Zeitschrift für Praktische Philosophie
Band 11, Heft 1, 2024, S. 367–394
www.praktische-philosophie.org
https://doi.org/10.22613/zfpp/11.1.15

Caveat usor: Vertrauen und epistemische Wachsamkeit gegenüber künstlicher Intelligenz

Caveat usor: Trust and Epistemic Vigilance Towards Artificial Intelligence

Rico Hauswald, Dresden

Zusammenfassung: Die aktuelle Diskussion zu künstlicher Intelligenz und Vertrauen ist einerseits durch etwas geprägt, was man „Vertrauens-Enthusiasmus" nennen könnte. Dabei wird Vertrauen als eine Einstellung konzeptualisiert, die wir gegenüber KI-Systemen prinzipiell ausbilden können und – sofern und sobald diese Systeme entsprechend ausgereift sind – auch ausbilden sollten. Auf der anderen Seite wird diese Verwendungsweise des Vertrauens-Begriffs in einem signifikanten Teil der philosophischen Literatur mit großer Skepsis betrachtet. Zwei der in diesem Zusammenhang maßgeblichen Argumente lauten, erstens, dass Vertrauen in KI-Systeme nicht mit der für diese Systeme charakteristischen Intransparenz kompatibel sei, und zweitens, dass es auf eine Art Kategorienfehler hinauslaufe, zu sagen, man könne solchen Systemen „vertrauen". Ich möchte in diesem Aufsatz für die Auffassung argumentieren, dass sowohl die enthusiastische als auch die skeptische Position problematisch sind. Gegen die skeptische Position wende ich ein, dass weder das Intransparenz- noch das Kategorienfehler-Argument letztlich überzeugen, und argumentiere, dass es zumindest eine natürliche Verwendungsweise des Vertrauensbegriffs gibt – Vertrauen als Haltung nicht-hinterfragender Akzeptanz (unquestioning acceptance) –, die auch auf die Beziehung zu KI-Systemen angewandt werden kann. Andererseits wende ich gegen den Vertrauens-Enthusiasmus ein, dass dieser ein zu unkritisches Bild von Vertrauen zeichnet und dazu tendiert, dessen Risiken und Schattenseiten zu vernachlässigen. Ich setze dem enthusiastischen Bild das Prinzip *Caveat usor* entgegen und argumentiere, dass Vertrauen in KI-Systeme stets mit *epistemischer Wachsamkeit* einhergehen sollte.

Schlagwörter: Vertrauen, epistemische Wachsamkeit, künstliche Intelligenz, Intransparenz, nicht-hinterfragende Akzeptanz

Abstract: On the one hand, the current discussion on artificial intelligence and trust is characterised by what might be called "trust enthusiasm", which conceptualises trust as an attitude that we can, and should, have in relation to AI systems when they have reached a sufficient level of maturity. On the other hand, this use of the concept of trust is viewed with great scepticism in a significant part of the philosophical literature. Two of the relevant arguments in this context are, first, that trust in AI systems is incompatible with the opacity characteristic of these systems, and, second, that to say that one can "trust" such systems amounts to a kind of category mistake. In this paper, I will argue that both the enthusiastic and the sceptical positions are problematic. Against the sceptical position, I argue that neither the opacity argument nor the category mistake argument is ultimately convincing, and that there is at least one natural use of the concept of trust – trust as an attitude of unquestioning acceptance – that can be applied to the relationship with AI systems. On the other hand, I argue that trust enthusiasm paints an overly uncritical picture of trust and tends to neglect its risks and downsides. I counter the enthusiastic picture with the principle of *caveat usor*, arguing that trust in AI systems should always be accompanied by *epistemic vigilance*.

Keywords: trust, epistemic vigilance, artificial intelligence, opacity, unquestioning acceptance

1 Einleitung

Im politischen Diskurs, in Leitlinien, Werbetexten, aber auch in wissenschaftlichen Aufsätzen ist einerseits oft davon die Rede, dass künstliche Intelligenz „vertrauenswürdig" gestaltet werden sollte (vgl. z. B. Kerasidou et al. 2022, 852). Dabei wird Vertrauen als eine Einstellung konzeptualisiert, die wir gegenüber KI-Systemen prinzipiell ausbilden können und – sobald sie entsprechend ausgereift sind – auch ausbilden sollten (vgl. etwa die „Ethics Guidelines for Trustworthy AI" der High-Level Expert Group on Artificial Intelligence (HLEG) der Europäischen Kommission). „Vertrauen" erscheint in solchen Texten oft als kaum hinterfragtes „Buzzword" – als Sammelbegriff für alle als erstrebenswert erachteten Merkmale (vgl. Reinhardt 2023 für eine Analyse der allgegenwärtigen Vertrauens-Rhetorik im Kontext von KI). Man könnte diesbezüglich von einem „Vertrauens-Enthusiasmus" sprechen.

Auf der anderen Seite wird diese enthusiastische Haltung in einem signifikanten Teil der aktuellen philosophischen Literatur mit großer Skep-

sis betrachtet. Zwei der in diesem Kontext maßgeblichen Argumente lauten, erstens, dass ein Vertrauen in KI-Systeme nicht mit der für diese Systeme charakteristischen epistemischen Intransparenz kompatibel sei, und zweitens, dass KI-Systeme grundsätzlich keine geeigneten Objekte für Vertrauen seien und es auf eine Art Kategorienfehler hinauslaufe, zu behaupten, man könne solchen Systemen „vertrauen". Vertrauen, so heißt es, sei ein genuin interpersonaler Zustand, und da KI-Systeme keinen Personenstatus besitzen, könne man ihnen prinzipiell genauso wenig „vertrauen", wie man ihnen im eigentlichen Sinne „misstrauen" könne.

Ich möchte in diesem Aufsatz für die Auffassung argumentieren, dass beide Extrempositionen – der Vertrauens-Enthusiasmus einerseits und die Vertrauens-Skepsis andererseits – verfehlt sind. Es ist *nicht* so, dass die für KI typische epistemische Intransparenz die Ausbildung von Vertrauensbeziehungen grundsätzlich verunmöglichen würde; und es ist *kein* Kategorienfehler, von Vertrauen im Hinblick auf KI zu sprechen. Gleichwohl ist es fragwürdig, wenn ohne weitere Qualifikation ein Zustand als wünschenswertes Optimum propagiert wird, in dem wir KI uneingeschränkt oder blind vertrauen können.[1] Vernünftig eingesetztes Vertrauen sollte – so möchte ich deutlich machen – stets mit *epistemischer Wachsamkeit* einhergehen. Epistemische Wachsamkeit ist nicht nur *begrifflich kompatibel* mit Vertrauen; es ist vernünftigerweise auch *geboten*, gegenüber einem KI-System, dem man vertraut, zugleich epistemisch wachsam zu sein. Man kann diese Forderung mit dem Slogan *Caveat usor* auf den Punkt bringen: Ähnlich wie man beim Kauf einer Sache wachsam sein und aufpassen sollte, dass diese Sache nicht mit offensichtlichen Mängeln behaftet ist (wie es in der auf das römische Recht zurückgehenden Formel *Caveat emptor* zum Ausdruck gebracht wird), so sollte man auch bei der Benutzung eines KI-Systems wachsam sein und auf Anzeichen dafür achten, dass dessen Performance fehlerhaft ist.

Mein primäres Argumentationsziel besteht wohlgemerkt nicht in der kategorischen Behauptung, dass man KI in jedem Fall vertrauen *sollte*, oder dass Vertrauen rationalerweise eine Art *Default*-Einstellung gegenüber KI

1 Genau dies geschieht oft, wie Reinhardt (2023) in ihrer Analyse existierender KI-Leitlinien feststellt. Die Leitlinien betrachten Vertrauen überwiegend „as a ‚fundamental requirement for ethical governance' […] and as something ‚good' that shall be promoted. Whereas lack of trust is dominantly seen as something to be overcome […]. Ambivalences regarding trust are rarely discussed. Some hint at the fact that blind trust may be problematic […] but that trust is a rather ambivalent concept […] is rarely mentioned" (737).

ist. Mein *Caveat usor*-Prinzip hat vielmehr eine konditionale Struktur: *Wenn* man einem KI-System vertraut, *dann* sollte man zugleich auch epistemisch wachsam sein. Daraus ergibt sich folgende Gliederung des Aufsatzes. In Abschnitt 2 geht es zunächst um den Nachweis, dass das Antezedens des Konditionals erfüllt sein kann (denn wäre das prinzipiell unmöglich, so wäre ja jedes weitere Nachdenken über das Konditional müßig). Dazu untersuche ich kritisch die beiden genannten Vertrauens-skeptischen Argumente, beginnend mit dem Intransparenz- (2.1) und anschließend dem Kategorienfehler-Argument (2.2). In Abgrenzung dazu argumentiere ich anschließend (2.3) für die Auffassung, dass der Vertrauensbegriff facettenreicher ist, als in der Philosophie häufig angenommen wird, und dass zumindest eine Form von Vertrauen existiert – nämlich Vertrauen als Haltung nicht-hinterfragender Akzeptanz –, die sich auch in Bezug auf KI manifestieren kann. Abschnitt 3 versucht dann, das Konditional selbst zu etablieren, indem gezeigt wird, dass Vertrauen in KI stets mit epistemischer Wachsamkeit einhergehen sollte. Dazu argumentiere ich zunächst (3.1), dass Vertrauen einerseits positive Funktionen hat, andererseits aber auch mit – im Rahmen des Vertrauens-Enthusiasmus oft vernachlässigten – Risiken einhergeht, und dass *blindes* Vertrauen niemals angemessen ist. Als vernünftigen Mittelweg zwischen *gar keinem* und *blindem* Vertrauen führe ich schließlich (3.2) den Begriff der epistemischen Wachsamkeit ein und plädiere für das Prinzip *Caveat usor*, das besagt, dass Vertrauen stets mit epistemischer Wachsamkeit einhergehen sollte.

