Entwicklung von Erklärungen

Leitfaden

Erstellt im Rahmen der Masterthesis von
Miriam Chmil

Stand: 29.03.2023

Vorwort

Der Einsatz von künstlicher Intelligenz (KI) wird immer präsenter und damit einhergehend auch das Problem der Intransparenz von immer komplexer werdenden Modellen. Es ist jedoch wichtig, dass Menschen verstehen, wie diese Technologien funktionieren, um sie effizient nutzen zu können. Eine Methode, um geeignete Erklärungen zu konzipieren, ist der Fragen-basierte Ansatz. Bei dieser Herangehensweise liegt der Fokus auf dem Menschen, indem der Erklärungsbedarf in von Fragen erhoben wird.
Dieser Leitfaden soll Ihnen dabei helfen solche Erklärungen für KI-Systeme zu entwickeln. Der Leitfaden deckt nicht nur die Fragenerhebung ab, sondern stellt ein ganzheitliches Konzept zur Verfügung.
In diesem Leitfaden werden Sie durch die verschiedenen Schritte der Analyse, Entwicklung und Evaluation geführt. Dabei erhalten Sie Anleitungen sowie Tipps, die Ihnen helfen die Erklärungen zu Konzipieren und zu Verbessern.
Der Leitfaden kann während oder nach dem eigentlichen Designprozess des KI-Systems angewendet werden.

Es wird empfohlen den Leitfaden in Microsoft Word auszufüllen, da so Tabellen erweitert und ergänzt werden können. Der Leitfaden ist öffentlich unter GitHub[1] abrufbar.

[1] https://github.com/MiriamChm/Masterthesis_question_based_XAI

Überblick

System-analyse
- Ziele des Systems identifizieren
- Aufgaben des Systems identifizieren
- KI-Komponenten beschreiben

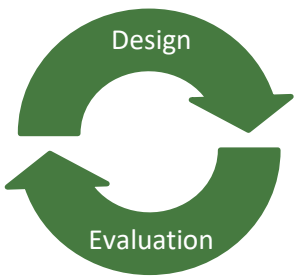
Analyse der Nutzenden
- Nutzende identifizieren
- Anwendungsszenarien identifizieren
- Ziele der Erklärungen identifizieren
- Kriterien der Gebrauchstauglichkeit aufstellen

Kontext-analyse
- Kontextfaktoren identifizieren
- Zusammenfassung der Analysen der Nutzenden und des Kontexts

Fragen-erhebung
- Fragenerhebung in Zusammenarbeit mit den Nutzenden
- Fragenerhebung ohne die Nutzenden

Fragen-analyse
- Kategorien der Fragen identifizieren
- Relevante Themen identifizieren
- Grad der Selbsterklärung einschätzen

Fragen-zuordnung
- Zuordnung der priorisierten Fragekategorien zu Erklärungstechniken



- Design der Erklärungen

- Intrinsische Evaluation
- Dialogische Evaluation
- Evaluation der Wirkung

Systemanalyse

In der Systemanalyse werden die Ziele des Systems untersucht und die zu erklärenden KI-Komponenten beschrieben. Die Nutzenden können in diesen Schritt mit eingebunden werden, indem durch Interviews und Fragebögen, deren Bedürfnisse und Erwartungen an das System evaluiert werden.

Ziele des Systems identifizieren

In diesem Schritt gilt es Ziele zu identifizieren, die mit dem System erreicht werden sollen.
Sind die Ziele des Systems bereits bekannt, können diese direkt in die Tabelle des dritten Schritts eingetragen werden. Wurde das System noch nicht entwickelt oder befindet sich in der Anfangsphase, können die Systemziele mithilfe der folgenden Anleitung ermittelt werden. Am Ende dieses Abschnittes ist ein Beispiel abgebildet.

1. Im ersten Schritt werden die Geschäftsziele identifiziert, die zur Wettbewerbsfähigkeit beitragen. Diese können sich z.B. auf die Produktivität, Flexibilität oder Zufriedenheit der Kund:innen beziehen.
   Bitte tragen Sie die Geschäftsziele der Priorität geordnet in die folgende Tabelle ein.

