Zeitschrift für Praktische Philosophie
Band 11, Heft 1, 2024, S. 395–418
www.praktische-philosophie.org
https://doi.org/10.22613/zfpp/11.1.16

Vertrauen (in Technik), Vertrauenswürdigkeit (von Technik), Vertrauensadjustierung (gegenüber Technik)

Mensch-Technik-Interaktion im Spannungsfeld von kommunikativen Fähigkeiten, einem kooperativen Miteinander und der Vertrautheit (mit Technik)

Trust in, Trustworthiness of and Trust Adjustment Towards Technology

Human-Machine-Interaction Within a Tension of Communicative Capabilities, Cooperation and Familiarity with Technology

Arne Sonar, Berlin & Christian Herzog, Lübeck

Zusammenfassung: Ziel dieses Beitrags ist es, die Bedeutung der zunehmend kommunikativen und kooperativen Fähigkeiten von innovativen, z. B. auf KI-Verfahren basierenden technischen Anwendungen für die Triade aus Vertrauen (in Technologie), Vertrauenswürdigkeit (von Technologie) und Vertrauensadjustierung (gegenüber Technologie) zu diskutieren. Zudem wird die Frage aufgeworfen, wie die Rolle der Vertrautheit (mit Technologie) in diesem Spannungsverhältnis einzuordnen ist. Vertrauen ist essenziell sowohl im zwischenmenschlichen Miteinander als auch für die spezifischen Vorgänge der Mensch-Technik-Interaktion. Insbesondere neue kommunikative Potenziale technischer Anwendungen in der unmittelbaren Interaktion mit Nutzer:innen können dabei das auf Vertrauen gegründete, kooperative Miteinander von Mensch und Technik in gänzlich neuen Formen fördern. Anwendungen, die beispielsweise direkte und auf die individuellen Fähigkeiten der Nutzer:innen eingehende Rückmeldungen zu den Funktionen und möglichen Unsicherheiten der Anwendung – z. B. bei diagnostischen Empfehlungen – geben können, könnten nicht nur das grundlegende Vertrauen in digitale Technologien an sich stärken. Als spezifischen Komponenten in der Mensch-Technik-Interaktion könnte diesen zugleich

zugesprochen werden, den Anspruch der Vertrauenswürdigkeit von Applikationen (und der Entwickler:innen) ebenso wie die Zuweisung dieser Vertrauenswürdigkeit seitens der Nutzer:innen zu fördern. Daneben können sie aber auch einen für die Interaktion mit Technik ebenso essenziellen wie kritischen Punkt unterstützen: Das kontinuierliche Adjustieren von Vertrauensgraden seitens der Nutzer:innen den Applikationen gegenüber kann einem Verlust kritischer Distanz vorbeugen, der beispielsweise durch ein ungerechtfertigtes allgemeines Zuweisen von Vertrauenswürdigkeit entstehen könnte. Hervorzuheben ist daher letztlich auch die unmittelbare Bedeutung des Faktors der Vertrautheit (mit Technologie) für die triadische Beziehung zwischen dem Vertrauen (in Technologie), der Vertrauenswürdigkeit (von Technologie) und der Vertrauensadjustierung (gegenüber Technologie).

Schlagwörter: Kommunikative KI, Kooperative KI, Vertrauen, Vertrauenswürdigkeit, Vertrauensadjustierung

Abstract: The aim of this paper is to discuss the importance of the increasingly communicative and cooperative abilities of innovative, e.g., AI-based technical applications for the triad of trust (in technology), trustworthiness (of technology) and trust adjustment (towards technology). It also raises the question of the role of familiarity (with technology) in this context. Trust is essential in both interpersonal interaction as well as in human-technology interaction. In particular, new communicative potentials of technical applications in direct interaction with users can promote the cooperative interaction of people and technology based on trust in completely new forms. Applications that can, for example, provide immediate feedback on their functions, as well as on possible uncertainties, e.g., in the case of diagnostic recommendations, could not only strengthen basic confidence in the technology itself. As specific components in the human-technology interaction, these kinds of feedback could at the same time promote the claim of trustworthiness of applications (and developers), as well as the assignment of trustworthiness on the part of users. In addition, however, they can also support a point critical for the interaction with technology: the continuous adjustment of trust levels on the part of the users towards the applications can prevent a loss of critical distance that could arise precisely due to an unjustified general assignment of trustworthiness. Consequently, emphasizing the direct importance of the factor of familiarity (with technology) for the triadic relationship between trust (in technology), trustworthiness (of technology) and trust adjustment (against technology) is essential.

Keywords: communicative AI, cooperative AI, trust, trustworthiness, trust adjustment

1 Einleitung

Für jede Form von Beziehungsstruktur, an der mindestens ein menschliches Individuum aktiv oder passiv beteiligt ist, ist Vertrauen eine grundlegende Komponente. Speziell wenn es um den Entwurf von Interaktionsmöglichkeiten mit technischen Anwendungen, und hierbei gerade solchen, die auf Verfahren der Künstlichen Intelligenz (KI) basieren, geht, ist Vertrauen (in Technik) ein entscheidendes Moment. Oder, um es mit den Worten von Ryan (2020, 2750) auszudrücken: „Trust is one of the most important and defining activities in human relationships, so proposing that AI should be trusted, is a very serious claim."

Zwar zeigt sich dabei in geisteswissenschaftlich-philosophischen Überlegungen die grundsätzliche Notwendigkeit eines normativen Anspruchs (prima facie) wie auch die ethische Bedeutung von Vertrauen in technische Applikationen. Oftmals stellt sich jedoch zugleich das Problem einer Übersetzung oder Übertragung entsprechender theoretischer Ansprüche in die unmittelbare Entwicklungspraxis, wie z. B. das konkrete Interface- und Interaktionsdesign. Gerade mit Blick auf KI-basierte Applikationen zeigt sich dabei aber, dass mit zunehmenden kommunikativen Leistungsmerkmalen der Anwendungen, z. B. über die Möglichkeit zum Dialog mit den Nutzer:innen, längst nicht mehr ausschließlich funktionale Faktoren wie die Klassifizierungs- oder Vorhersagegenauigkeit als vertrauensrelevant anzusehen sind (London 2019, 16). Betrachtet man das umfassendere soziotechnische System eines technischen Artefakts, so kann eine enge Fokussierung auf seine zuverlässige Funktionalität eine grobe Vernachlässigung der (zumeist menschlichen) Beziehungen, die Vertrauen ermöglichen und möglicherweise sogar begründen, implizieren (vgl. Rieder et al. 2021).

Hier sollen die nachfolgenden Überlegungen ansetzen und Zusammenhänge zwischen einer kommunikativen (und damit einhergehenden kooperativen) Dimension der Mensch-KI-Interaktion und einer unsererseits angenommenen Triade zwischen dem Vertrauen (in Technik), der Vertrauenswürdigkeit (von Technik) und der Vertrauensadjustierung (gegenüber Technik) erläutern. Wollen wir den Terminus der „Kooperation" vor allem auf die Zusammenarbeit zur Erreichung eines (bestimmten) gemeinsamen Ziels (z. B. Buxbaum und Sen 2018, 6) beziehen, soll der Terminus der „Kommunikation" zweierlei Dimensionen adressieren: Einerseits expliziert Kommunikation die Übertragung von Informationen von einem Sender zu einem Empfänger und deren Rekonstruktion seitens des Empfängers. Andererseits impliziert Kommunikation auch eine adaptive Dimension im Sinne des Ein-

flusses auf das Verhalten des Empfängers (Tjøstheim et al. 2020, 1–3). Insbesondere die zweite Dimension, d. h. mögliche Prozesse der Verhaltensveränderung mittels entwickelter Mechanismen, gilt es dabei in den Fokus der Betrachtung zu rücken (vgl. Tjøstheim et al. 2020, 3). Ein zentrales Anliegen dieser Arbeit besteht vor allem darin, relevante Momente für die Vertrauensbildung über einen spezifischen konzeptuellen Zugang zur kommunikativen und damit einhergehend auch kooperativen Dimension der Mensch-Technik-Interaktion nachzuzeichnen. Wir sind uns bewusst, dass eine umfassendere Thematisierung von Vertrauenswürdigkeit auch die entwickelnden, regulierenden, sozialen und konkurrierenden Bestandteile des sozio-technischen Gesamtsystems berücksichtigen muss (vgl. Rieder et al.: 2021). Für die hiesigen Zwecke wollen wir hauptsächlich die Dimension des Vertrauens in einer anwendungsimmanenten Umgebung fokussieren, d. h. wir ziehen die Grenze des sozio-technischen Systems sehr eng um den Kontext des beabsichtigten Anwendungsfalls einer Applikation.

