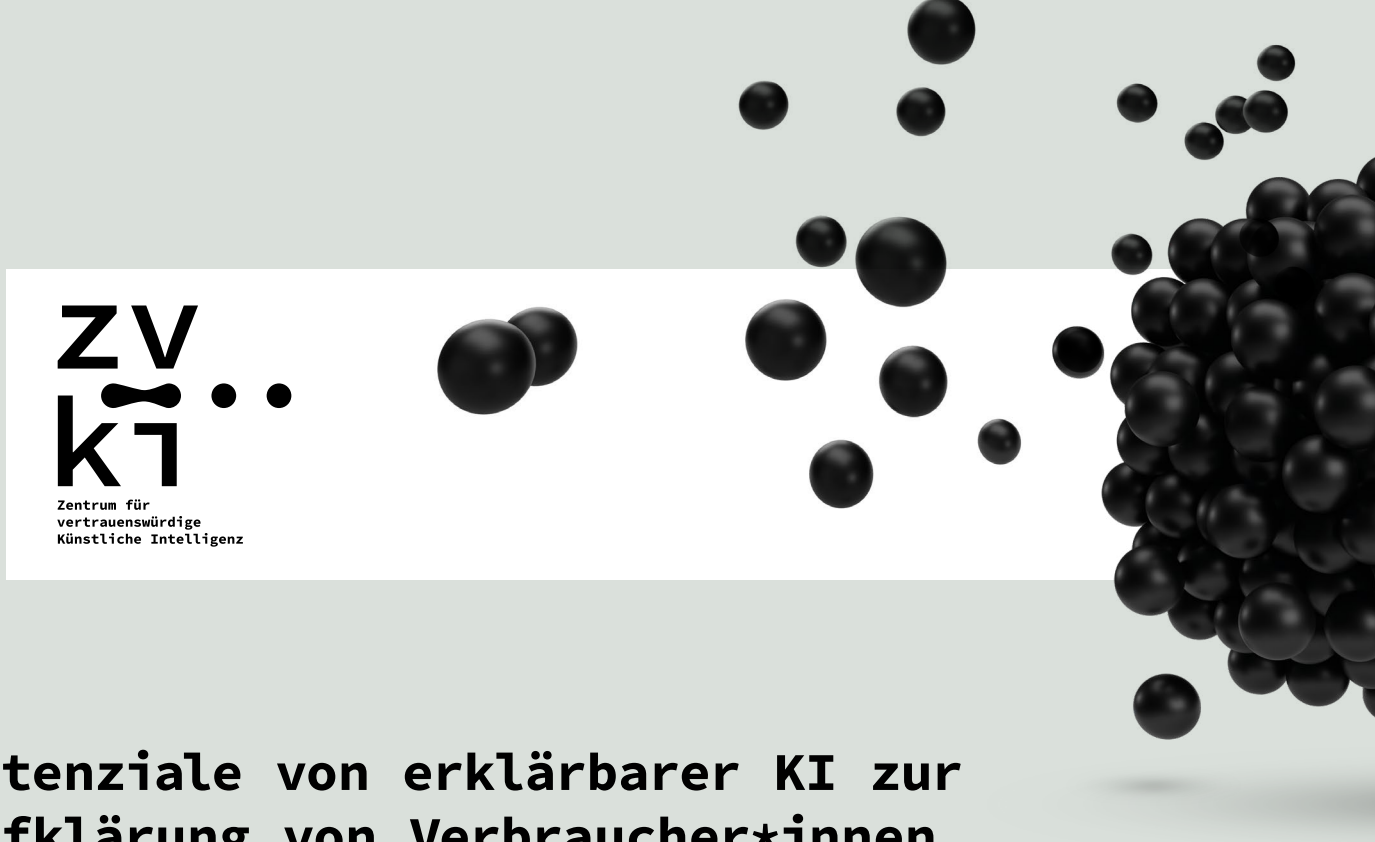

Zentrum für
vertrauenswürdige
Künstliche Intelligenz

Potenziale von erklärbarer KI zur Aufklärung von Verbraucher*innen

ZVKI-Facharbeitsgruppe Verbraucher*innen-Information

Thesenpapier, 2023

Zusammenfassung

Sind Methoden der erklärbaren KI (XAI-Methoden) hilfreich für die Aufklärung von Verbraucher*innen? Mit dieser Fragestellung befassten sich Expert*innen aus Wissenschaft und Zivilgesellschaft im Rahmen der Fach-Arbeitsgruppe Verbraucher*innen-Information des *Zentrums für vertrauenswürdige Künstliche Intelligenz (ZVKI).* Sie stellten mehrere Thesen auf, die zeigen, dass sich XAI-Methoden unter bestimmten Bedingungen eignen, um die Ergebnisse eines KI-Systems für Verbraucher*innen nachvollziehbar zu machen.