Nr.	Geschäftsziel

2. Im nächsten Schritt werden die Geschäftsziele auf das eigene System bezogen. Formulieren Sie Fragen nach Informationen, die Sie benötigen, um die Erfüllung der Geschäftsziele zu bewerten.

Geschäftsziel Nr.	Fragen

3. Nun gruppieren Sie bitte die Fragen nach Themenbereichen, auf die sie sich beziehen. Aus den Gruppen können Ziele für das System hergeleitet werden.

Ziele des Systems (Teilziele)	Fragen

Beispiel

Nr.	Geschäftsziel
1	Zufriedenheit der Kund:innen stärken

Geschäftsziel Nr.	Fragen
1	Ist die Produktqualität gut genug? Sind die Servicemitarbeitenden freundlich? Sind die Servicemitarbeitenden kompetent?

Ziele des Systems (Teilziele)	Fragen
Qualität des Service steigern	Sind die Servicemitarbeitenden freundlich? Sind die Servicemitarbeitenden kompetent?
Produktqualität steigern	Ist die Produktqualität gut genug?

Aufgaben des Systems identifizieren

In diesem Schritt werden die Aufgaben des Systems definiert, die zur Erreichung der Ziele notwendig sind. Tragen Sie bitte die Ziele des Systems in die folgende Tabelle ein und formulieren Sie Aufgaben, durch die eine Zielerreichung möglich ist.

Nr.	Ziele	Aufgaben des Systems
1		
2		
3		

4		

KI-Komponenten des Systems beschreiben

In diesem Schritt gilt es die zu erklärenden KI-Komponenten zu beschreiben. Bitte füllen Sie die folgende Vorlage für die einzelnen KI-Komponenten aus. Die Vorlage umfasst eine grobe Einordnung der KI durch Problemtyp und Paradigma, eine Beschreibung des Modells, die Realisierung sowie erste Erklärungspotentiale. Beispiele hierfür sind die intrinsische Erklärbarkeit einer KI oder typischerweise verwendete Erklärungstechniken.

Modellname	
Problemtyp	
Paradigma	
Beschreibung des Modells	
Realisierung	
Erklärungspotentiale	

Analyse der Nutzenden

Die Analyse der Nutzenden beinhaltet die Identifikation der Nutzenden, der Anwendungsszenarien in denen die Nutzenden mit der KI agieren, der Ziele, die durch die Erklärungen bei den Nutzenden bewirkt werden sollen sowie der Gebrauchstauglichkeitskriterien.
Es können Fragebögen, Interviews oder Gruppenbesprechungen eingesetzt werden, um die Bedürfnisse der Nutzenden zu identifizieren, den Arbeitsablauf mit dem System zu untersuchen und Probleme zu identifizieren.

Nutzende identifizieren

In diesem Schritt werden die Nutzenden des Systems und insbesondere die Empfangenden der Erklärungen identifiziert. Um alle Nutzenden abzudecken, können Sie folgende Kategorien verwenden:
Nutzende können in drei Kategorien eingeteilt werden:

- Primär Nutzende haben direkten Kontakt mit dem System und verwenden es als Endnutzende.
- Sekundärnutzende verwenden das System nur gelegentlich oder über eine Vermittlungsperson.
- Tertiärnutzende sind von den Auswirkungen der Nutzung des Systems betroffen oder treffen Entscheidungen über dessen Kauf.

Die Zielgruppen von Erklärungen können folgende sein:

- Entwickelnde und KI-Forschende: Die Erstellenden des Systems haben Interesse an der Überprüfung, um anhand der Erklärungen Fehler im System zu identifizieren.
- Expert:innen der Domänen und Nutzende des Systems: Durch Erklärungen kann Vertrauen in das Modell selbst gestärkt und neues Wissen erlernt werden.
- Nutzende, die durch die Modellentscheidungen beeinflusst werden: Die Erklärungen können dazu beitragen ihre Situation besser zu verstehen und faire Entscheidung zu belegen.
- Aufsichtsbehörden: Belegung der Rechtmäßigkeit des Systems.
- Führungskräfte und Vorstandsmitglieder: Bewertung der Einhaltung von Vorschriften und Förderung des Verständnisses von KI-Anwendungen im Unternehmen.