Obwohl nachfolgend vor allem mit konzeptionellen Überlegungen gearbeitet wird, entspringen diese einem anwendungsbezogenen Entwicklungsprojekt – Gegenstand ist ein KI-Algorithmus zur Unterstützung des Tiefvenenthrombosenausschlusses am Point-of-Care. Ein charakteristisches Spezifikum dieser Applikation liegt darin, über einen KI-Algorithmus nicht nur den Prozess der bildgestützten Diagnostik zu unterstützen, sondern gleichsam den vorhergehenden Untersuchungsvorgang zu begleiten bzw. gar auch anzuleiten. Für die Interaktionsvorgänge ebenso wie für den damit verbundenen Prozess der Vertrauensbildung werden derart sowohl die kommunikative als auch die kooperative Dimension relevant. Die konkrete Applikation ist Inspiration zur vorliegenden Arbeit, soll aber den Fokus nicht einschränken, sondern – wenn möglich – als Veranschaulichung dienen.

In Abschnitt 2 sollen zunächst die Grundlagen unserer Überlegungen zur Vertrauensrelevanz der kommunikativen und kooperativen Dimension der KI-Interaktion dargestellt werden. Mit Abschnitt 3 soll dann die eigentliche Triade von Vertrauen (in Technik), der Vertrauenswürdigkeit (von Technik) und der Vertrauensadjustierung (gegenüber Technik) näher betrachtet werden. Nachdem die einzelnen Punkte und ihre Zusammenhänge näher dargestellt wurden, widmet sich abschließend Abschnitt 4 dem Punkt der Vertrautheit (mit Technik). Abschnitt 5 fasst zusammen und führt mögliche Implikationen auf.

2 Zur Vertrauensrelevanz der kommunikativen und kooperativen Dimension in der Mensch-KI-Interaktion

Vor allem in Verbindung mit KI gibt es in der Literatur den feststellbaren Trend dazu, innovative technische Applikationen längst nicht mehr ausschließlich als nützliche Werkzeuge und funktionale Instrumente für bestimmte Aufgaben zu verstehen, sondern diese zunehmend als, in die Mensch-Technik-Interaktion eingebundene, proaktive Interaktionspartner, d. h. Kooperateure und Kommunikatoren aufzufassen (vgl. Gunkel 2017; Guzman und Lewis 2020; Hepp 2020; Bellet et al. 2011, 155). Das folgende Zitat soll hierfür exemplarisch stehen:

> Communicative AI technologies are not only designed to function as communicators but are also interpreted by people as such. (Guzman und Lewis 2020, 73)

Der unmittelbare Zusammenhang zwischen der kommunikativen als auch kooperativen Dimension der Mensch-KI-Interaktion und dem Aspekt des Vertrauens (in eine Applikation/Technik) offenbart sich hierbei vor allem dann, wenn es um die Gestaltung und Etablierung angemessener Vertrauensverhältnisse geht, z. B. durch Klärung möglicher Diskrepanzen zwischen den Erwartungen der Nutzer:innen und den tatsächlichen Leistungsmerkmalen einer spezifischen Anwendung (McCormack et al. 2020, 46f.) – dies sowohl im Allgemeinen vor jeder Nutzung, als auch während der unmittelbaren Interaktionsprozesse. Gerade auch weil eine kooperative Dimension im Bereich der Mensch-Technik-Interaktion die Einräumung zulässiger Grade an kooperativer Autonomie impliziert (vgl. Kirn 2002, 55f.), bedarf es einer sinnhaften, d. h. synergetischen Koordinierung der gegenseitigen Ergänzung (Komplementarität) (vgl. Hakli 2017, 644). Zusätzlich bedarf es ebenso einer jeweiligen, z. B. situativ-adaptiven Anpassung an das Verhalten des jeweiligen Kooperationspartners (Hakli 2017, 651). Speziell ein Verständnis von den Zielen einer Anwendung ist für die Verhaltensanpassung in konkreten Interaktionsprozessen und auch für die Vertrauensbildung wichtig. Die Transparenz von Zielfunktionen wird somit als Schlüssel für eine sichere, komfortable und koordinierte Interaktion zwischen Mensch und Technik zunehmend unverzichtbar (vgl. Huang et al. 2019, 310f.).

An dieser Stelle offenbart sich auch die Bedeutung der kommunikativen Dimension für Vertrauen in der Mensch-Technik-Interaktion: Die Kommunikation zwischen Nutzer:innen und einer technischen Anwendung

ist nicht nur wichtig, um einen kognitiven Bezugsrahmen zwischen beiden (Bellet et al. 2011, 170) zu schaffen. Sie ist auch Basis einer Verhaltensantizipation von Nutzer:innen gegenüber einer Applikation (vgl. Huang et al. 2019, 309, 324).

Die Vertrauensrelevanz der Kommunikations- und Dialogfähigkeit von Applikationen besteht also im Erkennbarmachen von Absichten, beispielsweise durch das Begründen von Empfehlungen, um so seitens der Nutzer:innen ein Hinterfragen zu ermöglichen – eine derartige Kommunikation eröffnet grundsätzlich neue, komplexere Ebenen der kooperativen Interaktion (vgl. McCormack et al. 2020, 48; Tjøstheim et al. 2020, 11). Die Vertrauenswürdigkeit einer Applikation könnte hierbei durch etwaige Faktoren der Transparenz, Nachvollziehbarkeit, Erklärbarkeit oder Interpretierbarkeit gestärkt werden.

Aktive Kommunikation von Applikationen kann auch einen unmittelbaren Einfluss auf Meinungen und das Verhalten ihrer Nutzer:innen ausüben (vgl. Guzman und Lewis 2020, 80) und die Vertrauensadjustierung gegenüber einer Anwendung bedingen. Hierbei sind sowohl die Möglichkeit antagonistischer Beziehungskonstellationen als auch die Möglichkeit und ggf. Notwendigkeit dynamischer Abhängigkeiten von beispielsweise situativen und anderen umweltbezogenen Bedingungen (u. a. Kompetenzen, Fähigkeiten, Stress) zu berücksichtigen (Kirn 2002, 56). Beispielsweise könnte ein Algorithmus die Optimierung einer speziellen Fairness-Metrik im Sinne einer Art Gemeinwohlorientierung vorsehen und kommunizieren, wohingegen dies den Eigeninteressen der direkten Nutzer:in zuwiderläuft. Im Beispiel des KI-gestützten Tiefvenenthrombosenausschlusses ließen sich unterstützende Informationen abhängig vom Fähigkeitsgrad des Nutzenden, der z. B. durch etwaige Charakteristika der Handhabung der Ultraschallsonde abgeleitet und transparent kommuniziert werden könnte, einblenden. Ist ein:e Nutzer:in geübt, wird sie ggf. durch zusätzliche, wahrscheinlich auch gelegentlich den eigenen Intentionen widersprechenden Unterstützungsversuchen das Vertrauen in die Nützlichkeit der Applikation verlieren.

Im oben Beschriebenen zeigt sich unseres Erachtens die unmittelbare Verknüpfbarkeit zwischen kommunikativen und kooperativen Anlagen von technischen Anwendungen und der Triade von Vertrauen, Vertrauenswürdigkeit und Vertrauensadjustierung im Rahmen von Mensch-Technik-Interaktionsprozessen: Wie beispielsweise Pols (2016, 1–3) beschreibt, ist die Funktionalität einer technischen Applikation, gerade solcher, die medizinische Prozesse unterstützen sollen, mitunter eng verknüpft mit den Men-

schen, welche sie nutzen. In diesem Sinne ist die Adaption von Applikationen im unmittelbaren Praxisgebrauch von grundlegender Bedeutung, speziell für deren konkrete Sinngebung. Erst derart werden diese medizinischen Prozesse als spezifische Mensch-Technik-Beziehungsgefüge umfänglich erschlossen. Grundsätzlich gilt es dabei zu bedenken, dass unterschiedliche Typen von Nutzer:innen stets auch divergente Adaptions-/Anpassungsgewohnheiten gegenüber Applikationen aufweisen können. Einige lassen sich vielleicht auf eine Art nachhaltig arbeitsteilige Mensch-Technik-Interaktion ein, während andere tendenziell das Ziel verfolgen, sich von unterstützenden Applikationen zu emanzipieren, indem sie eigene Expertise ableiten. Da die Möglichkeit einer derartig divergenten Adaption wiederum das unmittelbare Kooperationsdesign wie auch die damit verknüpfte allgemeine Vertrauensbildung erschwere, könnte ein Management der Technikadaption die Ausbildung effektiver Kooperationsbeziehungen und damit die grundlegende Vertrauensbildung unterstützen (vgl. Thomas und Bostrom 2008, 2–4).

