Missing Link

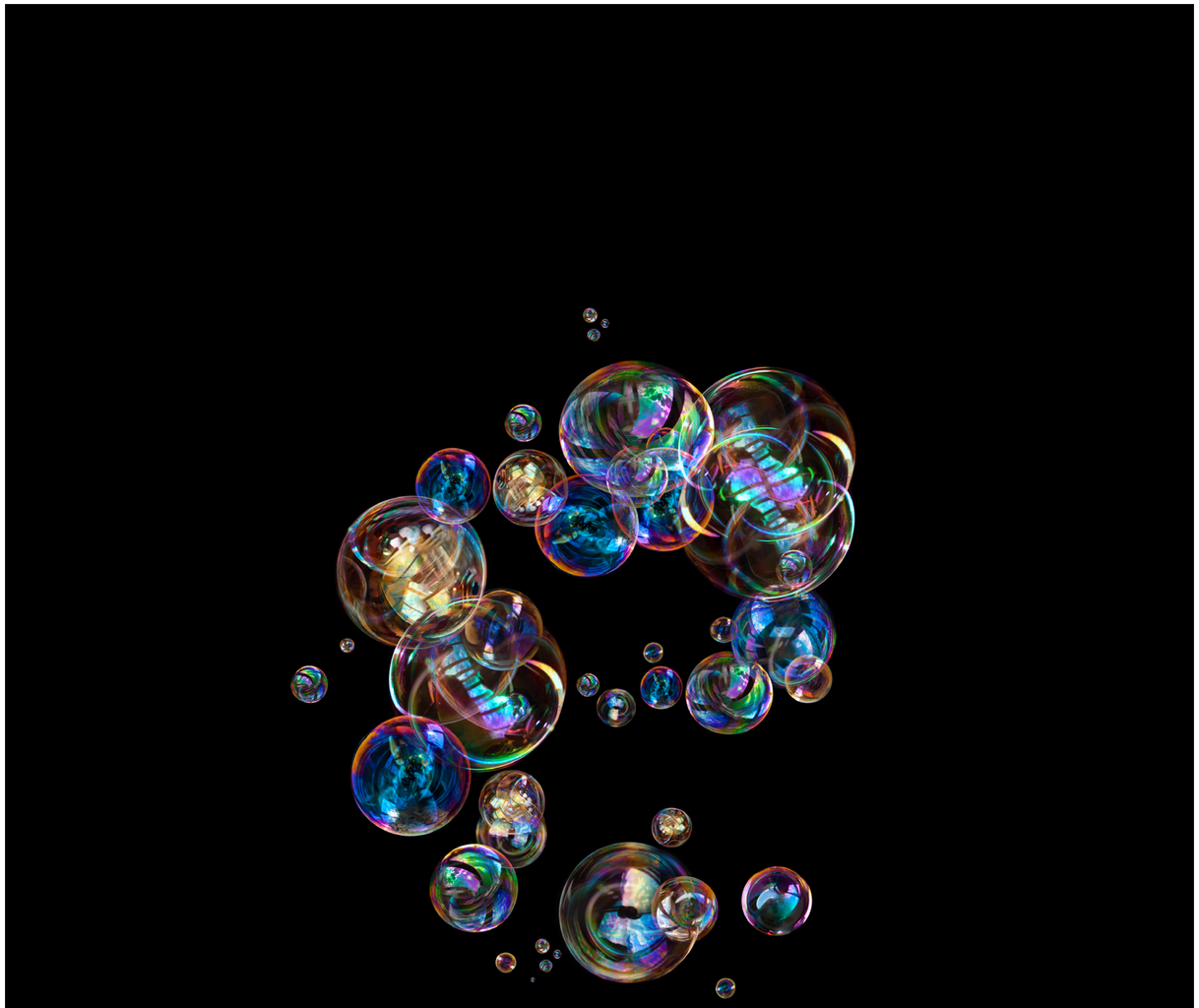
Magazin für vertrauenswürdige
Künstliche Intelligenz

Heft #2 / November 2022

Missing Link

Error 404

zvki
Zentrum für
vertrauenswürdige
Künstliche Intelligenz



WARUM BRAUCHEN WIR TRANSPARENZ?

Die Forderung nach transparenteren Anwendungen der Künstlichen Intelligenz (KI) ist mit der Erwartungshaltung verbunden, komplexe Software nachvollziehbar und überprüfbar zu machen. Ein Konsens darüber, was Transparenz in diesem Zusammenhang genau ausmacht und wie wir sie herstellen können, fehlt bislang. Doch die Auseinandersetzung mit möglichen Antworten hat längst begonnen und zeigt sich in Debatten über Informationsangebote und Kompetenzen, über Standards und Zertifikate sowie über Ansätze, um komplexe KI-Modelle zu erklären.

Vor allem Politiker*innen und zivilgesellschaftliche Organisationen, aber auch Wissenschaftler*innen fordern mehr Transparenz im Zusammenhang mit KI-Systemen und beschäftigen sich mit geeigneten Maßnahmen. Sie sehen darin eine Möglichkeit, Fehler oder Verstöße gegen Gesetze wie die Grundrechte zu erkennen und ihnen zu begegnen. Das kann im Ergebnis vertrauenswürdigere KI-Anwendungen bedeuten – so die Annahme. Mit Transparenzmaßnahmen ist ebenfalls die Hoffnung verbunden, dass sie Bürger*innen zu notwendigen Informationen verhelfen und sie auf Grundlage dessen mündigere Entscheidungen treffen können.

Die Notwendigkeit von Transparenz drängt sich in ähnlicher Weise auf wie die damit verbundenen Hoffnungen und Erwartungen: Seit einigen Jahren verbreiten sich KI-Anwendungen, die mithilfe von maschinellem Lernen funktionieren, in zahlreichen Bereichen unseres Lebens, Arbeitens und Wirtschaftens – Tendenz steigend.

Dabei kommen häufig komplexe KI-Anwendungen zum Einsatz, die weder für ihre Nutzer*innen oder Betroffene noch für ihre Entwickler*innen nachvollziehbar sind. Wie sie genau vorgehen, bleibt unklar. Klar ist indes, dass die Modelle häufig nicht fehlerfrei funktionieren und es im Prozess der Entwicklung zu Verzerrungen kommt, die folgenreich sein können. Das zeigt sich an Beispielen, bei denen Nutzer*innen die Fehler am Ergebnis feststellen konnten: Ein Bewerbungsfilter bevorzugte Männer.[1] Eine Gesichtserkennungssoftware erkannte das Gesicht einer Schwarzen Frau nur dann, wenn sie sich eine weiße Maske aufsetzte.[2] Doch nicht immer gibt das Ergebnis zu erkennen, dass das zugrunde liegende KI-Modell beispielsweise auf einer unzureichenden Datengrundlage basiert und mit folgenreichen Fehlschlüssen oder Verzerrungen arbeitet. Dadurch bleiben Fehler und unfaire Vorgehensweisen unentdeckt. Deshalb arbeiten Forscher*innen an Ansätzen der erklärbaren KI. Sie sollen dazu beitragen, die Vorgehensweisen von bislang häufig eingesetzten intransparenten Modellen besser verstehen zu können.

Mitmachen

Das *ZVKI* versteht sich als neutrale Schnittstelle zwischen Disziplinen und Akteur*innen, zwischen Nutzer*innen und Expert*innen. Treten Sie mit uns in Kontakt und in den Austausch. Sie erreichen uns per E-Mail unter *zvki@irights-lab.de*.

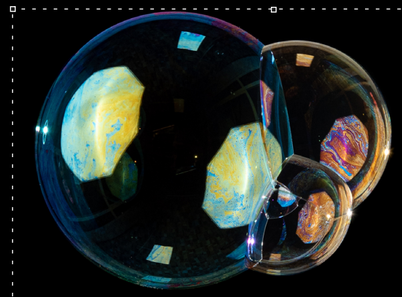
Mehr über unsere Aktivitäten und Themen erfahren Sie auf unserer Webseite www.zvki.de.

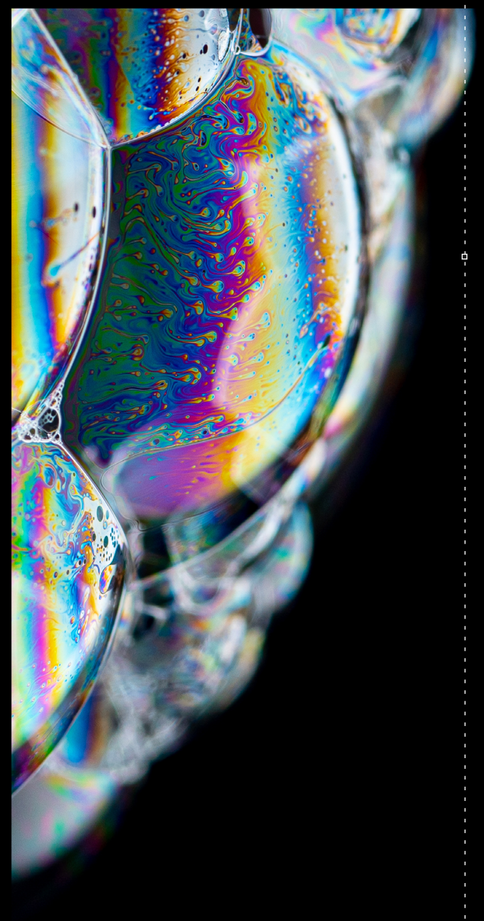
In den Debatten um vertrauenswürdige KI scheint das Ziel eindeutig zu sein: Wir brauchen mehr Transparenz. Jedoch fehlt bislang ein breiter Konsens darüber, was Transparenz genau bedeutet und ausmacht. Ein konkreteres Bild ergibt sich trotzdem, wenn wir Einzeldefinitionen, vage Annäherungen und die mit den Begriffen Transparenz und Nachvollziehbarkeit verbundenen Forderungen genauer in den Blick nehmen. Dazu zählt auch, sich mit den Ansätzen zu befassen, die gegenwärtig als geeignete Maßnahmen gelten, um mehr Transparenz zu schaffen. Hierzu gehören Prüfungen und Zertifizierungen, Erklärungsmethoden komplexer KI-Modelle und das Bereitstellen von Informationen für Bürger*innen. Vorschläge für solche Maßnahmen beziehen sich auf verschiedene Prozessschritte der Entwicklung, auf Einsatzkontexte, Verfahrensweisen und Systemeigenschaften.

Da diese Darstellungen unzureichende Verkürzungen sind, widmen wir der Transparenz ein ganzes Magazin. Darin stellen wir die Begriffe Transparenz, Nachvollziehbarkeit und Erklärbarkeit vor. Wir fragen uns, ob sich Bürger*innen tatsächlich mehr Transparenzmaßnahmen wünschen und welche das sind. Außerdem setzen wir uns mit vielversprechenden Maßnahmen in diesem Zusammenhang auseinander. Wir zeigen auf, wie weit sie bereits gediehen sind und welche Fragen bislang offenbleiben.

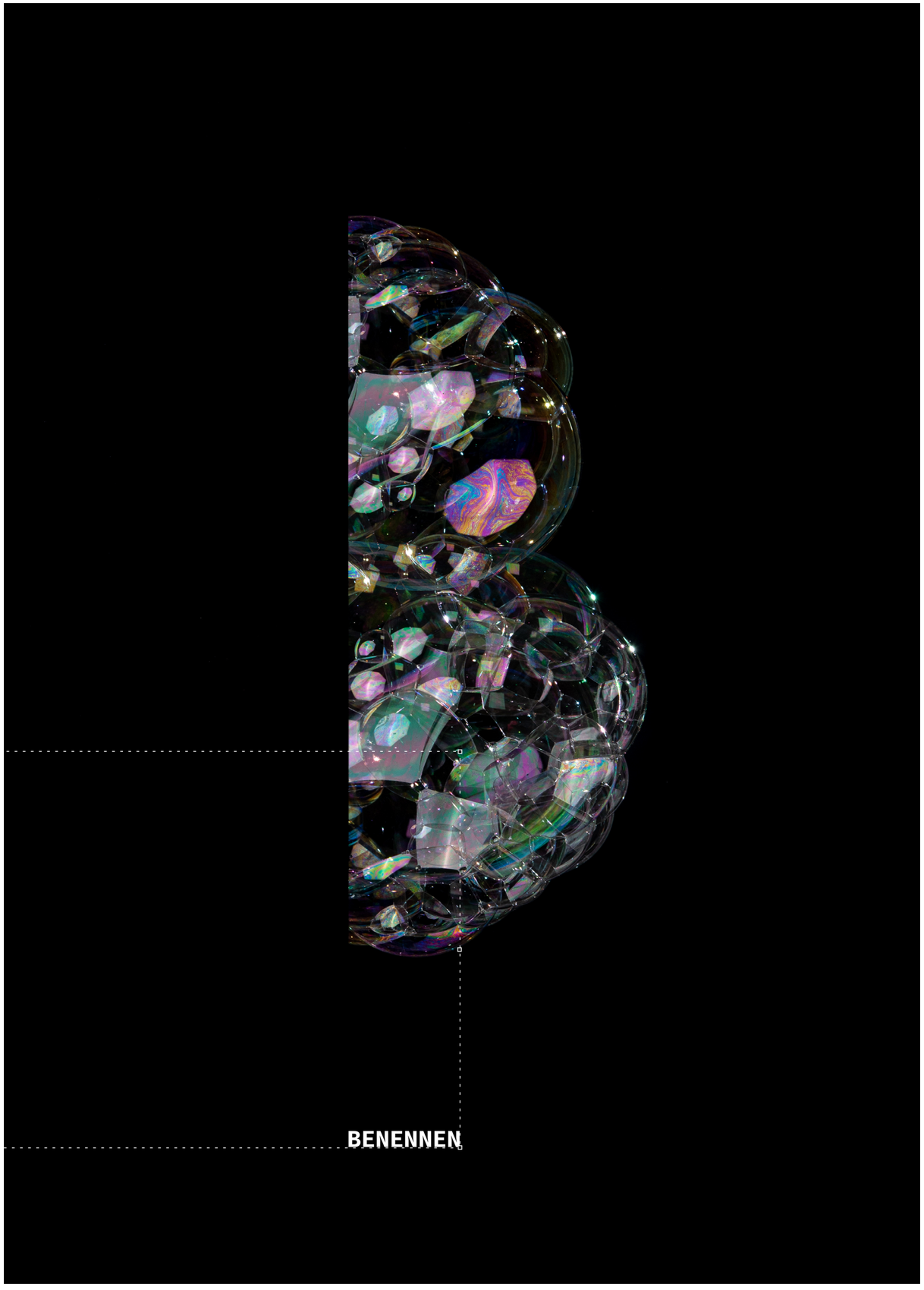
		Autorin
	Jaana Müller-Brehm	

1 Knobloch/ Hustedt, S. 15.
2 Buolamwini, o. S.





INHALT



BENENNEN

WAS IST TRANSPARENZ?

Das meinen wir, wenn wir von Transparenz, Nachvollziehbarkeit und Erklärbarkeit sprechen.

Ethische Richtlinien und Empfehlungen zur Gestaltung von KI-Anwendungen nennen Transparenz häufig als eine zentrale Voraussetzung für vertrauenswürdige Künstliche Intelligenz. Beispiele dafür sind der „KI-Prüfkatalog“ des *Fraunhofer-Instituts für Intelligente Analyse- und Informationssysteme IAIS* [1] von 2021, die Ethikleitlinien der *High-Level Expert Group on Artificial Intelligence (AI HLEG)*[2] der *Europäischen Kommission* von 2019 oder die Empfehlungen zur Ethik Künstlicher Intelligenz der *UNESCO*[3] von 2021. Diese Richtlinien betonen unterschiedliche Aspekte, die Transparenz auszeichnen. Transparenz zeigt sich als komplexes Konstrukt, weshalb eine einfache Definition des Begriffs nicht möglich ist.[4] Wir können uns dem Begriff lediglich annähern.

Im deutschen Sprachraum verwenden viele Autor*innen die Begriffe ‚Transparenz‘ und ‚Nachvollziehbarkeit‘ synonym.[5] Die AI HLEG der Europäischen Kommission führt ‚traceability‘ im Sinne von Nachvollziehbarkeit beziehungsweise Rückverfolgbarkeit als Komponente von Transparenz an. Demnach sollen die verwendeten Daten und „Prozesse, die zu der Entscheidung des KI-Systems geführt haben, […] so gut wie möglich dokumentiert werden.“[6]

Was zeichnet ein transparentes KI-System aus?

Zum einen bedeuten Transparenz und Nachvollziehbarkeit, dass Personen wissen, wann sie mit einer KI-Anwendung interagieren.[7] Wir erweitern dieses Verständnis, indem wir auch diejenigen Menschen einbeziehen, die von den Ergebnissen einer KI-Anwendung betroffen sind. Zum anderen beschreibt Transparenz, dass die Funktionsweise eines KI-Systems selbst nachvollziehbar ist.[8]

Insbesondere diese „funktionale“ Transparenz[9] können Entwickler*innen von KI-Systemen oft nicht sicherstellen: Bei vielen KI-Modellen liegt die Vorgehensweise, die zu einem Ergebnis führt, im Verborgenen. Der Forschungsbereich zu erklärbarer KI widmet sich Methoden, die die Ergebnisse solcher sogenannten Blackbox-Modelle nachvollziehbar machen sollen.

FAQ: Wie erkenne ich, ob ich es mit einer KI-Anwendung zu tun habe?

Bisher gibt es keine Kennzeichnungspflicht für den Einsatz von KI-Methoden. Bei freiwilligen Hinweisen geben die Hersteller*innen in der Regel nur grundsätzlich an, ob es sich um eine KI-Anwendung handelt oder nicht. Die konkrete Funktionsweise des KI-Systems behalten die Unternehmen für sich. Das macht Künstliche Intelligenz oftmals undurchsichtig und unverständlich – zu einer sogenannten Blackbox.

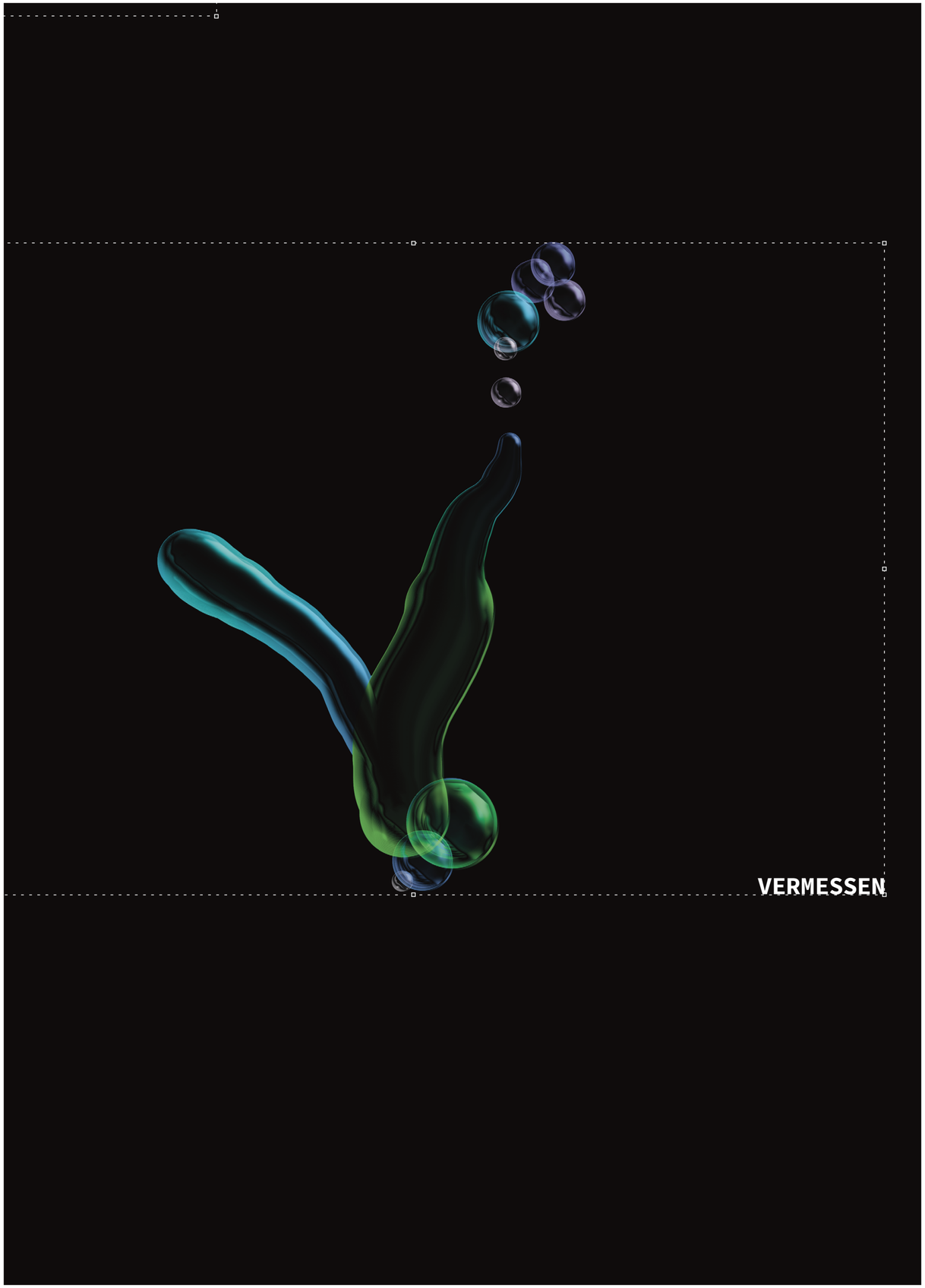
An diesem Zustand möchte die *Europäische Kommission* etwas ändern. In einer geplanten KI-Verordnung nennt sie eine Transparenzpflicht. Demnach sollen KI-Systeme, die mit Menschen interagieren, als solche erkennbar sein, sofern das nicht offensichtlich ist. Aktuell (Stand: September 2022) befindet sich das Gesetzesvorhaben im Entwurfsstadium.

