Was „denkt“ Künstliche Intelligenz? Wie wir sichtbar machen, wie intelligent KI wirklich ist

Katharina Weitz

Kurzzusammenfassung/Abstract

Die Diskussion um den Einsatz von KI in verschiedenen Anwendungsfeldern wird angeregt geführt. Wie funktionieren diese Systeme, die in der Lage sind, unsere Emotionen zu erkennen oder unsere Musikvorlieben herauszufinden? Und wie kann man sich sicher sein, dass diese Systeme auch das Lernen, was sie sollen? Um solche Systeme sicher einsetzen zu können, muss für Menschen nachvollziehbar sein, wie die KI zu ihren Entscheidungen kommt. Licht ins Dunkel bringt hier der Forschungsbereich der Erklärbaren Künstlichen Intelligenz. Er zeigt, wieso KI manchmal klüger scheint als sie ist.

1. Über Künstliche Intelligenz

Seit einigen Jahren ist der Begriff „Künstliche Intelligenz“ in aller Munde. Dabei ist er Begriff nicht neu. Wissenschaftler_innen setzen sich schon seit über 60 Jahren mit den Herausforderungen und Tücken Künstlicher Intelligenz auseinander. Die Geburtsstunde der Künstlichen Intelligenz (kurz: KI) war 1956 in den USA. In einem Sommer-Workshop am Darthmouth College einigten sich Wissenschaftler auf die Bezeichnung „Artificial Intelligence“ (AI), um Forschungsbereiche, die vorher mit „*thinking machines*“ oder „*complex information processing“* bezeichnet wurden, unter einem Begriff zu vereinen (McCorduck, P. 2004). Warum ist dies wichtig zu wissen? Weil die Namensgebung einen Einfluss auf die Wahrnehmung und Erwartungen an dieses Gebiet hat. Der Name gibt es bereits vor: Wir erwarten etwas, das intelligent ist. Intelligent handelt. Wo aber Erwartungen vorliegen, ist die Enttäuschung oft nicht weit entfernt. Davon zeugen zwei KI-Winter, in denen Künstliche Intelligenz fast schon abgeschrieben war. Das, was man als intelligent vermutete, war langsam, unflexibel und gar nicht mal so schlau. Und nun also seit einigen Jahren die große KI-Euphorie. Und nicht zu Unrecht: Es hat sich viel getan, die Fortschritte in der Forschung können sich sehen lassen, ein Durchbruch folgt dem nächsten. Die wissenschaftlichen Veröffentlichungen, die das Schlüsselwort „Artificial Intelligence“ beinhalten, haben sich seit dem Jahr 2000 mehr als verdreifacht. Es gibt inzwischen Systeme, die erfolgreich in unserem Alltag eingesetzt werden, zum Beispiel, um Gesichter zu erkennen, unsere Vorlieben bei Musik und Filmen zu lernen, um große Datenmengen zu analysieren (zum Beispiel Weitz et al. 2019a) oder unser Kaufverhalten aufzuzeichnen, um uns (scheinbar) perfekt zugeschnittene Werbung zu präsentieren. Doch wie kommen diese Systeme auf ihre Klassifikationsentscheidungen? Welche der Unmengen an Informationen waren relevant für das System? Hat das System das gelernt, was es sollte? Wo macht es noch Fehler? Alles Fragen, die so wichtig sind. Und gar nicht so einfach zu beantworten, bestehen die heutigen Systeme vor allem aus sogenannten „Tiefen Neuronalen Netzen“. Eines dieser Netze, das Krizhevysk et al. (2012) beschreiben, erkennt Objekte auf Bildern und verwendet hierfür 650.000 Neuronen und über 60 Millionen Parameter. Eine Komplexität also, die wir ohne Hilfsmittel weder überschauen noch erfassen können. Dies ist aber notwendig, wenn wir sichergehen wollen, dass KI das macht, was wir möchten. Daher müssen wir dafür sorgen, dass diese Systeme nachvollziehbar und transparent werden. Im Folgenden soll beleuchtet werden, wie wir dies bewerkstelligen können und welche Probleme und Chancen sich daraus ergeben.