Die Vertrauensentwicklung gegenüber einer Applikation – einerseits deren Vertrauenswürdigkeitswahrnehmung als auch andererseits die kontinuierliche Vertrauensadjustierung ihr gegenüber – ist unserem Verständnis nach in der konkreten Interaktion eng an die spezifische Kommunikation und auch unmittelbare Kooperation mit ihr gebunden. Greifen wir hierbei die obigen Darstellungen auf, entsteht Vertrauen in Technik zunächst einmal aus dem Wissen darüber, was eine Applikation kann und was nicht. Dieses Wissen unterstützt deren Adaption in der konkreten Praxis eines Anwendungsbereiches. Das innovative Moment einer solch direkten Kommunikation spezifischer Funktionen und Fähigkeiten, sowohl im Allgemeinen vor jeder Nutzung als auch im Spezifischen während konkreter Interaktionsprozesse, impliziert damit einen vermehrten kommunikativen Austausch, wodurch die kooperative Interaktion zu kontinuierlichen Prozessen der dynamischen Rückkopplung zwischen Nutzer:innen und Anwendungen führt (Hepp 2020, 1416). Die Relevanz kontinuierlicher Vertrauensadjustierung wird hierdurch nicht nur adressiert, sondern zugleich in ihrer unmittelbaren Bedeutung für das Vertrauen an sich und für die konkrete Vertrauensbildung gestärkt.

Nachdem wir nun den Zusammenhang zwischen der kommunikativen und kooperativen Dimension der Interaktion zwischen Mensch und Technik und der Triade von Vertrauen in, der Vertrauenswürdigkeit von und der Vertrauensadjustierung gegenüber technischen Anwendungen grundlegend erläutert haben, soll sich nachfolgend diesem triadischen Verhältnis und dessen einzelnen Komponenten näher gewidmet werden.

3 Vertrauen in, Vertrauenswürdigkeit von und Vertrauensadjustierung gegenüber Technik

Im Folgenden wird zwischen den drei Begriffen Vertrauen in Technik, Vertrauenswürdigkeit von Technik und Vertrauensadjustierung gegenüber Technik unterschieden. Dabei betrachten wir Vertrauen in Technik als einen übergreifenden Begriff, der Fragen des Vertrauens sowohl während der eigentlichen Nutzung einer Applikation als auch in Bezug auf das breitere sozio-technische System von der Inbetriebnahme bis zur Außerbetriebnahme umfasst. Daher werden wir den Begriff der Vertrauenswürdigkeit von Technik so verstehen, dass er sich auf vertrauensrelevante oder -auslösende Momente während der Konzeption, Entwicklung, Implementierung, Wartung und Stilllegung einer Technik bezieht.

Im Gegensatz dazu bezieht sich die Vertrauensadjustierung auf gerechtfertigte und ungerechtfertigte Niveaus vertrauensvoller Einstellungen von Personen gegenüber Technik für und während der direkten Interaktion. Wir werden diese Unterscheidung insbesondere in Bezug auf die Nutzer:innen selbst entwickeln. Es gilt also zum einen zu erklären, warum und inwieweit sowohl Vertrauenswürdigkeit als auch Vertrauensanpassung als grundlegende Komponenten für die Gestaltung von Vertrauen in der Mensch-Technik-Interaktion hervorzuheben sind. Darüber hinaus gilt es, mögliche Zusammenhänge zwischen den beiden Faktoren aufzuzeigen. Andererseits soll auch untersucht werden, welche normativen Prämissen und Prinzipien damit angesprochen werden könnten.

Um die in den hiesigen Überlegungen angenommenen etwaigen Verknüpfungen zwischen dem Vertrauen in, der Vertrauenswürdigkeit von und der Vertrauensadjustierung gegenüber Technik aufzuzeigen, ist es zielführend, die einzelnen Begriffe für sich aufzuarbeiten und zugleich in ihren jeweiligen Verbindungen zueinander zusammenzuführen. Darauf aufbauend sollen verschiedene normative Prinzipien reflektiert werden, die sich mit den einzelnen Begriffen, insbesondere aber mit den zwischen ihnen herzustellenden Beziehungen, verknüpfen lassen. Zudem sind weiterführend die damit verbundenen An- und Herausforderungen für die Gestaltung der Mensch-Technik-Interaktion zumindest aus einer theoretischen Perspektive zu diskutieren.

3.1 Vertrauen (in Technik)

Vertrauen in technische Anwendungen scheint bei zunehmendem Technisierungsgrad von Handlungen, Prozessen und Organisationsstrukturen grundlegend für kooperative Beziehungen zwischen menschlichen Nutzer:innen und technischen Anwendungen (vgl. McCormack et al. 2020, 49; Okamura und Yamada 2020, 1). Vertrauen ist also konstitutiv für erfolgreiche und langfristige Beziehungen (McCormack et al. 2020, 46) – sowohl in Bezug auf ein zwischenmenschliches Miteinander als auch bezogen auf die Mensch-Technik-Interaktion. Allgemein gefasst kann Vertrauen folgendermaßen verstanden werden: „[T]rust, in general, is taken as the belief (or a measure of it) that a person (the trustee) will act in the best interests of another (the trustor) in a given situation, even when controls are unavailable and it may not be in the trustee's best interests to do so" (Marsh und Dibben 2005, 19). Im Gegensatz dazu bieten Thomas und Bostrom (2008) ein entpersonalisiertes Konzept von Vertrauen an: Vertrauen wird hier aufgegriffen als „firm belief in the reliability, truth, ability, or strength of someone or something" (Thomas und Bostrom 2008, 2; basierend auf dem Oxford Dictionary). Dementsprechend bedeutet Vertrauen in eine Person oder einen Gegenstand, dass man dieser bzw. diesem auch zutraut, eine bestimmte Handlung auszuführen, bzw. davon überzeugt ist, dass diese von dem entsprechenden Gegenüber ausgeführt werden kann (Ryan 2020, 2752). Versuche einer Übertragung des Vertrauensbegriffs auf KI fokussieren vor allem auf eine leistungsbezogene Interpretation, wonach eine Erwartungshaltung gegenüber einem KI-System existiert, dass die von diesem zu vollziehende Aufgabe auch erfolgreich absolviert bzw. erledigt wird (Okamura und Yamada 2020, 3). Solche Vertrauensdefinitionen sind im Wesentlichen auf die funktionale Dimension der Zuverlässigkeit beschränkt. Im Fokus stehen dabei Aspekte wie Leistung (Verhalten), Prozesse (zugrundeliegende Mechanismen) und Zwecke (Absicht des Systems) (McCormack et al. 2020, 46).

Wie schon oben angedeutet, wird angesichts zunehmender kommunikativer Leistungsmerkmale von technischen Anwendungen dafür argumentiert, dass die kommunikative Dimension als vertrauensbildender Faktor stärker einbezogen bzw. hervorgehoben werden muss (vgl. McCormack et al. 2020, 48). Argumentiert wird hierbei, dass stabile Erwartungshaltungen auf Routine basieren (vgl. McCormack et al. 2020, 42). Die direkte Darstellung, Erklärung und Begründung von Unterstützungsleistungen durch eine technische Anwendung können vertrauensbildende Momente darstellen (vgl. London 2019, 16; McCormack et al 2020, 48). Sind die Empfehlungen oder

der allgemeine Systemoutput für die Nutzer:innen gänzlich oder zu einem gewissen Grad nachvollziehbar und ist die Systemfunktionen erklärbar, so kann dies einerseits die Vertrauensbildung in die Anwendung erlauben und andererseits – wie im nächsten Abschnitt diskutiert wird – zu einem gewissen Grad die Vertrauenswürdigkeit einer Applikation stärken.

