Eine sich irrende KI? – Ein Interview mit Gregor Schiele über Künstliche Intelligenz

Gregor Schiele (interviewt von Marcel Scholz)

Scholz: *Sehr geehrter Herr Schiele, zu Beginn würde ich gerne über den Begriff der „Künstlichen Intelligenz" sprechen. Was verstehen Sie als Informatiker unter „Künstliche Intelligenz"?*

Schiele: Das ist tatsächlich eine sehr gute Frage. „Künstliche Intelligenz" ist zunächst einmal ein Begriff, der sich im Laufe der Jahrzehnte in der Informatik weiterentwickelt hat. Grundsätzlich wäre die ursprüngliche Definition, dass man ein Computersystem hat, welches sich in einer Art und Weise intelligent verhält. Also wenn Sie dieses Verhalten des Computersystems einem Menschen zeigen würden, dass dieser Mensch zunächst einmal sagen würde: „Dieses Verhalten erscheint mir intelligent". Das ist eine rein phänomenologische Angelegenheit, die nichts damit zu tun hat, ob das System nun wirklich intelligent ist, sondern vielmehr ob es dem Menschen so erscheint. Als Informatiker denke ich hier an den sogenannten *Turing-Test*[1]. In diesem Test chattet ein Mensch mit einem Computer sowie mit einem anderen Menschen. Dabei wird getestet, ob der Mensch unterscheiden kann, welcher seiner Gesprächspartner der Mensch und welcher der Computer ist. Wenn er nicht zwischen beiden unterscheiden kann, hat der Computer den Turing-Test bestanden.

Aber in der Informatik hat man dann irgendwann angefangen zu sagen: Eine Künstliche Intelligenz charakterisieren wir eher als ein Computersystem, das eine Aufgabe lösen kann, die ein klassisches Computersystem nicht lösen kann. Zwar ist auch diese Definition sehr vage, jedoch sind dies zunächst grundsätzliche Definitionen von Künstliche Intelligenz, welche häufig verwendet werden.

Im Themenfeld der KI-Forschung geht es also um die Entwicklung „intelligenter Maschinen". Nun bezeichnen wir den Menschen als intelligent. Zwar können wir im Rahmen des Interviews nicht klären, was allgemein unter „Intelligenz" zu verstehen ist, doch was wäre unter einer „intelligenten Maschine" zu verstehen? Und was unterscheidet eine „intelligente Maschine" von einer „nicht intelligenten Maschine" und wie funktionieren diese?

1 Siehe hierzu auch: Turing, Alan M. (1950), Computing machinery and intelligence, in: Mind LIX 236, 433–460.

Eine intelligente Maschine wäre aus menschlicher Sicht eine Maschine, die sich intelligent verhält. Das können in Wirklichkeit aber ganz viele Dinge sein und damit ist dies auch das Problem dieser Definition. Es können nämlich ganz einfache, regelbasierte Systeme sein. Wenn Sie diese clever programmieren, dann erscheinen diese Maschinen einem Menschen intelligent. Was man aber normalerweise erreichen möchte ist, dass man eine Maschine hat, die sich in verschiedensten Situationen gut verhält.

Das wäre eine gute Abgrenzung: Eine nicht intelligente Maschine hat einen bestimmten Ablauf. Da wissen Sie als Mensch sofort, diese Maschine macht erst Schritt eins, dann Schritt zwei, Schritt drei und Schritt vier und so weiter. Zudem ist diese Maschine für einen spezifischen Kontext entwickelt, d. h. in einem bestimmten Nutzungskontext arbeitet diese Maschine gut, weil ein Experte sie daraufhin spezifisch programmiert hat. Demgegenüber würde ich von einer intelligenten Maschine erwarten, dass sie diesen intelligenten Rahmen, in welchem sie gut funktioniert, in der einen oder anderen Weise ein bisschen sprengen kann. Diese Maschine ist dann flexibler und kann damit umgehen, dass wenigstens in gewissem Maße Dinge passieren, an die der Entwickler vorher nicht gedacht hat. Und das wäre aus meiner Sicht auch das Prinzip, wie eine intelligente Maschine funktioniert. Eine solche Maschine hat irgendeine Variante, die es ihr erlaubt auch in unvorhergesehenen Situationen gut zu funktionieren oder zumindest gutes Verhalten zu realisieren. Heute ist dies typischerweise Maschinelles Lernen, sodass die Maschine jenes Verhalten lernt. Hierbei gilt es jedoch zu beachten, dass auch aktuelle Systeme mit Maschinellem Lernen allein für spezifische Fälle genutzt werden können. Für diese werden sie trainiert. In den letzten Jahren ist allerdings auch beobachtbar, dass jene Systeme zunehmend flexibler werden.

