Kann sich künstliche Intelligenz selbst erklären?

Wie Erklärungen aus rechtswissenschaftlicher und ethischer Sicht gestaltet sein sollten und was Psychologie und Informatik dazu beitragen können

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

Autoren:
Aike Horstmann, Nicole Krämer, Christian Geminn, Tamer Bile, Carina Weber, Arne Manzeschke, Lina Mavrina, Stefan Kopp, André Artelt, Barbara Hammer

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

Kontakt:
Prof. Dr. Nicole Krämer

Telefon: +49 203 379 - 2482
E-Mail nicole.kraemer@uni-due.de

Universität Duisburg-Essen
Forsthausweg 2
47057 Duisburg

Details zur Publikation:
DOI: 10.17185/duepublico/77378
ISBN (Print): 978-3-940402-73-8

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

1. Auflage, Februar 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

Nutzer*innen technischer Geräte sehen sich mehr und mehr mit Systemen konfrontiert, die Ihnen auf Basis von künstlicher Intelligenz (KI) Auskünfte geben, mit ihnen Konversationen führen und – je nach Anwendung – auch Verhalten oder Entscheidungen vorschlagen. Etwa das von OpenAI[1] zeigt hierbei erstaunliche Funktionalitäten mit einem potentiell sehr breiten Anwendungsspektrum: Es erlaubt Unterhaltungen zu Kochrezepten für spezifische Diäten, die Zusammenfassung von wichtigen in Fachartikeln enthaltenen Informationen, oder auch Verhaltensempfehlungen im Fall von beobachteten Krankheitssymptomen. Auch in Form von Sprachassistenten oder Chatbots bieten ähnliche Systeme Dienste an, die auf die spezifischen Bedürfnisse einer Person abgestimmt werden, indem Daten dieser Person erfasst, gespeichert und weiterverarbeitet werden.

Die Konversation im Sinne eines menschenähnlichen Dialogs wird hierbei als Königsweg der Interaktion auch mit technologischen Systemen gesehen. Dies führt zu spezifischen Erwartungen seitens der Nutzer*innen, birgt aber auch Gefahren in Bezug auf die Konsequenzen unüberlegter oder uninformierter Nutzung. Auch wenn die sprachliche Gewandtheit der mithilfe großer Datenmengen erstellten, konversationalen KI-Systeme auf einem menschenähnlichen Niveau liegen kann, können diese Systeme trotzdem unvollständige oder falsche Informationen liefern. Dabei ist zu beachten, dass dies auch ohne Absicht geschehen kann, bedingt durch die verwendeten Lernmethoden und -architekturen, die man als intransparente „Black Boxes" bezeichnen kann.

Etwa das ChatGPT-System basiert auf Sprachmodellen, welche öffentlich zugänglichen Texte wie Wikipedia-Artikel, Gesetzestexte, oder wissenschaftliche Artikel nutzen, um eine kompakte Repräsentation von Text zu generieren, die implizit typische Strukturen in Dokumenten abbildet. Um daraus sinnvolle Dialoge zu generieren, wurde zudem ein Training auf Basis von durch Menschen gezielt generierten Beispielen für gute Dialoge eingesetzt. An keiner Stelle im Prozess ist dabei das „Wissen", das im resultierenden KI-Modell implizit vorhanden ist, expliziert. Die Schritte und Inferenzen basieren auf aus den Trainingsbeispielen gelernten statistischen Zusammenhängen statt expliziten logischen Regeln oder dem Bewusstsein von Fakten. So werden zum Beispiel wissenschaftliche Quellen "erfunden", indem Autor*innennamen und Titel plausibel zusammengesetzt werden, aber nicht real existieren.

Wie lässt sich nun verhindern, dass Nutzer*innen aufgrund von falschen Annahmen oder fehlendem Wissen auf die natürlichsprachige Konversationsfähigkeit solcher Systeme mit unangemessen viel Vertrauen reagieren? Aus rechtswissenschaftlicher Sicht wurden zwar erste Regelungen formuliert, die Pflichten zur Aufklärung von Nutzer*innen beinhalten, aber im Folgenden soll unter anderem diskutiert werden, inwieweit diese als noch unvollständig bewertet werden müssen. Rechtswissenschaftliche Überlegungen zum Schutz der Nutzer*innen stellen daher sowohl den Ausgangspunkt als auch den Endpunkt der Betrachtungen in diesem Policy Paper dar. Eine besondere Rolle spielt dabei der Begriff der informationellen Selbstbestimmung und die Frage, inwiefern und wodurch sichergestellt

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

werden sollte, dass eine Einwilligung zur Nutzung der Technologie tatsächlich *informiert* erfolgt. Dabei soll es an dieser Stelle nicht vertieft um die Notwendigkeit gehen, dass über die Erhebung und Speicherung von Daten informiert wird (d.h. Transparenz im datenschutzrechtlichen Sinne, damit die betroffene Person nachvollziehen kann, ob die sie betreffende Datenverarbeitung rechtmäßig ist), sondern vorrangig darum, dass den Nutzer*innen beispielsweise erklärt wird, was mit Hilfe der Daten inferiert werden kann und wie beziehungsweise wodurch das System funktioniert. Um den Anforderungen der Informiertheit aus rechtlicher Sicht Genüge zu tun, greifen Anbieter*innen mitunter auf Formulierungen wie „Ihre Daten werden zur Verbesserung des Systems/der Dienste verwendet". Dies ist aber aus psychologischer Sicht nicht als Erklärung zu bezeichnen, durch die die Nutzer*innen wirklich etwas über das System lernen können. So können Nutzer*innen weder Ziel und Zweck der erhobenen Daten ermessen noch etwas darüber erfahren, wie das System funktioniert.

Aus Sicht der modernen interdisziplinären Forschung kann man Erklärungen als soziale Prozesse bezeichnen, an deren Gestaltung die teilnehmenden sozialen Agenten aktiv mitwirken. In einem solchen Prozess können Interaktionsmechanismen realisiert werden, die zur Erfüllung der (nicht nur informationellen) Selbstbestimmung der Nutzer*innen beitragen. Zum Beispiel können in einer Interaktion durch Austausch und Feedback die individuellen Informations- und Erklärbedarfe der Nutzer*innen ermittelt, ihr Verständnis überprüft und Erklärstrategien angepasst werden. Hierbei bieten Systeme wie Sprachassistenten oder Chatbots ein geeignetes Medium für die Realisierung solcher sozialen Erklärprozesse und stehen somit im Fokus dieses Policy Papers.