3.2 Vertrauenswürdigkeit (von Technik)

Speziell im Hinblick auf KI-basierte technische Anwendungen ist der Aspekt der Vertrauenswürdigkeit nicht zuletzt seit der Veröffentlichung der Leitlinien für vertrauenswürdige KI der High-Level Expert Group on Artificial Intelligence (AI HLEG) der Europäischen Kommission allgegenwärtig. Wie Rieder et al. (2021, 27) beispielhaft zeigen, verweist diese Omnipräsenz des Begriffs „vertrauenswürdige KI" (engl.: „Trustworthy AI") auf die implizierte politische Dimension der Vertrauensbildung an sich. Autor:innen wie Ryan (2020, 2752) betonen dabei, dass die Zuweisung von Vertrauen eine Ansicht oder einen Glauben an die Vertrauenswürdigkeit einer Entität voraussetzt:

> Trustworthy agents are those worthy of being trusted–whether or not we trust them. To be worthy of trust, though, one must be capable of being trusted. Trustworthy agents are those who have the competence to actually fulfil the trust that is placed in them. Ryan (2020, 2752)

Rieder et al. (2021, 31) knüpfen hier in gewisser Weise an:

> In philosophical analyses of trust, ‚trust' is ordinarily formulated as a three-place relation involving a trustor A, a trustee B, and either a domain of inter-action or a specific good P. ‚Trustworthiness' or ‚being trustworthy' is then defined by the character of being merit of trust. Rieder et al. (2021, 31)

Gemeinsam ist den zitierten Autor:innen die Kritik am normativen Wert des (auch politisch motivierten) Konzepts „vertrauenswürdige KI". Ein Fokus auf die etwaigen Dualismen von Vertrauenswürdigkeit und Verlässlichkeit (engl.: reliance/reliability) erscheint beispielsweise im Hinblick auf die operative Stabilität von Applikationen argumentativ vertretbar (vgl. Weydner-Volkmann 2021, 54). Relevante Kritikpunkte legen jedoch nahe, dass die der Vertrauenswürdigkeit zugrunde liegenden Konzepte nicht alleinig auf den Aspekt der Verlässlichkeit einer KI-Anwendung beschränkt werden sollten (vgl. Rieder et al. 2021, 31, 33; Ryan 2020, 2750). Vertrauen an sich und

eine damit verbundene Vertrauenswürdigkeit von KI basiert nicht nur auf einem Ansatz rational erwartbarer Zuverlässigkeit (Rational-Choice-Ansatz), sondern auch auf affektiven und normativen Ansätzen (Ryan 2020) oder auf moralischen und motivationsbezogenen Momenten (Rieder et al. 2021). So argumentiert Ryan beispielsweise, dass das, was als Vertrauenswürdigkeit von KI angesehen wird, nur unter der Gesamtheit rational erwartbarer Zuverlässigkeit und moralischer Momente gilt (Ryan 2020, 2750). Die Kritik zielt hier vor allem darauf ab, dass KI-Systeme zwar durchaus dem rationalen Moment, d. h. der erwarteten Leistung entsprechen. Innerhalb solcher Systeme seien jedoch weder affektive Momente, d. h. wohlwollende Handlungsmotive, noch normative Vertrauensmomente, d. h. eine Anerkennung von Vertrauen und ein Bewusstsein von Vertrauensmissbrauch gegeben (vgl. Ryan 2020, 2759–2761). Die Zugehörigkeit dieser affektiven Momente zu einer Vertrauensbeziehung ganz allgemein und unabhängig vom KI-Kontext ist jedoch kaum anzweifelbar, weshalb Ryan schlussfolgert, dass Verlässlichkeit und (deren) Erfahrung nur Teilbereiche, aber keine absoluten Merkmale der Vertrauensentwicklung darstellen. Technische Systeme seien jedoch nicht fähig, entsprechend dieser affektiven Momente dem Vertrauen in sie gerecht zu werden, weil ihnen das Bewusstsein für und die Sorge um die Vertrauenden fehle.

Rieder et al. (2021) stellen in ähnlicher Weise fest, dass Vertrauenswürdigkeit neben der Verlässlichkeit auch moralisch-motivationale Aspekte als konzeptionelle Grundlage benötigt. KI als Technologieform weise jedoch keine moralischen Fähigkeiten auf, was sich z. B. in ihrer Unfähigkeit äußere, auf die Werte und Interessen der ihr Vertrauenden einzugehen, oder gar die Tatsache des Vertrauens in sie als Anlass zu nehmen, im Sinne dieser zu „handeln“ (vgl. Rieder et al. 2021, 32).

Entsprechend kann es auch nach Ryan und Rieder et al. Vertrauen in KI nicht geben, sondern vielmehr nur eine Erfahrung der Verlässlichkeit von KI-basierten Anwendungen (Ryan 2020, 2759). Gerade weil lediglich die Organisationen, welche die KI-Systeme entwickeln, letztlich für Schäden verantwortlich gemacht werden könnten (Ryan 2020, 2761), könne Vertrauen und damit zugleich auch die Vertrauenswürdigkeit nur in Bezug auf die Organisationen und Individuen bestehen, welche die KI-Systeme entwickeln, kommissionieren oder kontrollieren. KI als Technologieform könne daher das Attribut der Vertrauenswürdigkeit per se nicht zugewiesen werden (Ryan 2020, 2763).

Ähnlich wie Ryan argumentieren Rieder et al. (2021, 32), dass es allenfalls so etwas wie ein indirektes Vertrauen in die technischen Objekte geben kann. Dies resultiere aber vor allem aus dem Vertrauen in die dahinterstehenden menschlichen Akteure, d. h. Designer:innen, Hersteller:innen, Administrator:innen und Betreiber:innen. Diese seien letztendlich diejenigen, welche auch die Werte und Interessen der Vertrauensgeber:innen (engl.: trustors) in ihrem Handeln berücksichtigen können. Vertrauenswürdigkeit ist dabei laut Rieder et al. durch zwei Kernprinzipien gekennzeichnet: Auf der einen Seite bedarf es, wie schon erwähnt, einer Vertrauensreaktion, d. h. der Berücksichtigung von Werten und Interessen der Vertrauensgeber:innen. Andererseits bedarf es der Kompetenz, die Erwartungen der Vertrauensgeber:innen zu erfüllen (Rieder et al. 2021, 33). Hierbei vertreten die Autor:innen die Position, dass die von der High-Level Expert Group on Artificial Intelligence (AI HLEG) im Jahr 2019 veröffentlichten Leitlinien zur vertrauenswürdigen KI die Erwartungen an ein notwendiges Kompetenzniveau für die Entwicklung von Vertrauenswürdigkeit definieren (ebd.). Die Entwicklung vertrauenswürdiger Interaktionsbeziehungen zwischen Nutzer:innen und KI-basierten Anwendungen, d. h. soziotechnischen Systemen, setzt voraus, dass die entwickelnden Instanzen (z. B. Unternehmen) diese Erwartungen erfüllen (Rieder et al. 2021, 33f.).

Wie schon weiter oben angesprochen, besteht in den hiesigen Überlegungen ein etwaiger Zusammenhang zwischen den kooperativen wie auch kommunikativen Fähigkeiten und dem Aspekt des Vertrauens in eine technische Anwendung. Der Aspekt des Vertrauens wiederum steht in unseren Überlegungen in einem triadischen Verhältnis zu den Aspekten der Vertrauenswürdigkeit (von Technik) und der Vertrauensadjustierung (gegenüber Technik). Den Aspekt der Vertrauenswürdigkeit (von Technik) verorten wir hierbei spezifisch im Moment der Entwicklung einer Anwendung – wenn auch nicht ausschließlich, da er sich letztlich auch über den gesamten KI-Lebenszyklus erstreckt.[1]

Ein grundlegender Punkt ist dabei in den Intentionen der Entwickler:innen und den Anforderungen an die Anwendung zu sehen – sowohl in Bezug auf die Applikation selbst als auch bezogen auf die Nutzer:innen. So sind u. a. die Form und Menge an Informationen, welche den Nutzer:in-

1 So könnte man hier z. B. argumentieren, dass gerade selbstlernende Systeme einen hohen Wartungsaufwand erfordern. Dieser könnte vielleicht genauso, wenn nicht sogar noch mehr, entscheidenen Einfluss auf die Vertrauenswürdigkeit haben ist als die Umstände der ursprünglichen Entwicklung an sich.

nen in bestimmten Situationen zur Verfügung gestellt, als auch wie diese durch eine Applikation unmittelbar den Nutzer:innen gegenüber präsentiert werden, für das Interaktionserlebnis relevant. Ebenso gilt dies für die Verständlichkeit und Nachvollziehbarkeit der präsentierten Informationen und Empfehlungen oder die Transparenz und Erklärbarkeit der sie generierenden Systeme. Dies sind Aspekte, für die entwickelnde Personen verantwortlich sind. Hier ist zugleich eine erste grundlegende Herausforderung für die Vertrauenswürdigkeit einer Anwendung zu finden: Die an der Entwicklung der Applikation selbst, wie auch an der Gestaltung der konkreten Interaktionsprozesse beteiligten Personen, müssen sich der Verantwortung stellen, die Vertrauenswürdigkeit einer technischen Anwendung herzustellen. Die Einbindung von Stakeholder:innen sowohl bei der Entwicklung als auch bei der Validierung von Anwendungen (vgl. Neri et al. 2020, 3) sind hierbei mögliche Maßnahmen, um die Motivationen prospektiver Anwender:innen näherungsweise zu berücksichtigen.

