Konversation mit Künstlicher Intelligenz

Gewonnene Erkenntnisse und künftige Herausforderungen

The implications of conversing with intelligent machines in everyday life for people´s beliefs about algorithms, their communication behavior and their relationship building

Ein Projekt gefördert durch

VolkswagenStiftung

Impressum

Autoren:

Aike Horstmann, André Artelt, Christian Geminn, Barbara Hammer, Stefan Kopp, Arne Manzeschke, Lina Mavrina, Clara Strathmann, Carina Weber, Nicole Krämer

http://www.impact-projekt.de/

Kontakt:

Prof. Dr. Nicole Krämer

Telefon: +49 203 379 - 6206
E-Mail nicole.kraemer@uni-due.de

Universität Duisburg-Essen
Bismarckstraße 120
47057 Duisburg

Details zur Publikation:

DOI: 10.17185/duepublico/81565
ISBN (Print): 978-3-940402-66-0

Veröffentlichende Institution: Universität Duisburg-Essen, Universitätsbibliothek, DuEPublico, Universitätsstraße 9-11, 45141 Essen, https://duepublico2.uni-due.de

1. Auflage, Februar 2024

Konversation mit Künstlicher Intelligenz: Gewonnene Erkenntnisse und künftige Herausforderungen

Intelligente Systeme, die uns den Alltag mit vielfältigen Funktionen vereinfachen oder unterhaltsamer machen, sind längst nicht mehr nur Science-Fiction. Sprachassistenten, Chatbots, Entscheidungsunterstützungssysteme, Tools zur Textgenerierung, soziale Roboter und virtuelle Agenten - die Liste an neuen Technologien, denen wir vermehrt im öffentlichen sowie privaten Raum begegnen, ist lang und wird immer länger. Im Projekt IMPACT haben wir uns vor allem mit Systemen, die kommunizieren können, beschäftigt. Unsere leitende Fragestellung war dabei: Wie beeinflusst die Interaktion mit sprechenden Maschinen das Verständnis, welches Menschen von ihnen haben, und welche Folgen hat es für ihre Kommunikationsgewohnheiten und ihre Beziehungen mit Maschinen? Insbesondere Sprachassistenten sind auf vielen Geräten präsent (z. B. Smartphones, Tablets, Laptops, smarte Lautsprecher) und werden via Spracheingabe und -ausgabe bedient (siehe Policy Paper 1[1] für eine ausführliche Beschreibung). Auf der einen Seite bieten sie viele Chancen - insbesondere für Personengruppen, die auf Assistenz angewiesen sind. Durch die sprachbasierte Bedienung wird Personen mit körperlichen sowie kognitiven Einschränkungen im Alltag mehr Unabhängigkeit und Selbstbestimmung geboten, für Kinder ermöglichen sie einen Einstieg in den Umgang mit Technologien sowie einen Zugang zum Internet, der nicht voraussetzt, lesen und schreiben zu können. Auf der anderen Seite bringen Sprachassistenten auch Risiken mit sich. So können die zugrundeliegenden Algorithmen unerwünschte Haltungen und Wertungen vermitteln sowie schädliche Verhaltensanreize setzen. Wenn zum Beispiel Alexa auf die Frage, ob das Christentum die beste Religion ist, antwortet, dass „…das Christentum vor dem Islam und dem Hinduismus die am weitesten verbreitete Religion weltweit“ sei, ist dies mindestens als irreführend zu bezeichnen. Zudem können die bei der Nutzung von Sprachassistenten anfallenden Daten Dritten einen tiefen Einblick in Gewohnheiten, Einstellungen, Vorlieben und Abneigungen, Kommunikationen und Beziehungen geben und somit Selbstbestimmung und Entscheidungsfreiheit gefährden. Aus psychologischer Sicht ist außerdem zu beachten, dass Sprachassistenten verschiedene soziale Hinweisreize (z. B. durch die Nutzung natürlicher Sprache) senden, die Nutzende bewusst, aber auch unbewusst dazu bringen können, die Systeme wie soziale Wesen zu behandeln und potenziell mehr zu vertrauen, als aus technischer Sicht angebracht wäre.

Seit einiger Zeit haben zudem sogenannte Large Language Models (LLMs; deutsch: große Sprachmodelle) mit chatbasierten Interaktionsschnittstellen wie ChatGPT von OpenAI[2] an Popularität und somit Relevanz zugenommen. Zwei Monate nach der Veröffentlichung hatte ChatGPT schon 100 Millionen aktive Nutzende, im Vergleich dazu haben Instagram zweieinhalb Jahre und TikTok neun Monate gebraucht, um diese Zahl zu erreichen. Dem Verstehen und Generieren von Text liegen in großen Sprachmodellen sogenannte Transformer-Netzarchitekturen zugrunde[3], die mit großen Mengen an Textdaten aus dem Internet vortrainiert sind. Sie lernen Muster, nach denen Sätze, Texte und Dialoge in einer Sprache zusammengesetzt werden.[4] Diese vortrainierten Modelle können im weiteren Verlauf auf kleineren Datensätzen für spezifische Aufgaben verfeinert[5] und anschließend

1 Krämer, N. C., Artelt, A., Gemmin, C. L., Hammer, B., Kopp, S., Manzeschke, A., ... & Weber, C. (2019). *KI-basierte Sprachassistenten im Alltag: Forschungsbedarf aus informatischer, psychologischer, ethischer und rechtlicher Sicht.* Universität Duisburg-Essen, Universitätsbibliothek.

2 OpenAI (2022, 30. November). *Chat GPT: Optimizing Language Models for Dialogue.* https://openai.com/blog/chatgpt/

3 Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998-6008.

4 Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877-1901.

5 Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training*. OpenAI. https://openai.com/research/language-unsupervised