Bitte tragen Sie die identifizierten Nutzenden in folgende Tabelle ein:

Gruppe von Nutzenden	**Beschreibung**

Anwendungsszenarien identifizieren

Anwendungsszenarien beschreiben die verschiedenen Interaktionen zwischen den Agierenden und dem System. Innerhalb von Rollen üben Agierende verschiedene Interaktionsarten mit dem System aus und sind somit nicht mit den Nutzenden gleichzusetzen.
Da in diesem Leitfaden Modelle der KI erklärt werden sollen, sind nur die Anwendungsfälle relevant, die sich auf diese Modelle beziehen. Am Ende des Abschnittes ist ein Beispiel abgebildet.

1. Für die Identifikation der Anwendungsszenarien müssen zuerst die Interaktionsarten identifiziert werden.

Interaktionsart	**Beschreibung**

2. Anschließend ordnen Sie bitte die Interaktionsarten zu Agierenden zu.

Agierende	**Interaktionsart**

3. Im dritten Schritt wird zugeordnet welche Gruppe von Nutzenden die Agierenden-Rollen ausüben.

Gruppe von Nutzenden	**Agierende**

4. Im Verlauf dieses Leitfadens sollen Fragen von Nutzenden an das System erhoben werden. Um die Erhebung zu strukturieren, ist es sinnvoll diese anhand von Anwendungsszenarien anzuleiten. Als Vorbereitung dafür sollen die Interaktionsarten visualisiert werden. Fügen Sie daher bitte eine Abbildung der Oberfläche der Interaktionsarten ein. Wenn noch keine Oberfläche vorhanden ist, reicht eine Beschreibung der Oberfläche.

Interaktionsart	**Abbildung**

Beispiel

Interaktionsart	**Beschreibung**
Vertrag anlegen	Nutzende dürfen einen Vertrag anlegen
Vertrag bearbeiten	Nutzende dürfen einen Vertrag bearbeiten

Agierende	**Interaktionsart**
Sachbearbeitung Stufe 2	Vertrag anlegen, Vertrag bearbeiten

Gruppe von Nutzenden	**Agierende**
Sachbearbeitung	Sachbearbeitung Stufe 1, Sachbearbeitung Stufe 2

Interaktionsart	**Abbildung**
Vertrag anlegen	*Screenshot der Eingabemaske*

Ziele der Erklärungen identifizieren

In diesem Schritt gilt es Ziele zu identifizieren, die mit den Erklärungen erreicht werden sollen. Die Erklärungen werden auf diese Ziele hin ausgerichtet und evaluiert.
Mit den Zielen der Erklärungen wird beschrieben, was die Erklärungen bei den Nutzenden bewirken sollen. Das wichtigste Ziel von Erklärungen ist es, das Verständnis der Nutzenden zu stärken. Daher ist dieses bereits in der folgenden Tabelle eingetragen. Die weiteren Ziele können je nach Gruppe von Nutzenden voneinander abweichen, weshalb eine differenzierte Betrachtung pro Gruppe notwendig. Bitte kopieren Sie daher die Tabelle für jede Gruppe von Nutzenden.

Gruppe von Nutzenden:	
Nr.	**Ziel**
1	Verständnis
2	
3	
4	

Beispiele

Ziel	**Bedeutung**
Transparenz	Den Nutzenden erklären, wie das System zu einer Schlussfolgerung gelangt ist.
Rechtfertigung	Das Vertrauen der Nutzenden in die Schlussfolgerung durch eine zusätzliche Unterstützung erhöhen (z.B. durch Hintergrundwissen).
Relevanz	Den Nutzenden vermitteln, wieso die verfolgte Strategie des Systems relevant ist.
Konzeptualisierung	Den Nutzenden die verwendeten Konzepte (u.a. das Vokabular) erklären.
Lernen	Das Verständnis der Nutzenden für die Domäne wird gestärkt, indem der Lösungsprozess an diesen angepasst erklärt wird.
Vertrauen	Das Vertrauen der Nutzenden in das System wird gestärkt.
Interaktion	Den Nutzenden wird eine Interaktion mit dem System ermöglicht.
Überprüfbarkeit	Die Nutzenden können die Schlussfolgerung überprüfen und dem System mitteilen, wenn sie falsch ist.
Effektivität	Die Nutzenden werden darin unterstützt gute Entscheidungen zu treffen.
Persuasivität	Überzeugen der Nutzenden etwas auszuprobieren, zu gebrauchen oder zu kaufen.
Effizienz	Den Nutzenden wird geholfen Entscheidungen schneller zu treffen.
Zufriedenstellung	Erhöhung der Gebrauchstauglichkeit oder dem Spaß an der Nutzung.
Fairness	Die Nutzung von Erklärbarkeit, um Gerechtigkeit zu erreichen.