Innerhalb der Literatur zeigt sich zudem eine interessante Dichotomie: Einerseits wird darauf hingewiesen, dass Vertrauen in technische Anwendungen eher eine Variable der Bewertung, Wahrnehmung oder Beurteilung der Systemleistung an sich und damit weniger ein Aspekt der Gestaltung im technischen Sinne sei (McCormack et al. 2020, 47). Andererseits argumentieren Autor:innen wie Weydner-Volkmann (2021, 54f.), dass ein wesentlicher Bezugspunkt für die Vertrauenswürdigkeit einer Anwendung in ihrer Zuverlässigkeit, z. B. im Hinblick auf ihre Betriebsstabilität, zu suchen ist. Wiederum andere verstehen die Systemzuverlässigkeit und -genauigkeit (im Vergleich zu bestehenden Alternativen) gar als Parameter für Vertrauen, welche sowohl messbar als auch (gestalterisch) verbesserbar seien (vgl. London 2019, 19).

Gemäß der *Guidelines on Trustworthy AI* der EU High-Level Expert Group on Artificial Intelligence (AI HLEG AI 2019) sind technische Sicherheit und Robustheit Kernanforderungen. Eine grundsätzliche Frage, die sich dabei jedoch für jede Anwendung spezifisch stellt, ist, wo bzw. in welchen Bereichen die entsprechenden Grundwerte der Sicherheits- und Funktionalitätsbewertung angesiedelt sind (und sein müssen). Sicherheitsmargen bei autonomen Mobilitätslösungen unterscheiden sich ggf. stark von Anforderungen an Falsch-Negativ- oder Falsch-Positiv-Raten in medizinischen Klassifikatoren, vor allem wenn Prozesse weitere Diagnoseschritte vorsehen. Auch hier zeigt sich die Bedeutung einer partizipativen Einbindung der avisierten Stakeholdergruppen, um praxisnahe bzw. praxisrelevante Parameter und Erfahrungswerte für die Leistungsbewertung auszuwählen.

Darüber hinaus sollten weiterführend die Möglichkeiten der proaktiven Kommunikation von Unsicherheiten bei der Generierung und Abgabe von systemischen Empfehlungen (vgl. Kompa et al. 2021, 4) sowie die Möglichkeit gegenteiliger, d. h. vertrauensmindernder Effekte (vgl. Okamura und Yamada 2020, 2) als Aspekte der Vertrauenswürdigkeit einer Anwendung untersucht werden. So gesehen sind ethische Prämissen wie das Gebot der Wohltätigkeit, der technischen Robustheit und Sicherheit, der Verlässlichkeit, der Verantwortlichkeit sowie der Schadensvermeidung als normativer Rahmen für die Gewährleistung der Funktionssicherheit und Genauigkeit einer Anwendung unverzichtbar. Hieran lassen sich weitere normativ relevante Aspekte wie die Transparenz, Nachvollziehbarkeit und Explizierbarkeit der anwendungsimmanenten Abläufe, des Outputs der Anwendung selbst (z. B. Handlungsempfehlungen) ebenso wie dessen Generierung anschließen. Nachvollziehbarkeit und Explizierbarkeit spielen beispielsweise in der KI-unterstützten Medizin eine hervorgehobene Rolle. Autoren wie London (2019) ziehen den Nutzen der Explizierbarkeit in Zweifel und behaupten, dass eine starke Forderung danach möglicherweise lebensrettende KI-Innovationen in der Medizin verhindere. Dagegen argumentieren Herzog (2022) wie auch Bjerring und Busch (2021), dass nur Erklärungen eine qualifizierte Bewertung und Abwägung etwa von Entscheidungsempfehlungen hinsichtlich ihrer Kompatibilität mit lebensweltlich oder wertebasiert begründeten Rahmenbedingungen erlauben. Domänenspezifische Begründungen – statistische Evidenz, Angaben zu Korrelationen oder gar Kausalitäten, Unsicherheiten und Fehlerwahrscheinlichkeiten, aber keine technischen Details zur Funktionsweise neuronaler Netze – sind hier insofern Beiträge zur Vertrauenswürdigkeit, als dass sie Handlungsspielräume der Nutzenden und Betroffenen eröffnen, anstatt diese zu verschließen. Nutzer:innen wird ein differenzierterer Umgang mit der Technologie ermöglicht. Derartig explizierbare Entscheidungsunterstützungssysteme wären entsprechend mit einer Sensibilität für Motivationen medizinischen Personals zu entwickeln, um „gute Medizin“ im Sinne der einvernehmlichen Entscheidungsfindung zu gestalten.

Gestaltungsbemühungen sollten dabei jedoch nicht auf eine einfache Maximierung von Vertrauen oder Vertrauensgraden abzielen, sondern vielmehr auf die Unterstützung der Ausbildung angemessener Vertrauensmaße bei den Nutzenden. Hierbei sollten sich die Entwickelnden stets der Grenzen ihrer eigenen gestalterischen Möglichkeiten als auch jener der Applikation, die sie entwickeln, bewusst sein. Das bedeutet, dass die an der Entwicklung

und Gestaltung Beteiligten nur dann als wirklich vertrauenswürdig gelten können, wenn sie versuchen, Vertrauen in ihre Entwicklungen in einem vertretbaren Ausmaß zu fördern. Es wäre demnach paternalistisch und ggf. ethisch nicht zu rechtfertigen zu versuchen, Vertrauen in eine Applikation zu maximieren, wenn klar ist, dass die der Applikation immanenten Werte von Nutzer:innen nicht geteilt werden. So kann es beispielsweise einen Manipulationsversuch darstellen, wenn den Entwickelnden mögliche Dissonanzen mit den Wertesystemen der Nutzenden zwar evident sind, aber weder die Zielfunktion einer KI-Anwendung, noch ihre zugrundeliegende wertebasierte Rationale transparent kommuniziert werden, stattdessen aber kommunikative Mittel (rhetorisch, grafisch, emotional usw.) genutzt werden, um eine Illusion von Vertrauenswürdigkeit zu vermitteln.

Abgleiche zwischen den Erwartungen an eine Anwendung und ihren tatsächlichen Fähigkeiten böten hierbei dienliche Ansatzpunkte, um im Sinne bestmöglicher Transparenz von System- und Zielfunktionen vertretbare Grade an Vertrauensmaßen zu gewährleisten (vgl. McCormack 2020, 46f.). Darüber hinaus kann Explizierbarkeit sowohl durch Menschen als auch in technischen Anwendungen an sich schon als ein Zeichen von disziplinärer Kompetenz der Entwicklungsteams angesehen werden – Explizierbarkeit fördert Vertrauen, insbesondere durch die Fähigkeit, Entscheidungen oder Empfehlungen zu verstehen und zu bewerten (London 2019, 18). Wie McCormack et al. (2020, 46) zudem im Umkehrschluss anmerken, läge ein interessanter Untersuchungsbereich zudem in der Frage, wodurch sich nichtvertrauenswürdige KI auszeichnen würde.

Die Frage der Vertrauenswürdigkeit (von Technik) ist unserer Meinung nach stets eng geknüpft an die Zuweisung eines solchen Attributs gegenüber einer Applikation seitens der Nutzer:innen selbst. Vertrauenswürdigkeit ist demnach nicht nur ein Anspruch, der an etwas gestellt werden kann, sondern der vom Gegenüber zwangsläufig auch erfüllt werden muss. Eine vermeintliche oder gar fälschlicherweise angenommene Vertrauenswürdigkeit von Technik kann wiederum zu falschen Maßen an Vertrauen in die Technik führen. Hierin zeigt sich die Bedeutung des dritten Bestandteils der unsererseits angenommenen Vertrauenstriade: Jener einer kontinuierlichen Vertrauensadjustierung gegenüber Technik zur Gewährleistung gerechtfertigter und Vermeidung ungerechtfertigter Vertrauensgrade in eine technische Anwendung. Diesem Punkt wollen wir uns im folgenden Teil dieser Ausarbeitung detaillierter widmen.