2 Vertrauen und künstliche Intelligenz

2.1 Intransparenz und Vertrauen

Wenn Fragen des Vertrauens in KI diskutiert werden, steht häufig die Problematik der Intransparenz von KI im Fokus. Auf maschinellem Lernen und künstlichen neuronalen Netzen beruhende Systeme weisen bekanntermaßen eine Form von epistemischer Opazität auf, die sie zumindest partiell als „black boxes“ erscheinen lassen: Typischerweise lässt sich nicht oder zumindest nicht vollständig nachvollziehen, wie das System einen bestimmten Input in einen bestimmten Output transformiert. Diese Intransparenz – so ein gängiges Argument – sei nicht mit Vertrauenswürdigkeit vereinbar. Nur in dem Maße, in dem die Funktionsweise eines Systems transparent und „erklärbar“ gemacht werden kann, könne das System vertrauenswürdig sein (vgl. z. B. von Eschenbach 2021). Das in den letzten Jahren zunehmend an

Fahrt aufnehmende Forschungsprogramm der *Explainable Artificial Intelligence (XAI)* bezieht eine zentrale Rechtfertigung aus dieser Überlegung; wobei die Frage, wie genau KI transparent gemacht werden kann und ob dies überhaupt in hinreichendem Maße möglich ist, allerdings nach wie vor offen ist (vgl. zu Dimensionen und Relevanz der Erklärbarkeit sowie verschiedenen Ansätzen, diese herzustellen, bspw. Zednik 2021 und Boge 2022).

Ich möchte hier durchaus nicht die Sinnhaftigkeit des XAI-Programms insgesamt infrage stellen. Gleichwohl scheint mir die Auffassung, dass epistemische Transparenz eine *Conditio sine qua non* für die Schaffung vertrauenswürdiger KI ist, aus mindestens zwei Gründen nicht überzeugend zu sein. Zum einen scheint epistemische Transparenz auch in herkömmlichen Zusammenhängen keine notwendige Vorrausetzung dafür zu sein, dass etwas oder jemand als vertrauenswürdig bewertet werden kann. Wenn wir mit anderen Menschen interagieren, dann sind deren Gehirnprozesse für uns nicht weniger intransparent als die in KI-Systemen stattfindenden Prozesse, und dennoch können diese Menschen für uns die Rolle von Vertrauenspersonen[2] spielen.[3] Man mag gegen diese Überlegung einwenden, dass dies nicht die Form von Transparenz sei, die für die Etablierung von Vertrauensbeziehungen relevant ist, und es vielmehr um Argumente, Rechtfertigungen oder das Geben und Nehmen von nachvollziehbaren Gründen ankomme (vgl. zur Kritik der Analogie KI-Mensch im Hinblick auf Intransparenz und Erklärbarkeit etwa Maclure 2021). Dieser Einwand überzeugt jedoch nur bedingt. So sind etwa die Beziehungen zu epistemischen Autoritäten ein paradigmatisches Beispiel für menschliche Vertrauensbeziehungen, die seitens der epistemisch inferioren Akteure (der „Laien") typischerweise auf der Annahme beruhen, dass die Autoritäten überlegene Kompetenzen in ihren jeweiligen Expertisedomänen besitzen und in der Lage sind, Informationen und andere wertvolle epistemische Güter aus diesen Domänen testimonial an die Laien weiterzugeben. Die Deliberationsprozesse epistemischer Autoritäten beruhen jedoch oft auf esoterischer Expertise und können nicht in einer für Laien

2 In der englischsprachigen Literatur wird oft zwischen „trustor" (der Person, die vertraut) und „trustee" (der Person, der vertraut wird) unterschieden. Letztere bezeichne ich hier als „Vertrauensperson" (wie etwa auch Lahno 2002, 34). Da es mir aber primär um KI-Systeme in der Rolle des „trustee" geht, werde ich auch von „Vertrauensobjekten" sprechen.

3 Hier sei auch auf die von Krügel, Ostermaier und Uhl (2022) diskutierten empirischen Belege dafür verwiesen, dass Versuchspersonen durchaus bereit sind, unter Bedingungen von Intransparenz zu vertrauen.

nachvollziehbaren Weise verständlich gemacht werden; die Laien übernehmen die Informationen bzw. anderen epistemischen Güter, ohne deren Zustandekommen oder entsprechende Rechtfertigungen verstehen zu können (für eine ausführliche Diskussion vgl. Hauswald 2024). Vertrauen scheint epistemische Transparenz also nicht zwingend zu erfordern.

Darüber hinaus lässt sich, zweitens, fragen, ob es Vertrauen in Kontexten, die für die vertrauende Person epistemisch transparent sind, überhaupt brauchen würde. Vertrauen scheint vor allem dort eine Funktion zu haben, wo Unsicherheit und Komplexität vorherrschen; Vertrauen dient dazu, diese Unsicherheit und Komplexität zu reduzieren und bewältigbar zu machen (Luhmann 2014). In diesem Sinne ist darauf hingewiesen worden, dass bei der Etablierung vertrauenswürdiger KI-Systeme in stärkerem Maße deren zuverlässige Funktionsweise in den Fokus gerückt werden sollte und die Frage der Transparenz oder Intransparenz ihrer internen Prozessabläufe bestenfalls eine untergeordnete Rolle spiele.[4]

Sicherlich wird die Frage, unter welchen genauen Umständen wir intransparenten KI-Systemen vertrauen können, weiterhin kontrovers diskutiert werden, und ich erhebe nicht den Anspruch, mit diesen knappen Überlegungen Knockdown-Argumente zu liefern, mit denen sich alle Details dieser Debatte entscheiden lassen. Gleichwohl scheinen mir diese Überlegungen hinreichend dafür, die These, dass sich Vertrauenswürdigkeit und Intransparenz prinzipiell und unter allen Umständen ausschließen, als sehr fragwürdig zurückzuweisen.

2.2 „Vertrauen in künstliche Intelligenz" – ein Kategorienfehler?

Ein noch fundamentalerer Einwand gegen das Konzept einer vertrauenswürdigen KI ergibt sich allerdings aus dem Argument, dass Vertrauen eine genuin interpersonale Einstellung sei, die prinzipiell nur zwischen menschlichen Personen bestehen könne. Auf KI-Systeme und andere nicht-personale Entitäten könne man sich bestenfalls „verlassen"; von „Vertrauen" gegenüber diesen Entitäten zu sprechen, liefe hingegen auf eine Art Kategorienfehler oder „begrifflichen Unsinn" hinaus.[5]

4 Durán und Jongsma (2021) nennen diese Position „computational reliabilism". Zur Diskussion der Potentiale und Grenzen dieser Position vgl. auch Grote (2021), Lang (2021), Mishra (2021) und Véliz et al. (2021).

5 Die Formulierung „conceptual nonsense" findet sich u. a. bei Freiman (2023); vgl. auch Braun, Bleher und Hummel (2021).

In der philosophischen Diskussion wird Vertrauen häufig als Einstellung des Sich-Verlassens *plus* X analysiert (vgl. z. B. Baier 1986; Hawley 2014). Worin besteht dieses X, das zum bloßen Sich-Verlassen noch hinzukommen muss, damit von genuinem Vertrauen gesprochen werden kann? Zu dieser Frage werden unterschiedliche Auffassungen vertreten, wobei sich – bei allen Unterschieden zwischen einzelnen Theorien im Detail – insbesondere affektive und normative Ansätze unterscheiden lassen. Affektive Ansätze gehen davon aus, dass die vertrauende Person auf den guten Willen der Vertrauensperson setzt und an diesen glaubt. Die vertrauende Person geht von der Erwartung aus, dass die Vertrauensperson durch den Gedanken, dass erstere auf sie zählt, positiv motiviert wird. Normative Ansätze nehmen dagegen an, dass die Vertrauensperson von normativen Erwartungen motiviert wird, d. h. davon, was sie tun soll. Erfüllt sie diese normativen Erwartungen *nicht*, kann es für die (bis dahin) vertrauende Person angemessen sein, mit reaktiven Einstellungen wie dem Gefühl, sich verraten zu fühlen, zu reagieren. Es bedarf an dieser Stelle keiner ausführlicheren Darstellung der klassischen Ansätze. Entscheidend ist, dass sowohl die affektiven als auch die normativen Ansätze Bedingungen postulieren, die KI-Systeme offensichtlich nicht erfüllen können. So kommt Ryan zu der Schlussfolgerung „AI cannot be something that has the capacity to be trusted according to the most prevalent definitions of trust because it does not possess emotive states or can be held responsible for their actions – requirements of the affective and normative accounts of trust" (2020, 2749).[6]

Diese Schlussfolgerung ist meines Erachtens gültig: Unter der Voraussetzung, dass der Vertrauensbegriff im Sinne der gängigen affektiven oder normativen Ansätze zu analysieren ist, lässt sich die Konklusion ableiten, dass KI-Systeme keine geeigneten Vertrauensobjekte sind. Es lässt sich zwar nicht die prinzipielle Möglichkeit ausschließen, dass der technologische Fortschritt einmal KI-Systeme hervorbringen wird, die über echte affektive Zustände verfügen oder die als „full ethical agents" (Moor 2006) einen normativen Status erlangen, der demjenigen menschlicher Personen vergleichbar ist. Doch aktuelle KI-Systeme sind davon ganz offenbar noch weit entfernt. Das Problem an Ryans Schlussfolgerung ist allerdings, dass sie von falschen Prämissen ausgeht. Wie ich im folgenden Abschnitt zeigen werde, ist der Vertrauensbegriff facettenreicher, als von Ryan und generell

6 Vgl. mit ähnlicher Stoßrichtung auch Hatherley (2020), Freiman (2023) und Kerasidou et al. (2022).

in der philosophischen Diskussion oft angenommen wird. Ich räume zwar ein, dass es Spielarten von Vertrauen gibt, die von den affektiven und normativen Ansätzen korrekt beschrieben werden. Das schließt aber nicht aus, dass es weitere Spielarten gibt, die auch eine Anwendung auf die Beziehung zu KI zulassen.