2. Wie „denkt“ eine KI?

Wie KI funktioniert, möchte ich exemplarisch an Reeti zeigen. Reeti ist einer unserer Roboter an der Universität Augsburg (s. Abb. 1). Reeti kann humorvoll sein (Ritschel & André, 2018), Nutzer bei gesundheitsbezogenen Themen wie gesunder Ernährung unterstützen (Ritschel et al., 2019a) und als Alltagsassistent für Menschen in ihrem Zuhause eingesetzt werden (Ritschel et al., 2019b). Um seine

Aufgaben erfüllen zu können, ist es nützlich, wenn Reeti einschätzen kann, wie sein menschliches Gegenüber empfindet. Helfen können ihm hier soziale Signale. Für das soziale Zusammenleben von Menschen waren soziale Signale schon immer zentral. Ein wichtiger Nutzen und Zweck dieser Signale liegt darin, dass wir schnell den Gemütszustand einer anderen Person einschätzen können, um passend zu reagieren. Ein wichtiges Signal ist die Gesichtsmimik. Diese Informationen soll Reeti nutzen, um den Emotionsausdruck seines Gegenübers einschätzen zu können. Um Reeti diese Aufgabe bewältigen zu lassen, statten wir ihn mit einem künstlichen Tiefen Neuronalen Netz aus, mit dessen Hilfe er Emotionsausdrücke anhand von Gesichtsbildern erkennen soll. Wir haben nun schon gehört, dass Tiefe Neuronale Netze aus vielen Hunderttausend Neuronen und Millionen von veränderbaren Parametern bestehen. Neuronen finden wir auch noch woanders, nämlich im menschlichen Gehirn. Wir Menschen besitzen noch viel mehr davon als eine KI, nämlich circa 100 Milliarden. Die Art und Weise, wie künstlichen Neuronen funktionieren, hat viel Ähnlichkeit mit der Funktionsweise der Neuronen in unserem Gehirn. In unserem Gehirn erhält ein Neuron Informationen in Form von Signalen. Diese werden mithilfe der Dendriten in das Neuron befördert. Sind diese Signale stark genug, wird ein Aktionspotenzial ausgelöst, dass dazu führt, dass das Neuron die Information mithilfe seiner Synapsen, an denen Neurotransmitter ausgeschüttet werden, an die Dendriten des nächsten Neurons weitergibt. In einem künstlichen neuronalen Netz senden die Neuronen Informationen (numerische Werte) an angrenzende Neuronen (s. Abb. 2). Die Gewichte (w für *weight*) gewichten, wie ihr Name schon sagt, die Information (x). Eine Aktivierungsfunktion (f) (s. Abb. 3) entscheidet dann in Abhängigkeit der Gewichte, wie stark die Information an das nächste Neuron weitergegeben wird.

Abbildung 1: Roboter wie Reeti werden an der Universität Augsburg verwendet, um soziale Interaktionen zwischen Mensch und Maschine zu erforschen. Foto: Hannes Ritschel
((Dateiname: Abb. Reeti.jpg;))

Abbildung 2: Aufbau eines Tiefen Neuronalen Netzes aus einzelnen Neuronen, die miteinander verbunden und in Schichten angelegt sind. Foto:Katharina Weitz
((Dateiname: Abb. NeuronalesNetz.png))

Abbildung 3: Schematischer Aufbau eines künstlichen Neurons. Foto:Katharina Weitz
((Dateiname: Abb. KünstlichesNeuron.png))

Um bei den Tausenden Neuronen nicht den Überblick zu verlieren, sind diese im künstlichen neuronalen Netz in Schichten angelegt (s. Abb. 2).

