Zeitschrift für Praktische Philosophie
Band 11, Heft 1, 2024, S. 325–342
www.praktische-philosophie.org
https://doi.org/10.22613/zfpp/11.1.13

Einleitung: Neue Vertrauensfragen? Digitalisierung und Künstliche Intelligenz

Introduction: New questions of trust? Digitization, Digitalization and Artificial Intelligence

Karoline Reinhardt, Passau & Johanna Sinn, Passau

Zusammenfassung: Vertrauen ist ein wesentlicher Bestandteil menschlichen Lebens und menschlicher Interaktionen. Der vorliegende Schwerpunkt untersucht, ob Digitalisierung und Künstliche Intelligenz neue Vertrauensfragen aufwerfen, ob das Konzept des Vertrauens auf digitale Technologien angewandt werden kann und inwiefern es modifiziert werden muss, um den begrifflichen wie praktischen Herausforderungen, die mit dem Einsatz dieser Technologien verbunden sind, gerecht zu werden. Die Einleitung stellt einführend dar, wie Vertrauen bislang in der philosophischen Debatte verstanden wurde und welche neuen Fragen durch digitale Technologien aufgeworfen werden. Es wird herausgearbeitet, dass, obwohl Vertrauen schon lange ein Thema der Philosophie war, insbesondere im Bereich digitaler Technologien Unsicherheit und Verletzbarkeit neue Formen annehmen, die Implikationen für Vertrauensbeziehungen haben und die es philosophisch zu reflektieren gilt. Die Beiträge des Themenschwerpunkts befassen sich dabei gleichermaßen mit übergreifenden Fragen zu Vertrauen im Kontext digitaler Technologien, wie auch mit Fragen des Vertrauens in konkreten Anwendungsfeldern wie der Medizin und Blockchaintechnologien.

Schlagwörter: Vertrauen, Digitalisierung, Künstliche Intelligenz, Blockchain, Unsicherheit, Vulnerabilität

Abstract: Trust is an essential component of human life and interactions. The special issue examines whether digitalization and artificial intelligence (AI) raise new questions of trust, whether the concept of trust can be applied to digital technologies, and

to what extent it must be modified to address the conceptual and practical challenges associated with these technologies. The introduction provides an overview of how trust has been understood in the philosophical debate so far and what new questions are raised by digital technologies. It highlights that although trust has long been a subject of philosophy, uncertainty, risk and vulnerability take on new forms in the field of digital technologies, which have implications for trust relationships that call for philosophical inquiry. The contributions to this special issue address overarching questions about trust in the context of digital technologies as well as issues of trust in specific fields of application, such as medicine and blockchain technologies.

Keywords: trust, digitization, digitalization, artificial intelligence, blockchain, vulnerability, risk

Vertrauen ist ein wesentlicher Bestandteil menschlichen Lebens und menschlicher Interaktionen: Wir vertrauen einander nicht nur in engen menschlichen Beziehungen, sondern auch in Alltagszusammenhängen. Im Verkehr vertrauen wir darauf, dass alle anderen überwiegend mit Umsicht und unter Beachtung der Regeln fahren oder Bus- und Taxifahrer:innen uns sicher befördern. Ebenso sind viele Bereiche von vielfältigen und zum Teil sehr spezifischen Vertrauensverhältnissen und -praktiken geprägt, etwa bei Vertragsschlüssen und ökonomischen Interaktionen[1], in pädagogischen Kontexten und Lehrsituationen[2], im Theater[3], in der Wissenschaft[4] oder in der Medizin[5]. Vertrauen ermöglicht, sich zu orientieren und zu handeln. Ohne Vertrauen wäre alles möglich: „Solch eine unvermittelte Konfrontierung mit der äußersten Komplexität der Welt hält kein Mensch aus“ (Luhmann 1968, 1).

Obwohl in alltäglichen Situationen weit verbreitet und Grundlage vielfältiger Interaktionen, ist Vertrauen „bekanntermaßen ein empirisch schwer zugängliches Phänomen“ (Hartmann 2001, 8). Dies liegt auch daran, dass eine Vielzahl von Verhältnissen als Vertrauensbeziehungen beschrie-

1 Fukuyama (1995).

2 U. a. Fisher/Tallant (2019), Reichenbach (2020).

3 Baier (2001, 45).

4 U. a. Hartwig (1991), Krämer (2009), Oreskes (2019), Rolin (2020).

5 O’Neill (2002), Calnan/Rowe (2008), Steinfath/Wiesemann (2016), Wolfensberger (2019), Nickel/Frank (2020).

ben werden kann und sich Vertrauen in unterschiedlichen Kontexten unterschiedlich konstituiert, zeigt und „anfühlt". Das Anvertrauen von Geheimnissen kann persönliche Nähe schaffen und mit Zuneigung verbunden sein. Das Vertrauen darauf, dass jemand eine mir wichtige Aufgabe übernimmt, kann dagegen eher ein inneres „Loslassen" bedeuten oder den Charakter einer rationalen Entscheidung annehmen, und Vertrauen in Situationen der Abhängigkeit kann auch mit Sorge verbunden sein.

Vertrauen ist zudem politisch relevant, weil es das Fundament für die Stabilität und Funktionsfähigkeit (demokratischer) Systeme bildet.[6] Ohne Vertrauen zwischen Bürger:innen und ihren politischen Repräsentant:innen sowie zwischen den verschiedenen gesellschaftlichen Gruppen erodiere, so wird häufig angenommen, die Basis für soziale Kohäsion. Politisches Vertrauen ermögliche es, dass Institutionen effektiv arbeiten und politische Entscheidungen legitimiert werden, da die Bürger:innen darauf vertrauen, dass ihre Interessen und Rechte berücksichtigt und geschützt werden. Darüber hinaus fördere Vertrauen die Bereitschaft zur Zusammenarbeit und Kompromissbereitschaft, was in pluralistischen Gesellschaften unerlässlich sei, um politische Konflikte friedlich zu lösen.[7] Gleichzeitig hat das politische Klima Auswirkungen auf personale Vertrauensbeziehungen: Wenn ein Staat seine Bürger:innen zu Denunziationen aufruft, schürt er Misstrauen zwischen Familienmitgliedern, Nachbar:innen oder Freund:innen und erschwert Vertrauensverhältnisse. Auch in politischen Situationen, in denen Ressourcen umkämpft sind, kann der Aufbau von dauerhaften Vertrauensverhältnissen beeinträchtigt sein. Politische Instabilität kann ein weiterer Faktor sein, der personale Vertrauensbeziehungen unterminiert.