2.3 Vertrauen als Haltung nicht-hinterfragender Akzeptanz

Wenn versucht wird, die These zu begründen, Vertrauen sei ein genuin interpersonaler Zustand, besteht eine oft verwendete Strategie darin, auf linguistische Befunde zu verweisen. Wir sagen beispielsweise, dass wir einem Freund vertrauen, dass er sein Versprechen einhält, oder einer Expertin, dass sie wahre Informationen bereitstellt; demgegenüber wäre es befremdlich – wenn nicht ungrammatisch –, einen Satz zu äußern wie „Ich vertraue der Sonne, dass sie morgen wieder aufgeht".[7] Ganz so eindeutig ist die linguistische Evidenz jedoch keineswegs. So scheint etwa der eben genannte Satz in nur leicht modifizierter Form – „Ich vertraue darauf, dass die Sonne auch morgen wieder aufgeht" – durchaus nicht ungrammatisch zu sein. Auch sprechen wir natürlicherweise davon, dass wir unseren Sinnen, kognitiven Fähigkeiten oder überhaupt uns selbst vertrauen (zu epistemischem Selbstvertrauen vgl. etwa Zagzebski 2012, Kap. 2); oder wir sagen, dass Blinde ihren Hunden vertrauen (Braun/Bleher/Hummel 2021). Auch die Rede von Vertrauen in technische Artefakte einschließlich KI-Systeme erscheint vielen Mitgliedern der deutschen Sprachgemeinschaft offenbar alles andere als unnatürlich (ähnliches gilt für das Englische und andere Sprachen); auf die Allgegenwart der entsprechenden Rhetorik hatte ich eingangs ja bereits hingewiesen.

Wenn Vertrauen ein genuin interpersonaler Zustand ist, der sich nach Maßgabe der affektiven oder normativen Ansätze analysieren lässt, müssen diese linguistischen Befunde mindestens als sehr bemerkenswert erscheinen. Angesichts der Vielfalt unterschiedlicher Objekte, von denen natürlicherweise gesagt wird, dass man ihnen vertraut (Sinnesorgane, Blindenhunde, technische Artefakte usw.), sowie der quantitativen Menge entsprechender Sprechakte, erscheint es als wenig vielversprechend, in all diesen Fällen das

7 Das Beispiel wird von Hatherley verwendet: „Although I am supremely confident that tomorrow the sun will rise in the east and set in the west, there is not familiar sense in which I could reasonably said to *trust* the sun to do so" (2020, 480).

Vorliegen von sprachlichen Fehlern zu konstatieren. Eine viel plausiblere Erklärung besteht aus meiner Sicht in der Diagnose, dass Vertrauen ein facettenreicher Begriff ist, der verschiedene Spielarten umfasst, von denen sich *einige* im Sinne der affektiven oder normativen Ansätze analysieren lassen mögen, wohingegen *andere* jedoch eine andere Analysestrategie erfordern.

Als vielversprechend erscheint mir insbesondere ein jüngst von Nguyen (2023) gemachter Vorschlag, der Vertrauen im Sinne einer *non-questioning attitude* analysieren möchte. Vertrauen in diesem Sinne zeichnet sich dadurch aus, dass man die Zuverlässigkeit der Vertrauensperson oder des Vertrauensobjekts sozusagen außerhalb des Raums der Deliberation und Bewertung stellt und ihm gegenüber eine „unhinterfragende Haltung" oder „Haltung des Nicht-Hinterfragens" einnimmt. Nguyen betont, dass man eine solche Haltung gegenüber Personen einnehmen kann, aber genauso eben auch gegenüber nicht-personalen Entitäten. Und offenbar kann es eine solche Form von Vertrauen auch gegenüber KI geben. Man könnte dann sagen, dass wenn man einem KI-System in diesem Sinne vertraut, man das System verwendet, ohne aktiv seine Funktionsfähigkeit zu hinterfragen oder seine Ergebnisse gegenzuprüfen. Das schließt nicht aus, dass man das System im *Vorfeld* gründlich auf seine Funktionsfähigkeit geprüft *hat*. Aber wenn man – aufgrund eines positiven Ergebnisses der Prüfung – einmal Vertrauen in das System gefasst hat, dann läuft dieses Vertrauen darauf hinaus, dem System gegenüber bis auf Weiteres eine nicht-hinterfragende Haltung einzunehmen.

Genauer lautet Nguyens Charakterisierung: „To trust X to P is to have an attitude of not questioning that X will P" (2023, 225). Darüber hinaus präsentiert Nguyen auch eine Formulierung für den epistemischen Spezialfall, d. h. für den Fall, dass das X als testimoniale Quelle, epistemisches Instrument[8] oder anderweitige Informationsquelle dient; sie lautet: „To trust X as an informational source in domain Z is to have an attitude of not questioning X's deliverances concerning Z" (ebd.).

Diese Formulierungen scheinen mir allerdings in einer Hinsicht noch defizitär oder zumindest unglücklich zu sein. Einer Person oder anderen Entität X zu vertrauen ist mehr, als sie einfach nur *nicht* zu hinterfragen. Die bloße Abwesenheit einer hinterfragenden Haltung ist noch nicht hinrei-

8 Goldberg versteht unter einem *epistemic instrument* „any instrument that is designed so that its outputs convey worldly information so as to be accessible to parties who learn to comprehend these outputs" (2020, 2785).

chend für Vertrauen. Denn ich kann ja einem X gegenüber auch dann eine nicht-hinterfragende Haltung haben, wenn ich komplett indifferent gegenüber X bin. Wenn mir X einfach egal ist (oder ich noch nicht einmal von X Kenntnis besitze), dann hinterfrage ich es nicht, aber ich vertraue ihm auch nicht. Kurz: Vertrauen ist mehr als bloße Indifferenz,[9] Nicht-Hinterfragen ist aber mit Indifferenz kompatibel.

Was ist es, was zum bloßen Nicht-Hinterfragen für Vertrauen noch hinzukommen muss? Ich denke, wir können der Antwort durch Vergleich der „Indifferenz-Fälle" und der „Vertrauens-Fälle" näherkommen. Wenn ich einer Person oder Sache X gegenüber indifferent bin, dann kümmere ich mich nicht um sie. Sie mag tun oder lassen, was sie will, ihr Verhalten wird keinen Einfluss auf meine Handlungen haben. Wenn ich X im Hinblick auf P aber vertraue, dann wird es einen solchen Einfluss sehr wohl geben, und zwar typischerweise einen positiven Einfluss.[10] Ich werde dann meine Handlungen bzw. Anschlusshandlungen so orientieren, dass ich P als gegeben annehme,[11] oder – im Fall epistemischen Vertrauens – dass ich die von

9 Diesen Aspekt betonen auch Hartmann (2011) und Budnik (2021). So schreibt Hartmann: „Kontrapunkt zum Vertrauen und zum Misstrauen ist folglich eher eine Gleichgültigkeit oder Indifferenz, die sich durch eine Abwesenheit von Interesse und Engagement kennzeichnen lässt" (2011, 58). Und Budnik schreibt: „Vertrauen hängt also nicht nur mit epistemischer Unsicherheit, sondern auch mit der praktischen Relevanz des Vertrauensgegenstands für die vertrauende Person zusammen. Wenn wir vertrauen, dann sind wir uns nicht ganz sicher, wie eine andere Person sich verhalten wird, und es steht dabei gleichzeitig etwas auf dem Spiel für uns" (2021, 48).

10 Dass der Einfluss „positiv" ist, soll heißen, dass die Aktivitäten von X wertgeschätzt werden. Diese Einschränkung ist nötig, um eine Abgrenzung zu möglichen Fällen vorzunehmen, in denen ich einer Person oder Sache nicht vertraue, ihr aber auch nicht indifferent gegenüberstehe und mich irgendwie darauf einstellen muss – Fälle, in denen ich die Aktivität geringschätze, z. B. weil sie für mich ungünstig oder schädlich ist, und sie einen negativen Einfluss auf mein Handeln hat in dem Sinne, dass sie mich z. B. zum Treffen bestimmter Vorkehrungen motiviert.

11 Wenn ich beispielsweise einem Kooperationspartner in einem Projekt vertraue und er eine bestimmte Zuarbeit zugesagt hat, dann könnte mein Vertrauen etwa darin bestehen, dass ich bestimmte weiterführende Arbeiten, die auf dieser Zuarbeit aufbauen, in Angriff nehme, noch bevor sie geliefert wurde, und auch ohne dass ich Vorkehrungen für den Fall treffen würde, dass sie nicht geliefert wird.

der Quelle bereitgestellten Informationen in meinen Deliberationsprozessen als Prämissen verwende, ähnlich wie ich die von meinen Sinnesorganen bereitgestellten Informationen normalerweise einfach als gegeben behandle.[12] Was also Vertrauen von bloßen Indifferenz-Fällen zu unterscheiden scheint, ist, dass das Verhalten der Vertrauenspersonen oder -objekte nicht nur nicht hinterfragt wird, sondern mit der Erwartung eines positiven Einflusses auf die Handlungen der vertrauenden Personen verbunden ist.[13]

Die von Nguyen beschriebene Form von Vertrauen weist also *zwei* Dimensionen oder Aspekte auf: die Abwesenheit von aktivem Hinterfragen bzw. Prüfen (ich werde kurz „Nicht-Hinterfragen" sagen) und einen positiven Aspekt (ich werde diesbezüglich wie Nguyen von „Akzeptanz" sprechen); entsprechend können wir die fragliche Form von Vertrauen als „nicht-hinterfragende Akzeptanz" beschreiben. Beide Aspekte – der positive und der negative – stellen notwendige (und gemeinsam hinreichende) Bedingungen dafür dar, dass Vertrauen im Sinne einer nicht-hinterfragenden Akzeptanz vorliegt.