Erklärungsbedarf - Was sollten Nutzer*innen konversationaler KI-Systeme verstehen können?

Bevor die Erfordernisse sowie mögliche Lösungen aus der Sicht relevanter Disziplinen wie Rechtswissenschaft, Ethik, Psychologie und Informatik erläutert werden, wird zusammengefasst, worin der Erklärungsbedarf besteht. Im Bereich der KI beziehen sich aktuelle Verfahren der Erklärung oft auf durch Bilder oder Tabellen beschriebene Daten, und es wird eine spezifische Funktionalität, meistens eine Klassifikation, erklärt. Allerdings sind die Funktionalitäten, die ein Sprachassistenzsystem erfüllen kann, enorm vielfältig, so dass Erklärungen sich auf unterschiedlichste Facetten beziehen können[2]: Warum bezeichnet eine konversationale KI gewisse Sachverhalte als wahr? Warum präsentiert die konversationale KI einem spezifischen Menschen gewisse Sachverhalte, aber nicht einem anderen? Welche vom Menschen geäußerten privaten Informationen speichert die KI und wozu können diese in Folge potenziell benutzt werden?

Dabei unterscheidet sich der Erklärungsbedarf bei konversationalen KI-Systemen durchaus dahingehend, zu welchem Zweck es hauptsächlich genutzt wird. Bei reinen Informationsanfragen ("Wie wird das Wetter morgen?") werden weniger Daten der Nutzer*innen erfasst und verarbeitet als bei Empfehlungen oder Entscheidungsvorschlägen im Rahmen sogenannter Decision Support Systems. Bei letzterem ist ein wichtiger Teil der rechtlich geforderten informationellen Selbstbestimmtheit betroffen: Zu verstehen, wie und auf Basis welcher Daten die konversationale KI Entscheidungen fällt beziehungsweise

2 Zini, J. E., & Awad, M. (2022). On the explainability of natural language processing deep models. *ACM Computing Surveys, 55*(5), 1-31.

Erklärungsbedarf - Was sollten Nutzer*innen konversationaler KI-Systeme verstehen können?

vorschlägt. An dieser Stelle müssen sowohl Rahmenbedingungen als auch Zusammenhänge im weiteren Kontext erläutert werden. Hier spielt auch eine Rolle, was KI-Verfahren aus scheinbar unverfänglichen Daten machen können, etwa indem sie zusätzliche Quellen (z.B. Social Media Accounts) und darin enthaltene Informationen nutzen.

Von KI-Systemen getroffene Entscheidungen können außerdem einen größeren Kreis an Personen jenseits der unmittelbaren Nutzer*innen betreffen, dessen Erklärungsbedarf sich unter Umständen in Bezug auf das gleiche KI-System unterscheidet.[3] Zum Beispiel kann sich der Erklärungsbedarf von Ärzt*innen bei der Nutzung eines medizinischen Decision Support Systems wesentlich unterscheiden vom Erklärungsbedarf der Patient*innen, die mithilfe dieses Systems diagnostiziert werden. Der direkte und erweiterte soziale Kontext sollte demnach bei der Auswahl und Präsentation von Erklärungen immer mitberücksichtigt werden, zum Beispiel indem man die folgenden Fragen stellt: (1) Warum soll das System erklärbar sein? (2) Für wen soll das System erklärbar sein? (3) Auf welcher Art der Interpretation soll die Erklärung basieren? (4) Wann soll die Erklärung präsentiert werden? und (5) Auf welche Art und Weise soll die Erklärung präsentiert werden?[4] Eine Erklärung sollte somit als soziales Aushandlungsgeschehen betrachtet werden, bei dem die Empfänger*innen immer mitbestimmen, ob die gegebene Erklärung zureichend ist. Jede Erklärung bewegt sich daher in folgendem Spannungsfeld: Welche Art von Erklärung wird gegeben, welche Art von Erklärung wiederum benötigt?

3 Suresh, H., Gomez, S. R., Nam, K. K., & Satyanarayan, A. (2021, May). Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. In Y. Kitamura, A. Quigley, K. Isbister, T. Igarashi, P. Bjørn, & S. Drucker (Hrsg.) *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems - CHI '21* (S. 1-16).

4 Rosenfeld, A., & Richardson, A. (2019). Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems, 33*, 673-705.

Was bedeutet informationelle Selbstbestimmung im Umgang mit konversationaler künstlicher Intelligenz? Rechtliche und ethische Erfordernisse

Ziel von Erklärungen aus verfassungsrechtlicher Sicht ist es zunächst bezogen auf die Verarbeitung personenbezogener Daten aufgeklärte Entscheidungen der betroffenen Personen zu ermöglichen, einen Missbrauch durch die und Übervorteilung der datenverarbeitenden Stellen zu verhindern und damit letztlich das Grundrecht auf informationelle Selbstbestimmung zu sichern. Das Bundesverfassungsgericht erkannte schon früh die fundamentale Bedeutung von Erklärungen für den grundrechtlichen Datenschutz. Im Volkszählungsurteil von 1983 erklärte es eine Ordnung mit dem Recht auf informationelle Selbstbestimmung für unvereinbar, „in der Bürger nicht mehr wissen können, wer was wann und bei welcher Gelegenheit über sie weiß".[5] In einer jüngeren Entscheidung sprach das Bundesverfassungsgericht Erklärungen auch explizit den Zweck zu, bloße Gefährdungen der Persönlichkeit zu vermeiden, die entstehen, „wenn personenbezogene Informationen in einer Art und Weise genutzt und verknüpft werden, die der Betroffene weder überschauen noch beherrschen kann".[6] Im Urteil zum „Großen Lauschangriff" des Bundesverfassungsgerichts wurde die Notwendigkeit von Rückzugsräumen für den Persönlichkeitsschutz gerade „in einer Welt, in der es technisch möglich geworden ist, so gut wie jede Bewegung und Kommunikation einer Person zu verfolgen und aufzuzeichnen"[7] betont. Hieraus lässt sich ableiten, dass betroffene Personen wissen müssen, ob sie in ihren Rückzugsräumen auch dann „sicher" sind, wenn sich dort digitale Technik befindet. Der Europäische Gerichtshof hat im Urteil zu Cookie-Bannern im Internet in Bezug auf Erklärungen konstatiert, dass betroffene Personen durch klare und umfassende Informationen dazu befähigt werden müssen, die Konsequenzen ihrer Einwilligung abschätzen und die Funktionsweise technischer Hilfsmittel verstehen zu können.[8]

Fügt man diese Elemente zu einer Synthese zusammen, so liegt die verfassungsrechtlich geforderte Leistung von Erklärungen darin, dass sie ein selbstbestimmtes Handeln ermöglichen sollen – auch jenseits von Fragen der Verarbeitung personenbezogener Daten. Selbstbestimmung setzt dabei stets die Möglichkeit voraus, Informationen über die Umwelt oder die Parameter einer Interaktion zu erhalten. Dies muss insbesondere für Interaktionen mit KI-Systemen gelten. Beeinträchtigungen der Persönlichkeitsentwicklung müssen soweit möglich verhindert werden.