Aufgrund der lebensweltlichen, aber vor allem der politischen Relevanz erfährt Vertrauen seit den 1980er Jahren verstärkte philosophische Aufmerksamkeit. In der anglophonen Welt lässt sich seitdem sogar eine „explosionsartige Veröffentlichungswelle" (Hartmann 2001, 7) beobachten. Die jüngere Debatte um Vertrauen hat in den letzten Jahren noch einmal eine neue Dynamik erhalten, welche unter anderem durch Entwicklungen im Bereich der Computerwissenschaften und der Informationstechnologie

6 Zu Vertrauen als notwendiger Grundlage für das Funktionieren von Gemeinschaften und Gesellschaften s. Hartmann (2011).

7 Zur Diskussion inwiefern und in welchem Ausmaß Vertrauen für das Funktionieren von demokratischen politischen Systemen notwendig ist, s. u. a. die Beiträge in Warren (1999), Lenard (2012), Budnik (2018).

angestoßen wurde und durch entsprechende Bemühungen zur Regulierung dieser neuen technischen Möglichkeiten weiteren Antrieb erhalten hat. Allen voran sind hier computergestützte Algorithmen und Anwendungen zu nennen, die auf Methoden des so genannten maschinellen Lernens basieren und häufig unter dem Begriff Künstliche Intelligenz (KI) gefasst werden. Diese neueren Entwicklungen bringen nicht nur technische, sondern auch eine Reihe ethischer Herausforderungen mit sich, die seit einigen Jahren umfassend diskutiert werden.[8] Seit der Veröffentlichung der Ethics Guidelines for Trustworthy AI (2019) durch die Highlevel Expertgroup on Artificial Intelligence (HLEG), als einem Versuch, mit diesen Entwicklungen regulatorisch umzugehen, hat der Vertrauensbegriff im Bereich der Ethik der Künstlichen Intelligenz (kurz: KI-Ethik) neue Aufmerksamkeit erfahren.

Wirft also das Zeitalter der Digitalisierung, werfen digitale Technologien und Künstliche Intelligenz neue Vertrauensfragen auf? Oder begegnen uns hier die gleichen Fragen unter anderen Vorzeichen? Verkomplizieren diese Technologien Vertrauensbeziehungen und Vertrauensbegriff oder können sie uns vielleicht helfen, manchen Sachverhalt klarer zu sehen? Diesen Fragen widmet sich der vorliegende Themenschwerpunkt.[9]

Vertrauen in der philosophischen Debatte

Das Thema Vertrauen war, entgegen manchen Einschätzungen, durch die europäische Philosophiegeschichte hindurch präsent: Neben Untersuchungen zur Rolle von Vertrauen in Gesellschaften[10] finden sich beispielswei-

8 Für einen Überblick über einige zentrale Fragestellungen s. Bostrom/Yudkowsky (2014) und Mittelstadt et al. (2016) sowie mit Hinblick auf konkrete Anwendungen probabilistischer Modelle O'Neil (2016). Daran hat sich eine Debatte angeschlossen, wie Maschinelles Lernen in einer Weise gestaltet werden sollte, um moralisch unproblematisch gegebenenfalls sogar förderlich für bestimmte gesellschaftliche Zielsetzungen zu sein s. bspw. Floridi (2019). Einen besonderen Schwerpunkt haben die Untersuchungen dabei zunächst auf Verzerrung und Diskriminierung gelegt. Vgl. hierzu etwa Barocas/Selbst (2016), Bozdag (2013), Friedman/Nissenbaum (1996), Hagendorff (2019), Heesen et. al. (2022), Veale/Binns (2017), Zuiderveen Borgesius (2018).

9 Der Titel dieses Schwerpunkts ist inspiriert von Freverts Studie Vertrauensfragen. Eine Obsession der Moderne (2013).

10 Etwa bei Hobbes, *Leviathan*, Kap. 14; Locke, *Two Treatises of Government*, §240; Hume, *A Treatise of Human Nature*; Hegel, *Phänomenologie des Geistes;* oder Rawls (1993).

se Überlegungen zum Selbstvertrauen,[11] zum Vertrauen auf Gott[12] und in Freunde[13]. Innerhalb der Entwicklungspsychologie ist der Begriff des Urvertrauens einschlägig[14] und auch die Ökonomie interessiert die Untersuchung von Vertrauen.[15]

In der philosophischen Debatte zu Vertrauen wird herausgestellt, dass es einerseits Ermöglichungsbedingung zahlreicher menschlicher Interaktionen und Lebensweisen ist, die ohne Vertrauen undenkbar wären, es aber andererseits auch verletzlich macht (Baier 1986): Mit Vertrauen ist die Möglichkeit eines individuellen Schadens verbunden, welcher den Vorteil, den der Vertrauenserweis möglicherweise zeitigen wird, durchaus überwiegen kann. Es handelt sich bei Vertrauen nicht allein um das Ergebnis eines Kosten-Nutzen-Kalküls, bei dem der Nutzen eindeutig überwiegt: „Vertrauen bleibt ein Wagnis" (Luhmann 1968, 31). Wir laufen Gefahr, auf vielfältige Weise verletzt zu werden, von einer enttäuschten Hoffnung bis hin zu physischer oder psychischer Verwundung. Diese Verletzungen können sowohl durch gebrochenes oder enttäuschtes Vertrauen selbst als auch durch darauffolgende Konsequenzen entstehen. In einem nahen, etwa familiären Vertrauensverhältnis ist von enttäuschtem Vertrauen vor allem die Beziehung zu anderen und sich selbst betroffen – man *fühlt* sich verletzt, der Selbstwert kann Schaden nehmen und es kann schwer sein, derselben Person bald wieder zu vertrauen. Die Vertraulichkeit einer Information in beruflichen Kontexten zu missachten, kann sowohl die Beziehungsebene im Team betreffen als auch zum Scheitern von Vorhaben oder rechtlichen Konsequenzen führen. Das in eine Ärzt:in gesetzte Vertrauen bei einer heiklen Operation führt bei einem schlechten Verlauf gegebenenfalls zu konkreten physischen Einschränkungen und in der Folge womöglich zu Misstrauen gegenüber medizinischen Institutionen.

