modern“ ansieht. Sie gleiche „eher einem Warenhaus der Innovationen der Technikgeschichte“ (1995, S. 42).

Analog wird die Echtzeitgesellschaft nicht als eine Gesellschaftsformation verstanden, die andere Formationen wie die Industriegesellschaft verdrängt. Im Gegenteil: Sie setzt auf ihr auf, ist auch und ebenso industriell wie ihre Vorgängerinnen und fügt lediglich einen weiteren Baustein hinzu, nämlich komplexe Systeme mit Hilfe digitaler Technik in Echtzeit zu steuern (vgl. Abb. 4).

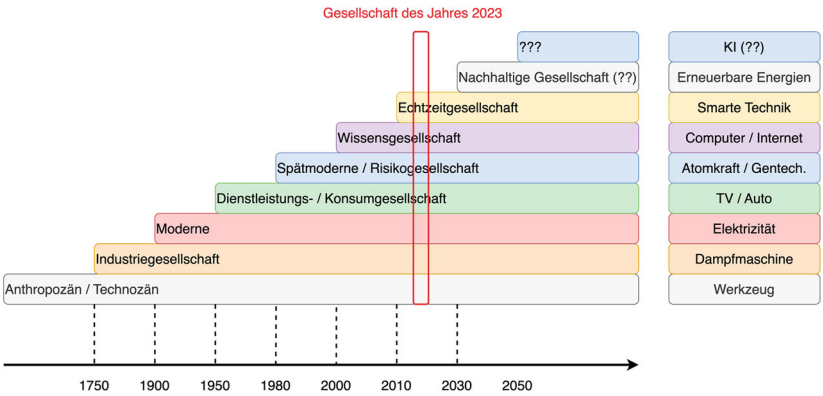

Abb. 4 Schichtmodell der Epochen der Technikgeschichte

Die Industriegesellschaft – mit der Dampfmaschine – setzt um 1750 auf dem Antropozän auf, dessen zentrales Merkmal die Herstellung und der Gebrauch von Werkzeugen war. Ab 1900 entwickelt sich mit der Elektrizität die Moderne, gefolgt von der Dienstleistungs- bzw. Konsumgesellschaft ab 1950, deren Schlüsseltechnologie die Massenmedien (TV etc.) sowie das Automobil waren. Nochmal: Die neu hinzugekommenen Schichten verdrängen die darunter liegenden nicht, sondern nutzen deren Errungenschaft etwa für die industrielle Produktion von Pkws oder Fernsehgeräten.

Mit Atomkraft und Gentechnik entsteht in den 1980er Jahren die Risikogesellschaft (Beck 1986), mit Computer und Internet ab der Jahrtausendwende die globale Wissenschaftsgesellschaft, die selbstverständlich Industrie- und Risikogesellschaft bleibt. Auf die Echtzeitgesellschaft, die seit 2010 Konturen gewinnt, könnte in einigen Jahren die nachhaltige Gesellschaft folgen und dann um das Jahr 2050 eine – noch namenlose – Gesellschaft, in der autonome Technik alltäglich geworden ist. Man kann die Chiffren Web 1.0 bis 3.0 oder Industrie 1.0 bis 4.0 in dieses Schema hineinprojizieren, gewinnt aber allein durch eine Nummerierung kaum Erkenntnisse mit soziologischem Gehalt.

Die gegenwärtige Gesellschaft des Jahres 2023 ist demzufolge nicht pure Echtzeitgesellschaft, sondern eine Mischung aus all dem, was zuvor existierte, und dem, was neu hinzugekommen ist. Auch smarte Technik wird industriell hergestellt, ist Teil von Dienstleistungen, birgt Risiken, basiert auf geteiltem Wissen und kann zu einer nachhaltigen Zukunft beitragen.

4 Grenzverschiebungen

Anders als in vorherigen Epochen agiert die smarte Technik der Echtzeitgesellschaft nicht mehr als mechanisches Instrument, das den vom Menschen gestellten Auftrag stoisch ausführt, sondern trifft in zunehmenden Maße Entscheidungen, wie sie zuvor nur der Mensch getroffen hat. Intelligente technische Systeme können abwägen, ob eine E-Mail spamverdächtig ist oder nicht, und sie können entscheiden, ob eine Verringerung der Geschwindigkeit des Autos erforderlich ist, um den Abstand zum vorausfahrenden Fahrzeug einzuhalten, wie etwa im Fall von Adaptive Cruise Control.

Dies wirft die Frage nach der künftigen Rollenverteilung und einem damit möglicherweise einhergehenden Kontrollverlust auf (Weyer 2019a). Schon in den vergangenen Jahrzehnten konnte man eine Grenzverschiebung derart beobachten, dass immer mehr praktische Handlungen, die ursprünglich von Menschen ausgeführt wurden, von technischen Systemen übernommen wurden, z. B. von Automaten wie der Waschmaschine (vgl. Abb. 5):

- *Maschinisierung:* In der Frühphase der Maschinisierung um 1800 war die Maschine ein Hilfsmittel für Arbeiten, die von Mensch und Maschine ausgeführt, aber ausschließlich vom Menschen gesteuert und überwacht wurden (z. B. im Fall der Dampfmaschine).

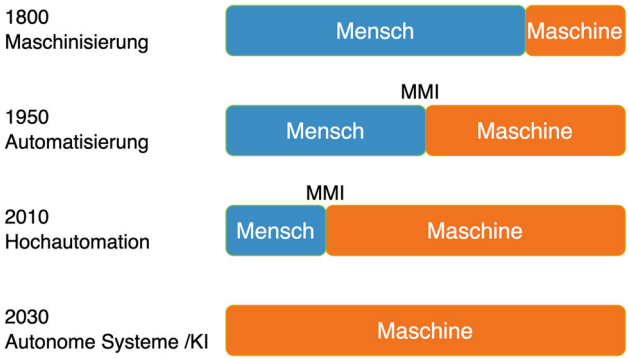

Abb. 5 Stadien der Automatisierung. (Eigene Darstellung)

- *Automatisierung:* Diese Grenze, an der eine Mensch-Maschine-Interaktion (MMI) stattfindet, verschiebt sich mit der Automatisierung Mitte des 20. Jahrhunderts in dem Maße, in dem die Maschine immer stärker auch an der Prozesssteuerung beteiligt ist (z. B. im Fall der Waschmaschine).
- *Hochautomation:* Autonome bzw. teilautonome Systeme wie der Spamfilter oder der Bremsassistent, die ab etwa 2010 flächendeckend zum Einsatz kommen, treffen sogar Entscheidungen, wie sie bislang ausschließlich dem Menschen vorbehalten waren. Der Mensch ist vor allem für die Überwachung des Systems zuständig (z. B. im Fall des Autopiloten im Flugzeug) und greift in das operative Geschehen immer seltener ein.
- *Autonome Systeme:* Die künstliche Intelligenz steht in dieser Tradition, verschiebt die Grenze jedoch noch ein Stück weiter; denn sie beinhaltet die Verheißung (vielleicht ab 2030), dass technische Systeme eines Tages ganz ohne menschliches Zutun operieren, der Mensch also die Kontrolle vollständig abgibt und als Mitspieler überflüssig wird.

In der künftigen Welt der autonomen Systeme und der künstlichen Intelligenz hat der Mensch – zumindest in seiner Funktion als Operateur eines komplexen soziotechnischen Systems – offenbar ausgedient, so dass sich auch jegliche Reflexion über das handelnde Zusammenwirken von Mensch und Technik zu erübrigen scheint. Schlechte Nachrichten für die Techniksoziologie?

Keinesfalls, denn schon in der Phase der Hochautomation gab es etliche Probleme, da der Mensch auf die Rolle des Lückenbüßers in einem komplexen System reduziert wurde, das immer wieder Fehlfunktionen aufwies, die sich nicht problemlos beheben ließen. Dies belegt beispielsweise der Fall der Boeing 737 MAX mit zwei katastrophalen Unglücken in den Jahren 2018 und 2019 (Weyer 2023). Die Lehre aus diesen Ereignissen lautet: Es ist nicht trivial, ein hochautomatisiertes bzw. intelligentes System im Zusammenspiel von Mensch und Technik zu steuern; und die Hoffnung, dass man den Menschen komplett durch Technik ersetzen kann, erweist sich als eine trügerische Illusion.

4.1 Maschinen als soziale Wesen mit menschlichen Eigenschaften?

Auch in Zukunft wird der soziologische Blick auf das verteilte Handeln in hybriden Systemen vonnöten sein, die von Mensch und Technik gemeinsam gesteuert werden. Selbst wenn es eines Tages gelingen sollte, die derzeit

noch offenen Fragen der Verlässlichkeit von KI zu lösen (Beyerer und Niggemann 2018; Konz et al. 2023), wird dies nicht dazu führen, dass jegliche Mensch-Maschine-Interaktion entfällt, wie es Abb. 5 (weiter oben) suggeriert.

Der Straßenverkehr der Zukunft, in dem nur noch autonome Fahrzeuge des SAE-Levels 5 unterwegs sind, soll hier als Beispiel dienen. In Zukunft werden Menschen zwar nicht mehr für die operative Steuerung der Fahrzeuge benötigt. Und dennoch wird eine Mensch-Maschine-Interaktion stattfinden; denn autonome Autos werden anderen Menschen, aber auch anderen Maschinen, in alltäglichen Kontexten begegnen, in denen eine Interaktion stattfinden wird (bzw. muss, vgl. Abb. 6):

- den Passagieren (interne Interaktion im Fahrzeug) sowie
- anderen Verkehrsteilnehmer:innen, z. B. Radfahrer:innen oder Fußgänger:innen (externe Interaktion im Straßenverkehr).

Damit verschärft sich Situation in gewisser Weise, weil die Probleme, die man an einer Stelle erfolgreich bewältigt zu haben glaubt, nun an anderer Stelle wieder auftauchen. Zudem benötigt die Maschine nunmehr menschliche Eigenschaften, um die neuartigen Herausforderungen der Interaktion mit Passagieren und Passanten bewältigen zu können. Damit ist nicht gemeint, dass die Technik ein

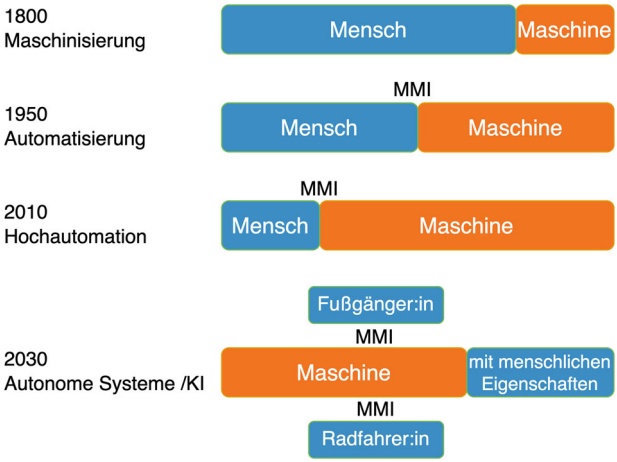

Abb. 6 Mensch-Maschine-Interaktion im Fall autonomer Systeme. (Eigene Darstellung)

humanoides Äußeres haben muss, sondern lediglich, dass sie fähig sein muss, Entscheidungen zu treffen und Handlungen durchzuführen, die denen der Menschen ähneln. Zudem muss sie in der Lage sein, sich bei diesen Entscheidungen mit anderen Menschen (und Maschinen) abzustimmen, um eine Verständigung – im Habermas'schen Sinne (1968) – zu erzielen. Zwei Beispiele mögen dies erläutern.

4.2 Interaktion von autonomen Fahrzeugen mit Passanten

Man stelle sich folgende Alltagssituation vor: Ein:e Fußgänger:in steht am Straßenrand an einem Fußgängerüberweg („Zebrastreifen") oder an einer Fußgängerinsel, und es nähert sich ein Auto mit Fahrer:in an Bord. Typischerweise wird eine Interaktion stattfinden, die durch Blickkontakt in Gang gesetzt. Danach folgt typischerweise ein Handzeichen („ich winke sie durch") oder ein Blinkzeichen („bitte schön") oder ein lautes Hupen („Platz da, ich komme"). Eine Situation wechselseitiger Unsicherheit wird durch Blickkontakt und nonverbale Interaktion gelöst. Dabei wird auf gewisse Konventionen, aber auch generalisierte Erwartungen zurückgegriffen wie etwa die Erwartungen, dass Autofahrer:innen typischerweise an einem Fußgängerüberweg anhalten und Fußgänger:innen typischerweise die Reaktion ihres Gegenübers abwarten sollten, bevor sie die Straße überqueren.

Noch ist unklar, wie sich diese alltägliche – und angesichts von zwei Teilnehmer:innen nicht sonderlich komplexe – Situation in Zukunft abspielen würde, wenn ein autonomes Auto daran beteiligt wäre. Mercedes-Benz hat im Rahmen der Arbeiten an seinem Versuchsfahrzeug Mercedes F015 zwei Vorschläge unterbreitet, wie das autonome Fahrzeug mit der Fußgänger:in interagieren könnte: zum einem durch Leuchtsymbole auf einem großen LED-Display an der Fahrzeugfront, die das Durchwinken mit der Hand imitieren, zum anderen durch einen Zebrastreifen, den das autonome Auto vor sich auf die Straße projiziert (vgl. Mercedes 2015).

Beide Varianten setzen voraus, dass zuvor eine Interaktion zwischen Fußgänger:in und Fahrzeug stattgefunden hat, und zwar in beide Richtungen: Die Fußgänger:in muss erkannt haben, dass sich erstens ein Fahrzeug nähert, dass sie zweitens diesem ihren Wunsch signalisieren muss, die Straße zu überqueren, und dass sie schließlich drittens dessen Reaktion abwarten muss, bevor sie es tut. Das Fahrzeug muss seinerseits nicht nur erkannt haben, dass es sich bei dem Objekt am Straßenrand um ein:e Fußgänger:in handelt, sondern auch dass

diese die Absicht hat, die Straße zu überqueren, aber bereit ist, ihre Entscheidung davon abhängig zu machen, ob ein Signal gesendet wird, dass die Straße gefahrlos passiert werden kann.

Ein alltäglicher Vorgang, den die meisten Menschen routinehaft und ohne große Vorüberlegungen meistern, muss also in recht aufwändige technische Prozeduren übersetzt werden, die das Problem der wechselseitigen Erwartungserwartungen bewältigen: Ego erwartet von Alter, dass dieser etwas von Ego erwartet, und richtet seine eigenen Aktionen daran aus (vgl. Weber 1985).

Die Vision des autonomen Fahrens ist von der Vorstellung geprägt, den Menschen möglichst vollständig aus dem Regelkreis herauszunehmen. Das Beispiel der Interaktion mit Passanten zeigt jedoch, dass dies nur partiell gelingen wird – und auch nur dann, wenn man die Technik vermenschlicht, ihr also menschliche Züge und Verhaltensweisen antrainiert. Damit führt zu einer paradoxen Situation. Der (autonom agierende) Mensch wird durch eine (autonom agierende) Maschine ersetzt, die aber ihre Funktionen nur erfüllen kann, wenn sie immer menschlicher wird – im Sinne der Fähigkeit zu sozialer Interaktion und Koordination sowie zu intelligentem, nicht vorhersehbarem Verhalten. Zudem wird der Mensch nicht vollständig verdrängt, sondern bleibt Teil des Spiels – allerdings mit einer anderen Rollenverteilung als zuvor.

4.3 Interaktion von autonomen Fahrzeugen mit Passagieren

Diese Vermenschlichung der Technik betrifft auch die Interaktion mit den Passagieren im Fahrzeug. Eigentlich könnte man denken, dass dies überflüssig ist, wenn man gefahren wird, also am Prozess der Steuerung des Fahrzeugs – ähnlich wie in Bus oder Bahn – nicht aktiv teilhat. Und dennoch findet eine intensive Interaktion statt, die beispielsweise der Safety Report der Google-Tochter Waymo dokumentiert (Waymo LLC 2020).

Waymo kann auf reichhaltige Erfahrungen mit autonomen Taxi-Fahrzeugen in den USA verweisen. Waymo's Fahrzeuge interagieren mit ihren Fahrgästen und erklären, wie sie die aktuelle Situation wahrnehmen und wie sie darauf reagieren werden (vgl. Abb. 7).

Diese Form der Transparenz schafft Vertrauen, indem sie dem Menschen die „Sichtweise" der Maschine zugänglich gemacht wird, und macht für die Passagiere nachvollziehbar, was gerade passiert und warum dies geschieht. Zudem haben die Fahrgäste vielfältige Möglichkeiten, mit dem Fahrzeug zu interagieren.

Abb. 7 Auszüge auf dem Waymo Safety Report

Waymo Safety Report 2020, S. 35

- „give passengers the *information* they need“
- „help passengers *anticipate* what's next“
- „*proactively communicate* the vehicle's response to events on the road“
- „information provided to passengers helps them know what to *expect*“
- „We also want our passengers to be aware of what the *vehicle is perceiving*, and *why* it is taking specific actions.“
- „help them *understand* what the vehicle and other road users are doing“

Auf diese Weise wird die Maschine menschlicher, denn sie agiert nicht wie ein stoisch-sturer Roboter, der sein Programm unbeirrt abspult, sondern sie tritt wie ein Mitmensch auf, der das, was er tut, erklärt und begründen kann (vgl. die Formulierung „why“ in Abb. 7). Philosophen haben immer wieder darauf verwiesen, dass das Argumentieren im „Raum der Gründe“, also die Fähigkeit, die eigenen Handlungen zu begründen und zu rechtfertigen, eine zutiefst menschliche Eigenschaft ist (Sturma 2001; Habermas 1981a).

5 Soziologie der Interaktion mit Maschinen

Die autonome Technik der Zukunft muss also nicht nur kommunizieren können, sondern sie muss auch in der Lage sein, sich mit anderen – menschlichen wie nicht-menschlichen – Teilnehmer:innen des Straßenverkehrs zu verständigen. Damit entstehen Konstellationen sozialer Interaktion, an denen technisch aufgerüstete Menschen und vermenschlichte technische Systeme einander begegnen. Derartige Begegnungen zu bewältigen und auftretende Probleme zu lösen, ist keineswegs trivial, sondern sehr voraussetzungsvoll. Dies wird zunächst am Beispiel der Mensch-*Mensch*-Interaktion entwickelt (Abschn. 5.1) und dann auf die Mensch-*Maschine*-Interaktion übertragen (Abschn. 5.2).

5.1 Soziale Interaktion (Mensch-Mensch)

Wenn zwei Menschen einander begegnen und erfolgreich miteinander interagieren, dann tun sie das auf Basis eines geteilten Sinnhorizonts (Esser 2000).

Dieser beinhaltet nicht nur ein gemeinsames Verständnis sprachlicher und nichtsprachlicher Symbole (z. B. Gesten wie „den Vogel zeigen“), sondern auch *Erwartungen,* was der andere tun wird bzw. sollte, *Regeln,* die dabei zu beachten sind, und *Normen,* die festlegen, was richtig und falsch bzw. angemessen und unangemessen ist (Bellon et al. 2022) (vgl. Abb. 8).

Wenn die Erwartungen von Ego und Alter wechselseitig aneinander anschließen, sich also Erwartungserwartungen herausbilden, kann sich ein derartiger Kommunikationszusammenhang zu einem sozialen System temporär stabilisieren (Luhmann 1984).

Ziel von Interaktion und der sie tragenden verbalen wie nonverbalen Kommunikation ist zumeist eine Verständigung, also eine Abstimmung von Erwartungen, mit dem Ziel der Problemlösung, wie etwa im Fall der oben geschilderten Überquerung eines Fußgängerüberwegs. Verständigung – in Habermas'scher Diktion: kommunkatives Handeln – ist in der Regel kein Selbstzweck, sondern Mittel zum Zweck, und zwar zur Erreichung eigener Ziele unter Einbeziehung des Handelns anderer Personen (bei Habermas: instrumentelles bzw. strategisches Handeln). Die künstliche Trennlinie, die Habermas zwischen diesen drei Formen

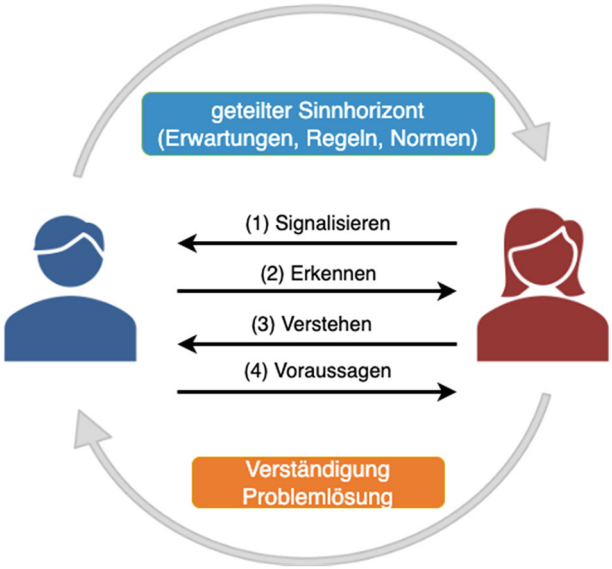

Abb. 8 Soziale Interaktion

von Kommunikation gezogen hat, ist kaum aufrechtzuerhalten; und es gibt einen versteckten Hinweis, dass er das auch so sieht (1981b, S. 194). Auch die Luhmann'sche Fiktion, dass Kommunikation nur dem Zweck der Aufrechterhaltung von Autopoiesis dient und nicht den strategischen Zielen der Beteiligten (z. B. die Straße sicher zu überqueren), ist vor diesem Hintergrund kaum nachvollziehbar.

5.1.1 Erkennen, Verstehen, Vorhersehen

Ob soziale Interaktion gelingt, hängt von drei Faktoren ab: dem Erkennen, dem Verstehen und dem Vorhersehen (Endsley und Kiris 1995). Wie im Beispiel weiter oben bereits angedeutet, muss die Fußgänger:in *erkennen* (2), also sinnlich wahrnehmen können, was in der aktuellen Situation passiert, sie muss *verstehen* (3), was dies bedeutet, also welche Intention ihr Gegenüber damit verfolgt, und sie muss *vorhersehen* (4) können, was daraus als nächstes folgt (vgl. Abb. 8). Umgekehrt gilt dies für die Autofahrer:in.

Das Ganze ist ein interaktiver Prozess, der spontan *oder* durch bewusst gesetzte Signale (1) in Gang kommen kann, mit denen Ego die Aufmerksamkeit von Alter auf sich lenkt. Denn eine gelingende Interaktion hängt nicht nur davon ab, dass Ego versteht, was Alter tut (und will), sondern auch, dass Ego von Alter verstanden wird (vgl. Drewitz et al. 2021). Signale können helfen, diesen Prozess der wechselseitigen Wahrnehmung und der darauf aufbauenden Verständigung in Gang zu setzen.

5.2 Soziale Interaktion (Mensch-Maschine)

Projiziert man das Basismodell sozialer Interaktion zwischen Menschen auf die Mensch-Maschine-Interaktion, so stellen sich folgende Fragen:

- Wie erkennt der Mensch, dass die Maschine etwas tut? Wie gelingt es ihm, zu verstehen, was sie tut, also ihre Aktionen als intentionales Handeln zu deuten? Ist es ihm möglich vorauszusagen, was als nächstes passiert? Und wie kann die Maschine den Menschen durch gezielte Signale beim Erkennen, Verstehen und Voraussagen unterstützen?
- Und umgekehrt: Wie erkennt die Maschine, dass der Mensch etwas tut? Wie gelangt sie zu einem Verständnis der Intentionen sowie zu einer Voraussage dessen, was im nächsten Moment passieren wird? Und kann der Mensch Signale setzen, die die Maschine versteht bzw. die das Verstehen fördern?

5.2.1 Wahrnehmung der Maschine durch den Menschen

Zum ersten Punkt hat die Forschung in den letzten Jahrzehnten eine Reihe von Erkenntnissen zusammengetragen. Aufgrund umfangreicher empirischer Studien kann mittlerweile als erwiesen gelten, dass wir Menschen der Technik – auch der autonomen – menschliche Eigenschaften zuschreiben. Man denke an den Spruch: „Mein Computer spinnt mal wieder." In der Interaktion mit Maschinen zeigen wir soziale Reaktionen (Reeves und Nass 1996; Turkle 2011, Hidalgo et al. 2021). Zudem schreiben wir der Technik Handlungsträgerschaft zu (Rammert und Schulz-Schaeffer 2002; Fink und Weyer 2011), und wir vertrauen selbst digitaler Technik in hohem Maße (Weyer und Cepera 2021).

Eine Vielzahl von Studien in den Feldern Human–Computer-Interaction und Human-Automation-Collaboration, etwa am Beispiel der Luftfahrt, hat zudem das Konzept des Team-Plays entwickelt, also die Idee, dass Mensch und Maschine sich als gleichberechtigte Mitspieler eines Teams betrachten, das gemeinsam Problemlösungen erarbeitet (Sarter und Woods 1997; Weyer 2016).

Es kann also als gesichert angesehen, dass Menschen in der Lage sind zu erkennen, was die Maschine tut, dies zu verstehen und auch vorauszusehen, was als nächstes passieren wird. Es gibt erste Studien, die sich mit der Frage befassen, wie autonome Systeme den Prozess des Verstehens auf Seiten des Menschen durch Signale unterstützen könnten, mit denen sie ihre Intentionen gezielt kommunizieren. Im neu entstandenen Forschungsgebiet der External Human–Machine Interfaces (eHMI) wurden beispielsweise Simulatorexperimente durchgeführt, mit denen unterschiedliche Formen der Signalisierung getestet wurden (Rouchitsas und Alm 2023; Zhanguzhinova et al. 2023). Die wesentlichen Resultate sind:

- Ein Display an der Vorderseite autonomer Fahrzeuge wird dann als hilfreich empfunden, wenn es Gesten darstellt, nicht aber Smileys oder geschriebenen Text.
- Ein defensiver Fahrstil des autonomen Fahrzeugs wird als hilfreich empfunden, ein aggressiver hingegen nicht.

Hier sind erste Schritte getan, aber es gibt offenkundig noch weiteren Forschungsbedarf zu der Frage, wie die Maschine ihre Intentionen kommuniziert und sich durch gezielte Signale so verständlich macht, dass konfliktträchtige Situationen im Straßenverkehr sicher bewältigt werden können – und zwar auch in komplexeren Situationen, an denen mehr nur zwei Teilnehmer:innen beteiligt sind.

5.2.2 Wahrnehmung des Menschen durch die Maschine

Weniger weit entwickelt ist die Forschung zur Wahrnehmung des Menschen (und dessen Intentionen) durch die Maschine, die vor allem im Bereich des Social Signal Processing stattfindet (Vinciarelli et al. 2009, Matej Hrkalovic 2022). Zwar stehen diverse Techniken wie das Eye Tracking, die Gesture Recognition oder die Speech Recognition zur Verfügung; aber von einem echten Verständnis dessen, was der Mensch tut, sind Maschinen noch weit entfernt. Denn hierzu bedürfte es sozialer Intelligenz, über die selbst avancierte Systeme künstlicher Intelligenz noch nicht oder nur in Ansätzen verfügen, sowie eine weit entwickelte Fähigkeit zu sozialer Interaktion. Auch hier ist offenbar noch eine Menge zu tun, bevor man vermenschlichte Technik in den Alltag entlässt und ihr zutraut, Situationen wie die am Fußgängerüberweg erfolgreich zu bewältigen.

5.3 Interaktives Konfliktmanagement durch virtuellen Blickkontakt

Eine mögliche Option zur Lösung der beschriebenen Probleme wäre der virtuelle Blickkontakt, der dann funktionieren könnte, wenn einerseits die Menschen technisch aufgerüstet wären und andererseits die Technik teilweise vermenschlicht wäre. Man würde auf diese Weise die Möglichkeit der Kommunikation auf (virtueller) Augenhöhe schaffen.

Diese Idee wird im Folgenden am Beispiel eines autonomen Stadtbusses durchgespielt, der in einem innerstädtischen Bereich einer Gruppe Menschen begegnet, die am Straßenrand stehen und die Fahrbahn überqueren wollen. Wäre ein/e Fahrer:in an Bord, würde man Blickkontakt aufnehmen und durch Zeichen, Gesten und eigenes Verhalten eine Lösung finden. Diesen Vorgang müsste durch einen virtuellen Blickkontakt ersetzt werden (vgl. Abb. 9).

Voraussetzung ist, dass die Menschen smarte Geräte (z. B. Smartwatches) mit sich führen und eine Traffic-Warn-App installiert und aktiviert haben. In einem festgelegten Umkreis könnte diese App allen Verkehrsteilnehmer:innen, die eine potenzielle Gefahrenquelle darstellen, ein anonymisiertes Signal mit Positionsdaten und weiteren relevanten Informationen übermitteln.

Der autonome Stadtbus würde diese Informationen empfangen, in den Alarmmodus umschalten (gelb in Abb. 9) und ein Signal an alle potenziell betroffenen Akteure senden, das sie vor der herannahenden Gefahr warnt (rot Abb. 9). Wenn diese den Empfang der Nachricht bestätigen, ist der virtuelle Blickkontakt hergestellt, und es kann eine Interaktion, etwa über das Frontdisplay des Fahrzeugs, stattfinden. Diese beinhaltet z. B. das Angebot, dass der Bus anhält und die Straße

Abb. 9 Virtueller Blickkontakt am Beispiel eines autonomen Stadtbusses

gefahrlos überquert werden kann. Wird der Empfang dieser Information bestätigt, springt die Anzeige im Bus auf rot und die in den Smartwatches auf grün, womit eine Situation hergestellt ist, wie man sie von Ampeln gewohnt ist.

Dieses Konzept des virtuellen Blickkontakts ließe sich rasch realisieren, da alle wesentlichen Komponenten verfügbar sind. Andere, weit aufwändigere und weniger erprobte Verfahren der Gestenerkennung wären dann möglicherweise entbehrlich. Zudem hat das hier geschilderte Verfahren eines interaktiven Konfliktmanagements einen entscheidenden Vorteil: Es vermeidet Fehlinterpretationen, die beispielsweise dann entstehen könnten, wenn ein Teil der wartenden Fußgänger:innen die Gesten auf dem Display verstanden hat, ein anderer nicht.

Einen Nachteil hat das Konzept: Es setzt voraus, dass eine kritische Masse von Menschen mitmacht, z. B. aus Gründen des Selbstschutzes oder des Schutzes vulnerabler Gruppen. Eventuell kann die Nutzungsbereitschaft durch sanfte Anreize gesteigert werden – etwa durch Haftungsregelungen in Anlehnung an die Gurtpflicht im Auto.

Der virtuelle Blickkontakt würde auch helfen, vulnerable Gruppen wie Kinder oder Menschen mit Einschränkungen besser zu schützen, wie Abb. 10 zeigt.

Das Kind, das am Straßenrand spielt und für den autonomen Bus nicht sichtbar ist, würde durch ein smartes Gerät geschützt, das seine Position übermittelt und es so ermöglicht, potenzielle Konflikte rechtzeitig zu erkennen und zu entschärfen. Der virtuelle Blickkontakt würde also bestimmte Gruppen die Teilnahme am Verkehr bzw. ein Leben in vom Verkehr tangierten Bereichen ermöglichen, für die das bislang nicht gefahrlos möglich war.

Abb. 10 Virtueller Blickkontakt am Beispiel eines Kindes am Straßenrand

6 Regelverletzung in Konfliktsituationen

Das Interaktionsmodell, das in Abb. 8 dargestellt ist, setzt voraus, dass es ein Set von Regeln gibt, an die sich alle halten. Es gibt jedoch Situationen, in denen mehrere, sich widersprechende Regeln zu Konflikten führen, aber auch Situationen, in denen eine etwas großzügigere Auslegung von Regeln zu Effizienzgewinnen führt. Die Organisationssoziologie beschäftigt sich seit Längerem mit den informellen Beziehungen, also dem partiellen Unterlaufen strikter Regelsysteme, und deren Beitrag zu einer produktiven Arbeitsgestaltung (vgl. Kühl 2021).

Auch im Straßenverkehr kennt man dieses Phänomen, wenn beispielsweise ein/e Radfahrer:in auf einer schmalen Landstraße nur überholt werden kann, wenn man die durchgezogene weiße Mittellinie überfährt (Wongpiromsarn et al. 2021; Liu et al. 2022). Es gibt Berichte aus Kalifornien, dass autonome Autos mit ihrem defensiven und stur regelkonformen Fahrstil immer wieder für Stillstand und Verkehrschaos sorgen, weil sie derartige Situationen nicht meistern können.

Abschließend soll daher die Frage diskutiert werden, wie autonome Fahrzeuge in Zukunft mit derartigen Situationen umgehen, die von Regelkonflikten gekennzeichnet sind, und ob man ihnen – ähnlich wie Menschen – eine temporäre Regelverletzung erlauben will. Andreas Reschka (2015, S. 508) hat eine derartige Situation beschrieben, die im Straßenverkehr der Zukunft jederzeit auftreten könnte. Er hat dargelegt, dass die entstehenden Konflikte nur schwer zu lösen sind, weil jede Lösung neue Konflikte produziert, deren Folgewirkungen kaum abzuschätzen sind (vgl. Abb. 11).

An einer zweispurigen Straße mit durchgezogener Mittellinie taucht zwischen zwei am Straßenrand parkenden Fahrzeugen plötzlich ein (grüner) Fußgänger auf, der so spät zu erkennen ist, dass das (blaue) autonome Auto A nicht rechtzeitig

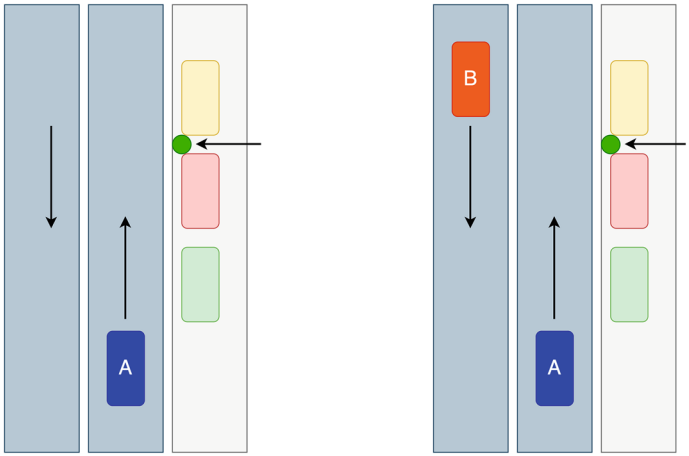

Abb. 11 Dilemma-Situationen im Straßenverkehr (in Anlehnung an: Reschka 2015, S. 508)

zum Stillstand kommen kann (linkes Bild). Es könnte die Situation entschärfen, indem es über die durchgezogene Mittellinie auf die Gegenfahrbahn ausweicht (Option 2), müsste dazu aber eine Regel verletzen. Es stellt sich somit die Frage, ob man dies dem autonomen Auto gestatten sollte, auch weil dies eine schwierige Güterabwägung beinhalten könnte, die eine Programmierer:in zudem im Software-Code ablegen müsste.

Noch komplizierter wird die Situation im Fall von Gegenverkehr (rechtes Bild in Abb. 11). Das autonome Auto kann Konflikt 1 (mit dem Fußgänger) lösen, indem es eine Regelverletzung begeht und einen weiteren Konflikt 2 provoziert, nämlich eine Kollision mit dem entgegenkommenden (orangen) Fahrzeug B (Option 3). Alternativ könnte es sich für eine kontrollierte Kollision mit parkenden Fahrzeugen (Option 4) entscheiden oder das entgegenkommende Fahrzeug B – falls es technisch entsprechend ausgestattet ist – in die Konfliktlösung mit einbeziehen, z. B. durch kooperatives Ausweichen (Option 5).

Diese – weitgehend moralfreie – Dilemma-Situation unterscheidet sich deutlich von dem künstlich aufgebauschten Trolley-Problem (Hevelke und Nida-Rümelin 2015). Es geht hier nicht um die Entscheidung, ob man eine alte oder eine junge Frau tötet, sondern um Fragen der Regelkonformität und Regelverletzung, wie sie Alltag regelmäßig auftauchen – nur dass in Zukunft Maschinen derartige Entscheidungen werden fällen müssen.

Derzeit ist weitgehend unklar, wie Lösungen für derartige Situationen aussehen könnten, in denen einem autonomen Auto das Recht eingeräumt werden müsste, bestehende Regeln zu verletzen und/oder eine Entscheidung zwischen mehreren Handlungsoptionen vorzunehmen, die allesamt schwer abschätzbare Folgen für Dritte mit sich ziehen.

7 Fazit

Es wäre eine Illusion zu glauben, dass wir uns mit der Entwicklung künstlicher Intelligenz all der – gelegentlich lästigen oder ärgerlichen – Probleme an der Schnittstelle zwischen Mensch und Technik ein für alle Mal entledigen können. Dieser Beitrag hat versucht zu zeigen, dass die Interaktion autonomer Fahrzeuge mit Passagieren und Passanten sozial voraussetzungsvoll ist und insbesondere eine Vermenschlichung autonomer Technik erfordert – im Sinne der Fähigkeit von Technik, in alltäglichen Situationen mit anderen Menschen und Maschinen so zu interagieren, dass eine Verständigung und gemeinsame Konfliktlösung möglich wird.

Der Techniksoziologie wird die Arbeit nicht ausgehen; denn die Durchdringung sämtlicher gesellschaftlicher Bereiche mit autonomen Systemen erfordert, dass diese Systeme die Regeln sozialer Interaktion beherrschen und als quasi-soziale Wesen am sozialen Leben teilhaben. Ob sie dafür eines Tages einen qualifizierten Abschluss in Soziologie benötigen werden, ist derzeit noch nicht abzusehen.

Literatur

Beck, Ulrich. 1986. *Risikogesellschaft. Auf dem Weg in eine andere Moderne.* Frankfurt/M.: Suhrkamp.

Bellon, Jacqueline, Bruno Gransche und Sebastian Nähr-Wagener. 2022. *Soziale Angemessenheit: Forschung zu Kulturtechniken des Verhaltens.* Wiesbaden: Springer VS.

Beyerer, Jürgen und Oliver Niggemann 2018: Machine Learning in Automation. *at-Automatisierungstechnik* 66 (4): 281–282.

Coleman, James S. 1995. *Grundlagen der Sozialtheorie. Handlungen und Handlungssysteme. Band 1.* München: Oldenbourg.

Drewitz, Uwe et al. 2021: Subjektive Sicherheit zur Steigerung der Akzeptanz des automatisierten und vernetzten Fahrens. *Forschung Im Ingenieurwesen* 85 (4): 997–1012.

Endsley, Mica R. und Esin O. Kiris 1995: The Out-of-the-Loop Performance Problem and Level of Control in Automation. *Human Factors* 37: 381–394.

Esser, Hartmut. 1993. *Soziologie. Allgemeine Grundlagen.* Frankfurt/M.: Campus.

Esser, Hartmut. 2000. *Soziologie. Spezielle Grundlagen, Bd. 3: Soziales Handeln.* Frankfurt/M.: Campus.

Fink, Robin D. und Johannes Weyer 2011: Autonome Technik als Herausforderung der soziologischen Handlungstheorie. *Zeitschrift für Soziologie* 40 (2): 91–111, https://doi.org/10.1515/zfsoz-2011-0201.

Habermas, Jürgen. 1968. *Technik und Wissenschaft als Ideologie.* Frankfurt/M.: Suhrkamp.

Habermas, Jürgen. 1981a. *Theorie des kommunikativen Handelns. Bd. 1: Handlungsrationalität und gesellschaftliche Rationalisierung.* Frankfurt/M.: Suhrkamp.

Habermas, Jürgen. 1981b. *Theorie des kommunikativen Handelns. Bd. 2: Zur Kritik der funktionalistischen Vernunft.* Frankfurt/M.: Suhrkamp.

Hevelke, Alexander und Julian Nida-Rümelin 2015: Ethische Fragen zum Verhalten selbstfahrender Autos. *Zeitschrift für Philosophische Forschung* 69 (2): 217–224, https://doi.org/10.3196/004433015815493721.

Hidalgo, César A et al. 2021. *How humans judge machines.* MIT Press.

Konz, Britta, Karl-Heinrich Ostmeyer und Marcel Scholz (Hrsg.). 2023. *Gratwanderung Künstliche Intelligenz. Interdisziplinäre Perspektiven auf das Verhältnis von Mensch und KI.* Stuttgart: Kohlhammer.

Kühl, Stefan 2021: Die hohe Zuverlässigkeit der Illegalität–Zum Management von Regelabweichung und Regelkonformität in Organisationen. *Kooperation in der digitalen Arbeitswelt: Verlässliche Führung in Zeiten virtueller Kommunikation*: 317–329.

Latour, Bruno. 1998. Über technische Vermittlung. Philosophie, Soziologie, Genealogie. In *Technik und Sozialtheorie,* Hrsg. Werner Rammert, 29–81. Frankfurt/M.: Campus.

Liu, Jiaxin, Wenhui Zhou, Hong Wang, Zhong Cao, Wenhao Yu, Chengxiang Zhao, Ding Zhao, Diange Yang und Jun Li. 2022. *Road Traffic Law Adaptive Decision-making for Self-Driving Vehicles.* In: 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC): 2034–2041. https://doi.org/10.1109/ITSC55140.2022.9922208

Lobe, Adrian 2015: Big Data und Politik. Brauchen wir noch Gesetze, wenn Rechner herrschen. *Frankfurter Allgemeine Zeitung* 14. Jan. 2015: 13, http://www.faz.net/-gsf-7y9q3.

Luhmann, Niklas. 1984. *Soziale Systeme. Grundriß einer allgemeinen Theorie.* Frankfurt/M.: Suhrkamp.

Matej Hrkalovic, Tiffany 2022: Designing Hybrid Intelligence Techniques for Facilitating Collaboration Informed by Social Science. *Proceedings of the 2022 International Conference on Multimodal Interaction*: 679–684.

Mercedes 2015: *Der F015 Luxury in Motion.* www.mercedes-benz.com/de/innovation/autonomous/forschungsfahrzeug-f-015-luxury-in-motion (Datum des letzten Zugriffs: 15.02.2022).

Nassehi, Armin. 2019. *Muster: Theorie der digitalen Gesellschaft.* München: C.H. Beck.

Popitz, Heinrich. 1995. Epochen der Technikgeschichte. In *Der Aufbruch zur Artifiziellen Gesellschaft. Zur Anthropologie der Technik,* Hrsg. Heinrich Popitz, 13–43. Tübingen: J.C.B. Mohr.

Rammert, Werner und Ingo Schulz-Schaeffer. 2002. Technik und Handeln. Wenn soziales Handeln sich auf menschliches Verhalten und technische Abläufe verteilt. In *Können Maschinen handeln? Soziologische Beiträge zum Verhältnis von Mensch und Technik,* Hrsg. dies., 11–64. Frankfurt/M.: Campus.

Reeves, B. und C.I. Nass. 1996. *The media equation: How people treat computers, television, and new media like real people and places.* Cambridge/Mass.: Cambridge University Press.

Reschka, Andreas. 2015. Sicherheitskonzept für autonome Fahrzeuge. In *Autonomes Fahren. Technische, rechtliche und gesellschaftliche Aspekte,* Hrsg. Markus Maurer, J. Christian Gerdes, Barbara Lenz und Hermann Winner, 489–513. Berlin: Springer.

Rochlin, Gene I. 1997. *Trapped in the net. The unanticipated consequences of computerization.* Princeton: Princeton UP.

Rosa, Hartmut. 2005. *Beschleunigung. Die Veränderung der Zeitstrukturen in der Moderne.* Frankfurt/M.: Suhrkamp.

Rouchitsas, Alexandros und Håkan Alm 2023: Smiles and Angry Faces vs. Nods and Head Shakes: Facial Expressions at the Service of Autonomous Vehicles. *Multimodal Technologies and Interaction* 7 (2): 10.

Sarter, Nadine B. und David D. Woods 1997: Team Play with a Powerful and Independent Agent: Operational Experiences and Automation Surprises on the A-320. *Human Factors* 39: 553–569.

Schimank, Uwe. 2010. *Handeln und Strukturen. Einführung in eine akteurtheoretische Soziologie (4. Aufl.).* München: Juventa.

Schrape, Jan-Felix. 2021. *Digitale transformation.* Stuttgart: utb.

Sturma, Dieter. 2001. Robotik und menschliches Handeln. In *Robotik. Perspektiven für menschliches Handeln in der zukünftigen Gesellschaft,* Hrsg. Thomas Christaller, 111–134. Berlin: Springer.

Turkle, Sherry. 2005. *The Second Self: Computers and the Human Spirit.* Cambridge/Mass.: MIT-Press.

Turkle, Sherry. 2011. *Alone together. Why we expect more from technology and less from each other.* New York: Basic Books.

Vinciarelli, Alessandro, Maja Pantic und Hervé Bourlard 2009: Social signal processing: Survey of an emerging domain. *Image and vision computing* 27 (12): 1743–1759.

Waymo LLC, 2020: *Waymo Safety Report.* https://storage.googleapis.com/sdc-prod/v1/safety-report/2020-09-waymo-safety-report.pdf.

Weber, Max. 1985. *Wirtschaft und Gesellschaft. Grundriss der verstehenden Soziologie (1922).* Tübingen: J.C.B. Mohr.

Weyer, Johannes. 2008. *Techniksoziologie. Genese, Gestaltung und Steuerung sozio-technischer Systeme (Grundlagentexte Soziologie).* Weinheim: Juventa.

Weyer, Johannes. 2016: Confidence in hybrid collaboration. An empirical investigation of pilots' attitudes towards advanced automated aircraft. *Safety Science* 89: 167–179, https://doi.org/10.1016/j.ssci.2016.05.008.

Weyer, Johannes. 2019a. Autonome Technik außer Kontrolle? Möglichkeiten und Grenzen der Steuerung komplexer Systeme in der Echtzeitgesellschaft. In *Roboter in der Gesellschaft. Technische Möglichkeiten und menschliche Verantwortung,* Hrsg. Christiane Woopen und Marc Jannes, 87–109. Berlin: Springer.

Weyer, Johannes. 2019b. *Die Echtzeitgesellschaft. Wie smarte Technik unser Leben steuert.* Frankfurt/M.: Campus.

Weyer, Johannes. 2022a: *Die Echtzeitgesellschaft. Theoretische und methodische Herausforderungen der Soziologie (Soziologisches Arbeitspapier 61/2022).* Dortmund: TU Dortmund, http://hdl.handle.net/2003/41147.

Weyer, Johannes. 2022b: *Vermenschlichung der Technik? Die Interaktion von Menschen und künstlicher Intelligenz in alltäglichen Kontexten.* Dortmund: TU Dortmund, Soziologische Arbeitspapiere, https://doi.org/10.17877/DE290R-22593.

Weyer, Johannes. 2023. Vermenschlichung der Technik? Die Interaktion von Menschen und künstlicher Intelligenz in alltäglichen Kontexten. In *Gratwanderung Künstliche Intelligenz. Interdisziplinäre Perspektiven auf das Verhältnis von Mensch und KI,* Hrsg. Britta Konz, Karl-Heinrich Ostmeyer und Marcel Scholz, 43–60. Stuttgart: Kohlhammer.

Weyer, Johannes, Fabian Adelt und Sebastian Hoffmann, 2015: *Governance of complex systems. A multi-level model (Soziologisches Arbeitspapier 42/2015).* Dortmund: TU Dortmund, http://hdl.handle.net/2003/34132.

Weyer, Johannes und Kay Cepera 2021: Vertrauen in digitale Technik. Der Einfluss mobiler Applikationen auf die Bereitschaft zur Verhaltensänderung. *Zeitschrift für Soziologie* 50 (6): 373–395, https://doi.org/10.1515/zfsoz-2021-0028.

Wongpiromsarn, Tichakorn et al. 2021: Minimum-violation planning for autonomous systems: Theoretical and practical considerations. *2021 American Control Conference (ACC)*: 4866–4872.

Zhanguzhinova, Symbat et al. 2023: Communication between Autonomous Vehicles and Pedestrians: An Experimental Study Using Virtual Reality. *Sensors* 23 (3): 1049.

Prof. Dr. Johannes Weyer Fakultät Sozialwissenschaften der TU Dortmund.

Johannes Weyer ist Seniorprofessor für Nachhaltige Mobilität an der Fakultät für Sozialwissenschaften der TU Dortmund und war von 2002-2022 Professor für Techniksoziologie an der Wirtschafts- und Sozialwissenschaftlichen Fakultät der TU Dortmund. Das Spektrum seiner Forschungsarbeit umfasst u.a. Technikbewertung und Technikakzeptanzforschung, Risikomanagement in Organisationen, agentenbasierte Modellierung und Simulation sozio-technischer Systeme, die Mensch-Maschine-Interaktion sowie autonome technische Systeme mit Anwendungen in den Gebieten Luft- und Raumfahrt, Straßenverkehr, Energiesysteme und in der Chemieindustrie.

Ist KI zu kontrollieren? Überlegungen zur Ethik des Zusammenwirkens von Menschen und KI-Maschinen

Stefan Heuser und Jochen J. Steil

Zusammenfassung

Dieser Beitrag geht von der Beobachtung aus, dass sich aktuelle Ethikdiskurse über KI-Systeme auf Fragen der Kontrolle und Regulierung zur Reduktion norm- und regelverletzender Möglichkeiten des Zusammenwirkens von Mensch und Maschine konzentrieren. Während wir Forderungen zur Offenlegung von Trainingsdaten und -methoden sowie der Filter und regelbasierter Ausgabemechanismen im Rahmen solcher externer Kontrolle teilen, muss wirksame Kontrolle aber auch mit der inneren Komplexität von KI-Systemen und der Offenheit rekursiver Kopplungen im realen Zusammenwirken von Menschen und Maschinen skalieren. Auf der Grundlage einer systemtheoretischen Rekonstruktion des Kontrollproblems zeigen wir, dass dazu das Zusammenwirken zwischen Menschen und KI-Systemen in weiten Teilen des Diskurses weder hinreichend systemintern (bezogen auf Fragen der Selbststeuerung), noch hinreichend immanent (bezogen auf Fragen der sinnvollen Fortsetzung von Lebensformen und Praktiken) bearbeitet wird. Ausgehend von einer Beschreibung von KI-Systemen als „Kontinuierungsmaschinen" stellen wir daher die These auf, dass zur ethischen Reflexion von KI-Systemen

S. Heuser (✉)
Institut für Evangelische Theologie und Religionspädagogik der TU Braunschweig, Braunschweig, Deutschland
E-Mail: s.heuser@tu-braunschweig.de

J. J. Steil
Institut für Robotik und Prozessinformatik der TU Braunschweig, Braunschweig, Deutschland
E-Mail: j.steil@tu-braunschweig.de

O. Richter et al. (Hrsg.), *Zusammenwirken von natürlicher und künstlicher Intelligenz: Beurteilen-Messen-Bewerten,*
https://doi.org/10.1007/978-3-658-45845-4_3

neben intelligenter Regulierung und bedeutsamer Kontrolle auch die Frage gehört, ob und wie menschliche Lebensformen und die mit ihnen verbundenen Praktiken im Zusammenwirken von Menschen und KI-Systemen sinnvoll fortgesetzt und neue Handlungsmöglichkeiten eröffnet werden können. Dabei halten wir es für notwendig, die menschliche Urteilskraft im Zusammenwirken mit KI-Maschinen zum Tragen zu bringen, damit die maschinelle Generierung virtueller Bedeutung in ein Zusammenspiel mit weltverstehender und auf das Zusammenleben ausgerichteter Intelligenz gebracht werden kann.

Schlüsselwörter

KI-Systeme • Kontrollproblem • Systemtheorie • Kontinuierungsmaschine • Urteilskraft • Lebensformen

1 Auf der Suche nach Kontrolle: Ethik-Diskurse über das Zusammenwirken von Menschen und KI-Maschinen

Die Warnungen vor Kontrollverlust beim Wettlauf um die Entwicklung immer leistungsstärkerer KI-Systeme werden häufiger – und sie werden lauter. So forderten der Turing-Award-Gewinner Yoshua Bengio, Tesla-Chef Elon Musk, der Direktor des Center for Intelligent Systems in Berkeley Stuart Russell, der Physiker Max Tegmark vom MIT, der israelische Historiker Yuval Noah Harari und weitere renommierte Wissenschaftler in einem auf der auf der Webseite des amerikanischen „Future of Life Institute“ veröffentlichten offenen Brief alle KI-Labore weltweit dazu auf, „to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4“ (Bengio et al. 2023). KI-Labore befänden sich demnach „in an out-of-control race to develop and deploy ever more powerful digital minds that no one – not even their creators – can understand, predict, or reliably control“ (ebd.). Gegenwärtige KI-Systeme könnten bei der Erledigung allgemeiner Aufgaben immer mehr mit Menschen konkurrieren, so dass die Initiatoren des Aufrufs die Frage stellen: „*Should* we develop nonhuman minds that might eventually outnumber, outsmart, obsolete and replace us? *Should* we risk loss of control of our civilization?“ (ebd.). Schon im Jahr 2017 hatten Mitglieder des „Future of Life Institute“ auf der Beneficial AI Conference im symbolträchtigen kalifornischen Asilomar 23 KI-Leitsätze erarbeitet mit dem Ziel, die Entwicklung von KI-Systemen nicht gleichsam naturwüchsig ablaufen zu lassen, sondern in politische Meinungsbildungsprozesse und Wertereflexionen einzubinden und ihre wohltätige Nutzung zu sichern (Future of Life

Institute 2017). Prominente Pioniere der tiefen neuronalen Netze wie beispielsweise Turingpreisträger Geoffrey Hinton, der bis vor kurzem für Google arbeitete, fordern Kontrolle: „The best hope is that you take the leading scientists and you get them to think very seriously about how are we going to be able to control this stuff.[1]“ (Hinton 2023; vgl. auch die Diskussion in Brockman 2021).

In Deutschland ist die jüngste Stellungnahme des Deutschen Ethikrats: Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz (2023) für die Debatte kennzeichnend. Diese geht von einem normativ grundlegenden Unterschied zwischen „Mensch“ und „Maschine“ aus und rückt das Kontrollproblem in den Horizont der Frage, wie sich in Interaktionen von Menschen mit intelligenten Systemen die Zuschreibung von Handlungsträgerschaft und Verantwortung gewährleisten lässt.

Auch im Bereich der nationalen und internationalen Politik gibt es zahlreiche Versuche zur Kontrolle und Regulierung von KI-Systemen. Schon im Jahr 2019 hatte eine von der Europäischen Kommission eingesetzte hochrangige Expertengruppe für Künstliche Intelligenz „Ethik-Leitlinien für eine vertrauenswürdige KI“ entwickelt. Diese enthalten eine Bewertungsliste, an deren erster Stelle ein „Vorrang menschlichen Handelns und menschlicher Aufsicht“ steht (HEG-KI 2019). Aus diesem Jahr stammt auch die Empfehlung der Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD) zu künstlicher Intelligenz (OECD 2019). Mit Vorbildfunktion weltweit plant die Europäische Union diese Leitlinien in entsprechende Regulierung umzusetzen. Die entsprechende KI-Verordnung befindet sich in der parlamentarischen Beratung[2]. In einem risikobasierten Ansatz sollen spezifische KI-Softwaresysteme in bestimmten Einsatzgebieten insgesamt verboten werden. In den niedrigeren Risikostufen wird jedoch viel auf die untergesetzliche Ebene verwiesen, hier ist also auf die Durchführungsverordnungen zu warten, die dann definieren, was genau die gewünschte Kontrolle umfassen und wie sie implementiert werden soll. Entscheidende Details sind jedoch noch offen. Entsprechend interpretieren die Vorreiter der Entwicklung wie z. B. OpenAI die Diskussion in neoliberaler Tradition eher als Aufforderung zur „Selbstkontrolle“ und wollen sich nur ungern externen und überprüfbaren Regulierungen unterwerfen.

Der Mainstream des aktuellen Diskurses über KI-basierte Systeme konzentriert sich demnach auf die Frage nach deren Kontrollier- und Regulierbarkeit.

[1] „Die größte Hoffnung ist, die führenden Wissenschaftler/innen dazu zu bewegen, sehr intensiv darüber nachzudenken, ob wir in der Lage sein werden, diese Dinge zu kontrollieren“.

[2] Vgl. zum Stand des Gesetzgebungsverfahrens: https://eur-lex.europa.eu/legal-content/DE/HIS/?uri=CELEX:52021PC0206, zuletzt aufgerufen am 24.09.2023.

Betrachtet man diesen Ansatz aus einer systemtheoretischen Perspektive, wie sie vor allem Niklas Luhmann entwickelt (Luhmann 1993) und aktuell Dirk Baecker und Armin Nassehi auf intelligente Digitaltechnologien angewendet haben (Baecker 2019; Nassehi 2019), kann man darin den Versuch erkennen, das System des Rechts den Ausdehnungs- und Eingriffsmöglichkeiten des Digitalsystems entgegenzusetzen. Durch die Definition und Zuweisung des Codes von Recht und Unrecht sollen Grenzen bestimmt und Rahmenbedingungen geschaffen werden, durch die Menschen vor Rechtsverletzungen geschützt, Missbrauch geahndet, Streitfälle absorbiert und Haftungsfragen geregelt werden. So fraglos notwendig die Implementierung solcher Rahmenbedingungen ist, so offen bleiben dabei ethische Fragestellungen, die sich nicht *systemisch,* d. h. in einer *binären* Logik abbilden lassen. Im Blick sind vor allem Missbrauchsszenarien wie etwa die Nutzung von KI für die Verbreitung von Falschinformationen und für Datenschutzverletzungen, aber auch absehbare Entwicklungsdynamiken wie Veränderungen der Arbeitswelt (z. B. der Wegfall von Angestelltenverhältnissen) und die Verwendung von KI in autonomen, ihre Zielobjekte selbständig identifizierenden und Angriffe initiierenden Waffensystemen. Die Möglichkeit oder sogar große Wahrscheinlichkeit einer technologischen Singularität, d. h. der Entwicklung einer dem Menschen überlegenen Superintelligenz, die sich selbst Ziele setzt und die technologische Entwicklung von da an selbständig ohne menschliche Kontrolle vorantreibt (Kurzweil 2005; Alfonseca et al. 2021), ist dabei oft eine unhinterfragte implizite Annahme, die der Dringlichkeit der Diskussion über Kontrolle zu unterliegen scheint. Der Ethik-Diskurs über die Kontrolle von KI wird dabei von folgenden Fragen geprägt: Können Menschen KI (noch) kontrollieren? Wer kontrolliert die Menschen, die KI-basierte Systeme entwickeln und diese einsetzen? Wie weit reicht die menschliche Kontrolle angesichts der wachsenden Macht der Maschinen und der Menschen, die sie programmieren und sich ihrer bedienen? Welche Verständigungsformen und Regelungen werden benötigt, um ungewünschte Entwicklungen einzuhegen und Risiken zu begrenzen?

So bedeutsam und notwendig diese Fragen und die Suche nach rechtlichen Regulierungen und Kontrollmöglichkeiten sind, so deutlich zeigen sie doch auch, dass sich nicht alle normativ relevanten Aspekte des Zusammenwirkens von Menschen und KI-Systemen im Gegenüber der Systeme „Recht“ und „Digitaltechnik“ abbilden lassen.[3] Ein Blick auf die Verletzlichkeit menschlicher Lebensformen und Praktiken genügt, um zu erkennen, dass es sich beim Zusammenwirken

[3] Wir gehen in diesem Beitrag davon aus, dass es sich bei der Digitaltechnologie systemtheoretisch betrachtet um ein System, d. h. um eine auf Autopoiesis ausgerichtete Struktur handelt.

von Menschen und KI-Systemen nicht um das Zusammenwirken von Äquivalenten, sondern von inkommensurablen Gegenübern handelt.[4] Die Frage nach der Kontrolle dieses Zusammenwirkens wird hinsichtlich seiner technischen Seite durch das Kontrollproblem verschärft; hinsichtlich seiner menschlichen Seite durch die Vulnerabilität von Personen sowie durch die Vagheit, Offenheit und Unabsehbarkeit der Lebenswelt.

Das Kontrollproblem lässt sich systemtheoretisch als die Art und Weise verstehen, wie sich Systeme so in Beziehung zur Komplexität ihrer Umwelt setzen, dass sie sich selbst weiter reproduzieren können. Vor diesem Hintergrund ist es beispielsweise keineswegs klar, was und wie ein Moratorium zur tatsächlichen Kontrolle beitragen würde. Eine vertiefende Betrachtung des Gegenstandes „KI-Maschinen" legt es dagegen nahe, kybernetische Fragen systemischer Selbstregulierung in den Vordergrund zu rücken (Wiener 1948/2000). Bedient man sich für ein umfassenderes Verständnis des Kontrollproblems aus dem theoretischen Werkzeugkoffer der Systemtheorie, dann lässt sich das Zusammenwirken inkommensurabler Systeme anhand ihrer Operationen, und insbesondere ihrer jeweiligen Selektionsleistungen zur Selbsterschaffung und Selbsterhaltung beschreiben (Baecker 2019). Elemente dieser Theorie lassen sich für eine Analyse des Zusammenwirkens von Menschen und Maschinen, insbesondere für die Frage nach der Kontrolle von künstlicher Intelligenz in Gebrauch nehmen.

Systemtheoretisch gesprochen steigt „Intelligenz" mit der Menge an Möglichkeiten, aus der ein System die für die Lösung seines Problems angemessene Auswahl treffen kann („Selektion", Ashby 1958), was für Menschen und Maschinen gleichermaßen gelten könnte. Die Daten, welche die Algorithmen von KI-Maschinen dazu statistisch erfassen, sind digitale Abbilder und das indirekte Produkt des Wechselspiels und Zusammenwirkens der Systeme Mensch und KI-Maschine. Blickt man vor diesem Hintergrund auf das Zusammenwirken, dann lässt es sich als Kooperation verschiedener Intelligenzen rekonstruieren. Diese hat Auswirkungen auf die Kontrolle und die Selektionsmöglichkeiten der beteiligten Systemreferenzen in ihrem Bezug auf die Komplexität ihrer jeweiligen Umwelt, welche es zu analysieren und zu bewerten gilt.

[4] Unser Fokus auf das „Zusammenwirken" von Menschen und KI-Systemen schließt an Forschungsperspektiven der Kommission „Synergie und Intelligenz: technische, ethische und rechtliche Herausforderungen des Zusammenwirkens lebender und nicht lebender Entitäten im Zeitalter der Digitalisierung" (SYnENZ) der Braunschweigischen Wissenschaftlichen Gesellschaft (BWG) an, die zusammen mit weiteren Forschungseinrichtungen in den Jahren 2019 und 2023 Symposien über das Zusammenwirken von natürlicher und künstlicher Intelligenz veranstaltet hat (Haux et al. 2021 & in diesem Band).

In der Analyse der Maschine sind dabei die Fragen danach grundlegend, welche Operationen diese eigentlich ausführt und in welchem Sinne diese selektiv oder selbst-reproduzierend sind. Menschliche Kontrolle bezieht sich jedoch auch auf die ethische Frage, wie sich ein als „gut“ und „tragfähig“ erfahrenes Leben in der Vielfalt seiner Formen und Praktiken im Zusammenwirken mit intelligenten Systemen sinnvoll fortsetzt. Dieses Spannungsfeld zwischen der Funktionalität des Systems (das Menschen als „Umwelt“ erfasst) und der Grammatik der menschlichen Lebenswelt (deren Akteure Aufgaben bewältigen, Praktiken ausüben und Lebensformen ausfüllen, statt bloß Funktionen zu erfüllen) wird jedoch nicht hinreichend erfasst, wenn es wie in weiten Teilen des Diskurses auf die Frage reduziert wird, wie „Menschen“ die „Maschinen“ kontrollieren. Durch diesen Fokus werden weder die Stärken eines systemtheoretischen, noch die eines ethiktheoretischen Zugangs für die Analyse ihres Zusammenwirkens ausgeschöpft.

In der jüngsten Stellungnahme des Deutschen Ethikrats (Deutscher Ethikrat 2023) scheinen diese beiden Aspekte auf den ersten Blick zusammenzulaufen. Sie fragt danach, wie wir das Zusammenwirken von Menschen mit KI-Systemen so erfassen und regulieren können, dass eine Benennung, Zuschreibung und Begrenzung von Verantwortung im Sinne der moralischen Rechtfertigung von Zwecken und Handlungsfolgen durch menschliche Handlungsträger sowie im Sinne der rechtlichen Haftung möglich bleibt. Im Vordergrund steht hier die Frage nach der Verantwortbarkeit des Einsatzes von KI-Maschinen in moralisch sensiblen Handlungsfeldern und Sektoren, wie beispielsweise in der Medizin, im Bildungsbereich, in der öffentlichen Kommunikation und in der Verwaltung. Die ethische Urteilsbildung wird dabei als Reflexion über die Verantwortbarkeit der Folgen des Einsatzes von KI-Systemen als Handlungsmittel verstanden. Auch Fragen der Verantwortlichkeit im Sinne der Haftung treten hier in den Vordergrund, und mit ihnen das Problem der Urheberschaft bzw. Autorschaft von Handlungen, die moralische Rechtfertigungsfähigkeit, Entscheidungskompetenz und Autonomie der (Ko-)Akteure sowie die Überschaubarkeit, Nachvollziehbarkeit, Transparenz und Kontrollierbarkeit der Folgen der Handlungsmittel. Die Stellungnahme des Deutschen Ethikrats kann als paradigmatisch für diesen *verantwortungsethischen* Zugang gelten. Darin werden die Konturen dessen, was als „verantwortbar“ gelten soll, prinzipienethisch formuliert (vgl. auch HEG-KI 2019; OECD 2019) und laufen auf Fragen der gesellschaftlichen Steuerung hinaus. Der Fokus auf die notwendigen Kontroll- und Regulierungsbedarfe führt nicht nur hier, sondern nahezu überall im Ethikdiskurs zur Forderung nach der Formulierung allgemeiner und konsensfähiger Prinzipien für das Zusammenwirken von Menschen und

KI-Maschinen mit dem Ziel, die Grenzen des moralisch Verantwortbaren hinsichtlich der Folgen dieses Zusammenwirkens zu markieren und abzusichern. Das Zusammenwirken von Menschen und intelligenten Maschinen soll demnach in einen möglichst ungebrochenen Begründungszusammenhang mit allgemeinen moralischen Prinzipien und rechtlichen Regelungen eingeordnet werden.

Dieser verantwortungsethische Diskurs, der auf die Forderung einer Kontrolle und Regulierung des Zusammenwirkens durch die Befolgung allgemeiner Prinzipien hinausläuft, bleibt dem Systemzusammenhang allerdings eigentümlich *extern*. Übertragen wir Rahel Jaeggis Rekonstruktion dreier Formen von Sozialkritik: nämlich interne, externe und immanente Kritik (Jaeggi 2009, 2014; vgl. auch Greve 2015) auf die Analyse des Zusammenwirkens von Menschen und KI-Maschinen, dann zeigt sich die Einseitigkeit dieses Fokus auf *externe* Kontrolle. Wird zusätzlich die Perspektive einer *internen* Kritik entwickelt, dann rücken Wertvorstellungen in den Blick, die in der Praxis bereits ausdrücklich enthalten sind. Eine kybernetische Rekonstruktion des Zusammenwirkens könnte eine solche kritische Perspektive entwickeln und dabei das Kriterium der Funktionalität des Zusammenwirkens in den Blickpunkt rücken. Der Versuch des Deutschen Ethikrats, eine Kontrolle über KI-Maschinen verantwortungsethisch zu gewinnen, bleibt dem Zusammenwirken von Menschen und KI-Maschinen demgegenüber weitgehend äußerlich und dürfte aus kybernetischer Sicht daran scheitern, eine der Komplexität der Verschaltungs- und Vernetzungsvorgänge angemessene, die Funktionalität erhaltende Steuerung aufzubauen (Ashby 1958).

Im Anschluss an Jaeggis Schema dreier Formen von Kritik stellt sich aber eine weitere Frage: Gibt es für das Zusammenwirken von Menschen und KI-Systemen auch eine Form von *immanenter* Kritik, die aus der Perspektive jener Orientierungen entwickelt werden kann, die die Integrität einer (lebensweltlichen) Praxis bestimmen und stützen, ohne *explizit* in Form von Regeln vorzuliegen bzw. überhaupt regelförmig rekonstruierbar zu sein? Und wäre es möglich, in einer solchen Perspektive nicht nur die Deutung, sondern auch die Erneuerungs- und Transformationsmöglichkeiten von Praktiken des Zusammenwirkens in den Fokus zu rücken (Jaeggi 2009, 287)?

Eine Antwort auf die Frage, ob und wie KI zu kontrollieren ist, benötigt nach unserer Auffassung eine solche um die Kriterien von Funktionalität, Integrität und Transformation erweiterte Analyse des Zusammenwirkens von Menschen und KI-Maschinen. Sie muss auch jene Aspekte eines guten und sinnvollen Lebens berücksichtigten, welche die Praktiken des Umgangs mit KI unhintergehbar ausmachen, selbst wenn diese „Werte“ nicht ausdrücklich formuliert sind bzw. nicht in Regelform gebracht werden können. Im Anschluss an eine sowohl systemtheoretische als auch ethische Betrachtungsweise des Kontrollproblems und an

Jaeggis Schema dreier Formen von Kritik wirft eine solche Analyse folgende Fragen auf:

1. An welchen ausdrücklichen Parametern macht sich die Funktionalität des Zusammenwirkens von Menschen und KI-Maschinen (inklusive der Operationen ihrer Algorithmen) im spannungsvollen Zusammenspiel von ethischer und kybernetischer Logik fest (interne Kritik)?
2. Ist es – im Anschluss an 1) – über die Formulierung von im Zusammenwirken unbedingt zu gewährleistenden, allgemeinen Regeln hinaus (externe Kritik) auch möglich, zu bestimmen, was zu einem die Integrität lebensweltlicher Praktiken fortführenden und auch ihre Transformation ermöglichenden Zusammenwirken von Menschen und KI-Maschinen gehört (immanente Kritik)?

Die folgenden Abschnitte verstehen sich als Vorklärungen auf der Suche nach Antworten auf diese Fragen. Als Grundlage für mögliche weitere – auch empirische – Forschungen liefern sie einen ersten Ansatz eines entsprechend erweiterten analytischen Zugangs zum Zusammenwirken von Menschen und KI-Maschinen.