An dieser Stelle möchte ich aber noch eine Sache anmerken: Über die Jahrzehnte gab es auch Experimente mit sehr simplen Systemen. Vielleicht kennen Sie den Chatbot „Eliza“. Dabei handelt es sich um ein sehr altes System, mit welchem man versucht hat, den Turing-Test zu bestehen. Da ist in Wirklichkeit nicht besonders viel Intelligenz drin, sondern solche Systeme sagen häufig viel mehr über uns Menschen aus als über den Computer, weil der Mensch mit einfachen „Tricks“ dazu gebracht wird, dass er ein intelligentes Verhalten impliziert, obwohl der Computer nur ganz simple Regeln befolgt. In diesem Beispiel fragt Eliza eine Frage, die nicht zum Kontext passt, wenn der Bot nicht weiterweiß.

*Expert*innen unterscheiden oft zwischen „schwacher“ und „starker“ KI. Wann handelt es sich also um eine „schwache“ und wann um eine „starke“ KI?*

Eine schwache Künstliche Intelligenz ist eine KI, die tatsächlich nur intelligent erscheint, während eine starke KI häufig verstanden wird als ein System, das tatsächlich eine echte Intelligenz entwickelt oder darstellt. Also das heißt, dass

eine schwache KI zum Beispiel auch ein System sein kann, das einfach nur für einen ganz spezifischen Anwendungszweck optimiert ist, z. B. bestimmte Bilder klassifizieren kann. Eine starke KI dagegen kann auch in unvorhergesehenen Situationen erkennen, wie sie ihr Verhalten ändern muss, um auch diese Situation zu bewältigen. Es wäre damit das System, welches wir heute eigentlich gerne haben möchten. Also tatsächlich ein wirklich ‚intelligentes' System.

Das, was sich die meisten Menschen vorstellen, wenn sie von Künstlicher Intelligenz hören oder davon sprechen, ist meistens ein Roboter aus dem Kino. Diese haben dann ein Bewusstsein, sind tatsächlich intelligent wie ein Mensch, aber halt künstlich. Auch das wäre dann eine starke KI. Die meisten Forscher hingegen, die von KI sprechen, meinen die schwache KI.

Ihr Vortrag im Rahmen der Ringvorlesung trug den Titel „KI – Irren ist nicht nur menschlich". Es liegt also die Annahme zugrunde, dass sich der Mensch irren kann. Zugleich klingt es, als hätte man die Erwartung, dass dies für die KI nicht gelten würde. Können Sie erklären, was Sie mit einer „sich irrenden KI" genau meinen?

Unter einer „sich irrenden KI" verstehe ich erstmal nur eine KI, die Fehler macht. Das ist alles. Die implizite Annahme von vielen Mensch ist, dass eine KI eben keine Fehler macht. In dem „Irren ist menschlich" liegt auch eine Anspielung auf einen Minderwertigkeitskomplex. Darin ist eine gewisse Frustration erkennbar. Menschen machen halt Fehler und das kann man dann verschieden interpretieren: Entweder als ein „naja ist halt so, muss man damit leben" oder aber als ein „es wäre doch auch schön, wenn wir die Fehler nicht machen würden." Von den „KI-Göttern" erhofft man sich dann, dass diese keine Fehler machen. Was ich damit ausdrücken möchte ist, dass das selbstverständlich nicht stimmt. So funktioniert das nicht, denn auch die Künstliche Intelligenz besteht ja aus einer Vielzahl von Algorithmen, die in bestimmten Situationen extrem gut funktionieren und die uns auch in manchen Anwendungsbereichen einige Vorteile bieten. Auch wenn wir da einen großen Sprung gemacht haben, sind es dennoch Algorithmen, die eine gewisse Fehlerrate haben und die werden sie auch immer haben.

Haben Sie diesbezüglich ein anschauliches Beispiel? Inwiefern irrt die KI?