Vertrauen ist wesentlich mit Unsicherheit verbunden: Ich muss nur dort vertrauen, wo ich nicht weiß, ob das, worauf ich vertraue, auch eintre-

11 Bspw. Cicero, *De Inventione* 2,163; Seneca *Ad Lucilium Epistulae Morales* 97, 13–16; Hobbes, *De homine* II, 2; Emerson (2003).

12 Etwa: Thomas von Aquin, *Summa theologica*, II-II, 128, 1, ad 2.

13 Bspw. Aristoteles, *Nikomachische Ethik*, VIII, 3 und 13; Aristoteles, *Eudemische Ethik*, VII; Kant, *Metaphysik der Sitten*, § 46–47.

14 Geprägt von Erikson (1950).

15 U. a. Fukuyama (1995).

ten wird. Ein bestimmtes Maß von Unwissen ist eine notwendige Bedingung von Vertrauen. Da, wo mir alle relevanten Informationen vorliegen, muss ich nicht vertrauen. Streng genommen kann ich dann gar nicht mehr vertrauen, weil ich bereits weiß. Unsicherheit ist also eine begriffliche Vorbedingung von Vertrauen.

Ein großer Bereich der philosophischen Debatte um Vertrauen beschäftigt sich daher mit der Frage, wann wir vertrauen *sollten* – und wann Vertrauen fehlgeleitet wäre. Onora O'Neill etwa argumentiert, dass Vertrauen für sich genommen ohne moralischen Wert sei, und dass es vielmehr darauf ankomme, wem oder was auf welcher Grundlage vertraut wird (O'Neill 2020). Idealiter solle nur jenen vertraut werden, die auch vertrauenswürdig seien.

Aber wann ist jemand – oder etwas – vertrauenswürdig? In der Einschätzung der Vertrauenswürdigkeit unterliegen wir zahlreichen epistemischen Einschränkungen. Zunächst einmal wissen wir oft zu wenig, um eine fundierte Einschätzung hinsichtlich der Vertrauenswürdigkeit des Gegenübers zu treffen. Dass eine Ärzt:in qua Approbation medizinische Fähigkeiten hat, ist relativ sicher, in einem Spezialfall ist ihre Expertise für Patient:innen dagegen schon schwerer einzuschätzen. Und ob eine neu gewonnene Freund:in vertrauenswürdig ist, stellt sich erst nach einiger Zeit heraus. Unsere Wahrnehmung von Vertrauenswürdigkeit ist darüber hinaus nicht besonders zuverlässig, da sie durch gesellschaftliche Rahmenbedingungen verzerrt wird, die uns manchen im Übermaß und anderen nicht hinreichend vertrauen lassen (Jones 2002, 2013). Es wird eher den Aussagen derer vertraut, die in Aussehen, Sprache und Auftreten jenen ähnlich sind, denen sonst auch Kompetenz in den entsprechenden Fragen zugesprochen wird – ohne dass dies mit deren Vertrauenswürdigkeit korrelieren muss. Dagegen können identitätsbezogene Vorurteile bewirken, dass Menschen aufgrund von für Vertrauen irrelevanten Merkmalen für nicht vertrauenswürdig gehalten werden. Hierfür wurde von Miranda Fricker der Begriff der epistemischen Ungerechtigkeit geprägt (Fricker 2007).

In der philosophischen Debatte um Vertrauen wird weiterhin untersucht, was Vertrauen eigentlich ist: Ist es eine Disposition? Oder ein Gefühl (Lahno 2001)? Hat Vertrauen eher kognitive oder eher affektive Momente (Jones 1996)? Ist es rational oder nicht (Taddeo 2010)? Das Vertrauen in das Gegenüber etwa bei Vertragsschlüssen als rein affektiv zu beschreiben, würde sicherlich in vielen Fällen den Sachverhalt emotional überfrachten. Gleichzeitig würde es dem Phänomen Vertrauen nicht gerecht werden, wenn

wir versuchen würden, alle Vertrauenssituationen nach dem Modell des Vertrages zu beschreiben: „Das Verhältnis zwischen Liebenden und Freunden, zwischen Eltern und Kindern, Kranken und ihren Pflegern, aber auch das zwischen Ehepartnern wäre falsch beschrieben, wenn man es in eine vertragsförmige Form gießen wollte“ (Hartmann 2001, 12).

Damit verbunden ist die Frage, ob wir uns entscheiden können, zu vertrauen; ob Vertrauen also volitional ist (Weckert 2011) oder nicht. Das Vertrauen von Kindern in ihre Eltern als volitional zu beschreiben, würde wohl an der Sache vorbeigehen. Gleichzeitig scheint es gerade in der Arbeit mit Kindern und Jugendlichen viele Formen von ‚pädagogischem Vertrauen‘ zu geben: Wir trauen Kindern und Jugendlichen etwas zu oder vertrauen ihnen etwas an, in der Hoffnung, dass diese, gerade weil wir ihnen vertrauen, wachsen und sich dann (im Nachgang) unseres Vertrauens als würdig erweisen. Sie werden dabei aber erst durch unser Vertrauen auch vertrauenswürdig. Dabei ist es für den pädagogischen Prozess wichtig, dass man tatsächlich vertraut und ein tatsächliches Risiko eingeht. In diesem Fall wäre das Vertrauen die Vorbedingung der Vertrauenswürdigkeit – und nicht andersherum. In so einem Kontext etwa vertraut man nicht unmittelbar, sondern entscheide mich in der Tat dazu, zu vertrauen.

Ein weiterer wichtiger Strang der Debatte zu Vertrauen behandelt die Frage, wer eigentlich als die angemessenen Akteur:innen von Vertrauensbeziehungen zu betrachten sind: Ist Vertrauen allein zwischen Menschen möglich? Lassen sich also komplexere Vertrauensbeziehungen letztlich immer auf interpersonales Vertrauen reduzieren? Wie können wir dann Vertrauen in Institutionen verstehen? Wenn zum Beispiel eine große Mehrheit in den entsprechenden Umfragen dem Bundesverfassungsgericht ihr Vertrauen ausspricht, dann ist damit nicht (allein) gemeint, dass sie den 16 Richter:innen und allen weiteren Mitarbeitenden jeweils persönlich vertrauen. Die wenigsten der Befragten können vermutlich überhaupt auch nur eine dieser Personen namentlich benennen, geschweige denn deren Vertrauenswürdigkeit einschätzen. Aber sie vertrauen dem ‚System Bundesverfassungsgericht‘ mit seinen Gesetzmäßigkeiten, Prozeduren und Strukturen. Beim Vertrauen in Institutionen scheint Vertrauen auf Verfahren und Prozesse, beispielsweise Auswahlverfahren und Überprüfungsprozesse, eine wichtige Rolle zu spielen. Gleichzeitig können diese Verfahren und Prozesse nicht aufrechterhalten werden, wenn es nicht auch Personen gibt, die diese verantwortungsvoll umsetzen. Vertrauen in Institutionen lässt sich also offenbar weder ohne weiteres auf personales Vertrauen reduzieren, noch vollständig von diesem lösen.