3.3 Vertrauensadjustierung (gegenüber Technik)

Der Prozess eines kontinuierlichen Vertrauensadjustierens (gegenüber einer technischen Anwendung) zeigt sich insofern als ein gestaltungsrelevanter Faktor für vertrauensvolle Mensch-Technik-Interaktionsbeziehungen, als dass er eine angemessene Vertrauensbildung der Nutzer:innen in den unmittelbaren Interaktionsvorgängen mit einer Applikation unterstützen und fördern kann. Beispielsweise sind Effizienzansprüche in der Medizin kaum einzulösen, wenn das medizinische Personal zu andauernder Skepsis gegenüber den Ausgaben von KI-basierten Diagnose- oder Therapiehilfen angehalten wird. Interaktionsdesigns benötigen daher möglicherweise dynamische Rückkopplungsmechanismen, die ein stetes Sammeln von Erfahrung und damit die Ausbildung von Vertrauen ebenso wie von kritischer Distanz zu unterschiedlichen Aspekten der Entscheidungsunterstützung ermöglichen.

Zwar ist Vertrauenswürdigkeit (von Technik) an sich manchmal auch unabhängig von subjektiven Vertrauenszuweisungen seitens der Nutzer:innen. Jedoch lassen sich häufig auch Kopplungen zwischen beiden Komponenten ausmachen, beispielsweise dann, wenn die Vertrauenseinstellung von Nutzer:innen gegenüber einer Applikation auch an die Erklärung einer subjektiv empfundenen Vertrauenswürdigkeit der Anwendung geknüpft ist. In der Literatur wird diesbezüglich einerseits betont, dass eine erfolgreiche Kooperation zwischen KI-basierten Anwendungen und menschlichen Nutzer:innen eine implizite, ständige Anpassung des Vertrauens der Nutzer:innen in die Zuverlässigkeit des Systems bedinge (Okamura und Yamada 2020, 1). Andererseits können komplexere Systeme aber auch Vertrauensasymmetrien fördern (McCormack et al. 2020, 46). Die Möglichkeit divergenter Systemfähigkeiten und -leistungen unter sich verändernden Bedingung und in divergierenden Umgebungen gilt es in einem solchen Zusammenhang ebenso zu berücksichtigen wie die Relevanz eines notwendigen Abgleichs zwischen Erwartungen und tatsächlichen Fähigkeiten von Applikationen.

Hierin zeigt sich zugleich ein unmittelbarer Zusammenhang zwischen der Vertrauenswürdigkeit einer technischen Anwendung und der Vertrauensadjustierung ihr gegenüber: Die Tatsache, dass Entwickler:innen auf eine angemessene Vertrauenseinstellung gegenüber einer Applikation abzielen, ist in sich selbst ein Beitrag zur Vertrauenswürdigkeit. So soll beispielsweise keine einfache Deklaration von Vertrauenswürdigkeit einer Anwendung erfolgen – wie dies vielleicht auch in politisch motivierten Papieren in Bezug auf ganze Innovationsökosysteme geschieht –, sondern vielmehr die Summe der Handlungen und Bedingungen (alternative Lösungen, nicht korrumpier-

bare Incentivierung, Transparenz etc.), aber auch das konkrete Interaktionsdesign zu einer individuellen Vertrauensadjustierung durch die jeweiligen Nutzer:innen gegenüber führen, die insgesamt zu einer verantwortbaren Nutzung gereicht.

Die Entwicklung von Applikationen ebenso wie die Gestaltung von Interaktionsprozessen an normativen Prämissen wie Transparenz, Verständlichkeit, Interpretierbarkeit, Kommunikation, Offenlegung und Darstellung zu orientieren, kann somit die Bildung von adäquaten, d. h. gerechtfertigten Vertrauensgraden der Nutzer:innen in KI-basierte Anwendungen grundlegend unterstützen. Zugleich wiederum können ungerechtfertigte bzw. unangemessene Vertrauensgrade (in der Literatur unter den Begriffen over/undertrust diskutiert) (vgl. Okamura und Yamada 2020, 2), ebenso wie hiermit potenziell einhergehende „errors of omission" oder „errors of commission" (Neri et al. 2020, 519) vermieden werden.

Wie bereits erwähnt, kann Transparenz der Funktionen und Ergebnisse einer Anwendung oder aber die Kommunikation von systemischen Unsicherheiten bei der Generierung von Empfehlungen nicht zwangsläufig mit ausschließlich positiven Nebeneffekten verbunden sein. Zwar kann dies einerseits helfen, die Unsicherheiten einer systemischen Empfehlung aufzudecken (Kompa et al. 2021, 4), und im Umkehrschluss normativ relevante Aspekte wie menschliche Kontrolle, Autonomie, Befähigung und Verantwortung adressieren und stärken. Andererseits kann ein offener Umgang mit potenziellen Fehlern in bzw. einer etwaigen Fehleranfälligkeit von Anwendungen dazu beitragen, mögliche, ggf. gar ungerechtfertigte Zweifel am Leistungsvermögen einer Applikation schüren (vgl. Okamura und Yamada 2020, 2) und derart vertrauensmindernd wirken. Darüber hinaus gilt es Erwägungen darüber anzustellen, inwieweit unterschiedliche, ggf. auch seitens der Nutzer:innen als unnötig empfundene Informationsmengen als potenziell störend in der Interaktionserfahrung wahrgenommen werden und sich dadurch nachteilig auf das Vertrauen auswirken könnten. Beispielsweise kann die Darstellung der Gesamtheit von Informationen aus forensischen Gründen motiviert sein. Eine Komplexitätsreduktion seitens einer Applikation impliziert Entscheidungen anhand diskussionswürdiger Kriterien, für die ein Unternehmen haftbar gemacht werden könnte. Den entwickelnden Personen eines KI-Systems – etwa in der Medizin – könnte nicht daran gelegen sein, Verantwortung auch für die anwendungsgerechte Komplexitätsreduktion zu übernehmen. Diese könnte aber notwendig sein, um medizinisches Personal dabei zu unterstützen, Wichtiges von Unwichtigem im akuten

Moment der Diagnose oder Behandlung zu unterscheiden – ein Moment, welches im Kern aus der grundlegenden Motivation entspringt, KI-Unterstützung einzusetzen, um „gute Medizin" zu praktizieren. Ähnliche Effekte haben sich in Studien bei der Verwendung elektronischer Patient:innenakten in den USA gezeigt, durch deren Einführung medizinisches Personal wegen der gestiegenen Informationsflut statt Entlastung eine größere Belastung erfahren hat (Shanafelt 2016).

In der Literatur werden darüber hinaus zwei weitere Punkte hervorgehoben, die im Zusammenhang mit der Vertrauensadjustierung ein spezifisches Spannungsverhältnis eröffnen: So wird mitunter darauf hingewiesen, dass fehlschlagende kooperative Interaktionsbeziehungen zu eklatanten Vertrauensverlusten führen können (vgl. Thomas und Bostrom 2008, 1). Dies kann im Umkehrschluss wiederum zu falsch-negativen Annahmen über die grundsätzliche Leistungsfähigkeit und damit auch die Vertrauenswürdigkeit einer technischen Anwendung seitens der Nutzer:innen führen. Andererseits wird auch das Risiko ungerechtfertigter Vertrauensgrade gegenüber KI-Systemen und daraus resultierender Abhängigkeiten bei deren Einsatz bzw. in deren Anwendung diskutiert (vgl. McCormack et al. 2020; Santoro 2020). So weist beispielsweise Santoro (2020, 5) auf die Gefahr einer Desensibilisierung gegenüber einer kritischen Reflexion von Systemempfehlungen hin, die eine objektive Vertrauenseinstellung gegenüber KI-basierten Anwendungen untergraben könnte. Auch in diesem Zusammenhang könnte die schon weiter oben thematisierte Problematik der „errors of omission" und „errors of commission" relevant werden.