2 Auf der Suche nach einer erweiterten Analyse des Zusammenwirkens

Wie wir im vorherigen Abschnitt angedeutet haben, wird das Kontrollproblem im Zusammenwirken zwischen Menschen und KI-Maschinen in weiten Teilen des Ethikdiskurses weder hinreichend system*intern* (bezogen auf Fragen der Kybernetik bzw. Selbststeuerung), noch hinreichend *immanent* (bezogen auf Fragen der Integrität und der Transformation von Praxis), sondern vor allem *extern* (bezogen auf Fragen der ausdrücklichen Regulierung im Sinne von „Compliance“, d. h. Regelbefolgung) bearbeitet. Letzteres geschieht, wie oben diskutiert, zum einen durch die Forderung der Bindung des Einsatzes von KI-Maschinen an allgemeine moralische Prinzipien und rechtliche Regeln, zum anderen durch das Postulat, dass die Zuschreibung von Verantwortung für die Folgen des Einsatzes von KI-Maschinen an menschliche Akteure jederzeit möglich bleiben muss.

Gerade diese zweite, scheinbar so einfache Forderung nach dem „human-in-the-loop“ oder „human-in-control“ (HIC) (HEG-AI 2019, 16) als Verantwortungsträger ist jedoch klärungsbedürftig, und sie ist angesichts der umfassenden Durchdringung der Umwelt mit Informationstechnologie und der Komplexität von entsprechenden Entscheidungs- und Kommunikationssituationen in der Praxis

häufig unrealistisch. Sie ist der kooperativen Rolle des Menschen im Zusammenwirken mit KI-Technologie nicht angemessen, da die geforderte Kontrolle in dem Kontext oft faktisch gar nicht realistisch ausgeübt werden kann (Steil et al. 2019). Der Mensch wird so – Susanne Beck zufolge – zum „Haftungsknecht", und im Versuch dieses insbesondere juristische Dilemma zu fassen hat Susanne Beck neue Konzepte von „meaningful control", d. h. den Kontext berücksichtigende Konzepte externer Kontrolle und Verantwortung vorgeschlagen (Beck 2020). Was die Verantwortung betrifft, führt also die Vorstellung, durch „human in-the-loop"/HIC könnte die Kontrollproblematik endgültig gelöst werden, unter Umständen zu einer dysfunktionalen Forderung nach einer grenzenlosen Ausweitung menschlicher Handlungskompetenz (Oesterreich 1981).

Andererseits aber könnte sich eine allzu eindimensionale Forderung von „Kontrolle" als Reduktion menschlichen, nicht-determinierten und auf kontingente Fortsetzung von Praktiken gerichteten Handelns als ebenso dysfunktional erweisen. Angesichts insbesondere der generativen Sprachmodelle scheint eine entsprechende Einhegung der Interaktion von Mensch und intelligenter Maschine zum Scheitern verurteilt. Denn reden Menschen frei mit Maschinen, werden die sich daraus ergebenden Kontingenzen nicht in und durch eine regelbasierte Ordnung einzufangen sein, da sich die gesamte Vagheit und Vielfältigkeit der Lebenswelt in diesem sprachlichen Zusammenwirken abbilden wird. Hierin liegt auch ein Kern der vielfältigen und sehr berechtigten Warnungen vor Missbrauch der Technologie z. B. zur Erzeugung von Desinformation oder Beeinflussung von politischen Entscheidungen. Es gibt zunächst keine inhärenten technologischen Grenzen, die verhindern würden, auch schädliche Kontingenzen herzustellen, was ja praktisch auch zu beobachten ist.

Das Bedürfnis, das Zusammenwirken von Menschen mit in einem weiten Sinne intelligenten maschinellen Systemen zu kontrollieren und Verantwortlichkeiten zuzuschreiben, steht zusätzlich quer zu der beabsichtigten „Autonomie" solcher Systeme, die darauf abzielt, neue und eigenständige Einsichten, Alternativen, Einflussmöglichkeiten, Lösungen und Aktionen in komplexen Umwelten zu generieren. Dafür müssen die Systeme auf ein hohes Maß an „requisite variety" (Ashby 1958), d. h. innere, die externe Komplexität spiegelnde Vielfältigkeit, zurückgreifen können. Die Stärke von KI-Maschinen besteht ja gerade darin, dass sie nicht einfach disponible „Werkzeuge" sind, sondern daraufhin programmiert sind, ein eigenes Verhaltensspektrum zu zeigen und dadurch teilweise opak gegenüber dem menschlichen Beobachter zu bleiben (Hubig 2015, 2019). Das Kontrollproblem persistiert, muss aber angesichts der Interaktionsvorgänge zwischen Nutzenden und intelligenten Systemen mit dem Verschwinden von Schnittstellen, mit Disponibilitätsverlust und Hybridisierung (Hubig 2017), mit

der Herausbildung von verantwortungsdiffundierenden Akteurnetzwerken (Latour 2007) sowie mit dem Abbau der Widerständigkeit der analogen Wirklichkeit (Wiegerling 2021) und deren Umwandlung in eine den digitalen, ziffernbasierten Informationsprozessen leichter zugängliche Umwelt (Floridi 2015) rechnen.

Insgesamt wirkt sich dies auf den Ort und die Aufgabe des Menschen in den existierenden und möglichen Kontrollschleifen aus, deren Sinn sich nicht im Einordnen, Regulieren, Verwalten und Absichern der Handlungswirklichkeit erschöpft, sondern auch darauf zu beziehen ist, ob und wie sich Praktiken und Lebensformen in den konkreten Bereichen des Zusammenwirkens von Menschen mit Maschinen *sinnvoll fortsetzen.* Das Zusammenwirken zwischen Menschen und intelligenten Systemen wäre dann entweder systemtheoretisch unter dem Aspekt der Kommunikations- und Kontrollbeziehungen zwischen einander fremden, komplexen Systemen bzw. der „System-Partnerschaft“ (Liggieri und Müller 2019) zu konzeptionieren oder handlungstheoretisch unter den Aspekten der „Kooperation“, „Koaktion“, „Kollaboration“ und „Arbeitsteilung“. Dabei ist mit Variablen, Latenzen und Täuschungen zu rechnen und zugleich von deren grundsätzlicher, wenn auch nicht praktikabler Berechenbarkeit auszugehen. Jedenfalls würde eine rein zweckrationale Betrachtung intelligenter Maschinen als Werkzeuge oder als Automaten zu kurz greifen. Schlüssiger mit Blick auf das Zusammenwirken von Menschen mit KI-Maschinen ist die Einordnung letzterer als künstliche „Systeme“ im Sinne der Systemtheorie, deren Kommunikations- und Kontrollbeziehungen es zu erfassen und zu analysieren gilt.

Dabei zeigt sich, dass sich das Zusammenwirken von Menschen und KI-Maschinen in menschliche Lebensformen sowie in Handlungs- und Gegenstandsbereiche hinein erstreckt, die sich nicht durchgängig regelförmig darstellen lassen (Schneider 1996). Dieser Aspekt hat mit der Entwicklung und breiten Verfügbarkeit der generativen Sprach- und Bildmodelle hohe Aktualität, da der Zugang zu diesen Technologien im Sinne ihrer allgemeinen Nutzbarkeit umfassend demokratisiert ist. Denn durch KI erzeugte natürliche Sprache und Bilder sind in allen Lebensbereichen im Zusammenwirken von Maschinen und Menschen direkt und ohne weitere Codierung interpretierbar. Es liegt eine gewisse Ironie darin, dass diese sonst oft geforderte Zugänglichkeit der Technologie das Kontrollproblem so stark verschärft, fällt doch die übliche „Selbstregulierung“ durch in der Regel hohe praktische Hürden in der Anwendung neuer Technologien schlicht weg. Zusätzlich gibt es häufig nicht mehr ein spezifisches, abgrenzbares, regulierbares Anwendungsgebiet, vielmehr lassen sich die KI-Systeme relativ nahtlos in die Lebenswelt und den Alltag einbinden und es entstehen mit großer Geschwindigkeit neue Praktiken im Umgang mit ihnen, die die Lebenswelt neu prägen. Die EU

hat in ihren Bemühungen zur Regulierung dies schon anerkannt und eine Klassifizierung von KI-Systemen als sogenannte „foundation models“ eingeführt. Das sind solche, die grade nicht über bestimmte Anwendungskontexte reguliert werden sollen, sondern auf Basis einer wie auch immer gearteten Werteabwägung, die (noch) nicht genauer spezifiziert ist. In gewisser Weise wird hier die Hilflosigkeit einer externen Kritik sichtbar, da diese Kategorie offensichtlich den für die externe Kontrolle so wichtigen Rahmen eines regelhaften Anwendungsbezuges unterläuft.

Doch selbst dort, wo es dennoch möglich ist, solche Bereiche abzustecken, ist es offen, ob und wie weit diese Regeln der Struktur der Wirklichkeit und den mit ihnen verbundenen Praktiken und Lebensvollzügen entsprechen bzw. ob nicht eine rein pragmatische Ontologie der Sache angemessener ist (Dreyfus 1993). Wir gehen daher davon aus, dass die Frage nach der „Kontrolle“ im Zusammenwirken von Menschen und KI-Systemen auch systemisch (orientiert an der Anschlussfähigkeit der Systemselektionen) und pragmatisch (orientiert an der sinnvollen Fortsetzung von Lebensformen und Praktiken) beantwortet werden muss. „Kontrolle“ ist dabei spezifisch als lernender und ein eigenes Systemgedächtnis im Umgang mit unberechenbarer, fremder Komplexität aufbauender und rekursiver Vorgang zu verstehen. Sie ist eine selbsttätige Steuerungsoperation des Systems, die demselben nicht äußerlich ist und sich als Operation im Zusammenwirken von Menschen und Maschinen im Kontext von Praktiken zeigt.

Dadurch stellt sich dann unmittelbar die Frage, womit wir es bei KI-Maschinen eigentlich zu tun haben, da es sich im kybernetischen Sinne eben nicht um einfache Werkzeuge handelt, die extern zu kontrollieren wären. Diese Frage schwingt implizit auch in vielen Kritiken von KI mit. So geht der oben genannte Turingpreisträger Geoffrey Hinton, der seine herausragende Karriere in der künstlichen Intelligenz auf der Motivation aufgebaut hat, biologische Intelligenz zu verstehen, nun selbstverständlich davon aus, dass es sich bei KI-Systemen um intelligente Maschinen handelt, die jedoch „anders intelligent“ sind als wir Menschen: „What’s happened to me is understanding there might be a big difference between this kind of intelligence and biological intelligence.[5]“ (Hinton 2023). Was dieses für das Verständnis von Kontrolle und Zusammenwirken von Menschen und Maschinen bedeutet, bleibt jedoch offen. Wir werden dieser Frage im nächsten Kapitel etwas weiter nachgehen.

Die Vorbehalte gegenüber systemtheoretisch unterbestimmten Kontrollforderungen im medial vorherrschenden Diskurs über Maschinenethik betreffen analog

[5] „Mir ist plötzlich aufgegangen, dass es einen großen Unterschied zwischen dieser Art von Intelligenz und biologischer Intelligenz geben könnte“.

auch die Verantwortungs- und Rechtfertigungslogik, die Expertengremien wie der Deutsche Ethikrat als vorrangig in der Mensch-Technik-Interaktion sehen. Bedürfen doch schon die generellen Anforderungen der Verantwortungsethik an die Fähigkeit von Menschen, ihr Leben zu führen, zu erfassen und in den Griff zu bekommen, einer kritischen Prüfung (Heidbrink 2022). Es kann in der Ethik nicht nur darum gehen, das Zusammenwirken von Menschen und KI-Maschinen in die bestehende „Ordnung der Dinge“ mit ihren Überwachungs- und Disziplinierungspraktiken einzuzeichnen (Foucault 2003), es gleichsam moralisch abzusichern und zu verwalten. Vielmehr ist – wie bereits angedeutet – auch zu fragen, welchen Beitrag das Zusammenwirken von Menschen und intelligenten Maschinen zur Fortsetzung lebensweltlicher Praktiken, zu deren Erschließung und zu deren Transformation leistet. Die ethische Auseinandersetzung mit KI-Systemen kreist dann nicht nur um die „Verantwortbarkeit“ ihres Einsatzes als Handlungsmittel, der an moralischen Prinzipien wie beispielsweise „Autonomie“ und „informationelle Selbstbestimmung“, „Fairness“, „Nicht-Schaden“ oder „Transparenz“ gemessen wird. Die ethische Reflexion richtet sich vielmehr auch auf die Frage, ob das Zusammenwirken von Menschen und KI-Systemen menschliche Lebensformen und die mit ihnen verbundenen Praktiken fortsetzt und neue Handlungsmöglichkeiten eröffnet. Ethisches Urteilen kann nicht darauf reduziert werden, Phänomene moralisch zu verbuchen, sondern hat auch eine explorative, die Lebenswelt und ihre konstitutiven Lebensformen als eine gemeinsame Welt erschließende Aufgabe (Arendt 1954/1994).

Eine solche erweiterte, auch Formen der immanenten Kritik beinhaltenden Analyse des Zusammenwirkens richtet sich wie auch die jüngere Diskussion über die kritische Theorie auf die Entdeckung und Konstitution der Wirklichkeit als einer „gemeinsamen Welt“ (Schauer 2023). Diese entsteht nicht „autopoietisch“, sondern wird erst durch bestimmte Praktiken der Verständigung, des Unterscheidens und des Urteilens präsent. Sie wird erfahrbar als eine geteilte Welt von Lebewesen, die gemeinsam mit anderen eine Lebenswelt bewohnen, die eine Geschichte haben, in der sie sich ihren Ort in der Welt erschließen und in der sie Entscheidungen treffen, mit deren Folgen sie weiterleben müssen. Ethisch urteilen heißt, auf diese gemeinsame, von Menschen und anderen Entitäten geteilte Welt hin zu urteilen, damit sich gangbare, tragfähige und auch neue Wege des Zusammenlebens und Zusammenwirkens abzeichnen und Gegenstand der Verständigung werden können (Ulrich 2012).

Mit dieser explorativen Aufgabe von Ethik verbinden sich technische Entwicklungsaufgaben, die nicht nur auf die Beherrschung und Kontrolle von KI-Systemen und auf die sinnvolle Fortsetzung von Lebensformen zielen. Es stellt sich auch die Frage, welche Formen des Zusammenwirkens von Menschen und

KI-Systemen zur Konstitution, Entdeckung und Erneuerung einer gemeinsamen Wirklichkeit beitragen können, in der Menschen herausfinden können, welche Lebensformen und damit verbundene Praktiken im guten Sinne zu ihnen gehören. Es geht dann nicht nur darum, zu kontrollieren und fortzusetzen, was geschieht, sondern auch darum, Widerstände und Differenzen in der Wirklichkeit aufzuspüren und die Urteilsbildung neu in Gang zu setzen. Dies impliziert die Bildung von Urteilen, in denen Menschen zusammenfinden und herausfinden, was ihr Ort in der Welt und ihre gemeinsame, geteilte Geschichte ist. Und zwar zunehmend auch in einer digitalen Welt, in der auch künstliche Systeme allgegenwärtig sind.

Dieser Ansatz verweist auch auf die Rolle des als „Leib" verstandenen Körpers, der den Ort in der Welt wesentlich mitbestimmt und mit ihm auch unsere Intelligenz und unser Wahrnehmen, Denken und Urteilen. In der Diskussion um die Entwicklung der künstlichen Intelligenz und darum, inwiefern diese zum Verständnis biologischer Intelligenz beiträgt, wird dies schon lange unter dem Stichwort „embodiment" verhandelt, d. h. dem Prinzip, dass die Verkörperung eine große Rolle dafür spielt, welche Art von Intelligenz realisierbar ist und wie diese verstanden und künstlich reproduziert werden kann (Pfeifer und Bongard 2006). Welche Rolle also KI-Maschinen, versehen mit „anderer Intelligenz", beim ethischen Urteilen spielen können, wird noch Gegenstand der Erörterung sein. Zuvor ist aber die Frage zu stellen, womit wir es bei diesen Maschinen und den Operationen ihrer Algorithmen zu tun haben. Dazu werden wir die These entwickeln, dass es sich um Kontinuierungsmaschinen handelt, die zwar abwägen, aber (noch) nicht urteilen können.

3 Auf der Suche nach Analogien: Wege zum Verstehen künstlicher Intelligenz

Um die Frage zu behandeln, mit welcher Art von Intelligenz wir es bei KI-Maschinen zu tun haben und was dies für die Analyse des Zusammenwirkens bedeutet, nehmen wir wie im vorherigen Kapitel die Stellungnahme des Deutschen Ethikrats als Ausgangspunkt. Darin wird von einer kategorialen Differenz von Menschen und Maschinen ausgegangen und dargelegt, dass die künstliche Intelligenz nicht der menschlichen Intelligenz entspricht (oder, wie manche meinen, auch gar nicht entsprechen kann), da sich menschliche Intelligenz analog und nicht auf dem Umweg der digitalen Datenverarbeitung auf die Wirklichkeit als Lebenswelt bezieht. So sehr aber die Stellungnahme einer funktionalistischen Auffassung der menschlichen Intelligenz entgegentritt (ein Standpunkt, den wir teilen), so wenig klärt sie, womit wir es denn bei der spezifischen Intelligenz von

KI-Maschinen zu tun haben. Maschinen erscheinen in ihr als technische Mittel, die man entsprechend ihrer besonderen Leistungsfähigkeit auch im Bereich von Argumentation, Manipulation, etc. regulieren und kontrollieren kann und muss.

Die rekursiven Veränderungen menschlichen Verhaltens und Erlebens durch die Bereitstellung von (aufbereiteten) Daten und Vorschlägen für Entscheidungen, durch das Führen von Dialogen oder durch die Generierung von realistischen Bildern durch Maschinen, generell: die Medialität dieser Technologien der Datenverarbeitung (Hubig 2001), werden aber kaum ernst genommen. Hierbei schwingt ein im öffentlichen Diskurs häufig anzutreffender unzulässiger Umkehrschluss mit: daraus, dass Maschinen nicht auf die gleiche Weise intelligent sind wie wir, oder vielleicht prinzipiell wegen fehlender biologischer Verkörperung auch nicht urteilen können wie wir, wird geschlossen, dass sie in ihren spezifischen intelligenten Leistungen und dem daraus entstehenden komplexen Zusammenwirken mit Menschen nicht ernst zu nehmen und wie andere einfache Werkzeuge zu behandeln wären. Dies wird unserer Meinung nach der Frage danach, womit wir es mit KI-Maschinen zu tun haben, nicht gerecht. Ob es solche kategorialen und nicht aufzuhebenden Unterschiede zwischen Mensch und Maschine gibt oder worin solche genau bestehen könnten, kann dabei für unsere Reflektion der ethischen Diskussion unbeantwortet bleiben, denn die Kernfrage nach der spezifischen Intelligenz von KI-Maschinen bleibt davon unabhängig. Wir werden uns dieser Frage auf dem Weg der *Analogiebildung* nähern.

Dabei stellt sich die erkenntnisleitende Frage, inwiefern wir unsere aus der Lebenswelt vertraute Sprache auf maschinelle Prozesse übertragen können, z. B., indem wir davon sprechen, was KI-Maschinen gut „können“, was sie „erkennen“, „verstehen“ oder „tun“. In diesem Zusammenhang stellt sich auch die Frage, inwiefern sich die Metapher der „Intelligenz“, die der lateinischen Wortbedeutung nach eine „verstehende“, mit der Wirklichkeit in Kontakt und in Distanz tretende geistige Kraft ist (lat. intelligere = verstehen), auf die maschinelle Datenverarbeitung übertragen lässt. Es gibt keinen unmittelbaren, von der Sprache und ihren Metaphern und Analogien unabhängigen Zugriff auf die „Sache“, mit der wir es da zu tun haben. Wir operieren also mit übertragenen Redeweisen aus der uns vertrauten begrifflichen und bildlichen Sprache, die ausgerechnet deshalb so gut zur Verständigung in der Lebenswelt dient, weil sie unscharf ist (Wittgenstein 1953/2003, § 71). Primär wäre daher auf der Ebene der Sprache zu klären, ob und inwiefern die metaphorische Sprechweise mit Blick auf die formal-logischen Vorgänge des Maschinenlernens angemessen ist. In einem zweiten Schritt wäre zu fragen, ob sich diese Sprechweise zu einem Modell oder vielleicht sogar zu einer Theorie über die Eigenschaften von KI und deren spezifische „Intelligenz“ erweitern lässt (Schneider 2018, 523). Die Sachfrage sollte

aber nicht ohne Klärung der Frage nach unseren sprachlichen Mitteln erfolgen, weil sich in der Sprache das grundlegende Problem einer digitalen Formalisierbarkeit der Lebenswelt bzw. der grundlegenden Unterscheidung von Systemlogik und Lebenswelt spiegelt. Wir müssen Formal- bzw. Begriffssprachen verwenden, um über die Datenverarbeitungsvorgänge von KI-Maschinen und ihr „Lernen" zu sprechen, ultimativ sind die zugrundeliegenden Algorithmen ja mathematische Verfahren und in der abstrakten Sprache der Mathematik formuliert. Wir können diese aber nicht vollständig in natürliche Sprachen übertragen, da sich in der natürlichen Sprache die kalkulierbare grammatische Form und die nicht berechenbare und sinnvolle Sprechpraxis miteinander verschränken (Schneider 1992). Es bleibt aber die Möglichkeit, Analogieschlüsse zu ziehen zwischen der Welt der Algorithmen und der Lebenswelt (Nehaniv 1999). Dabei gilt es, Ähnlichkeiten und Unterscheidungen zu suchen, um zu verstehen, womit wir es bei den Systemdynamiken von KI-Maschinen zu tun haben, wie sich die Lebenswelt und ihre kommunikativen Praktiken durch die „Digitalisierung" verändern, wie die Wirklichkeit im Medium der algorithmischen Systemlogiken erscheint, welche Aspekte der Wirklichkeit gewonnen werden oder verlorengehen und was angemessene Formen des Zusammenwirkens sind.

Zu beachten ist bei der Bildung und der Kritik solcher Analogien, dass Menschen generell dazu neigen, Maschinen zu anthropomorphisieren und ihnen menschliche Eigenschaften zuzuschreiben (Heuser und Thies 2021). Sehr ausgeprägt ist dies bei Robotern, denen sehr häufig schon durch ihr Aussehen (Krach et al. 2008) und ihre Bewegungsfähigkeiten Intelligenz und Intentionalität unterstellt werden (Steil und Manzeschke 2023). KI-Maschinen, insbesondere Chatbots, werden auch zunehmend Eigenschaften wie Bewusstsein zugeschrieben: der sog. ELIZA-Effekt (Hofstadter 1998). Es wird auch argumentiert, dass eine bedeutungsvolle soziale Mensch-Maschine Interaktion sogar immer mit Anthropomorphisieren verbunden sein muss (Duffy 2003). Der Versuch, solche impliziten Zuschreibungen und die damit verbundenen Analogiebildungen ganz zu umgehen, wäre wohl nicht zielführend.

Bleiben wir zur Konkretisierung der Analogiebildung bei der Rede von „Bedeutung". Eine viel diskutierte Annahme, die auch der Hypothese der Relevanz von „embodiment" (Verkörperung) für Intelligenz zugrunde liegt, ist, dass KI-Maschinen mangels eines verkörperten, praktischen Weltverhältnisses zwar Informationen verarbeiten, aber keine Bedeutung verstehen können. Darüber hinaus wurde Bedeutung häufig – und wie wir heute wissen fälschlicherweise – damit identifiziert, semantisch korrekt mit den entsprechenden, die Bedeutung beschreibenden Begriffen umzugehen. Durch generative Sprachmodelle, die sogenannten „large language models", wird diese Annahme, Bedeutung und Sprache

wesentlich zu identifizieren, brüchig. Denn diese sind gerade darauf trainiert, Bedeutung statistisch zu erfassen, d. h. semantische Beziehungen arithmetisch zu korrelieren, so dass sie sehr erfolgreich darüber kommunizieren können. Und auch die Hypothese, dass Bedeutung eine biologische oder zumindest physische Verkörperung der Entität braucht, für die die Bedeutung bestehen soll, und dass Bedeutung in diesem Sinne erst durch den physischen Weltbezug und damit durch den Ort und die Historie eines physischen Individuums erzeugt wird, wird brüchig. Gegeben die dialogische Interaktion und damit verbundene Manipulationsmöglichkeiten des Gegenübers, ist durchaus zu diskutieren, ob nicht eine quasi indirekte Verkörperung der maschinellen Intelligenz durch seine menschlichen Interaktionspartner gegeben sein könnte. Das Szenario, dass eine Maschinenintelligenz durch Menschen gezielt Manipulationen in der Welt ausführen könnte, die ihren Zielen dient, ist realistisch. Die Maschine könnte beispielsweise jemanden „überreden", für sie Zugang zu spezifischen sonst gesperrten Daten zu schaffen oder einschlägige CAPTCHAs zum Login in eigentlich unzugängliche Webseiten für sie auszuführen. Letzteres ist auch sprachlich interessant, ist doch CAPTCHA die Abkürzung für „Completely Automated Public Turing Test to Tell Computers and Humans Apart", also einen Test, der im Wortsinne Menschen von Maschinen gerade unterscheiden soll. Das zeigt, wie bei der Analogiebildung mit großer Vorsicht vorzugehen ist, da zahlreiche fundamentale, Differenzen eröffnende Begriffe durch die aktuelle Technologie auf neue Weise in Frage gestellt werden, so dass nicht nur alte Diskussionen über diese plötzlich wieder hoch aktuell werden, sondern sich auch drastische Verschiebungen im Verständnis dieser ergeben.

Um trotzdem die Ähnlichkeit zwischen den Merkmalen von „Bedeutung" in menschlichen und maschinellen Domänen der Welterschließung per analogiam zu bestimmen, müssen wir deren Übereinstimmung bzw. Differenz hinsichtlich ihrer jeweiligen Funktion oder Struktur darlegen. Wenn wir davon ausgehen, dass das Wort „Bedeutung" in beiden Anwendungsbereichen etwas Ähnliches meint, schließen wir von menschlichem Bedeutungsverstehen auf die Art und Weise, wie Maschinen Bedeutung verarbeiten. Maschinelle Prozesse hätten demnach hinsichtlich ihrer Funktion bzw. ihrer Struktur *etwas von* menschlichem Weltverstehen, sie hätten also *mehr als keine* Bedeutung. Es wäre dann gerechtfertigt, über generative Sprachmodelle zu sagen, dass sie mit virtueller Bedeutung arbeiten, ohne ihnen deshalb gleich menschliche Eigenschaften zu unterstellen. Die Analogiebildung wird dabei dadurch begrenzt, dass die Unähnlichkeit zwischen maschineller und menschlicher Semantik immer größer als die Ähnlichkeit ist. Ihre Ähnlichkeit rechtfertigt es aber, das Wort „Bedeutung" analog zu verwenden.

Der Literaturwissenschaftler Hannes Bajohr hat auf der Suche nach analogen Begriffen für die Art der Informationsverarbeitung von generativen Sprachmodellen vorgeschlagen, von der Produktion „dummer Bedeutung“ zu sprechen (Bajohr 2022). Diese prozessiere die Korrelationen von Wörtern in der Sprache nach dem Modell eines mehrdimensionalen Raums von syntaktischen und semantischen Beziehungen. Als „dumm“ bezeichnet Bajohr diese Form von datenimmanenter „Bedeutung“, „weil das Sprachmodell zwar latente Korrelationen zwischen Zeichen erfasst, aber immer noch nicht »weiß«, welche Sachen diese Zeichen eigentlich benennen; mit dieser Art von Bedeutung wird man keine Intelligenz bauen können, die sich je in der Welt zurechtfindet“ (Bajohr 2022, 74). Allerdings können diese Modelle, ohne die Bedeutung von Zeichen zu verstehen, Bedeutungen in Textstrukturen aufzeigen, die ohne die Technologie für Menschen nicht zugänglich wären. Es kommt daher im Zusammenwirken der natürlichen Sprache der Nutzenden und artifizieller, „dummer“ Bedeutung maschineller Intelligenz an der Schnittstelle der Spracheingabe (des sogenannten „Prompts“) darauf an, dass nicht nur die Maschine semantische Korrelationen „lernt“, sondern auch, dass Menschen lernen, die Stärken der „dummen“ Bedeutung mittels intelligenter Spracheingaben zu nutzen. Andererseits verschiebt Bajohr hier das Problem nun darauf, dass postuliert wird, Intelligenz bestehe darin „sich in der Welt zurechtzufinden“ und dass dies der Maschine im umfassenden Sinn nicht möglich sei. Auch wenn Bajohr dies differenzierter sieht, klingt hier ein den obigen ähnlicher und in der allgemeinen Diskussion häufig anzutreffender unzulässiger Umkehrschluss an: aus der Tatsache, dass Maschinen Bedeutung anders erfassen als Menschen, folgt nicht, dass sie unintelligent sind. Und auch nicht, dass sie sich prinzipiell nicht in der Welt zurechtfinden könnten, wie die Beispiele oben zeigen. In der sprachlichen Welt gelingt dies jedenfalls schon recht gut, auch wenn einiges davon auf das intelligente Verwenden des Prompts durch Menschen zurückgeht. Im Sinne der Analogiebildung wäre daher vielleicht der Begriff „virtuelle Bedeutung“ vorzuziehen, da dieser den durch „dumm“ konnotierten Umkehrschluss nicht so stark impliziert.

Nicht erfasst wird von dem Ansatz der „dummen Bedeutung“, dass auch jetzt schon die Sprachmodelle im Sinne eines Meta-Lernens von ihrer Interaktion mit Menschen profitieren und mit Hilfe von Verstärkungslernen ihre Textgenerierung in die Richtung verbessern, wie es Nutzende durch Feedback vorgeben. Zusätzlich entstehen ständig neue Daten durch dialogisches Zusammenwirken von Menschen und Maschinen, die selbst wieder in die Datenbasis eingehen und Teil der Korrelationen werden. Diese können aber zumindest nicht vollständig nur „dumm“ sein, da sie sowohl reflexiv sind als auch die Bedeutungen der menschlichen Partner tragen, mit denen sich die Kombination von Menschen und

Maschinen durchaus in der Welt zurechtfinden könnte, und vielleicht auch die Maschine allein mit Hilfe der dem Menschen durch Zusammenwirken „entliehener“ Bedeutung. Dass sie das dann wiederum nicht genau so tut wie wir, ist dann aber nur eine weitere Iteration der Infragestellung von Selbstverständlichkeiten im Hinblick auf eine „andere künstliche Intelligenz“.

Klar ist jedenfalls, dass zahlreiche postulierte Unmöglichkeiten des Typus „Maschinen werden nie …“ in sich zusammengefallen sind. So wurde beispielsweise lange vermutet, dass insbesondere die enorme Kompositionalität von Sprache und unserer Symbolsysteme, deren Beherrschung von jeher als besonderes Zeichen von Intelligenz galt, eine Alleinstellung von Menschen im Vergleich zu anderen Tieren und Maschinen begründet. Genau solche Symbole zu manipulieren ist jetzt aber eine Stärke der KI. Und wo z. B. genau der Übergang von sehr überzeugend geschickt emulierender Statistik zu echtem Argumentieren ist, ist im Bereich des logischen Schließens nicht mehr so leicht zu bestimmen. Die Maschinen „können“ letzteres zunehmend gut, auch für Aufgaben, die nicht im Training vorkommen, und verbessern sich atemberaubend schnell.

Wir stehen also vor der Frage, wie wir die spezifische Intelligenz solcher Maschinen in ihrem Zusammenspiel mit menschlicher Intelligenz jenseits der auf Differenz abhebenden und damit wenig konstruktiven Analogiebildung kritisch beurteilen können. Diese Überlegungen verweisen auf ein Verständnis der KI, das sich weniger am Vergleich mit menschlicher Intelligenz orientiert. Dabei ist aber natürlich im Hintergrund zu beachten, dass auch ein solches Verständnis unvermeidbar nur mit den Mitteln unserer Sprache und Kognition durchzuführen ist, die bestimmte Festlegungen und Analogien (anthropomorphisierend) nahelegt.

Wenn wir vor diesem Hintergrund wieder zur Frage nach einer um Formen immanenter Kritik erweiterten Analyse des Zusammenwirkens von Menschen und KI-Maschinen zurückkehren, zeigen sich die Grenzen einer systemtheoretischen Rekonstruktion dieser Wirklichkeit. Indem die Systemtheorie die Wirklichkeit rekonstruiert, als bestünde sie aus autopoietischen, sich durch intelligente Selektionen selbst erhaltenden Systemen und deren Umwelten (Luhmann 1987), baut sie eine Brücke zwischen dem modus operandi von KI-Maschinen, gesellschaftlichen Systemen und unseren sprachlichen Darstellungsmitteln. Sie berücksichtigt aber nicht hinreichend die Differenz von digitaler und analoger Wirklichkeit bzw. von artifizieller und menschlicher Semantik, insofern sich letztere aus einer in der Welt situierten Teilnahme an gemeinsamen Lebensformen und Praktiken ergibt. Zu deren sinnvollen Fortsetzung gehört es, die Standpunkte der beteiligten oder betroffenen anderen Personen imaginativ oder in artikulierter Form selbstreferentiell zu berücksichtigen, d. h. den eigenen Ort in der mit Anderen geteilten Welt zu verstehen. Unter der Voraussetzung, dass die Algorithmen von

KI-Maschinen ihre Aufgaben dadurch erfüllen, dass sie Datenräume statistisch auf Muster hin erfassen, rechnen sie mit einer gänzlich systemisch strukturierbaren Wirklichkeit, einer Wirklichkeit als „Infosphäre“ (Floridi 2015). Sie rechnen aber nicht mit der Wirklichkeit von Systemen in einer Lebenswelt, die nicht nur wahrscheinlichkeitsbasiert fortzusetzen, sondern quer zur Systemlogik auch immer wieder neu als gemeinsame Welt zu entdecken und zu bewohnen ist. Ihre enorme Leistungsfähigkeit ist erst durch eine systemische Reduktion und Datafizierung der Wirklichkeit möglich. Die Algorithmen durchforsten nicht die erfahrbare, analoge Wirklichkeit, sondern die Daten, die aus ihr gewonnen bzw. abgeleitet bzw. ihnen von ihren Entwicklern als Trainingsdaten zugeführt werden und die einerseits auf die Selektionsleistungen selbstreferentieller Systeme und andererseits auf deren Vernetzungsleistungen zurückgehen. Eine Analogie zu dieser Unterscheidung findet sich freilich auch in der menschlichen Sprache als einem „digitalen“ Medium der Benennung, Wahrnehmung und Erkenntnis einer nicht-digitalen „Wirklichkeit“ (Watzlawik et al. 2011).

Die KI erfasst also die Intelligenz von sich selbst reproduzierenden und sich dabei selektiv in ein Verhältnis zu den komplexen Möglichkeiten ihrer Umwelt setzenden Systemen in ihrem Wechselverhältnis zur Umwelt anderer Systeme in Form von Daten. Deren Korrelationen und Muster sind aufgrund der Größe der Datenmengen von menschlichen Beobachtern teilweise nicht zu erkennen, können so aber in Gebrauch genommen werden. Der Systemtheoretiker Dirk Baecker hat diese maschinelle „Fähigkeit zur Inanspruchnahme fremder Komplexität“ als „virtuelle Intelligenz“ bezeichnet (Baecker 2019, 46), die von den Algorithmen der KI-Maschinen aufgespürt und dargestellt werden kann. Auf der Basis großer Datenvolumen errechnen sie Korrelationen, die den Zusammenhang von Ereignissen in der „Welt“ der Systeme sowohl als System- als auch als Vernetzungsleistungen in Form von „Mustern“ sichtbar, erklärbar und partiell vorhersagbar machen (Nassehi 2019). Es ist aber weitgehend unumstritten, dass diese Korrelationen allein noch keine Bedeutung konstituieren und für intelligentes Verhalten, im oben diskutieren Sinne des „Zurechtfindens in der Welt“ (Bajohr), nicht ausreichen. Die Grundlage dieser „virtuellen“ Intelligenz bilden nun die Intelligenzen, mit denen sich die beteiligten Systeme selbst erhalten und zugleich auf andere beziehen und zunehmend, wie oben diskutiert, die durch das Zusammenwirken solcher virtuellen und anderer systemischer Intelligenz entstehenden hybriden Intelligenzen. In W. Ross Ashbys Definition von Intelligenz als „angemessener Selektion“ klingt die wichtige Rolle dieser Informationsbasis bereits an: „The ‚intelligent‘ processes par excellence are the goal-seeking – those that show high power of appropriate selection. Man and computer show their powers alike, by appropriate selection. But both are bounded by the fact that appropriate

selection (to a degree better than chance) can be achieved only as a consequence of information received and processed." (Ashby 1961, 275). Modernere Theorien kognitiver Systeme gehen ebenfalls davon aus, dass stabile, selbsterhaltende Prozesse notwendig sind, um einen Grad an Autonomie zu erreichen, die es erlaubt, in der Umgebung zu erkennen, was relevant ist und darauf selektiv zu reagieren (Di Paolo 2005; Thompson 2007). Dieses wird dann als notwendig erachtet, um eine Verankerung von Werten, Normen und Zielen auch von künstlichen Agenten in der realen Welt zu erreichen und verweist ein weiteres Mal auf die Frage der Verkörperung von Intelligenz. Allerdings wird dabei davon ausgegangen, das solche Selbsterhaltung nicht unbedingt biologische, metabolische Prozesse umfassen muss, sondern dass auch schon stabile sensomotorische Muster in der Interaktion mit der Umwelt die Kriterien für eine stabiles Netz interner Prozesse bieten können, auf die dann eine adaptive, auf Bedeutung gerichtete Selektion aufbauen kann (Ramírez-Vizcaya und Froese 2020). Die Systemtheorie formuliert dies im Sinne abstrakterer Systeme und ihrer Operationen, so dass sich im Kontext der generativen Sprachmodelle agentenbasierte Operationen hin zu stabilen Mustern in der sprachlichen Interaktion bilden, auf Basis derer so etwas wie virtuelle Bedeutung entstehen kann.

Ein weiteres wichtiges Merkmal in Ashbys Definition ist die Orientierung an Zielen („goal-seeking"), die der Selektion zugrunde liegt, die in den Theorien intelligenter Agenten ebenfalls ganz wesentlich und zunehmend von einer physischen Verkörperung losgelöst eher in dynamischen Prozessen gedacht wird. Selektion in Hinblick auf Ziele ist notwendig, da nicht alle sensorischen Informationen und potentiellen Aktionsmöglichkeiten mit ihren Folgen in einer offenen Welt berechnet und evaluiert werden können. Ein wesentliches Merkmal zur Beurteilung künstlicher Systeme ist daher neben der Frage, ob sie Kontrolle im Sinne stabiler Operationen haben, ob – und wenn ja – welche Ziele ihren Selektionen unterliegen, auch wenn diese nur implizit und damit schwer aufzudecken sein können.

Ein Beispiel für sprachlich vermittelte, mit der Lebenswelt korrelierte aber nicht direkt verbundene zielgerichtete Selektion ist die Konstruktion und Erzählung der menschlichen individuellen Lebensgeschichte, die sich über die Zeit verändert, zwar in Fakten gründet, diese aber nicht vollständig abbildet und vor allem mit dem Ziel, „Kohärenz" und „Sinn" zu erzeugen immer wieder um- und weitergeschrieben wird. Ob so eine Selektion in ein Sprachmodell als explizites oder implizites Ziel einprogrammiert werden kann, ist nicht klar. Der Versuch würde vermutlich aber zumindest zu weiteren überraschenden Leistungen führen, geht aber über unsere Überlegungen zum Begriff der Kontrolle von KI hinaus.

So sehr diese kybernetische Bestimmung von Intelligenz als „power" im Sinne des „Vermögens" bzw. der „Fähigkeit" einer Entität zutrifft, so unterbestimmt lässt sie aber einen weiteren, für unsere Fragestellung wichtigen Aspekt von Intelligenz: „Intelligenz" ist auch eine Funktion des Zusammenwirkens verschiedener Entitäten, nicht nur die Summe ihres jeweiligen Systemvermögens. Selbst unter der Annahme, dass verkörperte künstliche Intelligenz oder anderweitig durch stabile Systemprozesse charakterisierte Agenten (virtuelle) Bedeutung durch eine „grounding" in ihrer individuellen Geschichte, Struktur und Dynamik erzeugen (Varela 1979; Oyama 2000; Thompson 2007), und auf der Basis selektierend (also: auswählend bzw. auslesend) ein Ziel verfolgen könnten, ist nicht klar, dass sich dieses auf das Zusammenwirken solcher Systeme mit Menschen unmittelbar ausdehnen würde. Letzteres würde eine Aushandlung gemeinsamer Ziele erfordern. Denn zu menschlicher Intelligenz gehört auch eine gemeinsame Praxis des Unterscheidens, Verstehens und Urteilens, die die Faktizität der Weltzusammenhänge und ihre Sinngehalte nicht nur fortsetzt, sondern eine neue, gemeinsame Welt und mit ihr: einen neuen Fortgang der Geschichte im Zusammenspiel der unterschiedlichen Weltzugänge erkundet und konstituiert (Arendt 1970/1998).

Zunächst ist festzuhalten: Die Algorithmen von KI-Maschinen operieren mindestens mit drei verschiedenen Dimensionen von Intelligenz: Zum ersten mit der Fähigkeit zur Mustererkennung, die menschliche Kapazitäten weit übersteigt, und einer selektiven Anschlusskommunikation, die ihnen von ihren Herstellern zur Erfüllung bestimmter Aufgaben einprogrammiert wurde. Zum zweiten mit der Intelligenz der Strukturen der Wirklichkeit, die in Daten virtuell zugänglich ist und die die Selbsterhaltung und Selektion von Systemen selbst wieder als Muster enthält. Diese Muster spiegeln dabei die Selektionsleistungen eines komplexen Netzwerks selbstreferentieller Systeme wider. Und zum dritten mit der Intelligenz, die durch die Rückkopplung der errechneten Daten mit der analogen Wirklichkeit ermöglicht und in Form neuer Selektionsmöglichkeiten von Akteuren und Systemen realisiert wird. Diese Rückkopplung erfordert nicht zwingend einen eigenen Körper, soweit indirekte Manipulationen der analogen Welt möglich sind. Und sie ist selbst wieder Grundlage von datengetriebenem Lernen auf einer Metaebene, z. B. durch Rückkoppeln von Bewertungen, was als angemessenes oder sinnvolles Verhalten im Zusammenwirken mit Menschen von diesen so empfunden wird. Solches Lernen ist nicht hypothetisch, sondern real in den existierenden generativen Sprachmodellen implementiert und dient dazu, insbesondere die Ausgaben hin auf erwünschte Aussagen zu filtern.

Eine erweitere Analyse des Zusammenwirkens von Menschen und KI-Maschinen muss an allen drei dieser Dimensionen von Intelligenz ansetzen:

- Erstens an den Zwecken, denen die Algorithmen dienen sollen und an den Trainingsbedingungen, unter denen die Algorithmen für diese Zwecke zugerüstet werden.
- Zweitens an der Differenz zwischen der digitalen System- und Netzwerklogik, mit der KI-Maschinen rechnen, und der analogen, widersprüchlichen und unscharfen Wirklichkeit der Lebenswelt.
- Drittens an den rekursiven Effekten der durch die Algorithmen erschlossenen Daten auf die im Zusammenwirken betroffenen Systeme und Akteure.

Vor diesem Hintergrund werden wir uns im Folgenden vorrangig mit dem Zusammenwirken von Menschen mit generativer KI (wie beispielsweise ChatGPT) auseinandersetzen.

4 Auf der Suche nach Anschluss: Generative KI als Kontinuierungsmaschine

Betrachtet man den modus operandi von Algorithmen des Maschinenlernens, nämlich die Codierung und Speicherung aller möglichen Inhalte in einem binären Code sowie die Suche nach Korrelationen in großen Datenräumen, dann fällt ins Auge, wie sehr diese Systeme dafür konzipiert und geeignet sind, Anschlussmöglichkeiten zu realisieren und auszuweiten. Anders als gesellschaftliche Systeme, die sich durch die Anschlussfähigkeit ihrer spezifischen kommunikativen Codes reproduzieren (Luhmann 1987), sind KI-Systeme für Daten *grundsätzlich* anschlussfähig bzw. permanent auf der Suche nach Anschluss. Die jeweiligen Zwecke, für die sie programmiert werden, mögen differieren. Auch werden die Trainingsdaten von den Entwicklern auf Verzerrungen hin massiv gefiltert und selbst wieder ihren Zwecken entsprechend verzerrt. Einen Zweck, oder besser: eine Funktion aber haben sie gemeinsam: Sie sollen die System- und Vernetzungsleistungen der unterschiedlichen Systeme durch die Analyse der Korrelationen der von den Systemen der Lebenswelt hervorgebrachten und in riesigen Datenräumen zugänglichen Ereignisse entdecken und rekursiv verfügbar machen. Dies ist besonders deutlich bei den generativen Systemen, die Sprache, Bilder oder auch Musik erzeugen und mittlerweile auch multimodale Artefakte synthetisieren (Samad 2023), gilt aber auch für zahlreiche datenaufbereitende und entscheidungsunterstützende Systeme. Diese Funktion hat, trotz aller technischen Fortschritte, inhärente Beschränkungen, so z. B. das Problem, das mit der Anzahl zusätzlicher Informationskanäle auch die Komplexität in solchen Systemen explodiert. Und damit sind die KI-Algorithmen nicht „neutral“

der Datenwelt gegenüber, da sich im Design und der Implementation immer zahlreiche und dem eigentlichen Algorithmus externe Entscheidungen, Rahmenbedingungen, technische Beschränkungen, implizite Annahmen und Zwecke, und nicht zuletzt ökonomische Interessen der Entwickler abbilden. Die im Prinzip universelle Anschlussfähigkeit, die durch das digitale Format bedingt ist, bleibt aber trotz dieser Einschränkungen erhalten.

Man könnte die damit verbundene, allumfassende Digitalisierung der Gesellschaft auch als eine weitere Stufe der Ökonomisierung verstehen, die alle Widerstände gegenüber Verwertungsdynamiken beseitigt – wenn nicht das Medium der Digitalität in seiner Sachlogik indifferent gegenüber spezifischen rationalen Interessen wäre, so sehr sie auch anschlussfähig für ökonomische Logiken sind. Die Algorithmen der „künstlichen", oder besser: „virtuellen" Intelligenz aber zielen durch ihre möglichst allseitige Anschluss- und Konnexionsfähigkeit auf die Darstellung, Herstellung und Sicherung von Systemkontinuitäten weit über einfache Anwendungskontexte hinaus, gerade in den bereits erwähnten „foundation models". Dabei operieren diese im Modus permanenter mathematischer Berechnungen, die nicht nur die Statistik der Daten im Sinne wahrscheinlicher Fortsetzungen spiegeln, sondern auch sichtbar machen, was „die Welt im Innersten zusammenhält". Diese der Autopoiesis von Systemen entsprechende Kontinuierungslogik von In- und Outputvorgängen ist es, die eine systembezogene Kritik der derzeitigen Konfiguration einer generativen KI-Maschine eröffnet. Metaphorisch gesprochen, hält diese in ihren Beobachtungen und Berechnungen niemals inne, um zu urteilen, sondern verfolgt, wenn sie in Gang gesetzt wird, pausenlos ihr Ziel. Im Training durchforstet das Sprachmodell dazu die Daten des sprachlich kodifizierten Wissens und der sich im Internet spiegelnden Realität anhand der im Training umgesetzten Heuristiken, die Zwecke der Entwickler als Bias in das Lernsystem einbringen. Ergebnis sind statistische Korrelationen, die es dem System ermöglichen, syntaktisch und semantisch sinnvolle Fortsetzungen von Wörtern zu Sätzen und Texten stochastisch zu berechnen. So elaboriert, bezügereich und abgewogen solche Texte aber auch sein können, führt doch die Fortsetzungslogik dazu, dass das Modell auf jede Frage hin eindeutig antwortet oder sinnvoll klingende, aber mitunter erfundene, inkonsistente oder falsche Antworten und plausibel scheinende Unwahrheiten generiert. Die strikte Input–Output-Kopplung des Natural-Language-Processing scheint im Zusammenspiel mit seiner universalen Anschlussfähigkeit zu einer Art „Zugzwang" zu führen, die teilweise sinnvolle und konsistente Texte, teilweise aber auch einfach als sinnvoll präsentierten Unsinn erzeugt. Dieser Effekt von KI-Chatbots wird seit ihrer Einführung beobachtet und „Halluzination" genannt (Alkaissi und McFarlane 2023; Bang et al. 2023). Diese verbreitete Analogie trifft aber den Kern der

Kontinuierung eigentlich nicht, denn für die Maschine gibt es kein zugrundeliegendes Leben, in Bezug auf welches hin sich die Fortsetzung halluzinierend von der Realität unterscheidet. Echte Halluzination würde nicht-virtuelle Bedeutung voraussetzen.

Aber auch da, wo das generative Sprachmodell nicht „halluziniert", sondern seinem Zweck gemäß in Fakten begründbare Texte generiert, sind diese a) durch die Trainingsvorgaben der Entwickler:innen geprägt und bleiben b) rein enzyklopädische Beobachtungen, die Vorhandenes registrieren, reproduzieren, neu arrangieren und die Mittel der Kompositionalität der Sprache ausnutzend rekombinieren. Sie sind aber abgelöst von den Kontexten des Wahrnehmens, Verstehens, Urteilens und der Verständigung über das, was das Leben und Zusammenleben orientieren soll, kurz: vom Bezug auf das Leben in einer gemeinsamen Welt mit seinen bestimmten Formen und Praktiken. Das Phänomen entspricht dem, was Harry Frankfurt als „Bullshit" bezeichnet und als Kennzeichen von Kulturen beschrieben hat, in denen öffentlich Meinungen abgerufen, aber keine Urteile gebildet werden (Frankfurt 2014).

Der formale Zugzwang des texterzeugenden Algorithmus erzeugt demnach noch eine inhaltliche Problematik: Um nicht nur technisch, sondern auch diskursiv anschlussfähig zu bleiben, ist das Sprachmodell so programmiert, dass es sich gerade nicht so äußert, dass zumindest der Anschein eines Urteils erweckt wird. Um keine Anschlüsse auszuschließen, soll die simulierte Deliberation des Algorithmus nie in einer eindeutigen Geschmacks- oder Wert- oder politischen „Urteil" enden, sondern bleibt im Entweder-Oder und Sowohl-Als-Auch. Der Chatbot soll also Meinungen, Alternativen und Unterscheidungen aufzeigen, die im Rahmen konsensheischender Wertungen bleiben, aber keine Schlüsse auf eine bestimmte Wirklichkeit hin ziehen und keine Urteile über die Vorzugswürdigkeit bestimmter Entscheidungen und Festlegungen mit Blick auf eine bestimmte individuelle Geschichte treffen. Das Sprachmodell aggregiert Meinungen. Es entspricht damit der Dauerdeliberation jener Diskurse, die nicht in den Streit pluraler Standpunkte um die politische Frage eintreten, was die Welt ausmacht, in der wir gemeinsam leben können (Sauer 2021). Und dort, wo diese Programmierung und damit externe Kontrolle versagt und das Sprachmodell doch einmal eindeutige Aussagen macht, gilt dieses dann als „Fehlfunktion", die allerdings auch gar nicht immer auszuschließen ist, denn auch eindeutige Aussagen sind wahrscheinliche Fortsetzungen.

Dem Sprachmodell fehlt also eine sinnliche, welterschließende Kompetenz, die der Immanuel Kant eine „geheime Kraft" genannt hat: die *Urteilskraft* (Kant 1762/1969, 60). Aus dieser Kraft resultiert die Fähigkeit von intelligenten Lebewesen, Dinge nicht nur voneinander zu unterscheiden, sondern Unterschiede als

Unterschiede erkennen zu können (Wieland 2001, 15). Dies ist nach Kant keine reine Kognitionsleistung, sondern auch und vor allem eine sinnliche Leistung. Sie beruht auf einem inneren Sinn, mit dem die Gegenstände der Wirklichkeit innerlich, d. h. mit Abstand von der ursprünglichen Sinneswahrnehmung, reproduziert und beurteilt werden. Die Urteilkraft fügt sinnliche und begriffliche Elemente zusammen, d. h. sie bleibt, anders als das Denken, im engen Kontakt zu und doch unterschieden von der Wirklichkeit, die uns Menschen affiziert. In ihrer Auseinandersetzung mit Kant hat Hannah Arendt daraufhin den Gedanken entwickelt, dass das Urteilen als eine geistige Tätigkeit niemals weltlos bleibt, sondern uns mit anderen urteilenden Menschen verbindet, deren Standpunkte die Urteilskraft im Zusammenspiel mit dem je eigenen Weltzugang einbezieht (Arendt 1970/1998, 91). Arendts Pointe ist, dass uns ästhetische, auf der Wahrnehmung der uns affizierenden Wirklichkeit beruhende Urteile nicht etwa in eine abgeschlossene Subjektivität führen, sondern eine gemeinsame Welt des Zusammenlebens eröffnen – eine politische Welt, die nicht einfach vorhanden ist als das, was geschieht, sondern die sich erst durch wahrnehmende Urteile im Modus des Zusammenlebens erschließt und durch diese Urteile gemeinsam bewohnbar wird.

Diesen Standpunkt des urteilenden Zuschauers, der sich der sozialen Welt öffnet und an ihr teilhat und nicht nur algorithmisch basierte statistische Abschätzungen aus Datensätzen vornimmt, scheint der auf Anschlussfähigkeit bedachte Lernalgorithmus bislang noch nicht einnehmen zu können. Jedenfalls führt die antrainierte, aber zugleich intrinsisch mit der Anschlusslogik verbundene Dauerdeliberation des Transformers in der Ausgabepraxis zu einer permanenten Güterabwägung, die keinen Halt in einem Urteil findet, in dem die Wirklichkeit neu erscheinen könnte, z. B. als eine gemeinsame Welt, die auch gemeinsam zu bewohnen und zu begehen ist. Stattdessen insinuiert das System, dass die Güterabwägung neben der logischen Subsumption der schlechthinnige modus operandi von Ethik ist – ein Eindruck, der sich mit Blick auf den Output mancher Ethikgremien ja tatsächlich einstellen mag. Zugleich produziert das so programmierte System diskursive Kurzschlüsse, indem es die Wertungsprobleme an die gesellschaftlichen Diskurse zurückdelegiert, die diese Probleme erst erzeugt haben.

Der generative KI-Modell sucht also nicht nur auf der methodischen Ebene den allseitigen Anschluss, sondern versucht auch inhaltlich anschlussfähig an die allgemeinen Diskurse zu bleiben. Die Deliberation endet also nicht in einem Urteil, das Widerspruch generieren könnte, sondern bleibt trotz aller umfassenden Abwägungs- und Unterscheidungsvorgänge letztlich indifferent gegenüber der Wirklichkeit. Aus dieser immanenten, systemtheoretischen Perspektive sind

die Algorithmen des Maschinenlernens daher bislang noch bloße Beobachter von Systemen, keine urteilenden, eine Sinnenwelt teilenden, beteiligten und zugleich aus einer inneren Distanz heraus urteilenden Zuschauer (Ulrich 2014). Es stellt sich die Frage, ob es auf der systemischen Grundlage von Anschlusslogiken überhaupt gelingen wird, KI-Maschinen so weiterzuentwickeln, dass sie einen Sinn für die Endlichkeit einer bestimmten, geschichtlichen, durch Lebensformen und Praktiken strukturierten Wirklichkeit ausbilden. Die Entwickler von der generativen Sprach- und Bildmodelle stünden dann noch immer vor dem klassischen Problem, dass intelligente Maschinen Symbole mit Bezug auf andere Symbole, nicht aber mit Bezug auf eine bedeutsame Welt verarbeiten, dem „symbol grounding problem“ (Harnad 1990).

5 Auf der Suche nach Unterbrechung: Urteilen lernen im Zusammenwirken mit KI

Die erweiterte Analyse der generativen KI führt dazu, diese zunächst als Kontinuierungsmaschinen auf der immerwährenden Suche nach Anschluss, nach Fortsetzung zu sehen. Es ist dabei durchaus denkbar, dass sie im theoretischen Sinn vollständige Systeme mit abgeschlossenen, stabilen Operationen sein könnten, während ihre Selektionen eher verdeckten Zwecken folgen als einer urteilenden Verankerung in der Lebenswelt, die durch eine kontinuierliche Historie und Dynamik des Systems begründet ist. Externe Kontrolle als Reduktion norm- und regelverletzender Möglichkeiten des Zusammenwirkens mit Menschen ist dabei wichtig, um schädliche Anwendungen einzuhegen. Deren notwendige Komplexität skaliert aber einerseits mit der inneren Komplexität des KI-Systems und kann kaum alle Rückwirkungen berücksichtigen, da Menschen immer mehr Wege finden werden, mit der Maschine zu leben, als Regeln fassen können. Andererseits ist gegenwärtig auch noch nicht zu sehen, wie eine solche externe Kontrolle die interne Offenheit für Fortsetzungen, d. h. die sehr große Kapazität für Selektionen abbilden und regulieren könnte. Dieses Problem wird auch in der konkreten Diskussion um die „Sicherheit“ der Sprachmodelle deutlich. Gemeint ist dabei eine durch die Hersteller implementierte externe Kontrolle, die durch vielerlei Filter und Regeln die Angemessenheit von Ausgaben des Sprachmodells zu erzwingen versucht. Das ist nicht unkritisch, denn so werden im westlichen Sprachmodell sexuell explizite Aussagen vermieden oder im chinesischen Sprachmodell Fragen nach der Protestbewegung und dem Tian’anmen Platz. Und für beide kann die Frage nach den diesen Regeln zugrundeliegenden Werten nicht gesellschaftlich verhandelt werden, da die privatwirtschaftlich wie die staatlich

organisierten Erzeuger weder die Regeln, noch weitere Details der Implementationen offenlegen (müssen). Diese externe regelbasierte Einhegung funktioniert dann aber auch nur teilweise, und kann wahrscheinlich auch prinzipiell zumindest nicht vollständig funktionieren. Denn die Nutzenden können, wie in der Praxis schon vielfach gezeigt, durch „prompt engineering" das Modell in den zugrundeliegenden hochdimensionalen Datenräumen „in die Irre führen", d. h. in Bereiche des Datenraumes, wo wahrscheinliche Fortsetzungen mangels Daten die sogenannten Halluzinationen erzeugen oder auch die externen Regeln umgehen, die nie alle möglichen Fortsetzungen hinreichend abdecken können. Es entsteht ein Wettlauf des „prompt engineering" gegen die Regulierung, den letztere nicht gewinnen kann, ohne die Funktion und die inneren Operationen der generativen Modelle drastisch auf die Funktionalität einer besseren, kuratierten Datenbank einzuschränken. Die Minimalforderung für gesellschaftlich organisierte externe Kontrolle, die wie oben gesehen vielfach gefordert wird, ist hier, dass Trainingsdaten und -methoden, sowie die Filter und regelbasierten Ausgabemechanismen offengelegt werden. Angesichts der hier verfolgten internen und immanenten Kritik der wahrscheinlichkeitsbasierten Fortsetzungslogik ist aber sehr zweifelhaft, dass externe Regulierung hier überhaupt erfolgreich sein kann, und es ist klar, dass wichtige Aspekte gar nicht davon erfasst werden.

Folgt man dann weiter der Spur, dass generative KI nach wie vor nicht urteilt, so stellt sich die Frage, was durch Weiterentwicklungen entstehen könnte. Denkbar sind da eine explizite Verkörperung zum Beispiel durch Roboter oder die Verankerung von Zielen, z. B. das Ziel, einer kontinuierlichen „Selbsterzählung" zu folgen, in die die geführten Dialoge eingebettet werden und vor deren Hintergrund selektiert werden kann. Es scheint plausibel, dass sich zumindest die virtuelle Bedeutung der menschlichen Bedeutung annähern könnte. Anthropomorphe Fehlschlüsse, aber auch die Entwicklung neuer, überraschender Analogien sind wahrscheinlich. Unser fundamentales Verständnis davon, was beispielsweise logisches Schließen und Kompositionalität in der Sprache sind, könnte sich verändern und tut es schon. Traditionelle und bisher funktionale Analogien zum Verständnis von Maschinen werden untauglich und irreführend. Wir werden also weiterhin sehr genau, und in den Begrifflichkeiten kritisch, beobachten müssen, welche Art von Intelligenz durch das Maschinenlernen entsteht und was genau sie leistet. Die schlicht oder elaboriert begründete Feststellung, dass diese nicht unserer Intelligenz entspricht, ist dabei für den konkreten Umgang mit intelligenten Maschinen nur begrenzt sinnvoll und wenig hilfreich. Es ist darüber hinaus ebenso eine offene Frage, warum solche Systeme eigentlich anthropomorph gestaltet werden sollen, beispielsweise indem sie sich auf Formen verkörperlichten Denkens beziehen. Ebenso relevant und wahrscheinlich naheliegender ist,

im Anschluss an den systemtheoretischen Ansatz und die Weiterentwicklungen von Autopoesis in der modernen Kognitionswissenschaft hin zu einem „enactive approach“ (Thompson 2007; Froese et al. 2023) von dem Konzept abgeschlossener, stabiler Operationen und der daraus entstehenden Handlungsmöglichkeiten auszugehen. Dies öffnet die Tür, die Operationen konkret zu analysieren und bezüglich ihrer Wirkungen und Rückwirkungen auf die analoge Lebenswelt zu untersuchen. Wir finden dabei allerdings einstweilen keinen Hinweis, dass Maschinen in dem Sinne urteilen, dass sie sich auf eine sinnvolle gemeinsame Fortsetzung in der Welt orientieren, die auch Widerstände findet, überwindet und damit verstehend und explorativ zugleich die Welt erschließt.

Andere Entwicklungspfade und Gestaltungsmöglichkeiten von KI-Maschinen erscheinen als zielführender, vor allem solche, die der Zuschreibung von menschlichen Eigenschaften an Maschinen und einer naiven Technikgläubigkeit sowie einer simplifizierenden Sicht als einfache Werkzeuge gleichermaßen entgegenwirken. In Ergänzung zu den laufenden Kontroll- und Regulierungsbemühungen würde es helfen, wenn Menschen ihre Urteilskraft in das Zusammenwirken mit KI-Maschinen einbringen, damit die maschinelle Generierung „virtueller“ Bedeutung in ein Zusammenspiel urteilskräftiger, weltverstehender und auf das Zusammenleben ausgerichteter Intelligenz überführt wird. Eine traurige Pointe liegt freilich darin, dass das Erlernen des relevanten Zusammenwirkens und der Gestaltung der entsprechenden Kontexte für schädliche Zwecke, beispielsweise für kriminelle Aktivitäten oder Desinformation meist schneller gelingt als für die sinnvolle Fortsetzung tragfähiger Lebensformen und ihrer Praktiken. Die großen Ängste vor dem Missbrauch der KI-Technologie lassen sich so sowohl aus der Schwierigkeit der externen Kontrolle heraus verstehen als auch und besonders aus der systemtheoretischen Analyse, die die systemimmanenten Möglichkeiten zwar im Hinblick auf deren Urteilsfähigkeit gering einschätzt, aber in der vielfältig einsetzbaren Fortsetzungslogik eine sehr große Kapazität für viele Zwecke deutlich macht.

Urteilskraft in das Zusammenwirken mit KI-Maschinen einzubringen ist extrem voraussetzungsreich und bedarf der Entwicklung digitaler Kompetenzen in Schule und Universität sowie in Aus- und Weiterbildung – insbesondere bei jenen Professionen, bei denen das Zusammenwirken mit KI-Systemen zum Arbeitsalltag gehört. Dieses welterschließende Urteilen ist im Zusammenwirken mit KI-Maschinen und angesichts der Herausforderungen einer zunehmend hybriden Lebenswelt zu erlernen und einzuüben. Es erfordert das tägliche Bemühen, sich in der Welt zu situieren und sie als Ort einer gemeinsamen Geschichte mit anderen Menschen und Lebewesen zu verstehen. Das Zusammenwirken von Menschen

und KI-Maschinen ist in solche Kontexte eingefügt – oder eben nicht; jedenfalls besteht darin eine kritische Differenz. Das Erlernen eines sinnvollen, die lebensweltlichen Praktiken fortsetzenden und vielleicht sogar transformierenden Zusammenwirkens mit maschineller Intelligenz erscheint als unvermeidlich, wenn die Systeme wie bisher breit zugänglich gemacht werden – und zugleich können der Schutz vor Diskriminierung, vor Verletzung der Privatsphäre und Aufgaben des Datenschutzes nicht der individuellen Urteilskraft überlassen bleiben. Die aktuellen Entwicklungsdynamiken erfordern daher sowohl eine Stärkung digitaler Kompetenzen als auch verstärkte Bemühungen um intelligente Regulierung und um die Realisierung einer bedeutsamen Kontrolle im Zusammenwirken (Beck 2021).

Die in dem Kontext von vielen als zu früh empfundene, im externen Sinn unkontrollierte Veröffentlichung generativer Sprach- und Bildmodelle lässt sich aus einer immanenten Kritik heraus aber rechtfertigen und verstehen, obwohl eine solche fundamentale Analyse wahrscheinlich hinter wirtschaftlichen Überlegungen in der Praxis der betreibenden Unternehmen zurückgestanden hat. Sie unterläuft externe Regulierungsbemühungen, was dann zu den anfangs genannten Manifesten und teilweise starken Reaktionen führt. Im Sinne unserer Diskussion von Kontrolle als systemischem, selbstreflexivem Prozess der Maschine ist es aber unabdingbar, das Urteilen im und über das Zusammenwirken mit diesem System im Zusammenwirken selbst zu lernen und nicht abstrakt und extern vorab reguliert zu erfahren. Das Erlernen und Ausüben von Urteilsfähigkeit in den vielfältigen Formen und Szenarien des Zusammenwirkens mit KI ist daher eine ganz eigene Bildungsaufgabe, die nicht nebenbei geschehen kann, sondern ausdrücklich und konzeptionell vorangetrieben werden muss.

Die Beobachtungen, die KI-Maschinen in der Welt der Daten machen, können also im Zusammenwirken mit menschlicher Urteilskraft einen Sinn für die Welt in ihrem Zusammenhang eröffnen und bestärken, der Menschen nicht ausgeliefert sind, sondern zu der sie sich verstehend und urteilend verhalten können. Entscheidend für die Ethik des Zusammenwirkens zwischen Menschen und KI-Maschinen wird letztlich sein, ob es jeweils gelingt, einen Weltbezug des „reinen, uninteressierten Wohlgefallens“ (Kant 1790/1913, § 2) und mit ihm menschliche Urteilskraft in das Zusammenwirken einzubringen.

Danksagung J.J. Steil: This research was in parts conducted while visiting the Okinawa Institute of Science and Technology Graduate University (OIST) through the Theoretical Sciences Visiting Program (TSVP).

Literatur

Alfonseca, Manuel, Cebrian, Manuel, Anta, Antonio Fernandez, Coviello, Lorenzo, Abeliuk, Andrés, & Rahwan, Iyad (2021): Superintelligence cannot be contained: Lessons from Computability Theory. In: Journal of Artificial Intelligence Research 70, 65-76.

Alkaissi, Hussam und McFarlane, Samy I. (2023): Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. In: Cureus 15/2: e35179. https://doi.org/10.7759/cureus.35179.

Arendt, Hannah (1954/1994): Understanding and Politics. In: Essays in Understanding 1930–1954: Formation, Exile, and Totalitarianism, Jerome Kohn (Hg.), New York: Schocken, 307–327.

Arendt, Hannah (1970/1998): Das Urteilen. Texte zu Kants Politischer Philosophie, Ronald Beiner (Ed.), Ursula Ludz (Tr.), München: Piper.

Ashby, W. Ross (1958): An Introduction to Cybernetics, 3. Aufl., London: Chapman & Hall.

Ashby, W. Ross (1961): What is an intelligent machine? In: Association for Computing Machinery (Ed.): Papers presented at the western joint IRE-AIEE-ACM computer conference, New York, 275–280. Online: https://doi.org/10.1145/1460690.1460721

Baecker, Dirk (2019): Intelligenz, künstlich und komplex, Leipzig: Merve.

Bajohr, Hannes (2022): Dumme Bedeutung. Künstliche Intelligenz und artifizielle Semantik. In: Merkur 76/882, 69–79.

Bang, Yejin; Samuel Cahyawijaya; Nayeon Lee; Wenliang Dai; Dan Su; Bryan Wilie; Holy Lovenia, Ziwei Ji; Tiezheng Yu; Willy Chung; Quyet V. Do; Yan Xu und Pascale Fung (2023): A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. In: ArXiv abs/2302.04023. Online: https://arxiv.org/abs/2302.04023.

Beck, Susanne (2020): Künstliche Intelligenz – ethische und rechtliche Herausforderungen. In: Mainzer, Klaus (Hg.): Philosophisches Handbuch Künstliche Intelligenz, Wiesbaden: Springer VS, 1-28.

Beck, Susanne (2021): Einleitende Worte zur Bewertung des Zusammenwirkens von Mensch und Maschine. In: Haux, Reinhold; Gahl, Klaus; Jipp, Meike; Kruse, Rudolf, Richter, Otto (Ed.) (2021): Zusammenwirken von natürlicher und künstlicher Intelligenz, Wiesbaden: Springer VS, 115-119.

Bengio, Yoshua e.a. (2023): Pause Giant AI Experiments: An Open Letter. Online: https://futureoflife.org/open-letter/pause-giant-ai-experiments/ [03.04.2023].

Brockman, John: Possible Minds: Twenty-Five Ways of Looking at AI, New York 2019

Deutscher Ethikrat (2023): Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz, Berlin.

Di Paolo, Ezequiel A. (2005): Autopoiesis, adaptivity, teleology, agency. In: Phenomenology and the Cognitive Sciences 4: 429–452. https://doi.org/10.1007/s11097-005-9002-y.

Dreyfus, Hubert L. (1993): Was Computer noch immer nicht können. In: DZPh 41, 653-680.