Ein einfaches Beispiel für eine sich irrende KI ist, z. B. dass Sie in den heutigen Warenketten häufig eine Vorhersage haben, was die Kunden kaufen möchten, damit Sie die nachgefragten Produkte auf Lager haben, jedoch nicht zu viel Lagerplatz verschwenden. Dazu muss man dann halt wissen, wie oft ein Produkt gekauft wird. Das funktioniert aber nicht immer. Ein gutes Beispiel dafür ist die Corona-Pandemie, in welcher die vielen Prognosen des menschlichen Kaufverhaltens von einem Tag auf den anderen falsch waren. Während man also die KI zunächst als gut funktionierend einstufen würde, war sie plötzlich nutzlos. Das

lag daran, dass jene KI mit bestimmten Eingabedaten gefüttert wird und diese eben kein echtes Bewusstsein oder eine echte Intelligenz dafür hat, dass da draußen in der Welt jetzt etwas passiert, was wir so noch nicht erlebt haben. Da ist jetzt ein außerordentliches Ereignis eingetroffen und die KI hat keine Chance, dieses Ereignis auf irgendeine Art und Weise mit einzukalkulieren oder zu verstehen, dass die eigene Prognose aus der Vergangenheit ab sofort ungültig ist. Während also die KI weiterhin von der eigenen Prognose ausging, haben sich die Menschen hingegen schon gedacht, dass sich das Kaufverhalten ändern wird, wenn ein solches Ereignis eintrifft.

Vielleicht noch ein anderes Beispiel: Diesmal handelt es sich um ein Bilderkennungssystem. In China gibt es zum Beispiel ein System, welches erkennt, wenn jemand die Verkehrsregeln missachtet. Hierzu gibt es ein bekanntes Beispiel einer Managerin eines chinesischen Unternehmens, welche eine Verwarnung bekommen hat. Das war im Jahr 2018 und sie bekam diese, weil die Kamera erkannt hat, dass sie bei Rot über die Straße gelaufen ist. Später hat man dann allerdings herausgefunden, dass diese Managerin zu dem Zeitpunkt gar nicht in der Nähe gewesen ist. Bei der Untersuchung des Falls fand man heraus, dass die Kamera ein Bild auf einem vorbeifahrenden Bus erkannt hat, auf welchem sie die Person auf dem Bild als jene Managerin identifiziert hat. Die Bilderkennung und damit die KI hat im Wesentlichen super funktioniert. Die Problematik lag dann eher darin, dass das Wissen über den Kontext gefehlt hat. Die KI hat nicht erkannt, dass es sich nur um ein Bild handelt.

Nun erwarten wir Menschen ja, dass die KI eben nicht irrt. Das wäre in manchen Anwendungsfeldern auch fatal. Denken Sie, dass das Ziel einer irrtumslosen KI theoretisch möglich ist?

Da müsste man jetzt unterscheiden. Für spezifische Anwendungen, also wenn es jetzt zum Beispiel darum geht, dass die KI wirklich Personen auf Computern erkennen soll, dann glaube ich, dass das möglich ist. Ja, das können wir erreichen.

Ob eine starke KI keine Fehler begeht, ist eine philosophische Frage. Ich persönlich halte das für theoretisch nicht möglich. Ich denke, wir werden uns darauf einstellen müssen, dass die KI Fehler macht und machen wird. Von einem pragmatischen Standpunkt aus sind wir auch noch sehr weit weg von Systemen, bei denen sich diese Frage überhaupt stellen würde. Daher tun wir auch sehr gut daran, aktuell unser ganzes Verhalten darauf abzustimmen, dass eine KI Fehler macht.

Aus der Perspektive eines Informatikers: Welche Aspekte wären Ihnen für einen gesellschaftlich verantwortlichen Umgang mit KI besonders wichtig? Und warum?