Beide Dimensionen lassen offenbar Abstufungen zu. Das kann man sich gut am epistemischen Fall vor Augen führen. Wenn mir eine Informationsquelle bestimmte Informationen bereitstellt, dann kann ich diese Informationen mehr oder weniger intensiv prüfen, bevor ich sie akzeptiere. Und die Akzeptanz kann dann wiederum mehr oder weniger stark sein (mein Glaubensgrad, dass die fragliche Information wahr ist, kann mehr oder weniger hoch sein). Vertrauen ist aber nicht beliebig mit Abstufungen auf den beiden Dimensionen kompatibel. Wenn wir uns beide Dimensionen in einem zweidimensionalen System grafisch dargestellt vorstellen, dann ergibt

12 Vor diesem Hintergrund erscheint es als kein Zufall, dass wir natürlicherweise auch vom „Vertrauen in unsere Sinne" sprechen.

13 Zu Nguyens Gunsten sei betont, dass er sich des Umstands, dass bloßes Nicht-Hinterfragen nicht für Vertrauen ausreicht, durchaus bewusst ist. So finden sich an mehreren Stellen seines Textes Formulierungen wie: „To trust an informational source wholeheartedly is to *accept* its claims without pausing to worry or evaluate that source's trustworthiness" (2023, 214). Hier ist von *beidem* die Rede: der Nicht-Hinterfragung und dem positivem Aspekt (der Akzeptanz). An einer Stelle beschreibt er seinen Ansatz sogar explizit als den „unquestioning acceptance account" (228f.). Misslich ist allerdings, dass dieser positive bzw. Akzeptanz-Aspekt in seiner zentralen Begriffsbestimmung (in der lediglich von einer „attitude of not questioning" die Rede ist) keine Berücksichtigung findet.

sich lediglich ein bestimmter, enger Bereich, in dem wir von „Vertrauen" sprechen würden. Meine Akzeptanz muss hinreichend stark ausgeprägt sein und das Ausmaß, in dem ich die Informationen gegenprüfe, bevor ich sie akzeptiere, darf nur sehr gering sein (oder umgekehrt: das Ausmaß, in dem meine Haltung der eines Nicht-Hinterfragens entspricht, muss hinreichend groß sein). Je nachdem, in welchem Maße beide Dimensionen realisiert sind, kann mein Vertrauen als stärker oder schwächer beschrieben werden.

Wir können eine Haltung nicht-hinterfragender Akzeptanz grundsätzlich sowohl gegenüber KI-Systemen einnehmen, die bestimmte praktische Funktionen erfüllen, als auch gegenüber KI-Systemen, die wir als epistemische Instrumente verwenden. Ein Beispiel für den ersteren Fall – und damit ein Beispiel für *praktisches* Vertrauen – ist die Verwendung eines KI-gesteuerten selbstfahrenden Autos. Einem solchen Fahrzeug als Insasse zu vertrauen, läuft darauf hinaus, eine Haltung nicht-hinterfragenden Akzeptierens in Bezug auf die erfolgreiche Durchführung der dem Fahrzeug überantworteten Aufgabe – die Insassen sicher von A nach B zu bringen – einzunehmen und beispielsweise nicht beunruhigt zu reagieren oder Gegenmaßnahmen zu ergreifen, wenn das Fahrzeug eine Route einschlägt, die man nicht erwartet hat. Ein Beispiel für den zweiten Fall – und damit für *epistemisches* Vertrauen – ist die Verwendung eines Systems zur medizinischen Diagnostik (etwa zur Hautkrebserkennung). Dem System zu vertrauen, heißt hier, seine Outputs (dass in einem konkreten Fall Hautkrebs vorliegt oder nicht vorliegt) zu akzeptieren, ohne sie noch einmal durch Durchführung weiterer unabhängiger Tests oder die Konsultation menschlicher epistemischer Autorität gegenzuprüfen.[14]

Aber – so könnte eingewandt werden – handelt es sich in diesen Fällen wirklich um „Vertrauen"? Eine mögliche alternative Beschreibung könnte wie folgt aussehen: Was hier eigentlich geschieht, ist, dass sich die Personen lediglich auf die KI-Systeme *verlassen*. Wir mögen zwar manchmal auch in solchen Fällen von „Vertrauen" reden, aber das ist lediglich eine verkürzte, metaphorische, uneigentliche oder deflationäre Verwendungsweise dieses Wortes, das sich in der Alltagssprache manchmal schlicht auch syno-

14 Speziell im Hinblick auf die Verwendung von KI im medizinischen Kontext haben Ferrario, Loi und Viganò einen Ansatz vorgeschlagen, der eine gewisse Verwandtschaft mit demjenigen von Nguyen aufweist. Ihre Behauptung lautet: „[T]o trust a medical AI is to rely on it with little monitoring and control of the elements that make it trustworthy" (2021, 437).

nym mit „sich verlassen“ verwenden lässt.[15] Ich denke allerdings, dass diese alternative Beschreibung zu kurz greift. Wenn wir uns nochmals die Zweidimensionalität unserer Vertrauenskonzeption vor Augen führen, dann lässt sich veranschaulichen, dass sich Vertrauen – verstanden als nicht-hinterfragende Akzeptanz – sehr wohl sinnvoll von bloßem Sich-Verlassen abgrenzen lässt. Bloßes Sich-Verlassen entspricht genau *einer* der beiden Dimensionen, nämlich der Akzeptanz. Warum entspricht es der Akzeptanz und nicht dem Nicht-Hinterfragen? Zum einen, weil ich eine Haltung des Nicht-Hinterfragens an den Tag legen kann, ohne mich auf die Person oder das Objekt zu verlassen, z. B. in Fällen, in denen ich gegenüber der Person oder dem Objekt indifferent bin. Zum anderen kann ich mich auf etwas verlassen und mich insofern darauf stützen und in meinem Handeln oder meiner Deliberation darauf einrichten, was aber noch nicht heißt, dass ich ihm gegenüber eine Haltung des Nicht-Hinterfragens einnehmen müsste. Hier sind besonders Fälle instruktiv, in denen ich gar keine andere Wahl habe, als mich *nolens volens* auf etwas oder jemanden zu verlassen, während ich zugleich der Sache oder Person gegenüber eine zögerliche, skeptische Haltung aufrechterhalte. Nguyen gibt dafür Beispiele wie das folgende: „If I am walking across a muddy field, riddled with gopher holes, I'll examine the ground carefully before each step. And even when I do force myself to rely on that muddy ground – when I commit my weight to it – I don't, as yet, trust it. My reliance is tentative and demands constant reassurance. But when I trust the ground, I stop worrying about it“ (2023, 219). Vertrauen beginnt dann, wenn ich meine skeptische Haltung ablege und beginne, die Sache oder Person nicht mehr zu hinterfragen. Das oben erwähnte klassische Definitionsschema für Vertrauen, dem zufolge Vertrauen als ein Sich-Verlassen *plus* X zu analysieren ist, ist hier unmittelbar anwendbar: Vertrauen ist Akzeptieren (Sich-Verlassen) *plus* Nicht-Hinterfragen.

15 Diese Strategie, die alltagssprachliche Redeweise vom „Vertrauen“ in technische Artefakte zu erklären, findet sich etwa bei Lahno, der „nicht bestreiten [will], dass das Verb ‚vertrauen‘ in manchen Zusammenhängen korrekterweise synonym zu ‚sich verlassen‘ benutzt wird“ (2002, 38). Ähnlich argumentiert Budnik, „dass das Vokabular des Vertrauens zwei Begriffe abdeckt, die zwar im Gegensatz zum Fall von Homonymen wie ‚Bank‘ nicht völlig unabhängig sind, aber doch voneinander unterschieden werden sollten“ (2021, 15), nämlich zum einen genuines Vertrauen und zum anderen bloßes Sich-Verlassen, und dass wenn von Vertrauen in technische Artefakte die Rede ist, dies stets im Sinne des zweiten Falls interpretiert werden müsse.

Nun lässt sich fragen, welchen Wert der Aspekt des Nicht-Hinterfragens eigentlich haben sollte, der nicht schon durch das Akzeptieren bzw. Sich-Verlassen realisiert wäre? Nguyens aus meiner Sicht plausible Antwort auf diese Frage lautet, dass sich dieser Wert aus dem Beitrag ergibt, den das Nicht-Hinterfragen für unsere Funktionsfähigkeit als handelnde Akteure liefert. Einer Person oder Sache zu vertrauen, läuft demnach darauf hinaus, sie in das eigene Handeln zu integrieren, um dadurch ein höheres Maß an Handlungsfähigkeit zu erreichen. Wenn wir jeden potentiellen Beitrag, den eine Sache oder eine andere Person für unser Handeln liefern kann, immer erst nach ausführlichem Prüfen akzeptieren würden, würde uns das schnell paralysieren (ähnlich wie es uns paralysieren würde, wenn wir alle Informationen, die uns unsere Sinnesorgane liefern, erst explizit prüfen wollten). Wir wären nicht mehr in der Lage, uns effizient in der Welt zu bewegen. Vertrauen – und dabei insbesondere der Aspekt des Nicht-Hinterfragens – ermöglicht uns eine Form von Handlungsfähigkeit, die anderweitig kaum erreichbar wäre.

Es sei nochmals betont, dass die Analyse von Vertrauen als Haltung nicht-hinterfragender Akzeptanz nicht ausschließt, dass auch Spielarten des Vertrauens existieren, die im Sinne der affektiven oder normativen (oder vielleicht auch weiterer) Ansätze analysiert werden können. Die verschiedenen Formen von Vertrauen können sich vielleicht auch überlagern und wechselseitig verstärken (beispielsweise könnte mein Vertrauen in eine Person die Form einer nicht-hinterfragenden Akzeptanz annehmen, weil ich die Person affektiv mir gegenüber als positiv motiviert wahrnehme, was für die Person wiederum eine zusätzliche Motivation sein kann, sich mir gegenüber vertrauenswürdig zu erweisen). Aber die Formen können auch separat auftreten, und genau dies ist der Fall, wenn wir KI-Systemen oder anderen technischen Artefakten vertrauen, die die von den affektiven und normativen Ansätzen postulierten Bedingungen zwar nicht erfüllen, wohl aber Objekte einer nicht-hinterfragenden Akzeptanz sein können.