Mit dem Ziel eine informierte Entscheidung der betroffenen Personen zu fördern[9], setzt der in Art. 5 Abs. 1 lit. a Alt. 3 DSGVO normierte Transparenzgrundsatz zudem voraus, dass personenbezogene Daten „in einer für die betroffene Person nachvollziehbaren Weise" verarbeitet werden. Dies beschränkt sich nicht auf ein Auskunftsrecht der betroffenen Personen, sondern erstreckt sich auf alle Informationen und Informationsmaßnahmen, die erforderlich sind, damit die betroffene Person überprüfen kann, ob die sie betreffende Datenverarbeitung rechtmäßig ist und sie ihre Rechte wahrnehmen kann.[10]

5 BVerfGE 65, 1, 43.
6 BVerfGE 118, 168 (184).
7 BVerfGE 109, 279 (382 f.).
8 EuGH – C-673/17, Rn. 74; s. auch für einen ähnlich gelagerten Fall EuGH – C-61/19, Rn. 40.
9 Jaspers u.a., in Schwartmann u.a.: DSGVO/BDSG-Kommentar 2020, Art. 5, Rn. 34.
10 Roßnagel, in: Simitis/Hornung/Spiecker gen. Döhmann, Datenschutzrecht-Kommentar 2019, Art. 5 DSGVO, Rn. 50.

Informationen/Erklärungen sind gem. Art. 12 Abs. 1 DSGVO „in präziser, transparenter, verständlicher und leicht zugänglicher Form in einer klaren und einfachen Sprache zu übermitteln"[11]. Dieser Grundsatz betrifft insbesondere die Informationen über die Identität des Verantwortlichen und die Zwecke der Verarbeitung sowie sonstige Informationen, die eine faire und transparente Verarbeitung im Hinblick auf die betroffenen Personen gewährleisten, sowie deren Recht, eine Bestätigung und Auskunft darüber zu erhalten, welche sie betreffende personenbezogene Daten verarbeitet werden.[12]

Es besteht zudem die Pflicht des Verantwortlichen bei einer automatisierten Entscheidung einschließlich Profiling (Verwendung personenbezogener Daten, um bestimmte persönliche Aspekte zu bewerten, zu analysieren oder vorherzusagen; s. Art. 4 Nr. 4 DSGVO) gemäß Art. 22 Abs. 1 und 4 DSGVO aussagekräftige Informationen über die involvierte Logik sowie die Tragweite und die angestrebten Auswirkungen einer derartigen Verarbeitung für die betroffene Person bereitzustellen.[13] Allerdings besteht die Gefahr, dass die automatisierten Entscheidungen vom menschlichen „Letztentscheidenden" nicht selten unreflektiert übernommen werden oder zumindest die Letztentscheidung beeinflussen, dabei aber angesichts einer fehlenden Informationsverpflichtung auf Seiten des Verantwortlichen eine Information/Erklärung jedoch unterbleibt.[14]

Darüber hinaus setzt Art. 4 Nr. 11 DSGVO voraus, dass eine Einwilligungserklärung der betroffenen Personen in „informierter Weise" abgegeben werden wird. Die erforderliche Informiertheit ist gegeben, wenn die betroffene Person vor Erklärung der Einwilligung in die Verarbeitung ihrer personenbezogenen Daten Kenntnis über alle entscheidungsrelevanten Risiken und Vorteile hat. Dies ist der Fall, wenn die betroffene Person vorab über die Identität des Verantwortlichen der Datenverarbeitung informiert wird und ihr verständlich gemacht wird, welche ihrer personenbezogenen Daten zu welchem Zweck verarbeitet werden.[15] Ferner muss ihr unter anderem erläutert werden, wofür ihre Daten im Rahmen automatischer Entscheidungen (insbesondere Profiling) genau verwendet werden und welche Folgen und mögliche Risiken die Datenverarbeitung für sie birgt.

Auch jenseits des Datenschutzrechts finden sich relevante Vorgaben zu Transparenz und Erklärungen oder sie werden zumindest diskutiert. Hier kann zunächst auf § 18 Abs. 3 MStV[16] verwiesen werden. Die Norm stellt eine kommunikative Selbstbestimmung, aber auch die informationelle Selbstbestimmung sicher, indem sie etwa verhindern, dass bewusst oder unbewusst bestimmte Informationen preisgegeben werden, die sonst zurückgehalten würden. Der Staat kommt durch die Regelung seiner generellen Verpflichtung Rechnung, "eine funktionierende Kommunikationsordnung zu gewährleisten, um die Wahrnehmung der Freiheiten durch die Bürger zu ermöglichen".[17] Standards zur Herstellung von Transparenz müssen sich in diesen Bereichen aber erst noch etablieren. Ein anderes Beispiel findet sich im Entwurf eines Gesetzes über künstliche Intelligenz.[18] Nach Art. 52 Abs. 1 des Entwurfs stellen die Anbieter „[…] sicher, dass KI-Systeme, die für die Interaktion mit natürlichen Personen bestimmt sind, so konzipiert und entwickelt werden, dass natürlichen

11 S. auch Erwägungsgrund 58 Satz 1 DSGVO.
12 Erwägungsgrund 39 Satz 4 DSGVO.
13 Art. 13 Abs. 2 lit. f sowie Art. 14 Abs. 2 lit. g DSGVO.
14 S. Roßnagel/Geminn 2020, 83 ff., 129 f.
15 S. Erwägungsgrund 42 Satz 4 DSGVO.
16 S. Löber/Roßnagel, MMR 2019, 493.
17 Löber/Roßnagel, MMR 2019, 493 (494).
18 S. zum Entwurf Geminn, ZD 2021, 354 (358); Kalbhenn, ZUM 663 (669 ff.).