Besonders wichtig wäre mir zunächst einmal, dass wir offen über das Thema „Künstliche Intelligenz“ sprechen. Das bedeutet zum einen, dass wir die Men-

schen besser darüber aufklären, was eine KI wirklich ist und was dahintersteckt. Es ist nämlich nur ein ausgefeilter Algorithmus, ein Computerprogramm. Wir sollten daher etwas diesen in der Filmwelt produzierten Mythos und die Ehrfurcht vor der KI reduzieren, sodass die Menschen generell besser mit KI zusammenarbeiten können. Ich denke, dass wir in Zukunft sehr viele Fälle haben werden, wo wir die menschliche Intelligenz haben werden, die dann gemeinsam mit künstlich intelligenten Systemen operiert. Als Beispiel dient das autonome Fahren: Wir werden sicherlich nicht von einem auf den anderen Tag ausschließlich autonom fahrende Fahrzeuge haben, sondern zunächst einen Mischbetrieb. Viele menschliche Autofahrer, Fußgänger und Radfahrer, welche dann zusammen mit den autonom fahrenden Autos am Straßenverkehr teilnehmen. Die Menschen müssen die autonomen Fahrzeuge dann besser verstehen und die Grenzen der Künstlichen Intelligenz einschätzen können. Das halte ich für ein großes Problem, welches wir in Zukunft lösen müssen.

Ich bleibe bei dem Beispiel autonomes Fahren: Wenn Sie zum Beispiel mit dem Auto unterwegs sind, dann behaupte ich jetzt mal, dass, wenn Sie im Straßenverkehr unsicher sind, wie sich der andere Teilnehmer verhalten wird, Sie versuchen werden, den anderen Teilnehmer anzuschauen und Blickkontakt herstellen. Sie versuchen also als Mensch, den anderen Menschen anzuschauen, um zu kommunizieren. Sie versuchen abzuschätzen, was er tun wird. Ist dieser unsicher oder weiß er, was er tut? Wird er abbiegen? Aus meiner Sicht ist in Bezug auf KI aktuell völlig unklar, wie sowas funktionieren soll. Sie haben als menschlicher Fahrer keine Chance zu erkennen, ob die KI vielleicht was falsch machen könnte. Sie sehen dann ein autonom fahrendes Auto, das fährt und sie sehen ja keine Unsicherheiten. Das Auto fährt und wenn es einen Fehler macht, dann kommt dieser Fehler für Sie völlig unvorhergesehen. Das ist etwas, woran wir uns erstmal gewöhnen müssten.

Ich denke als Gesellschaft müssen wir uns auch viel mehr mit der Frage beschäftigen, wie wir erkennen, dass die KI einen Fehler gemacht hat. Ein verantwortungsvoller Umgang mit KI bedeutet für mich, dass wir überprüfen, ob die KI Fehler macht. Wenn wir annehmen, dass eine KI auch Fehler macht, dann müssen wir eine Möglichkeit haben, die von ihr verursachten Fehler einzusehen. Es ist wie im Rechtssystem: Falls Sie vor Gericht in Berufung gehen, dann sagen Sie im Wesentlichen ja auch, dass der Richter einen Fehler gemacht hat und sie das jetzt überprüfen möchten. Das Gleiche brauchen wir auch bei der KI. Wir brauchen eine Möglichkeit, zu überprüfen oder überprüfen zu lassen, ob die KI einen Fehler gemacht hat. Die Problematik liegt jetzt darin, dass Sie oft gar nicht verstehen, dass da eine KI involviert war, die einen Fehler gemacht haben könnte und dass Sie gar nicht wissen, dass Sie fälschlicherweise in einer bestimmten Art und Weise behandelt wurden. Insbesondere wenn wir KI in Bereichen einsetzen, die unser Zusammenleben betreffen. So gibt es diese Vorfälle bei Bewerbungen, wo Sie fälschlicherweise durch eine KI aussortiert werden könnten, ohne dass Sie es jemals erfahren werden. Das war jetzt nur ein Beispiel.

Grundsätzlich stellt dies für mich ein Problem dar, wenn da eine KI arbeitet und keiner überprüft, ob die KI möglicherweise einen Fehler gemacht oder wirklich korrekt gearbeitet hat.

Halten Sie es für möglich, dass Sie Ihren Vortrag in Zukunft in „KI – Irren ist nur menschlich" ändern müssten oder wäre ein solches Ziel in ferner Zukunft zu verorten?