Weiterhin wird diskutiert, ob es so etwas wie Vertrauen in Artefakte geben kann – oder ob hier allein, wie von einigen vorgeschlagen, die Zuverlässigkeit zählt (Weckert 2011): Vermutlich vertrauen wir darauf, dass das Flugzeug in der Luft bleibt, wenn wir darinsitzen, aber vertrauen wir dem Flugzeug? In der philosophischen Debatte wird daher der Unterschied zwischen Vertrauen und Verlässlichkeit (*reliance*) insbesondere im Kontext von Technologien intensiv diskutiert, da diese Unterscheidung fundamentale Implikationen für unsere Interaktion mit und Abhängigkeit von technologischen Systemen hat. Vertrauen impliziere, so einige Autor:innen eine tiefergehende Erwartung von Wohlwollen (Jones 1996, Baier 2001), was bei menschlichen Interaktionen durchaus von Bedeutung sei, während Verlässlichkeit eher auf die konsistente und vorhersehbare Erfüllung bestimmter Funktionen durch ein System abzielt. Im technologischen Kontext stellt sich daher die Frage, ob wir Maschinen und Algorithmen tatsächlich vertrauen können, oder ob wir lediglich ihre Verlässlichkeit anerkennen.[16] Es wird daher hinterfragt, ob es gerechtfertigt ist, den Begriff des Vertrauens auf Technologien anzuwenden, oder ob dies eine anthropomorphisierende Fehlinterpretation darstellt, die unsere Erwartungen gegenüber diesen Systemen verzerrt (Nida-Rümelin/Weidenfeld 2018, Fuchs 2020, Nyholm 2020, Frühbauer 2021).

Andere gehen davon aus, dass Vertrauen in Technologien sich nicht (allein) auf diese selbst und auch nicht nur auf die an ihrer Produktion beteiligten Personen bezieht, sondern, da es sich bei Techniken um soziotechnische Systemen handelt, auch bspw. auf Elemente der politisch-rechtlichen Regulierung und Steuerung (Mittelstadt et al. 2016): Wir vertrauen nicht deswegen darauf, dass das Flugzeug in der Luft bleiben wird, weil wir an die Expertise der beteiligten Ingenieur:innen glauben, sondern auch, weil wir Vertrauen in die Zulassungs- und Wartungsregularien haben. Vertrauen in Technologien habe also sowohl personale Aspekte als auch Aspekte, die dem Vertrauen in Institutionen nahekommen.

Vertrauen, Digitalisierung und Künstliche Intelligenz

Vertrauen ist ein vielfältiges Phänomen, welches eine nicht minder vielfältige Zahl von philosophischen Fragen aufwirft. Als Ermöglichungsbedingung zahlreicher menschlicher Interaktionen auf der einen Seite und aufgrund

16 Für einen Überblick über die wichtigsten Stränge in der Diskussion zu Vertrauen in Informationstechnologien s. Ess (2020).

der Verletzlichkeit, die mit ihm einhergeht, auf der anderen Seite, ist Vertrauen ein ambivalentes Phänomen.

Im Bereich der Digitalisierung und der Künstlichen Intelligenz hat man sich sicherlich auch deshalb für Vertrauen interessiert, weil die Annahme besteht, dass es einen Zusammenhang zwischen Vertrauen und Nutzung gibt: kein Vertrauen, keine Nutzung („no trust, no use"). In diesem Themenschwerpunkt soll es aber nicht um die Frage gehen, inwiefern (instrumentell) Vertrauen in diese Technologien hergestellt werden kann, sondern primär darum, welche (neuen) Vertrauensfragen durch Digitalisierung und Künstliche Intelligenz aufgeworfen werden, denen sich die gegenwärtige Philosophie in vielfältiger Weise annimmt.

Denn insbesondere neuere Entwicklungen wie KI-Anwendungen bieten viele Momente der Unsicherheit. Gleichzeitig werden automatisierte digitale Anwendungen in vielen sensiblen Lebensbereichen, wie etwa in Medizin und Gesundheitsvorsorge, Kreditvergabe, der Allokation von Sozialleistungen, eingesetzt. Damit liegen für diese Technologien viele Bedingungen vor, die Vertrauen als Einstellung ihnen gegenüber möglich und erforderlich erscheinen lassen.

Zunächst müssen wir festhalten, dass digitale Technologien in einer spezifischen Weise immer ‚lückenhaft' sind: Sie basieren im Wesentlichen auf Binärcodes, d. h. auf zwei gegensätzlichen Zuständen (1 oder 0), die in bestimmten Wechseln auftreten. Um analoge Signale in digitale Signale zu verwandeln, müssen jene in diskrete, d. h. voneinander getrennte Einzelwerte transformiert werden, die in einem solchen Binärcode darstellbar sind. Im Unterschied zu analogen Signalen sind digitale Signale daher nicht durchgängig, sondern immer ‚unterbrochen'. Hinzu kommt, dass, um die Datenmenge zu reduzieren, vieles herausgefiltert wird, was als ‚nicht relevant' eingestuft wird: Etwa bei digitalisierten Audiosignalen jene Frequenzen, die für das menschliche Ohr nicht hörbar sind – aber bei vielen digitalen Formaten durchaus auch hörbare Aspekte. Das heißt, die Digitalisierung analoger Eingaben ist in einer spezifischen Form verlustbehaftet. Je stärker Daten dabei komprimiert werden, desto mehr Informationen gehen verloren. Die Unsicherheit steigt.