Personen mitgeteilt wird, dass sie es mit einem KI-System zu tun haben, es sei denn, dies ist aufgrund der Umstände und des Kontexts der Nutzung offensichtlich“. Die Europäische Kommission spricht sich bezogen auf den Einsatz von künstlicher Intelligenz damit für letztlich “nur minimale Transparenzpflichten” aus.[19] Sie “gelten für “Systeme, die i) mit Menschen interagieren, ii) zur Erkennung von Emotionen oder zur Assoziierung (gesellschaftlicher) Kategorien anhand biometrischer Daten eingesetzt werden oder iii) Inhalte erzeugen oder manipulieren („Deepfakes“)”; Ziel ist es zu ermöglichen, dass “bewusste Entscheidungen getroffen oder bestimmte Situationen vermieden werden” können.[20]

Es scheint mittlerweile illusorisch, dass wir in der heutigen digitalen Gesellschaft in der Lage sind, unsere Daten und ihre Verarbeitung durch Algorithmen selbst zu kontrollieren.[21] Hier muss das Konzept der informationellen Selbstbestimmung ausgeweitet und angepasst werden, wie es beispielsweise der Deutsche Ethikrat vorschlägt: Der Begriff der informationellen Freiheitsgestaltung ist an dieser Stelle anschlussfähig, denn er gründet nicht in einem „eigentumsanalogen Ausschlussrecht“ – vielmehr tritt die Befugnis und damit die Fähigkeit in den Mittelpunkt, selbst darüber bestimmen zu können, mit welchen Informationen und Inhalten Nutzer*innen in Beziehung zu ihrer Umwelt treten.[22]

Erklärungen zu einem Sachverhalt, der ethische Implikationen aufweist, müssen außerdem noch einmal genauer betrachtet werden. Hier stellt sich die Frage, ob das, was ethische Deliberation ausmacht, in Interaktion mit beispielsweise einem Decision Support System, überhaupt erreicht werden kann – welche Bedingungen müssten erfüllt sein, um an dieser Stelle von einer ethischen Erklärung sprechen zu können? In diesem Sinne ist eine Erklärung auch als eine Rechtfertigung für eine bestimmte Handlung beziehungsweise Entscheidung zu verstehen. Ethische Deliberationen bestehen in einem ausführlichen Austausch über Gründe, idealerweise wird dies so lange fortgeführt, bis eine Partei die Argumentation der anderen hinreichend nachvollziehen kann. Hierzu ist eine diskursive Form des Austauschs inklusive Nachfragen und Klärungen nötig.

19 S. die Begründung in COM(2021) 206 final, 1.1.; s. auch ebd., 2.3: “Für andere, KI-Systeme, die kein hohes Risiko darstellen, werden nur sehr wenige Transparenzpflichten auferlegt, etwa dahingehend, dass bei der Interaktion mit Menschen der Einsatz von KI-Systemen angezeigt werden muss.”

20 S. COM(2021) 206 final, 5.2.4.

21 Baumann, M. O. (2016). Privatsphäre als ethische und liberale Herausforderungen der digitalen Gesellschaft. *Information-Wissenschaft & Praxis, 67*(1), 1-6.

22 Deutscher Ethikrat (2018). *Big Data und Gesundheit – Datensouveränität als informationelle Freiheitsgestaltung. Stellungnahme.* Deutscher Ethikrat.

Was macht eine Erklärung hilfreich? Psychologische Erkenntnisse zu relevanten Mechanismen

Um eine informierte Einwilligung geben zu können, bedarf es aus psychologischer Sicht einer entsprechend den Bedürfnissen der betroffenen Personen gestaltete Erklärung. Ziel ist es, dass die Personen über die wesentlichen Aspekte des Systems im Sinne der Datenerhebung und Datenverarbeitung aufgeklärt werden und in der Lage sind, diese Information zu verstehen und angemessen zu bewerten[23], um eine autonome Entscheidung treffen zu können[24] [25]. Generell umfassen autonome Entscheidungen neben physischen Voraussetzungen vor allem kognitive und affektive Aspekte [25] [26], welche unterschiedliche psychologische Anforderungen an eine hilfreiche Erklärung nach sich ziehen. Darüber hinaus sollte beachtet werden, wann eine Erklärung präsentiert wird, sodass sie wahrgenommen und verarbeitet werden kann. Eine Erklärung kann noch so gut gestaltet sein, sie verfehlt ihren Zweck, wenn sie nicht zur Kenntnis genommen wird. Da die Beziehung zwischen Mensch und System fundamental interaktional ist[27], sollten auch Erklärungen möglichst dynamisch gestaltet werden, sodass sie natürlich im Dialog integriert werden können.

Bezüglich der Anforderungen auf kognitiver Ebene sollte eine Erklärung Personen beim Verständnis der relevanten Informationen, bei der Einsicht in die Situation und die möglichen Konsequenzen sowie bei der rationalen Verarbeitung von Informationen unterstützen[28]. Eine hilfreiche Erklärung sollte Informationen über die erwartbaren Vor- und Nachteile der jeweiligen Entscheidungssituation vollständig offenlegen.[25] Dabei muss eine hilfreiche Erklärung nicht unbedingt die Wahrnehmung erzeugen, dass ein System leicht zu verstehen ist. In einer unserer Projektstudien wurde der Algorithmus eines KI-basierten Empfehlungssystems im Falle der erfolgreicheren Erklärung sogar als weniger leicht interpretierbar empfunden obwohl faktisch das Verständnis höher war.[29]

Bei Betrachtung der kognitiven Ebene sollte zudem der Aspekt der minimalen Ablenkung beachtet werden, welcher oft im Kontrast zur vollständigen Offenlegung steht. Denn wenn Erklärungen zu sehr von der aktuellen Aufgabe ablenken (z.B. durch große Informationsmengen, viele Anfragen, komplexe Funktionalität), tendieren Nutzer*innen dazu, sich von Prozessen der informierten Einwilligungen zurückzuziehen. Dies kann dazu führen, dass sie routinemäßig einwilligen oder ablehnen, ohne sich intensiver mit dem aktuellen Kontext zu beschäftigen und ohne die für eine informierte Einwilligung essenziellen Informationen bewusst wahr- und aufzunehmen.[25] Als Lösungsansatz wird hier „Informieren

23 Appelbaum, P. S. (2010). Consent in impaired populations. *Current Neurology and Neuroscience Reports, 10*(5), 367–373.