Es wurde bereits gesagt, dass künstliche neuronale Netze Ähnlichkeiten mit den Neuronen in unserem Gehirn aufweisen. Das heißt aber nicht, dass beide identisch sind. Das menschliche Gehirn diente als Inspiration für die Entwicklung künstlicher neuronaler Netze, ist aber nicht gleichzusetzen mit ihnen. Einige Funktionsweisen von künstlichen Neuronen sind im menschlichen Gehirn nicht zu finden. Auch ist unser Gehirn optimiert darauf, viele verschiedene Aufgaben zu lösen. Ein künstliches neuronales Netz ist meist darauf trainiert, eine bestimmte Aufgabe zu bewältigen. Im Falle von Reeti ist es die Aufgabe, den Emotionsausdruck im Gesicht von Menschen zu erkennen. Um diese Aufgabe zu bewältigen, muss das neuronale Netz von Reeti trainiert werden. Hierfür verwendet man eine Vielzahl von Bildern, die die Emotionen zeigen, die Reeti später erkennen soll. Nehmen wir zum Beispiel an, dass Reeti lernen soll, Menschen mit glücklichen, traurigen oder wütenden Gesichtern zu erkennen. Dann wird Reeti im Training ein Bild nach dem anderen gezeigt. Dem neuronalen Netz von Reeti ein Bild zu zeigen, bedeutet, das Bild in seine einzelnen Bestandteile, die sogenannten Pixel, zu zerlegen und der ersten Schicht des neuronalen Netzes jedem Pixel ein Neuron zuzuordnen (s. Abb. 4). Ein Bild, das zum Beispiel aus 224 × 224 Pixeln besteht, hat in der ersten Schicht 50.176 Neuronen. Bei jedem Bild wird das neuronale Netz Reetis dann eine der möglichen Klassifikationen treffen. Anschließend bekommt er die Information, ob seine Klassifikation richtig oder falsch war. Dadurch passt er seine

Gewichte an, die steuern, welches Neuron wann aktiviert wird. Diese Prozedur wird viele Male wiederholt, bis Reetis neuronales Netz die zu lernenden Emotionen bei den gezeigten Bildern gut unterscheiden kann. Dann kommt die Testphase. Hier werden Reeti andere Bilder als die im Training vorgelegten gezeigt. Bilder, die er zuvor noch nie gesehen hat. Damit möchte man sichergehen, dass Reeti nicht nur die Bilder, die er im Training gesehen hat, auswendig gelernt hat, sondern die für den Emotionsausdruck wichtigen Veränderungen im Gesicht.

Abbildung 4: Ablauf bei der Klassifikation von Emotionsausdrücken durch ein künstliches neuronales Netz. Ein Bild mit einem Emotionsausdruck wird in seine Pixel zerlegt, durch das neuronale Netz gesendet, um am Ende eine Klassifikation zu erhalten. Foto: Katharina Weitz ((Dateiname:NeuronalesNetzAnwendung.png))

3. Sollte man der KI vertrauen?

In unserem Emotionsbeispiel hat das neuronale Netz von Reeti bereits in der ersten Schicht über 50.000 Neuronen. Die Verschaltungen zwischen den Neuronen und die unterschiedlichen Aktivierungen dieser übersteigen das, was wir erfassen können. Man kann auch sagen: Das künstliche neuronale Netz ist zu komplex, als dass wir auf einen Blick nachvollziehen können, ob es das gelernt hat, was es lernen soll. Was also tun in dieser Situation? Reeti vertrauen, dass er schon alles richtig macht und auf die richtigen Aspekte achtet, wenn er Emotionsausdrücke klassifiziert? Aber möchte man wirklich Systemen, die man nicht versteht und bei denen nicht sicher ist, ob und welche Fehler sie machen, vertrauen? Ist die Alternative, solch eine KI gar nicht einzusetzen? Die Vorteile, die uns diese Systeme bescheren, würden wir dann einfach wegwerfen. Die dritte Variante scheint vielversprechender zu sein: nachvollziehbare und erklärbare KI zu schaffen.