Weitere Antworten auf häufig gestellte Fragen gibt es hier:

www.zvki.de > KI-Navigator > Unsere Inhalte > FAQ.

Verweis	Seite	
Mehr zum Begriff der Transparenz finden Sie in der Rubrik VERSTEHEN unter „Führt Transparenz zu Kompetenz?“.	13	
Weitere Informationen zum Thema erklärbare KI finden Sie in der Rubrik VERSTEHEN unter „Trägt erklärbare KI zu mehr Nachvollziehbarkeit bei?“ und im Interview mit Wojciech Samek in der Rubrik NACHFRAGEN unter „Mit welchem Ziel wollen wir erklären?“.	20 und 24	
	Autorin	Franziska Busse

1 Poretschkin et al.
2 High-Level Expert Group on AI.
3 UNESCO.
4 Felzmann et. al, S. 3335.
5 z. B. Poretschkin et. al; Schaaf/ Wiedenroth/ Wagner; Kraus et al.; Waltl.
6 High-Level Expert Group on AI, S. 22.
7 Rohde et al., S. 36.
8 Walmsley, S. 589; Poretschkin et al., S. 63.
9 Walmsley, S. 589.



WELCHE INFORMATIONEN FEHLEN?

Wenn Informationen zu mehr Transparenz beitragen sollen, müssen sie Verbraucher*innen, Nutzer*innen und Betroffene erreichen und gleichzeitig zu neuen Erkenntnissen führen. Einblicke darin, ob das gegenwärtig der Fall ist, geben repräsentative Umfragen. Sie zeigen, welche Informationen sich die Bürger*innen in Deutschland wünschen und welche Transparenzmaßnahmen sie als sinnvoll erachten.

Algorithmen und KI sind für zunehmend mehr Menschen in Deutschland geläufige Begriffe. Das zeigen repräsentative Befragungen wie „Was Deutschland über Algorithmen und Künstliche Intelligenz weiß und denkt“ der *Bertelsmann Stiftung* aus dem Jahr 2022. Die Befragten geben im Vergleich zu den letzten Jahren vermehrt an, in etwa zu wissen, was sich hinter diesen Begriffen verbirgt.[1] Bei Fragen nach der Akzeptanz von KI und dem Vertrauen in KI zeigen sich aber Unsicherheiten: Ein Drittel der Befragten kann nicht genau sagen, ob sie der Technologie in verschiedenen Anwendungskontexten vertraut oder misstraut.[2] Ein möglicher Grund für solche Antworten ist der Wissensstand zu KI in der Bevölkerung, denn den meisten Menschen fehlt ein präzises Verständnis davon, wie KI-Systeme funktionieren.

Wie viel wissen die Bürger*innen in Deutschland über KI?

97 Prozent aller Befragten können laut einer repräsentativen Umfrage des *ZVKI* von 2022 eine richtige Beschreibung von KI erkennen (Abbildung 1).[3] Jedoch nur 66 Prozent wissen, dass Algorithmen wesentliche Bestandteile von KI-Anwendungen sind. Andere Studien kommen zu ähnlichen Ergebnissen: 80 Prozent der Bürger*innen glauben laut einer Befragung des *Bitkom* zumindest ungefähr zu wissen, was KI bedeutet.[4] Nur 60 Prozent fühlen sich nach der Befragung der *Bertelsmann Stiftung* in ihrem Wissen über Algorithmen einigermaßen sicher.[5]

→ Auch wenn Basiswissen mittlerweile weit verbreitet ist, ist vertieftes Wissen seltener vorhanden.

Diese Wissenslücken hängen mit der Bereitschaft zusammen, sich umfassend über KI-Technologien zu informieren, wie verschiedene Umfragen zeigen:

- Nur 22 Prozent der Befragten fühlen sich ausreichend gut über KI informiert.[6]
- Nur ein Viertel der Befragten haben ein großes Interesse an Informationen rund um KI.[7]
- Knapp 30 Prozent der Befragten sind bereit, sich aktiv über KI-Verfahren zu informieren.[8]

Verbraucher*innen, Nutzer*innen und Betroffene

Im traditionellen Verständnis sind Verbraucher*innen Personen, die Waren oder Dienstleistungen für den privaten Verbrauch erwerben. Übertragen auf KI-Anwendungen lässt dieser Begriff viele Aspekte außer Acht und riskiert dadurch, wesentliche Themen des Verbraucher*innenschutzes zu vernachlässigen – beispielsweise, dass persönliche Daten von Verbraucher*innen in die Entwicklung eines KI-Systems einfließen. An dieser Stelle sind Privatpersonen reine Datensubjekte. Verbraucher*innen können sogar von KI-basierten Ergebnissen betroffen sein, obwohl sie keine konkrete Anwendung nutzen. Das ist beispielsweise der Fall, wenn Unternehmen KI-Systeme in Bewerbungsprozessen einsetzen.

Das *ZVKI*-Essay „Wer sind die Verbraucher*innen im Kontext von KI-Systemen?“ diskutiert solche Fragen und formuliert Anforderungen an einen zeitgemäßen Verbraucher*innen-Begriff in Bezug auf KI.

Das Essay steht hier zur Verfügung:

www.zvki.de > ZVKI Exklusiv > Fachinformationen.

1 Overdiek/ Petersen, S. 15ff.
2 Busse/ Baeva, o. S.
3 Ebd.
4 Berg/ Dehmel, S. 2.
5 Overdiek/ Petersen, S. 17.
6 Busse/ Baeva, o. S.
7 CAIS, o. S.
8 Busse/ Baeva, o. S.

Welche Informationen zu KI interessieren Sie?

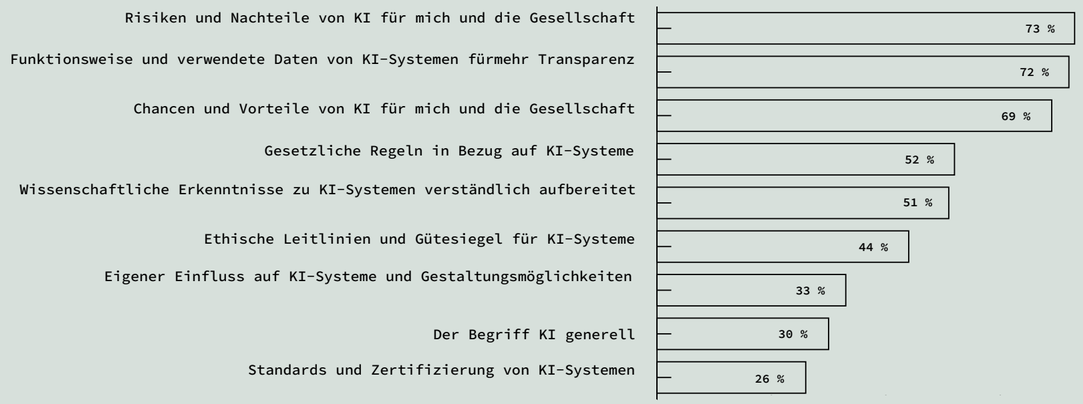


Abbildung 1	Basis: Online-Bevölkerung ab 16 Jahre, n=578	
	Frage: Wenn Sie mehr Informationen wünschen, was genau interessiert Sie denn?[11]	

Wie groß ist Ihr Vertrauen in KI-Systeme in folgenden Bereichen?

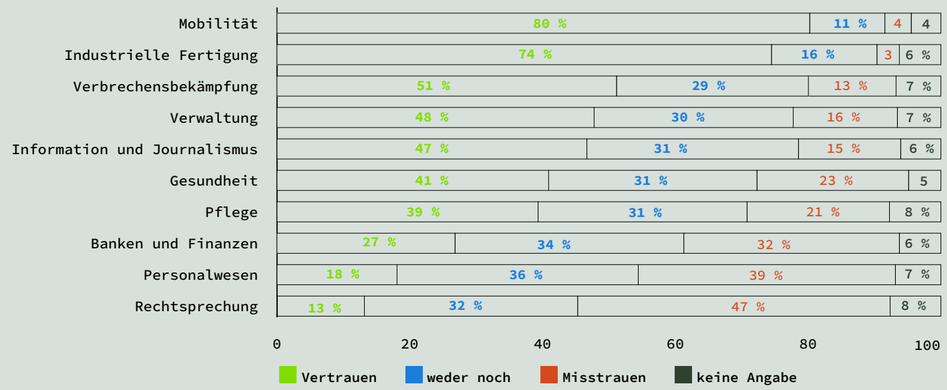


Abbildung 2	Basis: Online-Bevölkerung ab 16 Jahre, n=1.007	
		Alle Angabne in Prozent. Abweichungen von 100 Prozent sind auf Rundungsungenauigkeiten zurückzuführen.[17]

Welche Informationen wünscht sich die Bevölkerung?

Es wäre ein Fehlschluss, von einem Desinteresse an KI in der Bevölkerung auszugehen, denn 78 Prozent derjenigen, die sich schlecht informiert fühlen, wünschen sich mehr Informationen zu KI.[9] Dass vorhandene Informationsangebote viele der Bürger*innen nicht erreichen, mag an der Art und Weise liegen, wie die meisten ihre Nachrichten zu KI beziehen wollen – nebenbei und am liebsten aus journalistischen Quellen.[10]

Die Befragten haben klare Vorstellungen davon, welche Informationen sie in Bezug auf KI-Systeme erhalten möchten. Sie wünschen sich verständlich aufbereitete Berichte, mit denen sie die KI-Anwendungen nachvollziehen können und die ihnen helfen, Risiken und Chancen einzuordnen.[11] Eines der wichtigsten Themen für Bürger*innen ist Transparenz über die Funktionsweise und die verwendeten Daten bei KI-Systemen.

Wie viel Transparenz und Nachvollziehbarkeit wollen Bürger*innen?

Transparenz und Nachvollziehbarkeit sind aus Sicht der Teilnehmer*innen verschiedener Umfragen zentrale Anforderungen an KI:

- Über 80 Prozent der Befragten erachten eine klare Kennzeichnung von KI-Anwendungen sowie eine unabhängige Zertifizierung als wichtig.[12]
- Eine verpflichtende Zertifizierung wird von 60 Prozent der Befragten als eine wirksame Maßnahme eingeschätzt, um gegen Diskriminierung beim Einsatz von KI vorzugehen.[13]
- Verständliche Erklärungen, warum KI zu bestimmten Ergebnissen kommt, sind für 18- bis 30-Jährige europaweit ein zentrales Kriterium für deren Akzeptanz. Noch wichtiger sind jedoch Beschwerdestellen und menschliche Aufsicht.[14]

→ Die Teilnehmer*innen repräsentativer Umfragen möchten KI-basierte Ergebnisse besser verstehen können und identifizieren Kennzeichnungen, Zertifikate, Erklärungen sowie eine externe Aufsicht als geeignete Maßnahmen.

***ZVKI*-Befragung: Was sind die richtigen Zutaten für vertrauenswürdige Künstliche Intelligenz?**

Mithilfe einer repräsentativen Online-Befragung untersuchte das *ZVKI* im Mai 2022, was die deutsche Bevölkerung über KI-Systeme weiß, wie viel Vertrauen sie KI-Anwendungen entgegenbringt und welche Maßnahmen der Gestaltung sie fordert.

Die zentralen Erkenntnisse gibt es hier:

www.zvki.de > ZVKI Exklusiv > Fachinformationen.

Welche Zusammenhänge bestehen zwischen Wissen und Vertrauen?

Das Vertrauen in KI und die Akzeptanz von KI-Systemen unterscheiden sich je nach Bereich, in dem KI-Modelle zum Einsatz kommen. So vertrauen die Befragten KI-Anwendungen in den Bereichen industrielle Fertigung und Mobilität sowie im Gesundheitssektor grundsätzlich mehr als in der Rechtsprechung oder im Personalwesen.[15] Die Akzeptanz von KI-Anwendungen ist in den Bereichen Verkehr und industrielle Produktion entsprechend größer als bei Gericht oder bei politischen Entscheidungen.[16]

Diese Ergebnisse können darauf hindeuten, dass das Vertrauen in KI mit dem vorhandenen Wissen und somit der Nachvollziehbarkeit von KI-Systemen zusammenhängt, da Themen wie autonomes Fahren medial präsenter sind als KI-Anwendungen in anderen Bereichen. Auf einen Zusammenhang von Akzeptanz und Wissen deuten auch die Befragung der *Bertelsmann Stiftung* von 2022[18] sowie der „Bosch KI-Zukunftskompass 2020“ des Unternehmens *Bosch*[19] hin.

Es zeigt sich ein Zusammenhang zwischen den Einsatzbereichen von KI-Systemen und der Skepsis gegenüber KI-Einsätzen: So sind die Befragten misstrauischer, wenn KI-Systeme in Kontexten mit hohen Risiken zum Einsatz kommen. Für Bereiche, die mit mehr und schwerwiegenden Risiken für die Betroffenen verbunden sind, wünschen sich die Befragten, dass Menschen und nicht KI-Anwendungen entscheiden.[20] Unklar ist jedoch, welche Rolle das Wissen über KI-Systeme, die damit verbundenen Risiken und geeignete Maßnahmen dabei spielt, wie vertrauenswürdig KI-Systeme erscheinen. Wissenschaftler*innen diskutieren diese bislang kaum erforschte Frage derzeit kontrovers.[21]

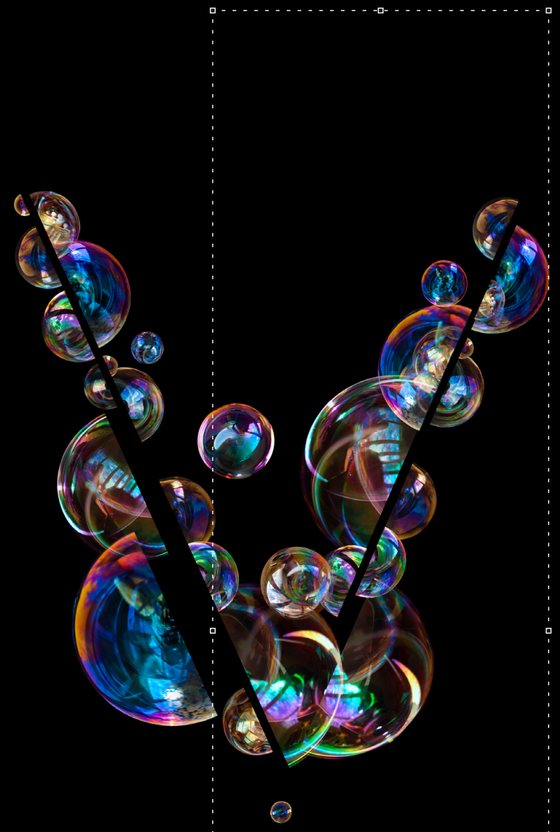
→ Verständlich aufbereitete Informationsangebote über KI-Systeme können eine Grundlage dafür sein, dass Vertrauen entsteht. Zu mehr Transparenz gehören aus Sicht der Bürger*innen jedoch nicht nur Erklärungen zur Funktionsweise, sondern auch verständliche Informationen über die damit verbundenen Risiken. Die Bürger*innen wünschen sich zudem Hinweise darauf, wie nachteilige Auswirkungen von KI-Anwendungen verhindert werden können – zum Beispiel durch Beschwerdestellen oder menschliche Aufsicht.

Verweis	**Seite**
Mehr über Informationsangebote für Verbraucher*innen, Nutzer*innen und Betroffene finden Sie in der Rubrik VERSTEHEN unter „Führt Transparenz zu Kompetenz?“.	13
Weitere Informationen zu Prüfungen und Zertifikaten als Transparenzmaßnahmen finden Sie in der Rubrik VERSTEHEN unter „Wie zertifizieren wir Transparenz?“.	16

Autorin
Gergana Baeva

9 Ebd.
10 Ebd.
11 Ebd.
12 Ebd.; TUV-Verband e. V.
13 Kieslich, S. 4.
14 Gagrčin et al., S. 54.

15 Busse/ Baeva, o. S.
16 CAIS, 2022, o. S.
17 Busse/ Baeva, o. S.
18 Overdiek/ Petersen, S. 23.
19 Robert Bosch GmbH, S. 7.
20 Ebd., S. 21.
21 Chazette/ Brunotte/ Speith, S. 9; Langer et al., S. 16f.



VERSTEHEN

WIE SCHAFFEN WIR TRANSPARENZ?

Transparenz sorgt für überprüfbare KI-Anwendungen, schafft Vergleichbarkeit und ermöglicht Erklärungen. Sie gleicht Informationsasymmetrien zwischen Anbieter*innen sowie Verbraucher*innen und Bürger*innen aus. Und sie trägt zu mündigen Nutzer*innen bei. Diese Erwartungen lassen die Forderung nach Transparenz sinnvoll und konsensfähig erscheinen. Bei genauerem Hinsehen ergeben sich jedoch zahlreiche Fragen, etwa, wie sich Transparenz umsetzen lässt, wer daran beteiligt sein sollte und welche Maßnahmen tatsächlich die erwartete Wirkung erzielen. Die Debatte um diese Fragen zeigt sich jünger und kontroverser als der in den digitalpolitischen Echokammern etablierte Ruf nach Transparenz.

Führt Transparenz zu Kompetenz?

Die Forderung von Transparenz ist ein Kernelement des digitalpolitischen Diskurses um KI. Es ist das am häufigsten genannte Prinzip in entsprechenden Richtlinien. Diese sollen gemeinsam mit Regulierungsvorhaben die Selbstregulierung von Unternehmen ergänzen. Doch das Konzept der Transparenz ist nicht frei von Kritik und auf weitere Maßnahmen angewiesen, damit Verbraucher*innen, Nutzer*innen und Betroffene ihr Wissen im Alltag anwenden können.

Der Versuch, Transparenz herzustellen, indem möglichst viele Informationen angeboten werden, führt nicht zwingend zu gut aufgeklärten Nutzer*innen. Vielmehr gilt es, die Frage zu berücksichtigen, welche Informationsangebote für eine bestimmte Zielgruppe hilfreich sind. So können sich beispielsweise textbasierte Informationen negativ auswirken, wenn sie die Leser*innen durch ihre Menge, Tiefe und fehlende Vergleichbarkeit überfordern (dann handelt es sich um einen sogenannten „information overload“).[1] Schwer verständliche oder gar unverständliche Informationen, die Anbieter*innen zur Verfügung stellen, können zum Beispiel zur Folge haben, dass Verbraucher*innen vermuten, die Informationen seien absichtlich verwirrend gestaltet. Dies könnte den Anbieter*innen einen Vorteil verschaffen.[2]

Woher stammt das Konzept der Transparenz?