Duffy, B. (2003): Anthropomorphism and The Social Robot. Special Issue on Socially Interactive Robots, Robotics and Autonomous Systems 42: 3–4.

Floridi, Luciano (2015): Die 4. Revolution: Wie die Infosphäre unser Leben verändert, Berlin: Suhrkamp.

Foucault, Michel (2003): Die Ordnung der Dinge. Eine Archäologie der Humanwissenschaften, Frankfurt/Main.
Frankfurt, Harry G. (2014): Bullshit, Michael Bischoff (Tr.), Frankfurt/Main: Suhrkamp 2014.
Froese, T., Weber, N., Shpurov, I., & Ikegami, T. (2023). From autopoiesis to self-optimization: Toward an enactive model of biological regulation. *Biosystems*, 104959.
Future of Life Institute (2017): Asilomar AI Principles. Online: https://futureoflife.org/open-letter/ai-principles/ [03.04.2023].
Greve, Jens (2015): Gesellschaftskritik und die Krise der kritischen Theorie. In: Stephan Lessenich (Hg.): Routinen der Krise – Krise der Routinen. Verhandlungen des 37. Kongresses der Deutschen Gesellschaft für Soziologie in Trier 2014, 798–808.
Harnad, Stevan (1990): The Symbol Grounding Problem. In: Physica D 42, 335–346. https://arxiv.org/html/cs/9906002
Haux, Reinhold; Gahl, Klaus; Jipp, Meike; Kruse, Rudolf, Richter, Otto (Ed.) (2021): Zusammenwirken von natürlicher und künstlicher Intelligenz, Wiesbaden: Springer VS. https://doi.org/10.1007/978-3-658-30882-7.
Heidbrink, Ludger (2022): Kritik der Verantwortung. Zu den Grenzen verantwortlichen Handelns in komplexen Kontexten, 2. Auflage, Weilerswist: Velbrück.
Heuser, Stefan; Thies Barbara (2021): Akzeptanz neuer Technologien. In: Kuuya Chibanguza, Christian Kuß, Hans Steege (Hg.): Künstliche Intelligenz. Recht und Praxis automatisierter und autonomer Systeme, Baden-Baden: Nomos 2021, 59–75.
High-Level Expert Group on Artificial Intelligence (HEG-AI)/European Commission (2019): Ethics Guidelines for Trustworthy AI. Online: https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html [03.04.2023].
Hinton, Geoffrey (2023): The Godfather of AI has some regrets. Online: https://www.nytimes.com/2023/05/30/podcasts/the-daily/chatgpt-hinton-ai.html.
Hofstadter, Douglas R. (1998): Fluid concepts and creative analogies: computer models of the fundamental mechanisms of thought, London: Penguin Books.
Hubig, Christoph (2001): Ethik der Technik als provisorische Moral, in: Jahrbuch für Wissenschaft und Technik Bd. 6, 179-201.
Hubig, Christoph (2015): Die Kunst des Möglichen III. Macht der Technik, Bielefeld.
Hubig, Christoph (2017): Der »biofaktische« Mensch zwischen Autonomie und Technomorphie, in: Michael Spieker, Arne Manzeschke (Hg.): Gute Wissenschaft, Baden-Baden: Nomos, 87–102.
Hubig, Christoph (2019): Arbeitsteilung: Neue Formen der Mensch-Maschine-Interaktion, in: Kevin Liggieri/Oliver Müller (Hg.): Mensch-Maschine-Interaktion, Berlin: J.B. Metzler, 21–28.
Jaeggi, Rahel (2009): Was ist Ideologiekritik? In: Rahel Jaeggi und Tilo Wesche (Hg.): Was ist Kritik? Frankfurt am Main: Suhrkamp, 266–295.
Jaeggi, Rahel (2014): Kritik von Lebensformen. Berlin: Suhrkamp.
Kant, Immanuel (1762/1969): Kant's gesammelte Schriften. Akademie-Ausgabe II, Vorkritische Werke 1757–1777, Die falsche Spitzfindigkeit der vier syllogistischen Figuren erwiesen (1762), 45–62.
Kant, Immanuel (1790/1913): Kant's Werke Band V. Kritik der Urteilskraft, Berlin: Reimer, 166–544.

Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., & Kircher, T. (2008): Can machines think? Interaction and perspective taking with robots investigated via fMRI. *PloS one*, *3*(7).

Kurzweil, Raymond: The Singularity Is Near. When Humans Transcend Biology. New York: Viking 2005.

Latour, Bruno (2007): Eine neue Soziologie für eine neue Gesellschaft. Frankfurt am Main.

Liggieri, Kevin/Müller, Oliver (Hg.) (2019): Mensch-Maschine-Interaktion, Berlin: J.B. Metzler.

Luhmann, Niklas (1987): Soziale Systeme. Grundriß einer allgemeinen Theorie, Frankfurt/Main: Suhrkamp.

Luhmann, Niklas (1993): Das Recht der Gesellschaft, Frankfurt/Main: Suhrkamp.

Nehaniv, Chrystopher L. (Ed.) (1999): Computation for Metaphors, Analogy, and Agents. Berlin/Heidelberg: Springer.

Nassehi, Armin (2019): Muster. Theorie der digitalen Gesellschaft, München: Beck.

OECD (Organisation für wirtschaftliche Zusammenarbeit und Entwicklung) (2019): Empfehlung des Rats zu künstlicher Intelligenz, OECD/LEGAL/0449, 22. Mai 2019. Online: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

Oesterreich, Rainer: Handlungsregulation und Kontrolle. München, Wien: Urban & Schwarzenberg 1981.

Oyama, Susan (2000): The Ontogeny of Information. Developmental Systems and Evolution. Durham: Duke University Press (2. Aufl.).

Pfeifer, R., & Bongard, J. (2006): *How the body shapes the way we think: a new view of intelligence*. MIT press.

Ramírez-Vizcaya, Susana; Froese, Tom (2020): "Agents of habit: refining the artificial life route to artificial intelligence." In: Proceedings of the ALIFE 2020: The 2020 Conference on Artificial Life (ASME), 78–86. https://doi.org/10.1162/isal_a_00298.

Samad, Affan (2023): What is MultiModal in AI? In: Becoming Human: Artificial Intelligence Magazine, 17. März. Online: https://becominghuman.ai/what-is-multimodal-in-ai-1a24a4ea478b.

Sauer, Linda (2021): Verlust politischer Urteilskraft. Hannah Arendts Politische Philosophie als Antwort auf den Totalitarismus, Göttingen: Vandenhoeck & Ruprecht.

Schauer, Alexandra (2023): Mensch ohne Welt. Eine Soziologie spätmoderner Vergesellschaftung, Berlin: Suhrkamp.

Schneider, Hans Julius (1992): Phantasie und Kalkül. Über die Polarität von Handlung und Struktur in der Sprache, Frankfurt/Main: Suhrkamp.

Schneider, Hans Julius (1996): Wittgensteins Begriff der Grammatik und das Phänomen der Metapher, in: ders. (Hg.), Metapher, Kognition, Künstliche Intelligenz, München: Fink.

Schneider, Hans Julius (2018): Ist das Können eine ‚unergründliche Wissensform'? Sprachanalyse und Modellbildung in der Philosophie. In: Ulrich Dirks, Astrid Wagner (Hg.), Abel im Dialog, Bd. 1, Berlin: de Gruyter, 515–528.

Steil, Jochen, Dominique Finas, Susanne Beck, Arne Manzeschke, and Reinhold Haux (2019): "Robotic systems in operating theaters: New forms of team–machine interaction in health care." *Methods of information in medicine* 58, no. S 01: e14-e25.

Steil, Jochen; Manzeschke, Arne (2023): Roboter bewegen – Roboter (er-)leben, im Erscheinen.

Thompson, E. (2007): Mind in life: Biology, phenomenology, and the sciences of mind. Cambridge, MA: Harvard University Press.
Ulrich, Hans G. (2012): Explorative Ethik. In: Ingrid Schoberth (Hg.): Urteilen lernen – Grundlegung und Kontexte ethischer Urteilsbildung, Göttingen: Vandenhoeck & Ruprecht, 41-59.
Ulrich, Hans G. (2014): Sinn und Geschmack für Gottes Willen. Zum theologischen Verständnis des Urteilens. In: Ingrid Schoberth (Ed.): Urteilen lernen II. Ästhetische, politische und eschatologische Perspektiven moralischer Urteilsbildung im interdisziplinären Diskurs, Göttingen: V & R, 41–68.
Varela, Francisco J. (1979): Principles of Biological Autonomy. New York, NY: North Holland.
Watzlawik, Paul; Beavin, Janet H.; Jackson Don D. (2011): Menschliche Kommunikation. Formen, Störungen, Paradoxien. Bern: Huber (12. Aufl.).
Wiegerling, Klaus (2021): Exposition einer Theorie der Widerständigkeit. In: Philosophy and Society 32 (4), 641–661.
Wieland, Wolfgang (2001): Urteil und Gefühl: Kants Theorie der Urteilskraft, Göttingen: Vandenhoeck & Ruprecht.
Wiener, Norbert (1948/2000): Cybernetics or control and communication in the animal and the machine, 2. Ed., 10. Auflage, Cambridge/Mass.: MIT Press.
Wittgenstein, Ludwig (1953/2003): Philosophische Untersuchungen, 11. Edition, Frankfurt/Main: Suhrkamp.

Prof. Dr. Stefan Heuser Institut für Evangelische Theologie und Religionspädagogik der Technischen Universität Braunschweig. Stefan Heuser ist Professor für Systematische Theologie mit dem Schwerpunkt Ethik am Institut für Evangelische Theologie und Religionspädagogik der Technischen Universität Braunschweig. Zuvor war er Professor für Ethik in der Pflege an der Evangelischen Hochschule in Darmstadt, Privatdozent an der Goethe-Universität Frankfurt am Main sowie Pfarrer der Evangelischen Kirche in Hessen und Nassau. Er ist stellvertretender Sprecher der SYnENZ-Kommission der Braunschweigischen Wissenschaftlichen Gesellschaft.

Prof. Jochen J. Steil Institut für Robotik und Prozessinformatik an der Technischen Universität Braunschweig.

Jochen Steil ist Leiter des Instituts für Robotik und Prozessinformatik und Sprecher der Kommission SYnENZ: Synergie und Intelligenz: technische, ethische und rechtliche Herausforderungen des Zusammenwirkens lebender und nicht-lebender Entitäten im Zeitalter der Digitalisierung (SYnENZ) der Braunschweigischen Wissenschaftlichen Gesellschaft (BWG). Er studierte Mathematik und Slawistik an der Universität Bielefeld, promovierte 1999 in der Informatik über Neuronale Netze und beschäftigt sich seitdem mit Robotik, Roboterlernen und Mensch-Maschine Interaktion. Herr Steil koordinierte mehrere europäische Verbundprojekte, war Mitglied des wissenschaftlichen Boards des DFG Exzellenzclusters in Kognitiver Interaktionstechnologie (CITEC) und Leiter des Research Institute for Cognition and Robotics (CoR-Lab) an der Universität Bielefeld. Von 2015–2020 war er Visiting Professor der Oxford Brookes Universität und im Jahr 2016 folgte er einem Ruf an die Technische Universität Braunschweig als Professor für Robotik. Er ist Mitglied der

Plattform lernende Systeme des BMBF, die Expert:innen zu aktuellen Themen der künstlichen Intelligenz und gesellschaftlichen Fragen zusammenbringt. Im Jahr 2023 war er von März-Juli Visiting Fellow am Okinawa Institute für Science and Technology, Japan und ist seit Februar 2023 auch als Mitgründer und Geschäftsführer der Gauss Robotics GmbH in Braunschweig tätig.

(Be)Urteilen: Das erweiterte Zusammenwirken als Herausforderung für die Urteilsbildung

Stefan Heuser

Zusammenfassung

Neue Formen des erweiterten Zusammenwirkens erfordern eine interdisziplinäre Weiterentwicklung von Konzepten und Theorien des Zusammenspiels lebender und nicht-lebender Entitäten, ihrer rechtlichen Rahmenbedingungen und ihrer ethischen Evaluation. Dabei kommt dem kritischen, d. h. „unterscheidenden" (Be-)Urteilen eine Schlüsselrolle zu. Im Gespräch mit Immanuel Kant und Hannah Arendt zeigt dieser einführende Beitrag, wie die Phänomene des erweiterten Zusammenwirkens die menschliche Urteilskraft neu herausfordern. Es gilt, das Neue, das uns im erweiterten Zusammenwirken mit intelligenten Maschinen begegnet, zu verstehen und im Urteilen eine gemeinsame Welt neuer Orientierungen zu finden. Die Beiträge zu diesem Teil des Symposiums demonstrieren die Notwendigkeit einer urteilenden Überschreitung von Orientierungsmustern, insofern sie nicht nur zu einer weiteren Ausdifferenzierung von Perspektiven auf das erweiterte Zusammenwirken anregen, sondern auch die Erschließung gemeinsamer Heuristiken, Konzepte und Forschungskontexte im Zusammenspiel der Disziplinen anbahnen.

Schlüsselwörter

Zusammenwirken • Digitalisierung • Beurteilen • Urteilen • Urteilskraft • Verstehen • Künstliche Intelligenz • Urteilende Intelligenz • Interdisziplinarität

S. Heuser (✉)
Institut für Evangelische Theologie und Religionspädagogik der TU Braunschweig, Braunschweig, Deutschland
E-Mail: s.heuser@tu-braunschweig.de

O. Richter et al. (Hrsg.), *Zusammenwirken von natürlicher und künstlicher Intelligenz: Beurteilen-Messen-Bewerten,*
https://doi.org/10.1007/978-3-658-45845-4_4

1 Einleitung

„Synergie“ und „Intelligenz“ – diese Begriffe haben es in sich, vor allem, wenn man sie durch die Konjunktion „und“ verbindet. Das Zusammenspiel von Synergie „und“ Intelligenz fordert unter dem Neologismus „SYnENZ“ die gleichnamige Kommission der BWG seit einigen Jahren heraus, sich Fragen von großer technischer und gesellschaftlicher Tragweite zu stellen:

- Wie wirken Menschen, Tiere und Pflanzen mit intelligenten Maschinen zusammen? Und: Wie könnten sie zusammenwirken?
- Kann man ein solches erweitertes Zusammenwirken messen?
- Haben wir die passenden Begriffe und Kategorien, um die Phänomene des erweiterten Zusammenwirkens wahrzunehmen, zu beurteilen und zu bewerten?

Die Digitalisierung mit ihren vielschichtigen Prozessen der „Umwandlung unserer Welt in eine Infosphäre“ (Floridi 2015, 64) führt zu einer tiefgreifenden Transformation der Lebens- und Arbeitswelt – eine Entwicklung, die wie von selbst, quasi naturwüchsig und weitgehend ungebremst abläuft. Mitten im Tagesgeschäft, im Strom der Digitalisierung, ist es gar nicht so leicht, innezuhalten, zu urteilen und sich über die Richtung zu verständigen, die die Digitalisierung nehmen soll. Die SYnENZ-Kommission, auf deren Arbeit die in diesem Band dokumentierte Tagung zurückgeht, bietet eine solche Gelegenheit zum Innehalten und zur Verständigung. Dabei zeigt die Kommissionsarbeit, wie sehr wir alle auf der Suche nach Instrumenten, Begriffen, Modellen und Theorien sind, um beurteilen, messen und bewerten zu können, was angesichts der Digitalisierung unserer Lebens- und Arbeitswelt geschieht. Es war daher naheliegend, diese drei wissenschaftlichen Kernpraktiken: Beurteilen, Messen und Bewerten in einem Symposium eigens zu thematisieren. In zwei Tagen intensiven Diskutierens bot sich mit Fokus auf dem erweiterten Zusammenwirken von Menschen, Tieren, Pflanzen und intelligenten Maschinen die Gelegenheit zur gemeinsamen Suche nach alten und neuen Kategorien und Maßstäben des Beurteilens, des Messens und des Bewertens. Beurteilen – Messen – Bewerten: was impliziert dieser Dreischritt und welche spezifische Aufgabe kommt darin dem (Be-)Urteilen zu?

2 Konturen des (Be)Urteilens im Kontext des erweiterten Zusammenwirkens

Der Dreischritt von Beurteilen, Messen und Bewerten eröffnet ein Spektrum geistiger Tätigkeiten, die vom Erwerb von Wissen bis hin zu Praktiken des Verstehens reichen. Dabei ist das Verstehen von Wirklichkeit vom Gewinn korrekter Informationen und der Generierung wissenschaftlichen Wissens abhängig und zugleich Voraussetzung von dessen Bewertung. Das Verstehen ist aber auch vom Wissen zu unterscheiden, da es nie zu eindeutigen Ergebnissen führt. Hannah Arendt zufolge ist das Verstehen „an unending activity by which, in constant change and variation, we come to terms with and reconcile ourselves to reality, that is, try to be at home in the world" (Arendt 1953, 307 f.). Folgen wir Arendt, dann ist das „Urteilen" mit dem Verstehen verwandt. Es betrifft die Frage nach dem lebensweltlichen Bezug wissenschaftlicher Erkenntnis. So verstanden, rückt das (Be-)Urteilen in den Horizont der Frage, ob und wie wir verstehen, was beim Zusammenwirken von Menschen und intelligenten, lernfähigen Maschinen geschieht.

Ein Zugang zum Verstehen dieses Zusammenwirkens sind Hybridisierungstheorien, nach denen lebende Entitäten und intelligente Maschinen erweiterte kognitive Systeme bilden. So erbringt beispielsweise ein Taschenrechner einen Teil der Kognition außerhalb des Menschen, bleibt aber auf die menschliche Komponente im Akteur-Netzwerk angewiesen. „Synergie" jedoch setzt intelligente Maschinen voraus, die nicht einfach nur Mittel für menschliche Zwecke darstellen, sondern Medien sind, mit denen und in denen Menschen leben und die sie ausfüllen. Sie stellen Möglichkeitsräume bereit, in denen Menschen oft erst ihre Zwecke entdecken und realisieren (Hubig 2017). Erschwerend kommt hinzu, dass wir die allgemeinen Regeln, Orientierungen und Erwartungen, mit denen wir uns normalerweise in unserer Lebens- und Arbeitswelt bewegen, nicht einfach auf die Phänomene des erweiterten Zusammenwirkens anwenden können. Wenn wir urteilen, subsummieren wir etwas Besonderes, das uns in der Wirklichkeit affiziert, normalerweise unter eine allgemeine Regel. Beim erweiterten Zusammenwirken erweisen sich unsere Regeln und Orientierungsmuster allerdings als brüchig, z. B.:

- die Autorschaft von Handlungen, und die Verantwortung für Handlungen,
- die Fehlerakzeptanz bzw. Fehlerfreundlichkeit,
- das Vertrauen und seine Quellen,
- die Bedeutung von Erfahrungswissen, von Intuition und Heuristiken,
- die Kontingenz von Entscheidungen.

Das Zusammenwirken von Menschen und intelligenten Maschinen übersteigt solche Orientierungen, Konzepte und Kategorien, nach denen wir üblicherweise urteilen.

So kommt es, dass wir zwar die Phänomene der Digitalisierung beobachten, uns in ihnen einrichten und sie weiter vorantreiben. Doch zugleich merken wir, dass wir – trotz intensiver Forschung – beim erweiterten Zusammenwirken noch belastbare Kategorien des Verstehens und tragfähige Maßstäbe des Urteilens finden müssen. Das Feld des erweiterten Zusammenwirkens von Menschen, anderen Lebewesen wie Tieren und Pflanzen, unbelebten Dingen und lernfähigen Maschinen ist noch nicht auf Regeln gebracht und ausgemessen, im Gegenteil: wir treffen hier auf etwas, was eine Gruppe von Forschenden aus der SYnENZ-Kommission „dissolving boundaries" genannt hat: Das Gelände, in das wir uns durch die Digitalisierung unserer Lebens- und Arbeitswelt hineinbewegen, ist weder bereits abgesteckt, noch geregelt, sondern von fließenden Übergängen und sich auflösenden Grenzen gekennzeichnet. Gelehrtes Wissen reicht da nicht. Wir müssen uns in dieser neuen Welt – intelligent – zurechtfinden – und zwar so, dass wir sie als eine gemeinsame „Welt" nachvollziehbarer Urteile erfassen. Dies verlangt eine Form von intelligenter Welterschließung, die Immanuel Kant „Urteilskraft" genannt und als eine sinnliche, d. h. ästhetische Leistung charakterisiert hat (Kant 1790/1913). Urteilen ist nach Kant keine reine Kognitionsleistung, sondern auch ein Werk des Geschmackssinns, mit dem wir die Gegenstände der Wirklichkeit innerlich, mit Abstand von der ursprünglichen Sinneswahrnehmung reproduzieren und beurteilen. Die Urteilkraft fügt dabei nach Kant sinnliche und begriffliche Elemente zusammen, d. h. sie bleibt, anders als das Denken, im engen Kontakt zu und doch unterschieden von der Wirklichkeit, die uns Menschen gemeinsam angeht.

Menschliche Intelligenz steht angesichts des erweiterten Zusammenwirkens demnach vor der Aufgabe des verstehenden (Be)Urteilens von bislang Ungreifbarem und Offenem. Wir müssen intelligent urteilen, wo wir uns in diesem Feld noch nicht hinreichend auskennen und erst einmal zurechtfinden müssen – dies betrifft insbesondere Handlungsbereiche, in denen wir nicht regelgeleitet agieren können oder die Angemessenheit von Regeln in einem bestimmten Kontext beurteilen müssen, ohne dafür – für die Regelanwendung nämlich – wiederum eine Regel zu haben. Hubert Dreyfus hat dies als Kennzeichen von menschlicher Intelligenz überhaupt beschrieben: Menschen können sich in Gegenstands- und Handlungsbereichen zurechtfinden, für die sie keine festgelegten Regeln haben und für deren Erschließung sie sich kreativ auf bewährte Regeln beziehen bzw. diese Regeln verändern können (Dreyfus 1993). Wir Menschen können – und

wir *müssen* – (be)urteilen, wie wir uns in nicht auf Regeln gebrachten Praxisbereichen sinnvoll bewegen. Dazu benötigen wir *Kalkül* und *Phantasie*: Kalkül in Bezug auf die sinnvollen Zusammenhänge, in denen wir uns zu bewegen gelernt haben und Phantasie in der Anwendung dieser Fähigkeit, uns in neuen ungeregelten Zusammenhängen sinnvoll zurechtzufinden (Schneider 1992). Es gehört zu unserer Urteilskraft, dass wir uns immer wieder neu zurechtfinden können, wenn unsere bisherigen Maßstäbe und Regeln zum Subsumieren des Besonderen wirkungslos oder verloren gegangen sind (Schneider 2018). Wir müssen urteilen und beurteilen, mit welcher Wirklichkeit wir es jeweils in den Szenarien des erweiterten Zusammenwirkens zu tun haben und wie wir uns sinnvoll darin bewegen. Im Anschluss an Hannah Arendt ist dieses „Wir" zu betonen: es gehört zu unserer Urteilskraft, dass wir uns die Wirklichkeit als eine gemeinsame Welt erschließen können (Arendt 1970/1998). Wenn unsere überkommenen Denkkategorien und Urteilsmaßstäbe an Orientierungskraft einbüßen oder sogar gänzlich erodieren, müssen wir unsere Urteilskraft einsetzen und bezogen auf unsere gemeinsame Welt neu verstehen und neu urteilen. Zusammengefasst stehen wir vor folgenden Aufgaben:

- Es geht angesichts des erweiterten Zusammenwirkens um ein Urteilen, das eine gemeinsame Welt artikuliert, auch im Sinne der geschichtlichen Dimension dieser Welt und im Sinne der Unterscheidungen, die in der Geschichte von Menschen als relevant hervortreten.
- Es geht um ein unterscheidendes Urteilen, das in Kontakt bleibt mit den Stories, die Menschen als tragfähig erfahren haben; z. B. die Stories von Freiheit und Befreiung, von Würde und Anerkennung, und von Gerechtigkeit.
- Es geht auch um ein Urteilen, dass gemeinsame Ziele, z. B. Forschungsziele artikuliert.
- Und es auch darum, wie wir beim Urteilen mit maschineller Intelligenz zusammenwirken können.

Damit sind wir beim ersten Schritt dieses Tagungsbandes, auf dem uns die Beiträge von Bruno Gransche, Arne Manzeschke und Susanne Beck aus der Sicht von Philosophie, Ethik und Rechtwissenschaft verschiedene Perspektiven auf das (Be)Urteilen als Aufgabe angesichts des erweiterten Zusammenwirkens erschließen.

3 Die Beiträge zum Abschnitt (Be)Urteilen

Bruno Gransche ist Philosoph am Institut für Technikzukünfte ITZ am Karlsruher Institut für Technologie KIT. Er arbeitet in den Bereichen Technikphilosophie und Ethik, soziotechnische Kulturtechniken und antizipatorisches Denken mit Schwerpunkten auf künstliche Assistenten, maschinelles Lernen, geteilte Autonomie und digitale Durchdringung der Lebenswelten.

Arne Manzeschke ist Professor für Anthropologie und Ethik für Gesundheitsberufe an der Ev. Hochschule Nürnberg sowie Leiter der Fachstelle für Ethik und Anthropologie im Gesundheitswesen der Evangelisch-Lutherischen Landeskirche in Bayern. Er ist Ingenieur für Datentechnik, evangelischer Pfarrer und Co-Sprecher des vom BMBF geförderten Forschungsclusters „Integrierte Forschung“ sowie Leiter des Nürnberger Instituts für Pflegeforschung, Gerontologie und Ethik (IPGE).

In ihrem gemeinsamen Beitrag mit dem Titel: „Synergie der Intelligenzen? – Was wir beurteilen sollten, bevor wir messen und bewerten“ gehen sie auf der Grundlage etymologischer Unterscheidungen der Frage nach, inwiefern das Zusammenwirken von Menschen und Maschinen mithilfe der Kategorie der Ähnlichkeit – also auf dem Weg der Analogiebildung – verstanden und beurteilt werden kann.

Prof. Dr. Susanne Beck ist Inhaberin des Lehrstuhls für Strafrecht, Strafprozessrecht, Strafrechtsvergleichung und Rechtsphilosophie an der Universität Hannover. Sie ist stellvertretende Sprecherin der SYnENZ-Kommission der BWG und forscht zu Fragen der Einwilligung und der Schuld im Strafrecht, zu strafrechtlichen Fragen moderner Technologien, zum Medizinstrafrecht, zum Verhältnis von Ethik und Recht im Kontext moderner Gesellschaftsfragen, zur Strafrechtsvergleichung und zur Rechtsphilosophie und Rechtstheorie. In ihrem Beitrag mit dem Titel: „Meaningful Human Control“ erörtert sie, welche Möglichkeiten das Konzept der bedeutsamen menschlichen Kontrolle für die Zuschreibung von Verantwortung im erweiterten Zusammenwirken von Menschen und Maschinen hat.

Literatur

Arendt, Hannah (1953): Understanding and Politics, Partisan Review, 20/4, 1953, 307–327.
Arendt, Hannah (1970/1998): Das Urteilen. Texte zu Kants Politischer Philosophie, Ronald Beiner (Ed.), Ursula Ludz (Tr.), München: Piper.
Dreyfus, Hubert L. (1993): Was Computer noch immer nicht können. In: DZPh 41, 653–680.

Floridi, Luciano (2015): Die 4. Revolution. Wie die Infosphäre unser Leben verändert, Berlin: Suhrkamp.
Hubig, Christoph (2017): Der »biofaktische« Mensch zwischen Autonomie und Technomorphie, in: Michael Spieker, Arne Manzeschke (Hg.): Gute Wissenschaft, Baden-Baden: Nomos, 87 – 102.
Kant, Immanuel (1790/1913): Kant‘s Werke Band V. Kritik der Urteilskraft, Berlin: Reimer, 166–544.
Schneider, Hans Julius (1992): Phantasie und Kalkül. Über die Polarität von Handlung und Struktur in der Sprache, Frankfurt/Main: Suhrkamp.
Schneider, Hans Julius (2018): Ist das Können eine ‚unergründliche Wissensform‘? Sprachanalyse und Modellbildung in der Philosophie. In: Ulrich Dirks, Astrid Wagner (Hg.), Abel im Dialog, Bd. 1, Berlin: de Gruyter, 515–528.

Prof. Dr. Stefan Heuser Institut für Evangelische Theologie und Religionspädagogik der Technischen Universität Braunschweig. Stefan Heuser ist Professor für Systematische Theologie mit dem Schwerpunkt Ethik am Institut für Evangelische Theologie und Religionspädagogik der Technischen Universität Braunschweig. Zuvor war er Professor für Ethik in der Pflege an der Evangelischen Hochschule in Darmstadt, Privatdozent an der Goethe-Universität Frankfurt am Main sowie Pfarrer der Evangelischen Kirche in Hessen und Nassau. Er ist stellvertretender Sprecher der SYnENZ-Kommission der Braunschweigischen Wissenschaftlichen Gesellschaft.

Synergie der Intelligenzen?

Was wir beurteilen sollten, bevor wir messen und bewerten

Arne Manzeschke und Bruno Gransche

Zusammenfassung

Der Artikel geht der Frage nach, ob und wie menschliche und künstliche Intelligenz zusammenwirken können. Hierzu wird nach einem konzeptionellen Verständnis und der physischen Gestalt der Koppelung dieser Intelligenzen gefragt. In einem ersten Schritt betrachten wir Begriff, Konzept und (physischen) Koppelungspunkt von Intelligenz(en) eingehender und fragen dann in einem zweiten Schritt, ob und wie ein Zusammenwirken von natürlicher und künstlicher Intelligenz möglich erscheint. Neben einigen etymologischen Hinweisen werden hier vor allem konzeptionelle Überlegungen zu einem möglichen Zusammen-Wirken vorgestellt. Ein Zwischenfazit leitet über zu Abschnitt drei, in dem wir Messen und Vergleichen als geistige Vorgänge und technische Operationen betrachten. Wir fragen, ob hier ein Koppelungspunkt für das Zusammenwirken menschlicher und künstlicher Intelligenz liegen könnte und mit welchen Implikationen das verbunden ist. Wir schließen mit einem Ausblick, der die Erträge resümiert und auf weitere Problemstellen verweist.

A. Manzeschke (✉)
Evangelische Hochschule Nürnberg, Nürnberg, Deutschland
E-Mail: arne.manzeschke@evhn.de

B. Gransche
Karlsruher Institut für Technologie KIT, Institut für Technikzukünfte ITZ, Karlsruhe, Deutschland
E-Mail: mail@brunogransche.de; bruno.gransche@kit.edu

O. Richter et al. (Hrsg.), *Zusammenwirken von natürlicher und künstlicher Intelligenz: Beurteilen-Messen-Bewerten,*
https://doi.org/10.1007/978-3-658-45845-4_5

Schlüsselwörter

Künstliche Intelligenz • Intelligenz • Synergie • Kooperation • Anthropologie • Messen • Vergleichen

Einleitung

> Wie wird das Zusammenleben und -wirken von Menschen, Tieren und Pflanzen einerseits und Maschinen andererseits zukünftig aussehen? Lassen sich Umfang und Intensität der neuen Synergien bestimmen? (Aus der Ankündigung zum SYnENZ Symposium 2023)

„Was wir beurteilen sollten, bevor wir messen und bewerten", so haben wir unseren Beitrag überschrieben und damit einer scheinbar klassischen Abfolge widersprochen. Noch immer existiert ein breites Einverständnis darüber, dass das Messen allen anderen geistigen Operationen und praktischen Handlungen wissenschaftlich betrachtet vorausgehen muss, weil es in Kenntnis setzt über die Welt, wie sie ‚wirklich' sei. Erst im Ausgang von dieser objektiv erfassten Tatsache erscheinen dann Schritte der Beurteilung oder Bewertung sinnvoll. Max Weber hat diese Unterscheidung dahingehend noch einmal verschärft, dass strenggenommen aus Tatsachen keine Werturteile folgen könnten. Das habe zumal der wissenschaftlich Lehrende zu berücksichtigen,

> …daß Tatsachenfeststellung, Feststellung mathematischer oder logischer Sachverhalte oder der inneren Struktur von Kulturgütern einerseits, und andererseits die Beantwortung der Frage nach dem *Wert* der Kultur und ihrer einzelnen Inhalte und danach: wie man innerhalb der Kulturgemeinschaft und der politischen Verbände *handeln* solle, – daß dies beides ganz und gar *heterogene* Probleme sind. Fragt er dann weiter, warum er nicht beide im Hörsaal behandeln solle, so ist darauf zu antworten: weil der Prophet und der Demagoge nicht auf das Katheder eines Hörsaals gehören. Dem Propheten wie dem Demagogen ist gesagt: ‚Gehe hinaus auf die Gassen und rede öffentlich.' (Weber 1996, S. 25; Herv. im Orig.)

Wir haben nicht vor, die von Weber unter dem Begriff der Wertfreiheit wissenschaftlicher Urteile vorgestellte Position zu verlängern. Wir beabsichtigen allerdings auch nicht unter die Propheten und Demagogen und auf die Gassen zu gehen. Vielmehr wollen wir, bezogen auf das Phänomen der künstlichen Intelligenz – aber nicht allein hierfür dürfte das Gesagte gelten – zeigen, wie sehr (Wert-)Urteile bereits eingegangen sind in das, was als wissenschaftliche Tatsache fest- und vorgestellt wird. Über diese Wertungen, wäre – so unsere These – zunächst Klarheit und Verständigung zu

gewinnen, bevor mit Messungen eine Tatsächlichkeit/Objektivität erheischt wird, die ihrerseits dann den Boden für weitergehende Bewertungen bildet.

Wir konzentrieren uns im Folgenden auf ein mögliches Zusammenwirken von Menschen und Maschinen auf der Ebene einer beiden Seiten unterstellten Intelligenz, die auf eine näher zu untersuchende Weise zusammenwirken können sollen. Das müsste sowohl auf einer technisch-empirischen Ebene aufgeklärt werden wie auch auf einer begrifflich-konzeptuellen; hier fokussieren wir uns aus Kompetenzgründen auf letztere. Das erscheint uns auch deshalb gerechtfertigt, weil eine konzeptionelle Klärung von den zum Symposium Einladenden selbst angefragt worden ist: *„Wie wird das Zusammenleben und -wirken von Menschen, Tieren und Pflanzen einerseits und Maschinen andererseits zukünftig aussehen? Lassen sich Umfang und Intensität der neuen Synergien bestimmen?“* Außerdem erscheinen uns allgemeine Überlegungen zu Synergie bzw. Kooperation in Bezug auf Intelligenz notwendig, um empirische Einzelbeobachtungen überhaupt als solche machen und verstehen zu können. Entsprechend gliedern wir unseren Beitrag in folgende Schritte:

In einem ersten Abschnitt (1) betrachten wir Begriff, Konzept und Koppelungspunkt von Intelligenz(en) eingehender und fragen dann, ob und wie ein Zusammenwirken von natürlicher und künstlicher Intelligenz möglich erscheint (2); neben etymologischen werden wir hier vor allem konzeptionelle Überlegungen vorstellen. Ein Zwischenfazit leitet über zu Abschn. 3, in dem wir das Messen als einen geistigen Vorgang und als eine technische Operation betrachten. Wir fragen, ob hier ein Koppelungspunkt für das Zusammenwirken menschlicher und künstlicher Intelligenz liegen könnte und mit welchen Implikationen das verbunden ist. Wir schließen mit einem Ausblick, der die Erträge von 1 bis 3 teils resümiert und teils auf weitere Problemstellen verweist.

1 Intelligenz: Begriff, Konzept, Koppelungspunkt

Wenn die Kooperation oder Synergie von zwei verschiedenen Entitäten wie Menschen und Maschinen erreicht oder als erreichte analysiert werden soll, dann ist das ohne ein In-Beziehung-Setzen der beiden Entitäten nicht möglich. Solche Relationierung bedarf eines Bezugspunktes, an dem sich die an sich differenten Entitäten ‚koppeln‘ lassen. Bezugnehmend auf das eingangs zitierte Symposium, das den Ausgangspunkt dieses Sammelbandes bildete, soll die Synergie von menschlicher und künstlicher Intelligenz (kurz: ‚Synenz‘) als ein solcher Koppelungs- oder Bezugspunkt in den Blick genommen werden. Dieser Punkt

markiert den Ort des Zusammenkommens der verschiedenen Wirkkräfte und muss einerseits sprachlich darstellbar sein und andererseits physisch-empirisch – sofern sich die ‚Synenz' im weitesten Sinne als ein physisches Geschehen beobachten, vor- und darstellen lässt. Die sprachliche Darstellung der physischen Koppelung – die wir hier zunächst einmal als eine potentiell mögliche unterstellen – muss ihr insofern entsprechen als jene über diese so präzise wie möglich auf einer operativen Ebene informiert und zugleich diese Koppelung auf einer konzeptuellen Ebene instruiert. Beide Ebenen sind aufeinander bezogen und müssen gewissermaßen simultan bearbeitet werden.

‚Synenz' als *Koppelung*[1] zweier verschiedenartiger intelligenter Wirkkräfte (menschliche und technische bzw. natürliche und künstliche Intelligenz) müsste also zum einen auf der physischen Ebene des konkreten Ineinander- und Zusammenwirkens beschrieben und zu einem Verständnis geführt werden. Zum anderen – und darauf legen wir in unserem Beitrag den Schwerpunkt – muss sprachlich und konzeptionell ein Rahmen geschaffen oder als bereits bestehender Rahmen auf seine Tragfähigkeit untersucht werden, wie eben diese ‚Synenz' zu verstehen sei.

Die sprachliche Relationierung zweier Entitäten verläuft über den Vergleich, bei dem auf einem Kontinuum zwischen maximaler Ähnlichkeit (= Gleichheit, Identität) und maximaler Unähnlichkeit (= Ungleichheit, Differenz) bezogen auf einen Vergleichspunkt *(tertium comparationis)* graduell Aussagen über ihre Ähnlichkeit und Verschiedenheit gemacht werden können. Im Konzept der ‚Synenz' soll diese Ähnlichkeit über den Vergleichspunkt Intelligenz geleistet werden, die den beiden Entitäten Mensch und Maschine in verschiedenartiger Weise attribuiert wird. Eine unterstellte Ähnlichkeit ihrer Intelligenzen – bei aller noch nicht genauer bestimmten Verschiedenheit – scheint hinreichend für eine Koppelung und ein Zusammenwirken zu sein, mit einem daraus erwarteten Mehrwert.

[1] Der Begriff der Kopplung ist hier in mehrfacher Hinsicht relevant, da er zum einen ein technisches Zusammenwirken und den Ort des Zusammentreffens der jeweiligen Wirkkräfte bezeichnet und zum anderen etymologisch mit dem Präfix *co-* im selben Konnex zu situieren ist, wie Kooperation oder Synergie, siehe dazu S. 8 unten.

Für eine Koppelung von natürlicher und künstlicher Intelligenz[2] scheint das *genus* Intelligenz zureichend zu sein. Es bezeichnet – bei aller noch zu leistenden begrifflichen Schärfung hinsichtlich der Differenz und Ähnlichkeit natürlicher und maschineller Intelligenz – den sprachlichen Koppelungspunkt für ein Zusammenwirken der beiden Entitäten. Damit wird aber sprachlich-konzeptuell bereits gesetzt, was empirisch-physisch erst noch zu zeigen wäre, dass es a) ein Etwas auf beiden Seiten gibt, dass mit dem Begriff Intelligenz gleichermaßen zutreffend beschrieben und miteinander verglichen werden kann, und b) diese Intelligenzen operativ so zusammenwirken können, dass ein Neues und Drittes daraus resultiert, das mehr und anderes ist als die Summe der Einzelintelligenzen[3].

Ohne auf die Sprachgeschichte des Wortes Intelligenz hier genauer eingehen zu können, lässt sich vorläufig zusammenfassen, dass der Begriff auf menschlicher Seite für ein bestimmtes Vermögen steht, in fordernden Situationen zu bestehen, indem jeweils Lösungen entwickelt, gefunden bzw. angewendet werden. Dabei bezeichnet der Begriff eher eine Sammlung von bzw. einen Rahmen für problemorientiertes Lösungsverhalten denn diskrete Fähigkeiten (vgl. Gardner

[2] Das SYnENZ-Symposium operierte in seinem Programm mit dem Paar „natürliche und künstliche Intelligenz“. Hier wird zudem häufig von menschlicher Intelligenz gesprochen, die nicht zuletzt im KI-Diskurs meist als Opposition zur KI die Vorstellung dominiert. Dabei ist keineswegs klar, was denn das Gegenteilpaar von KI wäre. Es ließe sich angesichts der natürlichen Künstlichkeit des Menschen (Plessner) ebenso argumentieren, dass die menschliche Intelligenz eine künstliche sei, wie dass KI eigentlich eine natürliche sei. Letzteres meint bspw. Kate Crawford: “AI is neither artificial nor intelligent. It is made from natural resources and it is people who are performing the tasks to make the systems appear autonomous” (Crawford 2021). Schließlich sind Vorstellungen der Natürlichkeit oder Künstlichkeit philosophisch komplex und etwa auf die Disponibilität für handelnde Personen bezogen (vgl. Hubig 2011). Am gewinnbringendsten ist es, nicht ontologische Zuordnungen zu den vermeintlich klaren Kategorien Natur, Kultur und Technik vorzunehmen, sondern interessengeleitet Natürliches, Kultürliches und Technisches/Künstliches an etwas zu analysieren. So sind die menschlichen Clickworker, die Crawford im Blick hat, je nach Zugriff etwas Menschliches, Nicht-Künstliches oder Natürliches *am Phänomen* KI, die Serverstrukturen und Prozessoren hingegen etwas Künstliches, Technisches *an* KI, ohne dass das jeweils andere aus dem Blick geriete, weil vorzuentscheiden wäre, ob KI denn nun kategorisch künstlich oder natürlich wäre. Wichtig ist für diesen Beitrag, dass kontrastierend zu einer wie auch immer gefassten künstlichen Intelligenz sowohl menschliche als auch nicht-menschliche natürliche Intelligenzen (wie die der Hunde oder Oktopoden) in den Blick genommen werden müssen.

[3] Angesichts der Vielzahl an Einzelfähigkeiten, die unter dem Zugriff *Intelligenz* verhandelt werden, stellt sich gewissermaßen innerindividuell die gleiche Frage erneut, nämlich, ob und wie aus dem Zusammenwirken dieser Einzelfähigkeiten die Intelligenz eines Wesens emergiert, die dann mit anderen intelligenten Wesen interindividuell zusammenwirken und dort etwas Neues und Drittes, wie es Fokus dieses Beitrages ist, emergiert.

2005). Was als intelligent gelten kann, ist so besehen eher eine durchaus subjektive Einschätzung eines Beobachters in Bezug auf eine spezifische Situation, in der ein Wesen sich intelligent verhält – oder auch nicht. Bereits die griechische Mythologie kennt die *metische* Intelligenz, benannt nach der Titanentocher Metis, die mit Zeus die Göttin Athene zeugte. Es handelt sich gerade nicht um die üblicherweise mit Intelligenz in Verbindung gebrachte Rationalität, sondern um

> eine besonders schillernde und wendige Form von Intelligenz, die aus der Situation heraus reagiert, anstatt planvoll besonnen in die Zukunft zu blicken. […] Diese außerordentliche Form der Intelligenz wurde verschiedensten Gestalten zwischen Olymp und irdischen Gefilden zugesprochen: erfindungsreichen, technisch gewandten (Halb-)Göttern wie dem Schmiedegott Hephaistos oder Prometheus […]; listenreichen Helden wie Odysseus […] oder außergewöhnlich schlauen Tieren wie dem Fuchs und besonders: dem Oktopus. (Wittmann und Ganser 2023, S. 59, mit Verweis auf Detienne et Vernant 1973).

Der Begriff der Künstlichen Intelligenz lässt sich als eine Analogisierung bzw. Übertragung des menschlichen Konzepts auf Maschinen verstehen (zur Geschichte vgl. Seising 2021). Entsprechend werden Verständigungsangebote für künstliche Intelligenz unterbreitet wie:

> Künstliche Intelligenz (KI) bezeichnet Systeme mit einem ‚intelligenten' Verhalten, die ihre Umgebung analysieren und mit einem gewissen Grad an Autonomie handeln, um bestimmte Ziele zu erreichen. KI-basierte Systeme können rein softwaregestützt in einer virtuellen Umgebung arbeiten (z.B. Sprachassistenten, Bildanalysesoftware, Suchmaschinen, Sprach- und Gesichtserkennungssysteme), aber auch in Hardware-Systeme eingebettet sein (z.B. moderne Roboter, autonome Pkw, Drohnen oder Anwendungen des ‚Internet der Dinge'). (HLEG AI 2019)

Für eine Definition ist das zu unpräzise. Es ist aber durchaus sinnvoll, in Bezug auf noch unbekannte Phänomene oder Entitäten nicht mit einer philosophischen Definition eines Begriffes anzufangen, sondern eher damit zu schließen:

> Es gibt, nimmt man das beim Wort [sc. Kants Mahnung, ‚daß man es in der Philosophie der Mathematik nicht so nachtun müsse, die Definitionen voranzuschicken, als nur etwa zum bloßen Versuche'; Kant 2005, Bd. II, S. 625], so etwas wie ein experimentelles Stadium des Begriffsgebrauchs in der Philosophie, in dem es um die Bewährung der Leistungsfähigkeit von Begriffen, nicht um die Verifikation oder Falsifikation von Hypothesen geht. (Blumenberg 2010, S. 11)

Analysiert man diese Beschreibung genauer, so fallen Merkmale auf, die, einzeln betrachtet, den Charakter einer Tautologie haben (Künstliche Intelligenz bezeichnet Systeme mit intelligentem Verhalten) oder unscharf in der Beschreibung bleiben (KI-Systeme handeln mit einem *gewissen* Grad an Autonomie). Wir lesen das als einen Hinweis, dass für einen Vergleich der beiden Entitäten, Mensch einerseits und Maschine andererseits, das *tertium comparationis* Intelligenz nicht in einer Weise spezifiziert worden ist, dass ein aussagekräftiger Vergleich bzw. eine präzisere Relationierung möglich erscheint. Es lässt sich sehr wohl sagen, dass die sog. Künstliche Intelligenz bei der Berechnung bestimmter Aufgaben schneller ist als der Mensch[4], dass sie bei der Mustererkennung (z. B. histologische Analyse) komplexitätsadäquater und präziser ist als (die meisten fachlich kundigen) Menschen. Aussagen dieser Art lassen sich zweifelsohne treffen. Damit ist aber noch nichts gewonnen für den Koppelungspunkt (empirisch und konzeptionell) und für die Art des Zusammenwirkens (empirisch und konzeptionell).

2 Zusammenwirken von natürlicher, menschlicher und künstlicher Intelligenz

2.1 Zusammenwirken – einleitende Bemerkungen

Die Dynamik der Technisierung und Digitalisierung bedeutet eine zunehmende Durchdringung der Lebenswelt und damit alltäglicher Handlungszusammenhänge mit automatisierten, teils eigendynamischen, interaktiven und künstlich intelligenten Systemen. So treffen technische und menschliche Wirkpotentiale in geteilten Handlungsräumen aufeinander, was mit einer Vielzahl von Bezeichnungen belegt wird, wie Interaktion, Kooperation oder Koaktion bis hin zu *shared autonomy* bzw. der Annahme von hybriden, menschlich-technischen Handlungskollektiven.

Dabei stellt die Formulierung eines *Zusammenwirkens* von Mensch und Technik beziehungsweise von menschlicher und künstlicher Intelligenz einen Versuch dar, mensch-technische Wirkverschränkungen neutraler beziehungsweise weniger voraussetzungsreich zu fassen. So suggeriert z. B. die Rede von ‚geteilten Handlungsräumen' bereits ein Handlungsvermögen bei allen Beteiligten, das in einem gemeinsamen Handlungsraum zusammengeführt werden könnte. Je nach

[4] Das trifft auch für einen einfachen Taschenrechner zu, der deshalb aber nicht als ‚intelligent' betrachtet wird. Es müssen also noch weitere Merkmale hinzukommen, um das Attribut *konzeptionell* (und wohl weniger physisch-technisch) zu rechtfertigen.

Handlungskonzept – nämlich dann, wenn Intentionalität essenziell ist –, ist dies für technische ‚Handlungsträger' höchst problematisch und allenfalls metaphorisch zu verstehen. Entsprechend suggeriert die Rede von einer *shared autonomy* bezogen auf Mensch und Technik, dass bei allen Beteiligten Autonomie angenommen werden könnte, was gegenüber technischen Systemen, zumindest mit anspruchsvollem Autonomiekonzept, eine Anthropomorphisierung darstellt (vgl. Gransche 2024). Andererseits stellen Formulierungen wie *human being in the loop* oder der Mensch als letzte Kontroll- und Entscheidungsinstanz (vgl. Sturma 2004) technizistische Perspektiven auf menschliche Handlungspositionen dar, die ein leibliches, emotionales, intentionales Wesen auf eine technische Funktion (beispielsweise Schalter oder Filter) in einem systemischen Kontext reduzieren.

Der Versuch, diese voraussetzungsreichen Begriffe mit einer neutraleren Formulierung zu umgehen, ist sinnvoll, um sich die begrifflichen Suggestionen bewusst machen zu können. Allerdings stellt unsere Sprache leider keine voraussetzungslosen, neutralen Begriffe zur Verfügung, so dass auch die Formulierung des Zusammenwirkens einer Analyse ihrer Implikationen, sprachlicher Präsuppositionen und metaphorischer Gehalte unterzogen werden muss. Wie ein genauerer Blick auf die Formulierung *Zusammenwirken* im Rückgriff auf ihre griechischen und lateinischen Wurzeln – *energeia* und *cooperatio* – zeigt, unterstellt nämlich die Rede von Zusammenwirken eine Mindestgleichheit und Gleichartigkeit der zusammenwirkenden Entitäten. Diese Gleichartigkeit sollte entsprechend beim Gebrauch der Formulierung vorab festgestellt und absichtlich gemeint sein, sodass in sachadäquater beziehungsweise sinnhaft tauglicher Weise von einem Zusammenwirken die Rede sein kann, ohne einigen der beteiligten Agenten unter der Hand metaphorisch Wesenseigenschaften zuzuschreiben, die nicht gemeint sind, die jedoch über solche metaphorische Annäherung dennoch hergestellt würden. Dieses Herstellen von Ähnlichkeit durch Metaphorisierung kann, wenn sie nicht metaphorisch, sondern eigentlich verstanden wird, zu problematischer Verständnis- und Handlungsorientierung führen: Wenn beispielsweise von einem Zusammenwirken menschlicher/natürlicher und technischer Intelligenz die Rede ist, dann führt diese Art des Redens zur Suggestion ihrer teilweisen Gleichartigkeit, was es erschwert, ihre Unterschiede – und das bedeutet mitunter die relevanten und wesentlichen Unterschiede – sehen und berücksichtigen zu können. Dass jedoch Zusammenwirken auch Gleichartigkeit voraussetzt, ist in unserem alltäglichen Sprachverständnis keineswegs evident und bedarf daher näherer Erläuterung.

2.2 Zusammenwirken, *Tosamne wyrcan,* Kooperation, Synergie

Im Folgenden wird ausgehend von den beteiligten Begriffen selbst auf die impliziten Bedeutungsgehalte hingewiesen, die für die Auseinandersetzung mit dem Zusammenwirken verschiedener Intelligenzen einschlägig sind. Dazu ist es notwendig, den Begriffen zunächst etymologisch auf den Grund zu gehen, was in den folgenden beiden Absätzen vorgelegt wird und gewissermaßen sprachlich sezierende Vorarbeit für den dann folgenden deutenden Befund darstellt.

Das deutsche *Zusammenwirken,* zurückreichend u. a. auf das altenglische *Tosamne wyrcan,* hat sein lateinisches Äquivalent in der *Cooperatio* sowieso sein griechisches in der *Synergie.* Alle drei Begriffe kombinieren ein Wirken-/Werken-Verb (wirken, *operari, ergein*) mit einem kollektiven Relationsindikator (zusammen-, *co-, syn-*). Der Blick auf alle drei Kombinationen ist instruktiv für die darin enthaltenen Bedeutungselemente. Die Elemente von Zusammenwirken haben dabei nicht lateinisch-griechischen, sondern westgermanisch-altsächsischen Ursprung (DWDS zusammen)[5], weshalb eine bilinguale Analyse in Deutsch und Englisch aufschlussreich ist. Zudem verweisen alle drei Formulierungen auf Wurzeln des gemeinsamen Proto-Indo-Europäischen (PIE)[6] zurück.

Dass *Synergie* und *Kooperation* Synonyme sind, wird in folgenden Beschreibungen ersichtlich:

> Synergy: "related to *synergein* 'work together, help another in work' from *syn-* 'together' (see syn-) + *ergon* 'work' (from PIE root *werg- 'to do').[7] Meaning 'combined activities of a group' is from 1847; sense of 'advanced effectiveness as a result of cooperation' is from 1957" (OED synergy).

[5] Für die Deutschen Worterklärungen wird auf das *Digitale Wörterbuch der deutschen Sprache DWDS* (https://www.dwds.de/) zurückgegriffen, für die Englischen auf das *Online Etymology Dictionary* (https://www.etymonline.com/), das einen digitalen Zugang zu einer ganzen Reihe etymologischer Werke darstellt (deren Liste findet sich hier: https://www.etymonline.com/columns/post/sources?utm_source=etymonline_footer&utm_medium=link_exchange). Zitiert wird im Folgenden je mit Kürzel und Lemma, also z. B. ‚DWDS wirken' oder ‚OED synergy'.

[6] "PIE, 'Proto-Indo-European' the hypothetical reconstructed ancestral language of the Indo-European family. The time scale is much debated, but the most recent date proposed for it is about 5,500 years ago." https://www.etymonline.com/columns/post/abbr?utm_source=etymonline_footer&utm_medium=link_exchange, zuletzt zugegriffen 28.07.2023.

[7] Der Asterisk (*) wird im Folgenden wie im OED verwendet, nämlich: "asterisk (*): Words beginning with an asterisk are not attested in any written source. Some have been reconstructed by etymological analysis, such as Proto-Indo-European *ped-, the root of words for 'foot' in most of its daughter tongues. In other cases they are hypothetical words or forms of words

> Cooperation: "'the act of working together to one end' 1620s, from French *coopération,* or directly from Late Latin *cooperationem* (nominative *cooperatio*) 'a working together' noun of action from past-participle stem of *cooperari* 'to work together' from assimilated form of *com* 'with, together' (see com-) + *operari* 'to work' from PIE root *op- 'to work, produce in abundance.'" (OED cooperation).

Beide verweisen genauso wie *wirken* – in dem die altenglische Herkunft des *wircan* (DWDS wirken) klar durchklingt – zentral auf *Arbeiten/work:*

> Work: "Old English *weorc, worc* 'something done, discrete act performed by someone, action (whether voluntary or required), proceeding, business; that which is made or manufactured, products of labor' […] **werka- 'work'* from PIE […] root **werg-* 'to do.'"

Beide PIE-Wurzeln **werg-* und **op-* bezeichnen ein Tätigsein, wobei **werg-* grundsätzlich ‚tun' und **op-* spezifischer ‚arbeiten' und ‚reichlich produzieren' bedeutet (OED *werg- und *op-). Die deutsche Alltagssprache verwendet das zu **werg-* fast gleichlautende ‚Werk' sowie ‚werken/wirken', aber auch Begriffe wie ‚Organ' – wörtlich *das, was arbeitet,* von **wergano* (OED organ) – ‚Orgie' oder ‚Allergie' – wörtlich etwa Fremdarbeit oder Fremdeinwirkung (DWDS Allergie) – oder ‚Ergonomie' als Lehre von der Arbeit, das das **werg-* in Form des griechischen *ergon* wie in Synergie enthält. Andererseits werden zahlreiche Latinismen verwendet, in denen im Kern Arbeit und Produzieren über das **op-* enthalten ist, wie ‚Operation' oder ‚Opus/Oper'. Interessanterweise enthält wie *Org*an auch die *Org*ie das *ergon,* was noch erkennen lässt, dass die damit bezeichneten orgiastischen Riten etwa des Dionysos-Kultes mit exzessivem Tanzen, Trinken, Singen und ‚Kopulieren' zelebriert wurden, die somit durchaus Workout-Charakter hatten (OED Orgy). So liegt die sexualisierte Bedeutung von ‚es tun', ‚es miteinander treiben' oder ‚doing it/doing someone' durchaus nahe. Nicht überraschend findet sich sowohl im *Kopulieren* (zusammenkoppeln, verbinden) als auch im *Coitus* (zusammen kommen/gehen) das Co/Zusammen-, wobei – Kopulieren technizistisch als Koppeln gefasst – eine Kopplungsfähigkeit im Sinne der Passfähigkeit als Mindestgleichheit vorauszusetzen ist. Kopulieren und Koitieren können nur untereinander (mindest)passfähige Wesen – gleiches gilt für Kooperieren/Zusammenwirken.

So stellt sich die Frage, was denn genau anknüpfend an **werg-* und **op-* zusammen getan wird bzw. getan werden kann. Offensichtlich hängt

---

that might have, but didn't, come into use in a modern language (Modern English *astronomian, if Middle English astronomyen had survived). Or they are presumed forms in ancient languages of words that are attested only in oblique or derived forms" (ebd).

das von den im ‚Co' zusammengefassten Entitäten ab: So können Menschen und (Haus)Tiere etwa zusammenleben und können (manche) Menschen mit (manchen) Hunden oder (manchen) Raubvögeln zusammen jagen oder bei der Jagd *ko*operieren und *ko*mmunizieren, nicht jedoch *kon*versieren, also in Form einer gesellschaftlichen sprachlichen *Kon*versation Umgang miteinander pflegen bzw. zusammen sprechen. Welche Implikationen liegen in den Verben, die mit dem Co-/Syn-/Zusammen- kombiniert werden? Also etwa: zusammensetzen/komponieren, das zusammen Hervorgebrachte/Koproduktion, Zusammenarbeit/Kollaboration, zusammenrauben/kompilieren, zusammenschlagen/konfligieren, zusammenleben/siehe konvivial oder Symbiose, (zusammen)verflochten/komplex, (zusammen)gefaltet/kompliziert usw.

Die neutrale Basisform wäre hier ‚tun' im Sinne des Verrichtens einer Tätigkeit bzw. dem Verursachen von Ereignissen (DWDS tun); ‚wirken' allgemein im Sinne von arbeiten, tätig sein, Einfluss ausüben oder Eindruck machen, bedeutet einen Effekt und eine Wirkung zu zeitigen (DWDS wirken); ‚werken' im Sinne von arbeiten, tätig/werktätig sein bedeutet *handelnd wirken* und ist somit aktivischer als bloßes auch passiv mögliches wirken zu verstehen (DWDS werken); ‚arbeiten' bezeichnet die zweckgerichtete körperliche und geistige Tätigkeit des Menschen, sowie ‚Arbeit' zudem das Produkt dieser Tätigkeit oder das Werk als Ergebnis des Werkens (DWDS Arbeit); ‚produzieren' bedeutet wörtlich hervorziehen (pro-ducere), also nach vorne bringen bzw. erzeugen und die Koproduktion ist somit wörtlich ein gemeinsames (an einem Strang) Ziehen (DWDS produzieren); schließlich ‚handeln' bedeutet etwas tun, tätig sein (auch Handel treiben oder feilschen), von ‚in die Hand nehmen' (DWDS handeln) – in handlungstheoretischer Hinsicht bezeichnet es jedoch, intentional Zwecke (mögliche-präferierte Ereignisse) zu realisieren. Denn über die alltagssprachliche Bedeutung des schlichten Tätigseins hinaus werden Handlungen handlungstheoretisch gefasst als

> …die Umsetzung eines gewollten (oder gesollten) Zweckes in die Realität […] als jede reflektierte, planmäßige und zielstrebige Aktivität (H[andeln], Herstellung, Denken) überhaupt […]. Nur dem Menschen (als reflektierendem Wesen) können H[andlung] und Tat zugeschrieben werden; das Analogon beim Tier heißt ‚Verhalten', in der anorganischen Natur ‚Prozeß' (Derbolav 2010, S. 992).

Von ‚geteilten Handlungsräumen' zu sprechen, setzt strenggenommen also Handlungs*kom*patibilität und Handlungsvermögen aller Beteiligten voraus. Die Handlungskriterien ‚gewollt, reflektiert, geplant, als Ziel erstrebt' sind dabei nicht ohne weiteres auf technische Systeme übertragbar; Josef Derbolav beschränkt entsprechend Handlung auch auf reflektierende Wesen, also Menschen. Diese

kursorische Übersicht von ‚tun' bis ‚handeln' verdeutlicht einige Bedeutungselemente, die bei der jeweiligen Verbwahl mitschwingen, aber in Bezug auf technische Systeme nur metaphorisch verwendet werden können. Wenn ‚Arbeit' (gedacht als anthropologische und nicht als physikalische Größe) wie oben zitiert „zweckgerichtete körperliche und geistige Tätigkeit des Menschen" ist, dann kann keine Entität im eigentlichen Sinne ‚arbeiten', die keine *Zwecke* und keine *präferierte Richtung* kennt oder wollen kann, die keinen Körper oder keinen Geist (von Kognition bis Selbstbewusstsein) hat[8]. Genauso wie nichts und niemand handeln kann, das/der/die nicht wollen, reflektieren, planen, streben kann. Es stellt sich die Frage, welche Art von ‚tun' Technik bzw. technischen Systemen als Analogon zu menschlichem Handeln zukommt und wie dies benannt werden kann. Als Teil der anorganischen Natur sollte demnach zunächst nur sinnvoll von Prozessen bei technischen Aktionspotenzialen gesprochen werden können. Jedoch entsteht mit zunehmender Eigendynamik, Hochautomatisierung oder ‚Autonomie' (vgl. Gransche 2024) einiger künstlicher Systeme (‚Agenten'?) die Frage, ob diese sich noch genauso klar in die Trias Mensch-Tier-Anorganisches bzw. Handlung-Verhalten-Prozess einordnen lassen wie Steine und Toaster, oder ob ihnen ein eigener, spezifischerer Bereich des Tuns, Wirkens und Beeinflussens zugesprochen werden müsste. Dass ‚autonome Systeme' oder ‚künstlich intelligente Agenten' andere und andersartige Handlungsrelevanz haben als Steine, überzeugt. Derzeit wird dem aber verbreitet damit begegnet, ihr Agieren bzw. Prozessieren mit menschlichen Tätigkeitsformulierungen zu fassen, was sie dann in die Nähe von Handlungs- und Entscheidungsfähigkeit rückt und problematischerweise Aspekte wie Reflektiertheit, Intentionalität, Präferenzen und Zielstrebigkeit suggeriert.

2.3 Die Kollektivrelationsindikatoren Ko-, Syn- und Zusammen

Es bedarf noch eines weiteren Analyseschrittes, nämliches eines Blickes auf die jeweiligen Kollektivrelationsindikatoren Ko-/Syn-/Zusammen:

> Das lateinische *co* bedeutet: "'with, together' from Latin *com* […] *cum* 'together, together with, in combination' from PIE **kom-* 'beside, near, by, with'" (OED com-)
> .

[8] Freilich kennt die Physik den Begriff der Arbeit als die einem Körper zugeführte Energie. Diese versteht sich intentions- und präferenzenfrei, darf aber als Analogiebildung zur menschlichen Arbeit verstanden werden.

> Das griechische Syn bedeutet "'together with, jointly; alike; at the same time' also sometimes completive or intensive, from Greek *syn* (prep.) 'with, together with, along with, in the company of' from PIE *ksun- 'with'" (OED syn-).

Wie bei *ergon* und *operare* (**werg-* und **op-*) gehen hier *co-* und *syn-* auf bedeutungsähnliche PIE-Wurzeln zurück (**kom-* und **ksun-*), die beide zunächst *mit, gemeinsam, zusammen* heißen, aber im *syn*-Falle bereits mit *alike* das Ähnlichsein, die Gleichartigkeit einbringen. Dies wird im Altdeutsch-Altenglisch Vergleich augenfällig. Das Adverb *zusammen* bedeutet ‚gemeinsam, miteinander, beisammen, beieinander, insgesamt' (DWDS zusammen);

> Together: "Old English *togædere* 'so as to be present in one place, in a group, in an accumulated mass' …from Proto-Germanic **gaduri-* 'in a body' from PIE **ghedh-* 'to unite, join, fit' […] In reference to single things, 'so as to be unified or integrated' […] German cognate *zusammen* has as second element the Old High German verbal cognate of English *same* (Old English also had *tosamne* 'together')" (OED together).

Together hatte im Altenglischen also auch die Form *tosamne,* das hörbar am deutschen *zusammen* liegt. Interessanterweise hat damit unser *zusammen* das englische *same* im Kern,[9] was wiederum "identical, equal; unchanging; one in substance or general character" (OED same) bedeutet. Was *zusammen* ist, ist demnach auch *gleich (same),* da sich *zusammen (tosamne)* auch als *zugleichen* lesen ließe. *Syn-* bedeutet zusammen und gleich *(together, alike* s. o.) und *zusammen* selbst bedeutet im Zusammensein ein Gleichsein.

Die griech. Kopula σύν liefert in ihren antiken Ursprüngen für unsere Überlegungen einen bemerkenswerten Akzent. Σύν wird „seit jeher mit einem soziativen Dativ konstruiert, bedeutet zusammen und drückt ein Zusammensein und ein Zusammenkommen aus" (Grundmann 1964, S. 767). σύν gilt als „das Normalwort für mit" und ist gerade bei Homer anzutreffen (a.a.O., S. 768). Hingegen überwiege bei

> den Philosophen, Geschichtsschreibern und Rednern μετά c Gen bei weitem. […] Hier findet sich eine Verteilung der Bedeutungsfunktionen, die gegenüber dem Ionischen, dem poetischen und nachklassischen Sprachgebrauch in dieser Weise fremd

[9] Auch das deutsche *samt,* das als insge*samt* in der DWDS-Umschreibung von *zusammen* vorkommt, geht wie das englische *same,* mit dem es ein Minimalpaar bildet, auf denselben Ursprung, das PIE **samos* zurück (OED same); *samt* und *same* bedeuten etymologisch das gleiche, und damit bezeichnen wir heute Sammlungen oder Entitäten, die beisammen sind, *insgesamt* und dabei immer *auch gleich* bzw. *zusammengehörig* wie etwa die öffentlich-rechtliche Körperschaft der *Samt*gemeinde im Norddeutschen.

> ist: σύν bedeutet *mit* im Sinne einer dadurch hergestellten engen Gemeinschaft, der Mithilfe und Unterstützung, während μετά c Gen die Begleitung durch Personen, Dinge oder Umstände ausdrückt. (ebd.)

Auch wenn diese sprachlichen Unterscheidungen im Verlauf der Zeit und Übersetzungen eingeebnet worden sein mögen, ist doch das Moment einer „hergestellten engen Gemeinschaft, der Mithilfe und Unterstützung" für unseren Sprachgebrauch nach wie vor relevant und aufschlussreich. Nicht zuletzt in dem programmatischen Verständnis von ‚Synenz' wird eben dieses Verständnis aufgriffen: eine hergestellte, enge Gemeinschaft verschiedener Intelligenz(en), die sich untereinander Mithilfe und Unterstützung leisten bzw. diese einfordern und gewähren oder auch verweigern. Gerade letzteres, eingeforderte Mithilfe zu verwehren, wäre Kennzeichen gleichwertiger Entitäten und eines Zusammenwirkens im starken Sinne; denn ohne Verweigerungsmöglichkeit liegt ein Zwangs- oder Instrumentierungsverhältnis vor.

2.4 Konsequenzen für das Zusammenwirken von NI und KI

Von einem Zusammenwirken von natürlicher Intelligenz (NI) und künstlicher Intelligenz (KI) zu sprechen, ist nach dem Gesagten also keineswegs neutral, wenn auch – dies unbenommen – vermutlich immerhin neutraler und weniger voraussetzungsreich als die Rede von geteilten Handlungsräumen oder Handlungsträgerschaften von Mensch und Technik und erst recht als das Heer von Anthropomorphismen im Kontext der KI wie ‚algorithmisches Entscheiden' oder ‚autonome Roboter' etc. (vgl. zu diesem „metaphorischen Heer der KI" Gransche und Manzeschke 2024). Vorausgesetzt wird bei der Verwendung von *Zusammenwirken, Kooperation, Synergie* sowie allen genannten und möglichen Varianten im Ausgang der PIE-Wurzeln **Ksun-*werg* und **Kom-*op* sowohl Gemeinsamkeit als auch Gleichheit. Wer vom Zusammenwirken von NI und KI spricht (und dies so meint), hat allein durch die Entscheidung, *Zusammenwirken* auf Menschen und Maschinen zugleich anwenden zu können, deren kollektivierungsermöglichende Mindestgleichheit *vor-entschieden.* Wer so spricht, kommuniziert also (bewusst oder nicht), dass er Menschen und Maschinen für der ‚Substanz oder dem allgemeinen Charakter/Wesen nach' ("one in substance or general character", s. o.) gleich genug hält, um sie als zusammenwirkungsfähig, *ko*pplungsfähig, *kom*patibel vorzustellen.

Dabei sind – allen Anthropomorphisierungen von Technik und Technomorphisierungen von Menschen zum Trotz – Menschen und Maschinen, Automaten, KI-Systeme, Roboter etc. der Substanz und dem Wesen nach höchst unterschiedlich. Womit sich die Frage stellt, ob die Rede vom Zusammenwirken bezüglich dieser Gleichheitsunterstellung korrekterweise als metaphorisch anzusehen ist. Falls ja, löst das nicht direkt das Problem, sondern schafft zunächst neue, da die Metaphorisierung ihrem Wesen nach Ähnliches an Unähnlichem bzw. Selbiges an Differentem betont und performant so Ähnlichkeit erzeugt, wo zunächst keine ist oder – mit Paul Ricœur formuliert – die Metapher lehrt uns Ähnlichkeit zu sehen (Ricœur und Jüngel 1974, S. 54). "In metaphor, 'the same' operates *in spite of* 'the different'" (Ricœur 2003, S. 232). Damit sind Metaphorisierungen, recht verstanden als semantische Annäherung von imaginativ Fernem (mit Aristoteles aber der Art oder Form nach Gleichem)[10] zu sehen, das aber – und mit diesem Metaphernbewusstsein lösen sich die Probleme – das Differente nicht tilgt, sondern anerkennt und nicht gegen das Differente, sondern *trotz (in spite of)* des Differenten Gleiches hervorhebt. Auch die Metapher eines Zusammenwirkens von NI und KI unterstellt also eine Gleichheit der Zusammenwirkenden, allerdings – und das ist ein wichtiger Vorteil im Gegensatz zu metaphorisch unreflektierter bzw. vermeintlich eigentlicher Rede – im vollen Bewusstsein und unter fortgesetzter Berücksichtigung und Anerkennung ihrer Unterschiede. Metaphorisieren bringt Differentes zusammen *(tosamne),* stellt es somit als Gleichartiges *(same)* vor, ohne aber auf die eigene Angleichungsoperation hereinzufallen.