Gebrauchstauglichkeitskriterien aufstellen

Erklärungen sollen neben den betrachteten Zielen auch die Gebrauchstauglichkeit stärken. Daher müssen die Gebrauchstauglichkeitskriterien priorisiert und Ziele formuliert werden, um den Erfolg messbar zu machen. Bitte wählen Sie für Ihr System, pro Gruppe von Nutzenden, passende Kriterien aus und tragen Sie diese mit messbaren Zielen und ihrer Priorität in folgende Tabelle ein. Es können allgemeine Kriterien, wie die gelisteten Beispiele, oder domänenspezifische Kriterien verwendet werden.

Gruppe von Nutzenden:		
Priorität	**Kriterium**	**Ziel**

Beispiele

Gruppe von Nutzenden: Entwickelnde		
Priorität	**Kriterium**	**Ziel**
hoch	Erlernbarkeit	Erlernbarkeit innerhalb einer bestimmten Zeit
niedrig	Effizienz der Nutzung nach dem Erlernen	Erreichung einer bestimmten Durchlaufzeit
mittel	Fähigkeit der Nutzenden, die Nutzung bei unregelmäßiger Verwendung nicht erneut lernen zu müssen	Einfindung in einer bestimmten Zeit
hoch	Häufigkeit und Schwere der Nutzungsfehler	Erreichung einer Fehlerquote
hoch	Subjektive Zufriedenheit der Nutzenden	Erreichung einer Zufriedenheitsquote

Kontextanalyse

In diesem Schritt gilt es einerseits Kontextfaktoren zu identifizieren und andererseits die Ergebnisse aus den Analysen der Nutzenden und des Kontexts zusammenzufassen.

Kontextfaktoren identifizieren

Die Anwendung eines Systems geschieht immer in einem Kontext. Für die Entwicklung der Erklärungen ist es daher wichtig den gesamten Kontext zu betrachten, um alle wichtigen Einflussfaktoren ausfindig zu machen. Ein Teil des Kontexts wurde bereits durch die Analysen des Systems und der Nutzenden abgedeckt. In diesem Abschnitt werden die fehlenden Kontextfaktoren aufgezeigt. Ziel ist es, relevante Faktoren zu identifizieren, damit die Erklärungen auf diese abgestimmt werden können. Bitte tragen Sie die Kontextfaktoren, die berücksichtigt werden sollen, in die folgende Tabelle ein. Dabei kann eine Unterscheidung zwischen den Gruppen von Nutzenden sinnvoll sein. Auf der nächsten Seite finden Sie mögliche Kontextfaktoren. Die Kontextfaktoren sollten mit Hilfe der Nutzenden beispielsweise durch Gespräche identifiziert werden, da sie den direkten Kontakt zur Umgebung haben.

Gruppe von Nutzenden:	
Kontextfaktor	**Beschreibung**

Beispiele

Organisatorische Umgebung

- Struktur
    - Gruppenarbeit
    - Arbeitspraktiken
    - Hilfestellung
    - Unterbrechungen
    - Managementstruktur
    - Kommunikationsstruktur
    - Gehalt oder Bezahlung
- Einstellungen und Kultur
    - Politik zur Computernutzung
    - Organisatorische Ziele
    - Arbeitsbeziehungen
- Arbeitsplatzgestaltung
    - Funktionen des Arbeitsplatzes
    - Arbeitszeiten
    - Flexibilität bei der Arbeit
    - Leistungskontrolle
    - Leistungsrückmeldung
    - Arbeitsrhythmus, Autonomie und Ermessensspielraum

Physische Umgebung

- Arbeitsplatzbedingungen
    - Atmosphärische Bedingungen
    - Akustische Umgebung
    - Thermische Umgebung
    - Visuelle Umgebung
    - Umweltinstabilität
- Arbeitsplatzgestaltung
    - Raum und Möbel
    - Benutzerhaltung
    - Standort
- Sicherheit am Arbeitsplatz
    - Gesundheitsgefahren
    - Schutzkleidung und Arbeitsplatzgestaltung

Zusammenfassung der Analysen der Nutzenden und des Kontexts

In diesem Schritt werden die Analyseergebnisse zusammengefasst, damit diese bei der Auswahl von Erklärungstechniken und dem Design der Erklärungen nutzbar sind.