Das Wort ‚Transparenz‘ verweist zusammen mit Begriffen wie ‚Klarheit‘ auf das Konzept ‚Wissen‘. In unserem Sprachgebrauch leiten wir von „etwas sehen“ eine Erkenntnis und damit Wissen ab. Dies hat eine positive Bedeutung. Im Gegensatz dazu stehen Redewendungen wie „im Dunkeln tappen“, was sich im Diskurs um KI-Verfahren zum Beispiel in der Metapher der „Blackbox“ widerspiegelt.[3] Transparenz als Konzept ist zudem mit Informationsasymmetrien verbunden, bei denen eine Seite mehr Informationen als die andere besitzt.[4] Informationen zu erhalten, ist also konzeptuell mit der Konsequenz verknüpft, Wissen über etwas zu erlangen.

Transparenz gewann im gesellschaftlich-politischen Raum seit den 1990er-Jahren an Bedeutung. Damals wurde das Thema vorwiegend von zivilgesellschaftlichen Organisationen sowie supranationalen Institutionen besetzt und bezog sich auf Anti-Korruptionsbemühungen. Mitte der 1990er- sowie in den 2000er-Jahren wurde Transparenz im Hinblick auf die Finanzkrisen und -skandale dieser Zeit gefordert und mündete in Gesetzgebungen wie der „Transparency Directive“ der *Europäischen Union (EU)* aus dem Jahr 2004. Seit einigen Jahren wird Transparenz vermehrt in Bezug auf KI-Verfahren erforscht sowie in gesellschaftlichen und politischen Debatten gefordert.[5] Dabei ist es das meistgenannte Prinzip in ethischen Richtlinien zu KI – 73 von 84 in einer Studie untersuchten Quellen nennen Transparenz als Kriterium. 88 Prozent dieser Richtlinien wurden seit 2016 veröffentlicht.[6]

1 Sachverständigenrat für Verbraucherfragen, S. 400.
2 Oehler, S. 261.
3 Larsson/ Heintz, S 6.
4 Ebd., S. 4.
5 Heintz/ Larsson, S. 4.
6 Jobin et al., S. 3.

Was hat Transparenz mit „Ethics Washing“ zu tun?

Einige Akteur*innen aus Wirtschaft, Wissenschaft und Zivilgesellschaft betrachten die steigende Anzahl ethischer Richtlinien kritisch. Sie werfen vor allem Unternehmen vor, in diesem Zusammenhang scheinheilig zu handeln, da sie ethische Richtlinien nach außen kommunizieren, strukturelle wertegetriebene Veränderungen jedoch ausbleiben.

Dass diese Richtlinien zudem unter dem Stichwort „KI-Ethik“ statt „KI-Politik“ diskutiert werden, ist laut der Juristin Elettra Bietti symptomatisch dafür, dass Unternehmen gesellschaftspolitische Debatten gestalten. Das geschieht zum Beispiel auch dadurch, dass sie Konferenzen oder wissenschaftliche Institute finanziell fördern.[7] Weitere Kritik bezieht sich zudem auf Praktiken, die die Verantwortung für Risiken von Produkten auf deren Nutzer*innen abwälzen. Die Hersteller*innen unterlassen es, das Design der Apps anzupassen. Stattdessen erhalten die Nutzer*innen beispielsweise Benachrichtigungen, ihre Nutzungsdauer von Apps einzuschränken. Laut Bietti adressieren solche Vorgehensweisen die Risiken der Produkte nicht tiefergehend.[8] Ähnliches gilt bei Transparenzmaßnahmen, wenn Nutzer*innen Informationen erhalten, aber damit auch Verantwortung übertragen bekommen und weitere unterstützende Maßnahmen ausbleiben. So kommt es zum sogenannten „Ethics Washing“, bei dem der Fokus vor allem auf der Selbstregulierung des Marktes liegt.

FAQ: Kann ich verlangen, dass mir erklärt wird, auf welcher Grundlage eine KI eine Entscheidung getroffen hat?

Momentan fehlt eine gesetzliche Regelung, die ein solches Informationsrecht ausdrücklich vorsieht. Deshalb können Nutzer*innen die gewünschte Erklärung von privaten Unternehmen nicht verlangen. Wie das KI-System genau funktioniert, kann ein Betriebsgeheimnis sein, das rechtlich geschützt ist. Gleiches gilt für die Trainingsdaten.

Anders ist es bei Behörden. Sie sind unmittelbar an das Grundgesetz gebunden. Das regelt das Rechtsstaatsprinzip in Artikel 20 Absatz 3 des Grundgesetzes, zu dem auch das Transparenzgebot gehört. Danach müssen staatliche Entscheidungen nachvollziehbar sein, um das Handeln der Behörde (gerichtlich) kontrollieren zu können. Das ist wichtig, damit die (Grund-) Rechte der Bürger*innen geschützt beziehungsweise durchgesetzt werden können. Eine Behörde darf sich also nicht dahinter verstecken, dass eine Entscheidung von einem KI-System getroffen wurde. Sie muss offenlegen, wie das KI-System zu der Entscheidung gekommen ist und sie begründen.

Weitere Antworten auf häufig gestellte Fragen gibt es hier:

www.zvki.de > KI-Navigator > Unsere Inhalte > FAQ.

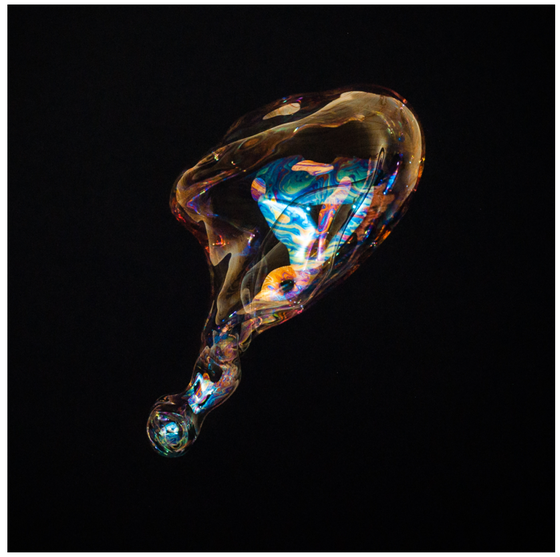
Welche Akteur*innen fordern Transparenz?

Zivilgesellschaftliche Organisationen auf nationaler und internationaler Ebene beschäftigen sich mit Möglichkeiten, KI-Verfahren transparenter zu gestalten, etwa *Mozilla.*[9] Darüber hinaus setzen sich Forschungseinrichtungen für transparente KI-Anwendungen ein wie das *Deutsche Forschungszentrum für Künstliche Intelligenz (DFKI)*. Es forderte bereits 2017 in einer gemeinsamen Publikation mit der *Bitkom*, dass „Algorithmen entwickelt werden [müssen], die das Vertrauen in das System verbessern und zur Transparenz beitragen.“[10] Das internationale Netzwerk *Global Network of Internet and Society Researchers (NoC)* fordert Transparenz von Algorithmen bei der automatisierten Moderation von Inhalten auf Plattformen.[11] Auch unter Politiker*innen gibt es schon länger Stimmen, die vor allem von großen Tech-Unternehmen mehr Transparenz fordern.[12] Vorschläge für eine tatsächliche Gesetzgebung wurden aber erst in den letzten Jahren konkreter.

Die „Ethic Guidelines for Trustworthy AI“ der *AI HLEG* der *Europäischen Kommission* von 2019 beschreiben Transparenz als eines der Kriterien für vertrauenswürdige KI. Es besteht aus drei Elementen: Nachvollziehbarkeit beziehungsweise Rückverfolgbarkeit (unter anderem durch Dokumentationen zu Datensätzen und Training des Systems), Erklärbarkeit (zum Beispiel die Möglichkeit, den Entscheidungsprozess des Systems darzulegen) und Kommunikation (wie Hinweise an Nutzer*innen zum Einsatz von KI-Systemen).[13] Der Entwurf des „Artificial Intelligence Act“ der *EU* schlägt Transparenzpflichten für Systeme vor, 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‘)“.[14] Je nach Risikostufe des KI-Systems ändern sich die Transparenzpflichten.[15] Auch die KI-Strategie der *Bundesregierung* sieht in der fortgeschriebenen Version von Dezember 2020 ein „risikoadäquates Maß an Transparenz und Nachvollziehbarkeit“[16] als eine der Rahmenbedingungen für „sichere und vertrauenswürdige KI-Anwendungen“[17] vor.

Welche Formen von Transparenz fördern Kompetenz?

Der *Sachverständigenrat für Verbraucherfragen (SVRV)* sieht Lücken in der digitalen Verbraucher*innenkompetenz und fordert daher in seinem letzten Gutachten, digitalpolitische Regulierungen wie die KI-Verordnung durch weitere Schritte zu flankieren. Hierzu zählen Maßnahmen der Verbraucher*innenbildung sowie Maßnahmen, die den wahrgenommenen Aufwand der Informationsbeschaffung für Verbraucher*innen reduzieren.[18] Dabei könnte die sogenannte ‚Meta-Bildung‘ helfen. Diese hat das Ziel, dass Verbraucher*innen, Nutzer*innen und Betroffene lernen, die passende Expertise zu finden anstatt selbst Expert*innen zu sein. Ein Beispiel für solche Expertisen sind zielgerichtete Angebote vermittelnder Einrichtungen wie der *Verbraucherzentralen*. Der Verbraucher*innenwissenschaftler Andreas Oehler beschreibt in diesem Zusammenhang, dass die Grenzen der Individualisierung erreicht sind. Wir können demnach nicht davon ausgehen, dass alle Menschen in allen Lebensbereichen stets kompetent sind.[19] Angebote von Informationslotsen müssen deshalb bei Bedarf die Bürger*innen unterstützen. Hilfreich sind sie vor allem dann, wenn sie auf alltägliche Probleme und Entscheidungen zugeschnitten sind oder am „Ort des Problems und der Betroffenheit“[20] zur Verfügung stehen.[21]



7 Bietti, S. 216.
8 Ebd.
9 Ricks et al., o. S.
10 Bitkom e. V./ Deutsches Forschungszentrum für Künstliche Intelligenz, S. 70.
11 Timm, o. S.
12 Schultz, o. S.

13 High-Level Expert Group on Artificial Intelligence, S. 18.
14 Europäische Kommission, S. 17.
15 Ebd.
16 Bundesregierung, S. 25.
17 Ebd.
18 Sachverständigenrat für Verbraucherfragen, S. 10.
19 Oehler, S. 269f.
20 Ebd., S. 270.
21 Ebd., S. 269f.

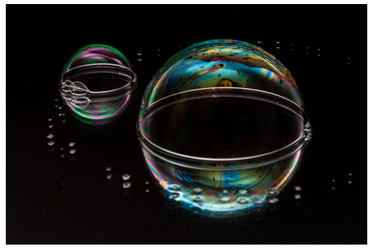
Verweis	Seite
Weitere Informationen zum Thema erklärbare KI finden Sie in der Rubrik VERSTEHEN unter „Trägt erklärbare KI zu mehr Nachvollziehbarkeit bei?“ und im Interview mit Wojciech Samek in der Rubrik NACHFRAGEN unter „Mit welchem Ziel wollen wir erklären?“.	20 und 24

Lisa Schmechel **Autorin**

Wie zertifizieren wir Transparenz?

Zertifikate gelten als eine Maßnahme, um einen gewissen Grad der Transparenz sicherzustellen. Dafür muss es jedoch klare Kriterien und Verfahren geben. Hierbei können Standards helfen. Doch wie müssen sie gestaltet sein, welcher Mehrwert entsteht dabei für Verbraucher*innen, Nutzer*innen und Betroffene und wie ist es möglich, KI-Systeme hinsichtlich Transparenzanforderungen unabhängig zu prüfen?

Die Prüfung von KI-Systemen gilt als eine Maßnahme, um rechtskonforme und ethisch akzeptable KI-Anwendungen zu ermöglichen.[1] Hierbei stellen fehlende Standards und Prüfmethoden eine Hürde dar,[2] auch wenn aktuell zahlreiche Normen in Arbeit sind.[3]



Welche Standards entstehen derzeit?

Um zu bewerten, wie Anbieter*innen von KI-Anwendungen Transparenz umsetzen, betrachten aktuelle Vorschläge der Standardisierung sowohl Eigenschaften des KI-Systems an sich als auch des Entwicklungsprozesses. Prozessevaluationen haben den Vorteil, dass sie in weiten Teilen aussagekräftig bleiben, auch wenn das KI-System später weiterentwickelt wird.[4]

- Das Projekt „VDE SPEC“ des *VDE Verband der Elektrotechnik Elektronik Informationstechnik e. V.*, an dem sich verschiedene Vertreter*innen aus Wissenschaft, Wirtschaft und Zivilgesellschaft beteiligen, legt den Fokus auf die Nachvollziehbarkeit eines KI-Systems sowie auf seine nachträgliche Überprüfbarkeit.[5] Zum einen soll eine KI-Anwendung sowohl für Nutzer*innen als auch für Betroffene nachvollziehbar sein. Zum anderen sollen Anbieter*innen Daten und Modelleigenschaften verständlich und vor allem zugänglich dokumentieren. Transparenz wird dabei in Beziehung zu anderen Kriterien für vertrauenswürdige KI wie Datenschutz oder Fairness gesetzt.

- Der „KI-Prüfkatalog“ des *Fraunhofer IAIS*[6] ermöglicht es, verschiedene Aspekte von Transparenz zu bewerten.[7] Dazu zählen die Fragen, ob eine KI-Anwendung für Nutzer*innen und Experten*innen angemessen nachvollziehbar ist und ob deren Ergebnisse begründet werden können. Die Evaluation orientiert sich an möglichen Risiken: Wie transparent eine KI-Anwendung sein muss, hängt vom Kontext ab. Eine Evaluation kann dies ermitteln und zudem überprüfen.[8] Auch künftige Veränderungen des Systems sind laut des Vorschlags regelmäßig zu bewerten.

- Der Kriterienkatalog „AIC4“ des *Bundesamts für Sicherheit in der Informationstechnik (BSI)* zielt darauf ab, die Sicherheit von KI-Cloud-Dienstleistungen zu bewerten. Er fordert darüber hinaus auch erklärbare KI.[9] Je nach Kontext, potenziellen Gefahren und menschlicher Kontrolle entstehen unterschiedliche Anforderungen an die Erklärbarkeit eines KI-Systems. Die jeweiligen Wissensstände unterschiedlicher Zielgruppen sollen dabei berücksichtigt werden. Falls KI-Ergebnisse nicht erklärbar sind, gilt es, auch diese Nicht-Erklärbarkeit transparent zu machen.[10] Ändert sich die KI-Anwendung, muss diese erneut überprüft werden.[11]

- Transparenz und Nachvollziehbarkeit sind auch im Entwurf für einen Prüfungsstandard des *Instituts der Wirtschaftsprüfer in Deutschland (IDW)* wichtige Kriterien.[12] So sollen verwendete Datensätze, KI-Algorithmen und KI-Modelle nachvollziehbar dokumentiert und für Anwender*innen verständlich gemacht werden.

„Wie gestalten wir vertrauenswürdige Künstliche Intelligenz?“

Die erste Ausgabe von *Missing Link* beschäftigt sich ausführlicher mit Empfehlungen und Richtlinien politischer Institutionen zur Regulierung von KI-Anwendungen und dem „Artificial Intelligence Act“ (Seite 10 bis 21).

Ausgabe 1 von Missing Link steht hier zur Verfügung:

www.zvki.de > ZVKI Exklusiv > Fachinformationen.

1 Mangelsdorf/ Gabriel/ Weimer, S. 2.
2 Beckert, S. 20; Adler et al. S. 19.
3 Nativi/ De Nigris, S. 54.
4 Heesen/ Müller-Quad/ Wrobel, S. 41.
5 VDE (2022a), S. 9f.
6 Poretschkin et al.
7 Ebd., S. 63.
8 Ebd., S. 65ff.
9 Bundesamt für Sicherheit in der Informationstechnik (BSI).
10 Ebd., S. 41.
11 Ebd., S. 10.
12 Institut der Wirtschaftsprüfer in Deutschland (IDW).

Wie ist Transparenz zu prüfen?

Für die Bewertung einer KI-Anwendung gibt es zahlreiche Voraussetzungen, da je nach Kontext unterschiedliche Kriterien ins Gewicht fallen.[13] Im Fall einer KI-Anwendung, die beispielsweise Mediziner*innen bei der Krebsdiagnose unterstützt, ist es besonders wichtig, dass diese die Funktionsweise nachvollziehen können.[14] In anderen Kontexten – wie in der industriellen Produktion – kann Transparenz für Nutzer*innen weniger bedeutsam sein.

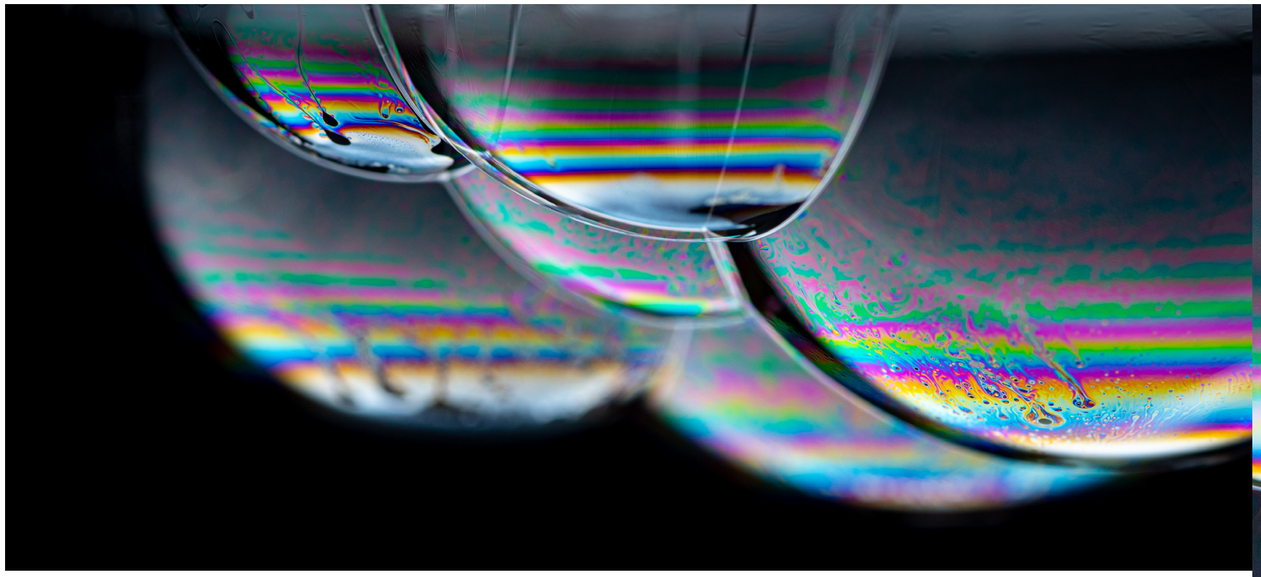
Je nach Anwendung sind zudem unterschiedliche Prüfmethoden angemessen. Die aktuellen Vorschläge sind hier unterschiedlich konkret: Lediglich der „KI-Prüfkatalog“ beschreibt und diskutiert technische Methoden der Prüfung.[15] Die anderen drei Vorschläge verzichten darauf, Prüfmethoden tiefergehend zu thematisieren.