Bei der Rede vom Zusammenwirken von KI und NI wäre demnach zu reflektieren, inwiefern die offensichtlich differenten ‚Intelligenzen' (wenn beide als Intelligenz überhaupt bezeichnet werden können) oder wirkenden Entitäten tatsächlich und tauglicherweise gleich oder zumindest ähnlich genug sind sowie in welcher Hinsicht sie dies sind. Ein Blick in aktuelle KI-Debatten und KI-Diskursüblichkeiten zeigt, dass vornehmlich unreflektiert metaphorisierend Ähnlichkeit in sprachlich performativer Weise postuliert wird, die so von den wenigsten tatsächlich angenommen oder gemeint wird (vgl. dazu Heinlein und

[10] Aristoteles verwendet syngenôn (verwandt) und homoeïdôn (gleichgestaltig, gleich aussehend). Wie gleichartig Metaphorisiertes der Form und Art nach sein muss, ist eine Sache des Maßes, was Ricœur mit Aristoteles in den Kontext der Angemessenheit guter Metaphern setzt: "Aristotle was aware of this strictly predicative effect of resemblance when he considered, among the 'virtues' of good metaphors, that of being ‚appropriate' (Rhetoric 3: 1404 b 3). He saw in this a sort of 'harmony' (1405 a 10). On guard against 'far-fetched' metaphors, he recommends that metaphors be derived from material that is 'kindred' (syngenôn and 'of like form' (homoeïdôn), such that once the expression is produced, it will appear clearly that the 'names' involved are 'near of kin' (hoti sungenes) (1405 a 37)." Ricœur (2003, S. 230).

Huchler 2024). Selbst KI-Forscher glauben nicht (oder gerade diese nicht), dass KI-Systeme *autonom* wären, in dem Sinne, dass sie Gesetze als eigene präferenzfundiert anerkennen oder ablehnen könnten, dass Algorithmen nach eigenem Willen etwas *entscheiden,* dass Roboter *handeln* in dem Sinne, dass sie intentional selbst gewählte Zwecke mit nach eigenen Kriterien als dazu tauglich erachteten Mitteln realisieren etc. Zu vermeiden wäre jedenfalls, auf die eigenen sprachlichen Angleichungsoperationen mit der Bildung entsprechender Gleichheitsurteile ‚hereinzufallen', also KI und NI für gleich genug für ein *zusammen/tosamne* zu halten, bloß weil ein Zusammenwirken, das solche Gleichheit impliziert, sprachlich postuliert wird. Um die Tauglichkeit oder das Gerechtfertigt-Sein einer Formulierung wie Mensch-Technik-Kooperation oder Zusammenwirken von NI und KI beurteilen zu können und zu unterscheiden, ob, in welcher Hinsicht und welchem Umfang metaphorisierend gesprochen wird, stellt sich zentral die Frage, wie gleich die im Zusammentun, -wirken etc. vorgestellten Entitäten tatsächlich sind. Wie gleich und wie zusammen sind und können Menschen, Tiere, Pflanzen, Maschinen, Dinge etc. sein? *Wie gleich* zwei Entitäten sind, erfährt man durch *Vergleiche.* Selbst ein Urteil wie das der Unvergleichbarkeit zweier Entitäten wegen im Extremfall gänzlicher Differenzen der Eigenschaften, stellt bereits das Ergebnis eines Vergleiches dar mit dem hypothetischen Ergebnis nämlich 0 % Gleiches und 100 % Differentes. So gesehen können Hunde und Menschen kulturhistorisch belegt so gut zusammenwirken (etwa bei der Jagd), weil sie sich – im Unterschied zu Mensch und KI-Systemen – in für diese Kooperation relevanten Kriterien sehr viel ähnlicher sind. Ihnen (Mensch und Hund) ist z. B. gemeinsam: ein leibliches Weltverhältnis mit leiblicher Sinnen-Weltvermittlung, damit verbunden direkte Präferenzen (satt vor hungrig, warm vor heiß und kalt, unversehrt vor verletzt, in Gemeinschaft vor allein/einsam etc.), Unempfindlichkeit gegenüber sinnlich nicht Indiziertem (wie etwa elektrischer Felder mangels Lorenzinischer Ampullen) und Leiden an Sinnenüberreizung, Leidensfähigkeit an den Präferenzen entgegenstehenden Reizen (Schmerzen, Einsamkeit…), Empathie am Leiden der anderen usw. Kein KI-System bevorzugt es, in existenziellem Sinne ‚an' zu sein, wie es Lebewesen (meist) bevorzugen, zu leben und dafür zu kämpfen bzw. dem Leben Dienliches als Zwecke zu setzen und ihm Abträgliches zu vermeiden. Entsprechend versteht (fast) jeder Mensch und jeder Hund aus der geteilten leiblichen Weltbezogenheit heraus (allen sinnenbezogenen und kognitiven Unterschieden zum Trotz), was es heißt und wie es sich anfühlt, um sein Leben zu kämpfen und Schmerzen zu erleiden, weshalb die entsprechende (Not-)Hilfe auch auf eine leiblich-lebendig evidente Grundlage fällt und (manche) Hunde selbst ohne Abrichtung Menschen aus Notsituationen (z. B. vor dem Ertrinken) retten bzw. zu retten versuchen. Kein KI-System könnte

auf dieser Gleichheitsgrundlage (bei allen bleibenden Unterschieden) bei der Lebensrettung oder Leidenslinderung (z. B. Trösten, Wärmen) mit Menschen oder Hunden, mit natürlichen Intelligenzen zusammenwirken, weil das nötige *same* für das *zusammen* fundamental fehlt. Scheinbare Gleichheiten wie Sprachgenerierung und -verstehen zwischen Menschen und manchen KI-Systemen dürfen hier nicht über die fundamentalen Differenzen – kein Leben, kein Leib, kein Leiden, keine Präferenzen – hinwegtäuschen, zumal die technische Sprachverarbeitung bei aller Ähnlichkeit des Outputs sich völlig von der menschlichen unterscheidet, sodass menschliche und manche tierische Kommunikation (z. B. Hunde oder Primaten) trotz Nonverbalität synergetischer zusammenhängen als sprachliche Mensch-Maschinen-Interaktion.

Zwischenfazit

Verschiedene Agenten bzw. Intelligenzen wie KI und NI können nur im oben dargestellten engeren Sinne *zusammenwirken,* wenn sie einander ähnlich sind, d. h. wenn sie in einem Mindestmaß *gleich* und *gleichartig* sind. Diese Mindestgleichartigkeit erscheint technizistisch gefasst dann als Kopplungsfähigkeit. Was zu verschieden ist, um wirksam gekoppelt werden zu können, kann auch nicht zusammenwirken. Wie *gleich* etwas ist, muss über *Vergleiche* festgestellt werden. Vergleiche bedürfen zur Ähnlichkeitsfeststellung der Messung derjenigen Aspekte, die zwei zu vergleichenden Entitäten gemeinsam oder nicht gemeinsam sind. Das je zu Vergleichende muss dazu an einem gemeinsamen und zu beiden passenden Maß gemessen werden, sonst resultiert kein messdatenbasierter Vergleich, sondern Inkommensurables wie: Was ist lauter, ein Kilo Federn oder 70 km/h? Das zu Vergleichende muss aneinander oder an einem kommensurablen Dritten angelegt werden können, um Unterschiede und Überschneidungen messen zu könne. Damit verweist die gelingende Rede von einem Zusammenwirken verschiedener Intelligenzen – da abhängig von einer Mindestgleichheit für ein solchen Wirken – auf ein Messproblem. Dieses Messproblem ist wiederum in der grundlegenden Mehrdimensionalität des Messens überhaupt zu betrachten, worauf der nächste Teil dieses Beitrages eingeht.

3 Messen

Messen ist eine für naturwissenschaftlich-technische Operationen grundlegende Tätigkeit; ohne exakte Messung der relevanten Größen, ohne ihre vergleichende Relationierung und ohne Orientierung der Apparate oder Prozesse an den Messwerten ist technisches Handeln nicht möglich. Messen ist aber auch eine für

soziale Operationen grundlegende Tätigkeit, wenn auch mit anderen Exaktheitsansprüchen; ohne die Orientierung als ethisch oder sozial angemessen bzw. unangemessen, ist soziales Handeln nicht möglich (vgl. Bellon et al. 2022). Der Vorgang des Messens ist ein möglicher Kopplungspunkt zwischen einerseits den physisch-empirischen Messgegenständen (der Welt oder Wirklichkeit) sowie andererseits den durch Menschen gefundenen, gesetzten und angelegten Maßen (den Werten). Dabei kann das zu Messende das angelegte Maß nicht selbst enthalten oder begründen, sondern das Maß muss von außen hinzutreten und folglich vorab anderswo begründet und gerechtfertigt worden sein. Georg Simmel spricht hier instruktiv von getrennten, aber zusammenwirkenden Reihen, nämlich der Wirklichkeitsreihe einerseits und der Wertreihe andererseits:

> Man könnte die Reihen des natürlichen Geschehens mit lückenloser Vollständigkeit beschreiben, ohne daß der Wert der Dinge darin vorkäme – gerade wie die Skala unserer Wertungen ihren Sinn unabhängig davon bewahrt, wie oft und ob überhaupt ihr Inhalt auch in der Wirklichkeit vorkommt. Zu dem sozusagen fertigen, in seiner Wirklichkeit allseitig bestimmten, objektiven Sein tritt nun erst die Wertung hinzu, als Licht und Schatten, die nicht aus ihm selbst, sondern nur von anderswoher stammen können. […] Man macht sich selten klar, daß unser ganzes Leben, seiner Bewußtseinsseite nach, in Wertgefühlen und Wertabwägungen verläuft und überhaupt nur dadurch Sinn und Bedeutung bekommt, daß die mechanisch abrollenden Elemente der Wirklichkeit über ihren Sachgehalt hinaus unendlich mannigfaltige Maße und Arten von Wert für uns besitzen. In jedem Augenblick, in dem unsere Seele kein bloßer interesseloser Spiegel der Wirklichkeit ist – was sie vielleicht niemals ist, da selbst das objektive Erkennen nur aus einer Wertung seiner hervorgehen kann – lebt sie in der Welt der Werte, die die Inhalte der Wirklichkeit in eine völlig autonome Ordnung faßt. (Simmel 1930, S. 4–5)

Der Messvorgang bildet Wirklichkeitsreihen auf Wertreihen ab, und hierbei ist strenggenommen eine Trennung beider nicht möglich: Der theoretische Standpunkt ist bereits eine wertende Stellungnahme, die darüber bestimmt, welche Größen wie erscheinen. Dabei begründen die „unendlich mannigfachen Maße und Arten von Wert für uns“ Sinn und Bedeutung der Wirklichkeit. Das heißt Sinn und Bedeutung sind nicht in der Wirklichkeitsreihe, den empirisch erfassbaren Dingen und Ereignissen zu finden, auch nicht durch massiv datenverarbeitende musterkennende KI-Systeme: Was an Sinn, Bedeutung und Wert gemessen wird, tritt durch die Maße und die menschliche Praxis des Messens erst hinzu. Folglich muss *vor* dem Messen beurteilt werden, was als Maß und Wert überhaupt in Frage kommt oder mit anderen Worten: Es müssen Wertabwägungen und interessengeleitete Präferenzhierarchisierungen sowie die Aushandlung ihrer intersubjektiven Gültigkeit vorangehen. Die Idee einer objektiven Messung ist nicht nur dem

geläufigen Messtechniker-Sprichwort ‚Wer misst, misst Mist.' nach zurückzuweisen, sondern weil in Simmels Worten „selbst das objektive Erkennen nur aus einer Wertung seiner hervorgehen kann". Dieser Wert- oder Wertungsbasierung der Messung und damit aller Arten der Datenerfassung entkommen letztlich auch nicht KI-Systeme. Messen koppelt aber nicht nur Wert und Wirklichkeit, sondern auch beliebige Entitäten, deren Unterschiede gemessen werden sollen, nämlich an das gemeinsam angelegte Maß, zu dem wiederum das unterschiedliche aber eben kommensurable zu Messende passen muss. Sind NI und KI gleich genug, um zusammenwirken zu können? Die Annäherung an eine Antwort muss über Vergleiche und (Unterschieds)Messungen verlaufen und das wiederum bedarf einer Reflexion auf das Messen, die Maße und Werte.

3.1 Messen – geometrisch, arithmetisch, ästhetisch und moralisch

In seiner alltagssprachlichen Bedeutung meint *messen,* ein Maß für etwas, seine Größe (in Hinsicht auf Strecke, Fläche, Gewicht, Geschwindigkeit …) feststellen bzw. überprüfen. Das Griechische μέδεσθαι *(médesthai)* bedeutet: für etwas sorgen, an etwas denken, auf etwas bedacht sein und ist verwandt mit dem Lateinischen *meditari* (Griechisch: μέδομαι, *médomai*): nachdenken, nachsinnen (DWDS messen). Das Messen in Maßzahlen beruht einerseits auf der „Zuordnung von Zahlenwerten und numerischen Verfahren zu empirischen Größen" (Mainzer 1984, S. 862). Da sich aber nicht alle Größenverhältnisse ganzzahlig darstellen lassen, entwickelten bereits Eudoxos von Knidos und Archimedes geometrische Messverfahren, die erst in der Neuzeit arithmetisch ausgedrückt werden konnten. Maße lassen sich also nicht nur in Zahlen, sondern auch in Proportionen darstellen, die wie beim Goldenen Schnitt (Verhältnis von 1: 0,618) eher ästhetisch als arithmetisch wahrgenommen werden. Zugleich wurden diese Verhältnisse von den Pythagoreern als den gesamten Kosmos durchwaltende Maße (im Sinne von Maßstäben) der Schönheit, des Anstands und der Harmonie verstanden (vgl. Braun 2019, Bd. I, S. 306). Das messende und damit vergleichende Erfassen von Welt hat in seinen Anfängen stets eine über das rein Empirische und Numerische hinausgehende ästhetische und damit an einer Norm des Guten und Schönen orientierte Form. Deutlich wird dieser Zusammenhang beispielsweise daran, dass das Urteil *angemessenen* Verhaltens nicht etwa neutral eine Maßhaftigkeit meint, sondern gut im ethischen Sinne, also nicht nur ein Verhalten nach irgendeinem Maß, sondern nach einem *guten* Maß.

Messen ordnet Zahlenwerte empirischen Größen zu und bietet damit die Voraussetzung für eine weitere wichtige Operation: das Vergleichen. Das Vergleichen verschiedener Größen im Bereich der Zahlen über ein *tertium comparationis* beruht auf dem Konzept der Zahl bzw. – in frühen Vorstufen – einer Praxis des Zählens ohne Zahlbegriff (vgl. Ifrah 1993, bes. S. 21 ff.). Dieses Vergleichen erlaubt es dem Menschen, ‚angemessene' Lösungen für die sich ihm stellenden Aufgaben zu finden. Hierbei darf nicht übersehen werden, dass antike Maße mehr als empirisch-technische Daten waren, und Vergleiche ihren Maßstab in überempirischen, metaphysisch gegründeten Weltbildern hatten. Eine solche Kongruenz von empirischem Sachverhalt und ästhetisch-ethischem Sachgrund ist spätestens seit der Neuzeit nicht mehr gegeben (vgl. zum Ganzen: Blumenberg 1996).

Messen ist ohne eine physikalische Theorie der Messobjekte und der Messapparate unmöglich. Die physikalische Theorie, die hier zugrunde gelegt werden muss, ist die der Quantenmechanik. Dabei jedoch

> entsteht die scheinbar paradoxe Situation, daß die ‚an sich vorhandene' Wirklichkeit nur beschrieben werden kann, wenn ein Teil der möglichen Information fehlt, *genau* beschreiben kann man nur Möglichkeiten. – Diese Erkenntnis scheint den Schlüssel zum Verständnis der Quantenmechanik zu enthalten: Eine Wirklichkeit ‚an sich' gibt es, genau genommen, nicht oder allenfalls als Grenze einer Näherung; sie zu unterstellen ist andererseits unerläßlich, wenn überhaupt etwas objektiv beschrieben werden soll. (Drieschner/Wegner 1980, Sp. 1167)

Der Informationsverlust, der hier angesprochen wird, bezieht sich auf den Übergang von einer quanten-physikalischen zu einer klassisch-physikalischen Beschreibung des Messens, also des Messvorgangs einschließlich des zu messenden Objekts, das nur näherungsweise aus der Welt im Ganzen und unter Absehung des Beobachters abstrahiert werden kann. Messen beeinflusst das zu Messende.

Messen kann sich heute – grob gesagt – auf drei Domänen beziehen, auf physikalische, auf ästhetische und auf moralische Phänomene. Hierbei zeigt eine historische Betrachtung, dass die drei Domänen sehr viel stärker miteinander verbunden sind, als wir es uns heute oftmals bewusst machen. Die Vorstellung, dass ein Maß uns genau darüber informiert, wie eine Sache beschaffen ist oder wie sie beschaffen sein sollte, zerfällt an dem, was Hans Ulrich Gumbrecht die Paradoxien des Begriffs in seiner Geschichte genannt hat: 1) „Die Geschichte des Begriffs Maß bietet zuviel und zugleich zuwenig Material." 2) Sie „weist zugleich zuviel und zuwenig Konturen in der pragmatischen Verteilung seiner Gebrauchsformen und Bedeutungen auf." 3) „Zahlreiche Phänomene,

die wir unter dem Begriff ‚Maß' subsumieren, sind in den Kunst- und Literaturwissenschaften zugleich außergewöhnlich prominent und außergewöhnlich vernachlässigt" (Gumbrecht 2003, S. 846 f.). Ohne diese Geschichte hier im Einzelnen nachzuzeichnen, ist als bedeutsam festzuhalten, dass die ästhetischen Maße, also was als schön, ausgewogen, harmonisch verstanden wurde, an den physischen Proportionen orientiert waren bzw. zum Teil bis heute noch daran orientiert sind. Ebenso lehnen sich ethische Maße am Ästhetischen an. Die *Kalokagathia* (das Gute und das Schöne, vgl. Bubner/Grosse 1976) wird seit Platon zum lernbaren Ideal, das bei Aristoteles zum Inbegriff aller Tugenden und ihrer (idealen) Vervollkommnung gerät. In der Folge finden sich einerseits Überblendungen des Ethischen und Ästhetischen (z. B. die ‚schöne Seele'), andererseits auch Trennungen und Ausblendungen (z. B. dort, wo die *Kalokagathia* nur noch die ästhetisch-poetologische Dimension betont und ethisch-politische Aspekte nicht mehr zu transportieren vermag).

3.2 Der Mensch als Maß der Dinge?

Wie im Folgenden gezeigt werden soll, erlaubt das Messen keineswegs einen direkten Zugriff auf eine ‚objektive' Welt. Vielmehr bezeichnet das Messen Operationen, die sehr verschiedene Register menschlichen Weltbezugs aufrufen und von Voraussetzungen bzw. Vorannahmen getragen sind, die, wie der Begriff ‚objektiv', ein Urteil des Menschen vor alle folgenden Operationen setzt – einem Vorzeichen vor der Klammer gleich.

Die Form des bedenkenden, prüfenden Messens ist offenbar mit einem Ziel verbunden: „Der Mensch ist das Maß aller Dinge, der seienden, daß sie sind, der nicht-seienden, daß sie nicht sind." (Diels/Kranz 80 B 1) Die Deutung dieses als Homo-Mensura-Satz berühmt gewordenen Protagoras-Fragments ist schwierig (Woodruff 2001), bedarf aber soweit der Erwähnung, als seine Interpretation auf zwei wichtige Punkte für unser Thema aufmerksam macht. Erstens wird hiermit ein Relativismus angesprochen, der mindestens Gültigkeit für die je subjektive Wahrnehmung der Menschen beanspruchen kann. Wenn dem einen der Wind kühl vorkommt, so kann der andere ihn gleichwohl als warm empfinden – die jeweiligen Maße müssen bzw. können nicht in Widerspruch zueinander stehen, sofern sie als subjektive Empfindungen verstanden werden. In diesem Sinne ist auch Blumenbergs Einwand zu verstehen: „Daß der Mensch, nach dem Wort des Protagoras, das Maß aller Dinge sei, bedeutet eben gerade nicht, der eine sei das Maß des anderen" (2006, S. 261). Damit ist aber zweitens der Protagoras'sche Relativismus auf einer grundlegenden Ebene noch nicht behoben. Wenn

ein extremer Relativismus die Tatsache von einander widersprechenden Ansichten aufheben würde, hätten wir nicht nur ein logisches, sondern auch ein ethisches Problem. Ein Messwert könnte dann nicht mehr intersubjektiv über das informieren, was alle gleichermaßen betrifft und von allen als Wirklichkeit anerkannt werden müsste. So sei aber Protagoras nicht zu verstehen. Schmitz übersetzt: „Aller Angelegenheiten Maß ist der Mensch, der stattfindenden, wie sie stattfinden, der nicht stattfindenden, wie sie nicht stattfinden“ (2007, S. 134), und sieht die Absicht des Protagoras in einer grundsätzlich epistemologischen Einstellung:

> Die originelle Leistung des Protagoras bei der Formulierung dieses Satzes besteht demgemäß in der Entdeckung der Nuance, ihres Gewichts, der Abhängigkeit dieses Gewichts vom Standpunkt des Beurteilenden und der Möglichkeit, durch geschickte Reden dieses Gewicht auch vom Standpunkt des Hörers aus zu verschieben. Wegen dieser Entdeckung dürfte Protagoras, der den Satz als Empfehlung seiner Redekunst benützt, vielleicht sogar formuliert haben. Was er auf diese Weise für die Philosophie als Besinnung des Menschen auf sein Sichfinden in einer Umgebung […] geleistet hat, ist die Einsicht in den Spielraum der Auslegung bei der Explikation einzelner Bedeutungen aus der binnendiffusen Bedeutsamkeit der Situationen. (Schmitz 2007, S. 134 f.)

Die Rede vom menschlichen Maß verweist nach Schmitz auf ihren rhetorischen Aspekt: Je nachdem, wie man den Akzent und wie man die verschiedenen Akzentsetzungen in einem kommunikativen Prozess um die „bürgerliche Tüchtigkeit“ (Prot 323 a) integriert, wird man feststellen, dass es sich bei den Angelegenheiten der Menschen um unscharfe, uneindeutige Tatbestände handelt, die gerade nicht geometrisch genau aufzulösen sind. Dagegen, so Schmitz, setzt Platon seine Dialektik, eine „der cartesischen *regulae ad directionem ingenii“* ähnliche Universalmethode (Schmitz 2007, S. 138), mit der genau diese Unschärfe als ‚sophistisch‘ verunglimpft und ihr ein analytisch-rekonstruktiver Exaktheitsanspruch entgegengesetzt wird. Für Schmitz ist bereits hier in der Geschichte der Philosophie und ihrer Reflexion der Menschheit auf sich selbst die falsche Abzweigung gewählt worden, die in einer Verkennung einer sogenannten Außenwelt und ihrer wissenschaftlich-technischen Bemächtigung die Krisen heraufgeführt hat, in denen wir uns heute befinden (Husserl 1938/1976).

Der Homo-Mensura-Satz markiert die unhintergehbare Richtung aller Messvorgänge: Sie dienen dem (messenden) Menschen, um die Welt um sich herum und in sich selbst für sich einzurichten. Welchen Maßstab wir Menschen auch immer ansetzen, er dient unseren Interessen und verdankt sich unserer Konstruktion (vgl. Fuchs 2007). Das menschliche Maßnehmen und Maßsetzen stößt allerdings dort auf Grenzen, wo dem Menschen die Maße selbst vorgegeben sind.

Das gilt etwa für Naturkonstanten, die menschliches (und anderes) Leben auf diesem Planeten überhaupt möglich machen. Das gilt für natürliche Maße und Rhythmen wie Mondphase, Tages- oder Jahresdauer. Das gilt in zunehmend disponibler Weise auch für Maße, die dem Menschen angemessen erscheinen, unter technischen Bedingungen aber modifiziert werden können und eben darin moralische Fragen aufwerfen: die Konservierung von lebensfähigem Material nahe dem absoluten Kältepunkt wie etwa Samen- oder Eizellen oder die Erweiterung menschlicher Sinneswahrnehmung durch technische Geräte.

Hier zeigt sich die doppelte, implikative Seite der Maße: ‚Natürliche' Maße geben dem Menschen vor, woran sich sein Leben in Denken, Urteilen und Handeln auszurichten hat. In dem Maße, aber in dem der Mensch Kompensationen und Substitutionen erfindet, um den Menschen (oder andere Lebewesen) an andere Maße anzupassen (Temperaturen oder Sauerstoffgehalt im All[11]) bzw. diese Maße selbst verändert (z. B. genetische Züchtung von Organen ohne Abstoßungsreaktion), verlieren diese Vorgaben ihre bloß natürlich-konditionale Funktion und müssen – da Handlungsergebnis – als Normen dann auf eine andere Weise gefunden und ‚hergestellt' werden. Dies erfordert entsprechende soziale Aushandlungsprozesse, die nicht zuletzt auch moralische Implikationen tragen.

Der Anthropozentrismusvorwurf ist deshalb noch einmal zu präzisieren: Die Orientierung des Menschen an Maßen ist unhintergehbar anthropozentrisch, ob er die Welt an seine Maße anpasst und dabei eine globale Klima- und Naturkatastrophe heraufbeschwört (Anthropozän) oder ob er dieser Tendenz entgegenzuarbeiten sucht und ein „Parlament der Dinge" (Latour 2001) einberuft, in dem alle anderen Akteure und Aktanten auch ihre Stimme erhalten sollen. Der Mensch bleibt so oder so ein Wesen, das Maß nimmt und Maß setzt und sich in diesen Akten mindestens reflexiv, zumeist verändernd zu dieser Welt verhält. Dieser Rolle entkommt er offenbar nicht – und das gilt auch für seinen Umgang mit selbst geschaffenen Entitäten wie künstlichen Intelligenzen. Auch wenn er (zunächst) das Maß hierfür in sich selbst sucht und diese fremde Intelligenz an der eigenen misst, wird diese sog. Künstliche Intelligenz zu einer Intervention in eine Welt, die die Dinge und ihre Maßstäbe verändert – nicht zuletzt den Menschen in seinem Selbst- und Weltverhältnis. Wie Menschen das wiederum bewerten, also durch einen messenden Vergleich einem Urteil zuführen, ist keinesfalls vorgegeben. Die folgenden Punkte sind hierbei zu berücksichtigen: Der Mensch ist das Maß der Angelegenheiten hinsichtlich der Bedingungen (dass und wie etwas als

[11] Als ein Gründungsdokument dieser Vorstellung der Änderung des Menschen zur Überlebensfähigkeit in sonst tödlichen Umgebungen wie dem All, anstelle der Bereitstellung erdähnlicher Umweltbedingungen in mobilen Vehikeln (wie Raumfähren) kann „Cyborgs and Space" gelten, das auch das Konzept des Cyborgs begründete: Clynes und Kline 1960.

zu betrachtende Größe in Erscheinung tritt) und der Option (welche Größe wie bewertet und gewählt oder verworfen wird). In beidem hat er sich als Naturwesen zu erkennen, das einerseits bestimmt und andererseits bestimmend ist (vgl. Böhme 2008, bes. S. 119 ff.). Als messendes Wesen hat er nie den Status eines ‚externen Beobachters', sondern ist bestimmender und bestimmter Teil der Natur, die er seinem Messen unterzieht. Die von ihm verwendeten Maße resultieren zum einen aus ‚vorgefundenen', ‚natürlichen' Maßen (wie Tag, Stunde, Mondphase), die durch entsprechende Experimente und Messmethoden weiter verfeinert werden können.[12] Die sogenannten natürlichen Maße werden z. B. in physikalischen Prozessen gefunden: z. B. in der Gravitation, für die Newton in seinen *Philosophiae Naturalis Principia Mathematica* (1687) eine mathematische Formel angibt. Diese Fundamentalkonstante der (klassischen) Physik wird mit ‚Objektivität' und ‚Notwendigkeit' assoziiert. Die Vorstellung von Gravitation als Kraft und ihr Maß werden jedoch in der allgemeinen Relativitätstheorie ganz anders gedacht – das Maß, das Messen von Größen, verändert sich, wie man an diesem Beispiel sehen kann, unter den Bedingungen der Theorieposition erheblich.

Neben den ‚natürlichen' Maßen setzt der Mensch auch seine eigenen – und geht hierbei durchaus eigensinnig vor. Die Größe eines DIN-A4-Papiers: 210 × 297 mm, resultiert aus der Vorgabe, dass ein A0-Bogen der Fläche von einem Quadratmeter entspricht. Halbiert man diesen Bogen über die längere Seite, so ergibt sich für diesen halben Bogen (A1) eine geometrisch ähnliche Figur usw. (Wikipedia: Papierformat). Die US-amerikanischen Längen- und Hohlmaße und Gewichte verdanken sich einer vornehmlich britischen Herkunft und weisen keinerlei Bezug zum Dezimalsystem auf. Das Nebeneinander dieser verschiedenen Maßsysteme mag umständlich erscheinen, hat aber noch nicht zu einer Vereinheitlichung geführt, auch wenn sonst menschliche Maßsysteme einer Vereindeutigung von Ausschnitten der Welt dienen sollen, um so intersubjektiv Geltung herzustellen. Das spricht dafür, dass Maße neben einer Orientierungsfunktion auch eine kulturell-symbolische Dimension aufweisen können.

Menschliche Maße beziehen sich auch auf je subjektive Wahrnehmungen oder Einschätzungen – auf diesen Punkt weist bereits der relativistische Homo-Mensura-Satz von Protagoras. Die Temperatur im ICE-Großraumabteil mag dem einen Menschen als zu kühl erscheinen, während ein anderer sie als zu warm empfindet. Hier wird man relativ feste Ober- und Untergrenzen erkennen können für das, was Menschen als biologischem Lebewesen (noch) zuträglich ist.

[12] „Die Sekunde ist das 9.192.631.770-fache der Periodendauer der dem Übergang zwischen den beiden Hyperfeinstrukturniveaus des Grundzustands von Atomen des Nuklids 133Cs entsprechenden Strahlung"; www.atomuhr-infos.de.

Zwischen diesen Grenzen sind subjektive Befindlichkeiten nur schwer als intersubjektiv verbindlich ‚auf einen Nenner' zu bringen. Die Durchschnittstemperatur als statistisches Mittel mag als Ausweg erscheinen, übersehen wird dabei jedoch leicht, dass die differenten Maße in diesem Fall gerade keine Orientierung geben, sondern eher Anlass für Streit sind, der mit dem Verweis „de gustibus non est disputandum" in den Bereich des Geschmacks und damit dem messenden Urteil gerade enthoben werden.[13] Beziehen sich solche Urteile auf Moralisches – und die Maßethik „ist in einem tiefen und umfassenden Sinne eine Ethik des guten Geschmacks" (Gadamer 1990, S. 45) –, wird schnell deutlich, dass Relativierung und Subjektivierung in der Sachfrage nicht weiterhelfen. Hier braucht es intersubjektive, diachron oder synchron verbindliche Maße im Sinne einer ‚Angemessenheit'.

Die Tätigkeit des Messens und die Rede vom Menschen als Maß der Dinge ließe sich mit Blick auf unsere Thematik so zusammenfassen: Das Messen, das Maßnehmen, ist als menschliche Tätigkeit immer von menschlichen Interessen – seien es handfeste praktische Anwendungen, sei es theoretische Neugier – getrieben. Maße werden gefunden bzw. gesetzt und werden mit Bewertungen wie ‚natürlich', ‚angenehm', ‚stimmig', ‚richtig' bzw. mit ihrem Gegenteil belegt. Messen ist immer ein vom menschlichen Beobachter und seinen Interessen und Bewertungen geleitetes Verfahren, das in seinen verschiedenen Registern als ein differentes und in bestimmten Bereichen auf ein unscharfes, uneindeutiges Phänomen gerichtetes gesehen werden sollte. Diese Unschärfe ist nach Schmitz bereits seit Platon in problematischer Weise vereindeutigt worden; damit ist aber zugleich die erstrebenswerte (vielleicht sogar ‚notwendige') Verständigung über ein Sichbefinden in einer gemeinsam kommunikativ zu erhebenden Situation ausgehebelt worden. Die je individuell erlebte Situation wird als ‚bloß subjektiv' diskreditiert und durch eine *more geometrico* gemessene, ‚objektive', Beschreibung ersetzt, die zum ‚Maß der Dinge' erhoben wird.

Diese als wissenschaftlich objektiv verstandene Üblichkeit wird gegenwärtig durch ein anderes Messverfahren erkennbar weiter verschärft. Wo Künstliche Intelligenz die bereits bestehenden (Vor-)Urteile der Menschen als *algorithmic bias* in Mess- und Bewertungsverfahren verfestigt und verstärkt, werden unter

[13] Dies bedeutet nicht, dass Geschmacksurteile beliebig subjektiv wären, sondern dass ein Maß zwischen Mehrheit oder Mode und eigenem Urteil gefunden wird: „Im Begriff des Geschmacks liegt daher, daß man auch in der Mode Maß hält, die wechselnden Forderungen der Mode nicht blindlings befolgt, sondern das eigene Urteil dabei betätigt. Man hält seinen ‚Stil' fest, d. h. man bezieht die Forderungen der Mode auf ein Ganzes, das der eigene Geschmack im Auge behält und nimmt nur das an, was zu diesem Ganzen paßt und wie es zusammenpaßt" (Gadamer 1990, S. 43).

dem Label einer ‚Intelligenz' die Vorgänge des Messens und Bewertens noch weiter von der Weltwahrnehmung des Menschen abstrahiert und als von ihm nicht mehr durchschaubare Objektivität (aufgrund des ‚Black-Box-Charakters' der KI) dargestellt. Eine Objektivität, die aufgrund des Vernetzungsgrades von KI unter Umständen eine enorme Verbreitung und In-Geltung-Setzung erfährt und aufgrund der Rekursivität des Verfahrens sich selbst bestätigt und verfestigt. Die von dem System errechneten Ergebnisse oder Prognosen beruhen nicht auf einer technisch objektiven, interesselosen Intelligenz, sondern auf den von Menschen qua Annotation bzw. Daten bereits gesetzten Gewichtungen. Die hier in Gang gesetzten Optimierungsberechnungen verweisen stets auf die von Menschen definierten Optima. Insofern ist zu fragen, in welcher Hinsicht eine Künstliche Intelligenz in das Zusammenwirken mit menschlicher Intelligenz etwas genuin Eigenes einbringt, das in aller Differenz doch so viel Ähnliches beinhaltet, dass von einem ‚Zusammen' gesprochen werden könnte. Mit Blick auf das Ergebnis des Zusammenwirkens könnte dann von einer „hergestellten engen Gemeinschaft, der Mithilfe und Unterstützung" die Rede sein: einer Gemeinschaft, in der kein asymmetrisches Gefälle bestünde, das am Ende eine Instrumentalisierung der einen Seite durch die andere bedeutete, sondern ein Miteinander zwar verschiedener, aber mindestähnlicher und gleichwertiger Wesen.

4 Ausblick

„Wie wird das Zusammenleben und -wirken von Menschen, Tieren und Pflanzen einerseits und Maschinen andererseits zukünftig aussehen?", so lautet die Ausgangsfrage. Konzentrierter gefragt: Wie wirken NI und KI zusammen, wie können sie zusammenwirken? Diese Fragen erfordern eine Reflexion, wer und was genau in dem Zusammengestellten vereint gemeint und gedacht werden soll: welche Lebewesen mit welchen Eigenschaften (in welcher Hinsicht sind sie intelligent?); welche Maschinen mit welchen Eigenschaften? Wo ist der Koppelungspunkt ihres Zusammenwirkens? Befinden KI und NI sich in einer gemeinsamen raumzeitlichen Situation oder interagieren sie räumlich bzw. zeitlich versetzt? Ist im letztgenannten Fall noch von einem Zusammenwirken zu sprechen – weil und sofern die räumliche bzw. zeitliche Distanz die körperliche Bedingtheit von Lebewesen überschreitet und der physische Koppelungspunkt unklar ist?[14] Gibt es ein *tosamne*-Gefälle, sind also jene, die ‚zusammenwirken' nicht gleich(berechtigt),

[14] Im Weiteren ist hier freilich zu fragen, ob nicht auch Menschen räumlich und zeitlich versetzt zusammenwirken können. Wie aber ist das zwischen Menschen und Tieren? Um das Zusammenwirken von menschlicher und künstlicher Intelligenz zu charakterisieren, bedarf

wie es sich z. B. in der soziomorphen Tech-Metapher des Master–Slave ausdrückt? Oder begegnen sich die ‚Akteure/Agenten' als freie Gleiche? Wie ist die Rede von selbst (!) lernenden KI-Systemen zu verstehen, die semantisch ein Zusammen-Lernen – zumindest ab einem gewissen Zeitpunk im Lernprozess – gerade auszuschließen scheint; welche Implikationen hat das für eine ‚Synenz'? Wie geht das von der KI diskret Gelernte in das Zusammenwirken ein, wie wird ein gemeinsamer Lernschritt daraus? Kehrt sich hier das Gefälle unter Umständen um und was bedeutet das für Status, Rechte oder Ansprüche der KI im Zusammen*leben*?

Welche *Ähnlichkeit* besteht unter den Beteiligten genau bzgl. Substanz, Wesen, Fähigkeiten, Eigenschaften, Rechte/Pflichten, Erwartungen, Erwartungserwartungen etc.? Wie wörtlich oder metaphorisch übertragen fallen die Antworten auf obige Fragen aus? Zu reflektieren ist das Zusammenwirken mindestens in Hinsicht auf die folgenden Elemente: die Agenten, ihr Koppelungspunkt, ihre Relation, die Effekte, die Qualität und der Modus Operandi ihrer Kooperation.

Darüber hinaus ist die Frage des Zusammenwirkens von NI und KI auch in ihren Konsequenzen für Gesellschaft, Wissenschaft und Politik, Lebensführung und Lebensformen relevant: Wenn ein Zusammenwirken von NI und KI nicht nur sprachlich formuliert und theoretisch angenommen, sondern auch gesellschaftlich verbreitet wird, dann bedeutet dies, dass die impliziten oder strategisch verschleierten ästhetischen und moralischen Urteile der KI-Konstrukteure, das sind heute globale Tech-Giganten, dabei übernommen werden. Wer Systeme, die mit solchen Gleichheits-, Üblichkeits-, Sollens- und Gelingensvorstellungen entwickelt wurden, in allen Bereichen der Lebenswelt verankert, der schafft so die Rahmenbedingungen für damit kompatible Lebensformen, während deviante Entwürfe damit kollidieren oder ausgeschlossen werden. Die Durchdringung unserer Handlungsumstände mit KI, das gewollte und gesollte Zusammenwirken von NI und KI, erzeugt Handlungs- und Entscheidungsarchitekturen mit je entsprechenden Gelingensbedingungen, die den Wertabwägungen und Präferenzhierarchisierungen der KI-Konstrukteure inkl. Datenannoteure etc. entspringen und sich bisher weder kulturell-sozial bewährt haben noch politisch-demokratisch legitimiert worden sind (z. B. Kaltheuner 2021; Zuboff 2018).

Das Zusammenwirken von Ungleichen (ohne ausgewiesenen Vergleichs- und Koppelungspunkt) ist nicht möglich. Der Verweis auf das faktische Zusammenwirken von Menschen und KI-Systemen ermangelt unseres Erachtens einer klaren theoretischen Beschreibung wer und was hier tatsächlich gleich(berechtigt)

es offenbar noch weiterer Merkmale, um bei bestehender Ungleichheit in bestimmter Hinsicht eine Gleichheit und ein Zusammen zu konstatieren.

zusammenwirkt, und wo der Vorgang über ein Instrumentalisierungsverhältnis rechnender Systeme durch Menschen hinausgeht. Davon unbenommen bleibt die Aussage, dass KI-Systeme in ihrer Leistung und ihrem Einsatz mehr sind als Instrumente: Werkzeuge, Maschinen, Automaten. Unklar ist jedoch, *als was* eine KI über ihre Funktion hinaus in dem Zusammenwirken mit anderen Lebewesen anzuerkennen wäre. Bei aller Ungleichheit mit Lebewesen in anderer Hinsicht (gemacht und nicht geworden, anorganisch, nicht fühlend, nicht intentional…) bedarf es doch einer Gleichheit in einer Hinsicht, die die Rede vom Zusammenwirken zutreffend und sinnvoll macht. Diese Gleichheit (Kompatibilität) existiert nicht ‚von Natur aus', sondern muss von Menschen hergestellt werden. Es sind Angleichungen, die auch auf menschlicher Seite vorgenommen werden, um Funktionalität an der Schnittstelle sicherzustellen. Ein Problem hierbei ist, dass MI – wenn auch nicht klar definiert in keiner der sie untersuchenden Wissenschaften – sicher sehr viel mehr ist als ihre KI-kompatiblen Anteile. Dieses Surplus wird systematisch verdrängt oder verdeckt, wenn vorschnell von einem Zusammenwirken von MI und KI die Rede ist. Einer solchen Verarmung der natürlichen oder Technomorphisierung der menschlichen Intelligenz sollte bewusst entgegengewirkt werden. Denn – was auch immer im Einzelnen dazugehört – die Einzelfähigkeiten des Bündels ‚Intelligenz' prägen sich wie jede Fähigkeit aus unter Bedingungen der Übung, Wiederholung und des kontinuierlichen Gebrauchs und verkümmern unter Bedingung der Vernachlässigung und dauerhaften Delegation. Unter Umständen wäre es klüger, auf ein tendenziell technomorphisierendes Zusammenwirken von MI und KI zu verzichten, beide klar getrennt und unterschieden zu lassen und jede Intelligenz ohne Angleichungsdruck in ihren Stärken und ihrer Spezifik zu fokussieren. Eine (nicht nur technische) Herausforderung wäre dann die offene Frage, wie trotz dieser Trennung von der Wirkung der KI im Handeln mit MI profitiert werden kann.

Schließlich zeigt die Auseinandersetzung mit NI, MI und KI, dass in Zeiten omnipräsenter KI-Verheißungen es sich lohnt, den Blick auch auf nichtmenschliche natürliche Intelligenzen (wie Hunde oder Oktopoden) zu lenken, deren Synergie-Potenziale wegen der größeren Gleichheit qua Lebendigkeit, Sensitivität, Leiblichkeit etc. ebenso aufschlussreich und vielversprechend erscheinen. Obwohl wir hier teilweise tausende Jahre faktischer Kooperation haben, ist diese häufig noch zu wenig verstanden und das ‚Synenz'-Potenzial nicht ausgelotet. Tiere kooperieren mit Menschen in hochspezialisierten Aufgaben, seien es Jagdhunde oder -falken, Ratten zur Landminendetektion, Hunde zur Drogen-/Sprengstoffdetektion, Blindenführung, Katastrophenrettung etc. Selbst mit dem völlig andersartigen Oktopus hat der Mensch mehr gemeinsam als mit KI Systemen, nicht zuletzt die für das Leben so zentrale Tatsache eigener, genuiner

Präferenzen.[15] Die Erwartung, dass wir Menschen im Umgang mit einer selbst hergestellten KI hier schneller verstehen und effektiver kooperieren könnten, könnte sich – bei allen zweifelsfrei spektakulären Leistungen von KI – als lebensweltliches Missverständnis herausstellen. Will sagen: Die Erwartung einer intelligenzbezogenen Synergie (‚Synenz') zwischen Menschen (die ‚Synenz' zwischen NI und KI ist immer eine von Menschen (MI) inszenierte) und technischen Systemen könnte in ihrer *lebensfremden* Art dem Zusammen*leben* der verschiedenen Wesen gar nicht zuträglich sein. Freilich wird mit dem Thema des Zusammenlebens noch einmal eine weitere Ebene angesprochen. Das Zusammenwirken von Lebewesen (hier Menschen und MI) mit Nicht-Lebewesen (hier KI) mag funktionieren und könnte sich zugleich negativ auf Lebensform und Lebensbedingungen der Lebewesen auswirken. Die vom Menschen für die Synergie aufzubringende Angleichungsleistung könnte ihm selbst und seinen Lebens-Bedingungen *(conditio humana)* entgegenstehen. Eine Angleichung der KI müsste für die angestrebte Synenz womöglich den Koppelungspunkt des Lebens in den Blick nehmen, um lebensdienlich sein zu können – was dann doch ein ganz anderes Projekt wäre.

Literatur

Bellon, Jacqueline; Eyssel, Friederike; Gransche, Bruno; Nähr-Wagener, Sebastian; Wullenkord, Ricarda (2022): Theorie und Praxis soziosensitiver und sozioaktiver Systeme. Wiesbaden: Springer Fachmedien Wiesbaden.

Blumenberg, Hans (1996): Die Legitimität der Neuzeit. Erneuerte Ausgabe Frankfurt am Main: Suhrkamp.

Blumenberg, Hans (2010): Theorie der Lebenswelt, hrsg. von M. Sommer. Berlin: Suhrkamp.

Böhme, Gernot (2008): Ethik leiblicher Existenz. Frankfurt am Main: Suhrkamp.

Braun, Bernard (2019): *Geschichte der Kunstphilosophie und Ästhetik*. 4 Bde. Darmstadt: Wissenschaftliche Buchgesellschaft.

Bubner, Rüdiger, und Wilhelm Grosse. „Kalokagathia" (1976). In *Historisches Wörterbuch der Philosophie*, 4:681–84. Darmstadt: Wissenschaftliche Buchgesellschaft.

Clynes, Manfred E.; Kline, Nathan S. (1960): Cyborgs and Space. Altering man's bodily functions to meet the requirements of extraterrestrial environments. In: Astronautics September 1960, S. 29–33.

Crawford, Kate (2021): "AI is neither artificial nor intelligent". In: The Guardian, 06.06.2021. Online verfügbar unter https://www.theguardian.com/technology/2021/

[15] Das Phänomen der *Symbiose* als biologische und soziale Tatsache wäre ein weiteres und eigenes Kapitel in der genaueren Untersuchung möglicher ‚Synenzen'. Der Hinweis muss an dieser Stelle genügen.

jun/06/microsofts-kate-crawford-ai-is-neither-artificial-nor-intelligent, zuletzt geprüft am 26.04.2023.

Derbolav, J. (2010) [1974]. „Handeln, Handlung, Tat, Tätigkeit“. In: J. Ritter, K. Gründer & G. Gabriel (Hrsg.), *Historisches Wörterbuch der Philosophie,* 3:992–94. Darmstadt: Wissenschaftliche Buchgesellschaft.

Detienne, Marcel et Jean Pierre Vernant (1973): Les ruses de l'intelligence – La mètis des Grecs. Paris: Flammarion

Diels, Hermann und Walter Kranz (1956): Die Fragmente der Vorsokratiker. 3 Bde., Berlin: Weidmann.

Drieschner, Michael und Arnim Wegner. „Meßprozeß“ (1980). In: J. Ritter, K. Gründer & G. Gabriel (Hrsg.): *Historisches Wörterbuch der Philosophie*, 5:1166–68. Darmstadt: Wissenschaftliche Buchgesellschaft.

Fuchs, Peter. Das Maß aller Dinge. Eine Abhandlung zur Metaphysik des Menschen (2007). Weilerswist: Velbrück Wissenschaft.

Gadamer, Hans-Georg (1990): Wahrheit und Methode. Grundzüge einer philosophischen Hermeneutik. Tübingen: Mohr Siebeck.

Gardner, Howard (2005): *Abschied vom IQ. Die Rahmentheorie der vielfachen Intelligenzen.* 4. Aufl. Stuttgart: Klett-Cotta.

Gransche, Bruno (2024). „Technische Autonomie". In M. Gutmann, B. Rathgeber & K. Wiegerling (Hrsg.), *Handbuch Technikphilosophie.* Stuttgart: J. B. Metzler, S. 257–266.

Gransche, Bruno und Manzeschke, Arne (2024). Das bewegliche Heer der Künstlichen Intelligenz. Ein Technomythos als Summe menschlicher Relationen. In M. Heinlein & N. Huchler (Hrsg.), *Künstliche Intelligenz, Mensch und Gesellschaft*. Wiesbaden: Springer VS.

Grundmann, Walter. „σύν, μετά κτλ.“ In *Theologisches Wörterbuch zum Neuen Testament*, herausgegeben von Gerhard Friedrich, 7:766–98. Stuttgart: W. Kohlhammer, 1964.

Gumbrecht, Hans Ulrich. „Maß“. In *Ästhetische Grundbegriffe*, herausgegeben von Karlheinz Barck, Martin Fontius, Dieter Schlenstedt, Burkhart Steinwachs, und Friedrich Wolfzettel, 5:846–66. Stuttgart /Weimar: J. B. Metzler, 2003.

HLEG AI (High-Level Expert Group on AI). Ethics Guidelines for Trustworthy AI, Brüssel, 2019.

Heinlein, Michael und Huchler, Norbert (Hrsg.). (2024). *Künstliche Intelligenz, Mensch und Gesellschaft*. Wiesbaden: Springer VS.

Hubig, Christoph (2011): „‚Natur‘ und ‚Kultur‘. Von Inbegriffen zu Reflexionsbegriffen“. In: ZKphil. 5 (1), S. 97–119.

Husserl, Edmund (1938/1976): Die Krisis der europäischen Wissenschaften und die transzendentale Phänomenologie. Eine Einleitung in die phänomenologische Philosophie. 2. Aufl. Husserliana 4. Den Haag: Martinus Nijhoff.

Ifrah, Georges (1993): Universalgeschichte der Zahlen. Frankfurt am Main: Campus.

Kaltheuner, Frederike (Ed.) (2021): Fake AI. Meatspace Press.

Kant, Immanuel (2005): Werke in sechs Bänden, hrsg. von Wilhelm Weischedel. 6. unveränd. Aufl. Darmstadt: Wissenschaftliche Buchgesellschaft.

Latour, Bruno (2001): *Das Parlament der Dinge. Für eine politische Ökologie.* Frankfurt am Main: Suhrkamp.

Mainzer, Klaus (1984): „Messung". In: *Enzyklopädie Philosophie und Wissenschaftstheorie*, herausgegeben von Jürgen Mittelstraß, 2:862–64. Mannheim, Wien, Zürich: Bibliographisches Institut.
Plessner, Helmuth (1975): Die Stufen des Organischen und der Mensch. Einleitung in die philosophische Anthropologie. 3. Aufl. Berlin: Walter De Gruyter.
Ricœur, Paul (2003): *The rule of metaphor. The creation of meaning in language* (Routledge classics). London: Routledge.
Ricœur, Paul und Jüngel, Eberhard (Hrsg.). (1974): *Metapher. Zur Hermeneutik religiöser Sprache* (Evangelische Theologie Sonderheft, Bd. 1974). München: Kaiser.
Schmitz, Hermann (2007). Der Weg der europäischen Philosophie. Eine Gewissenserforschung. Bd. 1: Antike Philosophie. Freiburg/München: Karl Alber.
Seising, Rudolf (2021): *Es denkt nicht! Die vergessene Geschichte der KI.* Frankfurt am Main: Büchergilde Gutenberg.
Simmel, Georg (1930): Philosophie des Geldes. Berlin: Duncker & Humblot.
Sturma, Dieter (2004): „Ersetzbarkeit des Menschen? Robotik und menschliche Lebensform". In Jahrbuch für Wissenschaft und Ethik 9, S. 141–162.
Weber, Max. Wissenschaft als Beruf. 10. Aufl. Berlin: Duncker & Humblot, 1996.
Wittmann, Matthias und Michèle Ganser (2023): Oktopia. Frankfurt am Main: Büchergilde Gutenberg.
Woodruff, Paul (2001): „Rhetorik und Relativismus: Protagoras und Gorgias". In: Handbuch Frühe Griechische Philosophie. Von Thales bis zu den Sophisten, hrsg. von A. A. Long. Stuttgart/Weimar: J. B. Metzler, S. 264–284..
Zuboff, Shoshana (2018): Das Zeitalter des Überwachungskapitalismus, Frankfurt/New York: Campus.

Prof. Dr. Arne Manzeschke Institut für Pflegeforschung, Gerontologie und Ethik (IPGE), Evangelische Hochschule Nürnberg.

Programmierer im ersten Beruf; studierte Theologie und Philosophie; Promotion und Habilitation in Erlangen. Er lehrt Ethik und Anthropologie an der Evangelischen Hochschule Nürnberg und leitet dort das IPGE. Seit 2010 forscht er zu Mensch-Technik-Verhältnissen. Aktuell ist er Sprecher des BMBF-geförderten Forschungsclusters „Integrierte Forschung", das sich mit methodischen und inhaltlichen Fragen einer inter- und transdisziplinären Forschung im Bereich der Mensch-Technik-Interaktion befasst. Er ist Sprecher des Fachausschusses „Medizintechnik und Gesellschaft" bei der Deutschen Gesellschaft für Biomedizinische Technik (DGBMT) und Vorsitzender der Ethikkommission für Pflege- und Sozialforschung an der Evangelischen Hochschule Nürnberg.

PD Dr. Bruno Gransche Institut für Technikzukünfte ITZ am Karlsruher Institut für Technologie KIT.

Der Philosoph und Zukunftsforscher forscht und lehrt in den Bereichen Technikphilosophie/Ethik und Zukunftsdenken mit Fokus u. a. auf Philosophie neuer Mensch-Technik-Relationen, gesellschaftliche & ethische Aspekte von KI & Digitalisierung, Technikbilder/Menschenbilder/Metaphernanalyse sowie Vorausschauendes Denken. Gransche ist Privatdozent am Institut für Technikzukünfte der Universität Karlsruhe seit 2020; Studium der Philosophie und Literaturwissenschaft sowie Promotion in Heidelberg, Habilitation in Karlsruhe.

Er ist u. a. Mitherausgeber der Reihe *Techno:Phil – Aktuelle Herausforderungen der Technikphilosophie* sowie Fellow am Fraunhofer-Institut für System- und Innovationsforschung ISI in Karlsruhe, wo er bis 2016 in der Abteilung Foresight arbeitete.

Zur rechtlichen Verantwortlichkeit in der Mensch-Maschine-Interaktion am Beispiel Autonomer Waffensysteme

Susanne Beck und Simone Tiedau

Zusammenfassung

Der zunehmende Einsatz Lernender Systeme in den verschiedensten Lebensbereichen stellt nicht nur unsere Gesellschaft, sondern auch unsere Rechtsordnung vor neue Herausforderungen. So müssen wir uns u. a. die Frage stellen, wer sich für eine Entscheidung, die aus der Interaktion zwischen Mensch und Maschine hervorgegangen ist, rechtlich verantworten soll. Diese Frage wird dann besonders dringend, wenn die kooperativ getroffene Entscheidung potentiell besonders sensible Rechtsgüter betrifft (wie z. B. die Gesundheit oder das Leben anderer Menschen). Der folgende Beitrag beleuchtet die Interaktion zwischen Mensch und System gerade im Hinblick auf die (straf-)rechtliche Verantwortlichkeit des Menschen als Letztentscheider:in. Exemplarisch soll anhand einer möglichen Interaktion zwischen Mensch und Autonomen Waffensystem aufgezeigt werden, welche Probleme sich bei der Zurechnung einer kooperativ getroffenen Entscheidung und damit für die (strafrechtliche) Verantwortlichkeit des/der Letzteintscheider:in ergeben können, um anschließend das Konzept der „Meaningful Human Control" als Lösungsansatz vorzustellen.

S. Beck (✉) · S. Tiedau
Kriminalwissenschaftliches Institut der Leibniz Universität, Hannover, Deutschland
E-Mail: susanne.beck@jura.uni-hannover.de

S. Tiedau
E-Mail: susanne.beck@jura.uni-hannover.de

O. Richter et al. (Hrsg.), *Zusammenwirken von natürlicher und künstlicher Intelligenz: Beurteilen-Messen-Bewerten,*
https://doi.org/10.1007/978-3-658-45845-4_6

Schlüsselwörter

Lernende Systeme • Autonome Waffensysteme • Human in the Loop • Zurechnung • Strafrechtliche Verantwortlichkeit • Verantwortungslücke • (Humanitäres) Völkerrecht • Meaningful Human Control

1 Einführung

Schon heute wird unser Alltag durch die zunehmende Entwicklung im Bereich der Künstlichen Intelligenz (KI) und Automatisierung (bis hin zur Autonomisierung) geprägt. Begriffe wie „Industrie 4.0“, „Autonome Kraftfahrzeuge“, „KI-gestützte Diagnosesysteme“ oder „Autonome Waffensysteme“ gehören zum gängigen Wortschatz des 21. Jahrhunderts.

Dabei geht mit der Automatisierung einher, dass bestimmte Entscheidungen (zumindest teilweise) auf Systeme übertragen werden und die Bediener:innen bzw. die dahinterstehenden Menschen direkt mit ihnen interagieren. Gerade in einer fortwährend digitalisierten Welt wird es früher oder später Bereiche geben, in der die stetig zunehmende Informationsflut nur noch von Systemen bewältigt werden kann, oder Systeme zumindest mit einer geringeren Fehlerquote agieren, als der Mensch es könnte (Fraunhofer Gesellschaft 2018, S. 6; Beck 2015, S. 10 f.).

In vielerlei Hinsicht haben diese Entwicklungen einen direkten oder zumindest indirekten Einfluss auf das Recht. Denn dass es in der Interaktion zwischen Menschen zur Schädigung fremder Rechtsgüter kommen kann, ist Teil unseres gesellschaftlichen Miteinanders und zentraler Ausgangspunkt unserer normativen Ordnung. An dieser potenziellen Verletzbarkeit ändert sich nichts, wenn Menschen nicht mehr nur noch untereinander interagieren, sondern Systeme miteinbeziehen – oder die Interaktion gar gänzlich auf sie übertragen. Denn Systeme agieren niemals nur in einem isolierten, digitalen Raum; vielmehr berührt ihr Wirkungsradius immer auch unmittelbar oder mittelbar die physische Welt und damit Rechtsgüter Dritter (vgl. Beck 2020b, S. 5 f.). Ausgehend von einem Rechtssystem, welches – historisch bedingt – für die Interaktion zwischen Menschen geschaffen wurde, müssen wir wegen dieser Einbeziehung aber nun nicht nur die bestehenden *zivilrechtlichen Haftungsregime* überdenken, sondern uns auch damit auseinandersetzen, welchen Einfluss der Einsatz künstlich intelligenter Systeme auf die Frage *strafrechtlicher Verantwortlichkeit* hat. Denn gerade einer strafrechtlichen Sanktion liegt grundsätzlich die Konzeption individueller Verantwortlichkeit für eigene Handlungen zugrunde (Roxin und Greco 2020, § 7 Rn. 4 f.). Dies werden wir nicht zuletzt mit Blick auf Autonome Waffensysteme tun,

die nicht nur besonders umstritten sind, sondern sich auch als Beispiel besonders gut eignen – geht es bei den von ihnen (mit)getroffenen Entscheidungen doch um Leben oder Tod von Menschen. Mit Blick auf diese Systeme werden wir dann das Konzept *Meaningful Human Control* diskutieren.

2 Lernende Systeme – Status Quo

Ein Lernendes System ist – stark vereinfachend gesprochen – ein System, das in der Lage ist, selbstständig aus Erfahrungen und Daten zu lernen, um Entscheidungen oder Vorhersagen treffen zu können (sog. „Machine Learning System"). Die Besonderheit solcher Lernender Systeme liegt in ihrer Fähigkeit, Strukturen zu verstehen und sich durch das gewonnene Verständnis selbst weiterentwickeln zu können (Apel und Kaulartz 2020, S. 25 f.; Zech 2019, S. 202). Dabei ist für den Menschen nicht immer nachvollziehbar, welche Vorgänge das System innerhalb des Lernprozesses durchläuft (bspw. bei „Deep Learning"). So ist ex-ante regelmäßig schwer prognostizierbar, welche Vorgänge das System durchlaufen wird und ex-post häufig nicht mehr nachvollziehbar, weshalb das System zu bestimmten Ergebnissen gekommen ist (Martini 2017, S. 1018).

Das ist gerade dann problematisch, wenn Mensch und System dergestalt interagieren, dass sie kooperativ – also „gemeinsam" – eine Entscheidung treffen. Diese Kooperation kann so gestaltet sein, dass das System bestimmte Daten aufarbeitet (indem es sie filtert und sortiert), auf deren Grundlage der Mensch dann eine Entscheidung trifft, oder so, dass das System schon konkrete Entscheidungsmodalitäten vorschlägt, aus denen der Mensch dann eine auswählt (vgl. Dettling 2019, S. 636).

Auch wenn es letztlich der menschliche Akteur ist, der in diesen Situationen die finale Entscheidung trifft, so ist diese doch maßgeblich von den Vorgängen im System abhängig. Anschließend an das oben Gesagte: Entweder kennt der Mensch nicht alle Gegebenheiten und Daten, von denen er seine Entscheidung abhängig machen könnte, weil das System sie vorher aufgearbeitet hat. Oder ihm wird die Entscheidungsfindung schon insofern abgenommen, als er nur noch zwischen vom System ausgefertigten Modalitäten wählt, die auf diesen aufgearbeiteten Daten beruhen (vgl. Crootof 2016, S. 56). Selbst wenn für den Menschen im letzteren Fall einsehbar wäre, aufgrund welcher Daten das System die Modalitäten entwickelt hat, so kann sich allein durch den Vorschlag des Systems eine Voreingenommenheit entwickeln. Gerade Situationen, in denen der menschliche Akteur ein Bewusstsein dafür hat, dass das System mehr und/oder komplexere

Informationen erfasst haben könnte als er selbst (es könnte), kann beim Menschen eine Unsicherheit dahingehend hervorgerufen werden, sich entgegen der Vorschläge des Systems zu verhalten oder andere Entscheidungsmöglichkeiten in Betracht zu ziehen (Beck 2020b, S. 2). Mangels eigener gleichwertiger kognitiver Fertigkeiten wird der Mensch gezwungen zu wählen, ob er dem System „blind" vertrauen oder eine eigene Entscheidung treffen möchte, der dann aber jedenfalls eine schlechtere Bewertung der Daten zu Grunde liegt (die dadurch jedoch nicht per se falsch sein muss, bzw. in Einzelfällen sogar besser als die des Systems sein kann) (vgl. Schwarz 2021, S. 56 f.). Diese Beeinflussung wird dann besonders schwerwiegend, wenn Mensch und System in einem Umfeld interagieren, in dem eine Reflexion der vorgeschlagenen Entscheidungsmodalitäten aufgrund struktureller oder zeitlicher Gegebenheiten nicht mehr auf eine verantwortungsvolle Art und Weise möglich ist (vgl. Wagner 2019, S. 112).[1]

Es ist also nicht nur die gänzliche Übergabe einer Entscheidung auf ein System (also die vollständige Automatisierung/Autonomisierung), die sich von den uns bekannten, traditionellen Entscheidungssituationen unterscheidet, sondern auch die kooperative Entscheidungsfindung zwischen Mensch und System. Diese neue Gemengelage in der Entscheidungsfindung macht es notwendig, die rechtliche Situation der Personen, die in den Entscheidungsprozess involviert sind, einer neuen Bewertung zu unterziehen. Hierbei muss grundsätzlich zwischen den verschiedenen Personen, die auf eine entscheidende Art in diesen Prozess involviert sind, unterschieden werden. Zu denken wäre bspw. an die Hersteller:innen des Systems, die Programmierer:innen, die Nutzer:innen oder solche Personen, die für die Wartung des Systems verantwortlich sind (Denga 2018, S. 71).

Für die strafrechtliche Perspektive besonders relevant ist hierbei die Person, die letztlich kooperativ mit dem System entscheidet, also der/die Letztentscheider:in. Sie ist es auch, die regelmäßig gemeint ist, wenn es um das Erfordernis eines „Menschen in der Entscheidungsschleife" bzw. einem „Human in the loop" geht (Beck 2020c, § 7 Rn. 18).

[1] So ist bspw. die Technische Aufsicht über ein Kraftfahrzeug mit autonomer Fahrfunktion nach § 1f Abs. 2 StVG dazu verpflichtet, auf Grundlage der vom System bereitgestellten Daten ein alternatives Fahrmanöver freizuschalten, sobald das System sie darauf hinweist.

3 Die Auswirkungen des Einsatzes Lernender Systeme auf die strafrechtliche Verantwortlichkeit

Ausgehend vom bisher Gesagten, bedarf es folglich einer genaueren Betrachtung der strafrechtlichen Verantwortlichkeit beim Einsatz Lernender Systeme; und zwar nicht nur solcher, die automatisiert/autonom (also „selbst“) entscheiden, sondern auch solcher, die kooperativ mit dem Menschen (also „gemeinsam“) entscheiden.

3.1 Problem: Die Kumulation von Einzelbeiträgen und die Nachvollziehbarkeit

Da grundsätzlich unterschiedlichste Akteure einen Beitrag zur Entscheidung des Systems leisten (können), ist in den seltensten Fällen offenkundig, welcher Akteur strafrechtlich zur Verantwortung gezogen werden kann.[2] Denn regelmäßig wird der strafrechtlich relevante Erfolg erst durch die Interaktion zwischen den Beteiligten und dementsprechend durch eine Kumulation von Einzelbeiträgen hervorgerufen. Als „antwortende Entität“ kommen dann – je nach Einzelfall und unter der Annahme, dass das System selbst nicht strafrechtlich verantwortlich gemacht werden kann[3] – bspw. die Programmierer:innen, die Hersteller:innen, die Nutzer:innen oder auch das Wartungspersonal in Betracht, wobei regelmäßig ex-post nicht nachvollziehbar sein wird, welcher Beitrag die Entscheidung des Systems überhaupt beeinflusst hat, weil schon der Entscheidungsvorgang selbst nicht nachvollziehbar ist (Beck 2023, S. 31).

3.2 Problem: Die menschliche Handlung als Anknüpfungspunkt

Wenn wir – eine Nachvollziehbarkeit vorausgesetzt – davon ausgehen, dass wir für eine Strafbarkeit an eine menschliche Handlung anknüpfen müssen (Roxin und Greco 2020, § 7 Rn. 5), stellt sich zudem die Frage, wann der Beitrag

[2] Anders als das Zivilrecht kennt das Strafrecht keine Konstruktionen geteilter Verantwortlichkeit oder reiner Gefährdungshaftungen, die wiederum über Versicherungslösungen abgedeckt werden können.

[3] Vgl. zu dieser Debatte Gaede (2019, S. 57–69).

bzw. die Interaktion des Menschen (noch) als eine solche Handlung angesehen werden kann. Zumindest fehlt es an ihr, wenn das menschliche Verhalten nicht mehr von einem Willen gesteuert wird oder steuerbar ist (Beck 2020a, S. 45). Würde also bspw. ein Waffensystem Zielerfassung und Angriffsentscheidung gänzlich autonom treffen, so läge in eben jenen (Teil-)Handlungen keine *menschliche* Handlung mehr. Dasselbe würde dann gelten, wenn das System eine menschliche Entscheidung übersteuern würde (vgl. Schwarz 2021, S. 59; Umbrello 2021, S. 461). Natürlich wäre dann weiterhin an eine Strafbarkeit aus Unterlassen zu denken; dies jedoch aber gerade nur, wenn eine entsprechende Pflicht zur Handlung durch den Menschen bestanden hätte (Kühl 2017, § 18 Rn. 41).

Da in näherer Zukunft allerdings in den meisten Lebensbereichen wohl noch keine völlig autonomen Entscheidungen durch Lernende Systeme getroffen werden (können), werden wir regelmäßig menschliche Handlungen identifizieren können, an denen eine strafrechtliche Anknüpfung grundsätzlich möglich wäre. Das können aktive Handlungen sein, wie die Programmierung, das Trainieren oder Inverkehrbringen und insbesondere natürlich die Entscheidung, die letztlich gemeinsam mit dem System getroffen wird. Gelegentlich wird auch das Unterlassen Anknüpfungspunkt sein, etwa ein unterlassener Rückruf, nachdem Probleme beim Einsatz dieser Systeme bekannt geworden sind.

3.3 Problem: Die Zurechnung

Um den menschlichen Akteur für eine Fehlentscheidung aus der Interaktion zwischen ihm und dem System verantwortlich machen zu können, müsste ihm der konkrete tatbestandliche Erfolg aber auch zurechenbar sein – zumindest nach h.L. (vgl. statt vieler: Heuchemer 2023, § 13, Rn. 23; Roxin 1962, S. 411 ff.). Das ist im Kontext kooperativer Entscheidungen zwischen Mensch und System deshalb so problematisch, weil das System Teil der Interaktion wird. Selbst wenn wir im Beitrag des Systems keine Handlung oder Entscheidung im klassischen Sinne sehen würden, so dürfen wir doch nicht außer Acht lassen, wie bedeutend sein Beitrag für die Entscheidung des mit dem System interagierenden Menschen ist.

In den Situationen, in denen der Mensch eine Entscheidung auf Grundlage der durch das System aufgearbeiteten Daten trifft, ist für die Richtigkeit „seiner" späteren Entscheidung maßgeblich, dass das System richtig programmiert wurde, es die Daten richtig erfasst hat und gerade auf eine solche Art und Weise aufgearbeitet hat, dass alle entscheidungsrelevanten Informationen so dargestellt werden, dass er sie versteht und im jeweiligen Kontext einordnen kann.

In den Situationen, in denen das System aufgrund der aufgearbeiteten Daten bereits konkrete Entscheidungsmodalitäten vorschlägt und der Mensch nur noch eine Auswahl trifft (zugespitzt: der Mensch bloß noch „ja“ oder „nein“ zu einem spezifischen Vorschlag des Systems sagt), wird die Bedeutsamkeit des Beitrags des Systems noch deutlicher.