Weiterhin sind Anthropomorphismen vertrauensrelevant. So problematisieren McCormack et al. (2020, 47), dass ein übermäßiger Einsatz von menschenähnlichen Komponenten (z. B. Stimme, Mimik) in der kommunikativen Interaktion das Risiko falscher Annahmen über die potenziellen und tatsächlichen Fähigkeiten des Systems in sich bergen könne. Wie von Santoro angesprochen, kann eine objektiv vorzunehmende Adjustierung des eigenen Vertrauens zunehmend durch subjektive Wahrnehmungen und Zuschreibungen beeinflusst werden. Eine unabhängige Einschätzung bzw. Bewertung der Vertrauenswürdigkeit einer KI-basierten Anwendung seitens der Nutzer:innen wäre daher, wenn überhaupt, nur sehr eingeschränkt möglich. Dies bedingt kritisch-reflexive Überlegungen bereits während des Entwicklungsprozesses, wie gerechtfertigte Vertrauensgrade sichergestellt und ungerechtfertigte Vertrauensgrade durch Anwendungs- und Schnittstellendesign vermieden werden können. Zudem könnten hieraus Überle-

gungen darüber resultieren, inwiefern eine Unterscheidung von verschiedenen Ordnungen der Vertrauenswürdigkeit von technischen Anwendungen, d. h. erster, zweiter oder höherer Ordnungen sinnvoll und praktikabel wäre. So wäre es beispielsweise denkbar, dafür zu argumentieren, dass eine Vertrauensanpassung erster Ordnung auf der Schnittstellenebene während der unmittelbaren Nutzung einer Technik stattfinden müsse, während eine Vertrauensanpassung zweiter Ordnung sich wiederum auf die Erfüllung von Standards/Audits oder normativen Anforderungen höherer Ordnung beziehen sollte (wie Transparenz, Unternehmensberichterstattung, Mittel zur Anfechtung und Meldung von Fehlern usw.).

Aus normativer Perspektive können Wege zur Vermittlung angemessener und gerechtfertigter Vertrauensniveaus nicht nur die Vertrauenswürdigkeit der Anwendung selbst und des dahinterstehenden Entwicklungsteams bzw. der Organisation stärken, sondern auch eine grundsätzliche, nicht ausschließlich technische Dimension der Betriebszuverlässigkeit und Sicherheit unterstützen – ohne dabei die unmittelbare Bedeutung der technischen Dimension vernachlässigen oder in Frage stellen zu wollen. In diesem Sinne ist die Betriebssicherheit nicht als ein ausschließlich technisch bedingter Faktor zu sehen, sondern impliziert auch das Momentum der Interaktion zwischen Menschen und Technik selbst, ebenso wie einen stetig kritischen Umgang seitens der Nutzer:innen mit der Applikation und ihrem jeweiligen Output.

4 Die Vertrautheit (mit Technik) als Ziel- und Verknüpfungspunkt

McCormack et al. (2020, 42) konstatieren, dass die Gestaltung von (Mensch-Technik-)Interaktionsprozessen stets darauf ausgerichtet sein sollte, Vertrauen an sich zu fördern, als auch eine Vertrautheit mit der Anwendung zu schaffen. Insbesondere in der Echtzeitinteraktion stelle Vertrauen ein kritisches Element dar und erfordere ein unabdingbares Bewusstsein für gemeinsame Ziele und Handlungsweisen (McCormack 2020, 47). Wie oben beschrieben, kann ein „Generieren“ von Vertrauen jedoch nicht bedingungslos befürwortet werden. Vertrautheit wiederum zielt hierbei auch auf die Bedeutung der schon weiter oben angedeuteten adaptiven Dimension der Interaktion und die Potenzialerschließung von Anwendungen ab (vgl. Pols 2016, 1–3). Entsprechend beruht die Vertrautheit im hier vertretenen Verständnis auf der unmittelbaren Erfahrungsbildung im interaktiven Umgang

mit einer Applikation. Wir stellen die Vertrautheit mit einer Applikation in ein zentrales Spannungsverhältnis mit den drei anderen Aspekten der hier angenommenen Vertrauenstriade. Vertrauen und Vertrautheit stehen dabei in einem grundlegend sich mitunter bedingenden Verhältnis zueinander: Vertrautheit (mit einer Applikation) ist eine notwendige epistemische Bedingung für die Ausbildung oder das Ablehnen einer Vertrauensbeziehung im Sinne eines gerechtfertigten Vertrauens. Vertraut mit einer Anwendung zu sein bedeutet beispielsweise das Wissen zu haben, dass die Änderung nur eines Parameters drastische Folgen haben könnte. Vertrautheit bedingt also auch etwaige Einsichten in die Labilität, die auch zu einer Vertrauensminderung führen kann. Die Vertrautheit mit einer Applikation impliziert daher einerseits ein grundlegendes Wissen über deren Funktionen und Fähigkeiten – sowohl im unmittelbaren Gebrauch, d. h. als Erfahrungswissen, als auch davor, d. h. im Sinne eines allgemeinen, theoretischen Wissens.

Hierin wiederum offenbart sich das Spannungsfeld zwischen dem Aspekt der Vertrautheit mit einer Applikation, deren Vertrauenswürdigkeit und der Vertrauensadjustierung ihr gegenüber. Die Vermittlung von Wissen über eine Applikation, sowohl vor als auch während der Nutzung, liegt vor allem im gestalterischen Verantwortungsfeld der Entwickler:innen. Diese haben die gestalterischen Möglichkeiten, einer allgemeinen als auch unmittelbaren Aufklärung der Nutzer:innen über die Applikation. Dies schließt funktionale Spezifika, intendierte Zwecke, aber auch etwaige Unsicherheiten im zu generierenden Output ein. Von grundlegender Bedeutung hierbei ist es, die mitunter hochgradig divergente Adaption einer Applikation durch spezifische Nutzer:innengruppen zu berücksichtigen (vgl. Thomas und Bostrom 2008, 3–4, 8). Wie eine Applikation adaptiert wird, hat potenziell auch Auswirkungen darauf, wie in konkreten Nutzungssituationen deren Vertrauenswürdigkeit eingeschätzt wird. Speziell unerfahrene Nutzer:innen könnten so beispielsweise dazu tendieren, spezifischen Applikationen ein höheres Maß an Vertrauenswürdigkeit zuzuweisen. Aber auch solche Aspekte wie das Veralltäglichen des Umgangs mit Applikationen und der Wahrnehmung einer Verlässlichkeit des Systems, d. h. die Ausbildung einer Vertrautheit mit einer Applikation in der alltäglichen Anwendungspraxis, können hierbei vermehrt zu Tendenzen des Verlustes von kritischer Distanz gegenüber der Zuverlässigkeit von Anwendungen führen. Das impliziert – insbesondere bei Anwendungen mit hohem Risiko – die Notwendigkeit einer kontinuierlichen, auch durch die Applikation selbst geförderten, Vertrauensadjustierung seitens der Nutzer:innen, gerade bei höheren Graden einer etwaigen Vertrautheit.

Vertrautheit (mit Technik) spielt also für das triadische Verhältnis von Vertrauen (in Technik), Vertrauenswürdigkeit (von Technik) und der Vertrauensadjustierung (gegenüber Technik) eine ambivalente Rolle. So bedarf es eines gewissen Maßes an Vertrautheit mit einer Applikation, um wiederum Vertrauen in eine Applikation zu haben; zugleich kann eine nicht weiter hinterfragte – daher vielleicht bloß vermeintliche – Vertrautheit mit einer Applikation aber auch zu falschen Annahmen über deren Vertrauenswürdigkeit führen, insbesondere, wenn sich Anwendungskontexte oder aber die Anwendung selbst verändern (bspw. in Form selbstlernender Systeme).

5 Schlussfolgerungen

Motiviert durch die europapolitische Agenda, im Wettlauf um die Innovationsführerschaft im Bereich KI eine eigene Marke der vertrauenswürdigen KI zu setzen, ist eine Auseinandersetzung mit der damit einhergehenden Gestaltungsaufgabe aktuell und bedarf einer belastbaren technikphilosophischen Grundlage. Rieder et al. (2020) haben hier, aufbauend auf Baier (1986) und Nickel et al. (2010), bereits wertvolle Arbeit geleistet: Sie haben moralisch relevante Bedeutungsaspekte von Vertrauenswürdigkeit jenseits der Verlässlichkeit identifiziert und Vertrauensaddressat:innen klar bei den menschlichen Akteuren lokalisiert. Dies verdeutlicht Verantwortlichkeiten und negiert die Möglichkeit des Vertrauens in Technik per se (vgl. Ryan 2020).