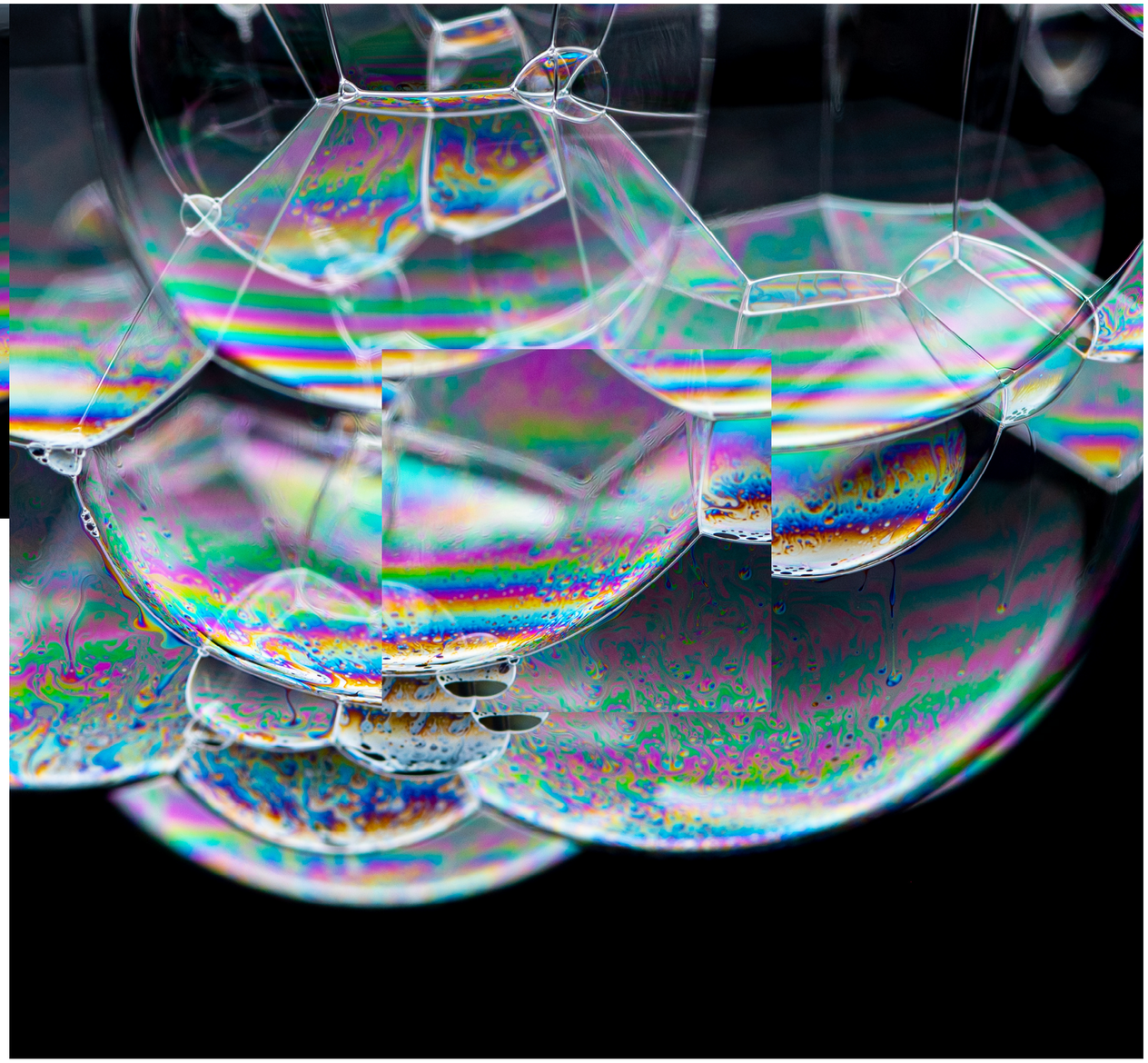
Die vorgeschlagenen Standards sehen meist eine freiwillige, oft interne Überprüfung von KI-Systemen vor. So hilft der „KI-Prüfkatalog“ vor allem dabei, umfassende technische Dokumentationen für die weiteren Prüfungen zu erstellen. Der *IDW*-Standard ermöglicht Prüfberichte durch externe Wirtschaftsprüfer*innen. „VDE SPEC“ soll interne oder externe Prüfungen anleiten, zukünftig aber um ein „AI Trust Label“ ergänzt werden.[16]

Eine Zertifizierung im Sinne einer Prüfung durch unabhängige Dritte[17] wäre auf dieser Basis möglich, ist jedoch nicht zwingend vorgesehen. Bei einer externen Prüfung ist es notwendig, dass die Prüfer*innen Zugang zu Daten und Dokumentationen haben. Eine Leitlinie von *Eticas Research and Consulting* empfiehlt, dass externe Prüfer*innen eng mit den Entwickler*innen zusammenarbeiten und die genutzten Daten und Methoden einsehen können.[18]

Dies sind erste Ansätze, um mithilfe von Standards mehr Transparenz und Nachvollziehbarkeit zu erzielen. Konkrete Anweisungen für verschiedene Anwendungskontexte fehlen jedoch. Außerdem sind Pilotprojekte notwendig, die Messverfahren für Transparenz weiter erforschen.[19]

13 Adler et al., S. 19.
14 Kraus et al., S. 43ff.
15 Poretschkin et al.
16 VDE (2022b), o. S.
17 Mock et al., S. 53, und Weiss, o. S.
18 Eticas Research and Consulting SL, S. 17f.
19 Adler et al., S. 45.





Welchen Mehrwert haben Zertifizierungen für Verbraucher*innen?

Wenn wir Transparenz herstellen und bewerten möchten, ist die Frage zentral, wer die Funktionsweise einer KI-Anwendung verstehen soll. Verschiedene Zielgruppen verfügen über unterschiedliches Wissen und benötigen dementsprechend aufbereitete Informationen. Daraus ergeben sich verschiedene Anforderungen an Transparenz. Auch wenn alle an dieser Stelle beschriebenen Standards diverse Zielgruppen im Blick haben, behandelt lediglich der „KI-Prüfkatalog“ des *Fraunhofer IAIS* Transparenzmaßnahmen für Expert*innen und Nutzer*innen umfassend und vor allem getrennt voneinander.[20]

Ob Zertifikate zu mehr Transparenz bei Verbraucher*innen, Nutzer*innen und Betroffenen beitragen können, hängt davon ab, ob sie für Lai*innen verständlich sind. Ansätze, ihnen Prüfergebnisse zu vermitteln, sind noch nicht weit entwickelt. Erste Vorschläge reichen von visuellen Labels, die dem Energieeffizienzklasse-Label ähneln – wie das „AI Ethics Label“ der *AI Ethics Impact Group*[21] –, bis hin zu kompakten Dokumentationen der Umsetzung ethischer Kriterien – wie dem „Aletheia Framework“ des Unternehmens *Rolls Royce*[22] oder dem „Explainability Statement“ der Gesundheitsanwendung *Healthily*.[23] Solche Darstellungen könnten für Zielgruppen mit geringen Vorkenntnissen jedoch schwer verständlich sein.

Zertifikate für KI-Systeme könnten zu mehr Transparenz für Verbraucher*innen, Nutzer*innen und Betroffene beitragen, wenn sie verschiedene Aspekte von Transparenz angemessen und auf unterschiedliche Zielgruppen zugeschnitten umsetzen. Das setzt jedoch etablierte Standards und Prüfverfahren voraus, die in weiten Teilen noch nicht vorhanden beziehungsweise noch nicht detailliert genug sind. Dabei sind Zertifikate einer von mehreren Bausteinen, die zu mehr Transparenz für KI-Anwendungen beitragen können.

Zertifizierung als Teil der Arbeit am *ZVKI*

Die Frage, wann KI-Systeme als vertrauenswürdig gelten können, bildet einen Schwerpunkt der Arbeit am *ZVKI*. Das Team des *ZVKI* erforscht die Grundlagen für zertifizierte vertrauenswürdige KI-Systeme und entwickelt Werkzeuge, die solche Prüfungen ermöglichen.

Zu unserer Arbeit in diesem Bereich zählt beispielsweise die Fach-AG Zertifizierung, in deren Rahmen wir uns mit Fragen der Standardisierung, Evaluation und Zertifizierung von KI-Systemen auseinandersetzen. Zudem wurden auf der *ZVKI*-Fachtagung im Juni 2022 Projekte und Initiativen vorgestellt und diskutiert, die aktuell an Prüfmethoden für eine Zertifizierung von vertrauenswürdiger KI arbeiten. Im Juli 2022 veröffentlichten wir darüber hinaus das Essay „Wie können Regulierung und Standards zu vertrauenswürdiger KI beitragen?“. Es befasst sich mit dem europäischen Ansatz einer KI-Regulierung und prüft, welchen Mehrwert solche Vorschläge für Verbraucher*innen haben. Im kommenden Jahr 2023 widmen wir uns verstärkt den Fragen der KI-Zertifizierung im Rahmen von interdisziplinären Workshops, Publikationen sowie eigenen Guidelines.

Mehr Informationen zu unseren Aktivitäten im Bereich Zertifizierung finden Sie auf unserer Webseite: www.zvki.de.

20 Poretschkin et al., S. 67ff. und 72ff.
21 Hallensleben et al.
22 Rolls Royce.
23 Healthily.

	Autorin
	Gergana Baeva

Trägt erklärbare KI zu Nachvollziehbarkeit bei?

Im Forschungsbereich zu erklärbarer KI entwickeln Wissenschaftler*innen Methoden, mit denen sich die Ergebnisse von KI-Systemen besser nachvollziehen lassen. Sollen Erklärungen zu mehr Wissen sowie zu überprüfbaren und vertrauenswürdigen Systemen beitragen, dürfen sie sich nicht auf die technische Ebene beschränken: Vielmehr müssen Entwickler*innen Erklärbarkeit im gesamten Entwicklungsprozess eines KI-Systems mitdenken und mit den Zielgruppen einer Anwendung erproben.

Erklärbare KI (Explainable AI; XAI) ist ein Forschungsbereich, in dem Wissenschaftler*innen Methoden entwickeln, mit denen sich die Ergebnisse und Vorgehensweisen von KI-Systemen nachvollziehen lassen. Die Metastudie „How to explain AI systems to end users“ von 2022 identifiziert fünf zentrale Ziele von erklärbarer KI: XAI soll das Verständnis, die Vertrauenswürdigkeit, die Transparenz, die Kontrolle und die Fairness einer KI-Anwendung erhöhen.[1] Generell gibt es zwei Möglichkeiten, die dazu beitragen, dass Menschen die Ergebnisse eines KI-Systems besser nachvollziehen können: Entwickler*innen verwenden von Beginn an sogenannte Whitebox-Modelle, die grundsätzlich eine höheres Maß an Transparenz in Bezug auf die verarbeiteten Daten und das Vorgehen eines KI-Modells aufweisen.[2] Alternativ wählen sie Blackbox-Modelle, die nachträglich erklärt werden.[3] Aktuell forschen Wissenschaftler*innen an verschiedenen Erklärmethoden (XAI-Methoden), um die Nachvollziehbarkeit von solchen Blackbox-Modellen zu erhöhen.[4]

Definition: Blackbox

KI-Systeme sind häufig so komplex, dass Menschen deren Entscheidungswege nicht mehr nachvollziehen können. Selbst Entwickler*innen können oft nicht nachvollziehen, warum das KI-System einen bestimmten Lösungsweg gewählt hat. Das trifft beispielsweise auf leistungsstarke künstliche neuronale Netze zu. Zudem weigern sich Organisationen oft, Informationen zum KI-Modell offenzulegen.

Das gilt sowohl für die Trainingsdaten und Zielvorgaben als auch für die Funktionsweise von KI-Modellen. Deshalb wissen wir etwa nicht, wie der Suchalgorithmus von *Google* genau funktioniert. Fehlende Transparenz erschwert die Kontrolle von KI-Systemen erheblich, beispielsweise um eine (un-)beabsichtigte Verzerrung (bias) aufzudecken oder um robuste KI-Modelle zu entwickeln. Transparenz und Nachvollziehbarkeit sind daher wichtige Bestandteile vertrauenswürdiger KI, die zum Beispiel mithilfe von Methoden erklärbarer KI (Explainable AI) erreicht werden können.

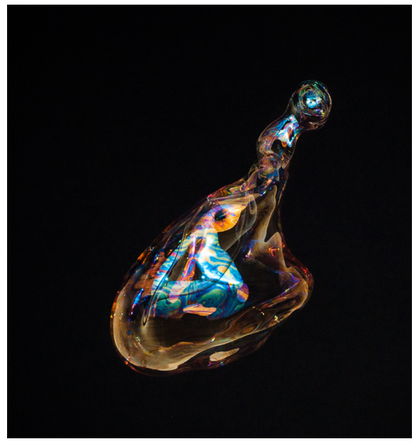
Weitere Begriffsdefinitionen rund um das Thema vertrauenswürdige KI gibt es auf unserer Webseite:

www.zvki.de > KI-Navigator > Unsere Inhalte > Glossar.

An wen richten sich XAI-Methoden?

XAI-Methoden richten sich derzeit vor allem an die Entwickler*innen von KI-Anwendungen. Die Modelle unterstützen sie zum Beispiel bei der Fehlersuche oder helfen ihnen dabei, ein KI-System zu verbessern.[5] Auch für Fachexpert*innen, die in der Praxis mit einer KI-Anwendung arbeiten, sind XAI-Methoden laut einer Umfrage des *Instituts für Innovation und Technik (iit)* von 2022 relevant. Dazu zählen etwa Mediziner*innen, die von KI-Modellen dabei unterstützt werden, Krebszellen zu erkennen. Gleichzeitig sehen die Befragten XAI-Methoden in der nahen Zukunft auch als Möglichkeit, um Endnutzer*innen und Endkund*innen sowie Patient*innen über die Funktionsweise eines KI-Systems aufzuklären.[6] Solche Erklärungen scheinen besonders wichtig für Personen zu sein, die von den Ergebnissen eines KI-Systems betroffen sind, die aber auf dessen Verwendung keinen Einfluss haben. Dies sind zum Beispiel Bewerber*innen um eine Stelle oder Anwärter*innen auf einen Kredit.

1 Laato et al., S. 10.
2 Kraus et al., S. 40.
3 Schaaf, Wiedenroth & Wagner, S. 15.
4 Für eine Übersicht siehe Holzinger et al.
5 Asghari et al., S. 12.
6 Kraus et al., S. 38.



Für Entwickler*innen existieren bereits etablierte XAI-Methoden, die jedoch technisches Wissen voraussetzen.[7] Auch für Fachleute, die in der Regel über kein IT-Wissen verfügen, gibt es erste verständliche Erklärmethoden.[8] Für Verbraucher*innen und Bürger*innen sind XAI-Methoden in der jetzigen Form unzugänglich und unverständlich. Damit Erklärmethoden den betroffenen Personen einen Mehrwert bieten, müssten sie speziell zu diesem Zweck aufbereitet werden.

Welche Grenzen haben XAI-Methoden?

Eine Herausforderung beim Einsatz von XAI-Methoden ist, dass Erklärungen entweder zu komplex und damit schwer interpretierbar oder zu einfach und dadurch irreführend sein können.[9] Bei einigen XAI-Methoden ist es zudem möglich, dass die Erklärung, wie ein KI-System vorgeht, von dessen tatsächlicher Vorgehensweise abweicht.[10] Außerdem ist eine vollständige Transparenz und Nachvollziehbarkeit von Blackbox-Modellen derzeit nicht möglich.[11]

Wenn XAI-Methoden einen Mehrwert für Bürger*innen und Verbraucher*innen haben sollen, sind begleitende Maßnahmen notwendig. Das Wissen darüber, wie ein KI-System zu einem Ergebnis kommt, ist die Grundlage dafür, dass Nutzer*innen entsprechend handeln und zum Beispiel falsche oder benachteiligende Ergebnisse anfechten können. Handlungsmöglichkeiten setzen jedoch zugängliche Erklärungen und Beschwerdestellen voraus. Lehnen beispielsweise Vermieter*innen potenzielle Mieter*innen aufgrund der Vorsortierung einer KI-Anwendung ab, müssen die Bewerber*innen wissen, welche Faktoren ausschlaggebend waren und wo sie sich im Zweifel beschweren können. XAI-Methoden können Blackbox-Modelle zwar nachvollziehbarer machen – sie sind für die Nachvollziehbarkeit von KI-Anwendungen seitens Verbraucher*innen, Nutzer*innen und Betroffenen allerdings nicht ausreichend.

Einblicke in die Arbeit der *ZVKI*-Partner*innen: Welche Ansätze gibt es, um KI-Systeme zu erklären?

Die *Fraunhofer-Institute AISEC* und *IAIS* sowie die *Freie Universität Berlin* erarbeiten aus verschiedenen Blickwinkeln wissenschaftliche Erkenntnisse zu den Fragen, was vertrauenswürdige KI ausmacht und wie sie gefördert werden kann. Einen Schwerpunkt der Forschungsarbeiten bilden Erklärmethoden für KI-Systeme. Sie spielen unter anderem eine zentrale Rolle im Verbraucher*innenschutz, zum Beispiel bei der Durchsetzung einer diskriminierungsfreien KI-Anwendung. Erklärmethoden können hauptsächlich in folgende Kategorien eingeteilt werden:

Modellspezifische oder modellagnostische Erklärmethoden:

Modellspezifische XAI-Methoden erfordern einen Zugang zur internen Funktionsweise des KI-Modells und sind nicht auf andere KI-Modelle übertragbar.

Modellagnostische XAI-Methoden können bei jedem Blackbox-Modell angewendet werden.

Lokale oder globale Erklärmethoden:

Lokale XAI-Methoden setzen das Ergebnis eines KI-Systems in Bezug zu den spezifischen Eingaben.

Globale XAI-Methoden erklären die Gesamtlogik des KI-Modells.

*Mehr über unsere Verbundpartner*innen gibt es auf unserer Webseite:*

www.zvki.de > ZVKI-Exklusiv > Netzwerk.

7 Ebd., S. 3.
8 Kraus/ Ganschow, S. 38; Kraus et al., S. 3.
9 Lossos/ Geschwill/ Morelli, S. 312, S. 314.
10 Schmid, o. S.
11 Walmsley, S. 592.

Wie kann die Nachvollziehbarkeit von KI-Systemen sichergestellt werden?

Aktuell beschränken sich die meisten Ansätze erklärbarer KI auf eine technische Perspektive. Doch KI-Anwendungen sind sozio-technische Systeme[12]: Entscheider*innen und Entwickler*innen müssen die Erklärbarkeit von KI-Systemen über den gesamten Entwicklungs- und Anwendungsprozess eines KI-Systems mitdenken. Erklärbarkeit darf sich deshalb nicht auf XAI-Methoden beschränken, die nachträglich komplexe KI-Systeme nachvollziehbar machen sollen.[13]

Erklärbare KI ist vielmehr ein Kommunikationsprozess zwischen verschiedenen Menschen.[14] Ein solcher ganzheitlicher Ansatz beinhaltet zum Beispiel, dass sich Auftraggeber*innen und Entwickler*innen zu Beginn des Entwicklungsprozesses fragen, ob ein KI-System nötig ist und wie es gestaltet sein muss: Welche Erklärmethoden und Kommunikationswege sind notwendig, damit die KI-Anwendung anschließend für Anwender*innen oder Betroffene nachvollziehbar ist? Dabei sind je nach Anwendungskontext unterschiedliche Erklärungen sinnvoll.[15] Verschiedene Zielgruppen stellen außerdem individuelle Ansprüche an die Erklärungen eines KI-Systems.[16] Deshalb müssen XAI-Methoden mit den Endnutzer*innen erprobt werden.[17]

Fach-AG Verbraucher*innen-Information

Die Fach-AG Verbraucher*innen-Information des *ZVKI* beschäftigt sich mit erklärbarer KI im Verbraucher*innenkontext und stellt die Frage: Wie können Nutzer*innen und Betroffene verstehen, warum ein KI-System ein bestimmtes Ergebnis ausgibt? Ziel ist es, Handlungsempfehlungen zu erarbeiten, die aufzeigen, wie XAI zur Aufklärung von Verbraucher*innen in der Praxis genutzt werden kann. Zu den Teilnehmer*innen zählen unter anderem Expert*innen aus Verbraucher*innenschutz- und Bildungsorganisationen sowie aus der Wissenschaft.

Mehr über die Fach-AG gibt es auf unserer Webseite:

www.zvki.de > Mitmachen.

Erhöht erklärbare KI das Vertrauen in KI-Systeme?