Dazu müssen XAI-Methoden niedrigschwellig aufbereitet sein und je nach Kontext um weitere Erklärformate ergänzt werden. Erklärungen allein reichen nicht aus, um die Vertrauenswürdigkeit von KI-Systemen zu fördern. Vielmehr sind weitere Maßnahmen notwendig, beispielsweise sollten Verbraucher*innen realisierbare Handlungsempfehlungen aufgezeigt werden. Zudem bedarf es unabhängiger Beschwerdestellen.

Unternehmen – als KI-Anbieter*innen – sind für die Umsetzung und Nachvollziehbarkeit von Erklärmaßnahmen hauptverantwortlich. Wissenschaftler*innen sollten weiterhin an zuverlässigen XAI-Methoden forschen. Dabei sollten Forschungsvorhaben einen interdisziplinären Ansatz verfolgen, um nicht rein technische, sondern am Menschen orientierte Erklärungsansätze zu erarbeiten. Politische Entscheidungsträger*innen sind in der Pflicht, ausreichend Ressourcen für die Verbesserung und Entwicklung solcher Methoden bereitzustellen.

Unter welchen Bedingungen eignen sich XAI-Methoden zur Aufklärung von Verbraucher*innen?

KI-Systeme sind häufig so komplex, dass Entwickler*innen und Nutzer*innen deren Ergebnisse und Vorgehensweisen nicht mehr nachvollziehen können. Im Forschungsbereich der erklärbaren KI (Explainable AI oder XAI) forschen Wissenschaftler*innen daher an Erklärmethoden, die nachträglich zeigen, welche Daten oder Annahmen zu einem bestimmten Ergebnis einer KI-Anwendung geführt haben. Erklärbare KI richtet sich derzeit vor allem an Entwickler*innen und Fach-Expert*innen. Doch auch Verbraucher*innen müssen die Ergebnisse von KI-Systemen nachvollziehen können – insbesondere, wenn sie von diesen direkt oder indirekt betroffen sind. Dies gilt beispielsweise für KI-Systeme, die Unternehmen als Unterstützung bei der Kreditvergabe einsetzen.

In der ZVKI-Fach-Arbeitsgruppe Verbraucher*innen-Information haben wir uns deshalb mit der Frage auseinandergesetzt, ob Methoden der erklärbaren KI hilfreich für die Aufklärung von Verbraucher*innen sind.

Zur Beantwortung der Fragestellung haben wir in mehreren Sitzungen der Fach-AG sieben Thesen aufgestellt. Diese beziehen sich vor allem auf die individuelle Ebene der Verbraucher*innen-Aufklärung. Unternehmen, zivilgesellschaftliche Organisationen oder Entwickler*innen können einzelnen Nutzer*innen mithilfe von XAI-Methoden Informationen über die Ergebnisse von KI-Systemen bereitstellen und diese damit zum Handeln befähigen. Die Voraussetzung hierfür ist jedoch, dass der Einsatz von Erklärmethoden mit weiteren Maßnahmen wie dem Hinweis auf Handlungsoptionen oder Beschwerdemöglichkeiten verknüpft ist.

Im Fokus der Fach-AG standen grundlegende Potenziale von erklärbarer KI für die Verbraucher*innen-Aufklärung. Die Möglichkeiten eines ethisch und juristisch vertretbaren Einsatzes von KI-Anwendungen zu bestimmten Zwecken thematisierten wir nicht.

Diskutierte Erklärmethoden

Die Fragestellung der Fach-AG haben wir anhand von *Saliency Maps* und *Counterfactual Explanations* diskutiert. Sowohl bei *Saliency Maps* als auch bei *Counterfactual Explanations* handelt es sich um Darstellungen lokaler XAI-Methoden. Sie zeigen, aufgrund welcher spezifischen Eingabefaktoren das KI-System zu einem bestimmten Ergebnis kommt.

Saliency Map

Eine *Saliency Map* ist eine Darstellung, die visualisiert, welche Faktoren in einer Eingabe für das Ergebnis eines KI-Systems entscheidend waren. Die zugrunde liegende XAI-Methode eignet sich vor allem für die Klassifizierung von Bild-Daten, beispielsweise, wenn eine KI-Anwendung Bilder mit Katzen von Bildern ohne Katzen unterscheiden soll. Als Erklärung werden bei einem Bild, das von einem KI-System als Katzenbild eingeordnet wurde, diejenigen Pixel farblich hervorgehoben, die für diese Einordnung entscheidend waren. Bei einem vertrauenswürdigen KI-System sollten das für den Menschen nachvollziehbare Merkmale, zum Beispiel die Augen oder Ohren, sein.