Ziel unseres Beitrags ist es, die Triade aus Vertrauen (in Technologie), Vertrauenswürdigkeit (von Technologie) und Vertrauensadjustierung (gegenüber Technologie) zu diskutieren und weitere Handlungsmöglichkeiten für das Interaktionsdesign aufzuzeigen. Anders als Rieder et al. (2020) haben wir uns auf eine anwendungsimmanente Betrachtung beschränkt, wohlwissend, dass Vertrauenswürdigkeit stark durch Wechselwirkungen innerhalb des gesamten sozio-technischen Systems beeinflusst wird (z. B. durch benannte Stellen, Konkurrenzen, informellen Erfahrungsaustausch etc.). Unserer Ansicht nach ist dieser eingeschränktere Blick wertvoll, um zu verdeutlichen, dass auch die direkt in der Entwicklung einer KI-Lösung involvierten Personen einen großen Gestaltungsspielraum bei der Unterstützung vertrauenswürdiger KI-Technologie besitzen. Andernfalls droht vielleicht sogar das Missverständnis, dass eine im Entwicklungsökosystem verteilte Verantwortung mit der Ohnmacht einzelner Entwickler:innen gleichzusetzen sei. Um Gestaltungsspielräume zu zeigen, differenzieren wir zwischen

gerechtfertigten und ungerechtfertigten Graden an Vertrauen, argumentieren für eine kontinuierliche Vertrauensadjustierung, die durch Interaktionsdesigns mediiert werden können, und stellen die Vertrautheit mit Technik als epistemischen Mediator innerhalb der Vertrauenstriade dar.

Eine Verkettung der Aspekte ist somit aus unserer Sicht unerlässlich für Entwicklungen von Applikationen an sich wie auch für die Gestaltung von Interaktionsvorgängen zwischen Nutzer:innen und Anwendungen. Ein Einbezug von Stakeholder:innen auch in die unmittelbare Gestaltung von Interaktionsvorgängen, erscheint hierbei unerlässlich, um konkrete Bedarfe für Unterstützungsanwendungen zu erschließen.

Literatur

AI HLEG (High-Level Expert Group on Artificial Intelligence). 2019. *Ethics Guidelines for TrustworthyAI*. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelinestrustworthy-ai.

Baier, Annette. (1986). Trust and Antitrust. *Ethics* 96 (2): 231–260.

Bellet, Thierry, Jean-Michel Hoc, Serge Boverie und Guy André Boy. 2011. „From Human-machine Interaction to Cooperation – Towards the Integrated Co-pilot". In *Human-Computer Interactions in Transport*, herausgegeben von Christophe Kolski, 129–155. London: Wiley-ISTE. https://doi.org/10.1002/9781118601907.ch5.

Bjerring, Jens Christian, und Jacob Busch. 2021. Artificial Intelligence and Patient-Centered Decision-Making. *Philosophy & Technology* 34 (2): 349–371. https://doi.org/10.1007/s13347-019-00391-6

Buxbaum, Hans und Sumona Sen. 2018. „Kollaborierende Roboter in der Pflege – Sicherheit in der Mensch Maschine Schnittstelle". In *Pflegeroboter,* herausgegeben von Oliver Bendel, 1–22. Wiesbaden: Springer. https://doi.org/10.1007/978-3-658-22698-5_1.

Gunkel, David J. 2017. „Communication Technology and Perception ". In *Handbook of Communication & Media Ethics*, herausgegeben von Patrick Plaisance, 453–470. Berlin/München/Boston: De Gruyter. https://doi.org/10.1515/9783110466034-024.

Guzman, Andrea L. und Seth C. Lewis. 2020. „Artificial intelligence and communication: A Human–Machine Communication research agenda". *New Media & Society* 22 (1): 70–86. https://doi.org/10.1177/1461444819858691.

Hakli, Raul. 2017. „Cooperative Human–Robot Planning with Team Reasoning ". *International Journal of Social Robotics* 9: 643–658. https://doi.org/10.1007/s12369-016-0377-4.

Hepp, Andreas. 2020. „Artificial companions, social bots and work bots – Communicative robots as research objects of media and communication studies". *Media, Culture & Society* 42 (7–8): 1410–1426. https://doi.org/10.1177/0163443720916412.

Herzog, Christian. 2022. „On the Ethical and Epistemological Utility of Explicable AI in Medicine“. *Philosophy & Technology* 35 (50). https://doi.org/10.1007/s13347-022-00546-y

Huang, Sandy H., David Held, Pieter Abbeel und Anca D. Dragan. 2019. „Enabling robots to communicate their objectives“. *Autonomous Robots* 43: 309–326. https://doi.org/10.1007/s10514-018-9771-0.

Kirn, Stefan. 2002. „Kooperierende intelligente Softwareagenten“. *Wirtschaftsinformatik* 44 (1): 53–63. https://doi.org/10.1007/BF03251465.

Kompa, Benjamin, Jasper Snoek und Andrew L. Beam. 2021. „Second opinion needed: Communicating uncertainty in medical machine learning“. *npj Digital Medicine* 4, 4. https://doi.org/10.1038/s41746-020-00367-3.

London, Alex J. 2019. „Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability“. *Hastings Center Report* 49 (1): 15–21. https://doi.org/10.1002/hast.973.

Marsh, Stephen und Mark R. Dibben. 2005. „The role of trust in information science and technology“. In *Trust Management, Proceedings for the Third International Conference iTrust 2005 Paris*, LNCS 3477, herausgegeben von Peter Herrmann, Valérie Issarny und Simon Shiu, 17–33. Berlin/Heidelberg: Springer. https://doi.org/10.1007/b136639.

McCormack, Jon, Patrick Hutchings, Toby Gifford, Matthew Yee-King, Maria Teresa Llano und Mark D'inverno. 2020. „Design Considerations for Real-Time Collaboration with Creative Artificial Intelligence“. *Organised Sound* 25 (4): 41–52. https://doi.org/10.1017/S1355771819000451.

Neri, Emanuele, Francesca Coppola, Vittoro Miele, Corrado Bibbolino und Roberto Grassi. 2020. „Artificial intelligence: Who is responsible for the diagnosis?“ *La Radiologia Medica* 125: 517–521. https://doi.org/10.1007/s11547-020-01135-9.

Nickel, Philip J., Maarten Franssen, und Peter Kroes. (2010). „Can We Make Sense of the Notion of Trustworthy Technology?“ *Knowledge, Technology & Policy* 23 (3–4), 429–444. https://doi.org/10.1007/s12130-010-9124-6.

Okamura, Kazuo und Seiji Yamada. 2020. „Empirical Evaluations of Framework for Adaptive Trust Calibration in Human-AI Cooperation“. *IEEE Access* 8: 220335–220351. https://doi.org/10.1109/ACCESS.2020.3042556.

Pols, Jeannette. 2016. „Good relations with technology: Empirical ethics and aesthetics in care“. *Nursing Philosophy* 18: e12154. https://doi.org/10.1111/nup.12154.

Rieder, Gernot, Judith Simon und Pak-Hang Wong. 2021. „Mapping the Stony Road toward Trustworthy AI: Expectations, Problems, Conundrums“. In *Machines We Trust: Perspectives on Dependable AI*, herausgegeben von Marcello Pelillo und Teresa Scantamburlo, 27–48. Cambridge, MA: MIT Press. https://doi.org/10.2139/ssrn.3717451.

Ryan, Mark. 2020. „In AI We Trust: Ethics, Artificial Intelligence, and Reliability“. *Science and Engineering Ethics* 26: 2749–2767. https://doi.org/10.1007/s11948-020-00228-y.

Santoro, Eugenio. 2020. „L'intelligenza artificiale in medicina: quali limiti, quali ostacoli, quali domande". *Recenti Progressi in Medicina* 108 (12): 500–502. https://doi.org/10.1701/2829.28580.

Shanafelt, Tait D., Dyrbye, Lotte N., Sinsky, Christine, Hasan, Omar, Satele, Daniel, Sloan, Jeff, & West, Colin P. (2016). „Relationship Between Clerical Burden and Characteristics of the Electronic Environment With Physician Burnout and Professional Satisfaction". *Mayo Clinic Proceedings*, *91*(7), 836–848. https://doi.org/10.1016/j.mayocp.2016.05.007

Thomas, Dominic M. und Robert P. Bostrom. 2008. „Building Trust and Cooperation through Technology Adaptation in Virtual Teams – Empirical field evidence". *Proceedings of the 41st Hawaii International Conference on System Sciences* (HICSS 2008): 423–423. https://doi.org/10.1109/HICSS.2008.82.

Tjøstheim, Trond A., Andreas Stephens, Andrey Anikin und Arthur Schwaninger. 2020. „The Cognitive Philosophy of Communication". *Philosophies* 5 (4): 1–18. https://doi.org/10.3390/philosophies5040039.

Weydner-Volkmann, Sebastian. 2021. „Trust in technology – Ethical contributions to technology assessment beyond acceptance and acceptability?" *Journal for Technology Assessment in Theory and Practice (TATuP)* 30 (2): 53–59. https://doi.org/10.14512/tatup.30.2.53.