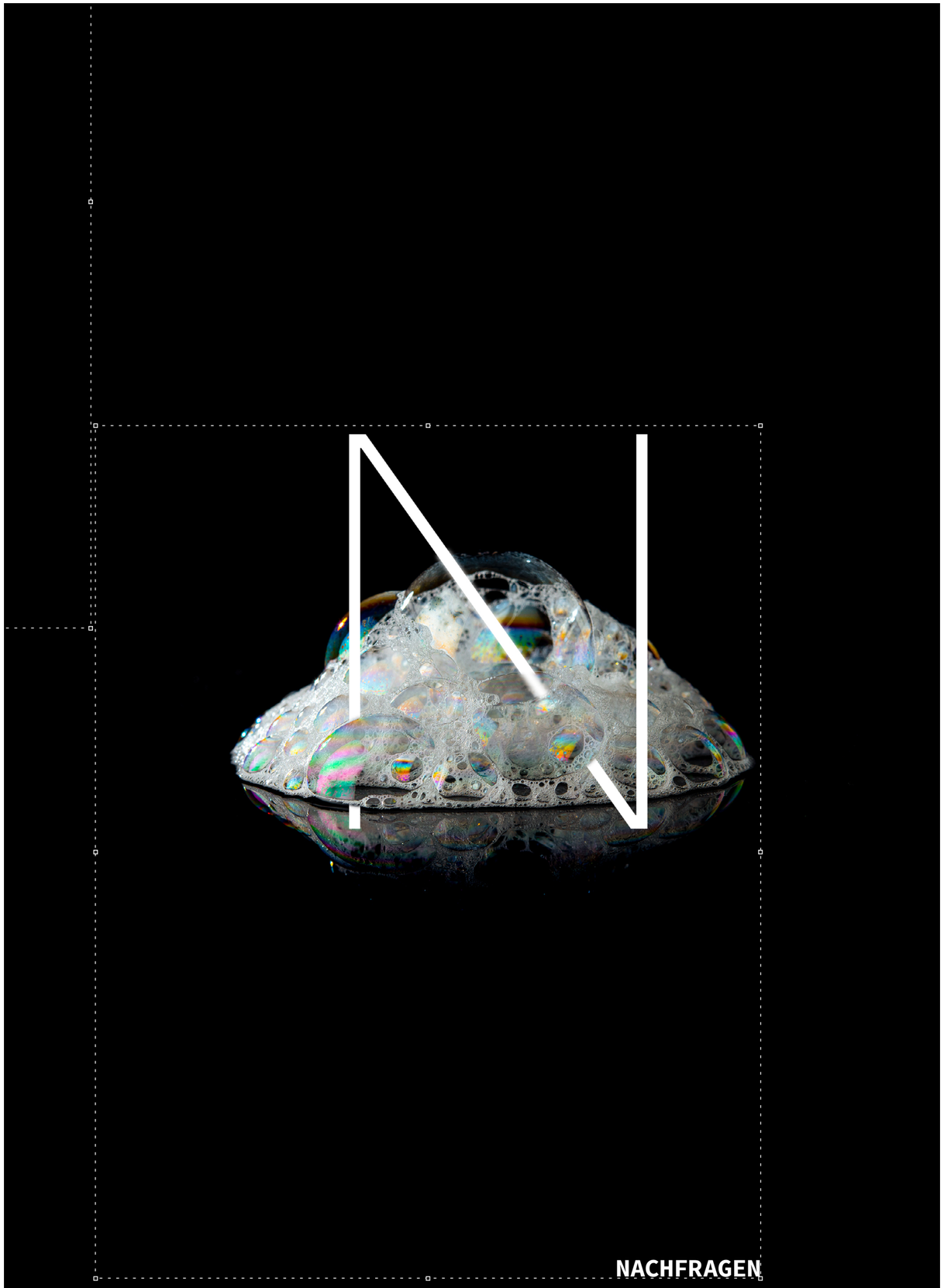
Erklärbare KI kann die Nachvollziehbarkeit und Transparenz von KI-Systemen erhöhen – vor allem, wenn Entwickler*innen die Erklärbarkeit über den gesamten Entstehungsprozess einer KI-Anwendung mitdenken. Doch steigert erklärbare KI auch die Vertrauenswürdigkeit von KI-Systemen?

Laut der Metastudie „How to explain AI systems to end users“ von 2022 ist es ein allgegenwärtiges Ziel, mithilfe von Erklärmethoden die Vertrauenswürdigkeit eines KI-Systems zu erhöhen.[18] Viele Wissenschaftler*innen nehmen an, dass Erklärbarkeit das Vertrauen von Menschen in KI-Anwendungen steigert.[19] Dabei kommt es unter anderem auf die Qualität und die Art der jeweiligen Erklärung an[20] – nicht jede Erklärung führt automatisch zu mehr Vertrauen. Andere Studien kommen hingegen zu dem Schluss, dass Erklärungen keinen oder sogar einen negativen Effekt auf das Vertrauen von Menschen in KI-Anwendungen haben können. Dies könne daran liegen, dass eine Erklärung Probleme und Fehler des Systems aufzeigt. Dadurch können Menschen Vertrauen in das System verlieren.[21]

Diese Erkenntnis widerspricht jedoch nicht automatisch dem genannten Ziel von erklärbarer KI, die Vertrauenswürdigkeit von KI-Systemen zu erhöhen: Lena Kästner, Professorin für Computer Science, und ihre Kolleg*innen argumentieren, dass es sinnvoll ist, zwischen dem Vertrauen in und der Vertrauenswürdigkeit von KI-Systemen zu unterscheiden. Wenn erklärbare KI die Transparenz von KI-Systemen erhöht, erleichtert das Entwickler*innen, Nutzer*innen und Betroffenen, die Systeme zu überprüfen. Auf dieser Grundlage können sie beurteilen, ob ein KI-System vertrauenswürdig ist oder nicht. Erklärbare KI führt also nicht automatisch dazu, dass Menschen KI-Anwendungen mehr Vertrauen entgegenbringen. XAI kann aber die Vertrauenswürdigkeit von KI-Systemen steigern, indem Menschen die Ergebnisse von KI-Anwendungen überhaupt überprüfen und damit legitimieren können.[22] Inwiefern diese Annahme in der Praxis haltbar ist, müssen weitere Studien zeigen.

12 Ehsan, S. 1.
13 Asghari et al., S. 2. Asghari et al. setzen sich in ihrem Bericht mit Systemen zur algorithmischen Entscheidungsfindung (ADM-Systemen) als sozio-technischen Systemen auseinander. Da auch KI-Anwendungen ADM-Systeme sind, übertragen wir die Erkenntnisse der Autor*innen auf diese.
14 Ebd., S. 1.
15 Ebd., S. 12.
16 Langer et al., S. 5.
17 Asghari et al., S. 1.
18 Laato et al., S. 10.
19 Kästner, S. 169.
20 Shin.
21 Kästner, S. 169.
22 Kästner, S. 170ff.

Verweis	Seite	
Mehr zur Funktionsweise und zum Einsatz von XAI-Methoden finden Sie im Interview mit Wojciech Samek in der Rubrik NACHFRAGEN unter „Mit welchem Ziel wollen wir erklären?“.	24	
Ein Beispiel für ein Erklärungskonzept einer KI-Spracherkennungssoftware finden Sie in der Rubrik VERARBEITEN unter „Wie erzeugen wir Verständnis?“.	28	**Autorin**
		Franziska Busse



NACHFRAGEN

MIT WELCHEM ZIEL WOLLEN WIR ERKLÄREN?

Es gibt zahlreiche Ansätze, wie Erklärungen von KI-Modellen aussehen können. Viele davon sind nur bedingt aussagekräftig. Wojciech Samek entwickelte einen dieser Ansätze gemeinsam mit Kolleg*innen weiter, um verständlichere Erklärungen zu erhalten. Im Interview schildert er, wie dieses neue Verfahren funktioniert, warum Erklärbarkeit wichtig ist, woran Forscher*innen weiterarbeiten müssen und warum es notwendig ist, dabei die Nutzer*innen und Einsatzkontexte zu berücksichtigen.

Die Erklärbarkeit von KI scheint in Fachkreisen einen regelrechten Hype zu erfahren. Warum ist das so?

An sich ist das Thema Erklärbarkeit in Bezug auf KI nicht neu. Bereits in den 1990er-Jahren beschäftigte es Wissenschaftler*innen. Auch damals wollten Entwickler*innen wissen, wie die Modelle funktionieren, mit denen sie arbeiten. Etwa seit der Jahrtausendwende, aber vor allem in den letzten Jahren haben sich Deep-Learning-Anwendungen stark verbreitet und damit wurden auch die Anwendungsbereiche vielfältiger. Gegenwärtig wirken KI-Modelle auf das Leben und Arbeiten von Menschen in der Medizin, im Mobilitätssektor oder in der Forschung. Dadurch rückt das Thema Erklärbarkeit stärker in den Fokus. Zu Recht wollen wir verstehen, wie die eingesetzten Modelle vorgehen und ob sie korrekt funktionieren. Wir wollen nachvollziehen können, wie sie eine gestellte Aufgabe lösen.

Wie sehen Erklärungen von KI-Modellen aus?

Wir haben ein Verfahren für Erklärungen entwickelt, das „layer-wise relevance propagation“ heißt. Es bezieht sich vor allem auf den Input – also auf die Daten, die ein System erhält. Die Erklärung enthält Informationen dazu, welche Inputdimensionen (zum Beispiel Pixel in einem Bild) durch das Modell als besonders wichtig bewertet werden. Damit erfahren wir mehr darüber, wie sich der Input zum Ergebnis verhält. Wenn es beispielsweise darum geht, auf Bildern einen Hund zu erkennen, erhalten wir als Erklärung eine sogenannte Heatmap. Das ist eine Darstellung davon, welche Pixel oder Bereiche eines Bilds als besonders relevant bewertet wurden. Im Fall des Hundes sind das etwa die Pixel, die die Schnauze oder die Ohren abbilden. Damit entsteht eine Verbindung zwischen dem Eingangsbild und dem Ergebnis.

Diese Art des Erklärens ist weit verbreitet und zugleich bleibt es ein recht allgemeines Verfahren. Ich möchte das an einem Beispiel verdeutlichen: Wir haben einen Klassifikator trainiert, der das Alter von Menschen anhand von Bildern schätzen soll. Das Modell klassifizierte ein Bild einer jungen lachenden Frau als jung. Dabei bewertete es die Bildausschnitte des Mundes als relevant. Das Bild einer alten lachenden Frau wurde als alt eingestuft. Hierbei wurden die Pixel, die den Mund abbildeten, als kontraindikativ für das Ergebnis (negative Relevanz) bewertet. Das lässt folgende Annahme zu: Das Modell hat gelernt, dass Menschen, die viel lachen, jünger sind als diejenigen, die weniger lachen. Dieses Beispiel stellte ich in einigen meiner Vorträge vor. Das führte häufiger zu der Frage, woher ich denn wüsste, ob wirklich das Lachen an sich ausschlaggebend sei. Es könnte doch auch sein, dass das Modell die Farbe der Zähne als relevant bewertete. Mit den Heatmaps allein können wir das nicht eindeutig beantworten. Wir wissen lediglich, welche Pixel als besonders bedeutsam bewertet wurden.

Definition: Deep Learning

(auf Deutsch: tiefes Lernen) Komplexe künstliche neuronale Netze, die mehr als drei Schichten künstlicher Neuronen umfassen

Der Begriff „deep“ bzw. „tief“ bezieht sich auf die Menge der Schichten künstlicher Neuronen. Mit jeder zusätzlichen Schicht erhöht sich der Abstrahierungsgrad des Systems und die Fähigkeit, komplexere Aufgaben zu bewältigen. Dazu gehört die Erstellung von realistischen Deep Fakes. Gleichzeitig verstärkt sich dabei das Problem von Künstlicher Intelligenz als Blackbox.

Weitere Begriffsdefinitionen finden Sie hier:

www.zvki.de > KI-Navigator > Unsere Inhalte > Glossar.

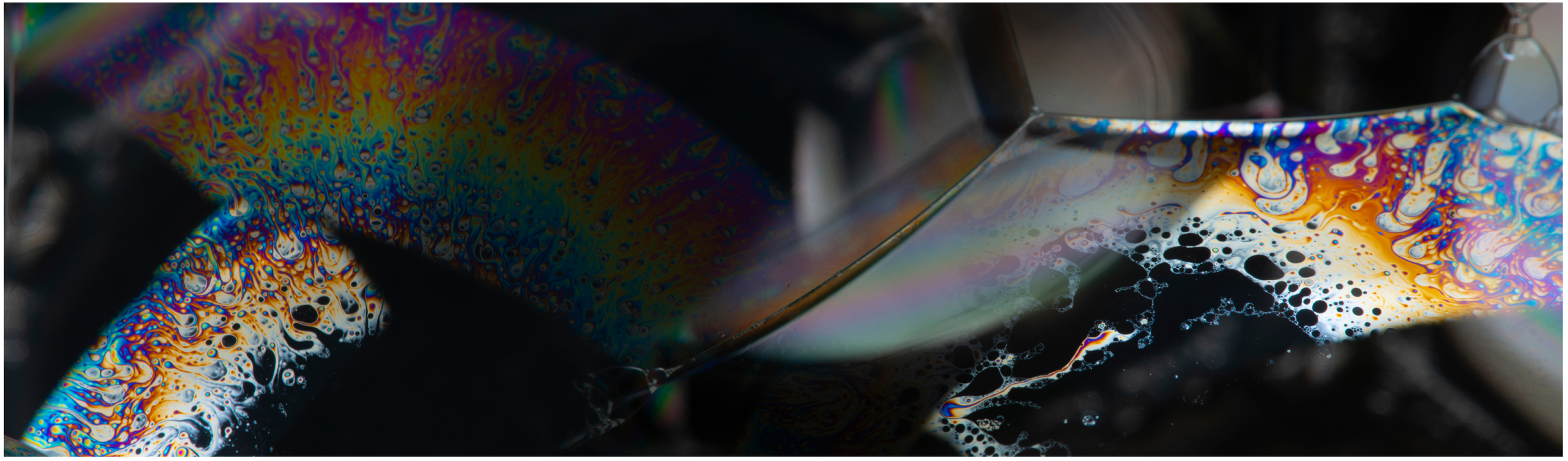
Wie ist es möglich, zu aussagekräftigeren Erklärungen zu kommen?

Genau daran arbeiten wir gerade. Eine Möglichkeit ist, die Erklärung von der Pixel- auf eine Konzeptebene zu heben, die für uns Menschen verständlicher ist. Im Fall des Beispiels mit der Altersschätzung würde das bedeuten, dass wir darstellen, ob das Konzept ‚Lachen‘ oder das Konzept ‚Farbe der Zähne‘ relevant für das Ergebnis ist. Dafür haben wir kürzlich eine neue Methode entwickelt. Sie nennt sich „concept relevance propagation“ und erlaubt uns, Entscheidungen stärker in menschliche Denkschemata zu übertragen. Dadurch sind sie genauer und verständlicher. Das Paper dazu ist im Juni erschienen.[1]

Ich bin davon überzeugt, dass wir Erklärungen so gestalten müssen, dass die jeweiligen Empfänger*innen in der Lage sind, sie zu verstehen. Bei Bildern ist das vergleichsweise einfach, weil wir Menschen sehr gut darin sind, Bilder zu lesen. Bei anderen Daten ist es deutlich komplizierter. Wenn beispielsweise ein EKG-Signal klassifiziert werden soll, müssen wir herausfinden, wie Kardiolog*innen das machen, damit die Erklärungen für sie verständlich sind.

„Ich bin davon überzeugt, dass wir Erklärungen so gestalten müssen, dass die jeweiligen Empfänger*innen in der Lage sind, sie zu verstehen.“

1 Lapuschkin/ Samek et al.



*Wir sprechen seit ungefähr einer halben Stunde miteinander. Sie haben meine Fragen beantwortet und mir Visualisierungen gezeigt. Dadurch und mithilfe von ein paar Vorkenntnissen sind die konzeptbasierten Erklärungen für mich nachvollziehbar und erscheinen hilfreich. Nutzer*innen ohne Vorkenntnisse und Unterstützung können diese Erklärungen wahrscheinlich nicht so schnell verstehen. Wenn es unser Ziel ist, KI-Modelle für sie nachvollziehbar zu machen, brauchen wir wohl noch verständlichere Erklärungen. Ist das realistisch?*

Ich denke, wir können diese konzeptbasierten Erklärungsansätze dahingehend weiterentwickeln, dass die Nutzer*innen am Ende besser verstehen, auf welchen Annahmen ein Ergebnis basiert. Momentan ist das noch nicht der Fall. Allerdings ist meiner Meinung nach die interessantere Frage, was die Nutzer*innen mit diesen Erklärungen machen. Eine Erklärung ist kein Selbstweck, sondern bringt einen zusätzlichen Nutzen. Damit das der Fall ist, müssen wir uns fragen, welches Ziel eine Erklärung erfüllen soll. Die Antwort darauf hängt stark von der Anwendung selbst und den Nutzer*innen ab. In diesem Zusammenhang ist Forschung wichtig, die sich damit befasst, wie Nutzer*innen mit den Erklärungen interagieren. Solche Arbeiten gibt es bereits. Manche davon zeigen, dass die bloße Existenz einer Erklärung zu einer falschen Sicherheit führen kann. Es entsteht der Eindruck, die vorhandene Erklärung zeige an, dass das Ergebnis richtig oder die Anwendung sicher sei.

An diesen Interaktionsfragen arbeiten derzeit einige Forscher*innen. Ihre Arbeiten sind wichtig, um in Zukunft Gesamtsysteme für Erklärungen zu entwickeln, in denen Erklärungen eines bestimmten KI-Modells in einem direkten Zusammenhang mit den Nutzer*innen stehen.

„Eine Erklärung ist kein Selbstweck, sondern bringt einen zusätzlichen Nutzen. Damit das der Fall ist, müssen wir uns fragen, welches Ziel eine Erklärung erfüllen soll.“

Dazu gehört auch, genauer zu bestimmen, was eine Erklärung im juristischen Sinn erfüllen muss. In diesem Zusammenhang ist noch vieles offen. Natürlich brauchen Entwickler*innen und Forscher*innen Erklärungen, um Modelle zu verbessern. Wenn wir noch einen Schritt weiterkommen möchten, sind auch Fragen danach wichtig, was eine gute Erklärung in einem juristischen Verständnis umfasst.

Ist es möglich, solche Eigenschaften, die eine Erklärung erfüllen muss, global für alle Anwendungen festzulegen?

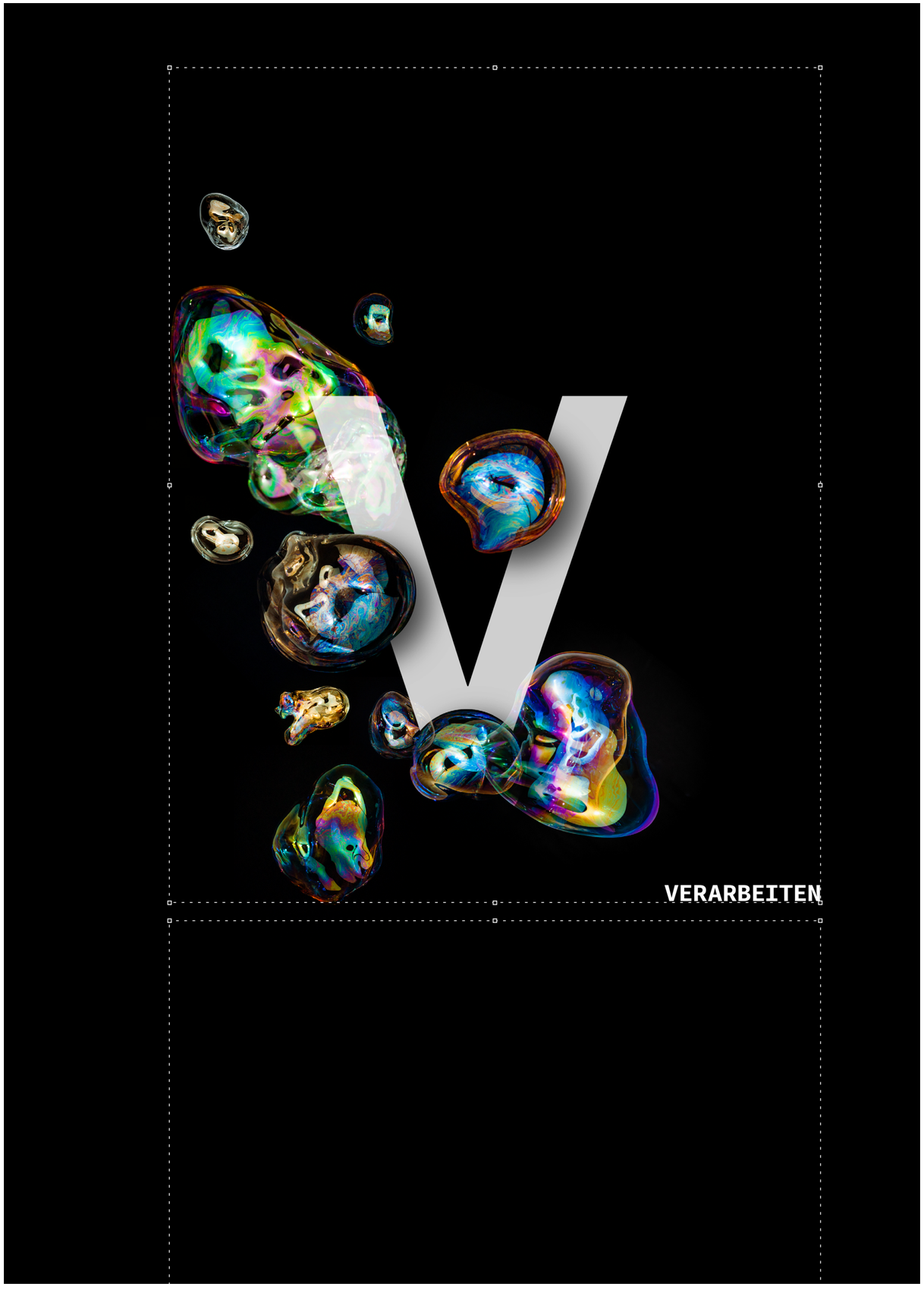
Das spielt auf die Prüfbarkeit an – also die Frage: Wie können wir mithilfe von Prüfungsmechanismen und -schemata sicherstellen, dass ein KI-Modell bestimmte Kriterien erfüllt, wie fehlerfrei zu funktionieren oder fair zu sein? Ich glaube nicht, dass wir hier ein einheitliches Maß oder Prüfkriterium für alle Erklärungen festlegen können. Deshalb dürfen wir aber nicht stehen bleiben. Wir können trotzdem daran arbeiten, die Erklärungen verständlicher zu machen.