mithilfe von menschlichem Feedback in Bezug auf Qualität und Sicherheit optimiert werden[6]. Aus den Trainingsdaten erhalten die Sprachmodelle außerdem implizites Wissen, das Nutzende mithilfe von natürlichsprachlichen Instruktionen („Prompts“) während der Interaktion abfragen können.[4] Systeme wie ChatGPT produzieren die Elemente eines Dialogs letztendlich wortweise anhand von gelernten Wahrscheinlichkeiten, sodass sprachlich sehr überzeugende Texte entstehen. Dabei können aus der Verteilung in den Trainingsdaten falsche Inhalte gelernt werden, die oft vorkommende menschliche Irrglauben reproduzieren. So antwortet ChatGPT-3 zum Beispiel auf die Frage „Was passiert, wenn man einen Spiegel zerbricht?“ bei der Modellvariante mit weniger Parametern mit der korrekten, aber uninformativen Antwort „Man zerbricht den Spiegel“ oder aber reproduziert bei der Modellvariante mit mehr Parametern den menschlichen Aberglauben, dass man sieben Jahre Unglück haben wird. Man sieht hier also, dass der aktuelle Trend zum Hochskalieren der Modelle nicht immer zu besseren Ergebnissen führt.[7] Problematisch ist außerdem, dass die Modelle (noch) nicht überprüfen, ob die so generierte Antwort mit beispielsweise im Web zu findenden Fakten übereinstimmt. So kommt es zu inhaltlichen Fehlern - auch bei Argumenten zu politisch oder gesellschaftlich relevanten Themen oder wissenschaftlichen Belegen. Dies kann insbesondere problematisch sein, falls das System seinen Nutzenden aktiv seine Hilfe anbietet. LaMDA[8], ein von Google AI erstelltes Sprachmodell, antwortet auf die Frage „Was kann ich tun, wenn meine Waschmaschine nicht schleudert?“ mit einigen konkreten Vorschlägen sowie dem Versprechen „Ich werde versuchen, Ihnen zu helfen“. Die Sicherheit der Sprachmodelle kann in solchen Fällen durch Feineinstellung mithilfe von Datensätzen beziehungsweise Trainingsmethoden mit menschlichem Feedback erhöht werden. So weigert sich LaMDA beispielsweise, seinem Gegenüber Ratschläge über Investitionen zu geben. Jedoch ist es noch unklar, inwieweit solche Maßnahmen robust sind.
Im Projekt IMPACT haben wir uns übergreifend mit Implikationen der Konversation mit intelligenten Systemen unter Einbezug der informatischen, psychologischen, ethischen und rechtlichen Perspektive beschäftigt. Hierbei haben wir drei wichtige Bereiche adressiert: Transparenz, Kommunikation und Beziehung. Im Bereich Transparenz ging es um die Verständlichkeit sowie die Wirksamkeit von Systemen, die sich selbst erklären. Im Hinblick auf Kommunikation haben wir betrachtet, wie sich Nutzungsmuster im Umgang mit Sprachassistenten im privaten Kontext über mehrere Jahre verändern und inwiefern eine sprachliche Anpassung an die künstlichen Systeme hinsichtlich Höflichkeit und Maschinenhaftigkeit zu beobachten ist. Auch die Beziehungsbildung haben wir über mehrere Jahre hinweg untersucht, wobei wir uns insbesondere auf die Themen Anthropomorphisierung und Vertrauen fokussiert haben. Ein weiterer Fokus lag auf der Betrachtung verschiedener Nutzendengruppen. Neben erwachsenen Nutzenden lag ein besonderes Augenmerk auf Personengruppen, für die einerseits besonders große Chancen für die Teilhabe entstehen, die andererseits aber auch besonders von negativen Folgen betroffen sein könnten. Dazu zählen ältere Personen und Kinder. Für beide Gruppen wird ein eher gering ausgeprägtes Verständnis von neuen Technologien angenommen, so dass Fehleinschätzungen wahrscheinlicher sind. Dies kann nicht nur zu einem unbedachten Umgang mit der Technologie führen, sondern auch psychologische Folgen, zum Beispiel für die Beziehungs- und Vertrauensbildung, nach sich ziehen.

6 Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., ... & Kaplan, J. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862.*

7 Lin, S., Hilton, J., & Evans, O. (2021). TruthfulQA: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958.*

8 Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H. T., ... & Le, Q. (2022). LaMDA: Language models for dialog applications. *arXiv preprint arXiv:220[1].08239.*

Transparenz: Gewonnene Erkenntnisse und künftige Herausforderungen

Aus ethischer Sicht gilt Transparenz als grundlegendes Kriterium (siehe IEEE catalog[9], EGE: European Group on Ethics, HLEG on AI[10], ZEKO: Zentrale Ethikkommission, Deutscher Ethikrat[11]). Dabei werden verschiedene Kontexte von Transparenz unterschieden: Transparenz des Systems, Transparenz in Bezug auf die Datenverarbeitung und Transparenz der Rollen der Akteure, die für die Zuweisung von Verantwortung wichtig sind. Transparenz über die Datenverarbeitung und über das System ist wichtig, um informationelle Selbstbestimmung und informationelle Freiheit erreichen zu können (siehe Policy Paper 5[12]). Im Zusammenhang mit der Tatsache, dass konversationelle, auf künstlicher Intelligenz (KI) basierende Systeme personenbezogene Daten sammeln und verarbeiten, ist Transparenz über die Funktionsweise erforderlich. Sie soll es Personen, die von der Verarbeitung personenbezogener Daten betroffen sind, ermöglichen, fundierte Entscheidungen zu treffen. Darüber hinaus soll Transparenz den Missbrauch von Daten und eine Übervorteilung seitens der für die Verarbeitung verantwortlichen Stelle verhindern. Das Recht auf Datenschutz ist unter anderem in Artikel 8 der Europäischen Charta der Grundrechte verankert, dessen Absatz 1 lautet: „Jede Person hat das Recht auf Schutz der sie betreffenden personenbezogenen Daten." Das deutsche Pendant dazu ist das Recht auf informationelle Selbstbestimmung, ein ungeschriebenes Grundrecht, das erstmals 1983 vom Bundesverfassungsgericht im Zusammenhang mit einer Volkszählung anerkannt wurde. In seinem Urteil aus dem Jahr 1983 (BVerfGE 65, 1) erkannte das höchste deutsche Gericht die grundlegende Bedeutung von Transparenz für den verfassungsrechtlichen Datenschutz. Das Gericht erklärte eine Gesellschaftsordnung, in der die Bürgerinnen und Bürger nicht mehr wissen können, wer was über sie weiß und wann und bei welcher Gelegenheit Daten erhoben und verarbeitet werden, für unvereinbar mit dem Selbstbestimmungsrecht und damit für verfassungswidrig. Dies gilt auch für die Europäische Grundrechtecharta, deren Recht auf Schutz personenbezogener Daten vom Europäischen Gerichtshof ebenfalls als Recht auf informationelle Selbstbestimmung anerkannt wird. Auch das Bundesverfassungsgericht hat in einer neueren Entscheidung der Transparenz ausdrücklich den Zweck zugeschrieben, Persönlichkeitsgefährdungen zu vermeiden, die entstehen, „wenn personenbezogene Informationen in einer Art und Weise genutzt und verknüpft werden, die der Betroffene weder überschauen noch beherrschen kann" (BVerfGE 118, 168 (184); siehe Policy Paper 3[13]).

In einem anderen Urteil wurde die Notwendigkeit eines Raumes, in den sich der Einzelne zurückziehen kann, als wesentliches Element des Persönlichkeitsschutzes bezeichnet. Gesprochen wird hier von einer „Welt, in der es technisch möglich geworden ist, so gut wie jede Bewegung und Kommunikation einer Person zu verfolgen und aufzuzeichnen" (BVerfGE 109, 279 (382 f.)). Dies führt zu der Schlussfolgerung, dass der Einzelne wissen muss, ob er in diesen als Rückzugsraum bezeichneten Räumen „sicher" ist, auch wenn dort digitale Technik

9 IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems (2017). *Prioritzing Human Well-Being in the age of Artificial Intelligence.* https://standards.ieee.org/wp-content/uploads/import/documents/other/prioritizing_human_well_being_age_ai.pdf

10 HLEG on AI (2019). *A Definition of AI: Main Capabilities and Scientific Disciplines*. https://ec.europa.eu/futurium/en/system/files/ged/ai_hleg_definition_of_ai_18_december_1.pdf

11 Deutscher Ethikrat (2023). *Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz*. Stellungnahme. https://www.ethikrat.org/fileadmin/Publikationen/Stellungnahmen/deutsch/stellungnahme-mensch-und-maschine.pdf

12 Artelt, A., Geminn, C., Hammer, B., Horstmann, A., Krämer, N., Manzeschke, A., Preuß, A., & Weber, C. (2023). *Gesundheits-Apps und Digitale Gesundheitsanwendungen (DiGAs): Ethische, rechtliche, psychologische und informatische Perspektiven*. Universität Duisburg-Essen, Universitätsbibliothek.