Counterfactual Explanation

Eine *Counterfactual Explanation* beschreibt die Umstände, unter denen das Ergebnis einer KI-Anwendung anders ausgefallen wäre. Nehmen wir als Beispiel den Einsatz eines KI-Systems bei einem Bewerbungsverfahren: Ein Unternehmen nutzt ein KI-System, um eingehende Bewerbungen hinsichtlich ihrer Eignung für eine Stelle zu sortieren. Die Bewerbung einer Person wird durch die KI-Anwendung als ungeeignet eingestuft. Eine *Counterfactual Explanation* analysiert, welche Faktoren sich ändern müssten, damit das KI-System ein positives Ergebnis für diese*n Bewerber*in ausgibt. Das kann zum Beispiel die Berufserfahrung sein: Hätte die Person zehn statt neun Jahre Berufserfahrung vorweisen können, hätte die KI-Anwendung als Ergebnis eine Einstellungsempfehlung ausgegeben.

These 1

XAI-Methoden sind grundsätzlich geeignet, um die Ergebnisse eines KI-Systems für Verbraucher*innen nachvollziehbar zu machen.

Bei der Forschung zu XAI-Methoden steht die Aufklärung von Verbraucher*innen selten im Vordergrund. Dennoch können Erklärmethoden auch dieser Gruppe helfen, die Ergebnisse von KI-Anwendungen nachzuvollziehen. Das ist insbesondere dann von Relevanz, wenn diese einen Einfluss auf die Lebensrealität der Verbraucher*innen haben. Werden KI-Anwendungen beispielsweise zur Unterstützung bei der Kreditvergabe eingesetzt, können *Counterfactual Explanations* den konkreten Einfluss bestimmter Faktoren auf das Ergebnis eines KI-Systems darstellen. Betroffene können unter anderem nachvollziehen, ob ihr Gehalt, Wohnort oder Alter zu einer Kreditabsage geführt haben. Der Einsatz solcher Erklärmethoden für die Aufklärung von Verbraucher*innen ist jedoch nur unter bestimmten Bedingungen, die wir im Folgenden ausführen, sinnvoll.

These 2

Die Erklärungen durch XAI-Methoden müssen niedrigschwellig aufbereitet sein und mit Nutzer*innen getestet werden.

Wenn der Einsatz einer *Counterfactual Explanation* zur besseren Nachvollziehbarkeit eines Kreditvergabeverfahrens einen Mehrwert für Verbraucher*innen haben soll, muss die Erklärung so aufbereitet sein, dass Verbraucher*innen diese tatsächlich verstehen und nutzen können. Ein Beispiel wäre ein interaktives, digitales Eingabeformular, in dem die Nutzer*innen die Eingabedaten wie Alter, Wohnort oder Gehalt beliebig ändern können. Auf diese Weise können sie nachvollziehen, welche Faktoren Einfluss auf das Ergebnis der KI-Anwendung und damit auf die Chancen zur Gewährung eines Kredits haben – ohne dass sie Fachwissen zur Funktionsweise der KI-Anwendungen besitzen müssen. Um einen Mehrwert der Erklärungen für Verbraucher*innen sicherzustellen, müssen diese mit Nutzer*innen evaluiert werden.

These 3

Zusätzlich zu den Erklärungen sollten den Verbraucher*innen realisierbare Handlungsempfehlungen aufgezeigt werden.

XAI-Methoden an sich geben Verbraucher*innen keine ausreichenden Handlungsoptionen. Daher sollten Anbieter*innen von KI-Systemen Verbraucher*innen ergänzend zu einer Erklärung darüber informieren, was sie selbst ändern können, damit das KI-System ein anderes Ergebnis ausgibt. Entscheidend ist dabei, dass die Handlungsempfehlung realisierbar ist.

Im Fall der Kreditvergabe kann es für Bewerber*innen beispielsweise hilfreich sein, zu wissen, ob bereits eine geringfügige Gehaltssteigerung ausreicht, damit sie einen Kredit gewährt bekommen. Die Empfehlung, das eigene Alter zu ändern, ergibt dagegen wenig Sinn. Die Umsetzung von Handlungsempfehlungen ist nicht allen Menschen im gleichen Maß möglich, sondern hängt von gesamtgesellschaftlichen Faktoren und Machtstrukturen ab: So ist es je nach Beruf und Anstellungsverhältnis beispielsweise nur manchen Menschen möglich, ein höheres Gehalt auszuhandeln.