„Ich glaube nicht, dass wir hier ein einheitliches Maß oder Prüfkriterium für alle Erklärungen festlegen können. Deshalb dürfen wir aber nicht stehen bleiben. Wir können trotzdem daran arbeiten, die Erklärungen verständlicher zu machen.“

KI-Modelle erklärbar zu machen, ist ein Ansatz. Ein anderer ist, sich gut zu überlegen, in welchen Zusammenhängen ihr Einsatz sinnvoll ist. Sollten wir an manchen Stellen lieber auf komplexe KI-Modelle verzichten, anstatt auf Erklärungen abzuzielen?

In manchen Kontexten ist es aus ethischen Gründen nicht sinnvoll, komplexe KI-Modelle einzusetzen, beispielsweise bei Gerichtsurteilen. Deshalb ist es wichtig, solche Bereiche zu definieren und an diesen Stellen auf andere Prozesse zurückzugreifen. In anderen Bereichen ist der Einsatz von komplexen KI-Verfahren zwar ethisch unbedenklich, aber nicht nützlich. Ein Beispiel dafür sind Methoden zur Videocodierung, um Videosignale zu komprimieren. Hier wurden über viele Jahre hinweg klassische Verfahren von Expert*innen entwickelt, die hervorragend und effizient funktionieren.

Verweis	**Seite**		
Ein Beispiel für ein Erklärungskonzept einer KI-Spracherkennungssoftware finden Sie in der Rubrik VERARBEITEN unter „Wie erzeugen wir Verständnis?“.	28		**Wojciech Samek** ist Professor am *Institut für Softwaretechnik und Theoretische Informatik* der *Technischen Universität Berlin* und Leiter der Abteilung für Künstliche Intelligenz sowie der Gruppe „Erklärbare KI“ am *Fraunhofer Heinrich-Hertz-Institut (HHI)*. Er hat mehrere Auszeichnungen für seine Arbeiten erhalten. Im Juni 2022 veröffentlichte er gemeinsam mit Kolleg*innen das Paper „From ‘Where’ to ‘What’: Towards Human-Understandable Explanations through Concept Relevance Propagation“, das einen neuen Ansatz der Erklärbarkeit von KI-Modellen vorstellt.
		© Fraunhofer Heinrich-Hertz-Institut	
		Interview von	Jaana Müller-Brehm



WIE ERZEUGEN WIR VERSTÄNDNIS?

Menschen nutzen KI-Software, die Sprache erkennen und verarbeiten kann, mittlerweile in vielen Anwendungen, etwa bei Sprachassistenten von Smartphones oder bei Smart Speakern zu Hause und im Auto. Doch wie funktionieren solche Anwendungen der Künstlichen Intelligenz? Das Programm „Talk to me“ macht dies für Lai*innen besser nachvollziehbar und ist ein Beispiel dafür, wie verständliche Erklärungen von KI-Modellen gestaltet sein können.

Wer?

„Talk to me“ basiert auf der Forschung der Doktorand*innen Valerie Krug und Jens Johannsmeier, die im *Artifical Intelligence Lab* der *Otto-von-Guericke-Universität Magdeburg* unter der Leitung von Professor Sebastian Stober tätig sind. Gemeinsam mit der gemeinnützigen Organisation *IMAGINARY*, die sich für die Vermittlung von Mathematik einsetzt, entwickelten sie eine erste Version des Programms weiter. Interessierte können es als Exponat der von *IMAGINARY* konzipierten Wanderausstellung „I AM A.I. – explaining artificial intelligence“ testen.[1]

Wann und wo?

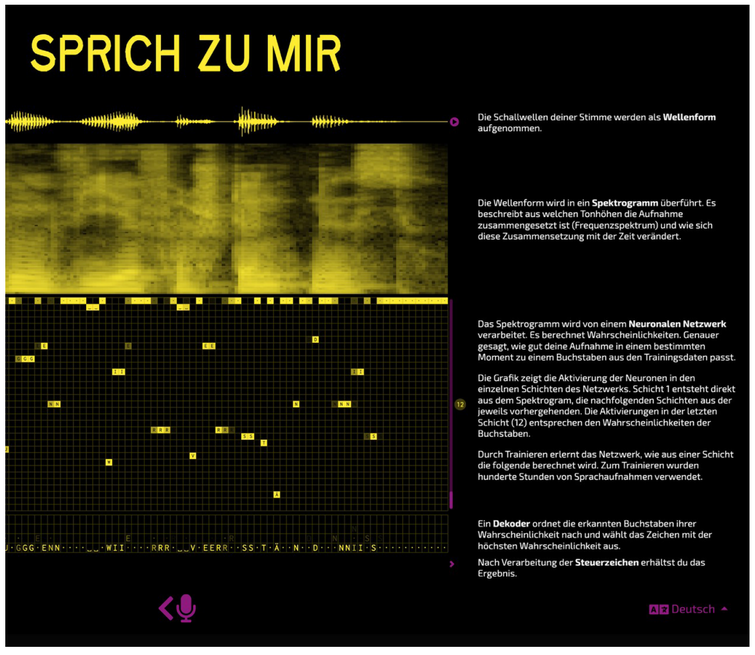
- **Forschung: Dezember 2016 bis April 2018 in Potsdam**
  Zu Beginn lag der Fokus auf dem Thema „transfer learning“, um ein Machine-Learning-Modell, das bereits auf die englische Sprache trainiert worden war, an die deutsche Sprache anzupassen. Mit dem Ziel, diesen Vorgang besser zu verstehen, kamen die Forscher*innen auf die Idee, XAI-Methoden anzuwenden.

- **Entwicklung: 2018 bis März 2020 in Potsdam, Magdeburg und Berlin**
  Eine erste Version von „Talk to me“ konzipierte das Team bereits 2018 für eine Ausstellung in Potsdam. Zusammen mit *IMAGINARY* entwickelten sie diese Version für „I AM A.I.“ weiter.

- **Ausstellung: seit 2022**
  Die erste Station der Wanderausstellung ist Heidelberg. „Talk to me“ ist bis Ende 2022 in der *MAINS (Mathematik-Informatik-Station)* zu sehen und zieht danach weiter zur *Friedrich-Schiller-Universität in Jena*.

Was?

Verfahren, um Sprache automatisiert zu verarbeiten, beschäftigen Forscher*innen schon seit einigen Jahrzehnten. Vor allem seit der Entwicklung von Sprachassistenten für Smartphones haben sich entsprechende Verfahren stark verbreitet. Dabei ist das Erkennen und Verarbeiten von Sprache ein komplexer Vorgang: Die Aussprache unterscheidet sich je nach Sprecher*in. Dialekte, Akzente und Störgeräusche können die Analyse unserer Sprache erschweren. „Talk to me“ veranschaulicht, wie KI-Software beim Erkennen und Verarbeiten von Sprache vorgeht. Das Ziel ist, dass Nutzer*innen die einzelnen Schritte dieses Prozesses verstehen können.



Bildschirmfoto der deutschen Version von „Talk to me“, Lizenz: IMAGINARY/talk-to-me, Apache License, Version 2.0, http://www.apache.org/licenses/

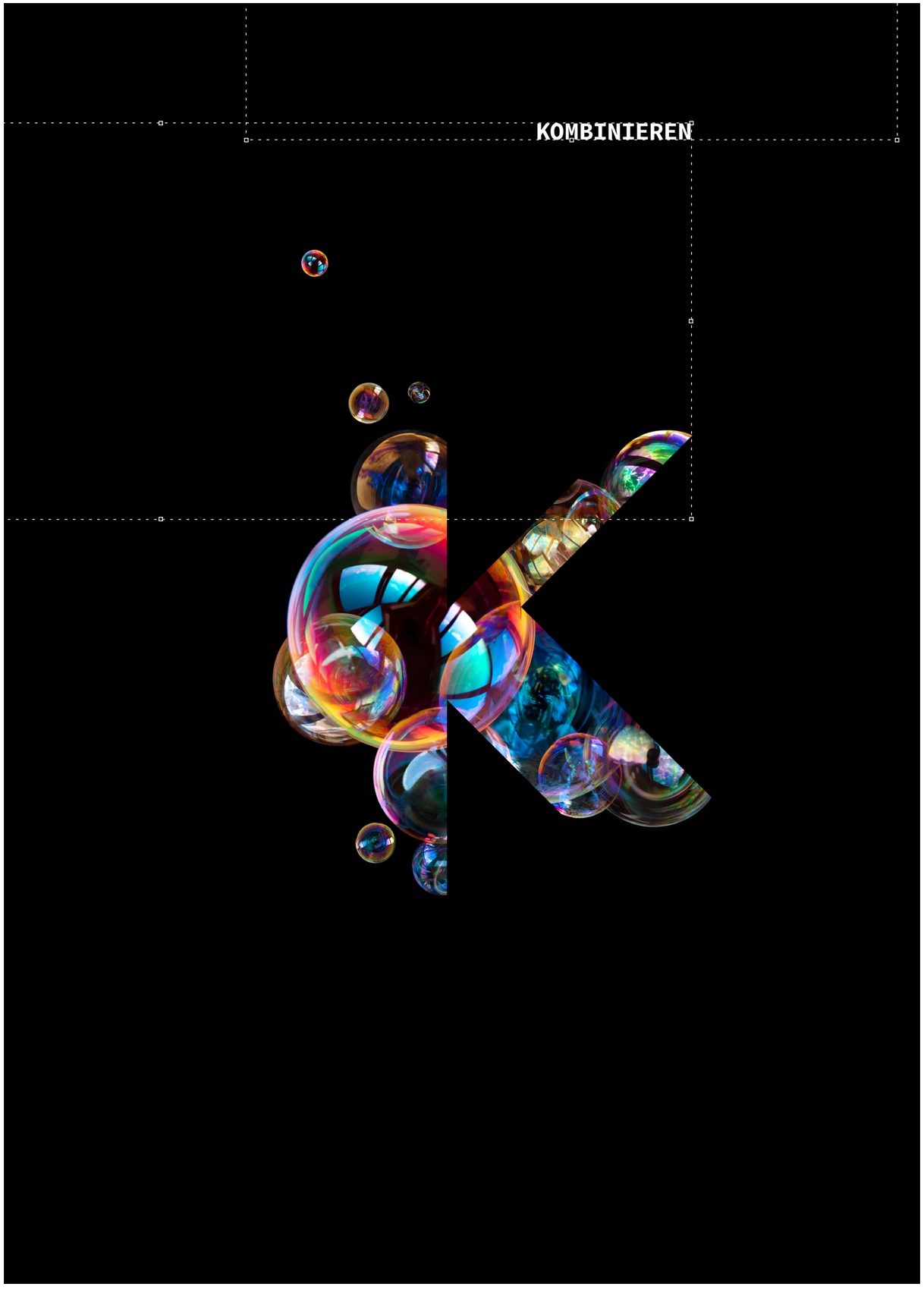
Wie?

Mit „Talk to me“ können Sprecher*innen ein Spracherkennungssystem testen. Sie erhalten Informationen darüber, wie die Software ihre Spracheingabe verarbeitet . Die Verarbeitung erfolgt in einer Reihe von Schritten: Zunächst sprechen die Nutzer*innen ein Wort oder einen Satz ein. Anschließend zeigt eine Visualisierung, wie das Gesprochene in ein sogenanntes Spektrogramm umgewandelt wird. In dieser Form kann ein künstliches neuronales Netzwerk die Daten verarbeiten, Muster darin erkennen und ermitteln, welche Buchstaben und Wörter die Sprecher*innen mit größter Wahrscheinlichkeit gesagt haben. Dieses Vorgehen beruht auf einer Trainingsphase, in der die Forscher*innen das Netz mit frei verfügbaren Sprachaufnahmen trainierten. „Talk to me“ bildet diese Schritte visuell ab und erklärt sie mithilfe von Texten. Zum Schluss stellt das Programm die erkannten Buchstaben und Pausen entsprechend der höchsten Wahrscheinlichkeit dar. So ergeben sich einzelne Wörter und ganze Sätze.

„Talk to me“ ist ein Open-Source-Programm und auf der Plattform *GitHub* verfügbar. Jede*r kann das Programm herunterladen, nutzen und weiterverwenden. Zudem können Interessierte auch den Quellcode über *GitHub* herunterladen.[2]

		Autorin
	Lisa Schmechel	

1 IMAGINARY.
2 IMAGINARY (2020).



WAS NEHMEN WIR MIT?

Die zweite Ausgabe von Missing Link widmet sich dem meistgenannten Prinzip in ethischen Richtlinien und Vorschlägen zur Gestaltung vertrauenswürdiger KI-Anwendungen – der Transparenz. Trotz einiger Jahre des Diskurses sind noch einige Fragen offen: An wen richten sich Transparenzmaßnahmen? Welchen Zweck sollen sie erfüllen? Wie können sich verschiedene Ansätze sinnvoll ergänzen? Informationsangebote, Zertifikate und Methoden der erklärbaren KI sind kein Selbstzweck, ebenso wenig wie der Einsatz von KI-Anwendungen an sich.

Transparenz, Nachvollziehbarkeit und Erklärbarkeit erfahren einen regelrechten Boom in digitalpolitischen Debatten und Fachdiskursen darüber, wie wir KI-Anwendungen rechtssicher und gesellschaftsorientiert gestalten und nutzen können. Auch der „Artificial Intelligence Act“ der *Europäischen Kommission* sieht Transparenzpflichten vor. Doch diese Themen sind im Zusammenhang mit Technik im Allgemeinen und Künstlicher Intelligenz im Speziellen nicht neu, sondern seit Jahrzehnten etabliert. Fragen danach, wie wir KI-Entwicklungsprozesse transparent und entsprechende Anwendungen nachvollziehbar machen können, bleiben dennoch aktuell und werden zunehmend bedeutsamer. Warum das so ist, beschreibt der Wissenschaftler Wojciech Samek im Interview: Seit der Jahrtausendwende verbreiten sich Deep-Learning-Anwendungen zunehmend. Diese Modelle werden immer komplexer und wir begegnen ihnen bereits in vielen Bereichen unseres Lebens. Das verleiht den Themen Transparenz, Nachvollziehbarkeit und Erklärbarkeit eine neue Brisanz. Transparenzmaßnahmen haben das Ziel, die Nachvollziehbarkeit von KI-Systemen für bestimmte Zielgruppen zu erhöhen. Dabei können unter anderem Erklärungen helfen. Transparenz kann sich auf drei Ebenen beziehen: auf das Kennzeichnen, dass eine KI-Anwendung zum Einsatz kommt, auf den Entwicklungsprozess und die daran Beteiligten sowie auf die Funktionsweise von KI-Modellen.

Ein Weg, um mehr Transparenz zu schaffen und etwaige zukünftige Rechtsvorgaben umzusetzen, sind Standards und darauf basierende Prüfprozesse. Sie bilden die Grundlage, um KI-Anwendungen und die Prozesse ihrer Entwicklung mit Zertifikaten zu versehen. Solche Maßnahmen können einerseits Unternehmen dabei helfen, Pflichten klar zu definieren und Rechtssicherheit zu schaffen. Andererseits ist damit die Hoffnung verbunden, dass Verbraucher*innen, Nutzer*innen und Betroffene besser beurteilen können, welche Anwendungen sie nutzen möchten und ob sie Ergebnisse von KI-Verfahren anfechten können. Derzeit arbeiten Akteur*innen aus Wirtschaft, Politik, Zivilgesellschaft und Wissenschaft gemeinsam an Vorschlägen, mit denen sich KI-Anwendungen überprüfen lassen. Bisher sehen die Vorschläge häufig eine freiwillige und oft nur interne Prüfung vor. Die genauen Methoden bleiben häufig offen. Ob Standards, Prüfprozesse und Zertifikate geeignet sind, mehr Transparenz für Lai*innen zu schaffen, bleibt allerdings unklar: Ansätze dafür, wie Prüfergebnisse verständlich an sie vermittelt werden können, sind bislang nicht weit gediehen. Indes ergeben repräsentative Umfragen, dass die Teilnehmer*innen Maßnahmen wie unabhängige Prüfungen als notwendig erachten.

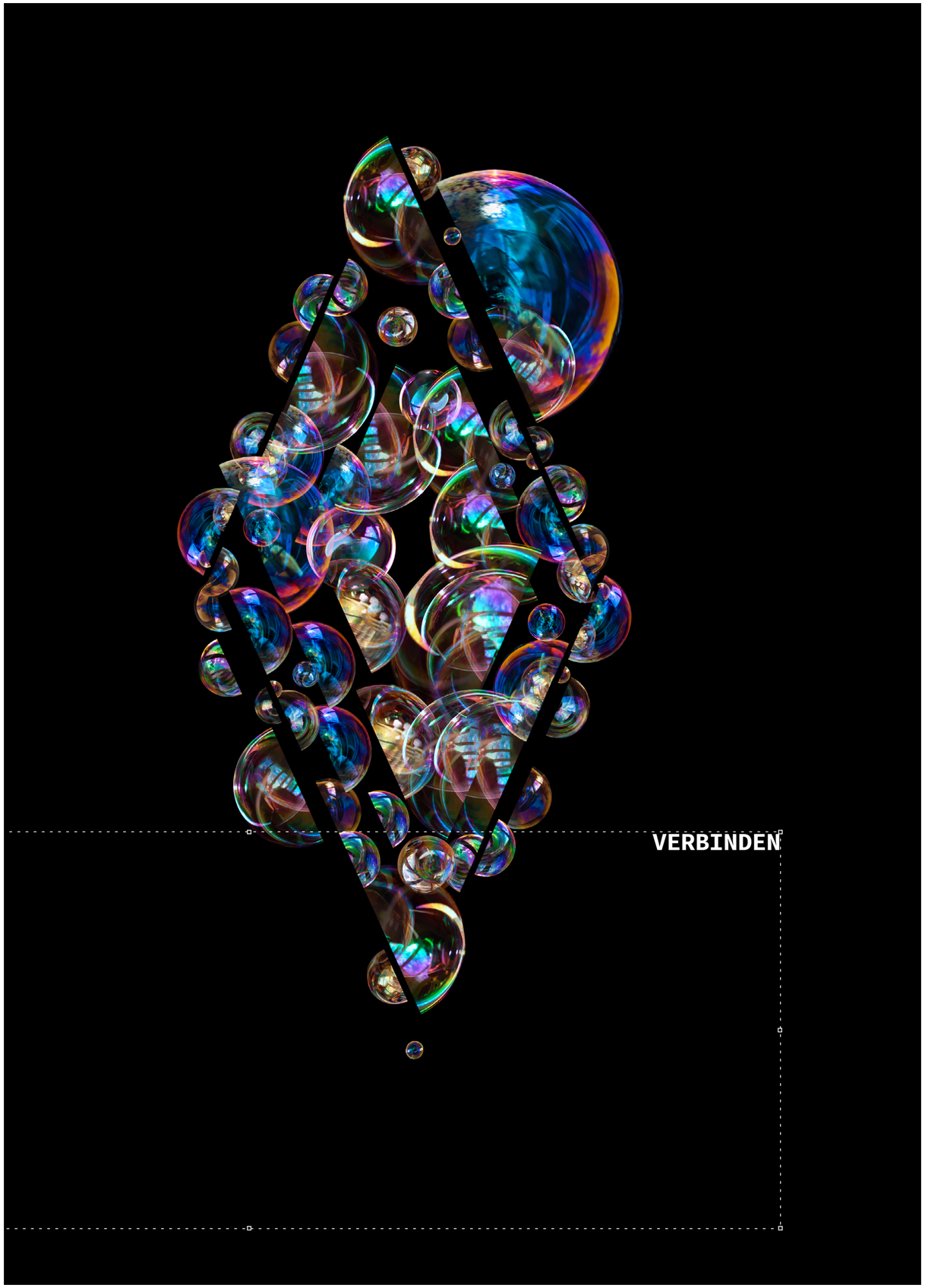
Einige Vorschläge für Richtlinien, Standards und Regulierungsempfehlungen nennen Erklärungen als eine Möglichkeit, damit sich KI-Anwendungen besser nachvollziehen lassen. Gegenwärtig forschen zahlreiche Wissenschaftler*innen an Erklärmethoden, um die Funktionsweise komplexer Blackbox-KI-Modelle besser verstehen zu können. Diese Arbeiten haben ihren Ursprung im Wunsch von Entwickler*innen, die Vorgehensweise ihrer eigenen KI-Systeme besser verstehen und beeinflussen zu können. Für Expert*innen mit dem notwendigen technischen Wissen existieren solche Methoden bereits. Sie beziehen sich auf einzelne Aspekte der Funktionsweise eines KI-Modells, können jedoch kein Blackbox-KI-Modell in Gänze nachvollziehbar machen. Für Verbraucher*innen, Nutzer*innen und Betroffene sind diese Erklärungen bislang unzugänglich und unverständlich. Dabei sind sie durchaus an Informationen zur Funktionsweise von KI-Systemen interessiert, wie repräsentative Befragungen zeigen. Verständliche Erklärungen sind für viele ein wichtiges Kriterium, um KI-Anwendungen zu akzeptieren.