24 Friedman, B., Felten, E., & Millett, L. I. (2000). Informed consent online: A conceptual model and design principles. *University of Washington Computer Science & Engineering Technical Report 00–12–2*.

25 Friedman, B., Howe, D. C., & Felten, E. (2002). Informed consent in the Mozilla browser: implementing value-sensitive design. *Proceedings of the 35th Annual Hawaii International Conference on System Sciences* (S. 10). IEEE.

26 Schütz, H., Heinrichs, B., Fuchs, M., & Bauer, A. (2016). Informierte Einwilligung in der Demenzforschung. Eine qualitative Studie zum Informationsverständnis von Probanden. *Ethik in der Medizin, 28*(2), 91–106.

27 Friedman, B., & Hendry, D. G. (2019). *Value Sensitive Design*. The MIT Press.

28 Grisso, T., & Appelbaum, P. S (1995). Comparison of standards for assessing patients' capacities to make treatment decisions. *American Journal of Psychiatry, 152*(7), 1033–1037.

29 Szczuka, J., Horstmann, A. C., Mavrina, L., Artelt, A., Strathmann, C., Szymczyk, N., Bohnenkamp, L. M., & Krämer, N. C. (under review). Enhancing the Understanding of Algorithms With Contrastive Explanations: An Experimental Study on the Effects of Explanations and Person-Likeness on Trust in and Understanding of Algorithms.

durch Interaktion“ vorgeschlagen: die dynamische Einbindung von Erklärungen in die Interaktion, um Unterbrechungen oder exzessive Ressourcenbeanspruchung zu vermeiden. So kann eine natürliche Wahrnehmung und Verarbeitung der Erklärung gefördert werden, ohne übermäßige Ablenkung. Friedmann et al. (2002) nennen dieses Vorgehen just-in-time intervention, wobei die Nutzer*innen in dem Moment, wo Daten erhoben werden, mit einer entsprechenden Erklärung versorgt werden.[25] Gleichzeitig sollte an dieser Stelle die Möglichkeit geboten werden, eine (neue) Entscheidung basierend auf den neu erhaltenen Informationen zu treffen. Das bestehende Wissen sowie Erklärungstexte können zudem durch in die Interaktion eingebettete visuelle Hinweise erweitert werden, welche das Verständnis von der Funktionsweise und Sicherheit eines Systems fördern.[30] Zum Beispiel markiert bereits ein Schloss-Icon links von der Adressleiste im Webbrowser, ob die Verbindung verschlüsselt und somit sicher ist und ein rotes Ausrufezeichen-Symbol kennzeichnet eine aus dem Internet heruntergeladene Datei, die ein mögliches Sicherheitsrisiko beinhaltet.

Zusätzlich zu den kognitiven Aspekten, haben auch affektive Aspekte wie Motivation, Emotion und sozialer Kontext einen Einfluss darauf, wie hilfreich eine Erklärung ist.[26] [31] Eine noch so verständlich formulierte Erklärung kann ihr Ziel verfehlen, wenn nicht auf die Motivation der Personen eingegangen wird. Um die Motivation der Nutzer*innen zu unterstützen, sollten ihre Ziele in Bezug auf die Nutzung von Erklärungen erkannt und berücksichtigt werden. Als Ergebnis einer Interviewstudie wurden drei Zielbereiche ermittelt: a) Erhalt weiterer Erkenntnisse oder Beweise, um eine informierte Entscheidung zu treffen, die Entscheidungssicherheit zu erhöhen, die eigene Voreingenommenheit zu vermindern, oder Kausalitätszusammenhänge zu verstehen, b) Evaluation der Leistungs- und Anpassungsfähigkeit sowie Restriktionen des Systems und c) Anpassung des Nutzungs- oder Interaktionsverhalten, um die KI besser zu nutzen.[32] Um zum Beispiel eine gute und aufgeklärte Entscheidung treffen zu können, wird eine Erklärung benötigt, die den Grund für eine Empfehlung oder Einschätzung offenlegt. Wird die Erklärung den Zielen der Nutzer*innen gerecht, steigert dies ihre Motivation diese auch zu nutzen.
Menschen beziehen zudem ihre eigenen Emotionen und die der anderen in ihre Erklärungen mit ein, um zum Beispiel ihre Absichten oder ihr Wohlbefinden beziehungsweise Unbehagen zu erklären (z.B. „Ich rannte weg, weil ich Angst vor einer bedrohlich wirkenden Person hatte”).[33] [34] Durch KI simulierte Emotionen können erstens als Heuristik dienen, um geeignete Inhalte zur Erklärung einer Handlung zu finden (z.B. „Ich rannte weg, weil eine Person eine Waffe in der Hand hielt”). Zweitens können Emotionen genutzt werden, um die

30 Friedman, B., Lin, P., & Miller, J. K. (2008). Informed consent by design. In L. F. Cranor & S. Garfinkel (Hrsg.), *Security and Usability: Designing Secure Systems that People Can Use* (S. 495–521). O'Reilly Media, Inc.

31 Vollmann, J. (2000). Einwilligungsfähigkeit als relationales Modell. Klinische Praxis und medizinethische Analyse. Der Nervenarzt, 71(9), 709–714.

32 Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: Informing design practices for explainable AI user experiences. In R. Bernhaupt, F. Mueller, D. Verweij, J. Andres, J. McGrenere, A. Cockburn, I. Avellino, A. Goguey, P. Bjørn, S. Zhao, B. P. Samson, & R. Kocielnik (Hrsg.), *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems - CHI '20* (S. 1–15). ACM.

33 Hudlicka, E. (2003). To feel or not to feel: The role of affect in human–computer interaction. *International Journal of Human-Computer Studies, 59*(1-2), 1–32.

34 Kaptein, F., Broekens, J., Hindriks, K., & Neerincx, M. (2017). The role of emotion in self-explanations by cognitive agents. *Proceedings of the Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos - ACIIW '17* (S. 88–93). IEEE.