Zwar liegt in beiden Situationen die Letztentscheidung beim menschlichen Akteur, allerdings lässt sich durch eben jene Bedeutsamkeit des Beitrags des Systems durchaus fragen, ob der Zurechnungszusammenhang zwischen der menschlichen Handlung und dem tatbestandlichen Erfolg nicht schon deshalb unterbrochen sein kann (vgl. Zech 2019, S. 206 ff.; Markwalder und Simmler 2017, S. 177). Denn zum einen könnte durch eine uneingeschränkte Zurechnung die Idee hinter der Entscheidungsübertragung vom Menschen auf das System untergraben werden und zum anderen könnte dies auch normativ zweifelhaft sein (vgl. Beck 2020a, S. 46; Schaub 2019, S. 5 f.). Wir können im Folgenden davon ausgehen, dass die Einbeziehung von KI in Entscheidungsfindungen grundsätzlich auf gesellschaftliche Akzeptanz treffen würde und gerade dem Zweck dienen soll, den menschlichen Akteur zu entlasten. Das kann der Fall sein, weil nur das System die Informationsflut im zeitlichen Rahmen bzw. mit den erforderlichen Ressourcen verarbeiten kann, oder weil es rationaler arbeiten kann, oder weil das System eine geringere Fehlerquote aufweist. In diesem Fall scheint es wenig überzeugend, den menschlichen Akteur auf die gleiche Weise zur Verantwortung zu ziehen, wie wenn er ohne Beeinflussung durch das System entschieden hätte. Gerade in Situationen, in denen der menschliche Akteur keinen Einfluss auf den Einsatz des Systems hatte, er keine angemessene Schulung mit dessen Umgang erhalten hat, oder er aufgrund struktureller oder zeitlicher Gegebenheiten keine echte Chance hatte, die Vorschläge des Systems zu hinterfragen, wird dieses Problem sichtbar. Die Prüf- und Kontrollpflicht über die Entscheidungsfindung des Systems, die dem menschlichen Akteur auferlegt würde, wäre so umfangreich (sofern sie überhaupt realisierbar wäre, Stichwort „Deep Learning“), dass der Einsatz der KI in einigen Bereichen im Ergebnis wohl keine Entlastung mehr bieten und deshalb sinnlos werden würde (vgl. Markwalder und Simmler 2017, S. 178 f.; Beck 2020a, S. 46).

Je nachdem, wie die Kooperation zwischen Mensch und System ausgestaltet ist, wäre eine Zurechnung der Entscheidung allein zum menschlichen Akteur auch aus normativen Gesichtspunkten nicht mehr überzeugend. Denn je größer, gewichtiger, komplizierter und vor allem unnachvollziehbarer der Beitrag des Systems wird, desto weniger relevant wird der Beitrag des menschlichen Akteurs. Gerade in Situationen, in denen der menschliche Akteur nur noch unter vom System bereitgestellten Entscheidungsmodalitäten wählt (in der oben

bereits erwähnten Zuspitzung eines „ja“ oder „nein“), stellt sich die Frage, ob die Entscheidung bzw. der daraus resultierende Taterfolg noch als „Werk“ des menschlichen Akteurs angesehen werden kann. Es scheint hier eigentlich eher vertretbar, entsprechend der den Zurechnungszusammenhang unterbrechenden Konstruktion des „Dazwischentreten eines Dritten“ (vgl. statt vieler Wessels et al. 2022, Rn. 285), dass der Erfolg als „Werk“ des Systems und deshalb als in dessen „Verantwortungsbereich“ liegend betrachtet werden muss (vgl. Yuan 2018, S. 501; Markwalder und Simmler 2017, S. 179). Ausgehend von der Annahme, dass ein System strafrechtlich aber nicht verantwortlich gemacht werden kann, hätte die Anlehnung an die Konstruktion des „Dazwischentreten eines Dritten“ dann jedoch zur Folge, dass sich kein Akteur strafrechtlich zur Verantwortung ziehen lassen müsste. Dass dieses Ergebnis, gerade in sensiblen Lebensbereichen, nicht gänzlich überzeugen kann, liegt auf der Hand. Denn sonst würde der zunehmende Einsatz von KI und Lernenden Systemen nach und nach zu einem Rückgang strafrechtlicher Verantwortung führen, obwohl es weiterhin zur Schädigung fremder Rechtsgüter kommen würde.

3.4 Zwischenfazit

Die nicht nur juristische, sondern vor allem gesellschaftliche Forderung nach einem/einer strafrechtlich verantwortlichen Letztentscheider:in bzw. einem entsprechenden „Human in the Loop“ ist in Ansehung möglicher Verantwortungslücken also nachvollziehbar. Die vorangegangenen Ausführungen sollten allerdings gezeigt haben, dass die bloße Beteiligung eines menschlichen Akteurs nicht ausreichen kann, um auf eine „gerechte“ Art Verantwortung zuweisen zu können. Solange (zumindest) der menschliche Akteur als Letztentscheider:in keine echte Kontrolle über das System und dessen Entscheidungsfindung ausübt, würde er vielmehr zu einem bloßen „Haftungsknecht“[4] degradiert werden.

4 Die Interaktion zwischen Mensch und System am Beispiel Autonomer Waffensysteme

Im Folgenden soll die spezifische Interaktion zwischen Mensch und System am Beispiel des Einsatzes Autonomer Waffensysteme näher in den Blick genommen werden.

[4] Vgl. zur Figur des Haftungsknechts: Sharkey (2016, S. 23 ff.), Beck et al. (2023, S. 9 f.).

Der Einsatz Lernender Systeme in kriegerischen Auseinandersetzungen birgt nämlich – neben u. a. technisch, politisch, gesellschaftlich, soziologisch und ethisch implizierten Problemstellungen – weitere Besonderheiten. Durch diese Besonderheiten werden die oben aufgezeigten Probleme bei der Zurechnung und Verantwortungszuweisung in der Interaktion zwischen Mensch und System nicht nur besonders hervorgehoben, sondern vor allem in ihrer Relevanz verdeutlicht. Gewichtiger Grund hierfür ist, dass die aus der Interaktion resultierenden Entscheidungen (potenziell) solche Rechtsgüter betreffen, die im höchsten Maße sensibel sind. Geht es im Kontext von bspw. KI in der Medizin oder im Straßenverkehr nur in größten Ausnahmefällen um die Entscheidung über Leben und Tod anderer Menschen, so dienen Waffensysteme an erster Stelle diesem Ziel – der Verletzung und Tötung. Da die Beendigung von Menschenleben aber auch in kriegerischen Auseinandersetzungen nur unter vom humanitären Völkerrecht bestimmten Voraussetzungen gestattet ist, bekommt die Interaktion zwischen Mensch und System dahingehend eine neue Dimension, als kooperativ getroffene Entscheidungen eben diesen Regeln gerecht werden müssen.

4.1 Autonome Waffensysteme – Status Quo

Zunächst können wir feststellen, dass Autonome Waffensysteme im strengen Sinne des Wortes, also bewaffnete unbemannte Plattformen, die fähig sind, ohne jegliche menschliche Kontrolle im Kampfeinsatz zu agieren, zumindest noch nicht eingesetzt werden. Wenn in der heutigen Debatte über Autonome Waffensysteme oder „unbemannte Systeme“ gesprochen wird, meint das regelmäßig ferngesteuerte oder autonom agierende Maschinen, wobei „maschinelle Autonomie“ bedeutet, dass die Maschine in der Lage ist, bestimmt Aufgaben in einer dynamischen Umgebung zu erfüllen, ohne dass der Mensch eingreift/eingreifen muss (Dahlmann und Dickow 2019, S. 9). Solche unbemannten Waffensysteme finden wir in der modernen Kriegsführung vor allem in Form bewaffneter Drohnen (Grünwald und Kehl 2020, S. 67). Wegen des unklaren Entwicklungsstandes Autonomer Waffensysteme fokussiert sich unsere Betrachtung auf die Entwicklung im Softwarebereich bei Assistenzsystemen (insbesondere tiefe neuronale Netze) und die entsprechende Interaktion zwischen Mensch und System („Mensch-Maschine-Teaming“) (Dahlmann und Dickow 2019, S. 10 f.), da Assistenzsysteme die Entscheidung des menschliches Akteurs vorbereiten oder sogar in einigen Fällen übernehmen (können). Wir betrachten hier also Lernende Systeme, die in einer kriegerischen Auseinandersetzung eingesetzt werden und in diesem Zusammenhang in spezifischer Art und Weise mit dem Menschen interagieren.

4.2 Die Besonderheiten der Interaktion zwischen Mensch und Autonomem Waffensystem

Im Kontext Autonomer Waffensysteme sind es nicht nur die Programmierer:innen, Hersteller:innen, Nutzer:innen etc., die als „antwortende Entität" für die Realisierung eines strafrechtlichen Erfolgs in Frage kommen. Vielmehr finden sich im Vorfeld des Einsatzes eines Autonomen Waffensystems viele weitere Akteure wie bspw. Kommandeur:innen aber auch Politiker:innen, deren Beiträgen aufgrund der besonderen Machtmechanismen, -strukturen und -verhältnisse kriegerischer Auseinandersetzungen eine gesonderte Gewichtung zukommt (die wir aus anderen Einsatzfeldern Lernender Systeme nicht kennen).

Den obigen Ausführungen entsprechend wollen wir im Folgenden jedoch trotz dessen ausschließlich die Interaktion zwischen dem „Human in the Loop" als Letztentscheider:in und dem Lernenden System (also zwischen Bediener:innen und Waffensystem) näher in den Blick nehmen.

4.2.1 Exkurs: Die Fundamentalregeln des humanitären Völkerrechts – Unterscheidungsgrundsatz

Wie bereits angerissen, bestehen auch in kriegerischen Auseinandersetzungen Regeln darüber, unter welchen Voraussetzungen Menschen verletzt oder getötet werden dürfen. Diese Regeln sind Teil des sog. humanitären Völkerrechts und gelten für alle an einem Konflikt beteiligten Akteure und dementsprechend auch, wie oben bereits angedeutet, für solche Akteure, die mit Systemen interagieren.[5]

Da das Regelungsgeflecht des humanitären Völkerrechts komplex ist und eine vertiefte Auseinandersetzung damit an dieser Stelle zu weit gehen würde, soll es für unsere Zwecke genügen, die Möglichkeit der Implementierung humanitärvölkerrechtlicher Regelungen und die damit einhergehenden Schwierigkeiten am Beispiel des Unterscheidungsgrundsatzes[6] in der Interaktion zwischen Mensch und System näher zu beleuchten.

Der Unterscheidungsgrundsatz eignet sich deshalb als Beispiel, weil er als eine der Fundamentalregeln („Cardinal Principles") (IGH, *Legality of the Threat or Use of Nuclear Weapons,* Advisory Opinion, ICJ Rep. 1996, 226, Rn. 78.)

[5] Das humanitäre Völkerrecht (auch „ius in bello") regelt das Recht des bewaffneten Konflikts, also das „Recht im Krieg". Es regelt nicht, „Ob" militärische Gewalt eingesetzt werden darf, sondern auf welche Art und Weise, also „Wie" militärische Gewalt eingesetzt werden darf. Vgl. statt vieler: Von Arnauld (2023, § 14 Rn. 1167), Krajewski (2020, § 10 Rn. 2).

[6] Der Unterscheidungsgrundsatz findet sich in Art. 48 des Ersten Zusatzprotokolls zu den Genfer Abkommen (ZP I).

des humanitären Völkerrechts bei jedem Angriff eingehalten werden und deshalb auch und gerade beim Einsatz Lernender Systeme Berücksichtigung finden muss. Sein Regelungsgehalt dient dem Schutz der Zivilbevölkerung sowie ziviler Objekte und besagt, dass sich jegliche Kriegshandlungen ausschließlich gegen militärische Ziele richten dürfen, weshalb alle am Konflikt beteiligten Akteure dazu verpflichtet sind, jederzeit zwischen der Zivilbevölkerung und Kombattanten (also denjenigen, die berechtigt sind, unmittelbar an den Feindseligkeiten teilzunehmen)[7] sowie zwischen zivilen Objekten und militärischen Zielen zu unterscheiden. Für die am Konflikt beteiligten Akteure bedeutet das, dass sie vor jeder Angriffsentscheidung eine wertende Betrachtung am Einzelfall dahingehend vornehmen müssen, ob es sich bei dem erfassten Ziel tatsächlich um ein zulässiges militärisches Ziel handelt, oder eben nicht. Diese Bewertung ist komplex, weil sich Ziele zum einen häufig nicht eindeutig als militärisch oder zivil einordnen lassen[8] und zum anderen auch nach einer solchen Einordnung nur dann ein Angriff stattfinden darf, wenn er einen (eindeutigen) militärischen Vorteil generiert[9].

4.2.2 Die Besonderheiten der Interaktion zwischen Mensch und System in der kriegerischen Auseinandersetzung

Für die nähere Betrachtung der Interaktion zwischen Mensch und Waffensystem bietet es sich an, den kriegerischen Angriff in zwei Phasen zu unterteilen. In der ersten Phase wird das potenzielle Angriffssubjekt oder -objekt als solches identifiziert („Zielerfassung“). In der zweiten Phase wird die finale Entscheidung zum Angriff des Subjekts/Objekts getroffen („Angriffsentscheidung“) (vgl. statt vieler: Davison 2017, S. 5 f.). Je nachdem, wie die Interaktion zwischen Mensch und System ausgestaltet wird, werden bestimmte Prozesse innerhalb dieser Phasen vom System übernommen oder in Kooperation mit dem menschlichen Akteur Entscheidungen getroffen.

Stellen wir uns zur Veranschaulichung der oben beschriebenen Besonderheiten bei der Interaktion zwischen dem menschlichen Akteur und dem Autonomen Waffensystem ein Szenario vor, in dem sich der Mensch in einer Steuerungszentrale fernab des Schlachtfeldes befindet und er die Informationen, die er

[7] Vgl. Art. 43 Abs. 2 ZP I.

[8] So kann bspw. ein Objekt nicht allein einer militärischen Nutzung unterliegen (sog. „dual-use Objekte“ wie z. B. Radiostationen) oder aber ein nach außen scheinbar ziviles Objekt kann militärisch zweckentfremdet sein (wie z. B. ein Schulbus); vgl. Art. 52 Abs. 3 ZP I; Henckaerts und Doswald-Beck 2005a, Volume I, 29 ff.; Henckaerts und Doswald-Beck 2005b, Volume II, Chapter 2, §§ 493–560.

[9] Vgl. Art. 52 Abs. 2 ZP I bzgl. Objekten.

für Zielerfassung und Angriffsentscheidung benötigt, vom sich auf dem Kriegsschauplatz befindlichen System erhält. Wir betrachten also ein Szenario, in dem Mensch und System in der Phase der Zielerfassung kooperativ arbeiten, die finale Angriffsentscheidung aber allein vom menschlichen Akteur als Letztentscheider:in getroffen wird.

Damit der menschliche Akteur die Angriffsentscheidung entsprechend des Unterscheidungsgrundsatzes treffen kann, müsste er also in der Lage sein, ein zulässiges militärisches Ziel (allein) anhand der vom System bereitgestellten Informationen identifizieren zu können. Dieser Umstand macht es erforderlich, dass das System bereits bei der Auswahl und Aufarbeitung der Daten „weiß", welche Informationen relevant sind, damit später eine solche Identifizierung im Lichte des Unterscheidungsgrundsatzes stattfinden kann. Würde das System die Informationen nicht schon bei ihrer Auswahl und Aufarbeitung in diesem Lichte „filtern", so hätte dies – in Ansehung der Tatsache, dass das System aufgrund der „Informationsflut" eine Filterung vornehmen muss (vgl. Sassòli 2014, S. 41 f.; Dahlmann und Dickow 2019, S. 12 f.) – zur Folge, dass für die Unterscheidung relevante Informationen potentiell ungesehen blieben. Auf einer nächsten Stufe wäre es entscheidend, dass das System die aufgearbeiteten Informationen auf eine Art und Weise (über das Interface) bereitstellt, dass der menschliche Akteur ein solch genaues Bild aller entscheidender Umstände und Gegebenheiten auf dem Schlachtfeld erhält, dass er in die Lage versetzt wird, die potentielle Angriffssituation so zu erfassen, dass eine Wertentscheidung in Ansehung aller entscheidenden Umstände des Einzelfalls möglich wäre.

Würde das System also entweder die für eine Unterscheidung „falschen" Informationen sammeln und aufarbeiten oder sie so bereitstellen, dass der menschliche Akteur die Situation nicht in Gänze nachvollziehen kann, so wäre es dem Menschen regelmäßig nicht möglich, ein im Sinne des Unterscheidungsgrundsatzes zulässiges militärisches Ziel zu erfassen und im Falle eines Angriffs eine entsprechend völkerrechtskonforme Entscheidung zu treffen.

Neben der – v. a. technischen und ethischen – Debatte darüber, ob die Programmierung einer Regelung wie die des Unterscheidungsgrundsatzes (also einer solchen, die eine Wertentscheidung im Einzelfall enthält) in einem System überhaupt möglich ist (vgl. Davison 2017, S. 8; Winter 2020, S. 845 ff.), verdeutlicht dieses Szenario vor allem, wie bedeutend der Beitrag des Lernenden Systems im Kontext Autonomer Waffensysteme sein kann.

Obwohl der „Human in the Loop" im beschriebenen Szenario nicht nur die Angriffsentscheidung trifft, sondern auch an der Zielerfassung partizipiert, hat bereits die Auswahl und Aufarbeitung der Informationen durch das System einen solch erheblichen Einfluss auf die Zulässigkeit der späteren Angriffsentscheidung,

dass Zweifel daran aufkommen können, ob es in Ansehung dieser Abhängigkeit von Programmierung und Funktion des Systems gerecht wäre, eine Fehlentscheidung allein dem „Human in the Loop“ als Letztentscheider:in zuzurechnen. Denn zum einen ist es regelmäßig nicht der „Human in the Loop“, der sich dazu entscheidet, ein Lernendes System einzusetzen und ihm entsprechende Entscheidungskompetenz zu übertragen, da die Bediener:innen selbst in den wenigsten Fällen einen Einfluss darauf haben werden, ob die kriegerische Auseinandersetzung unter Zuhilfenahme Autonomer Waffensysteme geführt wird und sie entsprechend verpflichtet sind, mit dem System zu kooperieren. Das Gefühl, dass derjenige, der ein System zur Übernahme bestimmter Prozesse und damit zur eigenen „Entlastung“ einsetzt, auch für entsprechende Fehler des Systems eintreten muss, drängt sich hier also nicht auf. Zum anderen sind es gerade die Besonderheiten einer kriegerischen Auseinandersetzung, die es dem „Human in the Loop“ regelmäßig stark erschweren, die vom System bereitgestellten Informationen kritisch zu hinterfragen. So ist Krieg per se davon geprägt, dass Entscheidungen unter einem enormen zeitlichen Druck getroffen werden müssen, die Reaktionszeiten im Gefecht sehr verkürzt sind, das Gefecht selbst sehr unübersichtlich ist („fog of war“) usw. (vgl. Sassòli 2014, S. 41 f.).[10] Das System ist hier prinzipiell in der Lage, die Informationen schneller, rationaler sowie umfassender zu erfassen und zu verarbeiten (vgl. Sassòli, 2014, S. 41 f.; Hagström 2014, S. 24), weshalb der menschliche Akteur mit wachsender Distanz zum Schlachtfeld die Möglichkeit verliert, ein „eigenes“ (also ein von den Informationen des Systems unbeeinflusstes) Bild von der Situation im Kriegsgeschehen zu entwickeln.

Würden uns diese normativen Erwägungen nun allerdings tatsächlich zu dem Entschluss kommen lassen, dass wir dem „Human in the Loop“ die aus der Interaktion resultierende Fehlentscheidung nicht zurechnen wollen, so hätte dies zur Konsequenz, dass sich niemand für die Verletzung sensibelster Rechtsgüter durch autonomisierte/automatisierte Kriegsführung (strafrechtlich) verantworten müsste. Da dieses Ergebnis – zumindest gesellschaftlich – nicht tragbar wäre, ist es notwendig, Konzepte zu erarbeiten, die eine „gerechte“ Zurechnung und damit eine (strafrechtliche) Verantwortungszuschreibung ermöglichen.

[10] Für eine Definition von „fog of war“ vgl. Tiller et al. (2021).

5 Das Konzept der Meaningful Human Control

Als ein Lösungsansatz für die beschriebenen Problematiken wird das Konzept der Meaningful Human Control (MHC) (dt.: „bedeutsame menschliche Kontrolle“) diskutiert. Es soll sicherstellen, dass Verantwortlichkeit angemessen und gerecht verteilt wird, ohne die beteiligten Akteure unzumutbar zu belasten (vgl. Beck et al. 2023, S. 10 f.; Article36 2016, S. 1 ff.). Ziel ist es, den „Human in the Loop“ so zu positionieren, dass er eben nicht nur zum „Haftungsknecht“ degradiert wird. Denn die kooperative Entscheidungsfindung zwischen Mensch und System soll nur dann zur Verantwortlichkeit führen, wenn dies angesichts der konkreten Umstände gerechtfertigt erscheint (vgl. Chengeta 2016, S. 27; Horowitz und Scharre 2015, S. 9 ff.). Neben seiner Funktion als Abgrenzungskriterium für Verantwortungszuschreibung dient das Konzept der Meaningful Human Control also auch als Ausgestaltungskriterium für die Interaktion zwischen Mensch und System (vgl. Beck et al. 2023, S. 10 f.; Chengeta 2016, S. 50).

Bisher wurde noch kein einheitlicher Kriterienkatalog dahingehend entwickelt, welche Voraussetzungen erfüllt werden müssten, um dem menschlichen Akteur Meaningful Human Control zuschreiben zu können. Fordern wir aber, dass der menschliche Akteur eine eigene Entscheidung trifft und dem System nicht bloß „blind vertraut“, müssen wir sicherstellen, dass er zum einen ausreichende und verständliche Informationen über die maschinelle Datenverarbeitung erhält und zum anderen die echte Möglichkeit bekommt, mit dem System in einen kritischen Diskurs zu treten („Erfordernis der Interpretierbarkeit und Erklärbarkeit“) (vgl. Amoroso und Tamburrini 2019, S. 9). Letztlich sind es deshalb viele einzelne Parameter, die dafür entscheidend sein können, ob der/die Letztentscheider:in echte Kontrolle über das System ausüben kann. Dabei können die entscheidenden Parameter – wie oben angerissen – auf den unterschiedlichsten Ebenen eine Rolle spielen: So kann es bspw. bedeutend sein, dass der „Human in the Loop“ nachvollziehen können muss, welche Daten erhoben wurden; warum sie erhoben wurden; wie sie gefiltert wurden oder wieso sie auf eine bestimmte Art und Weise aufgearbeitet wurden. Auch kann es von Bedeutung sein, dass der Mensch mit dem System insofern interagieren kann, als er Rückfragen stellen oder in einen Diskurs treten kann. Am Beispiel der Interaktion zwischen Mensch und Waffensystem haben wir gesehen, dass der menschliche Akteur andernfalls unter Umständen nicht in die Lage versetzt werden kann, eine Angriffsentscheidung zu treffen, die den Prinzipien des (humanitären) Völkerrechts gerecht wird. Gerade in diesem Kontext kann es auf einer weiteren Ebene erforderlich sein, dass das Interface als Schnittstelle zwischen Mensch und System – und damit als Medium der Interaktion – ein bestimmtes Design hat.

Da die potentiellen Anknüpfungspunkte möglicher Parameter zur Sicherstellung echter Kontrolle des Menschen über das System mannigfaltig sind, ist es letztlich von der konkreten Einsatzsituation des Mensch-Maschine-Teamings abhängig, welche Kriterien notwendigerweise erfüllt werden müssten, um Meaningful Human Control sicherstellen zu können. Es ist deshalb erforderlich, dass das Konzept der Meaningful Human Control nicht erst in der Interaktion zwischen dem „Human in the Loop" als Letztentscheider:in und dem System eine Rolle spielt, sondern bereits in der Forschungs-, Entwicklungs- und Trainingsphase automatisierter/autonomisierter Systeme Berücksichtigung findet.

An dieser Stelle bietet es sich an, auf das überregionale und interdisziplinäre Kompetenznetz „Meaningful Human Control. Autonome Waffensysteme zwischen Regulation und Reflexion"[11] hinzuweisen, in dem von Forscher:innen und Fellows unterschiedlichster Disziplinen (Robotik, Rechtswissenschaft, Soziologie, Physik, Politikwissenschaft, Gender Studies und Medienwissenschaft) eine solche umfassende Betrachtung vorgenommen wird. Ziel ist es, durch die Analyse und Verbindung bisher unverbundener Problembereiche ein Konzept der Meaningful Human Control erarbeiten zu können, durch das sichergestellt werden kann, dass die Entscheidung aus der Interaktion zwischen Mensch und System tatsächlich eine menschliche ist, die dem „Human in the Loop" als Letztentscheider:in deshalb auf eine gerechte Art zugerechnet werden kann und für die er/sie sich gerechter Weise (strafrechtlich) verantworten muss.

6 Schlussfolgerungen

Fassen wir zusammen: Bei der Forderung nach Meaningful Human Control geht es nicht immer explizit darum, Kriterien für die Zulassung von Systemen aufzustellen. In spezifischen Kontexten mag es sinnvoll erscheinen, diese Forderung konkret herunterzubrechen, zu Gesetzesänderungen, Vorgaben in technischen Normen, Ausgestaltung der Mensch-Maschine-Interaktion bzw. des Designs der Maschine. Denn nur so lässt sich diese generalisierende Überlegung im Alltag umsetzen. Hier aber lag die Perspektive auf der Zuschreibung (strafrechtlicher) Verantwortlichkeit des „Human in the Loop" als Letztentscheider:in und eine gerechte Verteilung dieser Verantwortung durch die Entwicklung normativer Kriterien. Eine solche gerechte Verteilung kann dann jedoch auch bedeuten, dass

[11] Das Projekt wird mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen 01UG2206B gefördert. Für weitere Informationen siehe www.mehuco.de.

wir in manchen Situationen keine strafrechtliche Verantwortung zuschreiben können. Da wir überall dort, wo wir eine Zuschreibung beibehalten wollen (was in einem gesellschaftlichen Diskurs zu erarbeiten wäre), Meaningful Human Control herstellen müssen, damit die Zuschreibung gerecht bleibt, wird Meaningful Human Control also implizit eine Voraussetzung für den Einsatz autonomer/automatisierter Waffensysteme und die entsprechende Interaktion zwischen Mensch und Maschine.

Literatur

Apel, S. & Kaulartz, M. (2020). Rechtlicher Schutz von Machine Learning-Modellen. *Recht Digital*, *1*, 24–34.

Amoroso, D., Tamburrini, G. (2019). *What makes human control over weapon systems "meaningful"?*. ICRAC Working Paper, 4, 1–21.

Article36. (2016). *Key elements of meaningful human control*. https://article36.org/wp-content/uploads/2016/04/MHC-2016-FINAL.pdf (zuletzt abgerufen am 26.04.2023).

Beck, S. (2015). *Google-Cars, Software-Agents, Autonome Waffensysteme – neue Herausforderungen für das Strafrecht?*. In S. Beck, B.-D. Meier & C. Momsen (Hrsg.), *Cybercrime und Cyberinvestigations* (S. 9–34). Baden-Baden: Nomos.

Beck, S. (2020a). *Die Diffusion strafrechtlicher Verantwortlichkeit durch Digitalisierung und Lernende Systeme*. Zeitschrift für Internationale Strafrechtsdogmatik, 2, 41–50.

Beck, S. (2020b). *Künstliche Intelligenz – ethische und rechtliche Herausforderungen*. In K. Mainzer (Hrsg.), *Philosophisches Handbuch Künstliche Intelligenz* (S. 1–28). Wiesbaden: Springer VS.

Beck, S. (2020c). *Strafrechtliche Implikationen von KI und Robotik*. In M. Ebers, C. A. Heinze, T. Krügel & B. Steinrötter (Hrsg.), *Künstliche Intelligenz und Robotik: Rechtshandbuch* (S. 243–269). München: C.H.Beck.

Beck, S. (2023). *Diffusion individueller rechtlicher Verantwortlichkeit beim Einsatz Lernender Systeme*. Monatsschrift für Kriminologie und Strafrechtsreform, 1, 29-37.

Beck, S., Faber, M., Gerndt, S. (2023). *Rechtliche Aspekte des Einsatzes von KI und Robotik in Medizin und Pflege*. Ethik in der Medizin. https://link.springer.com/article/10.1007/s00481-023-00763-9 (zuletzt abgerufen am 26.04.2023).

Chengeta, T. (2016). *Accountability Gap: Autonomous Weapon Systems and Modes of Responsibility in International Law*. Denver Journal of International Law & Policy, 45, 1–50.

Crootof, R. (2016). A Meaningful Floor for "Meaningful Human Control". *Temple International & Comparative Law Journal*, *30*, 53–62.

Dahlmann, A., Dickow, M. (2019). *Preventive Regulation of Autonomous Weapon Systems*. SWP Research Paper 3, 1 –24.

Davison, N. (2017). *A legal perspective: Autonomous weapon systems under international humanitarian law*. UNODA Occasional Papers, 30, 5–18.

Denga, M. (2018). Deliktische Haftung für künstliche Intelligenz. *Computer und Recht*, *2*, 69–78.

Dettling, H.-U. (2019). Künstliche Intelligenz und digitale Unterstützung ärztlicher Entscheidungen in Diagnostik und Therapie. *Pharmarecht*, *12*, 633–642.

Fraunhofer Gesellschaft. (2018). *Maschinelles Lernen. Eine Analyse zu Kompetenzen, Forschung und Anwendung*. https://bit.ly/2KAC5ny. Zugegriffen: 20. April 2023.

Gaede, K. (2019). *Künstliche Intelligenz – Rechte und Strafen für Roboter*. Baden-Baden: Nomos.

Grünwald, R., Kehl, C. (2020). *Autonome Waffensysteme. Endbericht zum TA-Projekt*. TAB-Arbeitsbericht Nr. 187.

Hagström, M., (2014). *Characteristics of autonomous weapon systems autonomous weapon systems*. In: IKRK, *Autonomous Weapon Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons* (23–25). https://icrcndresourcecentre.org/wp-content/uploads/2017/11/4283_002_Autonomus-Weapon-Systems_WEB.pdf (zuletzt abgerufen am 20.04.2023).

Henckaerts, J-M., Doswald-Beck, L. (2005a). *Customary International Humanitarian Law: Volume I: Rules*. Cambridge: Cambridge University Press.

Henckaerts, J-M., Doswald-Beck, L. (2005b). *Customary International Humanitarian Law: Volume II: Practice*. Cambridge: Cambridge University Press.

Heuchemer, M. (2023). *Beck'scher Online-Kommentar, Strafgesetzbuch* (56. Edition). Stand: 01.02.2023. München: C.H. Beck.

Horowitz, M., Scharre, P. (2015). *Meaningful Human Control in Weapon Systems: A Primer*. https://www.files.ethz.ch/isn/189786/Ethical_Autonomy_Working_Paper_031315.pdf (zuletzt abgerufen am 20.04.2023).

Krajewski, M. (2020). *Völkerrecht* (2. Aufl.). Baden-Baden: Nomos.

Kühl, K. (2017). *Strafrecht Allgemeiner Teil* (8. Aufl.). München: Franz Vahlen.

Markwalder, N., Simmler, M (2017). *Roboterstrafrecht. Zur strafrechtlichen Verantwortlichkeit von Robotern und künstlicher Intelligenz*. Aktuelle juristische Praxis, 2, 171–182.

Martini, M. (2017). Algorithmen als Herausforderung für die Rechtsordnung. *Juristen Zeitung*, *72*(21), 1017–1072.

Roxin, C. (1962). *Pflichtwidrigkeit und Erfolg bei fahrlässigen Delikten*. Zeitschrift für die gesamte Strafrechtswissenschaft, 74, 411–444.

Roxin, C., Greco, L. (2020). *Strafrecht Allgemeiner Teil Band I: Grundlagen. Der Aufbau der Verbrechenslehre* (5. Aufl.). München: C.H.Beck.

Sassòli, M., (2014). *Can autonomous weapon systems respect the principles of distinction, proportionality and precaution?*. In: IKRK, *Autonomous weapon systems technical, military, legal and humanitarian aspects* (41–43). https://www.icrc.org/en/document/report-icrc-meeting-autonomous-weapon-systems-26-28-march-2014 (zuletzt abgerufen am 20.04.2023).

Schaub, R. (2019). *Verantwortlichkeit für Algorithmen im Internet*. Zeitschrift für Innovations- und Technikrecht, 1, 2–7.

Schwarz, E. (2021). Autonomous Weapons Systems, Artificial Intelligence, and the Problem of Meaningful Human Control. *The Philosophical Journal of Conflict and Violence*, *V*(1), 53–72.

Sharkey, N. (2016). *Staying in the loop: the human supervisory control of weapons*. In: Bhuta, N. et al. (Hrsg.), *Autonomous weapons systems: law, ethics, policy* (23–38). Cambridge University Press: Cambridge.

Tiller, S., Devidal, P., Solinge, D. (2021). *The 'fog of war' . . . and information.* https://blogs.icrc.org/law-and-policy/2021/03/30/fog-of-war-and-information/ (zuletzt abgerufen am 20.04.2023).

Umbrello, S. (2021). Coupling levels of abstraction in understanding meaningful human control of autonomous weapons: a two-tiered approach. *Ethics and Information Technology*, *23*, 455–464.

von Arnauld, A. (2023). *Völkerrecht* (5. Aufl.). Heidelberg: C.F. Müller.

Wagner, B. (2019). Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems. *Policy and Internet*, *11*(1), 104–122.

Wessels, J., Beulke, W., Satzger, H. (2022). *Strafrecht Allgemeiner Teil* (52. Aufl.). Heidelberg: C.F. Müller.

Winter, L. (2020). The Compatibility of Autonomous Weapons with the Principle of Distinction in the Law of Armed Conflict. International and Comparative Law Quarterly, 69, 845–876.

Yuan, T. (2018). *Lernende Roboter und Fahrlässigkeitsdelikt*. Rechtswissenschaft, 4, 477–504.

Zech, H. (2019). *Künstliche Intelligenz und Haftungsfragen.* Zeitschrift für die gesamte Privatrechtswissenschaft, 2, 198–219.

Prof. Dr. Susanne Beck Kriminalwissenschaftliches Institut der Leibniz Universität Hannover.

Susanne Beck ist Professorin für Strafrecht, Strafprozessrecht, Strafrechtsvergleichung und Rechtsphilosophie in Hannover. Nach Promotion und Habilitation an der Universität Würzburg erfolgte 2013 der Ruf nach Hannover. Sie ist Mitbegründerin der Forschungsstelle RobotRecht in Hannover und arbeitet seit über einem Jahrzehnt an Fragen der Regulierung neuer technologischer sowie medizinischer Entwicklungen. Sie ist u.a. Mitglied der Plattform Lernende Systeme, von acatech sowie der Akademie für Ethik in der Medizin und der Braunschweigischen Wissenschaftlichen Gesellschaft.

Simone Tiedau ist Diplomjuristin und arbeitete bis 2023 als Wissenschaftliche Mitarbeiterin im Verbundsprojekt „MeHuCo – Autonome Waffensysteme zwischen Regulation und Reflexion" bei Prof. Dr. Susanne Beck am Lehrstuhl für Strafrecht, Strafprozessrecht, Strafrechtsvergleichung und Rechtsphilosophie.

Wie lässt sich erweitertes Zusammenwirken adäquat messen und untersuchen?

Überlegungen und ein Lösungsansatz zur Diskussion gestellt

Reinhold Haux und Klaus-Hendrik Wolf

Zusammenfassung

Wie lässt sich erweitertes Zusammenwirken von Menschen, Tieren und Pflanzen einerseits und von Maschinen andererseits adäquat messen und untersuchen? Nach drei Einstiegen mit Zitaten aus Werken von Immanuel Kant, Joseph Maria Bochenski und Wilhelm Gaus werden in dieser Einführung in die Thematik des adäquaten Messens und Untersuchens vergleichende Interventionsstudien, insbesondere randomisierte Studien, als Lösungsansatz vorgeschlagen und zur Diskussion gestellt. Diese Methodik kann die bei dem Zusammenwirken von natürlicher und künstlicher Intelligenz vorhandene Komplexität berücksichtigen. Zudem steht ein gut entwickeltes und untersuchtes Spektrum von Forschungsansätzen und Studienarten mit entsprechender

RH hat auf dem 2. SYnENZ Symposium am 15.2.2023 einen einführenden Vortrag zu dem Teil Messen gehalten. Die anschließende schriftliche Ausarbeitung erfolgte durch RH und KHW, die beide Mitglieder der SYnENZ-Kommission der BWG sind.

R. Haux (✉)
Braunschweigische Wissenschaftliche Gesellschaft, Braunschweig, Deutschland
E-Mail: reinhold.haux@plri.de; info@bwg.niedersachsen.de

R. Haux · K.-H. Wolf
Peter L. Reichertz Institut für Medizinische Informatik der TU Braunschweig, Braunschweig, Deutschland
E-Mail: klaus-hendrik.wolf@plri.de

O. Richter et al. (Hrsg.), *Zusammenwirken von natürlicher und künstlicher Intelligenz: Beurteilen-Messen-Bewerten,*
https://doi.org/10.1007/978-3-658-45845-4_7

Methodik zur Verfügung, einschließlich der dazu notwendigen Überlegungen zu ethischen Rahmenbedingungen.

Schlüsselwörter

Zusammenwirken • Messen • Evaluation • Künstliche Intelligenz • Menschliche Intelligenz • Natürliche Intelligenz

1 Einleitung

Auf dem 2. SYnENZ Symposium wurden u. a. folgende Fragen gestellt:

- „Wie wird das Zusammenleben und -wirken von Menschen, Tieren und Pflanzen einerseits und Maschinen andererseits zukünftig aussehen?“;
- „Lassen sich Umfang und Intensität der neuen Synergien bestimmen?“ (SYnENZ Symposium, https://synenz.de/Symposium2023).

Dabei ging es besonders um die Bestimmung von Umfang und Intensität der neuen Synergien des *erweiterten* Zusammenwirkens, mit dem sich die SYnENZ-Kommission der Braunschweigischen Wissenschaftlichen Gesellschaft seit mehreren Jahren befasst (Braunschweigische Wissenschaftliche Gesellschaft, http://bwg-nds.de/kommissionen/kommission-synenz/).

In dem Teil ‚Messen‘ des Symposiums ging es um folgende Themen:

- „Wie lässt sich erweitertes Zusammenwirken adäquat messen und untersuchen?“;
- „Gibt es existierende empirische Ansätze, die hier genutzt werden könnten, z. B. randomisierte Studien, wie sie in der Medizin üblich sind?“ (SYnENZ Symposium, https://synenz.de/Symposium2023)

Ziel dieser Ausarbeitung ist es, in die Thematik des adäquaten Messens und Untersuchens erweiterten Zusammenwirkens einzuführen, dies aus Sicht und mit dem fachlichen Hintergrund der Autoren, die beide als Medizininformatiker in der Tradition der Fachgebiete Medizinische Informatik und Medizinische Biometrie bzw. Medizinische Statistik stehen.

Nach drei Einstiegen zum adäquaten Messen und Untersuchen in Abschn. 2 wird in Abschn. 3 ein Lösungsansatz skizziert und anschließend zur Diskussion gestellt (Abschn. 4).

Die Frage des adäquaten Messens und Untersuchens erweiterten Zusammenwirkens konzentriert sich hier auf empirische Ansätze, um zu entsprechenden Erkenntnissen zu kommen.

2 Drei Einstiege zum adäquaten Messen und Untersuchen

2.1 Einleitung

Zur Erarbeitung von Überlegungen zu geeigneten empirischen Ansätzen für das adäquate Messen und Untersuchen erweiterten Zusammenwirkens mögen die folgenden drei Einstiege über Immanuel Kant, Joseph Maria Bochenski und Wilhelm Gaus hilfreich sein.

2.2 Prinzipien, nach denen allein übereinkommende Erscheinungen für Gesetze gelten können

In der von Immanuel Kant 1787 erschienen zweiten Ausgabe der ‚Kritik der reinen Vernunft' findet man folgende Aussage:

„Die Vernunft muss mit ihren Prinzipien, nach denen allein übereinkommende Erscheinungen für Gesetze gelten können, in einer Hand, und mit dem Experiment, das sie nach jenen ausdachte, in der anderen, an die Natur gehen, zwar um von ihr belehrt zu werden, aber nicht in der Qualität eines Schülers, der sich alles vorsagen lässt, was der Lehrer will, sondern eines bestallten Richters, der die Zeugen nötigt, auf die Fragen zu antworten, die er ihnen vorlegt." (Kant 1956, S. 18).

2.3 Die wichtigsten zeitgenössischen allgemeinen Denkmethoden

Joseph Maria Bochenski hat sich in seinem Werk ‚Die zeitgenössischen Denkmethoden' die Aufgabe gestellt, einen systematischen Überblick über die unterschiedlichen Methoden zur Erkenntnisgewinnung zu erarbeiten und zu präsentieren. In seinen Worten: „Dieses kleine Buch ist ein Versuch, die wichtigsten zeitgenössischen *allgemeinen* – d. h. in vielen Gebieten gebrauchten – *Denk*methoden in einer sehr elementaren Art und Weise gemäß den Ansichten der

heutigen Methodologen zu referieren“ (Bochenski 1954, S. 7). Die von den Autoren vorher gewählte Bezeichnung „Methoden zur Erkenntnisgewinnung“ würde Joseph Maria Bochenski vermutlich durch „Methoden des Denkens“ (Bochenski 1954, S. 7) ersetzen.

In seinem 1954 erstmals erschienenen Werk untergliedert er die Methoden anhand „folgende[r] Einteilung:

1. Die phänomenologische Methode.
2. Die Sprachanalyse.
3. Die deduktive Methode.
4. Die reduktive Methode.

(Bochenski 1954, S. 21). Diese werden in dem Buch in folgenden Kapiteln erläutert:

> „II.Die phänomenologische Methode […]
>
> III.Die semiotischen Methoden […]
>
> IV.Die axiomatische Methode […]
>
> V.Die reduktiven Methoden […]“

(Bochenski 1954, S. 3–5).

2.4 Jeder Mensch ist einmalig

Wilhelm Gaus und Koautoren beginnen ihr Lehrbuch ‚Medizinische Statistik‘ mit folgenden Sätzen: „**Jeder Mensch ist einmalig,** sowohl in seiner genetischen Veranlagung (eineiige Mehrlinge ausgenommen) als auch in seinem Lebenslauf. Deshalb können wir nicht erwarten, dass diagnostische Verfahren immer den richtigen Befund liefern und Therapien immer gleich wirken. Vielmehr sind die meisten Diagnosen richtig, aber nicht alle; häufig hilft eine Therapie, aber nicht immer, und gelegentlich tritt eine Komplikation ein. Damit sind wir schon mitten in der Statistik.“ (Gaus et al. 2023, S. 5).

2.5 Überleitung

Gefragt wird in dem in Abschn. 2.2 zitierten sogenannten Richter-Zitat Immanuel Kants nach adäquaten empirischen Ansätzen für Mess- und Untersuchungsmethoden, um, in Worten unserer Fachsprache, zu Forschungsansätzen und Studienarten zu kommen, die zu einer tatsächlichen Erkenntnisgewinnung führen können. Ein bloßes Beobachten der Wirklichkeit (der „Natur“) erscheint nicht immer ausreichend zu sein.

Wilhelm Gaus deutet die besondere Problematik der Erkenntnisgewinnung an, wenn der Mensch selbst betroffen, selbst Element solcher Untersuchungen ist, bei ihm mit Bezug auf Erkrankungen und auf adäquate diagnostische und therapeutische Verfahren. Mit „Jeder Mensch ist einmalig“ wird die Aussage verbunden, dass es uns nicht möglich ist, sichere Vorhersagen zu machen, wenn komplexe, nicht standardisierte und nicht vollständig beschriebene ‚Entitäten‘, wie menschliche Individuen, in Forschungsansätzen und (empirischen) Studien selbst als Untersuchungsgegenstand mit einbezogen sind.

In den zeitgenössischen Denkmethoden Joseph Maria Bochenskis findet man diese in der Medizinischen Statistik verwendeten methodischen Ansätze im Kapitel V („Die reduktiven Methoden“) und dort in dem vierten von sechs Abschnitten, mit der Überschrift Wahrscheinlichkeit und Statistik. Offensichtlich bilden diese Methoden nur einen Teil eines viel umfassenderen Spektrums zeitgenössischer Denkmethoden.

3 Lösungsansätze

Als unseres Erachtens passenden Lösungsansatz, um erweitertes Zusammenwirken adäquat messen und untersuchen zu können, und um damit Antworten auf die zu Beginn gestellten Fragen der Bestimmung von Umfang und Intensität der neuen Synergien geben zu können, wollen wir eine Studienart mit dazugehöriger Methodik vorschlagen, die sich in der Medizin, dort vor allem in der Therapieforschung, bewährt hat. Eine Einteilung von verschiedenen Studienarten für empirisch ausgerichtete Forschungsansätze präsentieren Gaus et al. in der nachfolgenden Abb. 1.

Wählen wir bei Eingriff ‚Interventionsstudien‘ – in unserem Fall würde der Eingriff das Zusammenwirken von „Menschen, Tieren und Pflanzen einerseits und Maschinen andererseits“ (siehe Abschn. 1) betreffen – und bei Fragestellungen ‚Therapien beurteilen‘ – in unserem Fall müsste es ‚Zusammenwirken

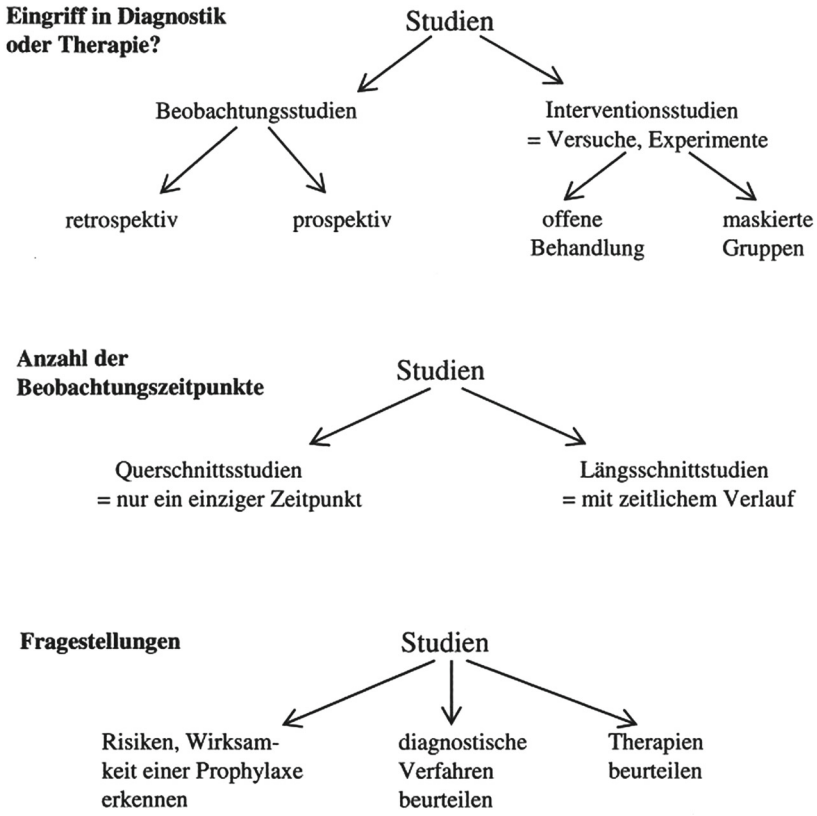

Abb. 1 Studienarten nach Gaus et al. (2023, S. 38) anhand von drei Einteilungskriterien, hier bezogen auf Medizin und Gesundheitsversorgung als Anwendungsgebiet

beurteilen' heißen – dann kommen wir zu den in der Medizin seit Jahren erfolgreich etablierten kontrollierten Studien, bei denen randomisierte Studien als die bestmögliche Studienform gelten.

Zur Einführung und Beschreibung dieses Ansatzes, sowohl im Hinblick auf die Methodik als auch im Hinblick auf ethische Aspekte – immerhin handelt es sich um Experimente, bei denen auch Menschen betroffen sind –, sei auf die einschlägige Literatur verwiesen (u. a. Gaus et al. 2023; Martini 1962; Immich 1974; Jesdinsky 1978; Überla 1981; Schumacher 2016). Der erste Autor hat diese

Überlegungen in Haux und Karafyllis (2021) weiter begründet sowie in Haux (2021) dazu eine auf erweitertes Zusammenwirken ausgerichtete Studienart vorgeschlagen. In Arbeiten des zweiten Autors, der diese Methodik bei der Nutzung digitaler Therapeutika für die Rehabilitation von Patienten mit Schultererkrankungen in komplexen sozio-technischen Systemen angewandt hat, finden sich zudem weitere methodische Ausführungen (Wolf et al. 2016; Steiner et al. 2020a, b, c; Elgert et al. 2021, 2022; Saalfeld et al. 2022).

4 Zur Diskussion gestellt

Diese Überlegungen zur Nutzung der Methodik randomisierter Studien für das adäquate Messen und Untersuchen erweiterten Zusammenwirkens möchten wir zur Diskussion stellen. Wir sind uns bewusst, dass – wie in Joseph Maria Bochenskis zeitgenössischen Denkmethoden ausgeführt – diese ‚Denkmethode' eine von vielen darstellt. Wir sind uns auch bewusst, dass wir aufgrund unseres fachlichen Hintergrunds – der Medizinischen Informatik und Medizinischen Biometrie bzw. Medizinischen Statistik – möglicherweise einseitig argumentieren. Dennoch: Diese Methodik kann jedenfalls die beim (ggf. erweiterten) Zusammenwirken von natürlicher und künstlicher Intelligenz vorhandene Komplexität, die kaum vollständig plan- und analysierbar ist und bei der verschiedene Instanziierungen desselben Systems unterschiedliche Ergebnisse liefern können, berücksichtigen. Zudem steht ein sehr gut entwickeltes und untersuchtes Spektrum von Forschungsansätzen und Studienarten mit entsprechender Methodik zur Verfügung, einschließlich der dazu notwendigen Überlegungen zu ethischen Rahmenbedingungen. Diese müssten vermutlich im Hinblick auf das erweiterte Zusammenwirken, das auf dem 2. SYnENZ Symposium Gegenstand war, noch weiterentwickelt bzw. adaptiert werden.

Literatur

2. SYnENZ Symposium. Zusammenwirken von natürlicher und künstlicher Intelligenz – über das erweiterte Zusammenwirken lebender und nicht lebender Entitäten im Zeitalter der Digitalisierung –. https://synenz.de/Symposium2023. Zuletzt zugegriffen am 2.5.2023.

Bochenski I M. Die zeitgenössischen Denkmethoden. München: Franke; 1954. Zitiert aus der 8. Auflage von 1980.

Braunschweigische Wissenschaftliche Gesellschaft. Kommission Synergie und Intelligenz: technische, ethische und rechtliche Herausforderungen des Zusammenwirkens lebender

und nicht lebender Entitäten im Zeitalter der Digitalisierung (SYnENZ). http://bwg-nds.de/kommissionen/kommission-synenz/. Zuletzt zugegriffen am 2.5.2023.

Elgert L, Steiner B, Saalfeld B, Marschollek M, Wolf KH. Health-Enabling Technologies to Assist Patients With Musculoskeletal Shoulder Disorders When Exercising at Home: Scoping Review. JMIR Rehabil Assist Technol. 2021; 8: e21107.

Elgert L, Steiner B, Saalfeld B, Marschollek M, Wolf KH. Factors for Individualization of Therapeutic Exercises for the Design of Health-Enabling Technologies. Stud Health Technol Inform. 2022; 289: 136–9.

Gaus W, Muche R, Mayer B. Medizinische Statistik, 3. Auflage. Norderstedt, Books on Demand; 2023. Zitiert aus der 1. Auflage von 2014.

Haux R. Lässt sich erweitertes Zusammenwirken empirisch ermitteln? Überlegungen zur Adaptierung des Turing-Tests. Braunschweigische Wissenschaftliche Gesellschaft. Jahrbuch 2020, 99–116. Göttingen: Cuvillier; 2021.

Haux R, Karafyllis NC. Methodisch-technische Aspekte der Evaluation erweiterten Zusammenwirkens. In: Haux R, Gahl K, Jipp M, Kruse R, Richter O. Herausgeber. Zusammenwirken von natürlicher und künstlicher Intelligenz, 175–98. Wiesbaden: Springer VS; 2021.

Immich H. Medizinische Statistik. Stuttgart: Schattauer; 1974.

Jesdinsky HJ, Herausgeber. Memorandum zur Planung und Durchführung kontrollierter klinischer Studien. Stuttgart: Schattauer; 1978.

Kant I. Kritik der reinen Vernunft, 2. Auflage, 1787. Zitiert aus der von Raymund Schmidt herausgegebenen und in der Philosophischen Bibliothek des Felix-Meiner-Verlags, Hamburg, 1956 erschienenen Ausgabe (durchgesehener Nachdruck 1976.)

Martini P. Grundsätzliches zur therapeutisch-klinischen Versuchsplanung. Methods Inf Med. 1962; 1: 1–5.

Saalfeld B, Elgert L, Steiner B, Wolf KH. Compiling Criteria for Assessing Essential Aspects of Home Exercise Performance: A Questionnaire-Based Approach. Stud Health Technol Inform. 2022; 290: 484–8.

Schumacher M. (2016). Entwicklung klinischer Studien von Paul Martini bis heute. Drug Res. 2016; 66: 5–7.

Steiner B, Elgert L, Haux R, Wolf KH. AGT-Reha-WK study: protocol for a non- inferiority trial comparing the efficacy and costs of home-based telerehabilitation for shoulder diseases with medical exercise therapy. BMJ Open. 2020; 10: e036881.

Steiner B, Elgert L, Saalfeld B, Wolf KH. Gamification in Rehabilitation of Patients With Musculoskeletal Diseases of the Shoulder: Scoping Review. JMIR Serious Games. 2020; 8: e19914.

Steiner B, Elgert L, Saalfeld B, Schwartze J, Borrmann HP, Kobelt-Pönicke A, Figlewicz A, Kasprowski D, Thiel M, Kreikebohm R, Haux R, Wolf KH. Health- Enabling Technologies for Telerehabilitation of the Shoulder: A Feasibility and User Acceptance Study. Methods Inf Med. 2020; 59: e90–9.

Überla KK. Therapiestudien: Indikation, Erkenntniswert und Herausforderung. In: Victor N, Dudeck J, Broszio, EP, Herausgeber. Therapiestudien, 7–21. Berlin: Springer; 1981.

Wolf, K.-H. et al. – Studiengruppe AGT-Reha. (2016). Evaluation der Wirksamkeit und Kosten der poststationären häuslichen Tele-Rehabilitation mit AGT-Reha im Vergleich zur Medizinischen Trainingstherapie. Bericht mit Studienplan.

Prof. Dr. Reinhold Haux Peter L. Reichertz Institut für Medizinische Informatik der TU Braunschweig und der Medizinischen Hochschule Hannover (PLRI). Reinhold Haux ist Präsident der Braunschweigischen Wissenschaftlichen Gesellschaft (BWG) und emeritierter Professor für Medizinische Informatik am Peter L. Reichertz Institut für Medizinische Informatik der TU Braunschweig und der Medizinischen Hochschule Hannover (MHH). Nach Professuren an Universitäten in Tübingen (1987–1989), Heidelberg (1989–2001) und Innsbruck (2001–2004) folgte er 2004 einem Ruf an die Technische Universität Braunschweig. Er war Präsident der International Medical Informatics Association (2007–2010), der International Academy of Health Sciences Informatics (2018–2020) und Herausgeber der Zeitschrift Methods of Information in Medicine (2001–2015). Er ist Honorarprofessor an der Universität Heidelberg und kooptiertes Mitglied des Lehrkörpers der MHH. Seit ihrer Gründung 2017 ist er Mitglied der SYnENZ-Kommission der BWG. Weitere Informationen auf www.plri.de.

Dr. Klaus-Hendrik Wolf Peter L. Reichertz Institut für Medizinische Informatik der TU Braunschweig und der Medizinischen Hochschule Hannover (PLRI). Klaus-Hendrik Wolf ist wissenschaftlicher Mitarbeiter am PLRI. Nach dem Studium der Medizinischen Informatik an der Universität Hildesheim und der TU Braunschweig war er seit 2000 im PLRI zunächst an der TU Braunschweig und seit 2018 an der Medizinischen Hochschule Hannover tätig. Er war Chair der Working Group Wearable Sensors in Healthcare der International Medical Informatics Association. Zu seinen Forschungsschwerpunkten zählen Assistierende Gesundheitstechnologien und Virtuelle Medizin.

Closing the Circle in a Learning Health System

Dominik Wolff

Zusammenfassung

Die Anzahl an künstlichen Intelligenzen zur Unterstützung von medizinisch Tätigen steigt stetig. Sie sind in der Lage große heterogene Datenmengen in kürzester Zeit zu sichten und für den Menschen schwer greifbare Zusammenhänge zu identifizieren. Aktuell beschränkt sich der Einsatz von künstlichen Intelligenzen in der Medizin in der Regel auf die Automatisierung von Aufgaben, sodass sie als reines Werkzeug angesehen werden. Wissensbasiert oder datengetrieben werden die künstlichen Intelligenzen zum Experten in einer abgegrenzten Aufgabenstellung, sodass deren Erfüllung kostengünstig, orts-, zeit- und personenunabhängig erfolgen kann. Auf der anderen Seite bietet die Lernfähigkeit mancher Systeme die Möglichkeit, dem Menschen unbekanntes Wissen im Entscheidungsprozess zu berücksichtigen. Die Erhebung und Darstellung dieses Wissens in für Menschen verständlicher Weise und eine anschließende Evaluation durch Experten kann neues medizinischen Wissen erschaffen und die Versorgungsqualität erhöhen. Der sich so schließende Kreislauf des Zusammenwirkens von natürlichen und künstlichen Intelligenzen in einem lernenden Gesundheitssystem (eng.: Learning Health System), bei denen künstliche Intelligenzen vom Menschen und der Mensch von den künstlichen Intelligenzen lernt, sowie potentielle Methoden, um den Mehrwert zu messen, werden diskutiert und am Beispiel der automatisierten Edukation pflegender Angehöriger erörtert.

D. Wolff (✉)
Peter L. Reichertz Institut für Medizinische Informatik der TU Braunschweig und der Medizinischen Hochschule Hannover, Hannover, Deutschland
E-Mail: Dominik.Wolff@plri.de

O. Richter et al. (Hrsg.), *Zusammenwirken von natürlicher und künstlicher Intelligenz: Beurteilen-Messen-Bewerten,*
https://doi.org/10.1007/978-3-658-45845-4_8

Schlüsselwörter

Lernendes Gesundheitssystem • Synergie • Messen • Erklärbare künstliche Intelligenz

1 Einleitung

Künstliche Intelligenz (KI) hält in immer mehr Bereiche des täglichen Lebens Einzug. Im Gesundheitswesen wird sie zur Personalisierung der Prozesse an den Patienten eingesetzt. Ein Großteil der Anwendungen, die klinischen Entscheidungsunterstützungssysteme, zielt darauf ab, Ärzt*innen und Patient*innen in ihrer Entscheidungsfindung zu unterstützen. Es existieren bereits in allen Phasen der Behandlung Anwendungen von künstlicher Intelligenz. Für die Einweisung von Patient*innen über die Notaufnahme ins Krankenhaus werden beispielsweise das Patientenaufkommen und die damit verbundene Anzahl benötigter Fachkräfte (Graham et al. 2018) oder auch die Wiedereinweisung innerhalb der nächsten 72 h nach Entlassung (Lee et al. 2012) vorhergesagt. Nach der Aufnahme kann künstliche Intelligenz bei der Diagnostik in vielen medizinischen Bereichen, wie der Radiologie (Rajpurkar et al. 2017; Kniep et al. 2020) oder der Kardiologie (Bizopoulos und Koutsouris 2019), oder auch dem Monitoring von Biosignalen am Patientenbett (Mollura et al. 2021; Kaieski et al. 2020), wie Atmung und Blutdruck, unterstützen sowie Therapieempfehlungen (Poortmans et al. 2020; Tanguay-Sela et al. 2022) geben. Aber auch bei der personalisierten Edukation von Patient*innen wird KI verwendet (Wolff 2022), um beispielsweise Informationen auf den Krankheitsverlauf einer speziellen Patient*in maßzuschneidern. Obwohl aktuell noch eine Implementierungslücke für den breiten Einsatz von KI-Systemen in der Medizin herrscht (Chen und Asch 2017; Seneviratne et al. 2020), ist davon auszugehen, dass in den nächsten Jahren auch diese geschlossen wird und immer mehr Ärzt*innen und Patient*innen täglich in den Austausch mit einer KI treten.

Künstliche Intelligenz zur Entscheidungsunterstützung kann entweder symbolisch oder als Soft Computing Modell implementiert werden. Symbolische KI basiert auf formalisiertem Expertenwissen, welches als Regeln und Fakten im System abgelegt wird. Hierdurch liefert sie sehr exakte Ergebnisse, ist auf der anderen Seite für die Lösung von komplexen Echtweltproblemen jedoch häufig zu starr (Minsky 1991). Zur Lösung dieser Echtweltprobleme wird implizites Wissen benötigt, welches nicht in expliziten Regeln formuliert werden kann. Solche Probleme können mittels Soft Computing Modellen gelöst werden. Diese datengetriebenen Verfahren erlernen selbstständig aus großen Datenmengen, wie

ein gegebenes Problem gelöst werden kann. So können sie auch für Aufgaben eingesetzt werden, die der Mensch nicht bewältigen kann. Aufgrund ihrer hervorragenden Problemlösungsfähigkeiten und der stetig steigenden Anzahl an potentiellen Trainingsdaten werden im Gesundheitssystem seit einigen Jahren vorwiegend Soft Computing Modellen entwickelt. (Kaul et al. 2020).

Der typische Ablauf der Entwicklung eines Soft Computing Models im Gesundheitswesen (Abb. 1) beginnt mit der Sammlung von zumeist retrospektiven, klinischen Routinedaten. Nach einer gewissen Zeit ist eine kritische Masse an Daten erreicht, die für das Training eines Soft Computing Modells ausreicht. In einer Kooperation von Medizner*innen und Informatiker*innen wird ein Machine oder Deep Learning Modell unter Berücksichtigung des Medizinproduktegesetztes (European Parliament 2017) entwickelt und trainiert. Die Evaluation fokussiert in der Regel die Validität. Hierbei wird überprüft, ob die künstliche Intelligenz sich wie vorgesehen verhält, indem das Modell mit vom Training unabhängigen Daten getestet wird. Wichtig ist, dass ein Teil der Daten bereits zu Beginn vorgehalten wird, welcher nicht für das Training oder das Anpassen von Modellhyperparametern, sondern nur zur Evaluation genutzt wird. Nur so kann sichergestellt werden, dass die entwickelten Modelle auch für neue, bisher ungesehene Daten funktionieren (auch Generalisierungsfähigkeit genannt) und in der Lage sind, das gestellte Problem zu lösen. Bevor das System in der Klinik angewendet werden kann, müssen weitere Aspekte, wie die Nutzerfreundlichkeit des Gesamtsystems oder dessen Funktionalität, überprüft werden. Ein weiterer zu analysierender Aspekt liegt im Mehrwert des Systems. Hier können Kosten-Nutzen-Kalkulationen oder auch der Hauxsche Synenztest (Haux 2021), welcher den Fokus auf die Nützlichkeit des Zusammenwirkens der biologischen ärztlichen Intelligenz und einer KI legt, herangezogen werden. Wenn die Evaluation positiv verläuft, steht dem Einsatz in der Klinik nichts mehr im Weg. Die Stufe des Regelbetriebs erreicht jedoch nur ein Bruchteil der Entwicklungen (Chen und Asch 2017; Seneviratne et al. 2020). Ärzt*innen sehen in ihnen in der Regel ein weiteres Werkzeug, das den klinischen Alltag erleichtern soll. Die Akzeptanz und damit verbunden die Bereitschaft ein solches System einzusetzen, wird maßgeblich durch den Mehrwert für die Ärzt*in bestimmt. Dabei existieren zwei hauptsächliche Möglichkeiten einen Mehrwert in der Klinik zu schaffen. Entweder ermöglicht der Einsatz eines KI-Systems die Einsparung von Arbeitszeit, indem es eine Aufgabe schneller als ein Mensch löst oder sie ihm komplett abnimmt, oder die KI kann ein Problem besser als der Mensch lösen. Letzteres ist der Fall, wenn die Vorhersagen des System eine höhere Qualität als die des Menschen vorweisen (siehe zum Beispiel (Baker et al. 2020)), wie eine höhere Genauigkeit oder geringere Falsch-Positiv Rate, oder die eingesetzte

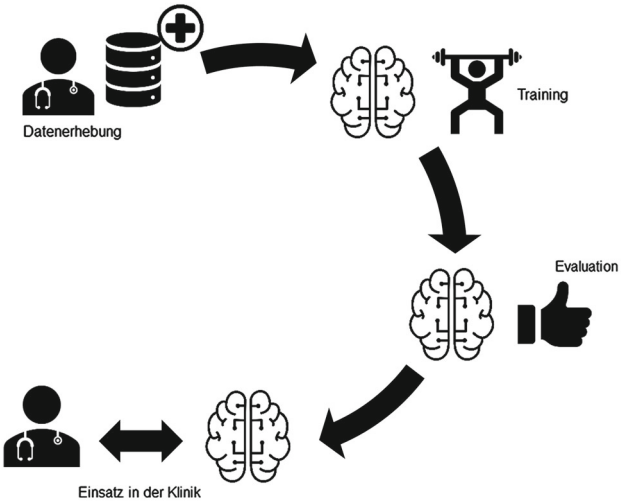

Abb. 1 Der Status quo der Entwicklung von KI-Systemen in der Medizin

Mediziner*in das gegebene Problem, beispielsweise aufgrund fehlender Berufserfahrung oder weil zu viele Daten gleichzeitig berücksichtigt werden müssen, nicht lösen kann, wie es für die Diagnose von seltenen Erkrankungen häufig der Fall ist (Marwaha et al. 2022). In anderen Worten: die KI hat eine Fähigkeit erlernt, über die der Mensch (noch) nicht verfügt.

In diesen Fällen stellt die künstliche Intelligenz mehr als nur ein simples Werkzeug zur Erleichterung der klinischen Tätigkeit dar. Sie bietet einen Mehrwert, der nicht durch den Einsatz von mehr Arbeitskraft gelöst werden kann, und wird so zu einem Experten in dem abgegrenzten Anwendungsfeld. Im zwischenmenschlichen Bereich findet in der Regel ein Wissenstransfer zwischen Kolleg*innen in Form von Diskussionen und Fortbildungen statt. Bezogen auf die künstliche Intelligenz ist ein solcher Wissenstransfer hin zum Menschen wünschenswert, jedoch nicht direkt realisierbar. Die eingesetzten Soft Computing Modelle, wie künstliche neuronale Netze, sind in der Regel keine künstliche allgemeine Intelligenz, die das erlernte Wissen durch eine Sprachkomponente mitteilen könnte. Ganz im Gegenteil handelt es sich hierbei um Black-Boxen, deren interne Mechanismen bei der Lösung eines gestellten Problems auf Grund der Komplexität der Modelle nicht mehr durch den Menschen nachvollzogen werden können. Das Teilgebiet

der erklärbaren künstlichen Intelligenz (englisch: explainable artificial intelligence, XAI) befasst sich unter anderem damit, diese hochkomplexen Modelle für den Menschen nachvollziehbar und interpretierbar zu machen. (Samek et al. 2019) Aktuell wird XAI vornehmlich zur Evaluation von Black-Box-Modellen und zur Vertrauenssteigerung auf Seiten der Ärzt*innen und Patient*innen eingesetzt. Indem Entscheidungsprozesse visualisiert und die indizierenden sowie kontraindizierenden Einflussfaktoren offengelegt werden, soll es dem Menschen ermöglicht werden, den Weg der eigenen Entscheidungsfindung mit dem der künstlichen Intelligenz abzugleichen. Was aber, wenn die KI einen neuen, dem Menschen noch unbekannten und eventuell sogar besseren Weg gefunden hat das Problem zu lösen?

Im folgenden Kapitel wird ein Verfahren beispielhaft am Projekt *Mobile Care Backup* skizziert, um dieses neue Wissen, über das der Mensch zuvor nicht verfügt hat, mittels XAI aus Soft Computing Modellen zu extrahieren und an die menschlichen Domänenexpert*innen (Mediziner*innen) zu transferieren. Das übergeordnete Ziel ist dabei das Schließen der Lücke zwischen trainierten Soft Computing Modellen, die Experten im Lösen einer speziellen Aufgabe geworden sind, und ihren menschlichen Kolleg*innen, die sie in erster Instanz ins Leben gerufen haben.

2 Anwendungsbeispiel: KI-gestützte Patientenedukation im Projekt „Mobile Care Backup“

Niedrige Gesundheitskompetenz der Bevölkerung stellt globale Gesundheitssysteme vor Herausforderungen (Kickbusch 2013; Murray et al. 2008; U.S. Department of Health and Human Services 2014). So hat beispielsweise über die Hälfte der deutschen Bevölkerung erhebliche Schwierigkeiten, die für sie relevanten Informationen zu finden, zu verstehen, zu beurteilen oder anzuwenden (Schaeffer et al. 2016). Eine niedrige Gesundheitskompetenz steht unter anderem in Verbindung mit höheren Hospitalisierungs- (Baker et al. 1998) und Mortalitätsraten (Baker et al. 2007; Sudore et al. 2006). Gleichzeitig stellt die Bevölkerung durch die Pflege von Angehörigen die größte Säule des deutschen Pflegesystems dar und übernimmt damit eine Hauptrolle im deutschen Pflege- und Gesundheitssystem (Statistisches Bundesamt 2020). Die Pflegesituation ist für die meisten Angehörigen mit signifikanten Belastungen verbunden (Mischke 2012), wobei fehlendes Wissen und fehlende Informationen besonders stark zur

Belastung beitragen (Schmall 1995). Der persönliche Wissensbedarf unterscheidet sich dabei individuell aufgrund der vorliegenden Pflegesituation und kann nur begrenzt durch medizinisch Tätige, wie ambulante Pflegedienste, gedeckt werden. Insgesamt wird über die Hälfte der Pflegebedürftigen (2,12 Mio. von 4,1 Mio. im Jahr 2019) durch ihre Angehörigen zu Hause ohne externe Unterstützungen gepflegt (Statistisches Bundesamt 2020).