3 Künstliche Intelligenz und epistemische Wachsamkeit

3.1 Vom Nutzen und Nachteil des Vertrauens für den Umgang mit künstlicher Intelligenz

Ich habe im vorigen Abschnitt zu zeigen versucht, dass es kein Kategorienfehler ist, von „Vertrauen in KI“ zu sprechen. Auch wenn aktuelle KI-Systeme nicht die von affektiven oder normativen Vertrauenskonzeptionen

postulierten Bedingungen erfüllen, so kommen sie grundsätzlich gleichwohl als Objekte von Vertrauen im Sinne einer Haltung nicht-hinterfragender Akzeptanz infrage.[16] Das heißt freilich noch nicht, dass jedes KI-System ein geeignetes Vertrauensobjekt ist. Unter menschlichen Personen gibt es vertrauenswürdige und nicht vertrauenswürdige, und für KI-Systeme gilt dasselbe; entsprechend kann ein Vertrauen in ein solches System angemessen oder unangemessen sein.

Was unterscheidet vertrauenswürdige von nicht-vertrauenswürdigen KI-Systemen? Ich denke, dass man ein Vertrauen in ein KI-System in dem Maße als angemessen bezeichnen kann, in dem es begründet ist, und gute Gründe für Vertrauen wiederum lassen sich, zumindest in typischen Fällen, in erster Linie durch hinreichend gründliches Prüfen des Systems generieren. Aktives Prüfen und Hinterfragen ist nach unserer Bestimmung zwar mit Vertrauen inkompatibel, diese Inkompatibilität gilt aber nur für den Fall einer zeitlichen Koinzidenz: Eine Person kann nicht zu *ein und demselben Zeitpunkt* prüfen und vertrauen; sie kann aber sehr wohl eine Prüfung durchgeführt haben, zu einem positiven Resultat gekommen sein und im *Anschluss* daran eine Haltung nicht-hinterfragender Akzeptanz ausgebildet haben.[17] Ich kann hier nicht auf die Details dessen eingehen, was es heißt, ein KI-System so zu prüfen, dass es angemessen ist, ihm zu vertrauen (zumal die Details je nach Typ des Systems erheblich variieren dürften). Ganz allgemein wird man aber wohl sagen können, dass die Prüfung eines Systems darin besteht, seine tatsächliche Performance im Hinblick auf bestimmte Parameter

16 Das heißt im Umkehrschluss nicht, dass Sprechakte der Form „Ich vertraue dem KI-System X“ unmöglich Kategorienfehler involvieren können. Würde man etwa eine Person, die einen solchen Sprechakt tätigt, fragen, was sie damit genau meint, und aus ihrer Antwort ginge hervor, dass sie unterstellt, das KI-System hätte emotionale Einstellungen oder ein normatives Verantwortungsbewusstsein, dann wird man ihr sicher durchaus einen Kategorienfehler vorwerfen können. Wenn die Person aber Vertrauen im Sinne einer Haltung nicht-hinterfragender Akzeptanz im Sinn hat, dann liegt kein Kategorienfehler vor.

17 Beispielsweise könnte es bei einem KI-System zur medizinischen Diagnostik eine Phase vor der eigentlichen Verwendung geben, in der es gründlich auf seine Zuverlässigkeit hin geprüft wird. Wenn es diese Prüfung mit positivem Resultat überstanden haben sollte, könnte es dann in den medizinischen Alltag integriert werden. Dem System zu vertrauen, könnte dann etwa bedeuten, seine Diagnosen zu akzeptieren, ohne in jedem Einzelfall wieder eine Prüfprozedur zu wiederholen.

mit der gewünschten abzugleichen. Ich möchte nicht ausschließen, dass es, je nachdem, welche Parameter für relevant befunden werden, hilfreich oder sogar erforderlich sein kann, die interne Funktion des Systems transparent zu machen. Die Pointe meiner Überlegungen in Abschnitt 2.1 war nicht, dass Vertrauen *niemals* Transparenz erfordert; sie war lediglich, dass es auch ohne Transparenz *möglich* sein kann, die Performance eines Systems soweit einzuschätzen, dass eine hinreichend präzise Beurteilung seiner Vertrauenswürdigkeit erfolgen kann. Man denke etwa an Fälle, in denen die relevanten Parameter lediglich die zuverlässige Transformation einer bestimmten Klasse von Input-Daten in eine bestimmte Klasse von Output-Daten betreffen, z. B. bei der Erkennung von Strukturen in visuellen Mustern. Hier könnte es für eine Beurteilung der Zuverlässigkeit des Systems u.U. genügen, seine Klassifikationsentscheidungen in einer hinreichend großen Anzahl von Fällen zu beobachten und auf ihre Korrektheit hin zu beurteilen.

Den gerade gemachten Ausführungen könnte eine gewisse individualistische Schlagseite vorgeworfen werden. Das Bild individueller User einzelner KI-Systeme ist freilich unterkomplex. Zur Korrektur möchte ich noch zwei Punkte ergänzen, um die (potentielle) soziale Dimension des Vertrauens in KI sichtbar zu machen. Erstens muss die gerade diskutierte Prüfung von KI-Systemen nicht zwingend eine individuelle Aufgabe sein; sie kann auch als arbeitsteiliger Prozess realisiert werden. Mit anderen Worten, wenn ich ein KI-System nutzen möchte, dann muss ich die Prüfung nicht unbedingt selbst durchführen, sondern könnte mich dabei auch auf Dritte stützen, die die Prüfung durchgeführt haben. Hier wäre etwa an Prüf- und Zertifizierungsstellen zu denken, die entsprechende Resultate und Empfehlungen bereitstellen. Das setzt freilich voraus, dass solche Stellen in hinreichendem Maße existieren und ihrerseits vertrauenswürdig sind. Die Europäische Kommission spricht in einem White Paper von 2020 davon, dass die Schaffung von Vertrauen in KI eine gesamtgesellschaftliche und politische Aufgabe sei und ein „Ökosystem des Vertrauens" („ecosystem of trust") geschaffen werden solle (Europäische Kommission 2020; vgl. auch Durante 2010, der die verwandte Idee eines „web of trust" verfolgt hat). Ich denke, dass *ein* wichtiger Aspekt eines solchen Ökosystems die Schaffung von solchen Stellen und die Gewährleistung, dass sie vertrauenswürdige Resultate liefern, sein könnte.

Zweitens kann die Redeweise von der „Vertrauenswürdigkeit eines KI-Systems" nahelegen, dass es sich dabei um eine Art absolute, intrinsische Eigenschaft solcher Systeme handeln würde, die in allen Kontexten und be-

zogen auf alle Perspektiven gleich zu beurteilen ist. Tatsächlich aber scheint die Vertrauenswürdigkeit eines KI-Systems eine akteursrelative Eigenschaft zu sein, die auf die Interessen der unterschiedlichen Stakeholder bezogen werden muss, die von dem System in unterschiedlicher Weise betroffen sein können. Betrachten wir zur Illustration ein Beispiel mit zwei Gruppen von Stakeholdern, einem Online-Shop (S) und seinen Kundinnen und Kunden (K). Angenommen, S setze einen Empfehlungsalgorithmus ein, der K, basierend auf ihrem bisherigen Kaufverhalten, weitere Produkte zum Kauf empfiehlt. Ob dieser Algorithmus als vertrauenswürdig eingeschätzt wird, kann nun davon abhängen, welche Interessen die Stakeholder haben, und diese müssen bei S und K nicht identisch sein. Während (wie wir annehmen wollen) das primäre Interesse von S etwa darin bestehen könnte, maximalen Umsatz zu erzielen, könnte das primäre Interesse von K darin bestehen, möglichst preiswerte und relevante Produkte empfohlen zu bekommen und nicht zu unnötigen Käufen verleitet zu werden. Wenn der Algorithmus aus Sicht von S perfekt funktioniert und in diesem Sinne vertrauenswürdig ist, muss er das nicht zwingend auch aus Perspektive von K sein (und umgekehrt).

Ich möchte noch auf eine grundlegende, auch für individualistische Settings relevante, kritische Dimension des Vertrauens eingehen, die auf ein Grundproblem des Vertrauens-Enthusiasmus verweist. Wie im vorigen Abschnitt betont, erfüllt Vertrauen einerseits wichtige positive Funktionen und ist in einem gewissen Maße für endliche Wesen wie uns Menschen schlicht unverzichtbar. Es würde auf eine Paralyse hinauslaufen, wollte man in jedem Moment hinterfragen, ob das, worauf man angewiesen ist, auch tatsächlich funktioniert. Vertrauen ist eine Strategie, um mit unserer Begrenztheit umzugehen und unsere begrenzten kognitiven Ressourcen sinnvoll einzusetzen. Indem wir die Überprüfung irgendwann für beendet ansehen und anfangen, im Sinne einer nicht-hinterfragenden Akzeptanz schlicht zu vertrauen, können wir die freigewordenen Ressourcen, die für die Überprüfung notwendig waren, nunmehr für andere Projekte einsetzen.

Jedes Vertrauen geht allerdings zugleich mit Risiken einher. Beispielsweise kann die Prüfung eines KI-Systems fehlerhaft gewesen sein: Es kann *irrtümlicherweise* als funktionsfähig und vertrauenswürdig eingeschätzt worden sein. Aber auch wenn die Prüfung zunächst zu dem *korrekten* Ergebnis gekommen ist, dass das System funktionsfähig ist, kann die Funktionsfähigkeit mit der Zeit nachlassen oder ganz verschwinden – sei es aufgrund technischer Defekte, sei es aufgrund sich verändernder äußerer Umstände.