Was macht eine Erklärung hilfreich? Psychologische Erkenntnisse zu relevanten Mechanismen

Erklärungen natürlicher (menschenähnlicher) zu formulieren, beispielsweise indem sie selbst Inhalt der Erklärung sind (z.B. „Ich rannte weg, weil ich Angst hatte, getötet zu werden"). Drittens können Emotionen helfen, Erklärungen verständlicher zu machen indem die zugrundeliegenden Beurteilungsprozesses erklärt werden (z.B. „Ich hatte Angst, weil da eine Person mit einer Waffe war und Waffen können tödlich sein."). [34]

Als letzter Punkt aus psychologischer Sicht ist es wichtig, den sozialen Kontext von Erklärungen zu berücksichtigen. Auch unbeabsichtigt kann ein sich selbst erklärendes System eine soziale Rolle einnehmen. Ein System, welches die Rolle eines*einer Arbeitskolleg*in inne hat, wird wahrscheinlich eher auf Augenhöhe und positiv wahrgenommen, was das Vertrauen in eine Erklärung dieses Systems steigern könnte.[35] Ein System in einer Tutor*inrolle mit entsprechender (wahrgenommener) Expertise wird wahrscheinlich als kompetenter und daher glaubwürdiger empfunden.[36] [37] [38] Auch wenn die Wirkung der verschiedenen potenziellen sozialen Rollen auf die Wahrnehmung einer Erklärung noch nicht erforscht wurde, ist es wichtig zu erwähnen, dass diese einen nicht zu unterschätzenden Einfluss haben könnten.

Zusammenfassend lässt sich sagen, dass aus psychologischer Sicht verschiedene Mechanismen in unterschiedlicher Intensität und in noch nicht gänzlich absehbaren Wirkungsgefügen die Qualität einer Erklärung beeinflussen können. Um eine möglichst hilfreiche Erklärung zu bieten, sollten kognitive sowie motivationale, emotionale und soziale Mechanismen im Zusammenspiel betrachtet und adressiert werden. Darüber hinaus ist auch das Timing für die Präsentation von Erklärungen wichtig. Am besten werden diese dynamisch in die Interaktion mit den Systemen eingebunden, um eine natürliche Wahrnehmung und Verarbeitung der enthaltenen Informationen zu begünstigen.

35 Nass, C., Fogg, B. J., & Moon, Y. (1996). Can computers be teammates? *International Journal of Human-Computer Studies, 45(*6), 669–678.

36 Dijks, M. A., Brummer, L., & Kostons, D. (2018). The anonymous reviewer: The relationship between perceived expertise and the perceptions of peer feedback in higher education. *Assessment & Evaluation in Higher Education, 43(*8), 1258–1271.

37 Fogg, B. J. (2002). Persuasive technology: Using computers to change what we think and do. *Ubiquity*(5), 89–120.

38 Krämer, N. C., Leiße, L.-M., Hollingshead, A., & Gratch, J. (2017). Evaluated by a machine: Effects of negative feedback by a computer or human boss. In J. Beskow, C. Peters, G. Castellano, C. O'Sullivan, I. Leite, & S. Kopp (Hrsg.), *Intelligent Virtual Agents: Proceedings of the 17th International Conference on Intelligent Virtual Agents - IVA '17* (S. 235–238). Springer.

Kann KI sich selbst erklären? Möglichkeiten und Ansätze aus der Informatik

In der KI-Forschung spielt aktuell die so-genannte erklärbare KI (auf Englisch XAI für „eXplainable AI") eine große Rolle.[39] Dieser Begriff bezieht sich dabei generell auf Methoden, die Teile des (Black-Box) KI-Mechanismus in einer für den Menschen zugänglichen Form darstellen. Um KI-Systeme zu erklären, wurden unterschiedliche Erklärmethoden entwickelt: Man unterscheidet zum Beispiel globale und lokale Erklärungen. Globale Erklärungen versuchen das gesamte System zu erklären, wohingegen lokale Erklärungen in der Regel nur einen konkreten Fall erklären. Für eine lokale Erklärungen kann zum Beispiel das Faktum herausgestellt werden, das für eine gegebene Empfehlung den größten Ausschlag gegeben hat. Darüberhinausgehend unterscheidet man Erklärmethoden dahingehend, wie und in welchem Medium sie Erklärungen ausdrücke. So können für Entscheidungen oder Modelle wichtige einzelne Aspekte hervorgehoben, visuelle Darstellungen genutzt, prototypische Fälle als Vergleich herangezogen oder logische Argumentationen geliefert werden.[39]

Im Fall von Sprachassistenten sind Erklärungen in der Regel in natürlicher Sprache formuliert oder mit Sprache assoziiert und damit für den Menschen unmittelbar zugänglich. Die Dialoge mit den heute vorhandenen kommerziellen Sprachassistenten wie Alexa oder Google Assistant werden normalerweise von Menschen entwickelt und gestaltet. Die Entwickler*innen definieren die Art der Eingabe, die von den Nutzer*innen erwartet wird und bestimmte Intentionen ausdrückt, sowie die Art und Weise, wie das System auf diese Intentionen reagiert. Die Intentionen werden hierbei mithilfe von KI erkannt, die Antworten des Systems dagegen werden oft manuell zusammengestellt. Hier haben die Entwickler*innen demnach einen großen Einfluss darauf, wie der Assistent sich verhält, zum Beispiel in Bezug auf die Präsentation der Informationen oder die Ausdrucksweise.
Allerdings existieren auch Systeme, die ohne von Menschen vorgefertigte Antworten Dialoge führen können, wie zum Beispiel ChatGPT[1] oder LaMDA[40]. Mithilfe von sogenannten Transformer-Architekturen werden diese Sprachmodelle mit großen Mengen an Textdaten vortrainiert und lernen somit Muster, nach denen Sätze, Texte und Dialoge in einer Sprache zusammengesetzt werden.[41] Außerdem erhalten die Sprachmodelle aus den Textdaten implizit gelerntes Wissen, welches in der Ausgabe reproduziert wird, um Fragen der Nutzer*innen zu beantworten oder von ihnen gestellte Aufgaben zu lösen.[40] [42] In dem folgenden Dialog kann man sehen, wie LaMDA auf die Anfrage eines*einer Nutzer*in reagiert, die Transformer-Architektur, das heißt das eigene Funktionieren, zu erklären.[40] [43]

39 Molnar, C. (2020). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.* Independently published.

40 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.*

41 Li, H. (2022). Language models: past, present, and future. *Communications of the ACM, 65*(7), 56-63.

42 Lee, N., Li, B. Z., Wang, S., Yih, W. T., Ma, H., & Khabsa, M. (2020). Language models as fact checkers? *arXiv preprint arXiv:2006.04102.*

43 *Original dialogue*: User: How do transformers work? LaMDA: They are giant robots from the planet Cybertron that can transform into vehicles, and some can transform into objects, such as a cassette tape. User: Sorry, I mean in the context of AI. LaMDA: If you mean artificial neural networks, well they're networks that simulate how human brains act, and work. User: How do brains work? LaMDA: They send and receive electrical signals and chemical messengers from our brain cells that make us think, feel, move, and live.