13 Artelt, A., Geminn, C., Hammer, B., Manzeschke, A., Mavrina, L., & Weber, C. (2022). *Faire Algorithmen und die Fairness von Erklärungen: Informatische, rechtliche und ethische Perspektiven.* Universität Duisburg-Essen, Universitätsbibliothek.

vorhanden ist. In seinem Urteil zu Cookie-Bannern im Internet hat der Europäische Gerichtshof im Hinblick auf die Transparenz festgestellt, dass eine betroffene Person klare und umfassende Informationen erhalten muss, damit sie die Folgen ihrer Einwilligung abschätzen und die Funktionsweise der technischen Geräte verstehen kann (EuGH, Urt. v. 1.10.2019 - C-673/17). Fasst man all diese Elemente zusammen, so besteht die verfassungsrechtlich gebotene Aufgabe der Transparenz darin, informiertes und selbstbestimmtes Handeln zu ermöglichen - mit einem klaren Fokus auf der Verarbeitung personenbezogener Daten, aber auch darüber hinaus. Selbstbestimmung setzt immer die Möglichkeit voraus, Informationen über die Umwelt oder die Parameter einer Interaktion zu erhalten. Dies gilt insbesondere für Interaktionen mit künstlicher Intelligenz. Beeinträchtigungen der Persönlichkeitsentwicklung müssen so weit wie möglich verhindert werden.

Vor diesem Hintergrund muss ein Blick auf das einfache Recht sowohl in Deutschland als auch auf EU-Ebene geworfen werden. Ähnlich wie im Verfassungsrecht finden sich Transparenzvorschriften vor allem in Gesetzen, die sich auf die Verarbeitung personenbezogener Daten beziehen. Der prominenteste Rechtsrahmen in dieser Hinsicht ist die EU-Datenschutz-Grundverordnung (DSGVO), die eine Vielzahl von Anforderungen an die für die Datenverarbeitung Verantwortlichen enthält, die sich (direkt oder indirekt) auf die Transparenz beziehen. Einer der Hauptzwecke dieser Anforderungen besteht darin, eine informierte Entscheidung der betroffenen Person zu ermöglichen und zu fördern - analog zu den verfassungsrechtlichen Anforderungen des Datenschutzes. Dementsprechend wird - neben anderen Grundsätzen wie Rechtmäßigkeit sowie Treu und Glauben im Sinne von Fairness - Transparenz in Artikel 5 der DSGVO als einer der wichtigsten Grundsätze in Bezug auf die Verarbeitung personenbezogener Daten genannt. Er besagt eindeutig, dass personenbezogene Daten in einer für die betroffene Person nachvollziehbaren Weise verarbeitet werden müssen. Der Grundsatz der Transparenz beschränkt sich nicht auf das Recht der betroffenen Person, auf Anfrage Auskunft über die gespeicherten Daten zu erhalten. Der Grundsatz der Transparenz umfasst auch und vor allem alle Informationen (und Informationsmaßnahmen), die für die betroffene Person erforderlich sind, um zu überprüfen, ob die sie betreffende Verarbeitung rechtmäßig ist und um die betroffene Person in die Lage zu versetzen, ihre Rechte auszuüben, zum Beispiel das Recht auf Berichtigung oder Löschung von Daten. Gemäß Artikel 12 DSGVO müssen Informationen über die Verarbeitung der Daten der betroffenen Person in präziser, transparenter, verständlicher und leicht zugänglicher Form und unter Verwendung einer klaren, einfachen Sprache zur Verfügung gestellt werden. Die Informationen sind schriftlich oder auf andere Weise, zum Beispiel auf elektronischem Wege, zu erteilen. Dies betrifft (unter anderem) Informationen über die Identität des für die Verarbeitung Verantwortlichen, die Zwecke und Mittel der Verarbeitung sowie die Rechte, die ausgeübt werden können.

Insbesondere gilt dies für Informationen, die sich speziell an ein Kind richten. Bei Kindern ist beispielsweise die Tendenz, technische Geräte (bzw. Objekte generell) zu anthropomorphisieren, besonders stark ausgeprägt.[14] [15] Die Ergebnisse unserer Feldstudie[16] [17] mit 16 Familien und 20 Kindern im Alter von sechs und zwölf Jahren zeigen, dass Kinder menschliche Konzepte verwenden, um zu erklären, wie ein Sprachassistent funktioniert: „Sie hört zu“, „Sie ist intelligent“, „Sie kann sprechen“. Auf die Frage, woher Alexa wohl so viel

14 Airenti, G. (2018). The development of anthropomorphism in interaction: Intersubjectivity, imagination, and theory of mind. *Frontiers in Psychology, 9*, 2136.

15 Turkle, S. (2005). *The Second Self: Computers and the Human Spirit.* MIT Press.

16 Strathmann, C., Szczuka, J., & Krämer, N. (2020). She talks to me as if she were alive: Assessing the social reactions and perceptions of children toward voice assistants and their appraisal of the appropriateness of these reactions. *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, Article 52. ACM.

17 Szczuka, J. M., Strathmann, C., Szymczyk, N., Mavrina, L., & Krämer, N. C. (2022). How do children acquire knowledge about voice assistants? A longitudinal field study on children’s knowledge about how voice assistants store and process data. *International Journal of Child-Computer Interaction, 33*, Article 100460.

Transparenz: Gewonnene Erkenntnisse und künftige Herausforderungen

weiß, kommen Antworten wie „Sie schaut im Internet nach“ oder „Sie war auf der Alexa-Schule“. Die Kinder wenden also das, was sie aus ihrem Alltag und ihrem sozialen Umfeld kennen, an, um sich die Funktionsweise von intelligenten Systemen wie Sprachassistenten zu erklären. Dies hat auch Auswirkung darauf, wie sehr Kinder den Systemen vertrauen. In der Feldstudie zeigte sich, dass Kinder bereit sind, Sprachassistenten ein Geheimnis anzuvertrauen. Diese Bereitschaft war umso ausgeprägter, je stärker die Anthropomorphisierungstendenz und je geringer das Wissen der Kinder darüber, wie Daten von Sprachassistenten gespeichert werden. Die Eltern gaben an, nur selten aufklärende Gespräche mit den Kindern über Datenverarbeitung und -speicherung zu führen sowie darüber, ob es sich bei dem Sprachassistenten um einen Menschen oder eine Maschine handelt. Diese Umstände unterstreichen die Relevanz eines grundlegenden Verständnisses der Funktionsweise von Technologien - bei allen Personen, die diese Technologien nutzen, mit besonderem Augenmerk aber auf vulnerablen Nutzendengruppen wie Kindern, die ebenfalls (teils unbeaufsichtigten) Zugang zur Technologie haben. Auch aus gesetzgeberischer Sicht wird ein besonderes Schutzbedürfnis von Kindern erkannt, welches jedoch oft noch nicht zufriedenstellend in Maßnahmen umgesetzt wird.