Was hätte diese nachvollziehbare und erklärbare KI dann für einen Einfluss auf unser Vertrauen? Dafür muss man zunächst klären, was wir unter Vertrauen überhaupt verstehen. Wenn man Vertrauen zwischen Menschen betrachtet, gibt es unterschiedliche Verständnisse, was Vertrauen hier eigentlich meint. Manche verstehen darunter eine dauerhafte Einstellung (Rotter 1967), andere wiederum sehen Vertrauen eher als zeitlich variabel und veränderbar (Driscoll 1978; Kee & Knox 1970). Wie sieht es beim Vertrauen zwischen Mensch und Maschine aus? Für die Mensch-Maschine-Interaktion ist folgende Definition sehr beliebt: Vertrauen wird als die Einstellung gesehen, dass ein Agent in einer von Unsicherheit und Verwundbarkeit geprägten Situation zur Erreichung der Ziele eines Individuums beitragen wird (Lee & See, 2004, S.51). Ein Agent kann zum Beispiel ein Roboter wie unser Reeti sein. Oder ein Sprachassistent wie Siri oder Alexa. Ob Menschen Maschinen vertrauen, hängt sehr stark von verschiedenen Aspekten ab. So entwickelten Hoff und Bashir (2015) einen theoretischen Ansatz, bei dem sie zwischen dispositonalen, situationalen und gelernten Vertrauen unterschieden. Dispositionales Vertrauen meint dabei die langfristigen Tendenzen, die ein Mensch, unabhängig von der jeweiligen Situation, hat. Hier spielen zum Beispiel das Alter der Person, aber auch Faktoren wie Geschlecht, kultureller Hintergrund und Persönlichkeit eine Rolle. Situationales Vertrauen beschreibt externale, also äußere Einflüsse wie zum Beispiel die Art des Systems, dem der Nutzer ausgesetzt ist, aber auch Aspekte wie die kognitive Beanspruchung in der Situation oder die Aufgabe, die in der Situation zu bewältigen ist, spielen eine Rolle. Neben den externalen Faktoren gibt es zusätzlich noch internale, also im Menschen verankerte, Aspekte, wie zum Beispiel die Stimmung oder die Selbstsicherheit des Nutzers. Gelerntes Vertrauen wiederrum bezieht sich auf das Vertrauen, dass jemand bereits aufgrund von Vorerfahrungen entwickelt hat.

Als wäre das alles noch nicht umfangreich genug, muss man dann noch unterscheiden, dass es neben dem Vertrauen auch noch das Nicht-Vertrauen sowie das Misstrauen gibt (Marsh & Dibben 2005). Nicht-Vertrauen meint dabei ein negatives Vertrauens-Level, während Vertrauen ein positives

Vertrauens-Level meint. Misstrauen kann eher als „fehlgeleitetes Vertrauen" verstanden werden, das durch Verrat oder Täuschung entsteht.

Durch viele Studien zeigt sich, dass zahlreiche Faktoren einen Einfluss haben, ob Menschen Agenten oder Robotern vertrauen.

In der Studie von Petrak et al. (2019) zeigte sich zum Beispiel, dass das Navigationsverhalten eines Roboters in virtueller Realität beim ersten Kennenlernen von Roboter und Mensch einen Einfluss auf das Vertrauen des Menschen hat. Der Roboter, der dem Nutzer folgte und mit ihm den Raum erkundete, wurde als vertrauensvoller wahrgenommen als ein Roboter, der selbstständig den Raum erkundete.

Wenn es nun darum geht, KI bei Entscheidungen, die sie trifft und erklärt, zu vertrauen, ist anzunehmen, dass die Art, wie die Erklärungen vermittelt werden, einen Einfluss auf das Vertrauen von Menschen hat. So konnten zum Beispiel Weitz et al. (2019c) in einer Studie zeigen, dass Menschen, die sich die Erklärungen für die Klassifikationen eines Spracherkenners anschauen, dem System mehr vertrauen, wenn ein virtueller Agent diese Erklärungen präsentiert.

4. Wie können wir sichtbar machen, was die KI denkt?

Wir haben gesehen, wie Reeti Emotionsausdrücke mithilfe von Bildinformationen erkennen kann. Nun könnte Reeti aber nicht nur sagen, welche Emotion er gerade erkannt hat, er könnte auch begründen, warum er denkt, einen bestimmten Emotionsausdruck erkannt zu haben. Der Aufgabe, KI nachvollziehbar und erklärbar zu machen, widmet sich der Forschungsbereich der Erklärbaren Künstlichen Intelligenz (auf Englisch: Explainable AI, kurz: XAI).