Viele computerisierte algorithmische Anwendungen, insbesondere Anwendungen, die auf dem sogenannten maschinellen Lernen basieren, benötigen dabei sehr große Datensätze (Big Data). Viele dieser Datensätze, die für diese Anwendungen genutzt werden, weisen dabei erhebliche Erhebungslücken und Fehler auf. So können trotz der enormen Datenmenge die

gesammelten Daten eine verzerrte Stichprobe darstellen, da sie häufig aus spezifischen Quellen stammen, die nicht repräsentativ für die gesamte Zielpopulation sind. Beispielsweise können soziale Mediendaten eine bestimmte demografische Gruppe überrepräsentieren. Auch kann es zu klassischen Messfehlern kommen, wenn etwa bei der Erfassung großer Datenmengen ungenaue Sensoren verwendet werden oder menschliche Eingabefehler auftreten. Diese Fehler können systematisch oder zufällig sein, in jedem Fall aber die Datenqualität erheblich beeinträchtigen. Big Data Datensätze stammen auch oft aus verschiedenen Quellen und Formaten, was zu Inkonsistenzen und Problemen bei der Datenintegration führen kann. Unvollständige oder widersprüchliche Daten müssten sorgfältig bereinigt und harmonisiert werden, um valide Analysen zu ermöglichen, was bei verschiedenen Anwendungen in unterschiedlichem Ausmaß erfolgt. Big Data wird weiterhin häufig in Echtzeit oder über längere Zeiträume hinweg gesammelt. Wenn es zu Änderungen in der Datenerfassungsmethodik oder äußerer Einflüsse über die Zeit hinweg kommt, können diese ebenfalls zu Verzerrungen führen, wenn diese Veränderungen nicht korrekt berücksichtigt werden.

Da digitale Technologien auf digitalen Eingaben basieren, können sie freilich außerdem auch nur mit Daten arbeiten, die digital vorliegen. Das heißt, was nicht bereits digitalisiert ist, kann nicht eingelesen werden und kommt in der digitalen Welt nicht vor. Aber manche Informationen, mit denen wir Menschen arbeiten, um uns Welt zu erschließen (bislang oder auch grundsätzlich) lassen sich nicht digitalisieren (Weizenbaum 1976), beispielsweise Gerüche. Digitale Repräsentationen von Welt (bzw. ihren Teilaspekten) sind also auch in dieser Hinsicht in spezifischer, wenn auch nicht immer offenkundiger, Weise unvollständig. Auch wenn diese Unvollständigkeit für jegliche Repräsentationen und auch für die menschliche Wahrnehmung vorliegt, gilt es doch, ihre Besonderheiten und Implikationen mit Hinblick auf digitale Technologien genauer zu verstehen.

Algorithmische Lösungen finden bereits in vielen Lebensbereichen Einsatz, was bedeutet, dass die Lücken- und Fehlerhaftigkeit der Datensätze immense Auswirkungen auf das Leben einzelner Menschen haben kann, wenn zum Beispiel auf dieser Grundlage jemandes Kreditwürdigkeit falsch eingeschätzt oder die Zahlung von Sozialleistungen eingestellt wird. Die Unsicherheit wird bei diesen Anwendungen noch einmal gesteigert – und die für die einzelne Betroffene möglichen Nachteile ebenso: Die epistemische Unsicherheit der Ergebnisse geht bei solchen Anwendungen mit einer deutlich erhöhten Vulnerabilität der Nutzer:innen einher.

Hinzu kommt, dass es sich bei vielen algorithmischen Anwendungen, die gegenwärtig im Umlauf sind, um sogenannte Black Boxes handelt. Als Black Boxes werden in den Sozialwissenschaften Systeme bezeichnet, bei denen man den Input und den Output beobachten kann, allerdings keine Erklärung zur Funktionsweise des Systems vorliegt. Man kann versuchen, aufgrund des Verhältnisses von Input und Output auf die Funktionsweise des Systems zu schließen, wie dieses also die Eingabe in die Ausgabe umwandelt, gesichertes Wissen über die inneren Prozesse hat man allerdings nicht. Einige algorithmische Anwendungen sind Black Boxes, weil sie proprietär sind, d. h. Eigentum bestimmter Personen oder Unternehmen, und beispielsweise als Betriebs- und Geschäftsgeheimnisse geschützt sind. Bei diesen Black Boxes wäre es prinzipiell möglich, zu verstehen, wie sie funktionieren, nur haben wir als Außenstehende keinen Zugang zu den notwendigen Informationen. Anders sieht es etwa mit Anwendungen aus, die auf so genanntem *deep learning* beruhen. Hier ist zwar eine prinzipielle Erklärung der Funktionsweise des Modells, aber keine Erklärung der Einzelergebnisse möglich. Nicht einmal die Eigentümer:innen oder die Programmierer:innen können genau sagen, wie ein spezifisches Ergebnis zustande kommt.

Digitale Technologien bieten also viele Momente der Unsicherheit. Hinzu kommt, dass bei einigen Anwendungen digitaler Technologien hohe persönliche Kosten entstehen können, sollten diese Anwendungen nicht leisten, was sie versprechen. Damit erfüllen sie wesentliche Elemente der oben erläuterten konzeptionellen Vorbedingungen von Vertrauen. Jedoch steht infrage, ob es sich bei diesen Anwendungen eigentlich um die angemessenen Adressat:innen einer Vertrauensbeziehung handelt. Letztlich sind auch diese Anwendungen Artefakte und daher ist es womöglich angemessener, so legt die philosophische Debatte nahe, wenn wir unser Verhältnis zu ihnen als ein ‚sich-Verlassen' (*reliance*) beschreiben. Gleichzeitig lässt sich auch hier der Vertrauensbegriff nicht ganz ausklammern. Wie oben schon erläutert, handelt es sich bei Werkzeugen und Technologien um soziotechnische Systeme. Das heißt, wir mögen vielleicht nicht dem Artefakt selbst vertrauen, aber Vertrauen in soziale Systeme und Institutionen spielt auch bei diesen durchaus eine Rolle. Darüber hinaus hilft uns insbesondere bei Anwendungen, die auf maschinellem Lernen basieren, der gegenwärtig so genannten Künstlichen Intelligenz (KI), der Alternativbegriff der Zuverlässigkeit (*reliability*) nur bedingt weiter: KI kann zwar im Großen und Ganzen oft beeindruckend gute Ergebnisse liefern, jedoch im Einzelfall ebenso vollständig daneben liegen. In einer bestimmten Hinsicht ist KI wesenhaft unzuverläs-

sig. Daher kann der für andere Technologien durchaus etablierte Gegensatz von Vertrauen (trust) und Vertrauenswürdigkeit (*trustworthiness*) auf der einen und Sich-Verlassen (*reliance*) und Zuverlässigkeit (*reliability*) auf der anderen Seite hier nur unter Einschränkungen angewandt werden.