Die Ergebnisse sollen komprimiert, je nach Bedarf, pro Gruppe von Nutzenden oder gesamt, aufgeschrieben werden. Die Zusammenfassung stellt die Anforderungen, die sich auf die Wirkungsweise bei den Nutzenden beziehen, dar. Die Vorlage dafür ist auf der nächsten Seite zu finden.

Die Zusammenfassung umfasst die Beschreibung der Nutzenden, die Anwendungsszenarien, die Ziele der Erklärungen, die Gebrauchstauglichkeitskriterien sowie die Kontextfaktoren.

Analyse: BLANKO

Beschreibung der Nutzenden

Gruppe von Nutzenden	Beschreibung

Interaktionsarten der Nutzenden

Gruppe von Nutzenden	Interaktionsart

Ziele der Erklärungen

Nr.	Ziele der Erklärungen für _____

Kontextfaktoren

Kontextfaktor	Hinweis

Gebrauchstauglichkeitskriterien

Priorität	Kriterium	Ziel

Fragenerhebung

In diesem Schritt wird ermittelt, welche Anforderungen und Erwartungen die Nutzenden an die Erklärungen haben.
Wenn der Zugang zu den Nutzenden besteht, wird eine gemeinsame Fragenerhebung empfohlen. Ansonsten kann die Ermittlung ohne die Nutzenden erfolgen.

Fragenerhebung in Zusammenarbeit mit den Nutzenden

Wenn der Zugang zu den Nutzenden besteht, können die Fragen mit diesen zusammen erhoben werden. Zur Durchführung der Erhebung eignen sich viele Formate. Es sind Einzelgespräche mit Nutzenden, gemeinsame Workshops, Gespräche getrennt nach den Gruppen von Nutzenden oder der Einsatz eines Chatbots möglich. Der Aufbau dieses Informationsaustauschs ist jedoch gleich:
Zu Beginn werden die Aufgaben des Systems vorgestellt, um ein gemeinsames Verständnis zu schaffen. Die darauffolgende Befragung gliedert sich in zwei Teile:
Im ersten Teil werden die Nutzenden aufgefordert Fragen an das System zu stellen. Um die Teilnehmenden anzuleiten, werden sie gefragt, welche Fragen das System beantworten müsste, um die definierten Ziele zu erreichen. Bei einem komplexeren System, sollte die Befragung für die einzelnen Anwendungsszenarien stattfinden.
Im zweiten Teil werden die Nutzenden aufgefordert Absichten zu formulieren, die hinter den Fragen stehen. Außerdem sollen sie ihre Erwartungen an die Antworten auf ihre Fragen beschreiben. Es sollte auch auf das Format der Erklärung eingegangen werden.
Die folgenden Tabellen sollen je nach Granularität der Befragung für jedes Anwendungsszenario, für jeden Nutzenden und/ oder Gruppe von Nutzenden ausgefüllt werden und muss dementsprechend kopiert werden.
Hinweis: Ein unterschiedliches Verständnis in der Tiefe des Systems, z.B. der zugrundeliegenden Modelle von künstlicher Intelligenz, kann die Ergebnisse stark beeinflussen. Es kann daher sinnvoll sein innerhalb einer Schulung weitere Einblicke in die Funktionsweisen zu geben.