Daher sind Informationsangebote derzeit das Mittel der Wahl, wenn Lai*innen KI-Verfahren (besser) verstehen oder mehr über bestimmte Anwendungen wissen möchten. Doch das Bereitstellen von Informationen führt nicht automatisch zu mündigen Bürger*innen und Verbraucher*innen. Ob Informationsangebote tatsächlich zu mehr Nachvollziehbarkeit beitragen, hängt davon ab, wie sie aufbereitet sind, wo sie zu finden sind und wer sie zur Verfügung stellt. Umfragen ergeben, dass sich die Befragten vor allem journalistisch aufbereitete Inhalte wünschen. Dabei sind sie nicht nur an der Funktionsweise von KI-Anwendungen und den verwendeten Daten, sondern auch an den damit verbundenen Chancen und Risiken interessiert. Expert*innen schreiben vermittelnden Dritten wie den *Verbraucherzentralen* eine wichtige Rolle zu. Sie können als Informationslotsen dazu beitragen, dass Verbraucher*innen, Nutzer*innen und Betroffene passende Informationen erhalten.

Wissenschaftler*innen wie Wojciech Samek sehen Potenziale, die Forschung zu erklärbarer KI so voranzutreiben, dass Erklärmethoden auch bei Lai*innen zu einem besseren Verständnis von KI-Modellen beitragen. Die Forschung an solchen technischen Lösungen ist jedoch kein isoliert zu betrachtendes Unterfangen. Der Nutzen von Erklärungen ist in hohem Maß kontextabhängig – in Bezug auf die jeweilige Anwendung, deren Einsatzbereich und die Zielgruppe. Diese Erkenntnis eint alle diskutierten Maßnahmen und Ansätze, um Transparenz zu schaffen: Maßnahmen sind dann wirksam, wenn wir sie auf die verschiedenen Kontexte abstimmen, in denen neben der Technik an sich vor allem die beteiligten Menschen und das strukturelle Setting eine Rolle spielen. Deshalb bedarf es verschiedener Maßnahmen, die sich gegenseitig ergänzen. Sie wiederum müssen auf Untersuchungen ihrer Wirksamkeit und ihrer Wechselwirkungen beruhen, die bislang unzureichend existieren oder beachtet werden.

Kontextabhängige Betrachtungen müssen in zukünftigen Auseinandersetzungen eine größere Rolle spielen. Zudem gilt es, darüber nachzudenken, zu welchen Zwecken wir KI-Anwendungen einsetzen sollen und wann es aus ethischen Gründen sinnvoll ist, auf KI-Verfahren zu verzichten. Gleichermaßen müssen wir uns fragen, wann Blackbox-Modelle tatsächlich notwendig sind und an welchen Stellen nachvollziehbarere Whitebox-Modelle zum Einsatz kommen können. Wenn wir die Forderung nach menschenzentrierter KI nicht aus den Augen verlieren wollen, müssen wir diese Fragen beantworten.

Jaana Müller-Brehm		
	Autorin	



WOZU ALL DAS?

Wozu gibt es dieses Magazin? Wozu dient dieses *Zentrum für vertrauenswürdige Künstliche Intelligenz (ZVKI)*? Wir möchten Wissen rund um den großen Themenkomplex Künstliche Intelligenz miteinander verbinden, um neue Erkenntnisse zu gewinnen. Im Fokus stehen dabei wir Menschen, unsere Grundrechte, unser Wohlergehen und unser Zusammenleben. Wir wollen herausfinden, wann die Zuschreibung ‚vertrauenswürdig' gerechtfertigt ist.

Algorithmische Systeme und Verfahren der Künstlichen Intelligenz sind Begriffe aus Fachdiskussionen und zugleich Teil unseres Alltags. Mit ihrer Hilfe werden Entscheidungen getroffen, die Auswirkungen auf unser Leben haben. Um diese Auswirkungen zu erfassen, verständlich darzustellen und mit ihnen umzugehen, müssen wir aus Echokammern ausbrechen, Silodenken hinter uns lassen und Brücken zwischen Insellösungen bauen.

Das *ZVKI* ist ein zentraler Ort der Debatte in Deutschland. Es macht die Entwicklungen rund um gesellschaftliche Fragen zu Künstlicher Intelligenz und algorithmischen Systemen greifbar. Zugleich ist es eine Schnittstelle zwischen Wissenschaft, Wirtschaft, Politik und Zivilgesellschaft, die gemeinsam mit ihren Partner*innen Instrumente entwickelt, um vertrauenswürdige KI zu bewerten.

Die Ziele des *ZVKI* sind unter anderem:

- Informieren, Wissen vermitteln und aufklären: Verständnis ist die Voraussetzung, um Vertrauen aufzubauen. Deshalb bündeln wir Informationen und bereiten sie für verschiedene Zielgruppen auf.
- Forschen und wissenschaftliche Erkenntnisse verständlich darstellen: Wir untersuchen unter anderem, welche Schritte unternommen werden müssen, um negative Auswirkungen von KI-Systemen zu erkennen und ihnen zu begegnen.
- Evaluieren und prüfen: Wir schaffen Konzepte, um zu überprüfen, ob KI-Systeme Kriterien der Vertrauenswürdigkeit entsprechen.
- Zertifizieren: Wir entwickeln Instrumente zur Bewertung von KI und erarbeiten Anforderungen für deren Zertifizierung.
- Netzwerken und unterstützen: Um möglichst viele Stakeholder*innen sowie deren Ansätze und Ideen zusammenzubringen, bieten wir verschiedene Formate des Austauschs an.

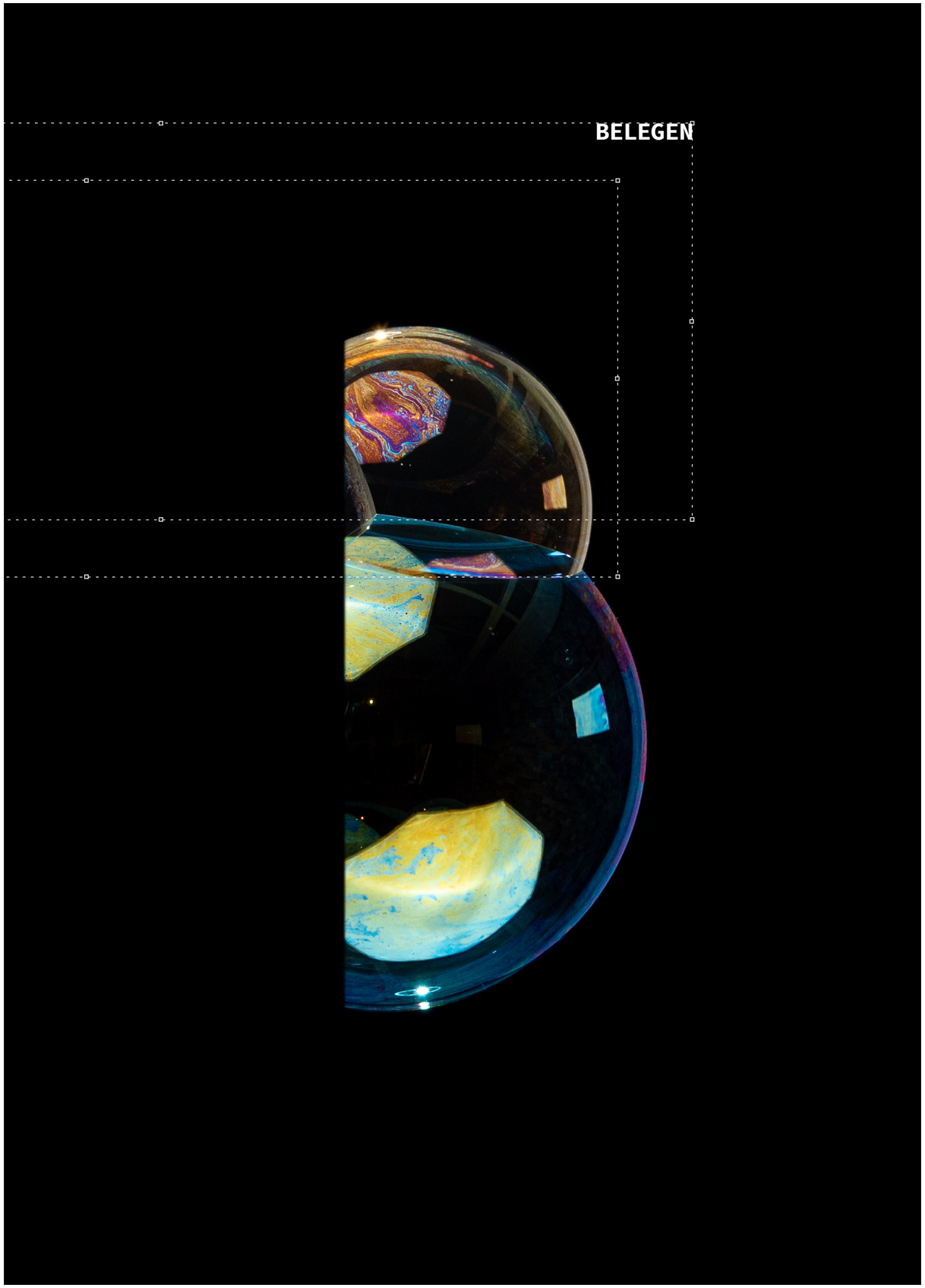
Um diese Ziele zu erreichen, arbeiten wir als interdisziplinäres Team und mit verschiedenen Partner*innen zusammen:

Mit Unterstützung des *Bundesministeriums für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz (BMUV)* baut der unabhängige Think Tank *iRights.Lab*, in Zusammenarbeit mit den *Fraunhofer-Instituten AISEC* und *IAIS* sowie der *Freien Universität Berlin*, das *ZVKI* auf.

Mitmachen

Das *Zentrum für vertrauenswürdige KI – ZVKI* versteht sich als Schnittstelle zwischen Disziplinen und Akteur*innen, zwischen Nutzer*innen und Expert*innen. Treten Sie mit uns in Kontakt und in den Austausch. Sie erreichen uns über *zvki@irights-lab.de*.

Mehr über unsere Aktivitäten und Themen erfahren Sie auf unserer Webseite www.zvki.de.



WOHER STAMMEN DIE INFORMATIONEN?

Quellen

Adler, Rasmus et al.: Abschlussbericht ExamAI – KI Testing und Auditing. Herausforderungen, Lösungsansätze und Handlungsempfehlungen für das Testen, Auditieren und Zertifizieren von KI. 2021. Hg. v. Gesellschaft für Informatik e. V. Online unter: https://testing-ai.gi.de/fileadmin/PR/Testing-AI/Abschlussbericht_ExamAI_-_KI_Testing_und_Auditing.pdf (letzter Aufruf: 26.09.2022).

Asghari, Hadi et al.: What to explain when explaining is difficult? An interdisciplinary primer on XAI and meaningful information in automated decision-making. 2022. Hg. v. Alexander von Humboldt Institute for Internet and Society gGmbH. Online unter: https://doi.org/10.5281/zenodo.6375784 (letzter Aufruf: 26.09.2022).

Beckert, Bernd: Vertrauenswürdige künstliche Intelligenz. Ausgewählte Praxisprojekte und Gründe für das Umsetzungsdefizit. 2021. In: TATuP – Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis, 30/3, S. 17–22. Online unter: https://doi.org/10.14512/tatup.30.3.17 (letzter Aufruf: 26.09.2022).

Berg, Achim/ Dehmel, Susanne: Künstliche Intelligenz. 2020. Hg. v. Bitkom e. V. Online unter: https://www.bitkom-research.de/system/files/document/Bitkom%20Charts%20Künstliche%20Intelligenz%2028%2009%202020_final.pdf (letzter Aufruf: 26.09.2022).

Bietti, Elettra: From ethics washing to ethics bashing: a moral philosophy on tech ethics. 2020. In: Proceedings of ACM FAT* Conference (FAT* 2019), S. 210–219. Hg. v. Association for Computing Machinery. Online unter: https://dl.acm.org/doi/abs/10.1145/3351095.3372860 (letzter Aufruf: 27.09.2022).

Bitkom e. V. / Deutsches Forschungszentrum für Künstliche Intelligenz GmbH: Künstliche Intelligenz. Wirtschaftliche Bedeutung, gesellschaftliche Herausforderungen, menschliche Verantwortung. 2017. Online unter: https://www.dfki.de/fileadmin/user_upload/import/9744_171012-KI-Gipfelpapier-online.pdf (letzter Aufruf: 27.09.2022).

Bundesamt für Sicherheit in der Informationstechnik (BSI): AI Cloud Service Compliance Criteria Catalogue (AIC4). 2021. Online unter: https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/CloudComputing/AIC4/AI-Cloud-Service-Compliance-Criteria-Catalogue_AIC4.pdf?__blob=publicationFile&v=4 (letzter Aufruf: 26.09.2022).

Bundesregierung: Strategie Künstliche Intelligenz der Bundesregierung. Fortschreibung 2020. 2020. Online unter: https://www.ki-strategie-deutschland.de/files/downloads/201201_Fortschreibung_KI-Strategie.pdf (letzter Aufruf: 27.09.2022).

Buolamwini, Joy: Gender Shades. O. D. Hg. v. mit media lab. Online unter: https://www.media.mit.edu/projects/gender-shades/overview/ (letzter Aufruf: 26.09.2022).

Busse, Franziska/ Baeva, Gergana: Was sind die richtigen Zutaten für vertrauenswürdige Künstliche Intelligenz? Ergebnisse der *ZVKI*-Online-Befragung: Wissen, Nachvollziehbarkeit und bewertbare Erfahrungen – Zutaten für vertrauenswürdige Künstliche Intelligenz (KI). Hg. v. Zentrum für vertrauenswürdige Künstliche Intelligenz (*ZVKI*). 2022. Online unter: https://www.*zvki*.de/*zvki*-exklusiv/fachinformationen/online-befragung (letzter Aufruf: 26.09.2022).

Center for Advanced Internet Studies (CAIS): MeMo:KI – Meinungsmonitor Künstliche Intelligenz. Beobachtung der Bevölkerungsmeinung und Berichterstattung über Künstliche Intelligenz in Deutschland. 2022. Online unter https://www.cais.nrw/memoki/#1621436314508-0d370012-5e39 (letzter Aufruf: 26.09.2022).

Chazette, Larissa/ Brunotte, Wasja/ Speith, Timo: Exploring Explainability: A Definition, a Model, and a Knowledge Catalogue. 2021. In: 2021 IEEE 29th international requirements engineering conference (RE), S. 197–208. Online unter: https://doi.org/10.48550/arXiv.2108.03012 (letzter Aufruf: 26.09.2022).

Cremers, Armin B. et al.: Vertrauenswürdiger Einsatz von Künstlicher Intelligenz. Handlungsfelder aus philosophischer, ethischer, rechtlicher und technologischer Sicht als Grundlage für eine Zertifizierung von Künstlicher Intelligenz. Hg. v. Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS. 2019. Online unter: https://www.iais.fraunhofer.de/content/dam/iais/KINRW/Whitepaper_KI-Zertifizierung.pdf (letzter Aufruf: 26.09.2022).

Ehsan, Upol et al.: Expanding Explainability: Towards Social Transparency in AI systems. 2021. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, S. 1–19. Hg. v. Association for Computing Machinery. Online unter: https://doi.org/10.1145/3411764.3445188 (letzter Aufruf: 26.09.2022).

Eticas Research and Consulting SL: Guide to Algorithmic Auditing. 2021. Online unter: https://www.eticasconsulting.com/wp-content/uploads/2021/04/Guide-to-Algorithmic-Auditing-English-Final-ALL-MZ-version-7.pdf (letzter Aufruf: 26.09.2022).

Europäische Kommission: Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für Künstliche Intelligenz (Gesetz über Künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union. 2021. Online unter: https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0019.02/DOC_1&format=PDF6 (letzter Aufruf: 27.09.2022).

Felzmann, Heike et al.: Towards Transparency by Design for Artificial Intelligence. 2020. In: Science and Engineering Ethics, 26, S. 3333–3361. Online unter: https://doi.org/10.1007/s11948-020-00276-4 (letzter Aufruf: 26.09.2022).

Gagrčin, Emilija et al.: We and AI. Living in a Datafied World: Experiences & Attitudes of Young Europeans. 2021. Hg. v. Weizenbaum-Institut e. V. Online unter: https://www.goethe.de/resources/files/pdf249/report_we_and_ai_20212.pdf (letzter Aufruf: 26.09.2022).

Hallensleben, Sebastian et al.: From Principles to Practice: How can we make AI ethics measurable? 2020. Hg. v. Bertelsmann Stiftung. Online unter: https://doi.org/10.11586/2020013 (letzter Aufruf: 26.09.2022).

Healthily: Explainability Statement. 2021. Online unter: https://assets.ctfassets.net/iqo3fk8od6t9/4Sy7OZIAdH65Kl2OmkAG9a/7e7e18ef63e464936b08f5c6cfc3fda7/FINAL_Short_Form_Explainability_Statement__-_17_Sep_2021.pdf (letzter Aufruf: 26.09.2022).

Heesen, Jessica/ Müller-Quade, Jörn/ Wrobel, Stefan et al.: Zertifizierung von KI-Systemen – Kompass für die Entwicklung und Anwendung vertrauenswürdiger KI-Systeme. Whitepaper. 2020. Online unter: https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG1_3_Whitepaper_Zertifizierung_KI_Systemen.pdf (letzter Aufruf: 26.09.2022).

High-Level Expert Group on AI: Ethics Guidelines for Trustworthy AI. 2019. Hg. v. Europäische Kommission. Online unter: https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html (letzter Aufruf: 26.09.2022).

Holzinger, Andreas et al.: Explainable AI Methods – A Brief Overview. 2022. In: Holzinger, Andreas et al.: xxAI – Beyond Explainable AI, S. 13–38. Online unter: https://doi.org/10.1007/978-3-031-04083-2_2 (letzter Aufruf: 26.09.2022).