Eine zusätzliche Herausforderung besteht darin, dass neuere digitale Anwendungen über ansprechend gestaltete Interfaces verfügen, oft auch über eine normalsprachliche Ausgabe von Ergebnissen, gegebenenfalls sogar über eine Stimme. Außerdem werden virtual reality. und holographische Anwendungen aktuell weiterentwickelt. Durch all diese technischen Entwicklungen wird die Immersion bei der Nutzung dieser Anwendungen verstärkt und gegebenenfalls ein stärkerer Eindruck quasi-persönlicher Interaktion erzeugt. Diese Anwendungen sind daher ganz anders in der Lage, Vertrauen einzuladen, als traditionelle Technologien: Bei Alexa und Siri stellt sich eher die Vertrauensfrage als bei einem Taschenrechner.

Insbesondere im Feld der KI wird daher der Vertrauensbegriff jüngst ausführlich diskutiert. Seit der Veröffentlichung der „Ethik-Leitlinien für eine vertrauenswürdige KI“ durch die Europäische Kommission (HLEG 2019), als einem Versuch, mit diesen Entwicklungen regulatorisch umzugehen, hat der Vertrauensbegriff im Bereich der KI-Ethik große Aufmerksamkeit erfahren. Dabei ist zu beobachten, dass das Vertrauen im Kontext von Künstlicher Intelligenz entemotionalisiert und einer rationalistischen Handlungslogik unterworfen wird: Es soll nur vertrauenswürdiger KI vertraut werden und es werden Standards entworfen, wann Künstliche Intelligenz als vertrauenswürdig zu betrachten sei (Reinhardt 2023). Ob diese Technologie überhaupt in angemessener Weise als eine Adressatin einer Vertrauensbeziehung betrachtet werden kann, wird allerdings, wie bereits erwähnt, in der philosophischen Debatte durchaus kontrovers diskutiert. Inwiefern Vertrauen und Vertrauenswürdigkeit darüber hinaus geeignete Begriffe für die politische und rechtliche Regulierung sind – zumal unter den politischen Prämissen freiheitlicher Demokratien – ist dabei eine weitere offene Frage.

Aber die Frage von Vertrauen stellt sich mit Hinblick auf ausgefeilte digitale Technologien nicht allein auf der Ebene der Mensch-Maschine-Interaktion, sondern auch mit Hinblick auf Mensch zu Mensch-Interaktionen in digitalen Handlungsräumen: Dass wir uns in digitalen Welten nicht als verkörperte (*embodied*) Entitäten begegnen, wir darüber hinaus nicht einmal mehr mit Sicherheit wissen, ob unsere Interaktionspartner:innen Menschen oder Bots sind, stellt die Frage nach den Bedingungen der Möglichkeit von Vertrauen in diesen Handlungsräumen noch einmal auf eine neue Weise.

Für die philosophische Diskussion um Vertrauen eröffnet sich in Bezug auf digitale Technologien also ein weites Feld für grundsätzliche Überlegungen und spezifische Untersuchungen zu konkreten und sich stetig verändernden digitalen Technologien und Anwendungen. Diesen und den dabei entstehenden neuen Fragen und Aspekten Raum zu geben und Rechnung zu tragen, ist das Anliegen dieses Schwerpunkts.

Zu den Beiträgen in diesem Schwerpunkt

Die ersten drei Beiträge des Schwerpunkts befassen sich mit übergreifenden Fragen zu Vertrauen im Kontext digitaler Technologien, drei weitere Beiträge befassen sich mit Fragen des Vertrauens in konkreten Anwendungsfeldern, der Medizin und Blockchaintechnologien:

Christopher Koska, Julian Prugger, Sophie Jörg und *Michael Reder* setzen sich mit der Unterscheidung von Verlässlichkeit und Vertrauen für digitale Technologien auseinander. In ihrem Beitrag „Die Verlagerung von Vertrauen vom Mensch zur Maschine“: Eine Erweiterung des zwischenmenschlichen Vertrauensparadigmas im Kontext Künstlicher Intelligenz verfolgen sie die These, dass die Nutzung selbstlernender Systeme eine schrittweise Verschiebung von interpersonalem Vertrauen hin zu Vertrauen zwischen Mensch und Maschine bewirkt. So entstünden neue Vertrauensverhältnisse, die auch begrifflich neu gefasst werden müssten, etwa im Hinblick auf neue Vulnerabilitäten bei der Nutzung von Künstlicher Intelligenz. Die Autor:innen diskutieren diese Transformation des Vertrauensbegriffs und geben einen Ausblick, wie dem philosophisch und politisch begegnet werden könnte.

Rico Hauswald untersucht in seinem Aufsatz „Caveat usor: Vertrauen und epistemische Wachsamkeit gegenüber künstlicher Intelligenz“ zwei gegensätzliche Tendenzen in der Diskussion um digitale Technologien und Vertrauen. Eine optimistische Haltung, die KI für prinzipiell vertrauenswürdig hält und auf die Verbesserung ihrer Funktionsweisen abzielt, stehe einer in der philosophischen Literatur stärker verbreiteten Skepsis gegenüber. Hauswald schlägt entgegen beider Haltungen ein Prinzip epistemischer Wachsamkeit für den Umgang mit Systemen Künstlicher Intelligenz vor. Ein solches Prinzip lasse Vertrauen als Einstellung gegenüber KI prinzipiell zu, sorge aber zugleich für die nötige kritische Distanz.

Arne Sonar und *Christian Herzog* stellen in ihrem Beitrag „Vertrauen (in Technik), Vertrauenswürdigkeit (von Technik), Vertrauensadjustierung (gegenüber Technik)“ die Bedeutung der kommunikativen und kooperati-

ven Fähigkeiten von technischen Anwendungen ins Zentrum. Sie diskutieren, welche Auswirkungen unmittelbare und individuelle Reaktionen einer Anwendung in der Interaktion mit Nutzer:innen auf Vertrauensverhältnisse zwischen Mensch und Technik haben können und wie dies in Interaktionsdesigns berücksichtigt werden sollte. Konzeptuell fassen sie die Überlegungen mithilfe der im Titel genannten Triade aus Vertrauen, Vertrauenswürdigkeit und Vertrauensadjustierung, und besprechen als weiteren Faktor die Rolle der Vertrautheit mit technischen Anwendungen.