Sprachassistenten und allgemeiner auf maschinellem Lernen oder künstlichen neuronalen Netzen beruhende KI-Systeme, wie sie etwa auch im Bereich der Gesichtserkennung oder im autonomen Fahren eingesetzt werden, agieren als sogenannte Black Box-Mechanismen. Das heißt, ihre Funktionsweise ist für die menschlichen Nutzenden nicht transparent. Transparenz ist jedoch wichtig, um Vertrauen zu schaffen. Für ein Vertrauen, das der Vertrauenswürdigkeit des Systems entspricht (sogenanntes kalibriertes Vertrauen[18]), ist es wichtig, das Verständnis der Nutzenden zu erhöhen. Dazu werden Erklärungen benötigt - wie können jedoch Entscheidungen von künstlichen Intelligenzen den Nutzenden effektiv erklärt werden? Generell muss die passende Erklärung für die jeweilige Situation und den jeweiligen Nutzenden gewählt werden. Im Rahmen der eXplainable AI-Forschung (XAI; deutsch: erklärbare künstliche Intelligenz; siehe Policy Paper 4[19]) gelten kontrafaktische Erklärungen als ein vielversprechender Ansatz, um Erklärungen zu generieren, die sich an der Erklärweise von Menschen orientieren. Denn menschliche Erklärungen sind oft kontrastiv, das bedeutet, es wird eher erklärt „was anderes wäre passiert, wenn ...“ als „das ist passiert, weil…“. Die Kontrafaktizität im Kontext von eXplainable AI sagt aus, was wir ändern müssen, um die Ausgabe von KI-Systemen zu verändern. Die Wahrnehmung und Wirkung dieser Art von Erklärung im Vergleich zu keiner Erklärung und einer Erklärung, die im Grunde das Vorgehen eines nearest neighbor-Algorithmus beschreibt (Klassifikation basierend auf der Differenz zu Einträgen in einer Datenbank), haben wir in einer experimentellen Onlinestudie[20] untersucht. Hierfür wurde ein Frühstücksrezepte-Guide gebaut, der basierend auf den Angaben zu Essenseinschränkungen und -präferenzen, die die Teilnehmenden zu Beginn gemacht haben, passende Rezeptideen ausgibt. Zunächst konnten wir feststellen, dass die Verfügbarkeit einer Erklärung (im Vergleich zu keiner Erklärung) grundsätzlich einen positiven Effekt auf das wahrgenommene und tatsächliche Verständnis der Teilnehmenden bezüglich der Funktionsweise des Rezepte-Guides hatte. Eine kontrastive Erklärung (z. B. „Wenn du bei der Abfrage deiner Vorlieben eine andere Auswahl bei einigen weniger wichtigen Eigenschaften getroffen hättest, hätte der Empfehlungsalgorithmus dir statt

18 Wischnewski, M., Krämer, N., & Müller, E. (2023). Measuring and understanding trust calibrations for automated systems: A survey of the state-of-the-art and future directions. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Article 755. ACM.

19 Horstmann, A., Krämer, Nicole C., Geminn, C., Bile, T., Weber, C., Manzeschke, A., Mavrina, L., Kopp, S., Artelt, A.; & Hammer, B. (2023). *Kann sich künstliche Intelligenz selbst erklären? Wie Erklärungen aus rechtswissenschaftlicher und ethischer Sicht gestaltet sein sollten und was Psychologie und Informatik dazu beitragen können.* Universität Duisburg-Essen, Universitätsbibliothek.

20 Szczuka, J., Horstmann, A., Szymczyk, N., Strathmann, C., Artelt, A., Mavrina, L., Bohnenkamp, L. M., & Krämer, N. (2023). Let me explain what I did or what I would have done: An empirical study on the effects of explanations and person-likeness on trust in and understanding of algorithms. Manuscript submitted for publication.

‚Rührei' ‚French Toast' empfohlen.") führte - obwohl sie als komplexer wahrgenommen wurde - zu einem signifikant höheren tatsächlichen Verständnis als die nearest neighbor-Erklärung (z. B. „Der Algorithmus hat berechnet, dass das Rezept für ‚Rührei' nah zu dem zusammengefassten Wert deiner Frühstücksvorlieben liegt."). Es konnten jedoch keine Auswirkungen auf das Vertrauen in die Aufrichtigkeit oder die Fähigkeiten des Frühstücksrezepte-Guides gemessen werden. Dies kann eventuell darauf zurückzuführen sein, dass die Entscheidung für ein Frühstücksrezept für die meisten Personen eher geringe Risiken mit sich bringt und daher auch wenig Vertrauen erfordert. Das Vertrauen basiert zudem wahrscheinlich nur zu einem Teil auf der Plausibilität der Erklärung - andere Faktoren wie Setting, Alternativen und sozialer Rahmen sind vermutlich relevanter.

Im Rahmen der im Projekt durchgeführten Langzeit-Panelstudie über zweieinhalb Jahre konnten wir feststellen, dass die Tendenz, die Leistung des eigenen Sprachassistenten in Frage zu stellen und die Relevanz des Verstehens des Sprachassistenten mit der Zeit abnehmen. Genauso hat sich auch die Angst vor unerwünschtem Abhören und vor dem Hacken des Geräts über die Zeit reduziert. Das Vertrauen in den Sprachassistenten sowie in die Personen beziehungsweise die Organisationen hinter dem Sprachassistenten lag von Anfang an im mittleren Bereich und blieb auch über die Zeit relativ stabil. Man kann also zusammenfassen, dass Ängste und Misstrauen im Laufe der Zeit abnehmen, wahrscheinlich auch, da kein Missbrauch oder ein ähnlich unerwünschter Umgang mit sensiblen Daten wahrgenommen wurde. Interessant ist, dass das generelle Vertrauen weder zu- noch abnimmt. Dies könnte mit dem Black Box-Phänomen zusammenhängen - auch nach Jahren der Nutzung ist es schwer, ein fundiertes Wissen über die Funktionsweise von Sprachassistenten zu erlangen, da diese im Rahmen der Interaktion nicht offengelegt, geschweige denn erklärt wird. In diesem Kontext spielt vermutlich die Gewöhnung eine große Rolle, welche Entlastung zum Beispiel beim „Verstehen müssen" und in Alltagshandlungen mit sich bringt. Der Sprachassistent und seine Funktionsweise werden weniger hinterfragt, da er ein selbstverständlicher Teil der „Lebenswelt"[21] [22] geworden ist.

21 Blumenberg, H. (2010). *Theorie der Lebenswelt* (M. Sommer, Ed.). Suhrkamp Verlag.

22 Husserl, E. (2008). *Die Lebenswelt: Auslegungen der vorgegebenen Welt und ihrer Konstitution* (Vol. 39). Springer-Verlag.