Nein, das denke ich nicht. Das wird nicht passieren. Die Beispiele werden dann sicherlich ausgefeilter sein, aber den Titel müsste ich nicht ändern. Ich würde gerne etwas dazu sagen, was sich vielleicht ändern würde: Es würde sich ändern, dass wir – wie vorhin auch erwähnt – weniger von diesen zwei Gegensätzen „Künstliche Intelligenz" und „natürliche Intelligenz" sprechen, sondern uns viel stärker darüber unterhalten, wie jetzt diese zwei in der Interaktion miteinander am besten funktionieren. Oder sich zu fragen, welche Effekte auftreten würden, wenn sie miteinander in Interaktion treten. Wie wir diese Interaktionen gestalten, ist eine Variante aus diesem Dilemma „die KI macht Fehler" herauszukommen. Typischerweise wird noch ein Mensch mit eingeschaltet, sodass die KI einen Vorschlag produziert, dieser aber anschließend von einem Menschen überprüft werden muss. Hier geht es dann auch um rechtliche Fragen.

Ich denke, dass wir noch einige Überraschungen erleben werden, denn wir haben auch noch nicht wirklich verstanden, wie Menschen Entscheidungen treffen. Wir könnten daher irgendwann feststellen, dass diese Sicherungsmechanismen nicht funktionieren. Wir werden Fälle erleben, wo die Absicherung durch den Menschen nicht funktioniert hat und uns dann überlegen, wie wir die Interaktion zwischen Mensch und KI optimieren. Optimieren könnte dann bedeuten, dass die künstliche und natürliche Intelligenz zusammen zu einem System finden, das weniger Fehler macht.

Ihre Antworten beziehen sich unter anderem auf die menschliche Wahrnehmung von Künstlicher Intelligenz. Nehmen wir die KI eventuell verzerrt wahr?

Kennen Sie das Moravec'sche Paradoxon? Es geht im Wesentlichen darum, dass Menschen die Komplexität von Problemen falsch einschätzen beziehungsweise gänzlich falsch einschätzen, was für einen Computer einfach und was schwierig ist. Die Leistungsfähigkeit von KI wird deswegen tendenziell überschätzt. Das liegt auch daran, dass Ereignisse wie „KI schlägt Schachweltmeister" oder „KI schlägt GO-Champion" öffentlich sehr ausführlich diskutiert werden. Das sind auch technisch spannende Dinge. Das Problem ist jedoch, dass der Mensch dazu neigt, einem Computer, der eine intellektuelle Leistung erbringen kann, Intelligenz zu attestieren und davon auszugehen, dass dieser Computer dann auch alles andere kann, was der Mensch kann. Was die Menschen dabei typischerweise ignorieren, ist dass sie unglaublich gute kognitive Fähigkeiten haben. Wir sind großartig und viel besser als jede KI, wenn es um Kognitionsaufgaben geht. Es

sind oft unbewusste Abläufe und weil wir diese nicht wahrnehmen, halten wir sie für einfach. Wir schauen uns zum Beispiel ein Bild einer Katze an und erkennen eine Katze, ohne dass wir diese Erkenntnisleistung als besonders bewerten. Erfahren wir aber von Schachweltmeistern, dass sie mehrere Züge vorausdenken, halten wir das für unheimlich kompliziert, vielleicht für uns persönlich sogar unmöglich. Für den Computer ist es genau umgekehrt. Computer sind viel besser in solchen berechenbaren und vorausdenkenden Dingen wie Schach. Das Erkennen der Katze auf dem Bild fällt einem Computer jedoch recht schwer. Für uns geschieht dies so schnell, dass wir den Prozess gar nicht wahrnehmen. Hierin liegt das Paradoxon. Ob etwas als schwierig bewertet wird, hängt nämlich davon ab, ob wir dabei Fehler machen und daher nehmen wir unsere Bilderkennung gar nicht wahr und denken, dass dies ganz normal sei.

Herr Schiele, ich danke Ihnen für das Gespräch.

Wie können wir autonomen KI-Systemen vertrauen?[1]

Eva Schmidt

1. Einleitung

In der aktuellen Debatte zur künstlichen Intelligenz (KI) in Philosophie, Informatik, Psychologie, Soziologie und verwandten Disziplinen wird häufig betont, dass es nur dann wünschenswert ist, immer komplexere und leistungsfähigere KI-Systeme einzusetzen, wenn wir ihnen vertrauen können (Lahijanian / Kwiatkowska 2016; Vellino / Alaieri 2016; Hoffman u. a. 2013; LaRosa / Danks 2018; Roff / Danks 2018; Holliday u. a. 2016). Solche Systeme sind häufig *opak*, wir können also nicht verstehen, warum sie einen bestimmten Output hervorbringen oder wie sie allgemein funktionieren (Holliday u. a. 2016; Roff / Danks 2018; Vellino / Alaieri 2016; Lahijanian / Kwiatkowsa 2016; Weller 2017). Entsprechend wird gefordert, dass KI *erklärbar* sein sollte – zukünftige KI-Systeme sollten so konstruiert werden, dass wir erklären und verstehen können, was sie tun oder wie sie entscheiden.