These 4

Erklärbarkeit für Verbraucher*innen benötigt Korrigierbarkeit.

Neben realisierbaren Handlungsempfehlungen sollten Verbraucher*innen die Eingaben und/ oder Ergebnisse von KI-Anwendungen korrigieren können. Die Korrigierbarkeit betrifft drei verschiedene Ebenen:

1. Verbraucher*innen sollten falsche Eingaben richtigstellen können. Das betrifft zum Beispiel die Eintragung eines falschen Alters bei einem KI-System, das eine Bank als Unterstützung bei der Kreditwürdigkeitsprüfung einsetzt. Erklärungen durch XAI-Methoden können solche falschen Eingabedaten offenlegen und eine Korrektur ermöglichen.

2. Nutzer*innen sollten die Ergebnisse einer KI-Anwendung personalisieren können, beispielsweise beim News Feed einer Social-Media-Anwendung. Eine mögliche Personalisierung erhöht die Benutzer*innen-Freundlichkeit solcher Anwendungen.

3. Verbraucher*innen sollte es möglich sein, sich über die Ergebnisse eines KI-Systems bei unabhängigen Stellen zu beschweren, beispielsweise bei einem Verdacht auf Diskriminierung durch eine Gesichtserkennungssoftware oder bei einer Entscheidung zur Kreditvergabe. An dieser Stelle wird deutlich, warum Erklärungen durch XAI-Methoden möglichst umfassend sein sollten und sich nicht nur auf die Faktoren beschränken dürfen, die die Nutzer*innen selbst beeinflussen können (vgl. These 3): Auch wenn Verbraucher*innen Eingabefaktoren – etwa zur Kreditvergabe – nicht ändern können, haben sie dennoch grundsätzlich die Möglichkeit, eine Benachteiligung aufzudecken und zu melden.

Ergänzend möchten wir an dieser Stelle anmerken, dass sich strukturelle Diskriminierungen in den Ergebnissen von KI-Systemen in der Regel erst zeigen, wenn größere Datenmengen untersucht werden. Für einzelne Nutzer*innen ist es schwierig, abzuschätzen, ob ein Kredit aufgrund individueller Gründe oder struktureller Diskriminierung abgelehnt wurde. Daher ist entscheidend, dass auch Forschungsinstitute, Antidiskriminierungsstellen oder Nichtregierungsorganisationen Zugang zu den Datengrundlagen von KI-Systemen und Erklärungen durch XAI-Methoden erhalten. Nur so können zivilgesellschaftliche Organisationen ihre gesellschaftliche Kontrollfunktion wahrnehmen.

These 5

Erklärbarkeit und Interpretierbarkeit sind für einfache KI-Modelle leichter umzusetzen.

Neben Erklärmethoden ist der Einsatz von sogenannten *White-Box-Modellen* eine Möglichkeit, um die Nachvollziehbarkeit von KI-Anwendungen zu erhöhen. Bei *White-Box-Modellen* programmieren Entwickler*innen KI-Anwendungen von Beginn an so, dass deren Ergebnisse leicht nachvollziehbar sind. Zur Interpretation und Erklärung solcher Modelle werden daher keine XAI-Methoden benötigt, was auch eine Vermittlung an Verbraucher*innen erleichtert. Wo immer es möglich ist, sollten Unternehmen von Beginn an inhärent nachvollziehbare KI-Systeme nutzen.

These 6

XAI-Methoden und Erklärformate müssen sich je nach Anwendung und Kontext unterscheiden.

Je nach Anwendungsfall kann eine Kombination von mehreren XAI-Methoden und Darstellungen hilfreich sein, um Nachvollziehbarkeit herzustellen. Im Fall der Kreditvergabe kann beispielsweise eine *Saliency Map* einen allgemeinen Überblick darüber geben, welche Eingabefaktoren für das Ergebnis (Kreditvergabe) relevant sind. Klicken Nutzer*innen einen Faktor an, zeigen *Counterfactuals,* welchen Einfluss dieser Faktor im jeweiligen Fall auf das Ergebnis der KI-Anwendung hat.

Zudem können manche Informationsbedarfe der Verbraucher*innen besser durch andere Formate als durch Darstellungen von XAI-Methoden abgedeckt werden: Informationen zu den gespeicherten individuellen Daten der Nutzer*innen einer Gesundheits-App könnten Unternehmen beispielsweise als grafisch aufbereitete Übersicht bereitstellen. Für eine solche Darstellung sind keine XAI-Methoden notwendig.