Wie reagieren Menschen auf Reeti, der erklärt, auf was er achtet, wenn er Emotionsausdrücke erkennt? Nun, das kommt einmal darauf an, wie die Erklärungen aussehen. Ein großer Bereich der Erklärbaren-KI-Forschung beschäftigt sich mit der Visualisierung von relevanten Bereichen, die für die KI bei der Klassifikation wichtig waren. Dabei können die Visualisierungen sehr unterschiedlich aussehen (s. Abb. 5). Das Verfahren LIME (das für Local Interpretable Model-Agnostic Explanations steht) visualisiert Bereiche im Gesicht, die für Reeti wichtig waren bei der Klassifikation einer bestimmten Emotion als grüne Superpixel, während Bereiche, die gegen die Klassifikation sprechen, als rote Superpixel dargestellt werden (Ribeiro et al. 2016). LRP (Layerwise-Relevance Propagation) nennt sich ein anderes Verfahren, dass relevante Pixel farblich hervorhebt. In dem Beispiel in Abb. 5 ist das die Farbe Rot. Je kräftiger die Farbe im Bild, desto relevanter waren diese Pixel für die Entscheidung (Bach et al., 2015). Selvaraju et al. (2017) hingegen stellen ihre Visualisierungen als Heatmaps dar. Hier stellen rote Flächen sehr relevante Bereiche dar. Je mehr der Farbverlauf in Blau- und Lilatöne verläuft, desto weniger relevant waren die Bereiche für die Klassifikation. Die Visualisierungen hängen also vom Algorithmus ab, den man verwendet (Weitz et al., 2019b). Daher kann es sein, dass auch das subjektive Vertrauen von Menschen davon abhängig ist, welche Art der Visualisierung man ihnen zeigt. Ein weiterer wichtiger Punkt ist, dass die in Abb. 5 gezeigten Bilder alleine noch keine Erklärung sind. Sie tragen zur Erklärung bei, müssen aber interpretiert werden. Es wäre also für eine Erklärung sehr hilfreich, wenn Reeti nicht nur zeigen könnte, wohin er geschaut hat, sondern auch noch sprachliche Interpretationen für die Visualisierungen liefert. Das könnte zum Beispiel sein, dass er sagt: „An deinem Mund habe ich erkannt, dass du gerade freudig schaust. Du hast deine Mundwinkel nach oben gezogen, das habe ich als ein Lächeln gedeutet.“ Forscher_innen arbeiten daran, solche sprachlichen Erklärungen zu generieren (Rabold et al. 2019; Stange & Kopp 2020).

Abbildung 5: Darstellung verschiedener Erklärbare-KI-Visualisierungen. Links: LIME, Mitte: LRP, Rechts: Grad-CAM. Foto:Katharina Weitz
((Dateiname: XAIMethoden.png))