Zwei von mir mitverfasste Aufsätze haben die Idee eingeführt, KI-Systeme mittels rationalisierender bzw. Gründe-Erklärungen erklärbar zu machen (Baum u. a. 2017; Baum u. a. 2022), d. h. durch Erklärungen, die sich auf die Gründe des Systems beziehen und die deutlich machen, was aus der Akteur*innen-Perspektive für eine bestimmte Reaktion sprach (Anscombe 1963; Davidson 1963). Ein Alltagsbeispiel: Dass Brokkoli gesund ist, ist ein Grund, der dafür spricht, dass ich jetzt dieses Stück Brokkoli esse. Wenn ich nun den Brokkoli esse, weil er gesund ist, erklärt dieser Grund meine Handlung und macht sie zugleich rational.

In diesem Beitrag nun verbinde ich Gründe-Erklärungen mit dem Ziel der vertrauenswürdigen KI. Die Forderung nach wohlbegründetem Vertrauen in autonome KI-Systeme *impliziert* eine Forderung nach Erklärbarkeit dieser Systeme

1 Danksagung: Ich habe diesen Beitrag auf dem Workshop Knowledge and the Management of Ignorance (Collegium Helveticum, Zürich 2018), an der TU Dortmund (2018), an der Universität des Saarlandes (2018), auf dem Workshop Ethics of Algorithmic Decision Making (Leverhulme Centre for the Future of Intelligence, Cambridge 2018), im Kolloquium für Theoretische Philosophie (Universität Zürich 2018), im Seminarkolloquium der Universität Neuchâtel (2018), auf der EuroCogSci (Bochum 2019) und im Rahmen des EIS-Kolloquiums (2021) präsentiert. Ich danke den Beteiligten für ihre Fragen und Kritiken, die wesentlich zur Verbesserung des Aufsatzes beigetragen haben.

durch Gründe-Erklärungen, so mein Argument. Um diesen Schritt zu motivieren, werde ich die folgenden Fragen näher beleuchten: Warum genau sollten wir autonomen KI-Systeme wohlbegründetes Vertrauen entgegenbringen können (Abschnitt 2)? Unter welchen Bedingungen können betroffene Akteur*innen solchen Systemen aus guten Gründen vertrauen (Abschnitt 3)? Und schließlich: Warum sollte wohlbegründetes Vertrauen Erklärbarkeit erforderlich machen, insbesondere Erklärbarkeit durch Gründe-Erklärungen (Abschnitt 4)? Ich beginne mit der ersten Frage.

2. Warum genau sollten wir autonomen KI-Systemen wohlbegründetes Vertrauen entgegenbringen können?

Ich beginne mit einigen Common-Sense-Beobachtungen. Wenn eine neue Technologie oder andere Innovation im Begriff ist, sich gesellschaftlich durchzusetzen, stellt sich die Frage, ob sie den Menschen nützt oder schadet. Ob Gasbeleuchtung oder die Pille, die positiven Auswirkungen einer Innovation können gegen die Risiken ihrer Anwendung abgewogen werden. Eine Innovation wird leichter angenommen, wenn Menschen entweder wissen, dass ihre Nutzung mit keinen oder minimalen negativen Auswirkungen verbunden ist oder dass zumindest der Nutzen den Schaden ausreichend überwiegt, oder wenn sie darauf vertrauen, dass dies der Fall ist. Jenseits der Frage, wie Menschen am besten dazu animiert werden können, eine neue Technologie anzunehmen, geht es hier auch um etwas *Normatives*: Es wäre unklug, eine neue Technologie zu verwenden, wenn wir nicht in der Lage wären, entweder ihren Nutzen und Schaden direkt abzuwägen (und zu wissen, dass der Nutzen den Schaden ausreichend überwiegt) oder zumindest aus guten Gründen auf sie zu vertrauen. Akteur*innen sollten neue Technologien nur dann einsetzen, wenn sie entweder über *Wissen* oder über *wohlbegründetes Vertrauen* dieser Art verfügen.