Eine Besonderheit von auf maschinellem Lernen beruhenden KI-Systemen gegenüber anderen technischen Artefakten ist etwa die Möglichkeit, dass ein solches System während seines Einsatzes irgendwann mit einer Klasse von Daten konfrontiert wird, die sich in bestimmten systematischen Hinsichten von der Klasse der Daten, mit denen es ursprünglich trainiert wurde, unterscheidet; obwohl es im Zusammenhang mit der ursprünglichen Datenklasse gut funktioniert hat, ist es möglich, dass es dadurch zu einer fehlerhaften Performance kommt. Eine weitere Hinsicht, in der es eine Diskrepanz zwischen der tatsächlichen und der gewünschten Performance geben kann, besteht darin, dass beim Design des KI-Systems bewusst bestimmte Funktionen implementiert wurden, die den Wünschen und Wertvorstellungen bestimmter das System verwendender Personen zuwiderlaufen.[18]

Diese Überlegungen legen nahe, dass man KI-Systemen niemals *blind* vertrauen sollte, wobei ich unter *blindem Vertrauen* eine Form von Vertrauen verstehen möchte, bei der ausnahmslos jeder Output oder jede Performance des Systems fraglos hingenommen wird. Lässt sich ein vernünftiger Mittelweg finden zwischen den beiden Extremen des blinden Vertrauens einerseits und einem gänzlichen Verzicht auf Vertrauen bzw. einem grundsätzlichen Misstrauen andererseits, bei dem man auf die oben skizzierten positiven Funktionen des Vertrauens verzichten müsste? Im nächsten Abschnitt argumentiere ich, dass eine Kopplung von Vertrauen und *epistemischer Wachsamkeit* ein vielversprechender Kandidat für einen solchen Mittelweg ist. Ich werde dabei zum einen zu zeigen versuchen, dass epistemische Wachsamkeit *begrifflich kompatibel* ist mit Vertrauen im Sinne einer Haltung nicht-hinterfragender Akzeptanz. Zum anderen möchte ich zeigen, dass Vertrauen tatsächlich auch stets mit epistemischer Wachsamkeit einhergehen sollte – eine Forderung, die ich mithilfe der Formel *Caveat usor* auf den Punkt bringen möchte.

3.2 Caveat usor

Aus der Tradition des römischen Rechts ist das Prinzip *Caveat emptor* überliefert: Wer eine Sache kauft, möge wachsam sein und darauf achten, dass diese Sache nicht mit offensichtlichen Mängeln behaftet ist. Denn wenn die

18 Hier könnte nochmals an das Beispiel des Empfehlungsalgorithmus des Online-Shops gedacht werden; ein weiteres Beispiel wäre die moralisch und/oder juristisch fragwürdige Sammlung und Weiterverwendung von Daten, die beim Betrieb des Systems anfallen.

erworbene Sache Mängel aufweist, die beim Kauf offensichtlich waren, so ist eine nachträgliche Reklamation ausgeschlossen. Ein ähnliches Gebot, Wachsamkeit walten zu lassen, lässt sich auch auf andere Bereiche übertragen.[19] Für den Umgang mit KI möchte ich hier das Prinzip *Caveat usor* formulieren: Wer ein KI-System verwendet, möge wachsam und für eklatante Fehlfunktionen oder fehlerhafte Resultate sensibel sein.

Ein möglicher Einwand gegen die mit dem Prinzip *Caveat usor* geforderte epistemische Wachsamkeit lautet, dass diese Wachsamkeit mit Vertrauen unverträglich sei, zumal wenn Vertrauen als Haltung nicht-hinterfragender Akzeptanz konzipiert wird. Die im vorigen Abschnitt verwendete Strategie, eine Kompatibilität von Vertrauen und (aktiver) Prüfaktivität mit Verweis auf die Möglichkeit einer zeitlichen Diskrepanz von beidem zu etablieren, ist hier nicht abermals anwendbar: Das Prinzip *Caveat usor* verlangt, wachsam zu sein, *während* man vertraut. Aber wie soll man wachsam sein, wenn man gleichzeitig vertraut und sich insofern in einer Haltung des Nicht-Hinterfragens befindet? Meine Erwiderung lautet, dass epistemische Wachsamkeit nicht mit einer *aktiven Suche* nach Anzeichen für Fehlfunktionen oder Unzuverlässigkeit zu verwechseln ist. Eine solche aktive Suche kann tatsächlich nicht zeitgleich mit Vertrauen erfolgen. Sie mag für die erste Phase der Benutzung eines KI-Systems charakteristisch sein – diejenige Phase, in der man das System prüft und noch kein Vertrauen gefasst hat. Epistemische Wachsamkeit muss demgegenüber als *Disposition* verstanden werden, die eine Haltung des Vertrauens durchaus begleiten kann. Genauer: Es handelt sich um die Disposition, eine Haltung nicht-hinterfragender Akzeptanz gegenüber dem System (zumindest vorübergehend) zu suspendieren, wenn bestimmte signifikante Anzeichen einer Fehlfunktion oder Unzuverlässigkeit des Systems auftreten.

Sperber et al., die die Diskussion zu epistemischer Wachsamkeit mit ihrem einflussreichen Aufsatz angestoßen haben, schreiben in diesem Sinne: „Vigilance (unlike distrust) is not the opposite of trust; it is the opposite of blind trust" (2010, 363). Sie illustrieren die Idee, dass Vertrauen und

19 McBrayer (2022) hat jüngst für den Kontext testimonialer Interaktionen das Prinzip *Caveat Auditor* formuliert. Zuboff (2020) hat die Formulierung „Caveat usor" im Kontext ihrer Kritik des Überwachungskapitalismus verwendet. Zuboffs primärer Fokus sind Internetuser, die sich die Bestrebungen von Digitalkonzernen stärker bewusst machen sollten, in oft wenig transparenter Weise Zugriff auf ihre Daten zu bekommen, um diese finanziell ausbeuten zu können.

Wachsamkeit kompatibel sind, ja sich sogar wechselseitig ergänzen, anhand folgender Analogie (2010, 364). Wenn wir einen belebten Bürgersteig entlanggehen, besteht ständig die Gefahr zufälliger oder gar absichtlicher Zusammenstöße. Dennoch vertrauen wir in der Regel unseren Mitmenschen auf der Straße und zögern nicht, uns unter ihnen zu bewegen. Zugleich beobachten wir den Weg der anderen („[w]e monitor the trajectory of others", ebd.) und achten auf eventuell abgelenkte, unachtsame oder aggressive Personen und passen unsere Wachsamkeit automatisch der Umgebung an. Meistens ist sie so niedrig, dass sie nicht in den Fokus des Bewusstseins gerät, aber sie steigt, wenn es die Situation erfordert. Dann sind wir in der Lage, einer anderen Person auszuweichen oder sie darauf aufmerksam zu machen, dass sie ihrerseits wachsamer sein möge.

Wichtig ist an dieser Stelle die Unabhängigkeit von Wachsamkeit und den beiden oben identifizierten konzeptuellen Dimensionen des Vertrauens (Akzeptanz und Nicht-Hinterfragen); aus dieser Unabhängigkeit ergibt sich die eigenständige funktionale Rolle der Wachsamkeit. Wir hatten in Abschnitt 2.3 die Vermutung zum Ausdruck gebracht, dass Vertrauen eine gewisse Abstufbarkeit zulässt, d. h. Vertrauen bezeichnet nicht nur den Zustand, bei dem beide Dimensionen *vollständig* realisiert sind, sondern es scheint kompatibel zu sein mit einer gewissen Reduktion des Akzeptierens und Nicht-Hinterfragens. Allerdings gilt das offenbar nur in engen Grenzen. Sobald die vertrauende Person beginnt, die Vertrauensperson oder das Vertrauensobjekt aktiv zu hinterfragen und zu prüfen, wird schnell ein Punkt erreicht, an dem die Haltung der (bis dato) vertrauenden Person nicht mehr als „Vertrauen" beschrieben werden kann. Aber Vertrauen ist sehr wohl kompatibel auch mit einem ausgeprägten Maß an Wachsamkeit, verstanden als Disposition, in bestimmten relevanten Situationen die vertrauende Haltung zu suspendieren (die Klasse der als relevant definierten Situationen kann unterschiedlich umfangreich sein; je nachdem ist die Wachsamkeit ausgeprägter oder weniger ausgeprägt). Hieraus ergibt sich die funktionale Rolle der Wachsamkeit: Sie vermag es, gewisse Risiken für das vertrauende Subjekt einzuhegen, die (wie wir im vorigen Abschnitt gesehen hatten) mit einer vertrauenden Haltung einhergehen können und die mit einer bloßen *Reduktion* von Vertrauen nicht oder bestenfalls ansatzweise eingehegt werden könnten. Denn eine Abstufung entlang der das Vertrauen konstituierenden Dimensionen würde sehr schnell schlicht zu einem Verschwinden des Vertrauens führen, was dann wiederum den Nachteil mit sich brächte, dass die Person auf die positiven Funktionen des Vertrauens verzichten müsste.