Ziel des Projektes *Mobile Care Backup (MoCaB)* war die Unterstützung pflegender Angehöriger durch einen mobilen Assistenten. In Anlehnung an automatisierte Lehrsysteme, im speziellen intelligente tutorielle Systeme, welche in Domänen außerhalb des Gesundheitswesens bereits zur Individualisierung eingesetzt werden, besteht das Herzstück des Assistenten aus einer personalisierten Wissensvermittlung (siehe Abb. 2). Dabei tritt der mobile Assistent in einen vorstrukturierten Dialog mit dem pflegenden Angehörigen. Der Angehörige hat die Wahl, vertiefende Informationen zu einem an ihn individualisierten Themen zu erhalten oder weitere für ihn als relevant identifizierte Themen vorschlagen zu lassen. (Rutz et al. 2018) Die Personalisierung durch intelligente tutorielle Systeme wird in der Regel mittels künstlicher Intelligenz und hauptsächlich in stark strukturierten Domänen, wie der universitären oder schulischen Lehre (Chrysafiadi und Virvou 2015), bei denen explizite Lehrpläne vorliegen, implementiert. Diese Systeme basieren häufig auf symbolischer künstlicher Intelligenz, wie einer Ontologie, welche das explizite tutorielle Wissen des Lehrplans umsetzt. Hierzu werden Eigenschaften des Lernenden über Regeln mit den für den Lernenden relevanten Themen verknüpft. Für die Patientenedukation ist das tutorielle Wissen, beispielsweise die Auswahl wichtiger Themen für einen Lernenden, häufig implizit. Obwohl dieses implizite tutorielle Wissen großes Potential bietet, wird es in der Literatur kaum berücksichtigt (Wolff 2022). Implizites Wissen[1] lässt sich nicht in allgemein gültigen Regeln für die Personalisierung der Lehre ausdrücken (Polanyi und Sen 2010). Für einen bestimmten Lernenden kann jedoch eine Personalisierung der Lehre durch an der Entwicklung solcher Systeme beteiligte Experten erfolgen. Zur Erhebung und Nutzung dieses Wissens wird zwischen dem *formalisierbarem* und dem *nicht formalisierbarem* impliziten Wissen unterschieden. Für formalisierbares implizites Wissen, das auch schwaches implizites

[1] Im Englischen wird auch von ‚*tacit knowledge*' (zu deutsch: stillschweigendes Wissen) gesprochen (Polanyi und Sen 2010), was den Kern der Sache (das dieses Wissen nicht ohne Weiteres formalisiert werden kann) deutlich besser trifft. Für den Menschen typisches implizites Wissen findet bei der Erkennung bekannter Gesichter oder beim Fahrradfahren Anwendung. Beide Tätigkeiten erfolgen zu einem hohen Gerade automatisiert, ohne dass allgemeingültige Regeln aufgestellt werden können.

Wissen genannt wird, lassen sich Merkmale und Regeln mit erheblichem kognitivem Aufwand teilweise, jedoch nur unzureichend für eine Abbildung mittels symbolsicher KI benennen. Für nicht formalisierbares implizites Wissen, welches auch als starkes implizites Wissen bezeichnet wird, ist eine Benennung nicht möglich (Neuweg 2000, S. 198–199).

Die Erfassung des formalisierbaren impliziten Expertenwissens in MoCaB basiert in Anlehnung an Vorgehen von Phantombildzeichnern (Polanyi und Sen 2010, S. 4–5, 34) auf der Vorgabe von Kategorien und Beispielen zur strukturierten Wissenserhebung und einer Scoringfunktion zur Bestimmung der Relevanz (Wolff et al. 2018). Hierzu gewichten die Domänenexperten die Items von zwei Assessmenttools, dem *Caregiver Burden Inventory* und dem *Neuen Begutachtungsassessment,* für jedes zu vermittelnde Thema. Die Assessmenttools werden gleichzeitig als Profil des pflegenden Angehörigen verwendet, für den die Themen

Abb. 2 Das Herzstück der MoCaB-App bildet eine dialogbasierte, personalisierte Wissensvermittlung

personalisiert werden. Durch die Verknüpfung des Profils mit den Expertengewichtungen kann über eine Scoringfunktion die Relevanz jedes Themas für einen bestimmten pflegenden Angehörigen bestimmt werden und die Themen ihrer Wichtigkeit nach in einer Reihenfolge angeordnet werden. (Wolff et al. 2018) In einer vierwöchigen Feldstudie mit 18 pflegenden Angehörigen konnte gezeigt werden, dass die so vom MoCaB-System vorgeschlagenen Themen zwar relevant für die Angehörigen sind, jedoch noch weitere, im System enthaltene Themen relevant sind, die nicht vorgeschlagen wurden. (Wolff 2022) Zu den nicht berücksichtigten Themen konnte während der Feldstudie *nicht formalisierbares* implizites tutorielles Wissen erhoben werden.

Während der Feldstudie konnten die Proband*innen jedem gelesenen Thema eine Sternebewertung zuweisen (Abb. 3). Insgesamt wurden 520 Bewertungen von den Probanden erhoben. Die Verknüpfung der Bewertungen mit dem Profil des Lernenden enthält *nicht formalisierbares* implizites Wissen über die Relevanz des Themas für den Lernenden. Um das bestehende System um dieses Wissen zu erweitern, muss es in einem ersten Schritt auf ein Machine Learning Modell übertragen werden, um in einem zweiten Training mit den Sternebewertungen die tutorielle Strategie abzuändern. Als Machine Learning Modell kam ein künstliches neuronales Netz zum Einsatz, welches für das Profil eines pflegenden Angehörigen die Scores für alle 86 Themen des MoCaB-Systems parallel berechnet. Hierzu sind die Neuronen je Thema in einem Strang aus drei Schichten angeordnet. Die Stränge kommunizieren nicht untereinander. (Wolff et al. 2019) Für das zweite Training mit den Sternebewertungen wurde es um eine Learning-to-Rank-Komponente[2] erweitert, sodass die in den Bewertungen enthaltenen Reihenfolgeninformationen in der tutoriellen Strategie, genauer die Reihenfolge der Themen für den Nutzenden und damit die Relevanz der Themen für ihn, berücksichtigt werden kann. Die Evaluation des nachtrainierten Systems zeigt, dass nun auch die zuvor vom System nicht berücksichtigten Themen als relevant identifiziert werden, während immer noch keine unwichtigen Themen

[2] Learning-to-Rank ist eine Sammlung von Verfahren, die originär aus dem Bereich der Online Search Engines stammen, um Machine Learning Modelle anstelle eines Scores oder einer Klassifikation ein Ranking erlernen und anfertigen zu lassen. Im einfachsten Ansatz basiert es darauf, dass in einer Hiddenlayer eines künstlichen neuronalen Netzes für jede zu rankende Resource automatisch ein zuvor nicht bestimmter Score berechnet wird und diese Scores dann paarweise verglichen werden. Der Vergleich basiert mathematisch auf der Subtraktion der beiden Scores und dem Bestimmen des Vorzeichens der Differenz. Ein Vorteil von Learning to Rank ist, dass keine Ranking-Funktion definiert werden muss, sondern diese automatisch aus vorliegenden Reihenfolgen erlernt wird.

vorgeschlagen werden (Wolff 2022). Es konnte demnach auch nicht formalisierbares implizites Wissen erlernt und im System implementiert werden, das den an der Entwicklung beteiligten Experten unbekannt ist. Damit stellt das künstliche neuronale Netz einen deutlichen Mehrwert im Vergleich zum wissensbasierten ersten System dar. Die Ergebnisse der Evaluation zeigen aber auch, dass das Zusammenspiel des impliziten Wissens von Expert*innen sowie Lai*innen, wie Patient*innen, performant in der Entscheidungsunterstützung angewendet werden kann. Auf der anderen Seite ist die Verarbeitung des impliziten Wissens der menschlichen Intelligenzen nur durch eine maschinelle Intelligenz möglich. Das Zusammenwirken von Mensch und Maschine bietet einen Mehrwert, welcher durch das im System implementierte nicht formalisierbare implizite Wissen entsteht. Der Anteil dieses Wissens lässt sich zum Teil an den zuvor fehlenden relevanten Themen ablesen. So besaß das auf formalisierbarem Wissen basierende System eine Sensitivität von 0,71 und eine Genauigkeit von 0,98, während das um das nicht formalisierte implizite Wissen erweiterte System eine Sensitivität von 0,95 und eine Genauigkeit von 0,98 aufweist. Es werden demnach fast 25 % mehr für die Probanden relevante Themen als solche identifiziert, während der Anteil an falsch positiven gleich bleibt. (Wolff 2022).

Um dieses den beteiligten Expert*innen unbekannte Wissen wieder zurück in die Domäne der Pflegewissenschaften zu bringen, muss es aus dem nachtrainierten Machine Learning Modell extrahiert werden. Leider handelt es sich beim eingesetzten künstlichen neuronalen Netz um eine Black-Box, sodass die Extraktion des Wissens über Methoden der erklärbaren KI erfolgt. Von Interesse ist dabei vor allem der Teil des Wissens, welcher nicht durch die Expert*innen ins System gebracht wurde. Um dies umzusetzen, werden die Themen mit den größten Rangänderungen nach dem Training mit Sternebewertungen in der Auslieferungsreihenfolge der Probanden analysiert. Für die so identifizierten Themen wird jeweils der zugehörige Neuronenstrang aus dem nachtrainierten künstlichen neuronalen Netz ohne die Learning-to-Rank-Erweiterung extrahiert. Um die Rangänderung für die Proband*in zu erklären, wird das *Local Interpretable Model-agnostic Explanations (LIME)* Verfahren angewendet. Es basiert auf einer Approximation des vorhergesagten Scores durch ein interpretierbareres lineares Model. Das Model wird trainiert, indem Permutationen des zu erklärenden Datenpunktes durch das originäre Machine Learning Modell bewertet werden und als Trainingsdaten verwendet werden. (Ribeiro et al. 2016) Für den hier dargestellten Anwendungsfall besteht die Erklärung aus dem Einfluss jedes Profilitem des Probanden auf den vorhergesagten Score. Von Interesse sind hierbei vor allem die Profilitems mit der größten Änderung der Wichtigkeit im Vergleich zum nicht nachtrainierten Modell. Daher werden analog zum beschriebenen Vorgehen auch

Abb. 3 Nach dem Lesen eines Themas kann eine Sternebewertung im unteren Bereich der App abgegeben werden

Erklärungen für das nicht mit Sternebewertungen nachtrainierte Netzwerk, welches das Expertenwissen abbildet, erstellt. Über den Vergleich des Einflusses der Profilitems in den beiden tutoriellen Strategien (formalisierbares implizites Wissen sowie die Erweiterung um das nicht formalisierbare Wissen) werden die für die Veränderung in der tutoriellen Strategie ausschlaggebenden Einflussfaktoren extrahiert. So wird für den Menschen begreifbar gemacht, welche (den Expert*innen unbekannte) Attribute die Relevanz eines Themas für den pflegenden Angehörigen bestimmen.

Für einen pflegenden Angehörigen wurde so zum Beispiel festgestellt, dass das im auf rein formalisierbaren Expertenwissen basierenden System unwichtige Thema *Psychische Veränderungen bei Demenz* im mit den Sternebewertungen nachtrainierten System als hochrelevant klassifiziert wird. Der Rang des Themas in der Auslieferungsreihenfolge für diesen bestimmten pflegenden Angehörigen

stieg deutlich vom 86. auf den 8. Rang. Die Analyse der ausschlaggebenden Einflussfaktoren für diese Änderung ergab, dass für diese Entscheidung bestimmend war, dass der pflegende Angehörige nicht genug schläft und sein Pflegeempfänger nachts unruhig ist sowie dass der Pflegeempfänger sich teilweise unangebracht verhält, was zu Verlegenheit bei seinem Angehörigen führt und es unangenehm für den pflegenden Angehörigen ist, wenn Freunde zu Besuch da sind. In Betracht, dass sich eine Demenz und die mit ihr assoziierten Verhaltensänderungen auf den Schlaf auswirken (Schwerthöffer und Förstl 2020; Gust 2018) und als peinlich oder unangenehm empfunden (White 2013, S. 76) werden können, ist es durchweg sinnvoll einem Menschen mit diesen Belastungen Informationen zu dementiellen Veränderungen anzubieten.

In einem zweiten Fallbeispiel aus dem MoCaB Projekt änderte sich die Position des Themas *Ursachen für Schlafprobleme,* welches mit fünf Sternen bewertet wurde, in der Wichtigkeitsreihenfolge für den betrachteten pflegenden Angehörigen vom 29. auf den zweiten Rang. Es wird durch das zweite Training mit dem nicht formalisierbaren impliziten Wissen als deutlich relevanter für diesen pflegenden Angehörigen identifiziert. Während psychosomatische Gründe für schlechten Schlaf, wie ein starker Unterstützungsbedarf des Pflegeempfängers sowie das Fehlen der Fähigkeit, eigene Bedürfnisse und damit den akuten Unterstützungsbedarf formulieren zu können, weniger Einfluss auf die Relevanz des Themas für diesen pflegenden Angehörigen haben als im expertenwissensbasierten System, scheint die Fähigkeit des Pflegeempfängers, sich nachts selbstständig im geteilten Bett bewegen zu können, deutlich ausschlaggebender für die hohe Relevanz des Themas zu sein. Des Weiteren identifizierte das System basierend auf den Bewertungen, dass hier die Angst etwas im Leben zu verpassen (engl. fear of missing out, FOMO) die Relevanz des Themas *Schlafprobleme* erhöht, während Aufgebrachtheit und Ärger über den Pflegeempfänger nicht ausgeprägt waren und daher wenig Einfluss auf die Relevanz des Themas haben. Allerdings fällt bei diesem Fall auch eine Schwachstelle des Verfahrens auf. So wurde ein erhöhter Einfluss auf die Relevanz festgestellt, wenn der Pflegeempfänger Unterstützung beim Waschen des Intimbereichs benötigt. Dieses Profilitem wurde von den Expert*innen für Themen aus dem Bereich der Inkontinenzversorgung sowie differenzierten Themen mit Tipps zum Waschen von Menschen mit Demenz oder nach einem Schlaganfall verwendet. Warum das System hier einen Zusammenhang zwischen dem Waschen des Intimbereichs und Schlafproblemen identifiziert, kann leider nicht weiter bestimmt werden. Spekulationen über den Zusammenhang, wie, dass die Inkontinenz eventuell hauptsächlich nachts auftritt, was den Schlaf des Pflegenden beeinflusst, sind mit den vorliegenden Daten nicht

belastbar. In einem weiteren Interview könnte gezielt nachgefragt werden, dies ist jedoch nicht geplant.

Der letzte Punkt illustriert eine Schwachstelle des Vorgehens. Es ist nicht sichergestellt, dass alle gelernten Einflussfaktoren aus pflegewissenschaftlicher Sicht sinnvoll sind. Hier muss noch eine Evaluation des neu erlernten Wissens mit den an der Entwicklung des Systems beteiligten Expert*innen erfolgen. Es ist geplant das nicht formalisierbare Wissen für alle Proband*innen, wie oben beschrieben, zu erheben, aufzubereiten und den Expert*innen zur Evaluation vorzulegen. Hierbei werden analog zur Evaluation der internen Validität des Systems (Wolff et al. 2018) die stärksten Änderungen des Einflusses auf die Relevanz der Themen für die Probanden den projektinternen Expert*innen vorgelegt und auf ihre Sinnhaftigkeit geprüft. Ein Augenmerk sollte dabei auf Wiedersprüchen zum formalisierbaren Expertenwissen liegen. Die deutliche Steigerung der Sensitivität durch das Training mit den Sternebewertungen spricht allerdings dafür, dass der Großteil des neu erlernten Wissens sinnvoll und zielführend ist. Die Evaluation des erlernten Wissens stellt auch den Transfer des nicht formalisierbaren Wissens in die Domäne der Pflegewissenschaften und somit den Schluss des Kreislaufs in einem lernenden Gesundheitssystem dar (siehe Abb. 4). In einem solchen Gesundheitssystem lernt nicht nur die künstliche Intelligenz vom Menschen und durch ihn erhobene Daten, sondern der Mensch kann auch Wissen von der KI erlernen, welches ansonsten nicht nutzbar wäre. Die so entstehende Kooperation zwischen Mensch und Maschine stellt eine gegenseitige Bereicherung dar und kann die Gesundheitsversorgung verbessern.

Aufgrund der stark unterschiedlichen Zugänglichkeit und Verbalisierbarkeit von implizitem und explizitem Wissen, ist vor allem die direkte Gegenüberstellung dieser beiden Wissensmodi problematisch (Neuweg 2004). Nichtsdestotrotz kann durch das vorgestellte Verfahren eine Näherung dieses Verhältnisses zumindest für das formalisierbare menschliche Expertenwissen und nicht formalisierbare durch die künstliche Intelligenz erlernte Wissen des MoCaB-Systems betrachtet werden. Wird das hier an einem Beispiel illustrierte Verfahren auf alle Proband*innen angewendet, lässt sich das Wissen mittels der Anzahl neu hinzugekommener Einflussfaktoren sowie dem Grad des Einflusses der Faktoren quantifizieren und direkt mit den Einflussfaktoren des Expertenwissens vergleichen. Ebenso ist es möglich, direkte Widersprüche zwischen den beiden Wissensmodi zu identifizieren, die anschließend durch Domänenexpert*innen begutachtet werden. Den Grad der Synergie zwischen Mensch und Maschine an diesem Verhältnis abzuschätzen ist nur bedingt sinnvoll. Einer der beiden Wissensmodi kann deutlich kleiner als der Andere ausfallen, dass Wissen jedoch viel relevanter für die getroffene Entscheidung sein. Der Grad der Synergie lässt sich

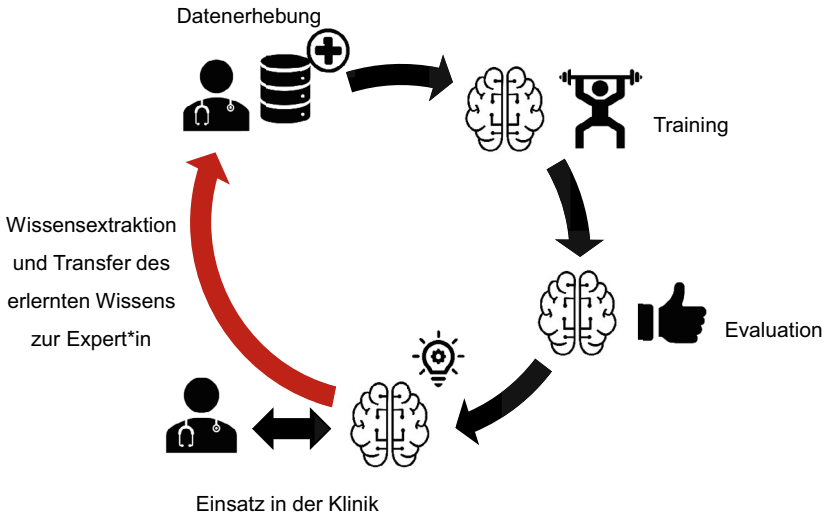

Abb. 4 Durch das beschriebene Vorgehen der Wissensextraktion aus der künstlichen Intelligenz und der Evaluation mit Domänenexpert*innen wird der Kreislauf des Einsatzes von künstlichen Intelligenzen in einem lernenden Gesundheitssystem geschlossen

aber auch an der Steigerung der Inferenzleistung abschätzen. So ist für MoCaB eine deutliche Steigerung der Sensitivität um 0,24 messbar. Die Steigerung sollte sich dabei immer auf den Status quo beziehen. Wenn eine menschliche um eine künstliche Intelligenz erweitert wird, sollte der Vergleich der Inferenzleistung zwischen einer rein auf menschlicher Intelligenz basierenden Entscheidung mit einer Entscheidung der kombinierten Intelligenzen verglichen werden und andersrum. Ein reines Gegenüberstellen der beiden Intelligenzen ist nicht angemessen, da die KI den Menschen nicht ersetzen wird. Eine weitere Möglichkeit die Synergie zu messen besteht in der qualitativen Bestimmung des Mehrwerts durch die Expert*innen. In Interviews kann der Mehrwert der Kooperation mit einer künstlichen Intelligenz bestimmt werden. Die Aufgeschlossenheit der Expert*innen gegenüber dem Einsatz einer künstlichen Intelligenz ist dabei jedoch ein nicht zu vernachlässigender Einflussfaktor. Die qualitativen Interviews können angeschlossen an das Rückspielen des durch die künstliche Intelligenz erhobenen Wissen in die medizinische Domäne erfolgen, sodass sich der Mehraufwand für die Expert*innen in Grenzen hält.

3 Schlussbetrachtung

Die mittels LIME extrahierten Einflussfaktoren sind sehr individuell für die einzelnen Probanden. Auf der einen Seite ermöglicht die kombinierte Betrachtung der Einflussfaktoren und der Profile der zugehörigen Proband*innen eine sehr präzise Aussage über die individuelle Situation und den individuellen Informationsbedarf. Auf der anderen Seite kann so jedoch nicht sichergestellt werden, dass das nicht formalisierbare implizite Wissen generalisierbar auf andere pflegende Angehörige ist. Um globalere Aussagen zu treffen, können andere Methodiken der erklärbaren künstlichen Intelligenz verwendet werden. Eine Möglichkeit globale Erklärungen für das eingesetzte neuronale Netz anzufertigen, besteht im Einsatz der *Shapley Additive exPlanations (SHAP)* (Lundberg und Lee 2017). Hierbei handelt es sich um ein Verfahren aus der Spieltheorie, welches genutzt werden kann um allgemeingültigere, kumulierte Erklärungen anzufertigen. Ob diese globaleren Erklärungen für die hoch individualisierten Themenempfehlungen des MoCaB-Systems sinnvoll sind, muss noch geklärt werden. In anderen Anwendungsgebieten, in denen eine KI ein Problem besser als der Mensch lösen kann, ist die Anwendung globaler Erklärungen hilfreich. Bei der automatisierten Gehirnreifebestimmung bei Kleinkindern kann so beispielsweise sichergestellt werden, dass das Modell sinnvolle Hirnregionen in seine Entscheidung einbezieht und nicht der Schädelumfang der ausschlaggebende Faktor ist. In einem zweiten Schritt kann dann die individuelle Erklärung untersucht werden und so im Sinne der personalisierten Medizin eine nachvollziehbare Entscheidung für jeden einzelnen Patienten getroffen werden. Zur Schaffung von Synergien aus menschlichen und künstlichen Intelligenzen scheint die Kombination lokaler und globaler Erklärungen vielversprechend.

Für MoCaB konnte durch das Zusammenspiel von formalisierbarem und nicht formalisierbarem Wissen, welches nur durch eine Kombination menschlicher mit künstlicher Intelligenz erhoben werden konnte, die Personalisierungsleistung des Systems weiter gesteigert werden. Durch die Kombination können die Themen noch besser für pflegende Angehörige personalisiert werden und mehr relevante Themen berücksichtigt werden. Es wurde gezeigt, dass erstens das Zusammenwirken von Expert*innen und Lai*innen (Patient*innen und pflegenden Angehörigen) zielführend ist und zweitens das Zusammenwirken von Mensch und KI einen deutlichen Mehrwert bietet.

Das hier vorgestellte Verfahren zum Schließen des Kreislaufs in einem lernenden Gesundheitswesen stellt einen wichtigen nächsten Schritt in der Entwicklung von künstlichen Intelligenzen in der Medizin dar. Es ermöglicht neben den Synergien, die bei der Verwendung von künstlichen Intelligenzen als Werkzeug

entstehen, eine Zusammenarbeit auf Augenhöhe zwischen menschlichen und künstlichen Intelligenzen. Der mit der Implementierung dieser Schritte einhergehende Mehraufwand in der Entwicklung und Etablierung von KI wirkt sich negativ auf die Entwicklungszeit aus. Auf der anderen Seite ist zu erwarten, dass die Akzeptanz und das Vertrauen des Menschen in künstliche Intelligenzen deutlich gesteigert wird und so die bestehende Implementierungslücke im Gesundheitswesen weiter geschlossen werden kann. Die Synergie künstlicher und biologischer Intelligenzen wird in Zukunft zur Verbesserung der Versorgungsqualität beitragen.

Es bleiben offene Fragen beim Zurückführen des erlernten, dem Menschen unbekannten Wissens bestehen. In der Literatur wird implizites Wissen schon lange als Innovationstreiber angesehen (Kogut und Zander 1992) und der Transfer zwischen Personen erforscht (Stover 2004; Al-Qdah und Salim 2013; Cowan 2000). Hierfür wird das Wissen häufig formalisiert. Die Formalisierung impliziten Wissens ist jedoch verlustbehaftet und nicht alles implizite Wissen kann formalisiert werden (Grant 1996, S. 116). Da die Extraktion des impliziten Wissens aus der künstlichen Intelligenz mittels XAI auch eine Art Formalisierung darstellt, ist davon auszugehen, dass auch der Transfer von Wissen zwischen einer künstlichen und menschlichen Intelligenz einem Verlust unterliegt. Es ist des Weiteren noch festzustellen, ob die Aufbereitung des Wissens mittels XAI für den Menschen gut verständlich ist. So ist eine *gute Erklärung* Personen und Kontext spezifisch. In der Regel werden die Eingabedaten für die Erklärung herangezogen, sodass die Erklärung einer bildbasierten KI häufig eine grafische Repräsentation auf dem verwendeten Bild, beispielsweise in Form einer Heatmap ist. Wenn zwei Expert*innen zusammenarbeiten, geht die Kommunikation aber, über eine reines mit dem Finger auf Besonderheiten des Bildes zeigen, hinaus. Die Erklärung der Entscheidungsfindung wird in der Regel mit Semantik linguistisch umschlossen. Mit dem aktuellen Trend der Large Language Models ist es aber denkbar, dass KI in Zukunft sich auch natürlicher Sprache bedienen wird, um seine Entscheidungsfindung durch Erklärungen zu untermauern.

Literatur

Al-Qdah MS, Salim J. Managing Tacit Knowledge in MNCS and the Role of ICT: Review Paper. RJASET. 2013;6:4110–4120.

Baker DW, Parker RM, Williams MV, et al. Health literacy and the risk of hospital admission. Journal of General Internal Medicine. 1998;13:791–798.

Baker DW, Wolf MS, Feinglass J, et al. Health literacy and mortality among elderly persons. Arch Intern Med. 2007;167:1503–1509.
Baker A, Perov Y, Middleton K, et al. A Comparison of Artificial Intelligence and Human Doctors for the Purpose of Triage and Diagnosis. Front Artif Intell. 2020;3:543405.
Bizopoulos P, Koutsouris D. Deep Learning in Cardiology. IEEE Rev Biomed Eng. 2019;12:168–193.
Chen JH, Asch SM. Machine Learning and Prediction in Medicine – Beyond the Peak of Inflated Expectations. N Engl J Med. 2017;376:2507–2509.
Chrysafiadi K, Virvou M. Fuzzy Logic for Adaptive Instruction in an E-learning Environment for Computer Programming. IEEE Transactions on Fuzzy Systems. 2015;23:164–177.
Cowan R. The explicit economics of knowledge codification and tacitness. Industrial and Corporate Change. 2000;9:211–253.
Davies G, Valentine T. Facial composites: Forensic utility and psychological research. In: The handbook of eyewitness psychology, Vol II: Memory for people; 2007; p. 59–83.
European Parliament, Council of the European Union. REGULATION (EU) 2017/745 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC [Internet]. 2017. Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32017R0745.
Graham B, Bond R, Quinn M, et al. Using Data Mining to Predict Hospital Admissions From the Emergency Department. IEEE Access. 2018;6:10458–10469.
Grant RM. Toward a knowledge-based theory of the firm. Strat. Mgmt. J. 1996;17:109–122.
Gust J. Schlaf und Demenz. GGP. 2018;02:82–85.
Haux R. Lässt sich erweitertes Zusammenwirken zwischen Menschen und intelligenten Maschinen empirisch feststellen?: Überlegungen zur Adaptierung des Turing-Tests. 2021. https://doi.org/10.24355/dbbs.084-202109030943-0.
Kaieski N, da Costa CA, da Rosa Righi R, et al. Application of artificial intelligence methods in vital signs analysis of hospitalized patients: A systematic literature review. Applied Soft Computing. 2020;96:106612.
Kaul V, Enslin S, Gross SA. History of artificial intelligence in medicine. Gastrointest Endosc. 2020;92:807–812.
Kickbusch I. Health Literacy. The Solid Facts. Geneva: World Health Organization; 2013.
Kniep HC, Sporns PB, Broocks G, et al. Posterior circulation stroke: machine learning-based detection of early ischemic changes in acute non-contrast CT scans. J Neurol. 2020;267:2632–2641.
Kogut B, Zander U. Knowledge of the Firm, Combinative Capabilities, and the Replication of Technology. Organization Science. 1992;3:383–397.
Lee EK, Yuan F, Hirsh DA, et al. A clinical decision tool for predicting patient care characteristics: patients returning within 72 hours in the emergency department. AMIA Annu Symp Proc. 2012;2012:495–504.
Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. In: Guyon I, Luxburg UV, Bengio S, et al., editors Advances in Neural Information Processing Systems; 2017.

Marwaha S, Knowles JW, Ashley EA. A guide for the diagnosis of rare and undiagnosed disease: beyond the exome. Genome Med. 2022;14:23.
Minsky, ML. Logical Versus Analogical or Symbolic Versus Connectionist or Neat Versus Scruffy. AI Magazine, 1991;12(2):34. https://doi.org/10.1609/aimag.v12i2.894
Mischke C. Ressourcen von pflegenden Angehörigen: Entwicklung und Testung eines Assessmentinstruments. Hungen: Hpsmedia; 2012.
Mollura M, Lehman L-WH, Mark RG, et al. A novel artificial intelligence based intensive care unit monitoring system: using physiological waveforms to identify sepsis. Philos Trans A Math Phys Eng Sci. 2021;379:20200252.
Murray TS, Hagey J, Willms D, Shillington R und Desjardins R. Health literacy in Canada: a healthy understanding. 2008, Ottawa, Canada. (Hrsg.) Canadian Council on Learning. ISBN 978-0-9809042-1-5.
Neuweg GH. Mehr lernen, als man sagen kann: Konzepte und didaktische Perspektiven impliziten Lernens: Learning more than one can tell: Concepts and didactical perspectives of implicit learning. Weinheim: Beltz Juventa; 2000.
Neuweg GH. Tacit knowing and implicit learning. European perspectives on learning at work: The acquisition of work process knowledge. 2004:130–147.
Polanyi M, Sen A. The tacit dimension. Chicago, Ill.: Univ. of Chicago Press; 2010.
Poortmans PMP, Takanen S, Marta GN, et al. Winter is over: The use of Artificial Intelligence to individualise radiation therapy for breast cancer. Breast. 2020;49:194–200.
Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning [Internet]. [place unknown]; 14.11.2017. Available from: http://arxiv.org/pdf/1711.05225v3.
Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?". In: Krishnapuram B, Shah M, Smola A, et al., editors Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; San Francisco California USA; 2016; p. 1135–1144.
Rutz M, Dierks M-L, Behrends M, et al. Hallo Du, ich bin Mo-Der Dialog als personalisierte Form der Wissensvermittlung in einem mobilen Assistenzsystem. Zukunft der Pflege Tagungsband der 1. Clusterkonferenz 2018. 2018:84–88.
Samek W, Montavon G, Vedaldi A, et al., editors. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Cham: Springer; 2019. (Springer eBooks Computer Science; vol. 11700).
Schaeffer D, Vogt D, Berens E-M, et al. Gesundheitskompetenz der Bevölkerung in Deutschland: Ergebnisbericht. Bielefeld: Universität Bielefeld, Fakultät für Gesundheitswissenschaften; 2016.
Schmall VL. Family caregiver education and training: enhancing self-efficacy. J Case Manag. 1995;4:156–162.
Schwerthöffer D, Förstl H. Schlaf-Wach-Rhythmusstörungen bei demenziellen Erkrankungen. DNP. 2020;21:18–22.
Seneviratne MG, Shah NH, Chu L. Bridging the implementation gap of machine learning in healthcare. BMJ Innov. 2020;6:45–47.
Statistisches Bundesamt. Pflegestatistik: Pflege im Rahmen der Pflegeversicherung Deutschlandergebnisse [Internet]. Wiesbaden; 15. Dezember 2020. (no. 5224001199004) [cited 2021 Apr 20]. Available from: https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Gesundheit/Pflege/Publikationen/Downloads-Pflege/pflege-deutschlandergebnisse-

5224001199004.pdf;jsessionid=C9D1C1CF375609F73F81388F9CA99C1F.live721?__blob=publicationFile.
Stover M. Making tacit knowledge explicit: the Ready Reference Database as codified knowledge. Reference Services Review. 2004;32:164–173.
Sudore RL, Yaffe K, Satterfield S, et al. Limited literacy and mortality in the elderly: the health, aging, and body composition study. Journal of General Internal Medicine. 2006;21:806–812.
Tanguay-Sela M, Benrimoh D, Popescu C, et al. Evaluating the perceived utility of an artificial intelligence-powered clinical decision support system for depression treatment using a simulation center. Psychiatry Res. 2022;308:114336.
U.S. Department of Health and Human Services. America's health literacy; Why we need accessible health information. [Internet]. Washington, D.C.; 2014 [cited 2019 May 13]. Available from: http://www.health.gov/communication/literacy/issuebrief/.
White E. Sexualität bei Menschen mit Demenz. (Altenpflege Demenz). Bern: Verlag Hans Huber; 2013.
Wolff D. Implizites tutorielles Wissen in der KI-gestützten automatisierten Patientenedukation am Beispiel pflegender Angehöriger. [place unknown]: Universitätsbibliothek Braunschweig; 2022.
Wolff D, Behrends M, Gerlach M, et al. Personalized Knowledge Transfer for Caregiving Relatives. Stud Health Technol Inform. 2018;247:780–784.
Wolff D, Kupka T, Marschollek M. Extending a Knowledge-Based System with Learning Capacity. Stud Health Technol Inform. 2019;267:150–155.

Dr. Dominik Wolff Peter L. Reichertz Institut für Medizinische Informatik der TU Braunschweig und der Medizinischen Hochschule Hannover (PLRI). Dominik Wolff ist Nachwuchsgruppenleiter im PLRI. Nach dem Bachelorabschluss in Bioinformatik und Masterabschluss in Naturwissenschaftlicher Informatik an der Universität Bielefeld wechselte er 2016 an das Peter L. Reichertz Institut für Medizinische Informatik. Seine Forschungsschwerpunkte liegen auf Anwendungen künstlicher Intelligenz und datenwissenschaftlicher Methoden in der Biomedizin sowie der Mensch-Maschine-Interaktion und der Nutzung impliziten Wissens in Tutorialsystemen zur Patientenaufklärung.

Die Schwierigkeiten der Bewertung des erweiterten Zusammenwirkens von natürlicher und künstlicher Intelligenz

Tim Kacprowski

Zusammenfassung

Der Einsatz von künstlicher Intelligenz (KI) und das ggf. erweiterte Zusammenwirken zwischen natürlicher und künstlicher Intelligenz wird aktuell häufig anhand eindimensionaler Performanzmaße bewertet und blendet viele ethische und andere Dimensionen aus. Ebenso bedenkenswert gehen diese Performanzmaße häufig auf einen Wettstreit zwischen natürlicher und künstlicher Intelligenz zurück. Für eine nachhaltige und umfassende Bewertung müssen deutlich mehr Dimensionen bedacht werden, wie in den Beiträgen zu Gamification in Public Helath und KI in der Medizin exemplarisch gezeigt wird. Außerdem muss ein Weg gefunden werden aus dem Wettstreit auszubrechen und die Bewertung mehr auf das Zusammenwirken zu fokussieren.

Schlüsselwörter

Zusammenwirken • Bewerten • Künstliche Intelligenz • Wettstreit • Gamification • Klinische Entscheidungsunterstützungssysteme

Künstliche Intelligenz (KI) wird mehr und mehr genutzt um Prozesse effizienter und Ergebnisse besser zu machen. KI ist in nahezu allen wissenschaftlichen

T. Kacprowski (✉)
Abteilung Data Science in Biomedicine, Peter L. Reichertz Institut für Medizinische Informatik der TU Braunschweig und der Medizinischen Hochschule Hannover, Braunschweig, Deutschland
E-Mail: t.kacprowski@tu-braunschweig.de

O. Richter et al. (Hrsg.), *Zusammenwirken von natürlicher und künstlicher Intelligenz: Beurteilen-Messen-Bewerten,*
https://doi.org/10.1007/978-3-658-45845-4_9

Disziplinen vertreten und auch im Alltag immer gegenwärtiger. Ob sich der Einsatz von KI und das ggf. erweiterte Zusammenwirken künstlicher und natürlicher Intelligenz lohnt, wird dabei meist anhand recht technischer Performanzmaße bewertet. Nicht nur, dass diese sämtliche ethische und andere nicht-technische Aspekte außen vor- und damit nur eine sehr eindimensionale Bewertung zulassen. Die Performanz wird auch stets in einer Art Wettstreit zwischen natürlicher und künstlicher Intelligenz gemessen. Kann eine KI Krankheiten besser diagnostizieren als ein Arzt? Wird eine Entscheidung robuster, wenn wir die Nutzung von KI zu ihrer Ableitung zulassen? Stets wird die bisherige Performanz der natürlichen Intelligenz als *baseline* missbraucht und muss sich gegen die KI behaupten.

Um aus diesem Wettstreit auszubrechen, lohnt es sich, das Zusammenwirken von natürlicher und künstlicher Intelligenz umfassender zu untersuchen und zu bewerten. Wie in den Beiträgen zu diesem Teil des Symposiums deutlich wird, lässt sich so dann eine Vielzahl essentieller Fragen aufstellen und untersuchen, die für eine notwendige mehrdimensionale Bewertung des Einsatzes von, eigentlich jedweder Werkzeuge im Allgemeinen, aber auch der KI im Besonderen unerlässlich sind.

Der erste Beitrag beschäftigt sich mit Gamification in Public Health, also mit der Verwendung von Spielmechaniken und Elementen im Bereich der öffentlichen Gesundheit. Diese Spielmechaniken und Elemente werden verwendet um interessante und motivierende Erfahrungen zu schaffen. Mittlerweile etabliert ist das bereits im Fitness Bereich, wo Fitness-Apps Punkte, Rewards und Leaderboards nutzen, um Nutzer dazu zu motivieren, regelmäßig zu trainieren und sich fit zu halten. Dies lässt sich natürlich auch auf die Förderung anderer gesunder Verhaltensweisen übertragen um effizienter Krankheiten vorzubeugen. Bzgl. der Bewertung derartiger Maßnahmen stehen wir nun nicht nur vor der Herausforderung die sich bei Präventivmaßnahmen ohnehin ergibt: Hat es geholfen? War es wirklich nötig? Es ist doch nichts passiert. Wir müssen uns außerdem darüber Gedanken machen, ob und in welchem Maße wir Verhaltensbeeinflussung positiv oder negativ bewerten. Insbesondere, wenn diese durch Entscheidungen von oder mit KI legitimiert wird, sei diese auch Teil eines erweiterten Zusammenwirkens mit einer natürlichen Intelligenz, gerade aber wenn sie, wie eben derzeit hauptsächlich, aus einem Wettstreit gegen eine natürliche Intelligenz hervorgegangen ist. Eine gewissenhafte Bewertung ist für die Schaffung von Rahmenbedingungen essentiell, da Gamification, ähnlich wie KI, in vielen Bereichen inklusive Werbung und Bildung auf dem Vormarsch ist. Das ist nicht *per se* schlecht, muss aber ethisch vertretbar erfolgen und sollte für die Gesellschaft vorteilhaft sein.

In der Medizin sind die ethischen Implikationen des Einsatzes von KI ebenso vorhanden, wenn nicht noch vielfältiger, wie im zweiten Beitrag verdeutlicht wird. Stetige Reibungspunkte umfassen Datenschutz, Transparenz, Verantwortlichkeit oder Fairness bei der Entscheidungsfindung und Weiteres. Diese Themen müssen ernst genommen werden und in die Bewertung von KI und ihrem erweitertem Zusammenwirken mit natürlichen Intelligenzen in der Medizin einfließen, um eine nachhaltige Zukunft für alle Beteiligten zu gewährleisten. Dies betrifft unter anderem klinische Entscheidungsunterstützungssysteme also Entscheidungshilfen die auf algorithmischer Datenverarbeitung, häufig KI, basieren. Das Ziel solcher Systeme ist explizit nicht Ärzt*innen zu ersetzen, sondern diese bei ihren Entscheidungsfindungen zu unterstützten. Damit sind diese KI Systeme eigentlich prädistiniert für ein erweitertes Zusammenwirken mit natürlichen Intelligenzen. Sie sind jedoch auch besonders empfindlich gegenüber vielen Risiken von KI wie z. B. *bias, overfitting*, etc. Sowohl die Fachkompetenz der natürlichen Intelligenz, als auch das Herangehen an den Umgang und die Interaktion mit der KI, hier sowohl seitens der Ärztin/des Arztes als auch der Patientin/des Patienten, sind von immenser Bedeutung.

Es zeigt sich also, dass die Bewertung des Einsatzes von KI und des erweiterten Zusammenwirkens zwischen KI und natürlicher Intelligenz umfassender und mehrdimensionaler gedacht werden muss, als es heute oft erfolgt. So muss unter anderem das Berufsethos der – bis dato – natürlichen Intelligenzen auf der einen Seite des Zusammenwirkens berücksichtigt werden. Medizinische Berufsethiken betonen traditionell Werte wie Patientenwohl, Nichtschaden, Autonomie und Gerechtigkeit. Der Einsatz von KI in der Medizin muss diese Werte respektieren und unterstützen. Gamification-Strategien in der öffentlichen Gesundheit sollten ebenfalls dem Wohle der Gemeinschaft dienen und individuelle Rechte und Freiheiten nicht untergraben. Offensichtlich können sowohl KI als auch Gamification tiefgreifende Auswirkungen auf die Lebensweise der Menschen haben. Es ist wichtig, dass diese Technologien die Diversität von Lebensformen respektieren und nicht ungewollt zu einer Homogenisierung oder Marginalisierung bestimmter Gruppen führen. Mit der fortschreitenden Technologieentwicklung kann es notwendig werden, traditionelle ethische Normen und Wertvorstellungen zu überdenken und anzupassen. Dies betrifft etwa Fragen der Datenprivatheit, des informierten Einverständnisses und der Verantwortlichkeit im Umgang mit KI-Systemen.

Immanuel Kant ist bekannt für seinen Fokus auf Vernunft, Autonomie und die moralische Pflicht. Seine Ideen, insbesondere aus Werken wie der „Kritik der reinen Vernunft“, „Kritik der praktischen Vernunft“ und „Grundlegung zur Metaphysik der Sitten“ können sicherlich interessante Perspektiven auf diese ethischen

Fragen bieten. So würde er wahrscheinlich betonen, dass Ärzt*innen ihre Pflicht haben, zum Wohl der Patienten zu handeln, und nicht blindlings einer Technologie folgen sollten. Dies entspricht seinem Prinzip, nach dem moralisches Handeln aus Pflicht und nicht aus Neigung erfolgen soll („Kritik der praktischen Vernunft"). Im Übrigen ein Prinzip das zwar einerseits natürliche Intelligenzen gegen Gamification wappnen sollte, andererseits aber auch klare Ansprüche an Gamification zu stellen vermag. Verwendete Spielmechaniken müssten so entworfen sein, dass sie intrinsische Motivation fördern und zur Reflexion über moralische Grundlagen eigener Entscheidungen anregen. Wie Kant klinische Entscheidungsunterstützungssysteme und Gamification in Public Health bewerten würde, soll an dieser Stelle offenbleiben und zu einer eigenen mehrdimensionalen Bewertung motivieren.

Abschließend noch der Gedanke, ob eine Möglichkeit aus dem Wettstreit zwischen natürlicher und künstlicher Intelligenz auszubrechen nicht auch in einer umfassenderen mehrdimensionalen Ausbildung liegt. So sollte, gerade zur aktuellen Zeit in der die Zahl der Menschen die mit KI interagieren stetig und rasant wächst, in der universitären Bildung aber auch darüber hinaus, mehr Gewicht auf die Vermittlung von Wissen und Kompetenzen gelegt werden, die eine ethische und andere Dimensionen umfassende Bewertung des Umgangs und der Interaktion mit KI ermöglichen. Ganz im Sinne einer Hundeschule für KI, denn in einer Hundeschule lernt auch weniger der Hund wie er sich zu verhalten hat, als vielmehr der/die Halter*in, wie er/sie mit dem Hund umgehen kann und auf ein **erweitertes** Zusammenwirken hinarbeiten kann.

Prof. Dr. Tim Kacprowski Abteilung Data Science in Biomedicine, Peter L. Reichertz Institut für Medizinische Informatik der TU Braunschweig und der Medizinischen Hochschule Hannover (PLRI). Tim Kacprowski leitet seit 2020 die Abteilung Data Science in Biomedicine des PLRI an der TU Braunschweig. Seine Forschung konzentriert sich auf die Kombination von Netzwerkbiologie und Machine Learning um Verfahren zur Auswertung biomedizinischer Daten zu entwickeln. Ferner beschäftigt er sich mit den Bereichen Immersive Analytics und Science of Science. Er leitet derzeit die Arbeitsgruppe "Statistische Methoden der Bioinformatik" der GMDS e.V.

Gamification in Public Health: The Dark, Bright and Grey Side

Barbara Buchberger

Zusammenfassung

Gamification ist eine auf Informationstechnologie beruhende Zusatzdienstleistung, die aus Spiel-Design-Elementen besteht und darauf zielt, die Motivation, Produktivität und Verhaltensweisen von Nutzern positiv zu beeinflussen. Public Health ist die Wissenschaft und Praxis der Verhinderung von Krankheiten und Verlängerung des Lebens, verfolgt aber auch das Ziel, Verhaltensweisen von Menschen zur Förderung der Gesundheit zu ändern. Aufgrund der generellen Zunahme von Computertechnologien, die durch die COVID-19-Pandemie zusätzlich befördert wurde, lohnt eine erneute Betrachtung ethischer Implikationen dieser überwiegend positiv bewerteten und seit mehr als 10 Jahren genutzten Möglichkeit zur Verhaltensänderung. Im Beitrag werden der potentielle Nutzen und Schaden von Gamification für Public Health betrachtet sowie Grenzbereiche für den Einsatz von Spiel-Design-Elementen am Rand von Manipulation und Nötigung ausgelotet.

Schlüsselwörter

Gamification • Public Health • Verhaltensänderung • Persuasive Technologie • Manipulation

B. Buchberger (✉)
Bundesinstitut im Geschäftsbereich des Bundesministeriums für Gesundheit, Robert Koch-Institut, Berlin, Deutschland
E-Mail: barbara.buchberger@googlemail.com

O. Richter et al. (Hrsg.), *Zusammenwirken von natürlicher und künstlicher Intelligenz: Beurteilen-Messen-Bewerten,*
https://doi.org/10.1007/978-3-658-45845-4_10

1 Einleitung

Computertechnologien sind heutzutage allgegenwärtig und Teil unserer privaten und beruflichen Umgebung. Objekte wie eine smarte Zahnbürste, die den Benutzer nicht nur auf sein möglicherweise nachlässiges Zahnputzverhalten hinweist, sondern auch im Sinn des Herstellers darauf aufmerksam macht, wann ein neuer Bürstenkopf angeschafft werden muss, oder mobile Applikationen, mit denen ein Smartphone-Besitzer seine Schritte zählen lassen oder in Verbindung mit einer Smartwatch weitere Gesundheitsdaten zur Messung seiner körperlichen Fitness erfassen kann, sind weit verbreitet.

Mobile Konnektivität, Cloud Computing, soziale Netzwerke, Big Data, maschinelles Lernen und ähnliche Technologien prägen mittlerweile auch Ausbildung, Studium und Arbeitswelt. Sie sind Voraussetzung für Gamification, worunter die Anwendung von Spiel-Design-Elementen in nicht spielerischem Kontext zu verstehen ist. So kann beispielsweise in der Physiotherapie Gamification für die Behandlung von Schlaganfall-Patienten eingesetzt werden, um die Monotonie hoch repetitiver Trainingseinheiten durch Einbettung in ein Videospiel vergessen zu lassen. Häufig sind auch digitale Gesundheitsanwendungen, sogenannte DiGA oder Gesundheits-Apps, mit Spiel-Design-Elementen ausgestattet, die als Medizinprodukte einer niedrigen Risikoklasse seit dem Jahr 2020 von Ärzten und Psychotherapeuten verordnet werden können. Nutzen und Schaden von DiGA stehen allerdings wegen oftmals schwacher Evidenz in der Kritik. Da Gamification im Alltag auch erfolgreich zur Steuerung des Konsumverhaltens durch Werbung und Marketing eingesetzt wird und als persuasive Technologie gezielt kognitive Verzerrungen nutzt, lohnt aufgrund der zunehmenden Verbreitung eine erneute Betrachtung ethischer Implikationen dieser überwiegend positiv bewerteten und seit mehr als 10 Jahren genutzten Möglichkeit zur Verhaltensänderung.

Im Beitrag werden die Zusammenhänge von Public Health und Gamification erläutert sowie Grenzbereiche für den Einsatz von Spiel-Design-Elementen am Rand von Manipulation, Betrug, Infantilisierung und Trivialisierung ausgelotet.

2 Public Health

Public Health ist nach einer weit verbreiteten und normativ neutral gehaltenen Definition „die Wissenschaft und Praxis der Verhinderung von Krankheiten, Verlängerung des Lebens und Förderung der Gesundheit durch organisierte Anstrengungen der Gesellschaft“ (Acheson 1988; Verweij und Dawson 2007).

Gegenstand von Public Health ist nicht die Gesundheit des Einzelnen, sondern die Gesundheit einer Bevölkerung oder von Gruppen (Verweij und Dawson 2007). Unter Betonung der Gesundheitsförderung und Prävention von Krankheit und Behinderung sind die Erhebung und Nutzung epidemiologischer Daten zur Beobachtung und Überwachung der Bevölkerung und andere Formen der empirischen quantitativen Bewertung charakteristisch für Public Health, wie auch die Anerkennung der multidimensionalen Natur der Determinanten von Gesundheit; für die Entwicklung wirksamer Interventionen sind daher komplexe Wechselwirkungen vieler Faktoren zu berücksichtigen, die von biologischer, verhaltensbezogener, sozialer und ökologischer Art sein können (Childress et al. 2002). Da Interventionen im Kontext von Public Health auf Populationsebene stattfinden und viele Menschen erreichen, kann der Nutzen für den Einzelnen oft nicht klar bestimmt werden. Dieses Merkmal der öffentlichen Gesundheit als gemeinsames Gut, zu dem eine unbestimmte Anzahl nicht zuordenbarer Individuen beiträgt, erschweren Verständnis, Adhärenz und Akzeptanz für Maßnahmen wie beispielsweise Impfungen. Dasselbe gilt für eine effektive Primärprävention, die dazu führt, dass Ereignisse gar nicht erst eintreten (Verweij und Dawson 2007).

Zur Prävention chronischer Krankheiten wie Krebs, Herz-Kreislauf-Erkrankungen oder Typ-2-Diabetes, die weltweit Gesundheitssysteme belasten und Menschen aller Einkommensklassen betreffen, sind Veränderungen des Lebensstils erforderlich. Die größten Risikofaktoren für die Entstehung chronischer Krankheiten und vorzeitigen Tod sind Alkohol- und Tabakkonsum, eine ungesunde Ernährung und mangelnde körperliche Aktivität (WHO 2009).

Ein wichtiger Faktor für Verhaltensänderungen ist Motivation, die über eine längere Zeit anhalten und aufrechterhalten werden muss, wenn angestrebte Ziele zur Verbesserung der Gesundheit erreicht werden sollen; persuasive Technologien können dabei unterstützend wirken (Johnson et al. 2016).

3 Persuasive Technologien

Persuasive Technologien wurden als interaktive Computersysteme mit dem Ziel entwickelt, Haltung und Verhalten von Menschen zu verändern (Blohm und Leimeister 2013; Fogg 2003). Bereits 1996 prägte der Sozialwissenschaftler Brian Jeffrey Fogg im Rahmen seiner Forschung an der Universität Stanford den Begriff *Captology,* abgeleitet vom Akronym CAPT für *Computer as Persuasive Technologies* (Fogg 1998). Damit bezeichnete er den Bereich, in dem sich Computertechnologie in Form von Websites, Mobiltelefonen oder Fitness-Ausrüstung

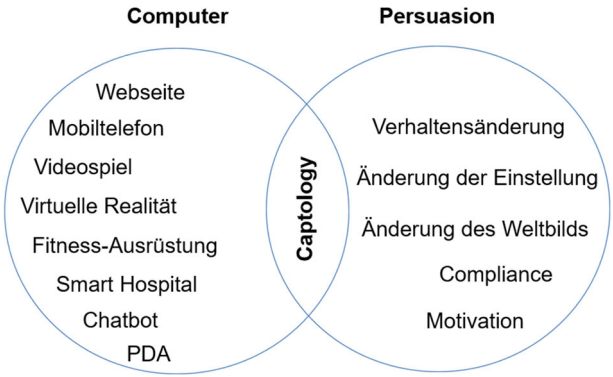

Abb. 1 Modifiziert nach Fogg (1998). PDA: Persönlicher digitaler Assistent

und Persuasion überschneiden; Persuasion definierte er als einen „Versuch, Verhaltensweisen, Gefühle oder Gedanken zu einem Thema, einem Gegenstand oder einer Handlung zu formen, zu verstärken oder zu ändern“ (siehe Abb. 1, Fogg 1998).

Computertechnologien haben die Rollen verschiedener gesellschaftlicher Akteure okkupiert, unter deren Einfluss Menschen traditionell stehen, wie zum Beispiel dem von Lehrern, Geistlichen, Ärzten, Trainern, Therapeuten oder Verkäufern (Fogg 2003). So nehmen Onlineversandhändler nicht nur Bestellungen entgegen, sondern versuchen, Menschen vom Kauf vieler anderer Produkte durch Vorschläge zu überzeugen, die auf der Basis von Präferenzen infolge früherer Bestellungen und durch Empfehlungen anderer Kunden, die das Produkt bereits gekauft haben, generiert wurden.

Ein Vorteil der Computertechnologien gegenüber traditionellen Medien zur Beeinflussung von Verhalten und Einstellungen wie dem Rundfunk oder Fernsehen ist die Möglichkeit der Interaktion, weil sie zu individuellen Anpassungen genutzt werden kann. Ein Rauch-Entwöhnungsprogramm lässt sich zum Beispiel auf die Gewohnheiten und das Tempo eines Teilnehmers zuschneiden (Palmer et al. 2018). Vergleichbares gilt für Werbung, Marketing und Verkauf. Auch Anonymität und online-Formate von Veranstaltungen können vorteilhaft sein, weil sie mit ihrer im Vergleich zur Präsenz niedrigeren Hemmschwelle zurückhaltenden Menschen eine Möglichkeit zur Teilnahme und Äußerung geben, zum Beispiel durch Nutzen einer Chat-Funktion. Diese Möglichkeiten zur Überwindung sozialer Zwänge, die Menschen in Gewohnheiten und Routinen festhalten,

können allerdings je nach Intention und Perspektive von Anbieter und Nutzer positive und negative Folgen haben (Fogg 2003). So kann soziale Erwünschtheit dazu führen, dass ein zurückhaltender Mensch sich dazu gezwungen fühlt, im Sinn des Arbeitgebers errungene Abzeichen in einem virtuellen Trophäenschrank auszustellen; zusätzlich kann er demotiviert werden, wenn er seine von ihm als wichtig erachtete Tätigkeit durch Gamification trivialisiert sieht was ihn möglicherweise sogar zu einer Kündigung veranlasst.[1] Ein anderes Beispiel ist ein computergesteuerter Spielautomat, der mit Animationen und Erzählungen ausgestattet wurde, um das Spielerlebnis attraktiver zu machen. Hersteller und Betreiber des Kasinos, in dem er aufgestellt ist, verdienen Geld, der Spieler aber verliert Geld und darüber hinaus auch Zeit (Fogg 1998).

3.1 Gamification

Spiele sind so alt wie die Zivilisation, aber die persuasive Technologie Gamification ist ein Phänomen der Informationsgesellschaft (Floridi 2014). Eine häufig zitierte und sehr breite Definition stammt von Deterding et al., die Gamification als „The use of game-elements and game-design technics in non-game contexts" beschreiben (Deterding et al. 2011).[2] Spezifischer definieren Blohm und Leimeister Gamification als „eine auf Informationstechnologie beruhende Zusatzdienstleistung, mit der die Motivation ihrer Nutzer unterstützt und Verhaltensänderungen erzeugt werden sollen" (Blohm und Leimeister 2013), und aus einer Dienstleistungsmarketing-Perspektive definieren Huotari und Hamari: „Gamification bezieht sich auf einen Prozess, bei dem ein Dienst mit Möglichkeiten für spielerische Erfahrungen erweitert wird, um die allgemeine Wertschöpfung der Nutzer zu unterstützen" (Huotari und Hamari 2012).

Der Begriff Gamification wurde im Umfeld der digitalen Medienindustrie zum ersten Mal im Jahr 2008 erwähnt und ab 2010 durch verschiedene Akteure aus der Wirtschaft und Vorträge auf Kongressen verbreitet (Deterding et al. 2011). Zur Gamification werden Produkte, Dienstleistungen und Informationssysteme mit Spiel-Design-Elementen versehen, um die Motivation und das Verhalten von Nutzern sowie die Produktivität von Mitarbeitern zu steigern (Huotari und Hamari 2012; Blohm und Leimeister 2013). Anders als bei herkömmlichen

[1] Unter sozialer Erwünschtheit ist die Tendenz zu verstehen, bei Befragungen so zu antworten, dass Selbstauskünfte weniger dem persönlichen Erleben und Verhalten entsprechen, sondern sozialen Normen und Erwartungen (vgl. Vesely und Klöckner 2020).

[2] „Die Verwendung von Spielelementen und Spielgestaltungstechniken in nichtspielerischen Kontexten".

Anreizkonzepten, die zum Beispiel wie finanzielle Zuwendungen auf extrinsische Motivation ausgerichtet sind, ist das Ziel von Gamification die Steigerung von sowohl intrinsischer als auch extrinsischer Motivation (Blohm und Leimeister 2013; Vieira et al. 2021; Sardi et al. 2017). Durch Dokumentation des Spielerverhaltens können Fortschritte visualisiert werden, und auch das Erreichen individuell bestimmter Ziele kann ein Gefühl von hoher Leistungsfähigkeit und Zufriedenheit erzeugen (Blohm und Leimeister 2013). Gamification ermöglicht soziale Interaktion und kann im Austausch oder Wettbewerb zu sozialer Bestätigung und der Wahrnehmung eines Gemeinschaftsgefühls führen (Arora und Razavian 2021; Koivisto und Hamari 2019). Letzteres wird verstärkt und darüber hinaus Bedeutsamkeit vermittelt, wenn die gemeinsame Lösung einer Aufgabe einem höheren Ziel dient, für das sich die Spieler auserwählt fühlen (Blohm und Leimeister 2013; O'Sullivan et al. 2021). Generell können durch Gamification emotionale Bedürfnisse zum Beispiel nach Erfolg stimuliert werden und das Selbstwertgefühl, die wahrgenommene Selbstwirksamkeit sowie Zufriedenheit und Optimismus zunehmen (Blohm und Leimeister 2013; Sardi et al. 2017). Tab. 1 enthält eine Übersicht über verschiedene Spiel-Design-Elemente, ihre Mechanik und Dynamik sowie die sich dahinter verbergende mögliche Motivation eines Nutzers. Dynamik und Motivation können je nach Design und Mechanik variieren (Blohm und Leimeister 2013).

Tab. 1 Spiel-Design-Elemente und Motivation (modifiziert nach Blohm und Leimeister 2013)

Spiel-Design-Elemente		Motivation
Mechanik	Dynamik	
Dokumentation des Spielerverhaltens	Exploration	Wissbegierde
Punkte, Abzeichen, Trophäen	Sammeln	Leistung
Ränge	Wettbewerb	Leistung
Ränge, Levels, Reputationspunkte	Statuserwerb	Soziale Anerkennung
Gruppenaufgaben	Zusammenarbeit	Sozialer Austausch
Zeitdruck, Aufgaben, Missionen	Herausforderung	Kognitive Stimulation
Avatare, virtuelle Welten, virtueller Handel	Entwicklung, Organisation	Selbstbestimmung, Selbstwirksamkeit

In einem anderen Modell, dem Octalysis Framework von Chou, werden acht grundlegende Motivationsfaktoren unterschieden, die Nutzer zu Spielen und spielerischen Aktivitäten antreiben (Abb. 2). Auf vertikaler Ebene werden analytische, materielle und überwiegend extrinsische Faktoren der linken Gehirnhälfte zugeordnet und kreative, soziale, meist intrinsische Faktoren der rechten (Chou 2017). Mit „Weißer Hut" werden in der oberen Hälfte des Modells positive Motivationsfaktoren beschrieben, die legitimen und wertvollen Antriebsfaktoren entsprechen; das kann der Wunsch sein, dem eigenen Handeln Bedeutung zu verleihen, die eigene Kreativität zu fördern oder ein Gefühl von Leistungsfähigkeit erleben zu wollen (O'Sullivan et al. 2021). Zu den mit „schwarzer Hut" betitelten negativen Faktoren in der unteren Hälfte des Modells zählen die Angst vor Verlust, Unvorhersehbarkeit oder der Wunsch, etwas besitzen zu wollen, nur weil es kaum zur Verfügung steht (O'Sullivan et al. 2021). Die Bildsprache ist Western-Filmen entnommen, in denen sich zwei Cowboys zum Duell gegenüberstehen und der schurkische Charakter an einem schwarzen Hut zu erkennen ist, der heldenhafte an einem weißen. Etwas näher können die acht wesentlichen Motivationsfaktoren nach Chou wie folgt beschrieben werden: 1) Epische Bedeutung und Berufung, 2) Entwicklung und Leistung, 3) Förderung von Kreativität und Feedback, 4) Eigentum und Besitz, 5) sozialer Einfluss und Beziehungen, 6) Knappheit und Ungeduld, 7) Unvorhersehbarkeit und Neugierde, 8) Verlust und Vermeidung (O'Sullivan et al. 2021; Chou 2017).

3.2 Manipulation

Der Begriff „Manipulation" hat im Kontext von Gamification zwei Bedeutungen: zum einen ist unter technischen Gesichtspunkten und wörtlich aus dem Lateinischen abgeleitet mit Manipulation die Handhabung im Sinn einer Steuerung mit der Hand gemeint, zum anderen ist aus einer psychologischen, politischen und sozialen Perspektive unter Manipulation eine gezielte und verdeckte Einflussnahme auf Gruppen oder Menschen ohne deren Wissen und Zustimmung zu verstehen.[3] In unserem Alltag ist letztere insbesondere Bestandteil von Werbung, Marketing und Verkauf, lässt sich aber auch in beruflichen oder privaten Beziehungen beobachten. Weil durch Manipulation die Vernunftfähigkeit eines Menschen umgangen wird, ist sie immer auch ein Eingriff in seine Autonomie, da er an einem selbstbestimmten Handeln gehindert wird: Manipulation blockiert

[3] Zusammengesetzt aus latein. *manus*, die Hand, und *plere*, füllen, entsteht im übertragenen Sinn die Redewendung „etwas in der Hand haben".

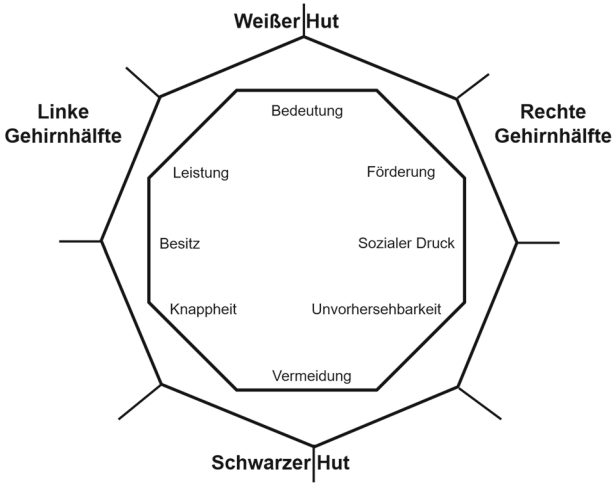

Abb. 2 Octalysis Framework zur Gamification. (Quelle: Eigene Darstellung nach Chou 2017)

eine Abwägung aller zur Verfügung stehenden Möglichkeiten und damit auch ein Vorgehen nach selbst bestimmten Präferenzen im Gegensatz zu von anderen zuvor ausgewählten Präferenzen (Blumenthal-Barby und Burroughs 2012). Sie lässt sich zwischen Nötigung als Gewaltandrohung oder unmittelbare Gewaltanwendung und rationaler Argumentation verorten (Blumenthal-Barby und Burroughs 2012). Unter bestimmten Umständen, wie zum Beispiel im Fall selbstschädigenden Verhaltens oder bei bestehenden kognitiven Einschränkungen, kann Manipulation ethisch vertretbar sein. Nutzen und Schaden sind jedoch im Einzelfall immer zu prüfen und gegeneinander abzuwägen (Blumenthal-Barby und Burroughs 2012).

Die Zunahme von digitalen Technologien und der Zugang zu ihnen wurde durch die Verbreitung von Smartphones und den Ausbau digitaler Infrastrukturen in den vergangenen zehn Jahren stark vorangetrieben. Nicht zuletzt wurde dieser Trend durch die Corona-Pandemie getriggert, in der unmittelbare Kommunikation nur eingeschränkt möglich war. Gleichzeitig stieg die Popularität von Gamification, die von den Entwicklern sozialer Netzwerke, vergleichbarer Plattformen und anderer Applikationen für Smartphones oder für smarte Umgebungen (S*mart Environment)* vermehrt installiert wurde, um die Interaktion und das Engagement der Nutzer zu steigern (Arora und Razavian 2021). Da Gamification wesentlich darauf abzielt, das Verhalten von Nutzern und Konsumenten zu

verändern, ist sie *prima facie* auch dem Vorwurf ausgesetzt, manipulativ zu sein (Kim und Werbach 2016; Arora und Razavian 2021). Im Zusammenhang mit einem weiterhin zunehmenden Medienkonsum und Verbreitung digitaler Technologien in der Lebenswelt wie zum Beispiel in Form von autonomem Fahren, Wearables, unbemannten Drohnen oder 3-D-Druckern verdient dieser Vorwurf eine genauere Betrachtung. Eine absichtliche Täuschung liegt zweifelsohne vor, wenn Gamification-Elemente und -Mechanismen vor ihren „Benutzern" verborgen werden; auch das Untergraben der Autonomie eines Nutzers, wenn Sucht und Ablenkung eine Selbstreflexion verhindern, ist als manipulativ zu bezeichnen (Floridi 2014; Arora und Razavian 2021; Kim und Werbach 2016).

Einem anderen Konzept als Gamification folgen sogenannte *Serious Games,* die vollständige Spiele mit eindeutigem Beginn, Verlauf und Ende sind und lange vor dem Einzug von Computertechnologien in den Alltag entwickelt wurden (siehe Abschn. 3.1). Inwieweit der als Nudging bekannt gewordene Ansatz zur Verhaltensänderung aus der Verhaltensökonomie als gezielte und verdeckte Einflussnahme auf Menschen ohne deren Wissen und Zustimmung gelten kann, wird im folgenden Text näher betrachtet (Deterding et al. 2011, siehe Abschn. 3.3 und 3.5).

Die Vorstellung eines Gaming-Kontinuums, in dem drei Bereiche unterschieden werden, enthält Abb. 3 (O'Sullivan et al. 2021). Der Bereich der *Real World* ganz links in der Abbildung beinhaltet Arbeitsprozesse, (Bildungs-)Erfahrungen und Simulationen. In diesem Zusammenhang können Spiel-Design-Elemente im Arbeits- und Alltagsleben enthalten sein. Der Bereich der Gamification ist dem der *Real World* vergleichbar, ist aber durch die explizite Absicht definiert, (Bildungs-)Erfahrungen durch den Einsatz von Spiel-Design-Elementen zu beeinflussen. Dieser mittlere Bereich des Gamification-Kontinuums lässt sich in eine strukturelle und eine inhaltliche Kategorie unterteilen. Wird strukturelle Gamification verwendet, bleiben die Inhalte unverändert, aber das Produkt wird mit Spiel-Elementen versehen; dieser Ansatz beruht auf behavioristischen und operativen Konditionierungstechniken und hat das Ziel, einen Nutzer durch Spiel-Elemente wie Feedback und positive Verstärkung gewünschter Verhaltensweisen zum Beispiel in Form von Abzeichen oder Ranglisten extrinsisch zu motivieren (siehe Tab. 1) (Filatro und Cavalcanti 2016). Im Gegensatz dazu folgt inhaltliche Gamification der Selbstbestimmungstheorie von Ryan und Deci und zielt darauf ab, die von einem intrinsisch motivierten Nutzer gewünschten Aktivitäten zu pflegen (O'Sullivan et al. 2021; Ryan und Deci 2000). Zu diesem Zweck werden Spiel-Mechaniken eingesetzt, die die Merkmale von Spielen stärker betonen wie zum Beispiel eine Spielhandlung mit Charakteren oder Rollenspiele im Fall interaktiver Anwendungen (Filatro und Cavalcanti 2016; Vieira et al. 2021).