5. Wie gehen Ethik und KI zusammen?

Nun ist es ganz unterhaltsam, sich vorzustellen, wie Reeti unser Roboter versucht, unsere Emotionsausdrücke zu erkennen und seine Entscheidungen zu erklären. Neben diesem unterhaltsamen Beispiel kann KI zur Emotionserkennung in Bereichen eine hilfreiche Unterstützung sein, in denen technische Systeme Nutzern als Assistent_in, Gefährt_in oder Lehrer_in zur Seite stehen sollen. Dies ist zum Beispiel der Fall bei Systemen mit KI, die ältere Menschen in ihrem Alltag unterstützen sollen, oder einer KI, die als Bewerbungstrainer eingesetzt werden kann. Wenn wir Reeti oder andere KI-Systeme in solchen Bereichen einsetzen möchten, kommen wir nicht darum herum, uns auch über ethische Aspekte Gedanken zu machen. So zeigt zum Beispiel eine Vielzahl von Studien, dass Menschen auf Computer ähnlich reagieren wie auf andere Menschen. Ohne dass es ihnen bewusst ist, erwarten viele Menschen, dass Computer soziale Normen erfüllen und beispielsweise Bedauern äußern, wenn die Software einmal nicht einwandfrei funktionieren sollte. Jeder hat wohl schon einmal eine Meldung wie „Entschuldigung, es ist ein Fehler passiert. Das Programm wird beendet“ erhalten, wenn er oder sie mit einem Programm auf dem Computer gearbeitet hat. Sprachassistenten wie Alexa oder Siri verstehen manchmal nicht, was wir sagen und entschuldigen sich dann bei uns. Wir Menschen kennen aber noch viele weitere soziale Normen, die in sozialen Interaktionen von Bedeutung sind. Die Frage ist, welche Normen und Werte möchten wir auch in unsere Roboter wie Reeti und andere KI-Systeme geben? Neben dieser Frage gilt es auch zu überlegen, ob wir vielleicht unbeabsichtigt Normen vermitteln, die auf eine fehlende Diversität in unserer Gesellschaft hinweisen. Ein anschauliches Beispiel ist hier die von den Medien als „Rassistische KI“ betitelte Objekterkennung von Google. Daten sind das Einzige, was auch das neuronale Netz von Google für sein Lernen heranzieht. Daher sollten diese Daten die Realität ziemlich gut abbilden. Wenn dies nicht der Fall ist, kommt es zu Fehlschlüssen. Das neuronale Netz von Google sollte Objekte erkennen und klassifizierte dunkelhäutige Menschen als Gorillas[1]. Der Fehler lag in dem Datensatz, der zum Lernen verwendet wurde. Hier waren kaum oder keine dunkelhäutigen Menschen in den Bildern vertreten. Dies führte dazu, dass das Netz eine Klasse, die ähnliche Merkmale zeigte, hier also die Gorillas, die ein dunkles Fell und eine dunkle Hautfarbe haben, für die Klassifikation verwendete. Was man nicht vergessen darf, ist das „Garbage-in- – Garbage-out“-Prinzip. Wenn die Daten, mit denen die künstlichen neuronalen Netze trainiert werden, schlecht sind, indem sie zum Beispiel Vorurteile oder Probleme der Gesellschaft (im Beispiel von Google eine fehlende Diversität) enthalten, wird auch die KI diese Fehler und beschränkte Sichtweise übernehmen. Sind wir nun mit erklärbarer KI für diese Problematiken gewappnet? Jein. Die derzeitigen Verfahren sind ein guter Anfang, aber es gibt noch viel zu tun. Wir brauchen Erklärungen, die möglichst aussagekräftige Informationen liefern. Visualisierungen sind da nicht ausreichend. Und das Ganze muss verständlich sein, nicht nur für KI-Experten, sondern auch für Endnutzer. Es gilt herauszufinden: Wie sieht eine gute Erklärung über KI-Systeme aus? Wann brauchen wir Erklärungen? Wie viele Details darf die Erklärung haben, ohne Endnutzer zu überfordern? Wie garantieren wir, dass die Erklärungen auch den Inhalt vermitteln, den sie transportieren sollen? Das ist eine der Herausforderungen der derzeitigen Forschung und Entwicklung von KI.

Den Wunsch nach einer nachvollziehbaren und menschenzentrierten KI sprach auch die Europäische Kommission aus. Sie veröffentlichte im April 2019 Richtlinien für einen vertrauenswürdigen, menschenzentrierten Einsatz von KI[2]. Diese Richtlinien zeigen einerseits, welche Werte und Normen

[1] https://www.spiegel.de/netzwelt/web/google-fotos-bezeichnet-schwarze-als-gorillas-a-1041693.html
[2] https://ec.europa.eu/futurium/en/ai-alliance-consultation

Europa für die Zukunft mit KI setzen will. Anderseits zeigte die Entwicklung des Dokuments unter Berücksichtigung von Wissenschaftler_innen, Firmen und Politikern in ganz Europa, dass viele Ansichten, Vorstellungen und Erwartungen aufeinanderstießen. Die Debatte ist wertvoll und die ungeklärten Unstimmigkeiten weisen darauf hin, dass noch viel getan werden muss. Diese Diskussionen und Debatten werden uns helfen, die Möglichkeiten, die uns KI gibt, zu nutzen und uns gleichzeitig umfassend mit den kritischen Aspekten dieser Thematik auseinandersetzen.