Was meine ich mit „Vertrauen aus guten Gründen" bzw. „wohlbegründetem Vertrauen" (McLeod 2021)? Nennen wir die Person, die vertraut, die vertrauensgebende Person und diejenige Person (bzw. das Objekt oder das System), der vertraut wird, die vertrauensnehmende Person (Objekt, System). Vertrauen aus guten Gründen hat dann zwei wesentliche Merkmale. Erstens: Die vertrauensnehmende Person muss tatsächlich vertrauenswürdig sein. Ich kann zum Beispiel meiner Ärztin nicht aus genuin guten Gründen vertrauen, wenn sie – ohne dass

ich es weiß – eine Hochstaplerin ist.[2] Zweitens muss die vertrauensgebende Person epistemischen Zugang zur Vertrauenswürdigkeit der vertrauensnehmenden Person (bzw. des Systems oder Objekts) haben, d. h. sie muss in der Lage sein zu wissen, dass die vertrauensnehmende Person (System, Objekt) vertrauenswürdig ist. Nehmen wir an, meine Ärztin ist sehr vertrauenswürdig. Wenn ich dies aber nicht wissen kann oder wenn ich irreführende Informationen habe, die darauf hindeuten, dass sie nicht vertrauenswürdig ist, kann ich ihr nicht aus guten Gründen vertrauen. Ich kann ihr in dieser Situation *blind* vertrauen – das mag sogar das Beste oder moralisch geboten sein – aber das heißt ja gerade nicht, jemandem aus guten Gründen zu vertrauen.

Daraus ergibt sich:

Wohlbegründetes Vertrauen
Eine vertrauensgebende Person vertraut einer vertrauensnehmenden Person (Objekt, System) genau dann wohlbegründet, wenn (a) die vertrauensnehmende Person (Objekt, System) vertrauenswürdig ist und (b) die vertrauensgebende Person in der Lage ist zu wissen, dass die vertrauensnehmende Person (Objekt, System) vertrauenswürdig ist.[3]

Um diese allgemeinen Überlegungen auf neue KI-gestützte Technologien anzuwenden, ist es erstens relevant, dass solche Technologien in vielen Kontexten eine immer wichtigere Rolle spielen, z. B. in Suchmaschinen, als Chatbots, im Personalwesen, bei der Interpretation medizinischer Bilder, in autonomen Waffensystemen oder in autonomen Fahrzeugen. Ihr Einsatz in diesen Kontexten ist insofern nützlich, als Menschen dadurch von bestimmten Aufgaben entlastet werden und sich auf wichtigere Dinge konzentrieren können, und insofern die Systeme solche Aufgaben besser erfüllen als Menschen. Ihr Einsatz kann jedoch auch schädlich sein: KI-Systeme können Fehler machen, sie können Bias und Diskriminierung perpetuieren oder sie können missbraucht werden, um z. B. persönliche Daten ihrer Nutzer*innen für kommerzielle Zwecke abzugreifen (Garcia 2016; Cadwalladr / Graham-Harrison 2018). Der Schaden, den KI-Systeme anrichten können, ist umso bedrohlicher, als sie zunehmend in Kontexten eingesetzt werden, in denen ihre Outputs massiven Einfluss auf Akteur*innen haben – darauf, ob diese beschäftigt oder arbeitslos, krank oder gesund, tot oder lebendig sind. Ich beschränke mich in der folgenden Diskussion auf KI-Systeme, die in

2 Gleichwohl kann es verständlich und sogar rational sein, ihr zu vertrauen, solange ich nicht weiß, dass sie eine Hochstaplerin ist.

3 Warum drücke ich den relevanten epistemischen Zugang durch „ist in der Lage zu wissen“ aus und nicht durch „weiß“? Wir müssen nicht tatsächlich herausgefunden haben, dass eine Vertrauensperson vertrauenswürdig ist, um ihr aus guten Gründen zu vertrauen; vielmehr scheint es ausreichend zu sein, dass wir dies herausfinden können, wenn sich die Frage stellt.