IMAGINARY gGmbH: An automatic speech recognition exhibit. O. D. Online unter: https://www.imaginary.org/de/node/2345 (letzter Aufruf: 26.09.2022).

IMAGINARY gGmbH: talk-to-me. 2020. Online unter: https://github.com/IMAGINARY/talk-to-me (letzter Aufruf: 26.09.2022).

Institut der Wirtschaftsprüfer in Deutschland (IDW): Entwurf eines IDW Prüfungsstandards: Prüfung von KI-Systemen (IDW EPS 861 (02.2022)). 2022. Online unter: https://www.idw.de/blob/134852/bf9349774314723f6246ba-

73fefc491f/idw-eps-861-02-2022-data.pdf (letzter Aufruf: 26.09.2022).

Jobin, Anna/ Ienca, Marcello/ Vayena, Effy: The global landscape of AI ethics. 2019. In: Nature Machine Intelligence, 1, S. 389-399. Online unter: https://doi.org/10.1038/s42256-019-0088-2 (letzter Aufruf: 27.09.2022).

Jung, Jae-Yoon/ Park, Donghyun: Chapter 1 - Are AI models explainable, interpretable, and understandable? 2022. In: Nam, Chang S./ Jung, Jae-Yoon/ Lee, Sangwon: Human-Centered Artificial Intelligence, S. 3-16. Online unter: https://doi.org/10.1016/B978-0-323-85648-5.00003-7 (letzter Aufruf: 26.09.2022).

Kästner, Lena et al.: On the Relation of Trust and Explainability: Why to Engineer for Trustworthiness. 2021. In: Yue, Tao/ Mirakhorli, Mehdi: IEEE 29th International Requirements Engineering Conference Workshops (REW), S. 169-175. Hg. v. The Institute of Electrical and Electronics Engineers. Online unter: https://doi.org/10.1109/REW53955.2021.00031 (letzter Aufruf: 27.09.2022).

Kieslich, Kimon: MeMo:KI - Factsheet Nr. 2 - August 2020. Künstliche Intelligenz und Diskriminierung. 2020. Hg. v. Center for Advanced Internet Studies (CAIS). Online unter: https://www.cais.nrw/wp-94fa4-content/uploads/2020/08/Factsheet-2-KI-und-Diskriminierung.pdf (letzter Aufruf: 27.09.2022).

Knobloch, Tobias/ Hustedt, Carla: Der maschinelle Weg zum passenden Personal. Zur Rolle algorithmischer Systeme in der Personalwahl. 2019. Hg. v. Stiftung Neue Verantwortung e. V./ Bertelsmann Stiftung. Online unter: https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/SNV_Robo_Recruiting_final.pdf (letzter Aufruf: 27.09.2022).

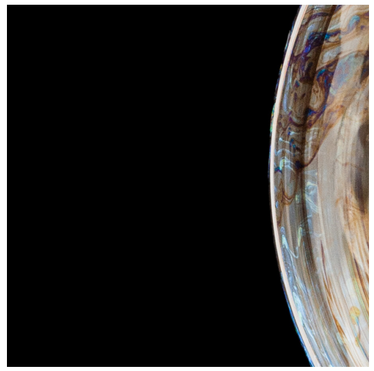
Kolain, Michael/ Baeva, Gergana/ Buchsbaum, Katharina: Wie können Regulierung und Standards zu vertrauenswürdiger KI beitragen? Hg. v. Zentrum für vertrauenswürdige Künstliche Intelligenz (*ZVKI*). 2022. Online unter: https://www.zvki.de/zvki-exklusiv/fachinformationen/essay-regulierung-standards (letzter Aufruf: 27.09.2022).

Kraus, Tom et al.: Erklärbare KI. Anforderungen, Anwendungsfälle und Lösungen. 2021. Hg. v. Technologieprogramm KI-Innovationswettbewerb des Bundesministeriums für Wirtschaft und Energie Begleitforschung/ iit-Institut für Innovation und Technik in der VDI/VDE Innovation + Technik GmbH. Online unter: https://www.digitale-technologien.de/DT/Redaktion/DE/Downloads/Publikation/KI-Inno/2021/Studie_Erklaerbare_KI.pdf;jsessionid=C94CE1AF540F11B6D64E3EB19A61A2AF?__blob=publicationFile&v=9 (letzter Aufruf: 27.09.2022).

Kraus, Tom/ Ganschow, Lene: Anwendungen und Lösungsansätze erklärbarer Künstlicher Intelligenz. 2021. In: Hartmann, Ernst A.: Digitalisierung souverän gestalten II, S. 38-50. Online unter: https://doi.org/10.1007/978-3-662-64408-9_4 (letzter Aufruf: 27.09.2022).

Laato, Samuli et al.: How to explain AI systems to end users: a systematic literature review and research agenda. 2022. In: Internet Research, 32/7, S. 1-31. Online unter: https://doi.org/10.1108/INTR-08-2021-0600 (letzter Aufruf: 27.09.2022).

Langer, Markus et al.: What Do We Want From Explainable Artificial Intelligence (XAI)? - A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research. 2021. Online unter:



https://doi.org/10.48550/arXiv.2102.07817 (letzter Aufruf: 16.09.2022).

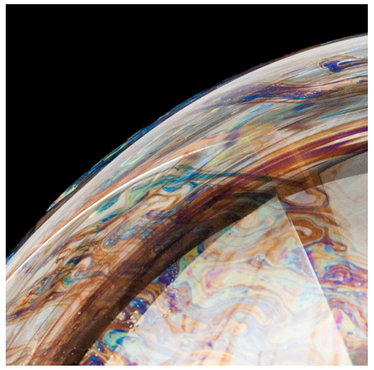
Lapuschkin, Sebastian/ Samek, Wojciech: From “Where” to “What”: Towards Human-Understandable Explanations through Concept Relevance Propagation. 2022. Online unter: https://doi.org/10.48550/arXiv.2206.03208 (letzter Aufruf: 27.09.2022).

Larsson, Stefan/ Heintz, Frederik: Transparency in artificial intelligence. 2020. In: Internet Policy Review, 9/2, S. 1-16. Online unter: https://doi.org/10.14763/2020.2.1469 (letzter Aufruf: 27.09.2022).

Lossos, Christian/ Geschwill, Simon/ Morelli, Frank: Offenheit durch XAI bei ML-unterstützten Entscheidungen: Ein Baustein zur Optimierung von Entscheidungen im Unternehmen? 2021. In: HMD Praxis der Wirtschaftsinformatik, 58, S. 303-320. Online unter: https://doi.org/10.1365/s40702-021-00707-1 (letzter Aufruf: 27.09.2022).

Mangelsdorf, Axel/ Gabriel, Peter/ Weimer, Martin: Die Zertifizierung von KI: Mehr Sicherheit für alle - oder unnötiger Ballast? 2021. Hg. v. Volker Wittpahl, Institut für Innovation und Technik (iit) in der VDI/VDE Innovation + Technik GmbH. Online unter: https://www.iit-berlin.de/wp-content/uploads/2021/04/2021_04_30_iit-perspektive_Nr-58_Zertifizierung_von_KI.pdf (letzter Aufruf: 27.09.2022).

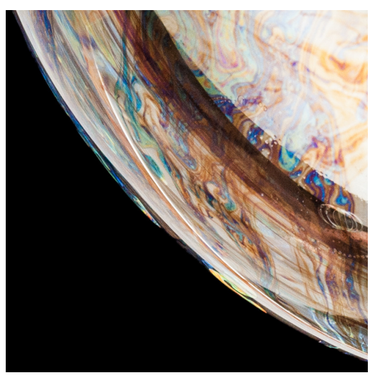
Mock, Michael et al.: Management System Support for Trustworthy Artificial Intelligence. A comparative study. 2021. Hg. v. Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS. Online unter: https://www.iais.fraunhofer.de/content/dam/iais/fb/Kuenstliche_intelligenz/Fraunhofer_IAIS_Study_%20MSS.pdf (letzter Aufruf: 27.09.2022).



Nativi, Stefano/ De Nigris, Sarah: AI Watch, AI standardisation landscape state of play and link to the EC proposal for an AI regulatory framework. 2021. Hg. v. der Europäischen Kommission, Gemeinsame Forschungsstelle. Online unter: https://op.europa.eu/de/publication-detail/-/publication/36c46b8e-e518-11eb-a1a5-01aa75ed71a1/language-en (letzter Aufruf: 27.09.2022).

Oehler, Andreas: Verbraucherinformation und Verbraucherbildung. 2021. In: Kenning, Peter et al.: Verbraucherwissenschaften, S. 259-273. Online unter: https://www.springerprofessional.de/verbraucherinformation-und-verbraucherbildung/18909056 (letzter Aufruf: 16.08.2022).

Overdiek, Markus/ Petersen, Thomas: Was Deutschland über Algorithmen und Künstliche Intelligenz weiß und denkt. Ergebnisse einer repräsentativen Bevölkerungsumfrage. Update 2022. 2022. Hg. v. Bertelsmann Stiftung. Online unter: https://www.bertelsmann-stiftung.de/de/publikationen/publikation/did/



was-deutschland-ueber-algorithmen-und-kuenstliche-intelligenz-weiss-und-denkt-all (letzter Aufruf: 27.09.2022).

Poretschkin, Maximilian et. al.: Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz. KI-Prüfkatalog. 2021. Hg. v. Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS. Online unter: https://www.iais.fraunhofer.de/content/dam/iais/fb/Kuenstliche_intelligenz/ki-pruefkatalog/202107_KI-Pruefkatalog.pdf (letzter Aufruf: 27.09.2022).

Ricks, Rebecca et al.: Creating Trustworthy AI. A Mozilla white paper on

challenges and opportunities in the AI era. 2020. Hg. v. Mozilla Corporation. Online unter: https://foundation.mozilla.org/en/insights/trustworthy-ai-white-paper/ (letzter Aufruf: 27.09.2022).

Robert Bosch GmbH: So denkt Deutschland über künstliche Intelligenz. Der Bosch KI-Zukunftskompass 2020. 2020. Online unter: https://www.bosch.de/news-and-stories/ki-zukunftskompass/ (letzter Aufruf: 27.09.2022).

Rohde, Friederike et al.: Nachhaltigkeitskriterien für künstliche Intelligenz. Entwicklung eines Kriterien- und Indikatorensets für die Nachhaltigkeitsbewertung von KI-Systemen entlang des Lebenszyklus. 2021. IÖW-Schriftenreihe, 220/21. Online unter: https://www.ioew.de/publikation/nachhaltigkeitskriterien_fuer_kuenstliche_intelligenz (letzter Aufruf: 27.09.2022).

Rolls Royce: The Aletheia Framework 2.0. 2021. Online unter: https://www.rolls-royce.com/~/media/Files/R/Rolls-Royce/documents/stand-alone-pages/aletheia-framework-worksheet.pdf (letzter Aufruf: 27.09.2022).

Sachverständigenrat für Verbraucherfragen beim Bundesministerium der Justiz und für Verbraucherschutz: Gutachten 2021. Gutachten zur Lage der Verbraucherinnen und Verbraucher. 2021. Online unter: https://www.svr-verbraucherfragen.de/wp-content/uploads/SVRV_Gutachten_2020.pdf (letzter Aufruf: 27.09.2022).

Schaaf, Nina/ Wiedenroth, Saskia Johanna/ Wagner, Philipp: Erklärbare KI in der Praxis. Anwendungsorientierte Evaluation von XAI-Verfahren. 2021. Hg. v. Bauernhansl, Thomas/ Huber, Marco/ Kraus, Werner. Online unter: https://www.ki-fortschrittszentrum.de/de/studien/erklaerbare-ki-in-der-praxis.html (letzter Aufruf: 27.09.2022).

Schmid, Ute: Explainable AI in der Medizin. 2021. Hg. v. science media center germany. Online unter: https://www.sciencemediacenter.de/alle-angebote/research-in-context/details/news/explainable-ai-in-der-medizin/ (letzter Aufruf: 27.09.2022).

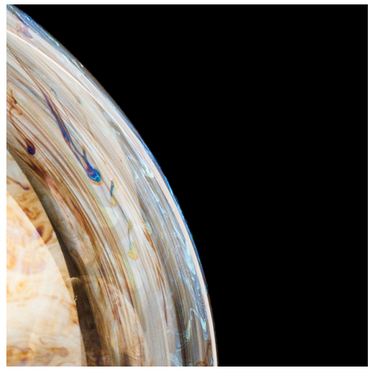
Schultz, Stefan: Maas hätte gerne, dass Google geheime Suchformel offenlegt. 2014. In: Spiegel Online, 16.09.2014. https://www.spiegel.de/wirtschaft/unternehmen/google-heiko-maas-fordert-offenlegung-von-algorithmus-a-991799.html (letzter Aufruf: 27.09.2022).

Shin, Donghee: The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. 2021. In: International Journal of Human-Computer Studies, 146, Art. 102551. Online unter: https://www.sciencedirect.com/science/article/abs/pii/S1071581920301531 (letzter Aufruf: 27.09.2022).

Timm, Moritz: International researchers call for accountability and transparency regarding algorithmic content moderation. 2020. Hg. v. Alexander von Humboldt Institut für Internet und Gesellschaft gGmbH. Online unter: https://www.hiig.de/en/international-researchers-call-for-accountability-and-transparency-regarding-algorithmic-content-moderation/ (letzter Aufruf: 16.08.2022).

TÜV-Verband e. V.: Sicherheit und Künstliche Intelligenz. Erwartungen, Hoffnungen, Risiken. Eine repräsentative Befragung der Bevölkerung in Deutschland im Auftrag des TÜV-Verbands. 2021. Online unter: https://www.tuev-verband.de/?tx_epxelo_file[id]=856779&cHash=1af8a3f0e6c845423fdd637c8dbcd080 (letzter Aufruf: 27.09.2022).

UNESCO: Draft text of the Recommendation on the Ethics of Artificial Intelligence. 2021. Online unter: https://unesdoc.unesco.org/ark:/48223/pf0000377897 (letzter Aufruf: 27.09.2022).



VDE Verband der Elektrotechnik Elektronik Informationstechnik e. V.: VCIO based description of systems for AI trustworthiness characterisation. VDE SPEC 90012 V1.0 (en). 2022. Online unter: https://www.vde.com/resource/blob/2177870/a24b13db01773747e6b7bba4ce20ea60/vde-spec-90012-v1-0--en--data.pdf (letzter Aufruf: 27.09.2022).

VDE Verband der Elektrotechnik Elektronik Informationstechnik e. V.: Kann Künstliche Intelligenz wertekonform sein? VDE SPEC als Grundlage künftiger Entwicklungen. 2022. Online unter: https://www.vde.com/de/presse/pressemitteilungen/ai-trust-label (letzter Aufruf: 27.09.2022).

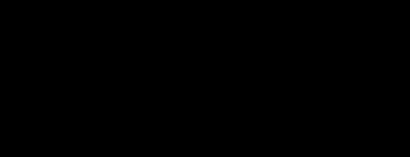
Walmsley, Joel: Artificial intelligence and the value of transparency. 2021. In: AI & Society, 36, S. 585-595. Online unter: https://doi.org/10.1007/s00146-020-01066-z (letzter Aufruf: 27.09.2022).

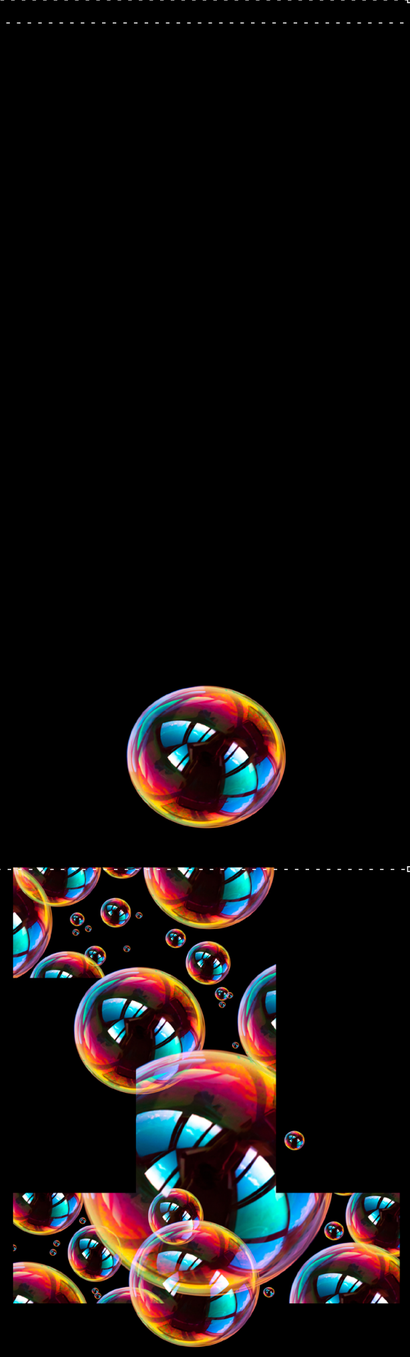
Waltl, Bernhard: Erklärbarkeit und Transparenz im Machine Learning. 2020. In: Mainzer, Klaus: Philosophisches Handbuch Künstliche Intelligenz, S. 1-23. Online unter: https://doi.org/10.1007/978-3-658-23715-8_31-1 (letzter Aufruf: 27.09.2022).

Weiss, Erik: Zertifizierung - Ein zentraler Baustein auf dem Weg zu einem vertrauenswürdigen Einsatz Künstlicher Intelligenz. 2022. Hg. v. libra. Das Rechtsbriefing. https://www.libra-rechtsbriefing.de/L/ki-zertifizierung/ (letzter Aufruf: 27.09.2022).

Wirth, Christian/ Schmid, Ute/ Voget, Stefan: Humanzentrierte Künstliche Intelligenz: Erklärendes interaktives maschinelles Lernen für Effizienzsteigerung von Parametrieraufgaben. 2021. In: Hartmann, Ernst A.: Digitalisierung souverän gestalten II, S. 80-92. Online unter: https://doi.org/10.1007/978-3-662-64408-9_7 (letzter Aufruf: 27.09.2022).

IMPRESSUM





iRights.Lab GmbH
Oranienstr. 185
D-10999 Berlin
Telefon: +49 (0)30 40 36 77 230
Fax: +49 (0)30 40 36 77 260
E-Mail: *zvki*@irights-lab.de

Geschäftsführer*in: **Philipp Otto, Dr. Wiebke Glässer**
Registergericht: Amtsgericht Berlin-Charlottenburg
Registernummer: HRB 185640 B
Finanzamt für Körperschaften II
USt-IdNr.: DE311181302

Projektleitung: **Philipp Otto**

Projektkoordination: **Verena Till**

Autorinnen: **Dr. Gergana Baeva, Franziska Busse, Jaana Müller-Brehm, Lisa Schmechel**

Chefredaktion: **Jaana Müller-Brehm**

Redaktion: **Vera Dünninger, Philipp Otto, Verena Till**

Inhaltliche Mitarbeit: **Raphael Hadadi, Michael Puntschuh**

Gestalterische Konzeption und Layout: **Christoph Löffler**

Lektorat: text|struktur

Die Online-Version von *Missing Link* und weitere Informationen zum Projekt *ZVKI* finden Sie unter:

www.*zvki*.de

www.instagram.com/*zvki*.de

www.twitter.com/*ZVKI*_de.

Verantwortung und Durchführung: iRights.Lab

Verbundpartner*innen: Fraunhofer AISEC, Fraunhofer IAIS, Freie Universität Berlin

Gefördert durch: *Bundesministerium für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz (BMUV)*

Das Projekt *ZVKI* wird vom unabhängigen Think Tank *iRights.Lab* verantwortet und durchgeführt. Das *iRights.Lab* entwickelt Strategien und praktische Lösungen, um die Veränderungen in der digitalen Welt vorteilhaft zu gestalten. Wir unterstützen öffentliche Einrichtungen, Stiftungen, Unternehmen, Wissenschaft und Politik dabei, die Herausforderungen der Digitalisierung zu meistern und die vielschichtigen Potenziale effektiv und positiv zu nutzen.

Weitere Informationen über das *iRights.Lab* finden Sie unter www.irights-lab.de.

Gefördert durch:

Bundesministerium
für Umwelt, Naturschutz, nukleare Sicherheit
und Verbraucherschutz

aufgrund eines Beschlusses
des Deutschen Bundestages

iRights.Lab
Think Tank für die
digitale Welt

ISSN (Online) 2751-2207