6. Wohin führt uns die KI-Forschung?

Unser Roboter Reeti, der Emotionen erkennt, war natürlich nur ein Beispiel. Aber wie wird es in Zukunft im Bereich der „sozialen" KI aussehen? In der Forschung wird an Simulationsmodellen gearbeitet, die menschliche emotionale Kompetenz, zum Beispiel Empathie (André 2014) nachzubilden versuchen. Dadurch können nicht nur Emotionen des menschlichen Gegenübers erfasst werden, sondern man ermöglicht der KI, angemessen auf die Emotionen der Person zu reagieren. Eine sehr einfache Methode, die man hier verwendet, ist das Spiegeln der vom Menschen gezeigten Emotion: Schaut eine Person traurig und ein Roboter, der KI verwendet, reagiert mit einem traurigen Blick als Antwort, kann dies als mitfühlende Geste interpretiert werden. Nicht immer ist es aber wünschenswert, Emotionen zu spiegeln. Der Roboter wird eine aggressive Person kaum mit einem ebenfalls aggressiven Emotionsausdruck beruhigen können. Deeskalationsstrategien wären hier ein besserer Ansatz. Zu entscheiden, wann welche Strategie am besten zum Einsatz kommt, ist eine große Herausforderung. Wir Menschen treffen solche Entscheidungen immer unter Bezugnahme der eigenen Erfahrungen und dem Kontext, in dem wir uns gerade befinden. Wir werden uns in einem Bewerbungsgespräch gegenüber einer Person, die unfreundlich zu uns ist, anders verhalten als im vertrauten Kreise unserer Familie. Um angemessen reagieren zu können, benötigt auch KI solche Informationen. Diese zu erfassen und auszuwerten, stellt bis heute noch eine große Herausforderung dar (Schiller et al. 2019). Forschungsprojekte wie EmmA[3] oder VIVA[4] widmen sich dieser Thematik.

((Box Overview start))

Forschungsprojekte

EmmA und **VIVA** sind Forschungsprojekte am Lehrstuhl für Multimodale Mensch-Technik-Interaktion der Universität Augsburg, die vom Bundesministerium für Bildung und Forschung gefördert werden.

Das Ziel des Projektes **EmmA** ist es, ein interaktives, mobiles, Assistenzsystem zu entwickeln, das bei psychischer Belastung individuell berät und darüber hinaus zur Gefährdungsbeurteilung am Arbeitsplatz sowie der betrieblichen Wiedereingliederung nach einer psychischen Erkrankung genutzt werden kann.

Das Projekt **VIVA** hat sich zum Ziel gesetzt, einen vertrauenswürdigen, lebendigen sozialen Roboter, der von Nutzern im privaten Umfeld als attraktive Bereicherung empfunden wird, zu entwickeln. VIVA soll das persönliche psychische Wohlbefinden der Nutzer verbessern und sie bei der Pflege von Sozialkontakten unterstützen.

((Box Overview end))

Ein weiterer Anwendungsbereich ist die Unterstützung von Krankenhaus- und Pflegepersonal beim Monitoring von Demenzpatienten, die nicht mehr in der Lage sind, Schmerzen verbal auszudrücken (Weitz et al. 2019b).

[3] https://www.uni-augsburg.de/de/fakultaet/fai/informatik/prof/hcm/forschung/emma/
[4] https://www.uni-augsburg.de/de/fakultaet/fai/informatik/prof/hcm/forschung/viva/

Es gibt also schon viele Ideen, wo wir Roboter wie Reeti, der mit einer KI ausgestattet ist, einsetzen könnten. Um in Zukunft solche Systeme mit dem Gedanken an eine menschenzentrierte KI zu nutzen, müssen Aspekte von Vertrauen, Nachvollziehbarkeit, Transparenz, Sicherheit und Ethik Berücksichtigung finden und noch weiter erforscht werden.