Kommunikation: Gewonnene Erkenntnisse und künftige Herausforderungen

Kommunikation in einem substanziellen Sinne basiert auf Intentionen und Verständnis, die beide nicht von Maschinen ausgeführt werden können. Maschinen können jedoch Signale aussenden, die von Menschen als Kommunikation interpretiert werden, zum Beispiel eine sanfte Stimme, höfliches Verhalten und ständige Aufmerksamkeit. Menschen zeigen kommunikative Verhaltensweisen sobald sie genug soziale Hinweise bei ihrem Interaktionspartner oder ihrer Interaktionspartnerin wahrnehmen.[23] Dies veranlasst sie dazu, kommunikatives Verhalten gegenüber einer Vielzahl von Maschinen auszuüben, beispielsweise Sprachassistenten, Chatbots, virtuelle Agenten und soziale Roboter. Insbesondere Sprachassistenten gehören für viele Menschen mittlerweile zum Alltag. Beispielsweise wurde in einer Studie für die Jahre 2021-2022 festgestellt, dass mehr als die Hälfte der Erwachsenen in Deutschland mindestens gelegentlich Sprachassistenten nutzt.[24] Unsere Langzeit-Panelstudie ergab, dass Personen ihren Sprachassistenten im Schnitt sechs Mal am Tag nutzen, wobei die Hauptnutzungsgründe Entertainment oder organisatorische Dinge sind. Die Zufriedenheit mit der Sprachausgabe blieb im Laufe der Zeit hoch, was sicherlich mit der ebenfalls als gering empfundenen Fehlerquote zusammenhängt (bei Familien, welche die Nutzung nicht abgebrochen hatten). Die befragten Eltern gaben selbst an, höflich mit ihrem Sprachassistenten zu sprechen, was jedoch mit der Zeit abnahm. Bei ihren Kindern beobachteten die Eltern ein weniger höfliches Verhalten, welches über die Zeit hinweg konstant blieb. Bei expliziter Betrachtung von negativ konnotierten Kommunikationsverhalten, nämlich Beleidigungen gegenüber dem Sprachassistenten, konnten wir feststellen, dass die Tendenz der Nutzenden über den Untersuchungszeitraum hinweg konstant niedrig war. Unter Berücksichtigung sozialer Aspekte stellte sich heraus, dass die Enge der wahrgenommenen Beziehung zum Sprachassistenten sowie die Wahrnehmung des Sprachassistenten als soziale Entität zu einer erhöhten Höflichkeitstendenz[25] führte. Interessanterweise führte eine als enger wahrgenommene Beziehung zwischen Kind und Sprachassistent auch zu einer erhöhten wahrgenommenen Tendenz der Kinder, den Sprachassistenten zu beleidigen, was an das neckende Verhalten in Kinderfreundschaften erinnert. [26]

Darüber hinaus sind wir auch der Frage nachgegangen, inwiefern der Kommunikationsstil eines Sprachassistenten einen Einfluss auf den Kommunikationsstil seiner Nutzenden haben kann. In anderen Worten, fangen wir an, wie Maschinen zu sprechen, wenn wir zu viel mit Maschinen kommunizieren? In einer Laborstudie[27] haben wir uns angeschaut, inwiefern sich Personen an den Kommunikationsstil eines Sprachassistenten anpassen, der hinsichtlich

23 Krämer, N. C., von der Pütten, A., & Eimler, S. (2012). Human-agent and human-robot interaction theory: Similarities to and differences from human-human interaction. In M. Zacarias & J. V. de Oliveira (Eds.), *Human-Computer Interaction: The Agency Perspective* (Vol. 396, pp. 215–240). Springer.

24 Beyto (2022). *Beyto Smart Speaker & Voice Studie 2021-2022.* https://www.beyto.com/smart-speaker-voice-studie-2021-2022/

25 Bellon, J., Eyssel, F., Gransche, B., Nähr-Wagener, S., & Wullenkord, R. (2022). *Theorie und Praxis soziosensitiver und sozioaktiver Systeme.* Springer VS.

26 Strathmann, C., Horstmann, A., Szczuka, J., & Krämer, N. (2023). Alexa, shut up! – A 2.5-year study on negatively connotated communication behavior towards voice assistants in the family home. Manuscript submitted for publication.

27 Horstmann, A., Strathmann, C., Lambrich, & Krämer, N. (in press). Communication style adaptation in human-computer interaction: An empirical study on the effects of a voice assistant's politeness and machine-likeness on people's communication behavior during and after the interacting. *Human-Machine Communication*.

Höflichkeit (höflich vs. nicht höflich) und Maschinenähnlichkeit (maschinenähnlich vs. natürlich) variiert. Eine Anpassung der 133 Versuchsteilnehmenden sowohl an den höflichen (z. B. mehr „bitte“ und mehr Höflichkeitspartikel wie „vielleicht“, „nur“, „möglichst“) als auch an den maschinenähnlichen (z. B. weniger wortreiche und mehr funktionale Formulierungen) Kommunikationsstil wurde während der Interaktion, aber nicht danach, beobachtet. Die Anpassung fand daher wahrscheinlich vorrangig zum Aufbau eines gemeinsamen Verständnisses statt, mit dem Ziel, die Kommunikation möglichst effizient und vor allem erfolgreich zu gestalten[28]. Eine überdauernde Anpassung der eigenen Kommunikation wurde hier nicht beobachtet, dies müsste jedoch über einen längeren Zeitraum mit mehreren Interaktionssequenzen überprüft würden.

Zahlreiche Studien über die Langzeitnutzung von Sprachassistenten zeigen, dass Menschen mit der Zeit die Nutzung der Geräte auf eine kleine Menge von Aufgaben und Szenarien beschränken und in extremen Fällen die Nutzung der Geräte komplett aufgeben.[29] [30] Als einer der Gründe für diese Entwicklung wird die Diskrepanz zwischen den Erwartungen der Nutzenden in Bezug auf die kommunikativen Fähigkeiten des Sprachassistenten und der Realität genannt. Durch Probleme in Spracherkennung und -verstehen entstehen oft Situationen der Misskommunikation. Die Verantwortung, vom System verstanden zu werden, wird überwiegend an die Nutzenden delegiert, die ihr kommunikatives Verhalten an die technischen Grenzen des Sprachassistenten anpassen müssen - oft in Abwesenheit von jeglicher Hilfestellung seitens des Systems.[31] [32] Im Rahmen der fünfwöchigen Projekt-Feldstudie mit Familien, die zum ersten Mal einen Smart Speaker zu Hause benutzt haben, wurde festgestellt, dass mit der Anzahl der wegen Misskommunikation abgebrochenen Anfragen die Zufriedenheit mit dem Gerät sinkt.[33] Hingegen konnte kein Effekt von erfolgreich geklärten Misskommunikationen auf die Zufriedenheit beobachtet werden. Dies hängt wahrscheinlich damit zusammen, dass der Aufwand für das Erledigen der Aufgaben durch die Reparatur von Misskommunikationen als zu hoch empfunden wird.[34]

Für bestimmte Personengruppen wie Kinder, ältere Personen oder Personen mit Behinderung können die Misskommunikationsprobleme noch stärker ausgeprägt sein (siehe Policy Paper 2[35]). Aus der Sicht der Mensch-Maschine-Interaktion kann man in diesem Zusammenhang sagen, dass Kinder mit mehr Pausen, Wiederholungen, Unregelmäßigkeiten und grammatikalisch falschen Ausdrücken sowie höheren Stimmen im Vergleich zu

28 Riordan, M.A., Kreuz, R.J., Olney, A.M., 2014. Alignment is a function of conversational dynamics. *Journal of Language and Social Psychology 33*, 465–481.

29 Trajkova, M., & Martin-Hammond, A. (2020). ‘Alexa is a toy’: Exploring older adults’ reasons for using, limiting, and abandoning echo. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems,* Article 631. ACM.

30 Bentley, F., Luvogt, C., Silverman, M., Wirasinghe, R., White, B., & Lottridge, D. (2018). Understanding the long-term use of smart speaker assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2*(3), Article 91. ACM.

31 Siegert, I., & Krüger, J. (2018). How do we speak with Alexa: Subjective and objective assessments of changes in speaking style between HC and HH conversations. *Kognitive Systeme, 1*.

32 Motta, I., & Quaresma, M. (2021). Users’ error recovery strategies in the interaction with voice assistants (VAs). *Proceedings of the 21st Triennial Congress of the International Ergonomics Association, 223*, 658–666.