Unternehmen sind für die Umsetzbarkeit und Nachvollziehbarkeit der Erklärmaßnahmen verantwortlich. Sie müssen einen möglichst einfachen Zugang zu unterschiedlichen Erklärungen durch XAI-Methoden schaffen und zudem begründen, warum sie ein bestimmtes KI-Modell verwenden. Unternehmen müssen ermöglichen, dass ihre KI-Anwendungen geprüft werden können und KI-Anwendungen als solche kennzeichnen.

These 7

XAI-Methoden zeigen keine allgemeingültige Wahrheit.

Je nach Anwendungsfall und Erklärmethode können verschiedene XAI-Methoden zu widersprüchlichen Erklärungen kommen. Verbraucher*innen müssen daher darüber informiert werden, dass auch XAI-Methoden keine allgemeingültige Wahrheit zeigen.

Wissenschaftler*innen sollten weiterhin an zuverlässigen XAI-Methoden forschen. Dabei sollten Forschungsvorhaben einen interdisziplinären Ansatz verfolgen, um nicht nur rein technische, sondern am Menschen orientierte Erklärungsansätze zu erarbeiten. Politische Entscheidungsträger*innen sind in der Pflicht, ausreichend Forschungsgelder und Förderungen für die Verbesserung und Entwicklung von solchen Methoden bereitzustellen.

Informationen zum Thesenpapier

Die Fach-AG Verbraucher*innen-Information

Das Thesenpapier ist das Ergebnis der Fach-AG Verbraucher*innen-Information des *ZVKI*. Die zwischen Juni und November 2022 monatlich stattfindenden Sitzungen leiteten Franziska Busse (*iRights. Lab*) und Prof. Dr. Eirini Ntoutsi (*Universität der Bundeswehr München*).

Die teilnehmenden Expert*innen waren:

- Dr. Miika Blinn, Verbraucherzentrale Bundesverband
- Yi Cai, Freie Universität Berlin
- Hania Elkersh, Freie Universität Berlin
- Dr. Julia Gerhards, Verbraucherzentrale Rheinland-Pfalz e.V.
- Dr. Alexander Goschew, DIN-Verbraucherrat
- Linus Henning, Freie Universität Berlin
- Dr. Malte Henningsen, Freie Universität Berlin
- Prof. Dr. Steffen Kroschwald, Hochschule Pforzheim
- Valerie Krug, AI Lab, Otto-von-Guericke-Universität Magdeburg
- Prof. Dr. Ute Schmid, Otto-Friedrich-Universität Bamberg
- Romy Stühmeier, Kompetenzzentrum Technik-Diversity-Chancengleichheit e. V.
- Prof. Dr.-Ing. Gerhard Wunder, Freie Universität Berlin
- Vita Zimmermann-Janssen, Institut für Verbraucherwissenschaften

Gefördert durch:

Bundesministerium für Umwelt, Naturschutz, nukleare Sicherheit und Verbraucherschutz

aufgrund eines Beschlusses des Deutschen Bundestages

iRights.Lab
Think Tank für die digitale Welt

Autorin:
Franziska Busse

Redaktion:
Dr. Gergana Baeva, Vera Dünninger, Tom Völkel

Lektorat:
Hannah Willing (text | struktur)

Satz:
Christoph Löffler

Veröffentlichung:
Februar 2023

Lizenz:
Dieses Thesenpapier steht unter der Lizenz Creative Commons CC-BY-SA-Lizenz 4.0 International.

Das Zentrum für vertrauenswürdige Künstliche Intelligenz (ZVKI)

Das *ZVKI* macht als zentraler Ort der Debatte in Deutschland die Entwicklungen rund um gesellschaftliche Fragen zu Künstlicher Intelligenz und algorithmischen Systemen greifbar. Als nationale und unparteiische Schnittstelle zwischen Wissenschaft, Wirtschaft, Politik und Zivilgesellschaft informiert das *ZVKI* über viele verbraucher*innenrelevante Aspekte, ermöglicht öffentliche Diskussionen und entwickelt Instrumente zur Bewertung und Zertifizierung von vertrauenswürdiger KI.

Weitere Informationen zum ZVKI:
Webseite: www.zvki.de
Twitter: @zvki_de
Instagram: @zvki.de

Kontakt:
zvki@irights-lab.de

Verantwortung und Durchführung:
iRights.Lab (www.irights-lab.de)

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

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