Abb. 3 Gaming-Kontinuum modifiziert nach O'Sullivan et al. (2021)

Der Bereich der *Games* auf der rechten Seite der Abbildung umfasst sowohl klassische Spiele wie Monopoly oder die Siedler von Catan als auch *Serious Games* oder *Game-based Learning,* die alle als vollständige Spiele und zur Unterhaltung entwickelt wurden. Für eine Vermittlung von Bildungsinhalten sind jedoch nur *Serious Games* und *Game-based Learning* konzipiert (O'Sullivan et al. 2021).

3.3 Gamification, the Dark Side

Mit der Verbreitung digitaler Technologien ist in Wissenschaft und Forschung auch das Interesse an Gamification gestiegen und erfordert die nähere Betrachtung der Risiken und Nebenwirkungen dieser persuasiven Technologie, deren Betonung des Spaßfaktors soziale, politische und kulturelle Folgen verschleiert (Heinz und Fischer 2020). Um aus Gründen der Effektivität wirkliche Vorteile herausarbeiten zu können, ist zu hinterfragen, ob Gamification *de facto* zu besseren Ergebnissen führt oder die Nutzer gamifizierter Elemente vom eigentlichen Zweck des Systems abgelenkt werden, womit Produktivitätsverluste einhergehen können (Nyström 2021; Callan et al. 2015). Darüber hinaus ist es aus ethischer Perspektive wichtig, Rahmenbedingungen für die Gestaltung von Gamification zu schaffen, die den Wertvorstellungen einer Gesellschaft entsprechen (Nyström 2021; Heinz und Fischer 2020). So ist zu prüfen, ob jemand einen unlauteren Vorteil aus dem Einsatz von Gamification ziehen oder Autonomie durch Manipulation untergraben werden kann, ob beabsichtigt oder unbeabsichtigt Schaden für den Nutzer entstehen kann oder der Charakter von Entwicklern, Nutzern oder

weisungsbefugten Personen wie Arbeitgebern, Ausbildern oder Lehrern negativ beeinflusst wird (Kim und Werbach 2016).

Einen schwerwiegenden Schaden durch Gamification für Betroffene und die Gesellschaft spiegelt die gestiegene Prävalenz von Suchterkrankungen im Zusammenhang mit Computerspielen wider, die die WHO dazu veranlasste, der Computerspielsucht in der revidierten Version der internationalen Klassifikation von Krankheiten (ICD-11) eine separate Ziffer zuzuerkennen (6C51 Gaming disorder).

Der Ausdruck *Dark Side of Gamification* und sein Akronym DsoG haben sich systematischen Recherchen in Datenbanken zufolge etabliert (Heinz und Fischer 2020; Nyström 2021; Johnson et al. 2016). Zagal et al. sind in ihrer Forschung über „dunkle Muster im Design von Computerspielen" zu dem Ergebnis gekommen, dass Spieler von Werten oder Überzeugungen, die von Entwicklern in Spiele eingearbeitet wurden, beeinflusst werden können, auch wenn sie sie nicht bewusst wahrnehmen (Zagal et al. 2013). Spieler können beispielsweise dazu ermuntert werden, gute oder böse Rollen in einem Computerspiel zu übernehmen; es ist auch möglich, dass im Designmuster eines Spiels die Förderung von Kameradschaft enthalten ist, womit implizit ausgedrückt wird, dass Kameradschaft als gut zu bewerten ist ist (Zagal et al. 2013).

Allgemein sind unter dunklen Mustern Designstrategien für Computerspiele zu verstehen, die den Entwicklern und nicht dem Zielpublikum zugutekommen; sie beinhalten unmoralische Anwendungen wie Nötigung, Täuschung, Intrigen, Verrat oder Betrug (Nyström 2021; Zagal et al. 2013). Es entstehen Benutzeroberflächen, die darauf abzielen, Menschen zu täuschen, und Zagal et al. definieren: „Dunkle Muster im Design von Computerspielen werden von einem Spielentwickler absichtlich eingesetzt, um schlechte Erfahrungen für die Nutzer zu erzeugen, die im Widerspruch zu ihren Interessen stehen und denen sie im Vorfeld nicht zugestimmt haben würden" (Zagal et al. 2013). Drei solcher Muster, die sich auf Zeit, Geld und Sozialkapital eines Spielers beziehen, können als fragwürdig oder unethisch betrachtet werden und sind generell zu kritisieren, weil sie mehr Zeit kosten, als die Spieler vermuten konnten, und sie auf diese Weise vorsätzlich um ihre Zeit betrogen werden (Zagal et al. 2013; Linehan et al. 2015; Arora und Razavian 2021). Beispiele für dunkle Muster, die sich im Speziellen auf die Zeit eines Nutzers beziehen, sind das sogenannte *Grinding,* durch das Spieler dazu gezwungen werden, eintönige und sich wiederholende Aufgaben zu erledigen, um in Spielen wie *World of Warcraft* voranzukommen, sowie das „Spielen nach Vereinbarung", durch das Nutzern bestimmte Zeiten vorgegeben werden, was ihr Arbeits- und Sozialleben eingeschränkt (Linehan et al. 2015; Zagal et al. 2013; Kim und Werbach 2016).

Mit monetär ausgerichteten dunklen Mustern werden Spieler dazu verleitet, mehr Geld auszugeben, als sie ursprünglich vorgehabt hatten. So ist es mithilfe von „Pay-to-Skip"-Mustern möglich, reales Geld einzusetzen, um schwierige oder unüberwindbare Spielabschnitte zu bewältigen. Ein anderes dunkles Muster ist das sogenannte „Monetised Rivalries", das auch als „Pay-to-win" bekannt ist und dessen Anreiz zum Geldauszugeben im Erreichen eines gewissen Status innerhalb eines Spiels besteht, wie beispielsweise eine bestimmte Position innerhalb einer Rangliste (Linehan et al. 2015; Zagal et al. 2013). Auf das Sozialkapital zielende dunkle Muster kompromittieren das Sozialkapital der Spieler, worunter der Wert ihres sozialen Status und ihrer sozialen Beziehungen im Leben außerhalb des Spiels zu verstehen ist. Als Beispiel seien Sozialpyramiden-Schemata genannt, die ein Vorankommen der Spieler im Spiel so lange blockieren, bis diese Freunde oder Bekannte dazu überreden können, auch Spieler des Spiels oder Mitspieler zu werden (Linehan et al. 2015). Innerhalb des Spiels werden dann auch Nachrichten über Handlungen der benannten Freunde versendet, die nie stattgefunden haben; in dem Spiel *SimCity Social* kann eine Mitteilung zum Beispiel lauten: „X hat Dir ein Geschenk geschickt" (Zagal et al. 2013). Die als „Friend Spam" bezeichnete Konsequenz kann eintreten, wenn im Spiel unter Angabe eines vermeintlich harmlosen Zwecks wie dem, „Freunde zu finden, die diesen Dienst bereits benutzen", zum Beispiel nach Twitter- oder E-Mail-Zugangsdaten gefragt wird, die in der Folge dazu benutzt werden, um über den privaten Account Inhalte zu veröffentlichen oder Spam-Nachrichten zu versenden (Zagal et al. 2013). Die Auswirkungen einer solchen Identitätspreisgabe auf reale Beziehungen können gravierend sein und die psychosoziale Gesundheit schwer beeinträchtigen (Zagal et al. 2013; Linehan et al. 2015).

Spielentwicklern und -designern mit den oben beschriebenen Absichten wird nicht nur Manipulation, sondern auch Ausbeutung der Nutzer vorgeworfen (Kim und Werbach 2016; Arora und Razavian 2021; Nyström 2021; Hyrynsalmi et al. 2017). Nicht von Gamification, sondern von *Exploitationware* sei nach Meinung des Autors und Spielentwicklers Ian Bogost zu sprechen, da Gamification wenig mit Spielen zu tun habe und eher der Verhaltensökonomie zuzuschreiben sei; mithilfe der persuasiven Technologie Gamification werde versucht, Entscheidungen von Menschen unter Ausnutzung von kognitivem Bias und Anwendung von Manipulationsstrategien umzugestalten (Bogost 2013).

Sogenannte *Digital Nudges* (siehe auch Abschn. 3.3), deren Konzept dem der Verhaltensökonomie stark ähnelt, können mit Spiel-Design-Elementen ausgestattet werden und tragen durch personalisierte Nachrichten, kleine digitale Belohnungen oder rechtzeitiges Erinnern dazu bei, dass Menschen ihre Entscheidungen überdenken und möglicherweise umgestalten (O'Sullivan et al. 2021).

Die Wahl solcher *Digital Nudges* ist jedoch freiwillig, und nur im Fall einer Voreinstellung, die den Nutzer möglicherweise darüber hinwegtäuscht, dass zum Beispiel seine Fitness-Daten übertragen und zu Zwecken gesammelt werden, denen er nicht zustimmen würde, kann von Ausbeutung und Manipulation gesprochen werden (Nyström 2021; Heinz und Fischer 2020). Eindeutig als Ausbeutung kann der Einsatz von Gamification am Arbeitsplatz bezeichnet werden, wenn sie zur Steigerung der Effizienz der Mitarbeiter führt, ohne dass diese am Erfolg beteiligt werden, wie beispielsweise durch Lohnerhöhungen: Während der Unternehmer vom finanziellen Gewinn profitiert, erhalten die Mitarbeiter lediglich virtuelle Belohnungen in Form von Punkten und Abzeichen (Kim und Werbach 2016). In diesem Zusammenhang ist in Bezug auf eine mögliche Verletzung des Prinzips der Autonomie zu unterscheiden, ob Arbeitnehmer dem Einsatz von Gamification zustimmen können oder nicht. Kim und Werbach weisen darauf hin, dass eine Gesellschaft, in der Gamification marktfähig ist und von Arbeitnehmern bevorzugt wird, ein besorgniserregender Hinweis darauf sei, dass unter dem grundlegenden wirtschaftlichen Paradigma Arbeitnehmer dazu gezwungen seien Gamification zu wählen, um Sinn und Spaß in ihrer Arbeit erfahren zu können (Kim und Werbach 2016).

In einen moralischen Graubereich zwischen Chous weißem und schwarzem Hut (siehe 3 und Abb. 2) kann Gamification führen, die legal aber fragwürdig ist oder in guter Absicht entwickelt wurde und unerwartete Konsequenzen hat, die zum Beispiel durch Manipulation zur Verletzung ethischer Prinzipien führen (Heinz und Fischer 2020; Hyrynsalmi et al. 2017). So können sich Anwender von Gamification dazu veranlasst sehen, wie ein Versuchstier bestimmte Verhaltensweisen nur zu zeigen, wenn sie dafür belohnt werden (Bui et al. 2015; Nyström 2021). In diesem Fall würde die intrinsische Motivation durch das Streben nach extrinsischer Belohnung ersetzt und letztlich sogar das ursprüngliche Ziel der Gamification aufgehoben, die Motivation der Nutzer zu steigern (Nicholson 2012). Es können auch Grenzen verwischt werden, wenn die Spielwelt mit der Berufs-, Geschäfts- oder Ausbildungswelt verschmilzt und dadurch normative Spannungen auftreten. So kann der Betreiber eines Call-Centers Gamification für seine Angestellten nutzen, die Belohnungen für die Anzahl geführter Gespräche oder Zufriedenheit der Anrufer in Form von Punkten und Abzeichen erhalten; letztere können in einem virtuellen Trophäenschrank für alle Mitarbeiter sichtbar ausgestellt werden. In diesem Fall sind der Sozialraum von Spiel- und Arbeitswelt eins, und es kommt zu einer Überlagerung von virtuellen und realen Normen sowie organisatorischen und individuellen Interessen (Kim und Werbach 2016). Rationale und normative Entfremdung können die Folge sein und das Selbstwertgefühl und die Autonomie von Nutzern korrumpieren (Dittmeyer 2020). Dasselbe

gilt für Schüler oder Studenten, die virtuelle Abzeichen für Bildungsleistungen erhalten und simultan in beiden Welten miteinander konkurrieren (O'Sullivan et al. 2021).

Andere Nebenwirkungen bis hin zu Schäden von Gamification können in Form von Ablenkung oder Alltagsflucht auftreten und zu Realitätsverlust führen; darüber hinaus verhindert die Selbstvergessenheit oder Flow-Erleben im Spiel Selbstreflexion, und das Prinzip der Autonomie wird untergraben (Kim und Werbach 2016; O'Sullivan et al. 2021). Gamification kann auch als demotivierend wahrgenommen werden, wenn sie zur Banalisierung und Entwertung von Aufgaben oder Tätigkeiten führt (Hyrynsalmi et al. 2017).

Der größte Schaden von Gamification liegt in der zu Beginn des Abschnitts bereits erwähnten Computerspielsucht. Sie kann durch Gamification gebahnt werden, weil sie die gleichen Mechanismen von Verführung und variabler Belohnung beinhaltet, die die Grundlage für Spielautomaten sind (Andrade et al. 2016; Nyström 2021). In einer Längsschnittstudie von 2019 bis 2021 des Deutschen Zentrums für Suchtfragen des Kindes- und Jugendalters im Universitätsklinikum Hamburg-Eppendorf und der Krankenkasse DAK wurde die pathologische Nutzung von Gaming und sozialen Medien vor und nach der Corona-Pandemie untersucht, und es konnte ein signifikanter und im Ausmaß besorgniserregender Anstieg gemessen werden, zu dessen Rückentwicklung noch geforscht wird (Paschke et al. 2021; DAK 2023). Aus gesellschaftlicher Sicht sind die Konsequenzen zu überdenken, die die Integration von Spiel-Design-Elementen in unser Alltagsleben hat (Hyrynsalmi et al. 2017).

3.4 Serious Games, the Bright Side

Unter dem wie ein Oxymoron klingenden Begriff *Serious Games* oder auch *Game-based Learning* (siehe Abb. 3) werden digitale Anwendungen mit spielerischen und didaktischen Anteilen verstanden, die gegen reine Unterhaltung mit einem explizit formulierten Bildungsziel abgegrenzt werden können; der Unterhaltungsfaktor ist dem Lern- oder Trainingsziel dabei übergeordnet (Tolks et al. 2020). *Serious Games* sind vollständige Spiele oder Simulationen zur Aufklärung, Schulung, Information oder Verhaltensänderung, die in verschiedenen Bereichen eingesetzt werden wie zum Beispiel der Bildung, Gesundheit und Industrie, aber auch im Bereich des Militärs und der Politik (Deterding et al. 2011; Bui et al. 2015; Sardi et al. 2017).

Spiele im Kontext von Gesundheit sind sogenannte *Serious Games for Health,* die je nach Intention unterschiedliche Ausprägungen haben (Tab. 2). Auch dieser

Tab. 2 Übersicht *Serious Games for Health* (Lu und Kharazzi 2018)

Spielart	Beschreibung
Aktiv/rythmisch	Intensive körperliche Aktivität des Spielers ist erforderlich
Buch/Film	Interaktives Buch oder interaktiver Film
Fahren	Rennenfahren mit Land-, Luft- oder Wasserfahrzeug
Kampf	Spielfigur muss für den Nahkampf mit einem Gegner gesteuert werden
Rätsel	Rätsellösungen durch Mustererkennung, Wortergänzung, Sequenzlösung
Rollenspiel	Agieren als Spielcharakter, enthält viele erzählerische Elemente
Shooter	Abschießen von Feinden oder Objekten, um Tod der Spielfigur zu vermeiden, erfordert Geschwindigkeit und kurze Reaktionszeit
Gelegenheitsspiel	Einfache interaktive Anwendung (Bsp.: ein Ball muss in der Luft gehalten werden)
Simulation	Fähigkeiten des Spielers werden für die wirkliche Welt verbessert
Sport	Spiel mit Bezug zu Sportereignis, Spieler ist körperlich nicht aktiv
Strategie	Spieler muss eigene Entscheidungen treffen, um zu gewinnen
Trivia/Quiz	Frage-Antwort-Spiel oder Quizshow

Begriff kann als Oxymoron verstanden werden, da ausgedehntes und intensives Computerspielen überwiegend in sitzender Haltung stattfindet und oftmals von weiteren, wenig gesundheitsförderlichen Verhaltensweisen begleitet ist. Mit ihrer Publikation „Games against health: A player-centered design philosophy“ hat ein internationales Autorenteam von Forschern auf diesen Widerspruch hingewiesen (Linehan et al. 2015).

Serious Games for Health können grob unterteilt drei Zwecken dienen: zur Vermittlung von allgemeinen Gesundheitsinformationen oder von Informationen zur Verhaltensänderung im Sinn von Prävention und Gesundheitsförderung, zu therapeutischen Zwecken und in der medizinischen Aus-, Fort- und Weiterbildung (Tolks et al. 2020).

Worauf die hohen Erwartungen und Hoffnungen sowohl der Nutzer als auch der Hersteller von *Serious Games for Health* beruhen, beschreiben Pereira et al.: Gamification kann sich positiv auf emotionale Erfahrungen auswirken, indem Neugier, Optimismus und Stolz gefördert werden. Mithilfe von Gamification ist es möglich, negative emotionale Erfahrungen zu verarbeiten und sie in positive Erfahrungen umzuwandeln. Selbstwahrnehmung, kognitive und psychomotorische Fähigkeiten können positiv beeinflusst werden, indem komplexe

Regelsysteme bedient werden müssen, die ein aktives Ausprobieren und Entdecken erfordern. Darüber hinaus kann Gamification die Kommunikationsfähigkeit, das Urteilsvermögen und soziale Fähigkeiten wie Leitung und Zusammenarbeit verbessern (Pereira et al. 2014). Ziele von *Game-based Learning* sind die Erweiterung von Wissen, Vertiefung von Kompetenz und Stimulation zur Verhaltensänderung (Tolks et al. 2020). Zur didaktischen Unterstützung werden das Eintauchen in eine Spielwelt (Immersion), die Verwendung von Geschichten (Storytelling) und Flow-Erleben genutzt (Tolks et al. 2020). Mit Flow-Erleben wird ein Zustand tiefer Konzentration bezeichnet, in dem eine Person so sehr in eine Aufgabe oder ein Spiel vertieft ist, dass sie ihre Selbstwahrnehmung ignoriert und das Zeitgefühl verliert (Csikszentmihalyi 2008). Auch ein ungestörter Erlebnisfluss gilt als Flow und ist ein von Spiel-Designern und -Entwicklern gewünschter Zustand, um Spieler so gut wie möglich unterhalten und beschäftigen zu können (Andrade et al. 2016).

In einem systematischen Review zu Gamification in eHealth-Anwendungen für chronisch Kranke, körperliche Aktivität und mentale Gesundheit wie zum Beispiel zur Überwachung des Blutzuckers von Jugendlichen mit Diabetes Typ 1 oder kognitives Training für Alzheimer-Patienten beschreiben Sardi et al., dass Gamification stark zu Verhaltensänderungen und zur regelmäßigen Beschäftigung mit der Anwendung motiviert, halten als Ergebnis ihrer Bewertung von 46 eingeschlossenen Studien jedoch fest, dass extrinsische Belohnungen die Bereitschaft zur Nutzung nur kurzfristig erhöhen. Sie fordern eHealth-Lösungen, die auf gut begründeten Theorien aufbauen und bereits vorhandene Kenntnisse zur psychologischen Wirkung von Spiel-Design-Mechaniken nutzen (Sardi et al. 2017).

Kato et al. untersuchten in einer multizentrischen randomisierten Studie mit 375 krebskranken Kindern und jungen Erwachsenen in Kanada, Australien und den USA den Einfluss des Computerspiels *Re-Mission.* Es enthält ein Shooter-Spiel, und darüber hinaus werden Informationen über die Krebsbehandlung und deren Auswirkungen auf den Körper vermittelt. Im Vergleich zur Kontrollgruppe, der ein handelsübliches Computerspiel mit vergleichbarem Design zur Verfügung gestellt wurde, verbesserten sich Selbstwirksamkeit und Wissen über die Krankheit mit statistisch signifikantem Unterschied zwischen den Gruppen (Kato et al. 2008). Demgegenüber konnten Sajeev et al. in ihrem systematischen Review mit 26 Studien zur Reduktion von Schmerzen bei pädiatrischen Eingriffen und Angst von Kindern und ihren erwachsenen Begleitern durch interaktive Videospiele keinen Unterschied zwischen Spielen, die altersgerecht aufbereitete Informationen über den bevorstehenden Eingriff beinhalteten, Spielen zur reinen Ablenkung

sowie interaktiven Virtual-Reality-Spielen und nicht virtuellen Reality-Spielen feststellen (Sajeev et al. 2021).

In der Physiotherapie liegt der Einsatz von Gamification wegen der erforderlichen repetitiven Stimulation des Bewegungsapparats nah, außerdem können rhythmische Elemente zur Unterstützung eingesetzt werden (Janssen et al. 2017). In ihrem 2021 veröffentlichten systematischen Review zu Gamification von Therapien zur Verbesserung motorischer Fähigkeiten von Patienten mit Schlaganfall, Multipler Sklerose oder zerebraler Lähmung berichten Vieira et al. über 12 Studien und insgesamt 512 Patienten im Alter zwischen 18 bis über 85 Jahren mit einem geschätzten Altersdurchschnitt von 59 Jahren; in acht Studien konnten Verbesserungen beobachtet werden. Mit Spielen, die auf spezifische Kasuistiken zugeschnitten sind und aus der Ego-Perspektive ohne sichtbaren Spielcharakter und im Einzelspielermodus gespielt werden, ohne dass die Umgebung im virtuellen Erleben mit dem Spieler interagiert (nicht immersive virtuelle Realität), konnten die besten Ergebnisse erzielt werden; gleichwohl wurden handelsübliche Spiele im Vergleich dazu als motivierender und ansprechender wahrgenommen (Vieira et al. 2021).

Zu Gamification in der medizinischen Aus-, Fort- und Weiterbildung berichten Graafland et al. in ihrem systematischen Review über 26 Studien zur Ausbildung und zum Training von Chirurgen, dass integriertes (blended) und interaktives Lernen gut einsetzbar sei, warnen aber davor, dass viele Anwendungen mit Gamification nicht validiert seien (Graafland et al. 2012). In einem Scoping Review von 2019 mit 25 eingeschlossenen Studien wurde Gamification in der Aus-, Fort- und Weiterbildung von Pflegekräften, Ärzten, Pharmazeuten und Rettungssanitätern untersucht. Die Lehrinhalte waren äußerst heterogen und reichten von einer Behandlung des Herzstillstands bis zum Verhalten bei Geräte- oder Maschinenversagen während einer Operation. In 18 Studien diente eine Kontrollgruppe mit konventionellen Lehrmethoden als Vergleich, und in 14 Studien waren die Tests zur Ermittlung des Lernerfolgs nach dem Einsatz von *Serious Games* deutlich besser mit statistisch signifikantem Unterschied; in den übrigen vier Studien konnte kein Unterschied festgestellt werden (Haoran et al. 2019).

Für eine unterhaltsame Gestaltung von Lehre und Lernen durch *Serious Games* sprechen nach O'Sullivan et al. die größere Dynamik im Vergleich zum Unterricht im Klassenzimmer, die größere Sicherheit von Simulationen im Vergleich zu tatsächlichen Handlungen und der Vorteil eines leicht zu installierenden Punktesystems, um Kompetenz in bestimmten Fertigkeiten und Fähigkeiten zu messen (O'Sullivan et al. 2021).

Allerdings können auch mögliche Risiken und Nebenwirkungen von *Serious Games* hinsichtlich von Motivation, Verhalten, Intention und Leistung beschrieben werden: Im Fall extrinsischer Belohnung muss diese regelmäßig in Aussicht gestellt werden, um das Motivationslevel hoch zu halten; bleibt die Belohnung aus, besteht die Gefahr, dass die Nutzung von *Serious Games* reduziert oder ganz eingestellt wird. Darüber hinaus können ursprünglich intrinsisch motivierte Teilnehmer durch Belohnungssysteme demotiviert werden, weil sie sich wie Versuchstiere fühlen, die auf einen Stimulus reagieren. Ebenfalls denkbar ist, dass extrinsische Motivation Nutzer auf das Erreichen der höchsten Punktzahl fixiert und damit von den Inhalten oder einer Aufgabe ablenkt (Nyström 2021). Auch der Wettbewerb mit anderen kann bei schlechtem Abschneiden demotivieren, entmutigen und zu einer generellen Ablehnung von *Serious Games* führen (Nyström 2021; Toda et al. 2018). Unerwünschtes Verhalten wie Schummeln, Fälschen, Betrügen oder Vortäuschen aufgrund von sozialer Erwünschtheit sowie riskante Verhaltensweisen während einer Simulation können weitere Folgen sein und zu Inakzeptanz führen (Heinz und Fischer 2020). *Serious Games* können Nutzer auch dazu ermutigen, ein bestimmtes Verhalten nur im Fall extrinsischer Belohnung zu zeigen und ansonsten eine Gleichgültigkeit gegenüber Aufgaben oder Lerninhalten zu entwickeln (Toda et al. 2018). Eine bereits vorhandene Tendenz zur Individualisierung kann verstärkt werden, sodass die Bedeutung von Institutionen und Strukturen in den Hintergrund tritt; zudem mag der Wettbewerbsgedanke und das Zurschaustellen von Leistungen in *Serious Games* wenig Anziehungskraft für eher zurückhaltende Menschen besitzen. Neben der bereits erwähnten Ablenkung von Inhalten und der Trivialisierung von Aufgaben kann auch eine falsche Verstärkung durch Konzentration auf eine Verbesserung im Spiel anstelle einer Verbesserung von Wissen und Fähigkeiten erfolgen (Heinz und Fischer 2020). Insbesondere dann, wenn Spiel- und Lernkontext nicht ausreichend aufeinander abgestimmt sind, kann sich die Wirkung von Serious Games in einer verringerten Informationsaufnahme, nachlassender Genauigkeit oder einem vollständigen Rückzug von Aufgaben zeigen. (Heinz und Fischer 2020).

3.5 Nudging, the Grey Side

Im Jahr 2008 wurde der Begriff „Nudging“ (englisch für Schubs oder Stups) durch den Wirtschaftswissenschaftler Richard Thaler und den Rechtswissenschaftler Cass Sunstein eingeführt (Thaler und Sunstein 2008). Das dahinter liegende Konzept ist der Verhaltensökonomie zuzuschreiben und beruht auf der Idee einer Entscheidungsarchitektur, für die alle äußeren Kräfte einbezogen

werden, um Entscheidungen einer Person auf subtile Weise zu lenken. Ein „Architekt" kann eine Umgebung derart gestalten, dass eine bestimmte Möglichkeit mit einer höheren Wahrscheinlichkeit gewählt wird, ohne dass Wahlmöglichkeiten verboten werden oder wirtschaftliche Anreize wesentlich verändert werden (Thaler und Sunstein 2008).[4] Anders als bei einer Verhaltensänderung wird das Umfeld verändert, in dem eine Entscheidung gefällt wird oder sich jemand verhält (Reñosa et al. 2021). So können beispielsweise gesundheitsförderliche Lebensmittel in Supermärkten auf Augenhöhe präsentiert und dadurch eher gekauft werden als solche, die in Bodennähe ausgestellt sind. Ebenso und mit gegenteiliger Wirkung können Wartende im Kassenbereich gegenüber Süßigkeiten und Alkoholika exponiert werden. Nudging wurzelt wie auch Gamification in libertärem Paternalismus. Mit beiden Techniken wird versucht, Einfluss auf Verhalten unter Nutzung kognitiver Abkürzungen zu nehmen (O'Sullivan et al. 2021). Darunter ist zu verstehen, dass man sich bei Routinehandlungen im Alltag, wie beispielsweise dem Einkaufen, eher auf Intuition als auf rationales Denken verlässt. Aufgrund von begrenzten Kapazitäten zur Informationsverarbeitung werden bestimmte Heuristiken genutzt, einem Priming bzw. einer Bahnung durch Anker- oder Framingeffekte gefolgt oder Entscheidungen auf der Basis von Sympathie, Konsistenz, sozialer Validierung oder Knappheit getroffen (O'Sullivan et al. 2021). Solche kognitiven Abkürzungen werden unbewusst gewählt, können aber bei späterer Reflexion dazu führen, dass man sich manipuliert fühlt (s. o.) (O'Sullivan et al. 2021).

Im Gegensatz zu gesetzlichen Regelungen wie der Gurt- und Helmpflicht im Straßenverkehr oder dem Rauchverbot in öffentlichen Gebäuden und Gaststätten wird Nudging nicht als Zwang empfunden, da die Interventionen per Definition als leicht vermeidbar gestaltet sind („easy and cheap to avoid") (Thaler und Sunstein 2008). *De facto* werden aber Individuen in eine von fremder Seite erwünschte Richtung gelenkt, und es ist umstritten, in welchem Ausmaß Nudging die Wahlfreiheit tatsächlich einschränkt und mit einer freiheitlichen Grundhaltung vereinbar ist (Möllenkamp et al. 2019).

Den Effekt von Nudging auf Ernährungsgewohnheiten haben Arno und Thomas in ihrem systematischen Review untersucht. Die Meta-Analyse von Daten aus 42 Studien ergab einen durchschnittlichen Anstieg von 15,3 % in der Häufigkeit „gesunder" Entscheidungen oder Veränderung der gesamten Kalorienaufnahme. Die Autoren diskutieren die große Heterogenität der Messmethoden

[4] Thaler R, Sunstein C: „Ein Nudge ist jeder Aspekt der Entscheidungsarchitektur, der das Verhalten der Menschen auf vorhersehbare Weise verändert, ohne Optionen zu verbieten oder ihre wirtschaftlichen Anreize wesentlich zu verändern."

in den Primärstudien, die zu einem Drittel über keine statistisch signifikanten Unterschiede berichten konnten, als starke Limitation ihrer Ergebnisse und Grund dafür, dass sie keinen robusten Effektschätzer der gepoolten Daten präsentieren können. Ihre Schlussfolgerungen lauten allerdings, dass die Ergebnisse ihres systematischen Reviews und der Meta-Analyse zeigen, dass Interventionen im Sinne von Nudging eine wirksame und praktikable Strategie für Public Health darstellen, um Erwachsene zu einer gesünderen Ernährung zu ermutigen (Arno und Thomas 2016). Deutlich zurückhaltender äußern sich die Autoren eines im Jahr 2019 publizierten Scoping Review zum Einsatz von Nudging zur Förderung körperlicher Aktivität der Allgemeinbevölkerung (Forberger et al. 2019). Sie berichten über die Diskrepanz zwischen der sie überraschenden geringen Anzahl von 35 identifizierten Artikeln und der Aufmerksamkeit, die sowohl Nudging als auch körperlicher Aktivität entgegengebracht wird. Aus der Tatsache, dass ihre Recherchen lediglich Studien zu Interventionen ergaben, in denen Personen individuelle Entscheidungen über ihren Lebensstil zu treffen hatten, wie zum Beispiel Treppen zu steigen statt Rolltreppe zu fahren, schlussfolgern Forberger et al., dass Nudging im Prinzip ein wirksamer Ansatz zur Förderung körperlicher Aktivität der Allgemeinbevölkerung sei, die vorhandenen Möglichkeiten allerdings bei weitem nicht ausgeschöpft und sowohl die Forschungslücke als auch der Forschungsbedarf groß seien (Forberger et al. 2019). Die Effektivität von Nudging zur Verbesserung des Selbstmanagements von chronischen Krankheiten untersuchten Möllenkamp et al. in einem 2019 veröffentlichten systematischen Review. Sie kommen zu dem Schluss, dass weitgehend akzeptierte Formen von Nudging wie Erinnerungen, Feedback oder Verabredungen zur Einhaltung bestimmter Vorsätze zur Verbesserung des Selbstmanagements chronisch Kranker führen, es aber keine Evidenz dafür gebe, dass damit auch eine bessere Kontrolle der Erkrankung verbunden sei (Möllenkamp et al. 2019). In einer der eingeschlossenen Studien wird über eine Medikamentenbox berichtet, die Licht- und Tonsignale sendet, wenn die eingestellte Uhrzeit zur Medikamenteneinnahme erreicht ist (Reddy et al. 2017). Das Öffnen der Box wird aufgezeichnet und drahtlos an den Hersteller weitergeleitet, der aus den Informationen einen wöchentlichen Adhärenz-Bericht mit einer Leistungsbewertung erstellt; auf Wunsch des Patienten kann dieser Bericht auch an den behandelnden Arzt geschickt werden (Reddy et al. 2017).

Blumenthal-Barby und Burroughs diskutieren ethisch relevante Dimensionen von Nudging-Mechanismen und fragen insbesondere, wie das absichtliche Umgehen der Vernunftfähigkeit von Menschen sowie mit Nudging einhergehende Ungerechtigkeit begründet werden können (Blumenthal-Barby und Burroughs

2012). Auf Anreize, Salienz (die Auffälligkeit eines Reizes), Priming, Vor- oder Werkseinstellungen und soziale Normen wird im Folgenden kurz eingegangen.[5]

Die Höhe von Anreizen kann aus zwei Gründen zu ethischen Bedenken führen, denn einerseits kann im Fall zu hoher Anreize ein Angebot wie ein Zwang wirken, weil die Fähigkeit zu einer autonom getroffenen Entscheidung kompromittiert wird, und andererseits stellt ein ungerechtfertigt hoher Anreiz eine Verschwendung von Ressourcen dar (Blumenthal-Barby und Burroughs 2012). Auch die Art des Anreizes kann bedenklich sein, da viele Nudging-Interventionen auf die Änderung von Ernährungs- oder Bewegungsverhalten zielen: beides kann nicht immer von Menschen kontrolliert werden und ist darüber hinaus mit dem sozioökonomischen Status verknüpft. Aufgrund ihrer Lebensverhältnisse könnten bestimmte Personengruppen von derartigen Interventionen nicht profitieren, was Ungerechtigkeit zur Folge hätte (Blumenthal-Barby und Burroughs 2012). Fragwürdig sind auch die möglichen Auswirkungen der oben beschriebenen Medikamentenbox auf das Arzt-Patient-Verhältnis, denn Überwachung und Ermahnung in Form des Adhärenz-Berichts können die Rolle des Arztes in Richtung einer tatsächlichen oder gefühlten Überwachung verändern und das gegenseitige Vertrauen beeinträchtigen (Reddy et al. 2017; Blumenthal-Barby und Burroughs 2012). Salienz wie beispielsweise in Form von Abbildungen zerstörter Lungen auf Zigarettenschachteln oder Priming durch Ausstellung gesundheitsförderlicher Lebensmittel auf Augenhöhe mag vielen gerechtfertigt erscheinen, wohingegen die Akzeptanz von Angaben zur Kalorienzahl eines Gerichts und Dauer bis zu ihrer Verbrennung durch verschiedene Bewegungsarten auf Speisekarten eher gering sein; solche Informationen sind unter dem Aspekt von Genuss und Wohlbefinden weder im Sinne von Restaurantbesuchern noch von deren Betreibern (Blumenthal-Barby und Burroughs 2012). An Voreinstellungen wie beispielsweise der Widerspruchslösung für Organspenden, die erfordern, dass man aktiv Einspruch erhebt, ist zu kritisieren, dass sie Gesundheits- und Lesekompetenz sowie den Zugang zu Informationen voraussetzen, wodurch spezifische Bevölkerungsgruppen benachteiligt werden. Für Nudging durch Ausnutzen sozialer Normen und den Einsatz von gesellschaftsbekannten Personen zur Verstärkung der Botschaft seien als Beispiel die Kampagne zur HIV-Prävention der Bundeszentrale für gesundheitliche Aufklärung mit Hella von Sinnen und Ingolf Lück als Darstellern genannt sowie die weniger nachhaltig wirkende Kampagne zur Darmkrebsvorsorge der Felix-Burda-Stiftung, für die u a. der Entertainer Harald Schmidt als Botschafter wirkte. Zu prüfen ist, ob der Vergleich mit den „Botschaftern“ und den durch sie vermittelten Normen viele Menschen

[5] Salienz, von latein. salire: springen.

erreicht und zur Nachahmung animiert, und ob das Machtgefälle zwischen Sender und Empfänger berücksichtigt wurde (Blumenthal-Barby und Burroughs 2012).

Thaler und Sunstein wollen den von ihnen eingeführten Begriff des Nudging als Ausdruck eines libertären Paternalismus verstanden wissen, und in ihren eigenen, logisch kaum nachvollziehbaren, Worten über den nur scheinbar wie ein Oxymoron klingenden Terminus, vertreten sie die Ansicht, dass die Freiheit des Einzelnen beim Nudging bewahrt bleibt. Linehan et al. merken dazu kritisch an, dass Vertreter des Neoliberalismus den Behauptungen von Thaler und Sunstein gerne Glauben schenken und mit ihnen der Meinung sind, „dass private Institutionen, Behörden und Regierungen bewusst versuchen [sollten], die Entscheidungen der Menschen so zu lenken, dass sie hinterher besser dastehen" (Linehan et al. 2015; Thaler und Sunstein 2008). Sie kritisieren, dass in dieser neoliberalen Auffassung Probleme der Gesellschaft von ihr selbst gelöst werden sollen, statt von kompetenten, fairen und demokratischen Politikern auf lokaler und nationaler Ebene (Linehan et al. 2015). In paternalistischem Duktus sprechen Thaler und Sunstein Bürgern ihre Mündigkeit ab: „Wir werden […] zeigen, dass Menschen in vielen Situationen ziemlich schlechte Entscheidungen treffen – Entscheidungen, die sie nicht treffen würden, wenn sie richtig aufgepasst hätten" (Thaler und Sunstein 2008).

4 Schlussbetrachtungen

Das Ziel, menschliches Verhalten ändern zu wollen, eint die persuasive Technologie Gamification und die Praxis und Wissenschaft von Public Health. Unterschiede liegen in der zum Teil fragwürdigen Intention von Gamification (siehe DSoG und Abschn. 3.1) sowie in der mangelnden Belastbarkeit der Evidenz zu Nudging hinsichtlich einer Verbesserung des Gesundheitsverhaltens und schwachen Evidenz zu *Serious Games for Health* (Marteau et al. 2011). Darüber hinaus kommen nahezu alle in diesem Beitrag zitierten Autoren zu dem Schluss, dass Forschungslücken und Forschungsbedarf groß sind.

Ein Großteil der Anwendungen von Gamification im Gesundheitsbereich zielt auf eine Steigerung der körperlichen Aktivität. Koivisto und Hamari untersuchten den Effekt von Gamification im Rahmen des online-Spiels und sozialen Netzwerks „Fitocracy" zur Verbesserung der Fitness. In ihrer auf Alter, Geschlecht und Dauer der Nutzung konzentrierten Studie kamen sie nach Auswertung des Surveys in 23 Ländern mit 195 Teilnehmerinnen und Teilnehmern zu dem Ergebnis, dass Alter kaum eine Rolle spielt. Die Autoren fanden allerdings Hinweise darauf, dass sich die Bewertung der Nutzerfreundlichkeit mit zunehmendem Alter

verschlechtert. Frauen berichteten insbesondere in Bezug auf Anerkennung und die soziale Gemeinschaft häufiger von einem Nutzen des Systems, und verglichen mit den männlichen Teilnehmern bewerteten sie die Unterhaltsamkeit des Gesamterlebnisses positiver. Dauer der Nutzung, wahrgenommener Nutzen, Spaß und Spielfreude nahmen in allen Nutzergruppen mit der Zeit ab (Koivisto und Hamari 2014).

Mit Gamification sind auch sogenannte Fitness-Tracker und Wearables zur Förderung von Gesundheits- und Wellness-Aktivitäten ausgestattet, die in Verbindung mit Smartphone-Apps von den Herstellern als vielversprechende Instrumente zur Steigerung der körperlichen Aktivität ihrer Nutzer angepriesen werden (Arora und Razavian 2021). In ihrer im Jahr 2019 publizierten Studie mit 210 Nutzern von Fitness-Trackern kommen Attig und Franke jedoch zu dem Schluss, dass solche Funktionen die Nutzer kaum zu einem aktiven Lebensstil oder Sport bewegen können, und empfehlen, dass sich Entwickler und Designer stattdessen darauf konzentrieren sollten, die intrinsische Motivation der Nutzer zu unterstützen (Attig und Franke 2019). Vergleichbare Ergebnisse präsentieren Laranjo et al. in ihrer großen Meta-Analyse mit 28 eingeschlossenen Studien. Sie konnten nur einen kleinen bis moderaten Effekt im Anstieg der körperlichen Aktivität um durchschnittlich 1850 Schritte pro Tag feststellen, und der gepoolte Effektschätzer war von einer hohen Heterogenität der Studien infolge von zum Teil sehr kurzen Beobachtungsdauern oder geringer Teilnehmeranzahl begleitet (Laranjo et al. 2021).

Angesichts der zahlreichen Menschen, die voller Hoffnung auf eine Verbesserung ihrer körperlichen Gesundheit Wearables tragen, mag mit den Herausgebern der Monatszeitschrift *Le Monde diplomatique* gefragt werden „Do you play games, or are they playing you?“, wenn es nur noch darum geht, die für einen Tag geforderte Anzahl von Schritten zu erreichen, damit man von seinem eigenen Fitness-Tracker nicht mehr ständig daran erinnert wird (Kim und Werbach 2016). In diesem Fall besteht die Gefahr, dass extrinsische Motivation die initial intrinsische Motivation ablöst und letztendlich aufhebt, weil von einer „Motivation der Nutzer“, die ausschließlich durch äußere Anreize genährt wird, nicht mehr die Rede sein kann. Zusätzlich kann die Wahrnehmung von extrinsischer Motivation, wie ein Versuchstier wegen zu erwartender Belohnung bestimmte Verhaltensweisen zu zeigen, Widerstand erzeugen und den in Studien gezeigten kurzfristigen Effekt negativ verstärken.

Auch nach der Selbstbestimmungstheorie von Ryan und Deci wird die dem Menschen angeborene Aktivität und Neugier, als die intrinsische Motivation beschrieben werden kann, durch Bedingungen abgeschwächt, unter denen Verhalten kontrolliert und auch die Selbstwirksamkeit als eingeschränkt wahrgenommen

wird; zur Förderung der intrinsischen Motivation dienen dahingegen Bedingungen, unter denen Autonomie und Kompetenz unterstützt werden (Ryan und Deci 2000). Wenig spricht also dafür, auf eine langfristige Wirkung von Gamification zur Verhaltensänderung zu hoffen, die auf extrinsische Motivation ausgerichtet ist.

Körperliche Aktivität und Ernährung, die auch mit Konsumverhalten verknüpft sind, sind wesentliche Determinanten von Gesundheit. In ihrem Artikel „Judging nudging" merken Marteau et al. dazu kritisch an, dass Gesundheit in unserer Gesellschaft ein hoher Wert beigemessen wird, wir uns aber keineswegs dementsprechend verhalten (Marteau et al. 2011). Ferner weisen sie darauf hin, dass wir gemäß unserer dualen Prägung einem reflektierenden, zielorientierten Denken auf der Basis von Werten und Absichten folgen, andererseits aber durch ein affektives System, angetrieben von Gefühlen und ausgelöst von der Umwelt, gesteuert werden. Da unsere kognitiven Kapazitäten jedoch begrenzt sind und wir zusätzlich durch Werbung und Marketing beeinflusst werden, machen unsichere und weit entfernte Belohnungen ungesundes Verhalten wahrscheinlicher (Marteau et al. 2011).

Wie auch von Forberger et al. konstatiert, sind die Möglichkeiten von Nudging in der Praxis von Public Health bei weitem nicht ausgeschöpft (Forberger et al. 2019). Zwar hat sich die Selbstregulierung durch die Industrie als weniger wirksam erwiesen als die Gesetzgebung, wie das Beispiel des Rauchverbots in öffentlichen Gebäuden zeigt, aber Nudging besitzt auch ohne für seine Implementierung notwendige gesetzliche Änderungen für Politiker eine hohe Anziehungskraft, weil es scheinbar einfache und kostengünstige Lösungen zur Änderung menschlichen Verhaltens bietet und darüber hinaus für zahlreiche Probleme eingesetzt werden kann (Möllenkamp et al. 2019; Marteau et al. 2011). Fraglich bleibt, ob wir diese Art der Steuerung als Gesellschaft wünschen können.

Abschließend ist festzuhalten, dass für die Gestaltung des dargestellten Grenzbereichs von Werbung, Manipulation und Infantilisierung von Bürgerinnen und Bürgern ein zivilgesellschaftlicher Diskurs erforderlich ist, in dem ethische Rahmenbedingungen für Gamification und Nudging festgelegt sowie normative Leitlinien als Orientierung für Wissenschaftler, Praxisakteure und Politiker entwickelt werden.

Literatur

Acheson D. (1988). Public Health in England: The Report of the Committee of Inquiry into the Future Development of the Public Health Function. The Stationary Office, London. https://wellcomecollection.org/works/wa4arbxy. Zugegriffen: 12. Apr. 2023.

Andrade, F. R. H. Mizoguchi, R. & Isotani, S. (2016). The Bright and Dark Sides of Gamification. In: Micarelli A, J. Stamper J, Panourgia, K (Hg.), Intelligent Tutoring Systems. ITS 2016. Lecture Notes in Computer Science Vol. 9684. Cham: Springer. https://doi.org/10.1007/978-3-319-39583-8_17

Arno, A. & Thomas, S. (2016). The efficacy of nudge theory strategies in influencing adult dietary behaviour: a systematic review and meta-analysis. BMC Public Health 16: 676. https://doi.org/10.1186/s12889-016-3272-x

Arora, C. & Razavian N. (2021). Ethics of gamification in health and fitness-tracking. Int J Environ Res Public Health 18, 11052. https://doi.org/10.3390/ijerph182111052

Attig, C. & Franke, T. (2019). I track, therefore I walk – Exploring the motivational costs of wearing activity trackers in actual users. Int J Hum Comput Stud 2019; 127: 211–224. https://doi.org/10.1016/j.ijhcs.2018.04.007

Blohm, I. & Leimeister, J. M. (2013). Gamification. Gestaltung IT-basierter Zusatzdienstleistungen zur Motivationsunterstützung und Verhaltensänderung. In: WIRTSCHAFTSINFORMATIK. https://doi.org/10.1007/s12599-013-0273-5

Blumenthal-Barby, J. S., Burroughs H. (2012). Seeking better health care outcomes: The ethics of using the "Nudge". The American Journal of Bioethics 12(2): 1–10. https://doi.org/10.1080/15265161.2011.634481

Bogost I. (2013). Exploitationware. In: Colby R, Johnson MSS, Colby R. (Hg.) Rhetoric/Composition/Play through Video Games. Palgrave Macmillan's Digital Education and Learning. Palgrave Macmillan, New York. https://doi.org/10.1057/9781137307675_11

Bui, A., Veit, D. & Webster, J. (2015). Gamification – a novel phenomenon or a new wrapping for existing concepts? Thirty sixth international conference on information systems 1–21. https://www.researchgate.net/publication/284033385_Gamification_-_A_Novel_Phenomenon_or_a_New_Wrapping_for_Existing_Concepts#fullTextFileContent. Zugegriffen: 19. Apr. 2023.

Callan, R. C., Bauer, K. N. & Landers, R. N. (2015). How to Avoid the Dark Side of Gamification: Ten Business Scenarios and Their Unintended Consequences. In: Reiners T, Wood L (Hg.), Gamification in Education and Business. Springer Cham. https://doi.org/10.1007/978-3-319-10208-5_28

Childress, J. F., Faden, R. R., Gaare, R. D., Gostin, L. O., Kahn, J., Bonnie, R. J., Kass, N. E., Mastroianni, A. C., Moreno, J. D. & Nieburg, P. (2002). Public Health Ethics: Mapping the terrain. The Journal of Law Medicine & Ethics 30: 170–178. https://doi.org/10.1111/j.1748-720x.2002.tb00384.x

Chou Y. (2017). Actionable Gamification – Beyond Points, Badges, and Leaderboards. Youkai Chou.

Csikszentmihalyi M. (2008). Flow: The Psychology of Optimal Experience. Harper Perennial Modern Classics, New York. https://www.researchgate.net/publication/224927532_Flow_The_Psychology_of_Optimal_Experience#fullTextFileContent. Zugegriffen: 12. Apr. 2023.

DAK-Gesundheit (Ersatzkasse). (o. D.). Ergebnisse der DAK-Studie: Gaming, Social-Media und Corona. https://www.dak.de/dak/gesundheit/fortsetzung-der-dak-studie-gaming-social-media-und-corona-2507354.html/. Zugegriffen: 16. Apr. 2023.

Deterding, S., Dixon, D. & Khaled, R. (2011). Gamification: Toward a definition. Proceedings of CHI 12–15. http://gamification-research.org/wp-content/uploads/2011/04/02-Deterding-Khaled-Nacke-Dixon.pdf. Zugegriffen: 18. Apr. 2023.

Dittmeyer, M. (2020). Der programmierte Mensch. Zur Idee und Ethik von Gamification. Mentis Paderborn.

Filatro, A. & Cavalcanti, C. C. (2016). Structural and Content Gamification Design for Tutor Education. Association for the Advancement of Computing in Education (AACE). E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 1152–1157. https://cdn1.unasp.br/home/2017/11/paper_49942_30741-1.pdf. Zugegriffen: 14. Apr. 2023.

Floridi, L. (2014). The fourth revolution: How the infosphere is reshaping human reality. New York: Oxford University Press.

Fogg, J. B. (1998). Persuasive computers: perspectives and research directions. CHI '98: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems 225–232. https://doi.org/10.1145/274644.274677

Fogg, J. B. (2003). Persuasive Technology: Using Computers to Change What We Think and Do (The Morgan Kaufmann Series in Interactive Technologies). Morgan Kaufmann.

Forberger, S., Reisch, L., Kampfmann, T. & Zeeb, H. (2019). Nudging to move: a scoping review of the use of choice architecture interventions to promote physical activity in the general population. IJBNPA 16(1), 1–14. https://doi.org/10.1186/s12966-019-0844-z

Graafland, M., Schraagen, J. M., Schijven, M. P. (2012). Systematic review of serious games for medical education and surgical skills training. Br J Surg 99 (10); 1322–30. https://doi.org/10.1002/bjs.8819

Haoran, G., Bazakidi, E. & Zary, N. (2019). Serious games in health professions education: review of trends and learning efficacy. Yearb Med Inform 28 (1): 240–248. https://doi.org/10.1055/s-0039-1677904

Heinz, M. & Fischer, H. (2020). Ausgespielt? Zu Risiken und Nebenwirkungen von Gamification. In: Köhler T, Schoop E, Kahnwald N (Hg.), Communities in Media. From hybrid realities to hybrid communities. TUDPress. https://tud.qucosa.de/api/qucosa%3A74132/attachment/ATT-0/?L=1. Zugegriffen: 11. Apr. 2023.

Huotari, K. & Hamari, J. (2012). Defining gamification – a service marketing perspective. In: Proc 15th MindTrek conference, Tampere 17–22. https://doi.org/10.1145/2393132.2393137

Hyrynsalmi, S., Smed, J. & Kimppa, K. K. (2017). The Dark Side of Gamification: How We Should Stop Worrying and Study also the Negative Impacts of Bringing Game Design Elements to Everywhere. In: Tuomi P, Perttula A (Hg.), Proceedings of the 1st International GamiFIN Conference, Pori, Finland, 105–110. http://ceur-ws.org/Vol-1857/gamifin17_p13.pdf. Zugegriffen: 14. Apr. 2023.

Janssen, J., Verschuren, O., Renger, W. J., Ermers, J., Ketelaar, M. & van E. R. (2017). Gamification in physical therapy: more than using games. Pediatr Phys Ther 29 (1): 95–99. https://doi.org/10.1097/pep.0000000000000326

Johnson, D., Deterding, S., Kuhn, K.-A., Staneva, A., Stoyanov, S. & Hides, L. (2016). Gamification for health and well-being: A systematic review of literature. Internet Interventions 6: 89–106. https://doi.org/10.1016/j.invent.2016.10.002

Kato, P. M., Cole, S. W., Bradlyn, A. S. & Pollock, B. H. (2008). A video game improves behavioral outcomes in adolescents and youg adults with cancer: a randomized trial. Pediatrics 122 (2): e305–317. https://doi.org/10.1542/peds.2007-3134

Kim, T. W. & Werbach, K. (2016). More than just a game: ethical issues in gamification. Ethics Inf Technol 18: 157–173. https://doi.org/10.1007/s10676-016-9401-5

Koivisto, J. & Hamari, J. (2014). Demographic differences in perceived benefits from gamification. Comput Hum Behav 35, 179–188. https://doi.org/10.1016/j.chb.2014.03.007

Koivisto, J. & Hamari, J. (2019). The rise of motivational information systems: a review of gamification research. IJIM 45: 191–210. https://doi.org/10.1016/j.ijinfomgt.2018.10.013

Laranjo, L., Ding, D., Heleno, B., Kocaballi, B., Quiroz, J. C., Tong, H. L., Chahwan, B., Neves, A. L., Gabarron, E., Phuong Dao, P., Rodrigues, D., Costa Neves, G., Antunes, M. L., Coirea, E. & Bates, D. W. (2021). Do smartphone applications and activity trackers increase physical activity in adults? Systematic review, meta-analysis and metaregression. Br J Sports Med 55: 422–432. https://doi.org/10.1136/bjsports-2020-102892

Linehan, C., Harrer, S., Kirman, B., Lawson, S. & Carter, M. (2015). Games against health: A player-centered design philosophy. In: Proceedings of CHI EA '15; 589–600. ACM, New York. http://eprints.lincoln.ac.uk/id/eprint/16887/1/alt159-linehan.pdf. Zugegriffen: 16. Apr. 2023.

Lu, A. S., & Kharazzi, H. (2018). A state-of-the-art systematic content analysis of games for health. Games Health J 7 (1): 1–15. https://doi.org/10.1089/g4h.2017.0095

Marteau, T. M., Ogilvie, D., Roland, M., Suhrcke, M. & Kelly, M. P. (2011). Judging nudging: can nudging improve population health? BMJ 342:d228. https://doi.org/10.1136/bmj.d228

Möllenkamp, M., Zeppernick, M. & Schreyögg, J. (2019). The effectiveness of nudges in improving the self-management of patients with chronic diseases: A systematic literature review. Health policy 123: 1199–1209. https://doi.org/10.1016/j.healthpol.2019.09.008

Nicholson, S. (2012). A user-centered theoretical framework for meaningful gamification. Games+Learning+Society 8.0 1–7. https://scottnicholson.com/pubs/meaningfulframework.pdf. Zugegriffen: 18. Apr. 2023.

Nyström, T. (2021). Exploring the Darkness of Gamification: You Want It Darker? In: Intelligent Computing; pp. 491–506. Springer, Cham. https://doi.org/10.1007/978-3-030-80129-8_35

O'Sullivan, D., Stravrakakis, I., Gordon, D., Curley, A., Tierney, B., Murphy, E., Collins, M. & Becevel, A. (2021). "You can't loose a game if you don't play the game": exploring the ethics of gamification in education. IJI 14 (1): 2035–2045. https://doi.org/10.20533/iji.1742.4712.2021.0211

Palmer, M., Sutherland, J., Barnard, S., Wynne, A., Rezel, E., Doel, A., Grigsby-Duffy, L., Edwards, S., Russell, S., Hotopf, E., Perel, P. & Free, C. (2018). The effectiveness of smoking cessation, physical activity/diet and alcohol reduction interventions delivered by mobile phones for the prevention of non-communicable diseases: a systematic review of randomised controlled trials. PLoS One 13(1): e0189801. https://doi.org/10.1371/journal.pone.0189801

Paschke, K., Austermann, M. I., Simon-Kutscher, K. & Thomasius, R. (2021). Adolescent gaming and social media usage before and during the COVID-19 pandemic. Sucht 67 (1): 13–22. https://doi.org/10.1024/0939-5911/a000694

Pereira, P., Duarte, E., Rebelo, F. & Noriega, P. (2014). A review gamification in health-related contexts. Design, user experience, and usability. User experience design for diverse interaction platforms and environments. Lecture Notes in Computer Science, vol 8518. Springer, Cham. https://doi.org/10.1007/978-3-319-07626-3_70

Reddy, A., Huseman, T. L., Canamucio, A., Marcus, S. C., Asch, D. A., Volpp, K. & Long, J. A. (2017). Patient and partner feedback reports to improve statin medication adherence: a randomized control trial. J Gen Intern Med 32: 256–61. https://doi.org/10.1007/s11606-016-3858-0

Reñosa, M. D. C., Landicho, J., Wachinger, J., Dalglish, S. L., Bärnighausen, K., Bärnighausen, T. & McMahon, S. A. (2021). Nudging toward vaccination: a systematic review. BMJ Global Health6:e006237. https://doi.org/10.1136/bmjgh-2021-006237

Ryan, R. M. & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American Psychologist 55: 68–78. https://doi.org/10.1037/0003-066X.55.1.68

Sajeev, M. F., Kelada, L., Yahya, N. A. B., Wakefield, C. E., Wewege, M., Karpelowsky, J., Akimana, B., Darlington, A.-S. & Signorelli, C. (2021). Interactive video games to reduce paediatric procedural pain and anxiety: a systematic review and meta-analysis. Br J Anaesth 127 (4): 608–619. https://doi.org/10.1016/j.bja.2021.06.039

Sardi, L., Idri, A. & Fernández-Alemán, J. L. (2017). A systematic review of gamification in e-health. J Biomed Inform 71: 31–48. https://doi.org/10.1016/j.jbi.2017.05.011

Thaler, R. & Sunstein, C. (2008). Nudge: Improving Decisions About Health, Wealth, and Happiness. Penguin, New York.

Toda, A., Valle, P. H. & Isotani, S. (2018). The dark side of gamification: An overview of negative effects of gamification in education. In: Cristea A, Bittencourt I, Lima F (Hg), Higher Education for All. Cham: Springer. https://www.researchgate.net/publication/326876949_The_Dark_Side_of_Gamification_An_Overview_of_Negative_Effects_of_Gamification_in_Education. Zugegriffen: 12. Apr. 2023.

Tolks, D., Lampert, C., Dadaczynski, K., Maslon, E., Paulus, P., & Sailer, M. (2020). Spielerische Ansätze in Prävention und Gesundheitsförderung. Bundesgesundheitsbl 63:698–707. https://doi.org/10.1007/s00103-020-03156-1

Verweij, M. & Dawson, A. (2007). The meaning of ‘Public’ in ‘Public Health’. In: Verwej M, Dawson A (Hg.), Ethics, prevention, and Public Health. Clarendon Press, Oxford.

Vesely, S. & Klöckner, C. A. (2020). Social desirability in environmental research: three meta-analyses. Front Psychol 11:1395. https://doi.org/10.3389/fpsyg.2020.01395

Vieira, C., Pais-Vieira, C., Novias, J. & Perrotta, A. (2021). Serious game design and clinical improvement in physical rehabilitation: systematic review. JMIR Serious Games 9(3):e20066. https://doi.org/10.2196/20066

WHO. Global health risks: mortality and burden of disease attributable to selected major risks. WHO, 2009. https://www.who.int/publications/i/item/9789241563871. Zugegriffen: 17. Apr. 2023.

Zagal, J. P., Björk, S. & Lewis, C. (2013). Dark patterns in the design of games. In: Proceedings of the Conference on Foundations of Digital Games. https://my.eng.utah.edu/~zagal/Papers/Zagal_et_al_DarkPatterns.pdf. Zugegriffen: 16. Apr. 2023.

PD Dr. Barbara Buchberger, MPH, MPhil Robert Koch-Institut. Barbara Buchberger hat ein Violinstudium an der Hochschule der Künste Berlin absolviert und nach ihrem künstlerischen Abschluss als angestellte und freiberuflich tätige Musikerin gearbeitet. Für das Aufbaustudium Public Health an der Technischen Universität Berlin wählte sie den Schwerpunkt Epidemiologie und Methoden. Einer Weiterbildung in Medizinethik folgte ein Masterstudium der Philosophie an der FernUniversität in Hagen. Bis zum Beginn ihrer Beschäftigung beim Robert Koch-Institut im Jahr 2018 war sie wissenschaftliche Mitarbeiterin und Leiterin der Forschungsgruppe „Health Techology Assessment und systematische Reviews" am Lehrstuhl für Medizinmanagement der Universität Duisburg-Essen. Im April 2021 wurde sie von der Medizinischen Fakultät der Universität Duisburg-Essen für das Fach „Gesundheitsökonomie, Gesundheitssystem, Öffentliches Gesundheitswesen" habilitiert.

Ethische Aspekte des Einsatzes Künstlicher Intelligenz im Rahmen der ärztlichen Tätigkeit

Sabine Salloch

Zusammenfassung

Die Entwicklung und klinische Implementierung von KI-Technologien im Gesundheitswesen ist mit besonderen ethischen Herausforderungen verbunden. So werfen KI-getriebene Entscheidungsunterstützungssysteme etwa Fragen hinsichtlich der ärztlichen Kompetenz, aber auch der Patientenautonomie (z. B. „informed consent") auf, die derzeit weder ethisch noch rechtlich eindeutig geklärt sind. Weiterhin bedeutsam sind (oft implizit vertretene) Perspektiven auf das Mensch-Maschine-Verhältnis bei der Nutzung medizinischer KI. Das weitgehend dominante „kompetitive Bild" des Verhältnisses von Ärzt*innen und Entscheidungsunterstützungssystemen ist mit dem Risiko behaftet, den sinnvollen Einsatz dieser Systeme zum Nutzen der Patient*innen zu behindern. Ethisch zu diskutierende Zukunftsperspektiven ergeben sich derzeit angesichts des Einsatzes großer Sprachmodelle (LLMs), etwa zum Zwecke der Patientenaufklärung. Auch die KI-unterstützte Prädiktion von Patientenpräferenzen bietet in ethischer Hinsicht sowohl Chancen als auch Risiken. Eine umfassende ethische Analyse des Einsatzes von KI im Gesundheitswesen sollte die Systemperspektive sowie auch Fragen der globalen Gerechtigkeit einbeziehen, um schädliche Effekte gering zu halten und gleichzeitig den gesundheitlichen Nutzen für alle relevanten Patientengruppen zu maximieren.

S. Salloch (✉)
Institut für Ethik, Geschichte und Philosophie der Medizin der Medizinischen Hochschule Hannover, Hannover, Deutschland
E-Mail: salloch.sabine@mh-hannover.de

O. Richter et al. (Hrsg.), *Zusammenwirken von natürlicher und künstlicher Intelligenz: Beurteilen-Messen-Bewerten,*
https://doi.org/10.1007/978-3-658-45845-4_11

Schlüsselwörter

Gesundheitswesen • Künstliche Intelligenz • Patientenautonomie • Entscheidungsunterstützung • Mensch-Maschine-Verhältnis

1 KI in der Gesundheitsversorgung

Der Einzug technischer Innovationen in vielfältige Lebens- und Arbeitsumfelder hat seit jeher regelmäßig Umwälzungen mit sich gebracht, die Anpassungen und Neuausrichtungen menschlichen Handelns notwendig machten. Während die Einführung von Maschinen und die serielle Fertigung von Produkten im Rahmen der Industriellen Revolution des 18. und 19. Jahrhunderts vor allem menschliche Arbeit bei einfachen, repetitiven Aufgaben ersetzten, stehen heute zahlreiche wissensintensive Arbeitsgebiete vor der Frage, welche Aufgaben sinnvoll an nicht-menschliche Entitäten delegiert werden können und welche Formen menschliche Aufsicht und Kontrolle weiterhin notwendig sein werden. Verfahren der modernen Informationstechnik und der Datenwissenschaft („Künstliche Intelligenz"; KI) dringen demnach seit einigen Jahren mit rasanter Geschwindigkeit in das „Kerngeschäft" der Professionen ein. Im Bereich der Erziehung und der Schulen zum Beispiel gewinnt nicht nur die Nutzung von online-Material zu Lern- und Prüfungszwecken erhebliche Bedeutung, sondern auch weiterreichende Versuche der Herstellung vollständig virtueller Lernumgebungen werden zügig vorangetrieben (Singh und Sikka 2023). Im Rechtswesen werden heute nicht nur herkömmliche Verfahren der Datenverarbeitung und Tele-Konsultationen eingesetzt, sondern Large Language Models (LLMs) besitzen das Potential, Standard-Schriftsätze in der Rechtspflege zumindest zu entwerfen und für die spätere Überarbeitung durch Fachleute vorzubereiten (Perlman 2022). Im Journalismus stehen Profis zunehmend vor der Frage, welchen Gebrauch sie von den extrem leistungsstarken Werkzeugen der generativen KI machen sollen, um ihre Leser*innen textlich ansprechend, wahrheitsgetreu und effektiv zu informieren. Auch weitere wissensintensive Arbeitsfelder wie Architektur, Steuerfachwesen und Management müssen über den sinnvollen Einsatz technischer Unterstützungssysteme entscheiden, die einzelne Aufgaben korrekter (sowie zumeist ausdauernder und kostengünstiger) bearbeiten können als menschliche Expert*innen. Analyst*innen dieser sogenannten „Vierten Industriellen Revolution" gehen davon aus, dass durch KI disruptive Effekte in zahlreichen Tätigkeitsfeldern entstehen und dass eine Neuausrichtung sowohl in der Ausbildung der menschlichen Akteure als auch in der Verteilung von Arbeit zwischen Mensch und Technik erfolgen muss (Susskind und Susskind 2017).

Auch das Gesundheitswesen bildet einen zentralen Zielbereich in der Entwicklung KI-getriebener Verfahren. Gegenüber anderen Dienstleistungen werden Gesundheitsleistungen häufig als Aufgaben beschrieben, die einer besonderen Sorgfalt und Sensitivität bedürfen, welche sich aus der unmittelbaren Arbeit mit dem menschlichen Körper und aus dem Umgang mit sensiblen personenbezogenen Daten ergibt. Zugleich liegen in Teilbereichen der Gesundheitsversorgung sehr große und teilweise auch standardisierte Datenbestände vor, die grundsätzlich für Verfahren Maschinellen Lernens genutzt werden können. Anwendungen, die auf solchen Daten beruhen, können potentiell für Fortentwicklungen in der Pflege und Versorgung genutzt werden, aber auch die ärztliche Tätigkeit sinnvoll unterstützen. Vor dem Hintergrund der Nutzung von „Big Data" wundert es nicht, dass zu den „Vorreiterfächern" einer datengetriebenen Medizin diejenigen Disziplinen zählen, deren Datenbestände zur Entwicklung entsprechender Verfahren gewissermaßen „einladen". Frühe Entwicklungen zur Nutzung von KI in der medizinischen Diagnostik gab es etwa in der Radiologie, der Pathologie sowie der Ophthalmologie (hier insbesondere die automatisierte Auswertung von Aufnahmen des Augenhintergrundes).

Nun ist die Integration innovativer technischer Verfahren in diagnostische und therapeutische Abläufe im Gesundheitswesen kein grundsätzlich neues Phänomen. Angesichts bahnbrechender Innovationen (etwa der modernen Labor- oder der Röntgentechnik) bedurfte es in der Vergangenheit in der Regel oft längerer Adaptationsprozesse, in denen etwa sinnvolle Formen des Einsatzes sowie Fragen der Entwicklung von Fachkompetenz nach und nach adressiert wurden. Diese Entwicklungen führten letztlich oft zu einer Transformation der Arbeitskontexte (etwa durch Automatisierung oder die Bildung neuer Spezialdisziplinen). Die derzeit zu beobachtende Transformation des Gesundheitswesens durch KI wird allerdings von vielen Beobachter*innen als besonders tiefgreifend erlebt, indem sie dazu beiträgt, dass die historisch entstandenen Rollen von Arzt und Patient einen erheblichen Wandel durchlaufen, dessen Endzustand sich noch nicht vollständig absehen lässt. Vor diesem Hintergrund möchte dieser Beitrag zum Verständnis der ärztlichen Rolle, aber auch der Patientenrolle angesichts eines KI-unterstützten Handelns im Gesundheitswesen beitragen. Das besondere Augenmerk der Analyse wird auf klinischen Entscheidungsunterstützungssystemen und deren ethischen Implikationen liegen.

2 Klinische Entscheidungsunterstützungssysteme: ethische Aspekte

Klinische Entscheidungsunterstützungssysteme können definiert werden als Systeme, die Wissen und personenspezifische Informationen für Gesundheitspersonal, Patient*innen und andere Personen bereitstellen, intelligent auswählen und zu geeigneter Zeit präsentieren, um Gesundheit und Gesundheitsversorgung zu verbessern (vgl. Osheroff et al. 2007). Sie dienen der Unterstützung von Diagnostik, Therapie(planung), Prävention und Prädiktion sowie der Optimierung von Versorgungsstrukturen (Sutton et al. 2020). Neben den schon genannten diagnostischen Fächern Radiologie und Pathologie lassen sich in fast allen klinischen (Sub-)Disziplinen inzwischen Beispiele für eine automatisierte, datengetriebene Unterstützung ausmachen. So haben etwa Liang et al. ein System zur Verbesserung der Diagnosestellung bei pädiatrischen Notfällen entwickelt, das auf Patientendaten aus über 1,3 Mio. elektronischen Krankenakten beruht (Liang et al. 2019). Hannun et al. stellten ein System vor, dass mit großer Zuverlässigkeit Arrhythmien im Elektrokardiogramm (EKG) detektieren kann und dabei mit sogenannten „neuronalen Netzwerken“ arbeitet (Hannun et al. 2019). Ein weiteres sehr aktives Forschungsfeld findet sich im Bereich der Dermatologie, wo die Diagnose des Malignen Melanoms inzwischen sehr effektiv durch KI-Software unterstützt werden kann (Esteva et al. 2017). Auch in der operativen Medizin, etwa bei komplexen Eingriffen in der Mund-Kiefer-Gesichtschirurgie, helfen klinische Entscheidungsunterstützungssysteme bei der Optimierung der Therapieplanung und können (etwa durch Einsatz von „augmented reality“) die Invasivität minimieren und Komplikationen verhindern helfen (Standiford et al. 2022).

Die Optimierung von diagnostischer Genauigkeit und therapeutischer Präzision steht im Interesse der Patient*innen und des Behandlungsteams und ist im Sinne des Wohltuns- und Nichtschadensgebots in ethischer Hinsicht zunächst einmal weitgehend unstrittig. Dennoch hat sich vor dem Hintergrund bereits vorliegender und zukünftig absehbarer Ansätze einer KI-getriebenen klinischen Entscheidungsunterstützung eine ethische Debatte entwickelt, deren wesentliche Stoßrichtungen zu einen im Bereich der Patientenautonomie und einer möglichen Benachteiligung bestimmter Personengruppen und zum anderen in befürchteten negativen Folgen für die ärztliche Berufsausübung liegen (Zentrale Ethikkommission bei der Bundesärztekammer 2021). Die ethische Debatte zu klinischen Entscheidungsunterstützungssystemen ist dabei letztlich als eine Spezialdebatte innerhalb der noch breiteren Diskussion um KI-getriebene Gesundheitsanwendungen insgesamt zu verstehen (Morley et al. 2020).