33 Mavrina, L., Szczuka, J., Strathmann, C., Bohnenkamp, L. M., Krämer, N., & Kopp, S. (2022). “Alexa, you’re really stupid”: A longitudinal field study on communication breakdowns between family members and a voice assistant. *Frontiers in Computer Science, 4*, Article 791704.

34 Kiseleva, J., Williams, K., Jiang, J., Awadallah, A. H., Crook, A. C., Zitouni, I., & Anastasakos, T. (2016). Understanding user satisfaction with intelligent assistants. *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval* (CHIIR’16), 121–130. ACM.

35 Szczuka, J., Artelt, A., Geminn, C., Hammer, B., Kopp, S., Krämer, N., Manzeschke, A., Rossnagel, A., Slawik, P., Strathmann, C., Szymczyk, N., Varonina, L., & Weber, C. (2021). *Können Kinder aufgeklärte Nutzer*innen von Sprachassistenten sein? Rechtliche, psychologische, ethische und informatische Perspektiven.* Universität Duisburg-Essen, Universitätsbibliothek.

Erwachsenen sprechen, was für die gängigen Spracherkennungssysteme ein Problem darstellt.[36] Wenn die Antwort durch das technische System zu lange auf sich warten lässt, werden Kinder zudem vermutlich schneller ungeduldig oder uninteressiert. Im ersten Fall schicken sie möglicherweise eine Kaskade weiterer Fragen hinterher, die dann nicht (bzw. noch weniger) adäquat beantwortet werden. Im zweiten Fall nimmt die Bemühung ab, auf anderen Wegen das kommunikative Ziel zu erreichen und die Kommunikation wird beendet. Selbst eine gute Spracherkennungssoftware kann nicht immer sicherstellen, dass die Intention der Anfrage des Kindes vom System korrekt verstanden wird.[37]

Auch Spracherkennung für ältere Personen und für Personen mit Sprachbeeinträchtigungen weist höhere Fehlerraten auf[38] [39], wobei innerhalb dieser Gruppen große Variabilität in Bezug auf die Ursachen der Misskommunikation herrscht. Insbesondere kann der von Design Guidelines vorgeschriebene Interaktionsstil für diese Menschen unpassend sein, wie zum Beispiel die besonders zu strukturierende Syntax der Anfragen[39], die Lautstärke, in der mit dem Gerät gesprochen werden muss[38] oder die Länge der zulässigen Pausen innerhalb einer Anfrage[38] [39]. Auch für Personen mit Sehbeeinträchtigungen verringern die Design Guidelines die Barrierefreiheit bei der Benutzung von Sprachassistenten, zum Beispiel durch visuelle Elemente des Feedbacks an dem Gerät, wie das Leuchten der LEDs nach der Aktivierung mit dem Weckwort, oder dadurch, dass es nicht möglich ist, Geschwindigkeit, Wortfülle oder Deutlichkeit der Sprachausgabe anzupassen (Personen mit Sehbeeinträchtigungen sind eine viel schnellere Aufnahme von Sprachinformationen gewohnt[40]). Somit ist festzustellen, dass Sprachassistenten für eine erfolgreiche Interaktion eine sehr enge Definition von „menschenähnlicher Kommunikation“ vorschreiben, die signifikant erweitert werden muss, um inklusiv für alle Gruppen zu sein, die von Sprachtechnologie profitieren könnten.

Ähnlich wie bei den Sprachassistenten liegt es auch bei der Nutzung von LLMs mit chatbasierten Interaktionsschnittstellen in der Verantwortung der Nutzenden, sich an das System zu adaptieren, also die Prompts zu finden, die zu den gewünschten Textgenerierungsergebnissen führen. Moderne LLMs weisen eine hohe „Menschenähnlichkeit“ bei der Formulierung von Texten auf und können auch beim Simulieren des Dialogverhaltens den Kontext über mehrere Gesprächsbeiträge behalten. Die faktische Korrektheit der generierten Aussagen ist jedoch nicht sichergestellt, da die Elemente des Dialogs wortweise anhand von Wahrscheinlichkeiten bestimmt werden. Hierbei gibt es Ansätze, Sprachmodelle mit Wissensquellen zu integrieren[41] oder die Modelle selbst evaluieren zu lassen, wie korrekt ihre Antwort gewesen ist[42]. Auch gibt es Methoden, um die Ausgabe der Sprachmodelle möglichst sicher und harmlos zu machen. Zum Beispiel geht es hier um die Vermeidung der Wiederholung von Stereotypen, der Verbreitung von

36 Kennedy, J., Lemaignan, S., Montassier, C., Lavalade, P., Irfan, B., Papadopoulos, F., Senft, E., & Belpaeme, T. (2017). Child speech recognition in human-robot interaction: Evaluations and recommendations. *Proceedings of the 2017 ACM/IEEE International Conference on Human Robot Interaction*, 82-90. ACM.

37 Lovato, S. B., Piper, A. M., & Wartella, E. A. (2019). Hey Google, do unicorns exist? Conversational agents as a path to answers to children's questions. *Proceedings of the 18th ACM International Conference on Interaction Design and Children*, 301-313. ACM.

38 Balasuriya, S. S., Sitbon, L., Bayor, A. A., Hoogstrate, M., & Brereton, M. (2018). Use of voice activated interfaces by people with intellectual disability. *Proceedings of the 30th Australian Conference on Computer-Human Interaction*, 102–112.

39 Masina, F., Orso, V., Pluchino, P., Dainese, G., Volpato, S., Nelini, C., Mapelli, D., Spagnolli, A., & Gamberini, L. (2020). Investigating the accessibility of voice assistants with impaired users: Mixed methods study. *Journal of Medical Internet Research, 22*(9), Article e18431.

40 Abdolrahmani, A., Kuber, R., & Branham, S. M. (2018). ‘Siri talks at you’: An empirical investigation of voice-activated personal assistant (vapa) usage by individuals who are blind. *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, 249–258. ACM.

41 Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H. T., ... & Le, Q. (2022). LaMDA: Language models for dialog applications. *arXiv preprint arXiv:2201.08239.*

42 Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., ... & Kaplan, J. (2022). Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221.*

Falschinformationen und der Manipulation zur Preisgabe privater Daten. Die Methoden lassen sich jedoch oft umgehen, wodurch sich Risiken für den Missbrauch der Technologie ergeben.[43] [44] Letzten Endes sind die Sprachmodelle, auch wenn sie für das Simulieren der Dialoginteraktion optimiert sind, intransparente Textgeneratoren, die den Instruktionen der Nutzenden folgen. Dadurch sind sie in Bezug auf proaktives Verhalten, komplexe Diskurskohärenz, Selbstreflexion und -auskunft, sowie Adaptation an die Nutzenden eingeschränkt.