Bei epistemischer Wachsamkeit gegenüber technischen Systemen lassen sich grundsätzlich zwei Dimensionen bzw. Formen unterscheiden, die ich „direkte“ und „indirekte epistemische Wachsamkeit“ nennen möchte. Die erstere betrifft direkt die Plausibilität der Performance oder des Outputs des Systems. Letztere betrifft Informationen, die in mittelbarer Weise nahelegen, dass die Funktionsfähigkeit oder Zuverlässigkeit des Systems negativ beeinträchtigt sein könnte. Betrachten wir, um diese Unterscheidung an einem simplen Beispiel deutlicher zu machen, die Verwendung eines Taschenrechners. Angenommen, ich führe mithilfe eines Taschenrechners eine Reihe von arithmetischen Operationen durch. Ich vertraue meinem Rechner, d. h. ich übernehme seine Resultate einfach, ohne sie nochmals durch schriftliches Rechnen, die Verwendung eines zweiten Rechners o.ä. zu überprüfen. Zugleich bin ich aber epistemisch wachsam, d. h. sensitiv für den Fall, dass es den Resultaten des Taschenrechners an einer gewissen Minimalplausibilität mangeln sollte. Wenn ich beispielweise zwei positive Zahlen addieren möchte, als Ergebnis aber eine negative Zahl angezeigt wird, dann übernehme ich nicht blind auch dieses Resultat, sondern suspendiere meine nicht-hinterfragende Akzeptanz und prüfe, ob der Taschenrechner womöglich eine Fehlfunktion oder ich einen Eingabefehler begangen haben könnte. Das erfordert wohlgemerkt nicht, dass mir die interne Funktionsweise des Taschenrechners transparent sein muss. Meine Wachsamkeit bezieht sich lediglich auf ein Monitoring seiner Outputs im Hinblick auf eine minimale Plausibilität. Dies wäre ein *direkter* Abgleich des Outputs mit dem zu erwartenden Resultat. Bei der *indirekten* Form von Wachsamkeit geht es dagegen um eine Sensitivität für Anzeichen, die mittelbar auf eine Unzuverlässigkeit des Taschenrechners hindeuten könnten. Wenn ich etwa bemerke, dass an dem Rechner ungewöhnliche Einstellungen aktiviert sind, die sich auf seine Funktionsweise auswirken könnten, dann scheint es ebenfalls rational zu sein, meine nicht-hinterfragende Akzeptanz vorübergehend zu suspendieren und zu prüfen, was es mit den Einstellungen auf sich hat. Ebenfalls in die Kategorie der indirekten Wachsamkeit gehört eine Sensitivität für Informationen, die eventuelle Fehlkonstruktionen oder Fehlfunktionen *anderer* Geräte derselben Baureihe betreffen, oder zum Beispiel auch solche Informationen, die die Vertrauenswürdigkeit der *Firma* betreffen, die das Gerät hergestellt hat.

Ich möchte noch einige Bemerkungen dazu anschließen, welche Voraussetzungen, Potentiale und Grenzen das Ausüben epistemischer Wachsamkeit hat. Zum einen muss bedacht werden, dass epistemische Wach-

samkeit nicht voraussetzungslos ist. In ihrer *direkten* Form setzt sie die Fähigkeit voraus, die Performance bzw. den Output des Systems auf eine Minimalplausibilität hin einschätzen zu können. Nun kann diese Fähigkeit mehr oder weniger gut bei einer Person ausgeprägt sein. Ich brauche zumindest arithmetische Grundkenntnisse, um zu wissen, dass bei der Addition zweier positiver Zahlen das Ergebnis nicht negativ sein kann. Sollte ich dieses Wissen nicht besitzen, oder wenn wir alternativ Operationen betrachten, die anspruchsvoller sind als einfaches Addieren, dann ist es schwieriger, einen direkten Abgleich des Outputs mit dem zu erwartenden Resultat vorzunehmen. Das heißt aber nicht, dass er gänzlich unmöglich ist und ich nicht doch epistemische Wachsamkeit walten lassen sollte.[20] Die Fähigkeit, epistemische Wachsamkeit (in ihrer direkten Form) walten zu lassen, ist abhängig von meiner speziellen Expertise, meiner Bekanntschaft mit dem betreffenden Bereich – im vorliegenden Beispiel der Arithmetik. Expertise und Bekanntschaft sind nicht entweder gar nicht oder vollumfänglich vorhanden, sondern sie sind typischerweise mehr oder weniger stark ausgeprägt. Wenn sie weniger stark bei mir ausgeprägt sind, dann verschiebt das die Grenze des Bereichs, in dem ich einen fehlerhaften Output bemerken würde, und nur dieser Bereich ist für die Wachsamkeit, die ich an den Tag legen kann, relevant (die minimale Abweichung zwischen dem Output und dem korrekten Resultat, ab der ich einen Fehler bemerke, ist umso kleiner, je mehr eigene Expertise bzw. Bekanntschaft mit dem betreffenden Bereich ich besitze). Man könnte dementsprechend auch sagen: Ob ich viel oder wenig eigene Expertise bzw. Bekanntschaft mit dem betreffenden Bereich besitze, so tue ich doch gut daran, epistemische Wachsamkeit (in ihrer direkten Form) walten zu lassen; aber die konkrete Ausgestaltung der Wachsamkeit, insbesondere die Größe des Fehlertoleranzbereichs, innerhalb dessen meine Wachsamkeit

20 Wie lässt sich der normative Status dieser Forderung, von der hier die Rede ist, verstehen? Ich denke, es wäre unplausibel, ihn deontisch zu interpretieren. Ich möchte keine deontische „Pflicht" postulieren, epistemisch wachsam zu sein. Aus meiner Sicht naheliegender wäre entweder eine prudentielle oder eine tugendtheoretische Lesart. In diesem Zusammenhang sei übrigens auf eine gewisse, in dieser Hinsicht bestehende Disanalogie zwischen den Prinzipien *Caveat usor* und *Caveat emptor* hingewiesen. Während letzteres einen Rechtsgrundsatz formuliert, möchte ich ersteres lediglich im Sinne einer Klugheitsregel oder eines Tugendgebots verstanden wissen. Der Status des *Caveat usor*-Prinzips entspricht insofern eher gewissen anderen Adaptionen des *Caveat emptor*-Prinzips, wie etwa dem bereits erwähnten *Caveat Auditor* von McBrayer (2022).

aktivierbar ist, ist sehr wohl vom Umfang meiner Expertise bzw. Bekanntschaft abhängig.

Je weniger Ressourcen mir zur Verfügung stehen, um epistemische Wachsamkeit in ihrer direkten Form walten zu lassen, umso wichtiger wird die indirekte Form. Indirekte epistemische Wachsamkeit betrifft nicht direkt die Performance oder den Output des Systems, sondern eine Sensibilität für Informationen, die indirekte Rückschlüsse auf dessen korrektes Funktionieren zulassen. Eine Möglichkeit wäre etwa, die Zuverlässigkeit der Firma, die das System hergestellt hat, als Indikator heranzuziehen. Vielleicht habe ich mit anderen Systemen, die von derselben Firma hergestellt wurden, gute oder schlechte Erfahrungen gemacht. Eventuell konnte ich die Performance *dieser* Systeme *direkt* einschätzen. Oder ich vertraue der Einschätzung anderer Personen: Wenn Personen, denen ich vertraue, ihrerseits die Firma oder die von ihr hergestellten Systeme als vertrauenswürdig einschätzen, dann ist das typischerweise ein Grund für mich, der Firma oder den Systemen meinerseits zu vertrauen.

Nun sind freilich auch solche Informationen nicht in jedem Fall in hinreichendem Maße zugänglich. Wie im vorigen Abschnitt angedeutet, kann die Bereitstellung entsprechender Informationen auch als gesellschaftliche Aufgabe und als politisches Desiderat begriffen werden. Wenn es vertrauenswürdige Zertifizierungsstellen gäbe, deren Aufgabe darin liegt, die Vertrauenswürdigkeit von KI-Systemen und von Firmen, die sie herstellen, zu beurteilen, dann könnte (indirekte) epistemische Wachsamkeit bei der Benutzung eines KI-Systems auch darin bestehen, sensibel zu sein für die von solchen Stellen bereitgestellten (und vielleicht laufend aktualisierten) Informationen. Die Vertrauenswürdigkeit solcher Zertifizierungsstellen sollte rationalerweise freilich ihrerseits nicht einfach blind hingenommen werden. Je größer die gesellschaftliche Bedeutung von Zertifizierungsstellen (in beliebigen Bereichen) ist, desto stärker sind sie typischerweise ein mögliches Ziel von unredlichen Beeinflussungsversuchen mit dem Hintergrund partikularer (ökonomischer, politischer usw.) Interessen; und auch wenn keine Beeinflussung dieser Art stattfindet, können in der besten Zertifizierungsstelle schlicht Fehler passieren. Ich hatte im vorigen Abschnitt auf das von der Europäischen Kommission geforderte „Ökosystem des Vertrauens" hingewiesen. Ich denke, dass komplementär dazu auch die Etablierung von epistemischer Wachsamkeit in Bezug auf KI als gesamtgesellschaftliches Ziel verfolgt werden sollte. Wir sollten auch daran arbeiten, ein „Ökosystem der Wachsamkeit" zu schaffen.

Epistemische Wachsamkeit ist kein Mittel, das die Risiken, die wir eingehen, wenn wir vertrauen, komplett zum Verschwinden bringen würde. Wie gerade gesehen, setzt epistemische Wachsamkeit sowohl in ihrer direkten als auch in ihrer indirekten Form Kompetenzen oder Informationen voraus, die häufig nur begrenzt verfügbar sind. Epistemische Wachsamkeit besteht nicht darin, sich auf unfehlbare Mechanismen zur Detektion von Fehlern zu stützen, sondern sie beruht auf einfachen, unsere kognitiven Ressourcen schonenden und falliblen Heuristiken, die sowohl falsch-positive als auch falsch-negative Resultate hervorbringen können. Wenn ich z. B. die Disposition habe, mein Vertrauen in den Taschenrechner zu suspendieren, wenn er ein negatives Resultat anzeigt, obwohl ich zwei natürliche Zahlen addieren wollte (wenn ich also die Heuristik *Wenn zwei natürliche Zahlen addiert werden, kann das Ergebnis nicht negativ sein* anwende), dann bewahrt mich dies davor, eine Reihe von Resultaten zu akzeptieren, denen es an einer minimalen Plausibilität fehlt. Aber es bewahrt mich nicht davor, *alle* denkbaren inkorrekten Resultate korrekt zu identifizieren (ich kann mit dieser Heuristik z. B. keine falschen Resultate im Bereich der positiven Zahlen identifizieren). Umgekehrt gilt, dass wenn die Heuristik tatsächlich „anschlägt" (d. h. wenn ich auf einen möglichen Fehler aufmerksam geworden bin), daraus noch nicht automatisch folgt, dass ich den Output des Systems zwangsläufig verwerfen sollte. Es heißt nur, dass es geboten ist, meine nicht-hinterfragende Akzeptanz (vorübergehend) zu suspendieren und den Output nochmals explizit zu prüfen. Dabei kann sich herausstellen, dass tatsächlich ein Fehler vorliegt, aber es kann sich auch herausstellen, dass der Output korrekt war und ich in meine nicht-hinterfragende Akzeptanz zurückfinden kann.