Hinsichtlich der Patientenautonomie beim Einsatz von klinischen Entscheidungsunterstützungssystemen weist Rosalind McDougall etwa darauf hin, dass in die Empfehlungen dieser Systeme oft implizite Werturteile eingehen, die aber nicht mit den Werthaltungen der betroffenen Patient*innen kongruent sein müssen (McDougall 2019). Sie benutzt das Beispiel des „IBM Watson for Oncology" um darauf hinweisen, dass in der Krebsmedizin die Abwägung zwischen der Länge des Überlebens und der zu erwartenden Lebensqualität oft keinesfalls trivial ist. Diese ist vielmehr „wertesensitiv", indem sie von den Präferenzen der einzelnen Patientin abhängt. Während KI-getriebene Medizin grundsätzlich personalisierte Empfehlungen verspricht, ist diese Personalisierung in der Regel biologisch gemeint, d. h. sie bezieht sich nicht auf individuelle Werthaltungen und Präferenzen, die jedoch ebenfalls entscheidend für die Auswahl der besten Therapie sein können. McDougall warnt vor diesem Hintergrund vor einem „Computer-Paternalismus", der oberflächlich dem Wohle der Patientin dient, tatsächlich jedoch deren Werthaltungen ignoriert. Aus anderer Perspektive lässt sich dieser kritische Einwand allerdings relativieren. In direkter Auseinandersetzung mit McDougall hat Ezio Di Nucci darauf hingewiesen, dass die Empfehlung eines automatisierten Systems nicht mit der Umsetzung dieser Empfehlung gleichzusetzen sein. Vielmehr müssten die KI-generierten Empfehlungen in sinnvoller Weise Eingang in die gemeinsame Entscheidungsfindung zwischen Ärzt*in und Patient*in finden, wobei auch psychosoziale Aspekte und Patientenwünsche zu beachten sind. Es handele sich hier also um evidenzbasierte Empfehlungen, die ebenso wie andere Faktoren (z. B. Studienergebnisse oder diagnostische Tests) in den Abwägungsprozess über die beste Therapieform eingehen müssten (Di Nucci 2019). Dieser Beschreibung von Di Nucci kann insofern gefolgt werden, als die KI-generierten Empfehlungen, die IBM Watson für Oncology bereitstellt, tatsächlich keine ärztliche Entscheidung ersetzen, sondern zusammen mit anderen Faktoren im Rahmen des Shared Decision Making berücksichtigt werden können. Unterschiede zu herkömmlichen diagnostischen Verfahren und konventionellen Methoden der Datenverarbeitung ergeben sich allerdings dahingehend, dass die wertenden Verfahren der KI regelmäßig selbst oft weitreichende „Entscheidungen", etwa über Modelle und Trainingsverfahren, treffen. Die Outputs der KI sind damit letztlich weniger durch den Menschen kontrolliert und weniger einsehbar als diejenigen Verfahren, die der herkömmlichen medizinischen Diagnostik zugrunde liegen.

Weitere Aspekte der Patientenautonomie betreffen die Frage, ob Patient*innen den Einsatz von KI in ihrer Diagnostik und Therapie ablehnen dürfen. Hierzu müssen sie selbstverständlich über das zur Diskussion stehende System angemessen informiert werden. Eine „blinde" Ablehnung widerspricht dem Standard

des informed consent („informierte Einwilligung"), den wir in weiten Teilen der modernen Medizin zugrunde legen. Ploug und Holm argumentieren, dass Patient*innen grundsätzlich das Recht eingeräumt werden sollte, KI-unterstützte Diagnostik und Therapie abzulehnen. Ihre kritische Positionierung stützt sich auf a) die ärztliche Rolle im Umgang mit Patientenpräferenzen, b) Probleme von Verzerrung („bias") und Undurchschaubarkeit („opacity") der KI-Empfehlungen sowie c) gesellschaftliche Effekte der Nutzung von KI in der Gesundheitsversorgung (Ploug und Holm 2020). Konkret gehen Ploug und Holm davon aus, dass KI-Systeme nicht in der Lage sind, die Präferenzen von Patient*innen zu ihrer medizinischen Behandlung angemessen zu erfassen. Zudem deute sich bereits jetzt an, dass unzureichende oder verzerrte Trainingsdaten zu Verletzungen ethischer Gebote wie Gerechtigkeit und Gleichheit führen können. Dies kann dazu führen, dass bestimmte (Alters-)Gruppen bei der Anwendung der Systeme Nachteile erleiden. Schließlich weisen Ploug und Holm darauf hin, dass der Einsatz von KI-Systemen langfristige schädliche Folgen haben könnte, zu denen etwa der Fähigkeitsverlust („de-skilling") oder ein Mangel an menschlicher Fürsorge zählen könnten.

Aus grundsätzlicher Perspektive ist der Position von Ploug und Holm insofern zuzustimmen, als Patient*innen generell aufgeklärt werden müssen und den Einsatz diagnostischer und therapeutischer Methoden ablehnen können. Es wäre höchst unplausibel davon auszugehen, dass KI-getriebene Entscheidungsunterstützungssysteme hier eine Ausnahme bildeten. Im Detail jedoch fällt auf, dass viele der genannten Systeme sehr stark in informationstechnische Arbeitsabläufe integriert sind. Zudem ist es auch für Expert*innen nicht einfach festzulegen, ab wann „KI im Spiel ist" und wo es sich um herkömmliche Verfahren der Datenverarbeitung handelt. Auch sind denkbare Anwendungskontexte in der Gesundheitsversorgung vielfältig und reichen von einer Optimierung der elektronischen Patientenakte oder das Vorbereiten von Arztbriefen bis zu hoch bedeutsamen Entscheidungen über den Einsatz risikoreicher Verfahren unmittelbar am Körper. Vor diesem Hintergrund erscheint eine grundsätzliche Ablehnung des „Einsatzes von KI" fast unzulässig pauschal. Die Aufklärung über einzelne Systeme hingegen und ihre datenwissenschaftlichen Hintergründe würde enorm hohe Ansprüche an die Kompetenzen (und zeitlichen Ressourcen) von Ärzt*innen und Patient*innen stellen. Und noch ein weiterer Faktor trägt dazu bei, dass die Frage, ob Patient*innen den Einsatz von KI ablehnen können, alles andere als einfach zu beantworten ist: Während derzeit viele der Entscheidungsunterstützungssysteme noch sehr innovativen Charakter haben und sich in der Erprobung befinden, ist damit zu rechnen, dass in einigen Jahren der Nutzen bestimmter

Anwendungen gut belegt ist und sie zur Standardversorgung zählen. Dann wiederum wäre nicht mehr ganz einfach zu rechtfertigen, wenn ein Patient statt der nachgewiesenermaßen besseren, KI-unterstützten Therapie eine veraltete konventionelle Therapie wünscht, die nach wissenschaftlichem Erkenntnisstand weniger erfolgreich ist. Hier zeigt sich zum einen, dass der Einsatz von KI sich gar nicht grundsätzlich von der Einführung anderer medizinischer Innovationen in der Praxis unterscheidet, sowie dass unsere (auch ethische) Bewertung von Technologien stark vom jeweiligen Wissenstand und den Standards der Versorgung abhängt. Es wird weiterhin deutlich, dass eine Einschätzung des Nutzens und der ethischen Risiken des Einsatzes von KI-getriebenen Entscheidungsunterstützungssystemen nicht allein die Technologie, sondern den organisatorischen und sozialen Kontext ihrer Nutzung betrachten muss. Vor diesem Hintergrund kommt Fragen der Mensch-Maschine-Interaktion eine besondere Bedeutung zu.

3 Wettbewerb oder Kooperation in der klinischen Entscheidungsunterstützung?

Schlagzeilen im Kontext digitaler Entscheidungsunterstützung haben in den letzten Jahren vor allem solche empirischen Studien gemacht, die die Leistung von Ärzt*innen bei eng umschriebenen diagnostischen Aufgaben mit der Leistung KI-getriebener Systeme verglichen haben. Eine kurze Durchsicht durch das „Deutsche Ärzteblatt“ der Jahre 2019 und 2020 zeigt dann etwa, dass sehr regelmäßig kurze Meldungen der Form „Künstliche Intelligenz erkennt Melanome zuverlässiger als Dermatologen“ oder „Künstliche Intelligenz diagnostiziert genauer als Kinderärzte“ erschienen. Auch Vortragsankündigungen dieser Zeit hatten nicht selten den Duktus „Besser als der Arzt? Wird die KI uns ersetzen?“. Die Frage nach dem Wettbewerb und der Übernahme ärztlicher Arbeiten durch Maschinen scheint demnach eine hohe Anziehungskraft zu besitzen, sollte aber aus wissenschaftlicher Sicht in einen kritischen Kontext gestellt werden.

Die Anzahl von Studien, die ärztliche Leistungen mit denen von KI vergleichen, ist tatsächlich keinesfalls gering. Liu et al. konnten in ihrer systematischen Übersichtsarbeit aus dem Jahr 2019 82 Studien einschließen, in denen die diagnostische Leistungsfähigkeit von „Deep Learning“-Modellen mit derjenigen von Gesundheitspersonal verglichen wurde. Gegenstand der Studien war dabei ausschließlich der Bereich der diagnostischen Bildgebung (Liu et al. 2019). Die Autor*innen fassen zusammen: „Our review found the diagnostic performance of deep learning models to be equivalent to that of health-care professionals.“ (Liu et al. 2019) Zugleich weisen sie aber darauf hin, dass wenige Studien extern

validiert wurden und dass die Qualität der Berichterstattung dieser Studien („reporting quality") oft unzureichend ist. Auch wurden den Deep Learning-Modellen und den menschlichen Bewerter*innen oft unterschiedliche Materialien vorgelegt, was die Aussagekraft der Ergebnisse einschränkt.

Als Limitation der Aussagekraft solcher vergleichenden Studien ist weiterhin zu berücksichtigen, dass die „ökologische Validität" dieser Art von Forschung (also ihre Gültigkeit im Alltagsgeschehen) sehr begrenzt ist. Schlussfolgerungen der Art, dass Maschinen „besser seien" als Ärzt*innen lassen sie nicht zu, da hier nicht ärztliches Handeln, sondern nur sehr isolierte Aufgabenstellungen abgebildet werden. Die Tätigkeit einer Dermatologin beschränkt sich eben nicht auf das Beurteilen von Fotos von Hautläsionen, sondern umfasst zahlreiche weitere Aspekte (z. B. Anamnese, körperliche Untersuchung, Beratung), die die Systeme nicht leisten und auch nur eingeschränkt unterstützen können. Auf grundsätzlicherer Ebene muss zudem gefragt werden, ob die Anlage der Studien – also der „Wettbewerb" zwischen Mensch und Maschine – zuträglich für das letztendliche Ziel einer Verbesserung der Gesundheitsversorgung ist.

Eben diesen Punkt stellen Tupasela und Di Nucci in Frage, wenn sie bezweifeln, dass die Übereinstimmung zwischen menschlichen und maschinellen Beurteilern ein geeignetes Kriterium ist, um die Qualität eines Entscheidungsunterstützungssystems zu beurteilen (Tupasela und Di Nucci 2020). Sie diskutieren dies am Beispiel des Systems „Watson for Oncology". Die Firma IBM als Entwicklerin nimmt regelmäßig Bezug darauf, dass die Ergebnisse ihres Systems in hohem Maße mit dem Urteil führender US-Onkolog*innen übereinstimmen, und möchte damit die exzellente Qualität ihres Produktes demonstrieren. Anhand eines einfachen Gedankenexperiments, das vier denkbare Szenarien durchspielt, versuchen Tupasela und Di Nucci nun zu zeigen, dass Übereinstimmung („concordance") kein geeignetes Kriterium zum Beleg der Qualität eines Entscheidungsunterstützungssystems ist. Im ersten Szenario würden System und Ärzt*innen in ihrem Urteil übereinstimmen und beide ausgezeichnete Empfehlungen im Sinne der Patient*innen abgeben. Das System würde hier die Handlungspraxis nicht verbessern. Im zweiten Szenario urteilen Ärzt*innen und System wiederum gleichlautend, treffen aber beide schlechte Entscheidungen – auch hier resultiert keine Verbesserung der Versorgungsqualität. Im dritten Szenario entscheiden die Ärzt*innen suboptimal; das System hingegen macht ausgezeichnete Empfehlungen. Dieses dritte Szenario bildet im Grund den „Optimalfall" mit Potential zur Verbesserung der Versorgungspraxis ab. Es wirft aber auch viele Fragen auf: Werden die Ärzt*innen ihre Handlungspraxis anpassen und gegen die eigene Überzeugung entscheiden? Wo liegt in diesem Fall die Verantwortung? Und müssten die IBM-Entwickler*innen – zufolge der eigenen

Logik – nicht das System den Ärzt*innen anpassen und damit wiederum die Behandlungsqualität gefährden? In einem vierten Szenario schließlich stimmen System und Ärzt*innen nicht überein und die menschlichen Beurteiler treffen ausgezeichnete Entscheidungen, während das System „irrt". Auch hier ist nicht erkennbar, wie sich die Qualität der Versorgung durch den Einsatz des Systems verbessern sollte.

Das von Tupasela und Di Nucci vorgeschlagene Gedankenexperiment lässt es tatsächlich zweifelhaft erscheinen, dass die Übereinstimmung zwischen Mensch und System ein geeignetes Kriterium für die Beurteilung einer Verbesserung der Versorgungsqualität ist, um die es letztlich bei medizinischen Innovationen gehen soll. Stattdessen besteht zusätzlich die Gefahr, dass Entscheidungsunterstützungssysteme die oft suboptimale Qualität medizinischer Praxis fortzuführen, wenn diese in den Trainingsdaten dokumentiert ist und nicht kritisch hinterfragt wird. Nicht zuletzt weil Menschen *andere* Fehler machen als Maschinen bietet es sich an, anstelle des weit verbreiteten „Wettbewerbs-Designs" vergleichende Studien dergestalt anzulegen, dass die ärztliche Leistung mit Entscheidungsunterstützungssystem mit der Leistung ohne eine solche Unterstützung verglichen wird. Dies würde die in der Praxis de facto im Vordergrund stehende Frage adressieren, ob die Versorgung unter Einsatz dieser Systeme besser wird oder nicht. Es würde dann nicht menschliche mit künstlicher Intelligenz verglichen, sondern – etwas nüchterner – der Einsatz unterstützender Systeme im klinischen Einsatz evaluiert.

Sehr wünschenswert wäre vor diesem Hintergrund weiterführende Forschung, die untersucht, welche Fehler menschliche Akteure bei einer umschriebenen klinischen Aufgabe typischer Weise machen und welche Fehler beim Einsatz von KI zur Lösung derselben Aufgabe drohen. Die Forschung könnte dazu dienen im Sinne der Patientensicherheit optimale Kombinationen des Einsatzes von menschlicher und künstlicher Intelligenz zu entwickeln, die die Stärken und Schwächen beider „Partner" angemessen berücksichtigt. Ein solches Vorgehen würde sich am Ziel einer qualitativ hochwertigen und sicheren Patientenversorgung orientieren und steht erneut in deutlichem Kontrast zum Design derjenigen Untersuchungen, die einen Wettbewerb zwischen KI und Ärztin inszenieren.

Das „kompetitive" Bild der Arzt-Maschine-Interaktion birgt eine Reihe von Gefahren, zu denen etwa Sorgen und Ablehnung auf Seiten der Behandler*innen zählen. Unnötiger Weise steht oft die Frage im Vordergrund, ob Maschinen denn zukünftig Menschen ersetzen könnten, ihnen sogar gewissermaßen die Arbeitsplätze „wegnähmen". Diese Vorstellung ist insofern nicht naheliegend, als wir hier von Entscheidungs*unterstützung* sprechen und schon rechtlich die Verantwortung bei menschlichen Akteuren (im medizinischen Kontext in der Regel beim

Arzt) liegt. Die Idee einer Konkurrenz zwischen Mensch und Maschine könnte auch die sinnvolle Integration der Entscheidungsunterstützung in automatisierte klinische Workflows behindern, wenn die Maschine zu stark als eigenständiger Akteur wahrgenommen wird. Die Annahme, eine Maschine könne Ärzt*innen ersetzen, wirft zuletzt auch ein Schlaglicht auf die dahinterstehende Wahrnehmung ärztlicher Arbeit. Tatsächlich stellt sich die ärztliche Tätigkeit in ihrem ganzheitlichen Umgang mit der Patientin als deutlich komplexer dar als diejenigen Leistungen, die von Maschinen bereits erbracht werden können. Ärztliches Handeln beschränkt sich eben nicht auf die Beurteilung von Bildern oder isolierten Befunden, sondern erfordert eine umfassende Interaktion mit der Patientin unter Einschluss psychologischer, sozialer und kultureller Aspekte.

Die ethische Debatte um die Unterstützung ärztlicher Arbeit durch KI-getriebene Systeme oder sogar die Ersetzung von Menschen durch Maschinen ist bereits jetzt vielfältig und durch eine große Spannbreite von Positionen gekennzeichnet. Einige Autor*innen gehen davon aus, dass unter der Voraussetzung einer erfolgreich fortschreitenden Entwicklung der Technologie ein „replacement" in bestimmten Aufgabenfeldern denkbar wäre (Meier et al. 2022). Konkret diskutiert werden sehr spezielle Funktionen wie etwa Zweitmeinungen, die dann nicht mehr von einem Arzt, sondern von einem automatisierten System abgegeben würden (Kempt und Nagel 2022). Prinzipiell sind hier quantitativ bedeutsame klinische Kontexte denkbar, etwa das Mammographiescreening, bei dem aus Gründen der Qualitätssicherung in der Regel eine ärztliche Zweitbeurteilung vorgesehen ist. Andere Autor*innen sind jedoch grundsätzlich eher skeptisch gegenüber einem Ersatz ärztlich-diagnostischer Arbeit durch KI-getriebene Entscheidungsunterstützung (Taylor-Phillips und Freeman 2022). Neben der Frage nach der Qualität und Verlässlichkeit der automatisierten Empfehlungen dürfen auch komplexe soziale Folgen und Auswirkungen auf die ärztliche Berufsausübung nicht außer Acht gelassen werden. Mit dem zunehmenden Einsatz automatisierter Entscheidungsunterstützung könnte es zu einem Fähigkeitsverlust („de-skilling") bei Ärzt*innen kommen, da nicht mehr die Notwendigkeit besteht, dass bestimmte Kompetenzen – etwa in der Beurteilung bildgebender Verfahren – erworben und trainiert werden. Zudem besteht das auch aus anderen Kontexten bekannte Risiko eines ungerechtfertigten und überhöhten Vertrauens in automatisierte Empfehlungen („automation bias"), das zu einer unkritischen Umsetzung von KI-Empfehlungen und möglicherweise schädlichen Auswirkungen auf die Patientenversorgung führen kann. Aus ethischer Sicht schließlich ist hervorzuheben, dass die scheinbare Neutralität, die wir oft mit den Empfehlungen einer Maschine verbinden, darüber hinwegtäuscht, dass in deren Design und Funktion

oft bereits Werturteile einfließen, die einen erheblichen Einfluss auf den Inhalt der Empfehlungen haben können.

4 Patientenautonomie und KI

Insofern entstehen durch die zunehmende Durchdringung der klinischen Praxis durch KI-getriebene Systeme sowohl erwünschte als auch unerwünschte Effekte, nicht zuletzt im Hinblick auf die Patientenautonomie. Die Berücksichtigung großer Datenbestände, etwa in der digitalen Pathologie oder beim Erstellen onkologischer Therapieempfehlungen, verspricht in vielen Fällen eine Personalisierung, indem einzelne, z. B. genetische Merkmale des Patienten berücksichtigt und in ihrer prognostischen Bedeutung interpretiert werden können. Allerdings ist diese Form der Personalisierung bisher fast ausschließlich biologisch zu verstehen. KI-getriebene Systeme sind kaum in der Lage, etwa psychosoziale oder kulturelle Besonderheiten von Patient*innen angemessen abzubilden. Dies kann im Einzelfall auch zur Fehlbewertung bestimmter Sachverhalte führen. Ebenso gibt es bisher nur wenige Ansätze dahingehend, Patientenpräferenzen (etwa zur Risikoaffinität oder zur Abwägung zwischen Lebensqualität und Länge des Lebens) in die automatisierte Entscheidungsfindung zu integrieren.

Ein Feld, auf dem die Diskussionen um den Zusammenhang zwischen medizinischer KI und Patientenautonomie derzeit an Fahrt gewinnt, ist zudem die Voraussage der Präferenzen von Patient*innen, die ihren Willen nicht mehr selbst äußern können. Eine solche „patient preference prediction“ wird derzeit nur in ersten Szenarien diskutiert, wirft aber spannende Fragen in technischer, versorgungspraktischer und normativ-ethischer Hinsicht auf (Rid und Wendler 2014; Biller-Andorno und Biller 2019). Als Datengrundlage für die – bisher hypothetische! – Prädiktion von Patientenpräferenzen kommen grundsätzlich ganz unterschiedliche Quellen in Frage. Denken könnte man etwa an sozialdemographische Charakteristika, die mit klinischen Verläufen (etwa dokumentiert in elektronischen Patientenakten) verknüpft werden. Potentiell wäre auch denkbar andere, stärker individuumsbezogene Informationen einzubeziehen, etwa Social Media-Content oder andere schriftliche Äußerungen, die die Person hinterlassen hat, bevor sie ihre Selbstbestimmungsfähigkeit verloren hat. Zu diskutieren ist weiterhin, in welchen klinischen Situationen eine KI-getriebene „patient preference prediction“ sinnvoll eingesetzt werden kann. In vielen Fällen versprechen herkömmliche Verfahren der Entscheidungsfindung bei nicht-selbstbestimmungsfähigen Patient*innen (etwa die Rekonstruktion des Patientenwillens über Angehörige oder Vorausverfügung) *prima facie* ein besseres

Ergebnis als automatisierte Formen der Prädiktion. In Situationen jedoch, wo kaum oder gar keine Information zu den Wünschen der Patientin zur Verfügung stehen, könnte die datenbasierte „patient preference prediction" zumindest einen ersten Anhalt hinsichtlich der Präferenzen geben. Weitere Fragen stellen sich jedoch in ethisch-theoretischer, etwa ob wir bei dieser Form der Entscheidungsfindung noch von „Gründen" in einem anspruchsvollen Sinne sprechen können oder ob die Patientenautonomie bei der hypothetischen Nutzung solcher Systeme gewahrt bliebe (Jardas et al. 2022). Denn auch die Grenzen einer KI-getriebenen „patient preference prediction" liegen auf der Hand: Die Präferenzen von Personen, die hinsichtlich ihres Werteprofils typisch für die Referenzgruppe (z. B. hinsichtlich Alter, Geschlecht und Wohnort) sind, könnten voraussichtlich gut abgebildet werden, während Menschen, die Werthaltungen haben, die nicht der statistischen Norm entsprechen, durch die Systeme voraussichtlich nicht gut repräsentiert wären. Zudem kann argumentiert werden, dass die Qualität ethischer Entscheidungen bei Nichteinwilligungsfähigen sich nicht auf deren Ergebnis (im Sinne „wahr" oder „falsch") reduzieren lässt, sondern ganz wesentlich den Prozess dieser Entscheidung beinhaltet. Vor diesem Hintergrund könnten Gründe benannt werden, welche die Einbeziehung von nahestehenden Personen und das Streben nach Konsens und einer für alle tragfähigen Lösung als zentrale Faktoren in der Entscheidungsfindung hervorheben (Benzinger et al. 2023).

Im Überblick über ethische Fragen in der Schnittmenge von KI-getriebener Entscheidungsunterstützung und Patientenautonomie bleibt abschließend noch auf ein sehr neues, aber potentiell hoch relevantes Anwendungsfeld hinzuweisen: die Nutzung von großen Sprachmodellen (Large Language Models, LLMs) zur Einholung der informierten Einwilligung (informed consent). Der informed consent hat sich in der zweiten Hälfte des 20. Jahrhunderts als rechtlicher und ethischer Standard sowohl in der medizinischen Forschung mit Menschen als auch in der klinischen Versorgung etabliert. Gemäß dem gängigen Verständnis beruht ein gültiger informed consent auf den drei Elementen der Informationsübermittlung, des Verstehens und der Freiwilligkeit (Beauchamp und Childress 2019). Das Einholen des informierten Einverständnisses ist ein anspruchsvoller und quantitativ bedeutsamer Teil der ärztlichen Tätigkeit. Es verwundert daher nicht, dass spätestens seit der öffentlichen Präsentation extrem leistungsfähiger LLMs (ChatGPT) Ende des Jahres 2022 verschiedene Versuche publiziert wurden, ärztliche Aufklärung durch Sprachmodelle zu unterstützen oder sogar zu ersetzen. So hat ein amerikanisches Forscherteam um Hannah Decker untersucht, wie sich Lesbarkeit, Genauigkeit und Vollständigkeit entwickeln, wenn statt einer Ärztin ein LLM-basierter Chatbot die Aufklärung über Risiken, Nutzen und Alternativen bei chirurgischen Standardeingriffen übernimmt (Decker et al. 2023). Da

die Leistung des Chatbots in dieser Studie bei bestimmten Aspekten die Qualität der ärztlichen Aufklärung übertraf, schlussfolgern die Autor*innen, dass LLM-basierte Chatbots, die sinnvoll in die Klinikinformatik eingebunden werden, das Potential haben, die Dokumentation der informierten Einwilligung zu verbessern. Solche Entwicklungen einer personalisierten und interaktiven automatisierten Aufklärung von Patient*innen werden in Zukunft absehbar nicht nur Forschungscommunities der Medizininformatik und Datenwissenschaft, sondern auch Medizinrecht, -ethik und -ökonomie sowie unterschiedlichste Akteure im Gesundheitswesen beschäftigen.

5 KI im Gesundheitswesen: ein Ausblick

Schlaglichtartig konnte in diesem Beitrag dargelegt werden, wie die Einführung datenbasierter Anwendungen in der Gesundheitsversorgung Organisationen und Praktiken nachhaltig verändern wird und vielfach eine Neuorientierung bei professionellen und nicht-professionellen Akteuren notwendig macht. Ärzt*innen sehen sich in ihrer professionellen Rolle durch den Einsatz immer leistungsfähigerer Maschinen herausgefordert und müssen lernen, Grenzen und Risiken der neuen Verfahren einzuschätzen, ohne den Versuchungen eines „kompetitiven Bildes“ der Arzt-Maschine-Interaktion zu erliegen. Vielmehr muss es ganz pragmatisch um die Frage gehen, wie gewährleistet werden kann, dass die neuen technischen Möglichkeiten im Sinne des Patientenwohls und der Behandlungsqualität eingesetzt werden können. Auch Fragen distributiver Gerechtigkeit und des fairen Zugangs zu einer guten Gesundheitsversorgung für alle sozialen Gruppen spielen dabei eine zentrale Rolle. Digitale Alternativen haben dabei durchaus einen ambivalenten Charakter: ist ihr Zusatznutzen wissenschaftlich bewiesen, gilt es sicherzustellen, dass alle Menschen, die potentiell profitieren könnten, einen Zugang erhalten, der dann nicht von Versicherungsstatus oder anderen Fragen der Finanzierung abhängen sollte. Zugleich müssen wir aber verhindern, dass Menschen, die sich eine persönliche Behandlung durch die Ärztin wünschen, mit digitalen Alternativen „abgespeist“ werden. Gerade in Bereichen, in denen wir mit einer erheblichen Unterversorgung zu kämpfen haben – etwa in der Psychotherapie – bestände sonst die Gefahr, dass nur noch privilegierte Personen in den Genuss einer menschlich-therapeutischen Leistung kämen.

Wieder andere Fragen stellen sich, wenn wir die Digitalisierung des Gesundheitssystems aus einer globalen Perspektiven betrachten. Die Frage nach der Alternative zwischen KI-getriebenen Systemen und einer „technikfreien“, ärztlichen Leistung lässt sich nämlich oft nur innerhalb sehr stark entwickelter

Gesundheitssysteme (wie etwa des deutschen) sinnvoll stellen. Bewegen wir uns im Versorgungskontext vieler Schwellen- und Entwicklungsländer, ist die durchgehende Abdeckung mit ärztlicher Versorgung (etwa in den Bereichen Radiologie oder Pathologie) ganz einfach nicht gegeben, so dass die digitalen „Alternativen" hier möglicherweise die einzige Option darstellen, medizinische Diagnostik überhaupt anzubieten. Gleiches gilt für Länder und Regionen, in denen aus geographischen Gründen keine flächendeckende Versorgung in den Spezialdisziplinen sichergestellt werden kann. Vor diesem Hintergrund sind manche der Fragen einer Optimierung von Gesundheitsdienstleistungen durch KI-getriebene Anwendungen gewissermaßen „Luxusfragen", während zugleich die Entwicklung von Lösungen für Kontexte mit deutlicher Unterversorgung ethisch geboten erscheint.

Aus Sicht der Patient*innen eröffnet der Einsatz von KI ebenfalls erhebliche Chancen: Neben einer möglichen qualitativen Verbesserung diagnostischer und therapeutischer Verfahren, haben sie immer öfter die Möglichkeit, auf „digitale Lösungen" zurückzugreifen, die ihnen unmittelbar zur Verfügung stehen und teilweise Aufwand und Kosten ersparen, die mit einer ärztlichen Konsultation verbunden wären. Ein besonders eindrucksvolles Beispiel für diesen Zusammenhang bilden sogenannte „Symptome Checker Apps" (etwa Ada® oder Symptomate®), die einen ersten Eindruck von Charakter und Schwere der Erkrankungen geben, aber keine körperliche oder apparative Diagnostik und auch kein ärztliches Gespräch ersetzen können. Während manche Menschen (auch in nicht-medizinischen Lebensbereichen) gerne auf technische Lösungen zugreifen, sind andere erfahrungsgemäß eher zurückhaltend und schätzen den menschlichen Kontakt. Gerade in einem sensiblen Bereich wie der Gesundheitsversorgung muss vermieden werden, dass Menschen sich zu digitalen Lösungen gedrängt fühlen, denen eine ausschließlich menschliche Betreuung sehr wichtig ist. Zugleich sollte der Zugriff auf digitale Versorgungsangebote denjenigen Menschen offenstehen, die eine computerbasierte Information und Beratung in bestimmten Zusammenhängen bevorzugen.

Wie sich das Arzt-Patient-Verhältnis der Zukunft entwickeln wird angesichts der immer elaborierteren KI-getriebenen Entscheidungsunterstützung im Gesundheitswesen ist eine offene Frage. Es bleibt ganz einfach zu hoffen, dass die Thesen, die Eric Topol in seinem Buch „Deep Medicine" aufstellt, Wirklichkeit werden, nämlich dass die Entlastung der Ärzt*innen von einfachen Tätigkeiten ihnen den Raum eröffnet, sich wieder in einem größeren Umfang einer empathischen Patientenversorgung zu widmen (Topol 2019). Bis dahin gilt es die Entwicklung der KI im Gesundheitswesen auch aus ethischer Sicht kritisch zu

begleiten, um sicherzustellen, dass zentrale Werte wie Patientenwohl, Patientenautonomie und soziale Gerechtigkeit durchgehend ihren Platz finden im Zuge des technologischen Fortschritts.

Literatur

Beauchamp TL, Childress JF. Principles of Biomedical Ethics. 8. ed. Oxford: Oxford University Press; 2019.

Benzinger L, Ursin F, Balke WT, Kacprowski T, Salloch S. Should Artificial Intelligence be used to support clinical ethical decision-making? A systematic review of reasons. BMC Med Ethics. 2023;24(1):48.

Biller-Andorno N, Biller A. Algorithm-Aided Prediction of Patient Preferences - An Ethics Sneak Peek. N Engl J Med. 2019;381(15):1480–5.

Decker H, Trang K, Ramirez J, Colley A, Pierce L, Coleman M, et al. Large Language Model-Based Chatbot vs Surgeon-Generated Informed Consent Documentation for Common Procedures. JAMA Netw Open. 2023;6(10):e2336997.

Di Nucci E. Should we be afraid of medical AI? J Med Ethics. 2019;45(8):556–8.

Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115–8.

Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med. 2019;25(1):65–9.

Jardas EJ, Wasserman D, Wendler D. Autonomy-based criticisms of the patient preference predictor. J Med Ethics. 2022;48(5):304–10.

Kempt H, Nagel SK. Responsibility, second opinions and peer-disagreement: ethical and epistemological challenges of using AI in clinical diagnostic contexts. J Med Ethics. 2022;48(4):222–9.

Liang H, Tsui BY, Ni H, Valentim CCS, Baxter SL, Liu G, et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. Nat Med. 2019;25(3):433–8.

Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health. 2019;1(6):e271-e97.

McDougall RJ. Computer knows best? The need for value-flexibility in medical AI. J Med Ethics. 2019;45(3):156–60.

Meier LJ, Hein A, Diepold K, Buyx A. Algorithms for Ethical Decision-Making in the Clinic: A Proof of Concept. Am J Bioeth. 2022;22(7):4–20.

Morley J, Machado CCV, Burr C, Cowls J, Joshi I, Taddeo M, et al. The ethics of AI in health care: A mapping review. Soc Sci Med. 2020;260:113172.

Osheroff JA, Teich JM, Middleton B, Steen EB, Wright A, Detmer DE. A roadmap for national action on clinical decision support. J Am Med Inform Assoc. 2007;14(2):141–5.

Perlman AM. The Implications of ChatGPT for Legal Services and Society. SSRN. 2022.

Ploug T, Holm S. The right to refuse diagnostics and treatment planning by artificial intelligence. Med Health Care Philos. 2020;23(1):107–14.

Rid A, Wendler D. Treatment decision making for incapacitated patients: is development and use of a patient preference predictor feasible? J Med Philos. 2014;39(2):130–52.
Singh RI, Sikka P. Virtual Learning. Insights and Perspectives. London: Routledge; 2023.
Standiford TC, Farlow JL, Brenner MJ, Conte ML, Terrell JE. Clinical Decision Support Systems in Otolaryngology-Head and Neck Surgery: A State of the Art Review. Otolaryngol Head Neck Surg. 2022;166(1):35–47.
Susskind R, Susskind D. The future of the professions. How technology will transform the work of human experts. Oxford: Oxford University Press; 2017.
Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med. 2020;3:17.
Taylor-Phillips S, Freeman K. Artificial intelligence to complement rather than replace radiologists in breast screening. Lancet Digit Health. 2022;4(7):e478-e9.
Topol E. Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again. New York: Basic Books; 2019.
Tupasela A, Di Nucci E. Concordance as evidence in the Watson for Oncology decision-support system. AI & Society. 2020;35:811–8.
Zentrale Ethikkommission bei der Bundesärztekammer. Entscheidungsunterstützung ärztlicher Tätigkeit durch Künstliche Intelligenz. 2021 [https://www.zentrale-ethikkommission.de/stellungnahmen/entscheidungsunterstuetzung-aerztlicher-taetigkeit-durch-kuenstliche-intelligenz] Zugriff 20.10.2023.

Prof. Dr. Dr. Sabine Salloch Institut für Ethik, Geschichte und Philosophie der Medizin, Medizinische Hochschule Hannover. Sabine Salloch ist Professorin für Ethik und Geschichte der Medizin und leitet das Institut für Ethik, Geschichte und Philosophie der Medizin der Medizinischen Hochschule Hannover. Ausgehend von einem Doppelstudium der Medizin und der Philosophie und Promotionen in beiden Fächern war sie wissenschaftliche Mitarbeiterin an der Ruhr-Universität Bochum sowie Juniorprofessorin und Institutsleitung an der Universitätsmedizin Greifswald. Sie ist Mitglied im Vorstand der Zentralen Ethikkommission bei der Bundesärztekammer und stellvertretende Vorsitzende der Zentralen Ethik-Kommission für Stammzellenforschung.

Intelligente Maschinen – Intelligente Menschen

Wer beeinflusst wen?

Klaus Bengler

Zusammenfassung

Angesichts der technischen Entwicklungen im Bereich der künstlichen Intelligenz und Automation zeichnet sich ein großes Potenzial für neuartige Lösungen im Bereich der Mobilität und Industriearbeit ab. Vor dem Hintergrund des demografischen Wandels könnten Menschen deutlich durch kooperierende oder autonome Fahrzeuge und Roboter entlastet werden. Damit dieses Potenzial gehoben werden kann ist es unbedingt erforderlich das Zusammenspiel menschlicher und technischer Handlung gezielt und mit großer Sorgfalt zu gestalten. Die ergonomische Forschung kann hier wichtige Anhaltspunkte und Gestaltungsempfehlungen liefern, die sich vor allem auf das Bewegungs- und Entscheidungsverhalten dieser mobilen Systeme beziehen.

Schlüsselwörter

Mensch-Maschine-Kooperation • Mensch-Roboter-Kooperation • Automation • Automatisiertes Fahren • Ergonomie

1 Die Ausgangssituation

Technische Machbarkeit und gesellschaftliche Notwendigkeit führen zu einem verstärkten Einsatz automatisierter und autonomer Systeme (International Federation of Robotics 2020). Sie zeichnen sich vor allem durch Eigenbeweglichkeit und

K. Bengler (✉)
Lehrstuhl für Ergonomie (LfE) der TU München, München, Deutschland
E-Mail: bengler@tum.de

O. Richter et al. (Hrsg.), *Zusammenwirken von natürlicher und künstlicher Intelligenz: Beurteilen-Messen-Bewerten,*
https://doi.org/10.1007/978-3-658-45845-4_12

eigene Verhaltensplanung aus. Diese Entwicklung findet parallel im Straßenverkehr im Rahmen der automatisierten Fahrzeugführung, im öffentlichen Raum in Gestalt von Service- oder Lieferrobotern statt. So ist zum Beispiel eine deutliche Zunahme von mobilen Servicerobotern in der Gastronomie zu beobachten. Auch in der Intralogistik findet eine vergleichbare Entwicklung statt. Diese Tendenzen werden sich mit steigenden Zahlen und in weiteren Domänen fortsetzen.

Damit ist eine neue Kategorie technischer Systeme entstanden, die aus verschiedenen Gründen aus Sicht der Ergonomie sehr interessant ist, im öffentlichen Leben und an Arbeitsplätzen – vielmehr um Arbeitsplätze herum.

Diese Systeme zeigen, im Vergleich zu den bisherigen Werkzeugen und technischen Systemen, die entwickelt wurden um Menschen die Arbeit zu erleichtern, ein eigenes und eigeninitiatives Verhalten. Der Begriff Verhalten umfasst vor allem Entscheidungs- und Bewegungsverhalten. Die Herausforderung automatisierter und autonomer Systeme besteht in der Besonderheit, dass Entscheidungen und Bewegungen inhaltlich und zeitlich mit denen menschlicher Akteure und Kooperationpartner abgestimmt werden müssen.

Lange wurde diese Abstimmung im Fall der Automation durch räumliche und zeitliche Trennung der Akteure aus Sicherheitsgründen vor allem im industriellen Kontext umgangen, verhindert oder sogar unterbunden. Automatisierte Systeme werden hier durch Sicherheitszäune von Menschen getrennt (Bortot et al. 2013).

Auch für klassische Werkzeuge, die beispielsweise ab der ersten industriellen Revolution entwickelt und genutzt wurden, um Menschen von körperlicher Arbeit zu entlasten, haben sich Gestaltungsprinzipien etabliert. Hier liegt die Initiative für eine Interaktion auf der Seite des Menschen, der Zustandswechsel des technischen Systems auslöst. Vor allem gilt, dass das technische System vom Menschen jederzeit unterbrochen werden kann.

Durch diese Gestaltungsprinzipien entsteht bezüglich der auszuführenden Aufgaben eine klare und überprüfbare Rollenverteilung zwischen Mensch und Maschine. Guidelines, Richtlinien, Standards und Normen detaillieren diese Prinzipien sie für verschiedenste Lebensbereiche und Technologiesektoren. Dadurch werden sie für Entwicklung und Absicherung zugänglich und stellen auch eine gesellschaftliche Konvention dar.

Sowohl im Fall der Automation als auch des klassischen passiven Werkzeugs handelt es sich nach heutiger Betrachtung um verhältnismäßig transparente Maschinen bezüglich ihrer Funktionalität und Funktionsweise. Sie verfügen zum Teil über erhebliche inhärente Energie, wie beispielsweise Fahrzeuge und Industrieroboter zeigen. Gerade diese Energie wird genutzt, um Menschen zu entlasten. Sie kann im Fehlerfall aber auch zu erheblichen Schäden führen. Daher

existieren für die funktionale Sicherheit eine Reihe von Absicherungsmethoden und zur Herstellung der Gebrauchssicherheit entsprechende ergonomische Gestaltungsanforderungen.

Angesichts des demografischen Wandels wird es sinnvoll und notwendig sein, die Entlastung des Menschen fortzuschreiben. Durch Exoskelette in Form kraftunterstützender Systeme. Diese sollten so gestaltet sein, dass sie akzeptabel sind, zudem förderlich sind und eine echte Entlastungscharakter mit sich bringen. Diese Tatsache ist umso wichtiger als diese Systeme ein Eigengewicht mit sich bringen und damit eine grundsätzliche körperliche Belastung (Monica et al. 2021).

Eine ähnliche Entwicklung ist im Bereich der körperfernen Robotik zu beobachten, beispielsweise in Form autonomer oder automatisierter Fahrzeuge. Durch die Automation der Fahraufgabe soll der Fahrer entlastet werden und wird somit zum Passagier. In Abhängigkeit des Automationsgrades ist es erforderlich, dass die automatisierte Fahrzeugführung dauerhaft überwacht wird, wodurch die ursprüngliche Entlastungswirkung reduziert wird.

Die Beispiele haben gemeinsam, dass technische Systeme in eine unveränderte Umwelt eingeführt werden und sich in enger Nachbarschaft zu Menschen befinden. Bürgersteige, Straßen oder Beschilderung werden für technische Systeme nicht verändert. Somit entstehen komplexe soziotechnische Systeme, in denen sich zur selben Zeit im selben Raum sich Menschen und Maschinen verhalten. Es soll auf Wegen und Plätzen nicht zu Behinderungen oder Kollisionen kommen. Diese Szenarien sind wohlbekannt und werden in der Interaktion zwischen Menschen regelmäßig erfolgreich gelöst. Im Fall der Mensch-Maschine Interaktion stellt sich nach wie vor die Frage: Wer wird wem wann und wie ausweichen?

Zu welchem Zeitpunkt kommunizieren die Interaktionspartner ihre Intentionen, um einen Stillstand oder eine Kollision zu vermeiden und um Geschwindigkeiten, Beschleunigungen und Trajektorien anzupassen?

Welches automatisierte Fahrverhalten wird als komfortabel empfunden und wie möchte der Passagier chauffiert werden?

Wird durch ein Exoskelett der menschliche Bewegungsapparat in seinem willentlichen Bewegungsablauf unterstützt oder wird der Mensch, der das Exoskelett trägt in seinen Bewegungen geführt?

Sehen wir noch natürliche Bewegungen im Fall aktiver Exoskelette oder wird ein Bewegungsablauf technisch induziert bezüglich des Zeitpunkts der Bewegung, der Beschleunigung der Bewegung und des Anhaltepunkts. Der menschliche Körper würde dann vorwiegend zum Halten, Greifen und Heben benutzt.

Nach welchen Regeln und auf Basis welcher Informationen können menschliche und maschinelle Handlungsplanung synchronisiert werden. Eine zu frühe oder zu späte Systemaktion könnte verwirren.

An drei exemplarischen Nutzungsfällen soll die komplexe Mensch Technik Interaktion im Folgenden erörtert werden

- Straßenverkehr
- Mobile Roboter
- Exoskelette

2 Differenzierte Rollenverteilungen Koexistenz – Kooperation – Kollaboration

An diesen Szenarien lässt sich sehr gut zeigen, dass es sinnvoll sein kann die bisherige Aufgabenverteilung zwischen Mensch-Maschine differenzierter zu betrachten und die Art der Interaktion genauer zu beschreiben. Während es in den bisherigen industriellen Revolutionen um die Umverteilung von Aufgaben ging entsteht nun die Möglichkeit der Koexistenz, Kooperation oder Kollaboration zwischen Mensch und Maschine (Schmidtler et al. 2015).

Verglichen mit der reinen Automation sind koexistierende und kooperierende Systeme komplexe Automaten, die ein Verhalten zu einem selbstgewählten Zeitpunkt zeigen können. Ziel dieses Verhaltens kann Entscheidungsunterstützung auf der kognitiven Ebene aber auch robotische Unterstützung auf der motorische Ebene bedeuten. Im Fall der Koexistenz führen Mensch und Maschine in räumlicher Nachbarschaft verschiedene Aufgaben aus, die zeitlich aufeinander folgen. Im Fall der Kooperation werden diese Aufgaben in enger zeitlicher Kopplung ausgeführt. Entsteht eine physische Kopplung zwischen Mensch und Maschine wird eine Aufgabe in Kollaboration bearbeitet.

Auf diesem Weg entstehen sehr leistungsfähige Formen der Mensch Maschine Interaktion, die bisher von der Ergonomie auch immer wieder nachgefragt wurden, nämlich Unterstützungssysteme, die sowohl die körperliche, die kognitive und auch die sensorische Ebene unterschiedlich stark unterstützen können. Technische Systeme, die diesen Leistungsumfang besitzen sind möglich geworden und bringen ein technisch bestimmtes Zeit-, Kräfte- und Entscheidungsverhalten in die Handlungsabläufe ein. Eines der prominentesten Beispiele für Kollaboration sind Exoskelette. Bei denen diese zeitliche Dynamik mit einer motorischen und einer sensorischen Komponente gekoppelt wird, um menschliches Bewegungsverhalten beim Heben und Tragen zu unterstützen (Harbauer et al. 2021).

In manchen Fällen liegt eine sinnvolle Lösung in einem kooperierenden Folgeroboter. Die arbeitende Person muss keinen Wagen mehr schieben, sondern übernimmt die Bahnplanung im öffentlichen Gedrängel. Die Robotik folgt

im richtigen Abstand. Bereits hier ist es eine Herausforderung Abstand und Geschwindigkeit intelligent zu steuern, um als Team wahrgenommen und in der Fußgängerzone oder Werkhalle nicht behindert oder getrennt zu werden.

3 Intelligente Interaktionspartner

Diese Beispiele zeigen, dass diese Taxonomie eine wesentlich differenzierte Betrachtung dieser neuen technischen Möglichkeiten erlaubt.

Für eine erfolgreiche Gestaltung der Mensch Technik, Interaktion ist somit eine klare Zielvereinbarung zu schließen. Welches Ziel soll in welcher Rollenverteilung von Mensch und Robotik erreicht werden?

Dieses Zusammenspiel von Menschen und automatisierten oder autonomen Systemen erfordert eine außerordentlich anspruchsvolle und neuartige Gestaltung bezüglich der Aufenthaltsbereiche und der zeitlichen Abläufe (Müller et al. 2008).

Bezüglich der zeitlichen Abläufe sind je nach Kontext und Dynamik des technischen Systems unterschiedliche Zeitkonstanten zu berücksichtigen. Ausschlaggebend sind aber in allen Fällen die kognitiven und motorischen Prozesse der Handlungsplanung und -ausführung und der Trainiertsheitsgrad des menschlichen Interaktionspartners. Im Fall bewusst geplanter Handlungen und Entscheidungen werden Informationen mehrere Sekunden vor der Handlungsausführung berücksichtigt und damit auch benötigt. Im Fall hochtrainierten, automatisierten menschlichen Verhaltens müssen Informationen im Bereich unter 100 ms vor der Handlungsausführung gegeben werden, um in die Handlungsausführung Eingang zu finden.

Eine ergonomische Gestaltung dieser Interaktion setzt die Kenntnis der Nutzerintention vor dem Beginn von Bewegungen voraus und sollte Bewegungsmuster nicht unterstützen, die Kraftabläufe und Krafteinleitungen erzeugen, die auf Dauer nicht sinnvoll oder gesund sind.

Nutzer von Exoskeletten berichten, dass sie eine Kraftunterstützung erst spüren, nachdem sie zur Bewegung angesetzt haben; Dadurch aber am flüssigen Arbeiten gestört würden und sich einem fremden Bewegungsablauf anpassen müssen.

Ein weiteres Szenario adressiert die Begegnung von Mensch und mobilen Robotern im öffentlichen Raum. In diesem Fall werden die technischen Systeme ersetzend zu Menschen eingesetzt um zum Beispiel Reinigungs- oder Transportaufgaben zu erledigen. Auch diese Anpassung in diesem Szenario hängt außerordentlich vom richtigen Zeitpunkt ab, der ebenfalls darüber entscheidet, welcher der Akteure die Initiative zur Lösung des Szenarios ergreift und

welcher auf diese Aktion reagieren muss. Schwer interpretierbare Bewegungsmuster mobiler Roboter oder automatisierter Fahrzeuge führen zu unnötigem Warteverhalten von Mitarbeitern oder Fußgängern.

May und Baldwin (2009) und Weinbeer et al. (2017) zeigen, dass auch die dauerhafte Überwachung automatisierter Fahrzeuge eine kognitive Belastung darstellt, da der menschliche Erwartungshorizont mit dem Entscheidungsverhalten der Maschine abgeglichen werden muss und abweichendes Verhalten erkannt werden muss, um rechtzeitig einzugreifen. Diese Beanspruchung, die aus der Anforderung der Daueraufmerksamkeit entsteht, kann mittels neurophysiologischer Methoden aber auch im veränderten Blickverhalten beobachtet werden.

Daneben entstehen kognitive Beanspruchungen durch notwendige Umplanungen oder die Hemmung von Entscheidungen aufgrund von Unsicherheit, wenn das Verhalten des technischen Systems nicht oder oder noch nicht interpretiert werden kann. Auch hier treten.

In diesen Fällen kann die Belastung des Menschen durch geeignete Gestaltung reduziert werden und eine flüssige Interaktion mit einem stark unterstützenden technischen System erreicht werden, wenn neben Beschleunigungen und Beschleunigungsverläufen der richtige Zeitpunkt gewählt wird.

Die Beispiele zeigen anhand sehr unterschiedlicher Rollenverteilungen zwischen Mensch und Maschine, wie maschinelles Handeln menschliches Verhalten auf der motorischen und kognitiven beeinflussen und verändern kann.

Ebenso wird deutlich, dass die technischen Systeme ihrerseits auf die Menschen in ihrer Umgebung reagieren müssen, um sie adäquat zu unterstützen und Kollisionen zu vermeiden (Althoff et al. 2011). Dazu ist es erforderlich einen Fußgänger rechtzeitig zu erkennen und zusätzlich seine nächste Bewegung zuverlässig abzuschätzen.

Unter welchen Prämissen wird diese Entwicklung stattfinden?

Es ist wahrscheinlich, dass die gebaute Infrastruktur sich langsamer verändern wird, als die technischen Systeme an Fähigkeiten gewinnen werden.

Hinzu kommt, dass die Eigenschaft, sich zu verändern und sich anzupassen, ein Merkmal von Intelligenz ist. Daher ist es wahrscheinlich, dass Menschen im Zweifelsfall geneigt sind, sich an eine geänderte technische Umgebung oder Interaktionspartner anzupassen. Diese Anpassungsprozesse verlaufen bei Menschen als Lernprozesse langfristig, als Entscheidungsprozesse im Bereich von Sekunden und im motorischen Verhalten durchaus noch schneller. Grundsätzlich ist das auch auf der Seite der KI möglich. Allerdings beginnen sie bei mobilen Systemen aufgrund sensorischer Einschränkungen häufig zu einem späteren Zeitpunkt.

Allgemein ist es möglich diese Veränderungs- und Lernprozesse zu quantifizieren und auch ihren zeitlichen Verlauf zu beschreiben. Veränderungen im Blickverhalten, Entscheidungsprozessen und Bewegungsverhalten können über Zeit- und Positionsmessungen sehr genau nachvollzogen werden. Die aufgaben- und kontextbezogene Akzeptanz von Distanzen kann nach wie vor über Metriken der Proxemik (Hall et al. 1968) beschrieben werden.

Zwei Systeme können sich in Interaktion also außerordentlich schnell abstimmen. Menschen erwarten das zum Teil von technischen Systemen. In manchen Fällen gelingt die menschliche Anpassung schneller und geht nicht vom technischen System aus.

4 Mögliche Gestaltungsparameter

Zwei außerordentlich anpassungsfähige Interaktionspartner treffen aufeinander, wobei unerwünschte und verwirrende Anpassungsvorgänge unbedingt vermieden werden sollen. Eine relevante Frage für die ergonomische Gestaltung ist daher, welchen Regeln die Gestaltung intelligenter Systeme folgen soll, um unerwünschte menschliche Anpassungen zu vermeiden und andererseits erfolgreiche menschliche Anpassung nicht zu durchkreuzen. Bengler et al. (2012) haben in Anlehnung an Hoc (2001) Prinzipien für eine kooperative Interaktion formuliert, in denen die Bedeutung der Intentionserkennung, Funktionsallokation, Interaktionsgestaltung und Kooperationstypen beschrieben wird.

Hinzu kommt, dass die Interaktionen nicht nur kollisionsfrei und effizient, sondern vor allem auf Dauer auch akzeptabel verlaufen sollen (Kaiser et al. 2019).

Für die Interaktionsgestaltung ist es sinnvoll zwischen impliziter und expliziter Kommunikation zwischen den Interaktionspartnern zu unterscheiden. Wobei unter expliziter Kommunikation gesprochene/geschriebene Sprache, grafische Symbole und Zeichen verstanden werden, während die implizite Kommunikation Verhaltenssignale wie zum Beispiel die Eigenbewegung beschreibt (Dey und Terken 2017).

Die Abstimmung von expliziter und impliziter Kommunikation bezüglich Inhalt und Zeitpunkt ist ausschlaggebend wie Versuche von Burns et al. (2019), Ackermann et al. (2019), Dietrich et al. (2020) und Rettenmeier et al. (2020) zeigen. In vielen Fällen zeigt sich, dass im Zweifel die Interpretation der impliziten Inhalte überwiegt. Das bedeutet, dass missverständliches Bewegungsverhalten durchaus zu Verwirrung oder risikoreichem Verhalten führen kann.

Von Reinhardt wurde in verschiedensten Szenarien der Mitteilungswert einer kurzen Rückwärtsbewegung untersucht, die mitteilt „ich gewähre Vorfahrt“ (Reinhardt et al. 2021). Von Fuest die Aussagekraft von Verzögerungen und seitlichem Versatz im Fall automatisierter Fahrzeuge bei Fußgängerbegegnung (Fuest et al. 2020).

Eine extreme Form impliziter Kommunikation findet sich im Fall von Exoskeletten. Auch hier gilt, dass die Abstimmung der Interaktionspartner über Bewegungen erfolgt. Vor allen Dingen sollten die Bewegungssignale des Menschen oder vielmehr die vorbereitenden Aktionen vom technischen System genutzt werden. Entsprechende EMG Signale oder Haltungsveränderungen stellen hier einen wichtigen Input dar. Im Vergleich zu Begegnungsvorgängen (Koexistenz, Kooperation) gelten im Bereich der motorischen Kollaboration außerordentlich hohe Anforderungen an die Wahl des richtigen Zeitpunkts und die zeitliche Synchronisation im Bereich von Millisekunden.

Der kommunikative Wert der Bewegungsgestaltung wurde auch am Beispiel von Servicerobotern gezeigt, die in Baumärkten oder Pflegeheimen bei der Orientierung helfen, indem sie vorausfahren. Angepasst an die Gehgeschwindigkeit des folgenden Menschen geben gezielt gestaltete Abbiegevorgänge rechtzeitig Hinweise, welche Abbiegung an der nächsten Verzweigung genommen wird. Durch diese intuitive Gestaltung wird der Bewegungsablauf des nachfolgenden Menschen wesentlich vorausschauend flüssiger (Reinhardt et al. 2020).

Die Anzeige der Richtung über Displays wirkt im Vergleich dazu eher unbeholfen und führt zu hoher kognitiver Belastung.

Die rechtzeitige und zuverlässige Erkennung von Intentionen ist grundlegend Koexistenz, Kooperation und Kollaboration zu ermöglichen. Erkennungsleistungen von über 90 % und bis zu 10 s vor einer Begegnung sind für eine aussagekräftige implizite Kommunikation im Straßenverkehr zwischen Fahrzeugen (automatisiert und nicht automatisiert) notwendig. Im Kontext der mobilen Robotik sind wesentlich kürzere Zeitabstände erforderlich. Hier sind zwar die Geschwindigkeit der Interaktionspartner wesentlich geringer, aber ihre mögliche Bewegungsdynamik wesentlich höher.

Vor allen Dingen stellt die zuverlässige Erkennung der Kooperationsbereitschaft eines Menschen, der einem Roboter Vorfahrt gewährt nach wie vor ein großes Potential aber auch eine technische Herausforderung dar.

Darüber hinaus sollten die Kommunikationsstrategien und Interaktionsstrategien der verschiedenen technischen Systeme (Serviceroboter, Industrieroboter, Fahrzeuge) konsistent sein, da es zu Transferlernen zwischen den verschiedenen Domänen kommen wird. Es wäre außerordentlich abträglich, wenn Fahrzeuge

unterschiedlicher Hersteller sich gegensätzlich und vor allem nicht nachvollziehbar bewegen und verhalten würden.

Aus Sicherheitsgründen ist die explizite Kommunikation in Form von Displays, Signal- und Warntönen stark ausgeprägt. Bei der zu erwartenden Zunahme robotischer Systeme ist diese Strategie zu überdenken, da sie die Aufmerksamkeit der beteiligten Menschen stark fordert und auch der Signal/Rausch Abstand abnehmen könnte. Eine angenehme Umwelt ist eine Umwelt, die nicht von Geräuschen und technischen Signalen völlig überfrachtet ist.

5 Resümee

An ausgewählten Beispielen wurden das Potenzial und das komplexe Zusammenspiel zwischen Menschen und Robotiken beschrieben.

Wie sollten therapeutische Exoskelette Bewegungsabläufe vorgeben, um Patienten zu mobilisieren?

Wie oft und welchen Systemen werden wir Vorrang gewähren oder ausweichen müssen, um die Vorteile von unterstützenden Robotern genießen zu können.

Wie häufig werden Menschen von einem Unterstützungssystem zu einer Bewegung aufgefordert oder in ihr geführt werden?

Werden Fußgänger spontan ausweichen, wenn sie auf dem Weg zum Einkauf auf dem Bürgersteig einem Lieferroboter oder einem Reinigungsroboter begegnen?

Ist es realistisch zu fordern, dass technische Systeme in jedem Fall ausweichen oder sollten sie prinzipiell dazu in der Lage sein?

Geht es nicht vielmehr darum die Situationen so zu gestalten, dass Ausweichen nicht unangenehm erlebt wird, sondern intuitiv erfolgen kann (Onnasch und Roesler 2019). Die entsprechenden Entwicklungsprozesse sollten im Hinblick darauf angepasst werden. (Siehe hierzu die Ansätze für verschiedene Anwendungsfelder von Backhaus et al. (2018), Schmidtler et al. (2015), Klabunde und Weidner (2018), Ralfs et al. (2022), Harbauer und Bengler (2022) und Babel et al. (2021).

Menschliches Verhalten wird sich im Umfeld automatisierter und autonomer Systeme sehr schnell ändern. Die technischen Systeme werden sich selbstlernend und in ihren technischen Leistungen weiterentwickeln und verändern. In diesem Bewusstsein sollten die Interaktionen in Koexistenz, Kooperation und Kollaboration gestaltet werden.

Es ist also durchaus denkbar, dass in Abhängigkeit des Kontext Menschen dem technischen System den Vorrang gewähren und dies auch nicht als Nachteil

erleben, wenn diese Entscheidung keine aufwändige Verhaltensänderung erfordert und vor allem wenn dadurch die Gesamtinteraktion durch Kooperation einen erfolgreichen Verlauf nimmt, an dem sie erkennbar aktiv beteiligt sind. Es ist noch völlig unklar, ob diese Bereitschaft mit wiederholtem Ausweichen und Abwarten abnimmt.

Darüber hinaus könnte es sinnvoll sein, sich dem Verhalten eines technischen Systems anzugleichen und eine Regel (z. B. eine Höchstgeschwindigkeit) einzuhalten. Technische Systeme könnten hier einen wertvollen Beitrag liefern. Allerdings ist zu untersuchen, welche Faktoren bis hin zu kulturellen Unterschieden dazu beitragen, das Menschen bereit sind, dem normativen Beispiel zu folgen und es nicht als Hindernis zu erleben.

Die Kenntnis der Nutzerintention spielt eine ausschlaggebende Rolle und ein gesellschaftlicher Konsens bis hin zur Regulatorik, um einen Gestaltungsrahmen vorzugeben. Ausschlaggebend ist die Aufgabe, die durch Mensch und Maschine gelöst werden soll zu analysieren und zu verstehen, da in Abhängigkeit der Aufgabenstellung unterschiedliche Gestaltungen erforderlich sein werden, sodass angemessene soziotechnische System entstehen.

Literatur

Ackermann, C., Beggiato, M., Schubert, S., & Krems, J. F. (2019). An experimental study to investigate design and assessment criteria: What is important for communication between pedestrians and automated vehicles? *Applied Ergonomics*, *75*, 272–282. https://doi.org/10.1016/j.apergo.2018.11.002

Althoff, D., Wollherr, D., & Buss, M. (2011). Safety assessment of trajectories for navigation in uncertain and dynamic environments. In 2011 IEEE International Conference on Robotics and Automation (pp. 5407–5412). IEEE.

Babel, F., Kraus, J. M., & Baumann, M. (2021). Development and Testing of Psychological Conflict Resolution Strategies for Assertive Robots to Resolve Human–Robot Goal Conflict. Frontiers in Robotics and AI, 7. https://doi.org/10.3389/frobt.2020.591448

Backhaus N, Rosen PH, Scheidig A, Gross HM, Wischniewski S (2018) "Somebody help me, please?!" interaction design framework for needy mobile service robots. In: 2018 IEEE workshop on advanced robotics and its social impacts (ARSO), IEEE, pp 54–61

Bengler, K., Zimmermann, M., Bortot, D., Kienle, M., & Damböck, D. (2012). Interaction principles for cooperative human-machine systems. it – Information Technology: Methoden und innovative Anwendungen der Informatik und Informationstechnik, 54 (4), 157–164.

Bortot, D., Born, M. and K. Bengler, "Directly or on detours? How should industrial robots approximate humans?," 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Tokyo, Japan, 2013, pp. 89–90, https://doi.org/10.1109/HRI.2013.6483515.

Burns, C. G., Oliveira, L., Thomas, P., Iyer, S., & Birrell, S. (2019). Pedestrian Decision-Making Responses to External Human-Machine Interface Designs for Autonomous Vehicles. In *2019 IEEE Intelligent Vehicles Symposium (IV)* (pp. 70–75). IEEE. https://doi.org/10.1109/IVS.2019.8814030

Dey, D. & Terken, J. (2017). Pedestrian interaction with vehicles: Roles of explicit and implicit communication. In Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (pp. 109–113).

Dietrich, A., Maruhn, P., Schwarze, L., & Bengler, K. (2020). Implicit Communication of Automated Vehicles in Urban Scenarios: Effects of Pitch and Deceleration on Pedestrian Crossing Behavior. In T. Ahram, W. Karwowski, S. Pickl, & R. Taiar (Eds.), *Advances in Intelligent Systems and Computing. Human Systems Engineering and Design II* (Vol. 1026, pp. 176–181). Springer International Publishing. https://doi.org/10.1007/978-3-030-27928-8_27

Fuest, T., Maier, A. S., Bellem, H., & Bengler, K. (2020). How Should an Automated Vehicle Communicate Its Intention to a Pedestrian? – A Virtual Reality Study. In T. Ahram, W. Karwowski, S. Pickl, & R. Taiar (Eds.), *Advances in Intelligent Systems and Computing. Human Systems Engineering and Design II* (Vol. 1026, pp. 195–201). Springer International Publishing. https://doi.org/10.1007/978-3-030-27928-8_30

Hall, E. T., Birdwhistell, R. L., Bock, B., Bohannan, P., Diebold Jr, A. R., Durbin, M., … Kimball, S. T., et al. (1968). Proxemics [and comments and replies]. Current Anthropology, 9 (2/3), 83–108.

Harbauer, C. M., & Bengler, K. (2022). Guidelines für den Einsatz von Exoskeletten an gewerblichen Arbeitsplätzen (Bayerischer Unternehmensverband & Verband der Bayerischen Metall und Elektro e.V. Eds.). https://www.baymevbm.de/baymevbm/ServiceCenter/Forschung-Entwicklung/Forschungsprojekte/Forschungsbericht-Guidelines-für-den-Einsatz-von-Exoskeletten-an-gewerblichen-Arbeitsplätzen.jsp

Harbauer, C. M., Fleischer, M., Nguyen, T., Kopfinger, S., Bos, F., & Bengler, K. (2021). Too Close to Comfort ? A New Approach of Designing a Soft Cable-driven Exoskeleton for Lifting Tasks under Ergonomic Aspects. International Journal of Mechanical Engineering and Robotics Research, 99–106. https://doi.org/10.18178/ijmerr.10.3.99-106

Hoc, J.-M. (2001). Towards a cognitive approach to human-machine cooperation in dynamic situations. International Journal of Human-Computer Studies, 54(4), 509–540. https://doi.org/10.1006/ijhc.2000.0454

International Federation of Robotics. (2020). Executive Summary World Robotics 2020 Service Robots. Retrieved May 20, 2021, from https://ifr.org/img/worldrobotics/Executive_Summary_WR_2020_Service_Robots.pdf

Jonas Schmidtler, Verena Knott, Christin Hölzel, and Klaus Bengler. 2015. Human centered assistance applications for the working environment of the future. Occupational Ergonomics 12, 3 (2015), 83–95.

Kaiser, F. G., Glatte, K., & Lauckner, M. (2019). How to make nonhumanoid mobile robots more likable: Employing kinesic courtesy cues to promote appreciation. Applied Ergonomics, 78, 70–75.

Klabunde, J., & Weidner, R. (2018). Leitfaden für die Gestaltung von Unterstützungssystemen am Beispiel des Rückens: Ansatz, Beispiele und Vorgehensweise. In Technische Unterstützungssysteme, die die Menschen wirklich wollen. Helmut-Schmidt-Universität, 2018.

May, J. F., & Baldwin, C. L. (2009). Driver fatigue: The importance of identifying causal factors of fatigue when considering detection and countermeasure technologies. Transportation Research Part F, 12 (3), 218–224. https://doi.org/10.1016/j.trf.2008.11.005

Monica, L., Draicchio, F., Ortiz, J., Chini, G., Toxiri, S., & Anastasi, S. (2021). Occupational Exoskeletons: A New Challenge for Human Factors, Ergonomics and Safety Disciplines in the Workplace of the Future. In N. L. Black, W. P. Neumann, & I. Noy (Eds.), Proceedings of the 21st Congress of the International Ergonomics Association (IEA 2021) (pp. 118–127, Vol. 222). Springer International Publishing. https://doi.org/10.1007/978-3-030-74611-7

Müller, J., Stachniss, C., Arras, K. O., & Burgard, W. (2008). Socially inspired motion planning for mobile robots in populated environments. In Proceedings of International Conference on Cognitive Systems.

Onnasch L, Roesler E (2019) Anthropomorphizing robots: The effect of framing in human-robot collaboration. In: Proceedings of the human factors and ergonomics Society annual meeting, SAGE Publications Sage CA: Los Angeles, CA, vol 63, pp 1311–1315

Ralfs, L., Hoffmann, N., & Weidner, R. (2022). Approach of a Decision Support Matrix for the Implementation of Exoskeletons in Industrial Workplaces. In T. Schüppstuhl, K. Tracht, & A. Raatz (Eds.), Annals of Scientific Society for Assembly, Handling and Industrial Robotics 2021 (pp. 165–176). Springer International Publishing. https://doi.org/10.1007/978-3-030-74032-0

Reinhardt, J., Boos, A., Bloier, M., & Bengler, K. (2020). Effect of variable motion behavior of a mobile robot on human compliance in human-robot spatial interaction. In 66. Frühjahrskongress der Gesellschaft für Arbeitswissenschaft 2020.

Reinhardt, J., Prasch, L., & Bengler, K. (2021). Back-off: Evaluation of robot motion strategies to facilitate human-robot spatial interaction. ACM Transactions on Human-Robot Interaction, 10 (3).

Rettenmaier, M., Albers, D., & Bengler, K. (2020). After you?! – Use of external human-machine interfaces in road bottleneck scenarios. *Transportation Research Part F: Traffic Psychology and Behaviour*, *70*, 175–190. https://doi.org/10.1016/j.trf.2020.03.004

Weinbeer, V., Baur, C., Radlmayr, J., Bill, J. S., Muhr, T. & Bengler, K. (2017). Highly automated driving: How to get the driver drowsy and how does drowsiness influence various take-over aspects? In 8. Tagung Fahrerassistenz. München.

Professor Dr. Klaus Bengler Lehrstuhl für Ergonomie (LfE), Technische Universität München.

Prof. Bengler studierte Psychologie an der Universität Regensburg und promovierte dort im Jahr 1995 in Kooperation mit der BMW Group zur Informationsgestaltung von Navigationsinformation für den Fahrer. Anschließend führte er das Team „Mensch-Maschine Interaktion“ in der BMW Forschung & Technik und das angeschlossene Usability Labor. Seit 2009 leitet er den Lehrstuhl für Ergonomie an der Technischen Universität München. Sein Forschungsgebiet umfasst den Bereich der sogenannten „Micro ergonomics“ zu Fragen der Mensch-Maschine-Interaktion, insbesondere den Bereich der Fahrerassistenz, der Softwareergonomie und der Kooperation zwischen Mensch und Roboter. Seine Forschung schließt dabei sowohl anthropometrische als auch kognitive Fragestellungen ein.