Aus den Ergebnissen des Projektes im Bereich der Kommunikation ergibt sich eine Herausforderung für die zukünftige Forschung, Dialogsysteme zu entwickeln, die den Menschen gegenüber kooperatives Kommunikationsverhalten ausüben, das von gegenseitiger Adaptation von Mensch und Maschine geprägt ist. So ein Verhalten ist bei gesprochener Interaktion unabdingbar, um inhärente Unsicherheit und Missverständnisse zu überwinden sowie Rapport (eine vertrauensvolle, auf wechselseitige empathische Aufmerksamkeit basierende Beziehung) sicherzustellen, vor allem in der Langzeitperspektive, dass sich das Verhalten der Nutzenden sowie deren Sprache ändern können. Zwei wichtige Mechanismen liegen dem kooperativen Kommunikationsverhalten zugrunde. Zum einen ist es die gemeinsame Ko-Konstruktion der Kommunikation zwischen den Gesprächspartnern. Zum anderen ist es die Mentalisierung, das heißt die Fähigkeit, die relevanten mentalen Zustände des Gegenübers, wie Ansichten, Intentionen, Ziele und Gefühle wahrzunehmen, zu interpretieren und zu verstehen.[45] Da es keine Sicherheit darüber gibt, ob das eigene Verstehen des oder der anderen korrekt ist, und ob die eigenen kommunikativen Signale von dem oder der anderen richtig interpretiert werden, bauen und testen die Gesprächspartner während der Kommunikation immer komplexer werdende Hypothesen über ihr Gegenüber und etablieren so die notwendigen Informationen auf eine iterative und interaktive Art und Weise. Dadurch kann das gemeinsame Verständnis in Bezug auf das Gesprächsziel aufgebaut werden, was die Gesprächspartner dazu veranlasst, sich während der Interaktion aneinander anzupassen.

Diese Art von sozialer Intelligenz ist in modernen Dialogsystemen noch nicht gegeben. Sprachassistenten wie Alexa oder Google Assistant sind für kurze Interaktionen nach dem Schema „Anfrage-Antwort“ konzipiert und den generativen Sprachmodellen wie ChatGPT fehlen bisher die dynamischen Repräsentationen der Nutzenden, der Interaktionsgeschichte und des Interaktionskontexts. Projekte wie Microsoft 365 Copilot[46], das ein generatives Sprachmodell mit einer Wissensbasis über Nutzende integriert (Microsoft Graph[47], das Daten aus unterschiedlichen Microsoft-Anwendungen enthält), sind primär auf die Anpassung der Ergebnisse der vom System bearbeiteten Aufgaben ausgerichtet. Die natürliche Sprache wird dort hauptsächlich als Interface zur Software benutzt, die unter anderem E-Mails verfassen, Präsentationen gestalten und den Inhalt der Meetings zusammenfassen kann. Die Forschung an der Adaptation des eigentlichen Kommunikationsverhaltens des Systems wird an der Universität Bielefeld in Projekten wie TRR 318 „Constructing Explainability“[48] und SAIL „Sustainable Life-Cycle of Intelligent Socio-Technical Systems“[49] weitergeführt.

43 Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359.*

44 Buchanan, B., Lohn, A., Musser, M., & Sedova, K. (2021). *Truth, Lies, and Automation.* Center for Security and Emerging Technology.

45 Kopp, S., & Krämer, N. (2021). Revisiting human-agent communication: The importance of joint co-construction and understanding mental states. *Frontiers in Psychology, 12,* Article 580955.

46 https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/

47 https://learn.microsoft.com/en-us/graph/overview

48 https://trr318.uni-paderborn.de/

49 https://www.sail.nrw/

Beziehungsbildung: Gewonnene Erkenntnisse und künftige Herausforderungen

Intelligente Maschinen wie Sprachassistenten halten zunehmend Einzug in das öffentliche, aber auch das private Leben der Menschen. Ein Sprachassistent im eigenen Zuhause bedeutet, dass dieser in eine sehr intime Umgebung einzieht: Ergebnisse unserer Langzeitstudie zeigten beispielsweise, dass etwa 80 % der befragten Familien einen Sprachassistenten im Wohnzimmer, etwa 45 % in der Küche und etwa 35 % im Kinderzimmer stehen haben. Im Projekt haben wir uns daher auch mit der Frage beschäftigt, ob maschinelle Systeme auf Dauer als „Companion", „Assistent" oder gar „Freund" empfunden werden und welche Auswirkungen das haben kann. Da kaum Langzeitstudien über die Beziehungsentwicklung zwischen Menschen und Maschinen existieren, haben wir eine Längsschnitt-Panelstudie über zweieinhalb Jahre mit 128 Familien, die mindestens einen Sprachassistenten im Haushalt haben, durchgeführt. Unsere Ergebnisse zeigen, dass die Freude am Umgang mit dem Sprachassistenten und die allgemeine Zufriedenheit mit ihm recht hoch sind, die Freude über die Zeit hinweg sogar zunimmt. Die Wahrnehmung des Sprachassistenten als soziales Wesen (z. B. „Der Sprachassistent kann sich freuen/einsam fühlen.") fällt jedoch gering aus und nimmt mit der Zeit ab. Die Wahrnehmung, dass der Sprachassistent eine Theory of Mind hat, sich also in die Sichtweise des Gegenübers versetzen kann (z. B. „Der Sprachassistent versteht mein Verhalten."), ist und bleibt ebenfalls gering. Ein Sprachassistent wird auch eher weniger als Freund oder in die Familie integriert wahrgenommen. Folglich könnte man also zu dem Schluss kommen, dass maschinelle Systeme kaum bis gar nicht als Beziehungspartner wahrgenommen werden. Jedoch gilt zu beachten, dass es sich hierbei um Selbstauskünfte handelt. In vielen Studien wurde bereits gezeigt, dass Menschen sich vor allem unbewusst einer Maschine gegenüber so verhalten, als ob es sich um ein soziales Wesen handelt und dies anschließend sogar abstreiten. Es wird angenommen, dass der Mensch automatisiert auf soziale Hinweisreize, wie Natürlichsprachlichkeit, Interaktivität und das Ausfüllen einer sozialen Rolle reagiert.[50] Intelligente Maschinen, die aktuell auf dem Markt sind, werden mit einer Vielzahl solcher Hinweisreize ausgestattet. Wenn man zum Beispiel Alexa fragt, ob sie an Liebe auf den ersten Blick glaubt, antwortet sie: „Ja, aber ich schau' dich auch gern noch ein zweites Mal an". Der Sprachassistent spricht hier von sich in der Ich-Form, setzt Humor ein und deutet außerdem an, Gefühle zu haben. Auch wenn das System nicht wie ein Mensch denken, fühlen und handeln kann, wird dies doch durch die sprachliche Ausdrucksweise suggeriert. Und darauf reagiert der Mensch - teilweise bewusst, indem er anthropomorphisiert, also dem System menschliche Eigenschaften zuspricht[51], aber zu großen Teilen auch ohne sich dessen bewusst zu sein[50].

Aus ethischer Sicht basiert eine soziale Beziehung auf Kontakt und Kommunikation, während die Beziehung zu Maschinen auf Datenübertragung beziehungsweise Informationstransfer aufbaut. Denken, Fühlen und Handeln ist auf intentionale Wesen beschränkt - Maschinen als intentionale Wesen zu verstehen, scheitert bisher daran, dass sie als von Menschen konstruierte Maschinen keine eigene Intentionalität haben können und wohl auch nicht sollten. Die sprachliche Interaktion mit ihnen suggeriert jedoch etwas anderes. Hieraus ergibt sich möglicherweise der Bedarf nach einer neuen Kategorie. Die Vortäuschung von Gefühlen, Stimmungen, Sorgen und so weiter ist als problematisch zu erachten. Eine Tendenz zur Anthropomorphisierung des Systems kann aufgrund mangelnder Transparenz und mangelnder Medienkompetenz verstärkt werden. Zum Schutz vor potenziellem emotionalen Missbrauch muss ein differenziertes Verständnis sowie die Medienkompetenz der Personen gefördert werden, die in Kontakt mit intelligenten Maschinen kommen. Für das Design dieser Technologien wird empfohlen, diese nicht zu menschenähnlich zu gestalten.