Trotz dieser Limitationen trägt epistemische Wachsamkeit zur Einhegung der Risiken bei, die wir eingehen, wenn wir vertrauen. Sie ist ein wichtiges Mittel des vernünftigen Umgangs mit diesen Risiken und reduziert unsere Verletzlichkeit im Vergleich mit Situationen, in denen unser Vertrauen *blind* ist.

4 Schluss

Ich habe mich in diesem Aufsatz kritisch mit dem Vertrauens-Enthusiasmus und der Vertrauens-Skepsis auseinandergesetzt. Beide sind in der gegenwärtigen öffentlichen bzw. philosophischen Diskussion zu KI verbreitete Positionen, die allerdings, wie ich zu zeigen versucht habe, mit grundlegen-

den Einwänden konfrontiert sind. Gegen die skeptische Position habe ich argumentiert, dass es kein Kategorienfehler ist, von „Vertrauen in KI-Systeme“ zu sprechen. Zwar gibt es Formen von Vertrauen, die auf interpersonale Beziehungen beschränkt sind; andere Formen von Vertrauen – insbesondere Vertrauen als Haltung nicht-hinterfragender Akzeptanz – sind jedoch auch gegenüber KI-Systemen möglich. Ferner habe ich gegen die Vertrauens-Skepsis eingewandt, dass auch die mit KI einhergehende epistemische Opazität kein prinzipieller Hinderungsgrund für Vertrauen ist.

Im Gegensatz dazu habe ich gegen den Vertrauens-Enthusiasmus eingewandt, dass dieser ein zu unkritisches Bild von Vertrauen zeichnet und dazu tendiert, dessen Risiken und Schattenseiten zu vernachlässigen. Ich habe dem enthusiastischen Bild das Prinzip *Caveat usor* entgegengesetzt und argumentiert, dass Vertrauen stets mit epistemischer Wachsamkeit einhergehen sollte. Epistemische Wachsamkeit bringt die mit Vertrauen verbundenen Risiken zwar nicht in Gänze zum Verschwinden. Aber sie hilft dabei, einen vernünftigen Mittelweg zu finden zwischen den beiden suboptimalen Extremen eines *blinden* Vertrauens in KI-Systeme einerseits und einer kategorischen *Abwesenheit* von Vertrauen in sie andererseits.

Literatur

Baier, Annette. 1986. „Trust and antitrust“. *Ethics* 96 (2): 231–60. https://doi.org/10.1086/292745.

Boge, Florian J. 2022. „Two Dimensions of Opacity and the Deep Learning Predicament“. *Minds and Machines* 32: 43–75. https://doi.org/10.1007/s11023-021-09569-4.

Braun, Matthias, Hannah Bleher und Patrik Hummel. 2021. „A Leap of Faith: Is There a Formula for ‚Trustworthy‘ AI?“ *Hastings Center Report* 51 (3): 17–22. https://doi.org/10.1002/hast.1207.

Budnik, Christian. 2021. *Vertrauensbeziehungen. Normativität und Dynamik eines interpersonalen Phänomens*. Berlin/Boston: Walter de Gruyter.

Durán, Manuel und Karin R. Jongsma. 2021. „Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI“. *Journal of Medical Ethics* 47 (5): 329–35. http://dx.doi.org/10.1136/medethics-2020-106820.

Durante, Massimo. 2010. „What is the model of trust for multi-agent systems? Whether or not e-trust applies to autonomous agents“. *Knowledge Technology & Policy* 23: 347–66. https://doi.org/10.1007/s12130-010-9118-4.

Europäische Kommission. 2020. „On Artificial Intelligence: A European Approach to Excellence and Trust“. Brüssel. https://commission.europa.eu/system/files/2020-02/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

Ferrario, Andrea, Michele Loi und Eleonora Viganò. 2021. „Trust does not need to be human: It is possible to trust medical AI“. *Journal of Medical Ethics* 47 (6): 437–8. DOI:10.1136/medethics-2020-106922.

Freiman, Ori. 2023. „Making sense of the conceptual nonsense ‚trustworthy AI‘“. *AI and Ethics* 3: 1351–60. https://doi.org/10.1007/s43681-022-00241-w.

Goldberg, Sanford C. 2020. „Epistemically engineered environments“. *Synthese* 197: 2783–802. https://doi.org/10.1007/s11229-017-1413-0.

Grote, Thomas. 2021. „Trustworthy medical AI systems need to know when they don’t know“. *Journal of Medical Ethics* 47 (5): 337–8. http://dx.doi.org/10.1136/medethics-2021-107463.

Hartmann, Martin. 2011. *Die Praxis des Vertrauens*. Berlin: Suhrkamp.

Hatherley, Joshua J. 2020. „Limits of trust in medical AI“. *Journal of Medical Ethics* 46 (7): 478–81. http://dx.doi.org/10.1136/medethics-2019-105935.

Hauswald, Rico. 2024. *Epistemische Autoritäten: Individuelle und plurale*. Berlin: Metzler.

Hawley, Katherine. 2014. „Trust, Distrust and Commitment“. *Noûs* 48 (1): 1–20. https://doi.org/10.1111/nous.12000.

High-Level Expert Group on AI. 2019. „Ethics Guidelines for Trustworthy AI“. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Kerasidou, Charalampia, Angeliki Kerasidou, Monika Buscher und Stephen Wilkinson. 2022. „Before and beyond trust: Reliance in medical AI“. *Journal of Medical Ethics* 48 (11): 852–6. http://dx.doi.org/10.1136/medethics-2020-107095.

Krügel, Sebastian, Andreas Ostermaier und Matthias Uhl. 2022. „Zombies in the Loop? Humans Trust Untrustworthy AI-Advisors for Ethical Decisions“. *Philosophy & Technology* 35 (17). https://doi.org/10.1007/s13347-022-00511-9.

Lahno, Bernd. 2002. *Der Begriff des Vertrauens*. Paderborn: mentis.

Lang, Benjamin. 2021. „Concerning a seemingly intractable feature of the accountability gap“. *Journal of Medical Ethics* 47 (5): 336. http://dx.doi.org/10.1136/medethics-2021-107353.

Luhmann, Niklas. 2014. *Vertrauen: Ein Mechanismus der Reduktion sozialer Komplexität*. 5. Aufl. Konstanz: UVK-Verl.-Ges.

Maclure, Jocelyn. 2021. „AI, Explainability and Public Reason: The Argument from the Limitations of the Human Mind“. *Minds and Machines* 31: 421–38. https://doi.org/10.1007/s11023-021-09570-x.

McBrayer, Justin P. 2022. „Caveat Auditor: Epistemic Trust and Conflicts of Interest“. *Social Epistemology*. https://doi.org/10.1080/02691728.2022.2078250.

Mishra, Abhishek. 2021. „Transparent AI: reliabilist and proud“. *Journal of Medical Ethics* 47 (5): 341–2. http://dx.doi.org/10.1136/medethics-2021-107352.

Moor, James H. 2006. „The Nature, Importance, and Difficulty of Machine Ethics“. *IEEE Intelligent Systems* 21 (4): 18–21. https://doi.org/10.1109/MIS.2006.80.

Nguyen, C. Thi. 2023. „Trust as an unquestioning attitude“. In *Oxford Studies in Epistemology,* Vol. 7, hg. v. Tamar Szabó Gendler, John Hawthorne und Julianne Chung, 214–44. Oxford: Oxford University Press.

Reinhardt, Karoline. 2023. „Trust and trustworthiness in AI ethics“. *AI and Ethics* 3: 735–44. https://doi.org/10.1007/s43681-022-00200-5.

Ryan, Mark. 2020. „In AI We Trust: Ethics, Artificial Intelligence, and Reliability“. *Science and Engineering Ethics* 26: 2749–67. https://doi.org/10.1007/s11948-020-00228-y.

Sperber, Dan, Fabrice Clément, Christophe Heintz, Olivier Mascaro, Hugo Mercier, Gloria Origgi und Deirdre Wilson. 2010. „Epistemic Vigilance“. *Mind and Language* 25 (4): 359–93. https://doi.org/10.1111/j.1468-0017.2010.01394.x.

Véliz, Carissa, Carina Prunkl, Milo Phillips-Brown und Theodore M. Lechterman. 2021. „We might be afraid of black-box algorithms“. *Journal of Medical Ethics* 47 (5): 339–40. http://dx.doi.org/10.1136/medethics-2021-107462.

von Eschenbach, Warren J. 2021. „Transparency and the Black Box Problem: Why We Do Not Trust AI“. *Philosophy & Technology* 34: 1607–22. https://doi.org/10.1007/s13347-021-00477-0.

Zagzebski, Linda T. 2012. *Epistemic Authority: A Theory of Trust, Authority, and Autonomy in Belief*. New York: Oxford University Press.

Zednik, Carlos. 2021. „Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence“. *Philosophy & Technology* 34: 265–88. https://doi.org/10.1007/s13347-019-00382-7.

Zuboff, Shoshana. 2020. „Caveat Usor: Surveillance Capitalism as Epistemic Inequality“. In *After the Digital Tornado*, hg. v. Kevin Werbach, 174–214. Cambridge: Cambridge University Press.