*Nutzer*in:* Wie funktionieren Transformer?
LaMDA: Das sind riesige Roboter vom Planeten Cybertron, die sich in Fahrzeuge verwandeln können, und manche können sich in Gegenstände verwandeln, zum Beispiel in eine Kassette.
*Nutzer*in:* Entschuldigung, ich meine im Zusammenhang mit KI.
LaMDA: Wenn du künstliche neuronale Netze meinst, dann sind das Netze, die simulieren wie menschliche Gehirne arbeiten und handeln.
*Nutzer*in:* Wie funktionieren Gehirne?
LaMDA: Sie senden und empfangen elektrische Signale und chemische Botenstoffe von unseren Gehirnzellen, die uns denken, fühlen, bewegen und leben lassen.

Dabei ist zu beachten, dass die Sprachmodelle über begrenzte Schlussfolgerungsfähigkeiten verfügen und somit bei komplexen Schlussfolgerungsaufgaben scheitern können, sowie keine expliziten Wissensmodelle über die jeweiligen Gesprächsthemen beinhalten.[40] [42] [44] [45] Die Erklärungen entstehen durch die Generierung von Text durch das Sprachmodell, das allgemein gesehen einen Imitationsvorgang auf Basis der zum Trainieren des Systems verwendeten Daten darstellt. Die Dialogelemente werden dabei wortweise anhand von gelernten Wahrscheinlichkeiten generiert und es wird nicht überprüft, ob die daraus entstehende komplette Antwort mit im Web zu findenden Fakten übereinstimmt. Wenn man also im Dialog mit einer KI nach einer Erklärung für eine Ausgabe fragt, erhält man eine Antwort, die jedoch nicht unbedingt mit den tatsächlichen Ursachen für die Ausgabe zusammenhängt. Die „Erklärung" wurde ebenfalls nur durch die zugrundeliegende Black Box erzeugt, welche auf der Basis von Daten trainiert wurde. Dadurch weisen die Erklärungen zwar eine hohe sprachliche Qualität auf und klingen menschenähnlich, können jedoch fehlerhaft oder unvollständig sein. In diesem Kontext gibt es Bemühungen, Sprachmodelle selbst evaluieren zu lassen, ob sie bestimmte Fragen korrekt beantwortet haben und wie korrekt ihre Antworten sind.[46] Neben der Präzisierung und Realisierung kognitiver, ethischer und rechtlicher Anforderungen an Erklärungen, existieren somit auch noch eine Reihe grundlegender mathematischer Herausforderungen.

Eine elementare Schlussfolgerung der interdisziplinären Betrachtungen ist, dass der Prozess des Verstehens und Erklärens interaktiv (d.h. konversational) und nicht in einem Schritt erledigt ist. Generell sind Erklärungen als soziale Interaktionen zwischen der Person, die erklärt, und der Person, der erklärt wird, zu verstehen.[47] [48] [49] Diese Personen stehen als soziale Agenten im Vordergrund des Erklärprozesses: ihre Ziele, Vorstellungen, Kenntnisse,

44 Zhou, X., Zhang, Y., Cui, L., & Huang, D. (2020, April). Evaluating commonsense in pre-trained language models. *Proceedings of the AAAI Conference on Artificial Intelligence* (Ausg. 34, Bd. 05, S. 9733-9740).

45 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.

46 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.*

47 Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267,* 1-38.

48 Dazeley, R., Vamplew, P., Foale, C., Young, C., Aryal, S., & Cruz, F. (2021). Levels of explainable artificial intelligence for human-aligned conversational explanations. *Artificial Intelligence, 299*, 103525.

49 Rohlfing, K. J., Cimiano, P., Scharlau, I., Matzner, T., Buhl, H. M., Buschmeier, H., ... & Wrede, B. (2020). Explanation as a social practice: Toward a conceptual framework for the social design of AI systems. *IEEE Transactions on Cognitive and Developmental Systems, 13*(3), 717-728.

Persönlichkeiten und Erwartungen beeinflussen aktiv die Gestaltung dieses Prozesses.[49] Diese Aspekte werden jedoch in aktuellen Ansätzen der erklärbaren KI (explainable AI) oft noch nicht (ausreichend) berücksichtigt.[50] Stattdessen betreffen Erklärungen hier in der Regel einfache Zuordnungen oder Klassifikationen – das heißt, die Erklärung wird jeweils nur für eine spezifische Entscheidung oder ein Entscheidungsmodell generiert. Es wird dabei aber nicht berücksichtigt, ob der den Menschen interessierende Aspekt erklärt wurde, oder eine ganz andere Funktionalität und ob der Mensch die Erklärung auch in der persönlich benötigten Tiefe verstanden hat. In der Realität beziehen sich Erklärungen oft nicht auf die Ziele der Endnutzer*innen.[51] Dazu kommt, dass Erklärungen von den in der Praxis eingesetzten KI-Systemen meist statisch sind und nach der Generierung nicht mehr beim System verdeutlicht oder hinterfragt werden können.[52] In diesem Kontext sind insbesondere nutzerspezifische (d.h. personalisierte) Erklärungen interessant und relevant.[53] Expert*innen schlagen daher interaktive natürlichsprachliche Erklärdialoge vor, in denen zielführende Nachfragen seitens des Menschen möglich sind, die das System kontext-abhängig interpretieren und bearbeiten soll.[52]

Zur Evaluation seiner Erklärungen sollte das System Feedbacksignale des*der Nutzer*in im Kontext der aktuellen Interaktionssituation interpretieren und unter Umständen auch Strategien entwickeln, um diese Signale den Nutzer*innen zu entlocken. Je nach Art des Systems können die Feedbacksignale verschiedene Formen annehmen. Oft geben Menschen nur kurze verbale oder nonverbale Signale während einer Konversation (z.B. „mhm" oder Nicken). Dieses Feedback hat eine wichtige Funktion: damit können Gesprächspartner*innen Informationen über die mentalen Zustände wie Wahrnehmung, Verstehen oder Zustimmung austauschen.[54] Es ist demnach wichtig für kommunikative Systeme, diese Signale bearbeiten und zuordnen zu können.[55] [56] Auf einer höheren Abstraktionsebene könnte das System durch die erhaltenen Feedbacksignale auch Inferenzen über die Vorstellungen und das Wissen (die sogenannten „Beliefs", also Annahmen) des*der Nutzer*in machen.[57] Beide Arten von Inferenzen können dazu verwendet werden, sowohl kurz- als auch langfristig Erklärungen und Erklärstrategien des Systems besser an die Nutzer*innen anzupassen.[58]

50 Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. *Entropy, 23*(1), 18.