Christian Budnik diskutiert in seinem Beitrag „Künstliche Intelligenz und Vertrauen im medizinischen Kontext", die Frage, ob KI vertraut werden kann und welche Konsequenzen das in diesem Anwendungsbereich hat. Ausgehend von einer Klärung des Vertrauensbegriffs in der Beziehung zwischen Ärzt:innen und Patient:innen setzt der Beitrag sich mit zwei problematischen Folgen bei der Beantwortung dieser Frage auseinander. Wenn KI vertrauenswürdig ist, könnte dies das interpersonale Vertrauen im medizinischen Bereich überflüssig erscheinen lassen; wenn KI dagegen nicht vertrauenswürdig ist, schädigt ihr Einsatz auch das interpersonale Vertrauensverhältnis. Budnik erwägt daher einen Kategorienfehler in der Diskussion um Vertrauen und Künstlicher Intelligenz und plädiert dafür, dass KI eher mit Ansprüchen der Verlässlichkeit begegnet werden sollte.

Ingrid Becker analysiert in ihrem Beitrag „Blockchain statt Vertrauen? Bedeutung der Blockchain-Technologie für Vertrauen und Sich-verlassen-auf" Bitcointransaktionen. Sie zeigt auf, dass der Begriff der Verlässlichkeit für die Funktionsweise von Bitcoin-Anwendungen wie deren Verschlüsselung passend erscheint, während das für den Vertrauensbegriff weniger klar ist. Becker untersucht, ob im Blockchain-Kontext weiterhin interpersonale Vertrauensverhältnisse bestehen und diese als solche bedacht werden sollten. Sie kommt dabei zu dem Ergebnis, dass das bestehende interpersonale Vertrauensparadigma mit den gesellschaftlichen Veränderungen durch digitale Technologien keineswegs unvereinbar ist.

Eva Pöll und *Katja Stoppenbrink* wenden sich in ihrem Beitrag „‚Trustless trust?' – Zum Begriff des Vertrauens im Rahmen von Blockchainanwendungen" gegen die These, dass die Blockchaintechnologie Vertrauen seitens der Nutzer:innen überflüssig werden lasse. Sie untersuchen vier verschiedene Positionen zu Vertrauen in Bezug auf Blockchainanwendungen und argumentieren dafür, dass Vertrauen in diesem Kontext am besten analog zu Vertrauen in Institutionen verstanden werden kann: Weder sei Vertrauen direkt auf die Technologie noch auf eine bestimmte Gruppe von

Verantwortlichen bezogen, sondern vielmehr auf deren strukturelles Zusammenspiel.

Dieser Themenschwerpunkt beleuchtet dabei freilich nur einen kleinen Bruchteil der vielfältigen Facetten und Anwendungsbereiche von Vertrauen im digitalen Zeitalter dar. Über die hier behandelten Fragen, würden besonders die Analyse von Plattformstrukturen und das Verhältnis von Vertrauens- und Aufmerksamkeitsökonomien zusätzliche Aufmerksamkeit erfordern. Ebenso bedarf die Einordnung generativer KI im Kontext von Vertrauen und Vertrauenswürdigkeit einer intensiveren philosophischen Betrachtung. Dieser Schwerpunkt soll daher als Anstoß für zukünftige Diskussionen dienen, die die komplexen Wechselwirkungen zwischen Digitalisierung, Künstlicher Intelligenz und Vertrauen weiter beleuchten.

Literatur

Aristoteles. *Nikomachische Ethik*. Übers. u. hrsg. v. U. Wolf. Reinbek bei Hamburg 2006: Rowohlt.

Aristoteles: *Eudemische Ethik*, übers. u. hrsg. v. F. Dirlmeier, Berlin 1962: Akademie Verlag.

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

Baier, A. 2001. „Vertrauen und seine Grenzen". In *Vertrauen. Die Grundlage des sozialen Zusammenhalts*, herausgegeben von M. Hartmann, 37–84. Frankfurt/New York: campus.

Barocas, S. und A. D. Selbst. 2016. „Big Data's Disparate Impact". *California law review* 104 (3): 671–732.

Bozdag, E. 2013. „Bias in algorithmic filtering and personalization". *Ethics and information technology* 15 (3): 209–227.

Bostrom, N. und E. Yudkowsky. 2014. „The ethics of artificial intelligence". In *The Cambridge Handbook of Artificial Intelligence*, herausgegeben von K. Frankish und W. Ramsey, 316–334. Cambridge: Cambridge University Press.

Budnik, C. 2018. „Trust, Reliance, and Democracy". *International Journal of Philosophical Studies* 26(2): 221–239.

Calnan, M. und R. Rowe. 2008. *Trust Matters in Health Care*. New York: McGraw-Hill Education.

Cicero: *Über die Auffindung des Stoffes/De Inventione*, herausgegeben von T. Nüßlein. Berlin 1998.

Emerson, R.W. 2003. „Self-Reliance". In Ders., *Nature and Selected Essays*, 175–203. London: Penguin Classics.

Erikson, E. H. 1950. *Childhood and Society*. New York/London: W.W. Norton & Co.

Ess, C. 2020. „Trust and Information and Communication Technologies“. In *The Routledge Handbook of Trust and Philosophy*, herausgegeben von J. Simon, 405–420. New York/London: Routledge.

Fisher, A., und J. Tallant. 2019. „Trust in education“. *Educational Philosophy and Theory* 52 (7): 780–790.

Floridi, L. 2019. „Establishing the rules for building trustworthy AI“. *Nature machine intelligence 1 (6): 261–262.*

Frevert, U. 2013. *Vertrauensfragen. Eine Obsession der Moderne*. München: C. H. Beck.

Fricker, M. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press.

Friedman, B. und H. Nissenbaum. 1996. „Bias in computer systems“. *ACM transactions on information systems* 14 (3): 330–347.

Frühbauer, J. 2021. „Künstliche Intelligenz, Autonomie und Verantwortung Erkundungen im maschinen- und roboterethischen Reflexionskontext“. In *Digitalisierung: Neue Technik – neue Ethik: Interdisziplinäre Auseinandersetzung mit den Folgen der digitalen Transformation,* herausgegeben von B. Held und F. van *Oorschot, 219–234*. Heidelberg.