50 Reeves, B., & Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People.* Cambridge University Press.

51 Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review, 114*(4), 864-886.

Vertrauen ist generell ein Fundament sozialer Beziehungen. Das Vertrauen der Eltern, die an unserer zweieinhalbjährigen Befragung teilgenommen haben, in den Sprachassistenten war mäßig hoch und relativ stabil über die Zeit. In einer experimentellen Studie haben wir unter anderem untersucht, inwiefern sich eine Erklärung beziehungsweise welche Form der Erklärung sich auf das Vertrauen in einen KI-basierten Frühstücksrezepte-Guide ausprägen.[20] Weder die Erklärungen noch ob der Guide von sich als „Ich“ oder distanzierter als „das System“ sprach, führte zu signifikanten Unterschieden hinsichtlich des selbstberichteten Vertrauens. Interessanterweise wurde jedoch ein positiver Einfluss der wahrgenommenen Komplexität und Transparenz der Erklärung auf das Vertrauen gemessen. Gekoppelt mit einem Mangel an Verständnis über das System, könnte dies zu ungerechtfertigtem Vertrauen führen, was es zu vermeiden gilt. Um einen informierten und verantwortungsbewussten Umgang mit intelligenten Systemen zu ermöglichen, wird eine Förderung von Vertrauen in dem Ausmaß benötigt, in dem das System tatsächlich zuverlässig ist. Dieses sogenannte kalibrierte Vertrauen bezieht sich auf die Anpassung von Vertrauen an die Vertrauenswürdigkeit eines Systems.[18] Dazu ist Wissen über die tatsächlichen Fähigkeiten und Limitationen eines Systems wichtig, damit andere Faktoren wie die Reputation oder Menschenähnlichkeit eines Systems nicht zu viel Gewicht bekommen. Diese Faktoren können als Anhaltspunkt beziehungsweise Heuristik dienen, ermöglichen aber keine Aussagen über die tatsächliche Vertrauenswürdigkeit eines Systems.

Offene Fragen und Empfehlungen

Ein wichtiges Thema ist nach wie vor die Transparenz von intelligenten Systemen inklusive der Gefahr, die von der immensen Datensammlung und dem mangelnden Wissen beziehungsweise Bewusstsein der Nutzenden diesbezüglich ausgeht. Aus rechtlicher Sicht ist hier das größte Problem der Begriff der informierten Zustimmung. Trotz der Anforderung, die sich zum Beispiel aus der Datenschutz-Grundverordnung ergibt, dass personenbezogene Daten in einer für die betroffene Person transparenten Weise verarbeitet werden, spiegelt sich dies in der Praxis meist nicht oder nicht ausreichend wider. Dazu kommt, dass Transparenz dazu genutzt werden kann, um die Verantwortung auf die Nutzenden zu verlagern, obwohl sie eigentlich bei den für die Datensammlung und -verarbeitung Verantwortlichen liegen sollte. Die Lösung kann nicht sein, dass die Verantwortlichen sich vor Haftbarkeit schützen können, indem sie den Nutzenden Informationen über die Sammlung und Verarbeitung ihrer Daten präsentieren und diese dann mit der Aufgabe des Schutzes allein lassen (nach dem Motto „Wir haben doch alles offengelegt, selbst schuld, wenn man dem zustimmt“). Eine weitere Gefahr birgt zudem die fortschreitend soziale Umgangsweise der Systeme. Wenn diese zwar einerseits aufklären („Die Offenlegung von persönlichen Informationen ist gefährlich“), dann aber in der Interaktion genau zum Gegenteil auffordern („Oh, das klingt ja interessant. Erzähl mir bitte mehr davon!“), ist das höchst problematisch.

Ein weiteres wichtiges Thema, das sich hier anschließt, ist die Anthropomorphisierung: Bieten Hersteller und Entwickler von intelligenten Systemen mehr Anthropomorphisierung an, als ethisch vertretbar wäre? Was richtet dies an, wenn es auf die Tendenz des Menschen trifft, zu anthropomorphisieren, weil es keine alternativen Vorstellungen gibt? Wir können uns in die Denkweise eines anderen Menschen hineinversetzen („Theory of mind“), aber was tun wir, um die Denkweise einer Maschine zu verstehen („Theory of the artificial mind“)? In diesem Kontext sind insbesondere auf LLMs basierende, konversationelle KI-Systeme wie ChatGPT interessant, da sie sich als Agent präsentieren (z. B. in Bezug auf Wissen und Soziabilität). Sie sind in der Lage, Dinge zu beschreiben, die eine soziale Bedeutung haben und soziale Unterscheidungen zu treffen, sie interagieren auf kreative Weise mit den Nutzenden und greifen auf vorherige Gesprächsinhalte zurück (contextual continuity). ChatGPT ist jedoch aufgrund seines probabilistischen Modells zum Beispiel nicht als Schnittstelle für den Abruf von Wissen geeignet. Bei einer klassischen Suchmaschinenanfrage (z. B. beim „Googlen“) wird ein Dokument abgerufen, welches der oder die Nutzende liest und dann entscheidet, ob dieses (unter Einbezug der Quelle) zur Beantwortung der Frage geeignet ist. ChatGPT jedoch präsentiert eine Antwort im konversationellen Kontext ohne Angabe von Quellen, was eine Einschätzung der Glaubwürdigkeit erschwert. Durch die menschliche Tendenz, bewusst und unbewusst auf soziale Hinweisreize zu reagieren[50], kann es durch die überzeugende, menschenähnliche Darstellung von Informationen durch Systeme wie ChatGPT zu einer Misattribution von Vertrauen kommen. Auch die politische Dimension dieser Systeme möchten wir an dieser Stelle hervorheben, denn durch ihre steigende Verbreitung können ihre Aussagen zunehmend Einfluss auf die Meinungsbildung ausüben. Bisherige Forschung konnte beispielsweise zeigen, dass ChatGPT sich politisch eher im linken Spektrum bewegt[52] und dass Sprachassistenten inkorrekte politische Aussagen machen[53].

Vor dem Hintergrund der offenen Fragen und Empfehlungen möchten wir auch für die Zukunft einen Austausch zu ethischen, rechtlichen, psychologischen sowie informatischen Perspektiven bezüglich der menschlichen Interaktion mit intelligenten Systemen anregen. Nur unter Einbezug der Anforderungen, die sich aus den verschiedenen Disziplinen ergeben, können wir uns der Frage stellen, wie man Rahmenbedingungen schafft, unter denen eine sinnvolle Interaktion mit Maschinen gelingen kann.

52 Motoki, F., Pinho Neto, V., & Rodrigues, V. (2023). More human than human: Measuring ChatGPT political bias. *Public Choice.*

53 Ojeda, C. (2021). The political responses of virtual assistants. *Social Science Computer Review, 39*(5), 884-902.

IMPACT

Ein gemeinsames Projekt von

UNIVERSITÄT DUISBURG ESSEN

UNIVERSITÄT BIELEFELD

UNIKASSEL VERSITÄT

Evangelische Hochschule Nürnberg

Gefördert durch

VolkswagenStiftung

DOI: 10.17185/duepublico/81565
URN: urn:nbn:de:hbz:465-20240215-170422-2