7. Literatur

André, Elisabeth (2014). Lässt sich Empathie simulieren? Ansätze zur Erkennung und Generierung empathischer Reaktionen anhand von Computermodellen. Nova Acta Leopoldina NF, 120(405), 81–105.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10(7).

Driscoll, J. W. (1978). Trust and Participation in Organizational Decision Making as Predictors of Satisfaction. Academy of Management Journal, 21, 44–56. https://doi.org/10.5465/255661

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. Human factors, 57(3), 407–434.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097–1105).

Lee, J. D., & See, K. A. (2004). Trust in automation: designing for appropriate reliance. Human Factors, 46, 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Marsh, S., & Dibben, M. R. (2005). Trust, untrust, distrust and mistrust–an exploration of the dark (er) side. In International conference on trust management (pp. 17-33). Springer, Berlin, Heidelberg.

McCorduck, P. (2004). Machines who think: A personal inquiry into the history and prospects of artificial intelligence. CRC Press.

Petrak, B., Weitz, K., Aslan, I., & Andre, E. (2019). Let me show you your new home: studying the effect of proxemic-awareness of robots on users' first impressions. In 2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN) (pp. 1-7). IEEE.

Rabold, J., Deininger, H., Siebers, M., & Schmid, U. (2019). Enriching Visual with Verbal Explanations for Relational Concepts--Combining LIME with Aleph. arXiv preprint arXiv:1910.01837

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135–1144).

Ritschel, H., & André, E. (2018). Shaping a social robot's humor with Natural Language Generation and socially-aware reinforcement learning. In Proceedings of the Workshop on NLG for Human–Robot Interaction (pp. 12–16).

Ritschel, H., Janowski, K., Seiderer, A., & André, E. (2019a). Towards a robotic dietitian with adaptive linguistic style. In: Emilio Calvanese Strinati, Dimitris Charitos, Ioannis Chatzigiannakis, Paolo Ciampolini, Francesca Cuomo, Paolo Di Lorenzo, Damianos Gavalas, Sten Hanke, Andreas Komninos and Georgios Mylonas (Ed.). AmI 2019: Poster and Workshop Sessions of AmI-2019; Joint Proceeding of the Poster and Workshop Sessions of AmI-2019, the 2019 European Conference on Ambient Intelligence, Rome, Italy, November 13-15, 2019. CEUR-WS, 16.

Ritschel, H., Seiderer, A., Janowski, K., Wagner, S., & André, E. (2019b). Adaptive linguistic style for an assistive robotic health companion based on explicit human feedback. In: Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments (pp. 247–255).

Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. Journal of Personality, 35, 651–665. https://doi.org/10.1111/j.1467-6494.1967.tb01454.x

Schiller, D., Weitz, K., Janowski, K., & André, E. (2019). Human-inspired socially-aware interfaces. In International Conference on Theory and Practice of Natural Computing (pp. 41-53). Springer, Cham.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618–626).

Stange, S., & Kopp, S. (2020). Effects of a Social Robot's Self-Explanations on How Humans Understand and Evaluate Its Behavior. In: Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction.

Weitz, K., Johren, F., Seifert, L., Li, S., Zhou, J., Posegga, O., & Gloor, P. A. (2019a). The Bezos-Gate: exploring the online content of the Washington Post. In Collaborative Innovation Networks (pp. 75–90). Springer, Cham.

Weitz, K., Hassan, T., Schmid, U., & Garbas, J. U. (2019b). Deep-learned faces of pain and emotions: Elucidating the differences of facial expressions with the help of explainable AI methods. tm-Technisches Messen, 86(7-8), 404–412.

Weitz, K., Schiller, D., Schlagowski, R., Huber, T., & André, E. (2019c). "Do you trust me?" Increasing user-trust by integrating virtual agents in explainable AI interaction design. In: Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents (pp. 7–9).