51 Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., ... & Eckersley, P. (2020). Explainable machine learning in deployment. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (S. 648-657).

52 Lakkaraju, H., Slack, D., Chen, Y., Tan, C., & Singh, S. (2022). Rethinking Explainability as a Dialogue: A Practitioner's Perspective. *arXiv preprint arXiv:2202.01875.*

53 Schneider, J., & Handali, J. (2019). Personalized explanation in machine learning: A conceptualization. *arXiv preprint arXiv:1901.00770.*

54 Allwood, J., Nivre, J., & Ahlsén, E. (1992). On the semantics and pragmatics of linguistic feedback. *Journal of Semantics, 9(*1), 1-26.

55 Kopp, S., Allwood, J., Grammer, K., Ahlsén, E., & Stocksmeier, T. (2008). Modeling embodied feedback with virtual humans. *Lecture Notes in Computer Science, 4930*, 18.

56 Buschmeier, H., & Kopp, S. (2018). Communicative listener feedback in human-agent interaction: Artificial speakers need to be attentive and adaptive. *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems* (S. 1213-1221).

57 Shvo, M., Klassen, T. Q., & McIlraith, S. A. (2020). Towards the role of theory of mind in explanation. *Explainable, Transparent Autonomous Agents and Multi-Agent Systems: Second International Workshop, EXTRAAMAS 2020* (S. 75-93). Springer International Publishing.

58 Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267,* 1-38.

Schlussfolgerungen und notwendige Weiterentwicklungen und Regulierungen

Zusammenfassend lässt sich festhalten, dass die Informationspflichten aus dem Datenschutzrecht zwar die informationelle Selbstbestimmung befördern, aber auch wesentliche Lücken hinsichtlich der Umsetzung lassen. Insbesondere wird nicht spezifiziert, wie umfassend das Verständnis sein muss, das informationelle Selbstbestimmung ermöglicht. Zum Beispiel wird nicht expliziert gefordert, dass Nutzer*innen in der Lage sein müssen, zu verstehen, welche Erkenntnisse aus der Auswertung von personenbezogenen (und auch anonymisierten) Daten inferiert werden können. Die Frage, ob zum Beispiel aus vermeintlich „harmlosen" Informationen ganz andere Schlüsse gezogen werden können (z.B. Ableitung des Gesundheitszustandes aus dem Einkaufverhalten), ist jedoch für die Entscheidung, einen Dienst zu nutzen oder nicht, mindestens genauso relevant wie die Frage, ob Informationen mit Dritten geteilt werden. Bei datenintensiven Diensten, die auf künstlicher Intelligenz basieren, verschärft sich diese Problematik noch. Besondere Defizite bestehen im Bereich der automatisierten Entscheidungsfindung nach Art. 22 DSGVO, da dem datenschutzrechtlich Verantwortlichen „Tür und Tor geöffnet" wird, sich von seinen Transparenzpflichten zu entziehen.[59] Insgesamt fordert das einfache Recht die Bereitstellung von Informationen im Kontext des Einsatzes von künstlicher Intelligenz nur allgemein bezogen auf die damit verbundene Verarbeitung personenbezogener Daten. Es wird dabei jedoch stets lediglich informiert, nicht notwendigerweise erklärt. Bezogen auf die verfassungsrechtlich geforderte Sicherung von Selbstbestimmung und den Schutz der Persönlichkeit, die eng mit der Möglichkeit des Verstehens einzelner Umstände wie auch von Gesamtzusammenhängen verknüpft ist, bleiben gerade im Kontext von künstlicher Intelligenz erhebliche Desiderata, die wohl auch durch das geplante Gesetz über Künstliche Intelligenz nicht adressiert werden.

Vor diesem Hintergrund ergeben sich zudem ethische, psychologische und informatische – mitunter äußerst komplexe – Aspekte, die berücksichtigt werden müssten, damit eine Einwilligung zur Nutzung einer Technologie tatsächlich *informiert* erfolgt. Mit großer Nachdrücklichkeit wird dafür plädiert, Erklärungen als soziale Prozesse zu betrachten, an deren Gestaltung die Nutzer*innen sowie das KI-System aktiv mitwirken. Durch Interaktion ist es am ehesten möglich, den individuellen Erklärbedarf der Nutzer*innen sowie spezifische kognitive, affektive und soziale Anforderungen zu ermitteln und berücksichtigen. Zudem bietet die Interaktion mit dem KI-System diesem die Möglichkeit das Verständnis der Nutzer*innen zu überprüfen und gegebenenfalls Erklärungen sowie Erklärstrategien durch Fragen und Feedback anzupassen. Erklärbarkeit interaktiv zu gestalten ist komplex und bringt einige Herausforderungen mit sich, es gibt jedoch bereits informatische Ansätze und Perspektiven. Somit lässt sich auch hinsichtlich dieser prospektiven Lösung feststellen, dass das Ziel der informationellen Selbstbestimmung diverse Anforderungen an verschiedene Fachbereiche mit sich bringt, welche nur durch interdisziplinäre Zusammenarbeit adressiert werden können.

59 S. Roßnagel/Geminn, Datenschutz-Grundverordnung verbessern, S. 83 ff., 129 f.

IMPACT

Ein gemeinsames Projekt von

UNIVERSITÄT DUISBURG ESSEN

UNIVERSITÄT BIELEFELD

UNIKASSEL VERSITÄT

Evangelische Hochschule Nürnberg

Gefördert durch

VolkswagenStiftung

DOI: 10.17185/duepublico/77378
URN: urn:nbn:de:hbz:465-20230308-093622-1