Fuchs, T. 2020. Verteidigung des Menschen. Grundfragen einer verkörperten Anthropologie. Berlin: Suhrkamp.

Fukuyama, F. 1995. *Trust*. New York u. a.: Free Press.

Hagendorff, T. 2019. „From privacy to anti-discrimination in times of machine learning“. *Ethics and information technology 21 (4): 331–343.*

Hartmann, M. 2001. „Einleitung“. In *Vertrauen. Die Grundlage des sozialen Zusammenhalts*, herausgegeben von M. Hartmann und C. Offe, 7–34. Frankfurt/New York: campus.

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

Hartwig, J. 1991. „The role of trust in knowledge“. *Journal of Philosophy* 88(12): 693–708.

Heesen, J., M. Friedewald, A. Roßnagel, N. Krämer und J. Lamla (2022). *Künstliche Intelligenz, Demokratie und Privatheit. Baden-Baden: Nomos.*

Hegel, G.W.F. [1807]. 1986. *Phänomenologie des Geistes*. Frankfurt am Main: Suhrkamp.

Hobbes, T. [1651]. 2017. *Leviathan*. London: Penguin Classics.

Hobbes, T. [1658]. 1959. *De Homine*. Herausgegeben von Günter Gawlick. Hamburg: Meiner.

Hume, D. [1739f.]. 1960. *A Treatise of Human Nature*, herausgegeben von L. A. Selby-Bigge. London: Clarendon Press.

Jones, K. 1996. „Trust as an Affective Attitude“. *Ethics* 107: 4–25.

Jones, K. 2002. „The Politics of Credibility". In *A Mind Of One's Own: Feminist Essays on Reason and Objectivity*, herausgegeben von L. Antony und C. Witt, 154–176. Boulder, CO: Westview Press.

Jones, K. 2013. „Distrusting the Trustworthy". In *Reading Onora O'Neill*, herausgegeben von D. Archard et al., 186–198. London: Routledge.

Kant, I. [1797/8]. 1900ff. *Die Metaphysik der Sitten*. In Kant's gesammelte Schriften, herausgegeben von der Königlich Preußischen Akademie der Wissenschaften, Bd. VI. Berlin.

Krämer, H. 2009. *Vertrauen in der Wissenschaft: zur kommunikativen Konstruktion von Vertrauen in wissenschaftlichen Publikationen*. Aachen: Shaker.

Lahno, B. 2001. „On the Emotional Character of Trust". *Ethical Theory and Moral Practice* 4: 171–189.

Lenard, P. T. 2012. *Trust, Democracy, and Multicultural Challenges*. University Park, PA: Penn State University Press.

Locke, J. [1663]. 1988. *Two Treatises of Government*. Cambridge: Cambridge University Press.

Luhmann, N. 1968. *Vertrauen. Ein Mechanismus der Reduktion sozialer Komplexität*. Stuttgart: Ferdinand Enke Verlag.

Nickel, P.J. und L. Frank. 2020. „Trust in Medicine". In *The Routledge Handbook of Trust and Philosophy*, herausgegeben von J. Simon, 367–377. New York/London: Routledge.

Mittelstadt, B. D. et al. 2016. „The ethics of algorithms: Mapping the debate". *Big data & society 3 (2): 1–21.*

Mühlfried, F. 2019: *Misstrauen. Vom Wert eines Unwertes*. Stuttgart: Reclam.

Nyholm, S. 2020. *Humans and robots. Ethics, Agency, and Anthropomorphism*. London/New York: Rowman & Littlefield.

O'Neill, O. 2020. „Questioning Trust". In *The Routledge Handbook of Trust and Philosophy*, herausgegeben von J. Simon, 17–27. New York/London: Routledge.

O'Neill, O. 2002. *Autonomy and Trust in Bioethics*. Cambridge: Cambridge University Press.

O'Neil, C. 2016. *Weapons of Math Destruction*. New York: Crown.

Oreskes, N. 2019. *Why trust science?* Princeton: Princeton University Press.

Rawls, J. 1993. *Political Liberalism*. New York: Columbia Univ. Press.

Reichenbach, R. 2020. *Pädagogische Autorität. Macht und Vertrauen in der Erziehung*. Stuttgart: Kohlhammer.

Reinhardt, K. 2023. „Trust and Trustworthiness in AI Ethics". *AI & Ethics* 3: 735–744.

Rolin, K. 2020. *Trust in Science, in:* In *The Routledge Handbook of Trust and Philosophy*, herausgegeben von J. Simon, 354–366. New York/London: Routledge.

Seneca. *Epistulae morales dd Lucilium/ Briefe an Lucilius*, Lat./Dt., übersetzt von H. Gunermann/F. Loretto/R. Rauthe, herausgegeben, kommentiert und Nachwort von M. Giebel. Ditzingen 2018: Reclam.

Simon, J. Hrsg. 2020. *The Routledge Handbook of Trust and Philosophy*. New York/ London: Routledge.

Steinfath, H. und C. Wiesemann. 2016. *Autonomie und Vertrauen: Schlüsselbegriffe der modernen Medizin*. Wiesbaden: Springer.

Thomas von Aquin. *Summa theologica*. In: Editio Leonina, Bd. 4–12, Rom 1888–1906.

Taddeo, M. 2010. „Modelling Trust in Artificial Agents". *Minds and Machines* 20 (2): 243–257.

Veale, M. und R. Binns. 2017. „Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data". *Big Data & Society* 4 (2): 1–17.

Warren, Mark E. Hrsg. 1999. *Democracy and Trust*. Cambridge: Cambridge University Press.

Weckert, J. 2011. „Trusting Software Agents". In *Trust and Virtual Worlds: Contemporary Perspectives*, herausgegeben von C. Ess und M. Thorseth, 89–102. New York: Lang.

Weizenbaum, J. 1976. *Computer Power and Human Reason. From Judgement to Calculation*. San Francisco: W. H. Freeman.

Wolfensberger, M. 2019. *Trust in Medicine: Its Nature, Justification, Significance, and Decline*. Cambridge: Cambridge University Press.

Zuiderveen Borgesius, F. 2018. „Discrimination, artificial intelligence, and algorithmic decision-making". Straßburg: Council of Europe.