Gruppe von Nutzenden:	
Anwendungsszenario:	
Nr.	**Frage**
1	
2	
3	
4	

Gruppe von Nutzenden:	
Anwendungsszenario:	
Nr.	**Erwartungen und Absichten**
1	
2	
3	
4	

Fragenerhebung ohne die Nutzenden

In Projekten mit begrenztem Zugang zu den Nutzenden, kann ein Einsatz eines fragendatenbankbasierten Fragebogens sinnvoll sein. Der Fragenkatalog sollte dabei an das verwendete KI-Modell angepasst werden. Der Fragenkatalog kann durch Sie ausgefüllt werden. In der rechten Spalte kann angegeben werden, ob die jeweilige Frage relevant ist und wie die Erklärung aussehen sollte. Ein Beispielfragebogen, der für das jeweilige Projekt angepasst werden kann, ist abgebildet. Dabei kann eine Unterscheidung zwischen verschiedenen Gruppen von Nutzenden und/oder Anwendungsszenarien sinnvoll sein und muss dementsprechend angepasst werden.

Fragen-kategorie	**Frage**	**Wenn relevant, wie sollte die Erklärung umgesetzt sein?**
Daten	**Mit welcher Art von Daten wurde das System trainiert?**	
	Was ist die Quelle der Trainingsdaten?	
	Wie wurden die Label/ Grundwahrheiten erzeugt?	
	Wie groß ist der Stichprobenumfang der Trainingsdaten?	
	Welche(n) Datensatz(e) verwendet das System NICHT?	
	Was sind die möglichen Einschränkungen/Verzerrungen der Daten?	
	Wie groß ist der Umfang, der Anteil oder die Verteilung der Trainingsdaten mit bestimmten Merkmalen/Merkmalswerten?	
Ausgabe	**Welche Art von Ausgabe liefert das System?**	
	Was bedeutet die Ausgabe des Systems?	
	Welchen Umfang hat die Fähigkeit des Systems? Kann es ...?	
	Wie wird der Output von anderen Systemkomponenten verwendet?	
	Wie sollte ich den Output des Systems am besten nutzen?	
	Wie sollte der Output in meinem Arbeitsablauf eingesetzt werden?	
Performanz	**Wie genau/präzise/zuverlässig sind die Vorhersagen?**	
	Wie oft macht das System Fehler?	
	In welchen Situationen ist es wahrscheinlich, dass das System richtig/falsch liegt?	
	Wo liegen die Grenzen des Systems?	
	Welche Art von Fehlern wird das System potentiell machen?	
	Ist die Leistung des Systems ausreichend für...?	
	Wie macht das System Vorhersagen?	
	Welche Merkmale berücksichtigt das System?	

Wie (globale modellweite Erklärung)	Wird [Merkmal X] für die Vorhersagen verwendet oder nicht verwendet?	
	Wie sieht die allgemeine Logik des Systems aus?	
	Wie gewichtet es verschiedene Merkmale?	
	Welchen Regeln folgt es?	
	Wie wirkt sich [Merkmal X] auf seine Vorhersagen aus?	
	Welches sind die wichtigsten Regeln/Merkmale, die die Vorhersagen bestimmen?	
	Welche Art von Algorithmus wird verwendet?	
	Wie wurden die Parameter festgelegt?	
Warum	**Warum/Wie wurde diese Instanz mit dieser Vorhersage versehen?**	
	Welche Eigenschaft(en) dieser Instanz bestimmen die Vorhersage des Systems für diese?	
	Warum erhalten [Instanz A und B] die gleiche Vorhersage?	
Warum nicht	**Warum wird dieser Fall NICHT als [ein anderes Ergebnis Q] vorhergesagt?**	
	Warum wird für diesen Fall [P anstelle eines anderen Ergebnisses Q] vorhergesagt?	
	Warum werden für [Instanz A und B] unterschiedliche Vorhersagen gemacht?	
Wie wird das (eine andere Vorhersage)	**Wie sollte sich diese Instanz ändern, um eine andere Vorhersage Q zu erhalten?**	
	Was ist die Mindeständerung, die für diese Instanz erforderlich ist, um eine andere Vorhersage Q zu erhalten?	
	Wie sollte sich ein bestimmtes Merkmal für diesen Fall ändern, um eine andere Vorhersage Q zu erhalten?	
	Für welche Art von Instanz wird [ein anderes Ergebnis Q] vorhergesagt?	
Wie bleibt das (die aktuelle Vorhersage)	**Wie groß ist der zulässige Änderungsbereich für diese Instanz, damit die Vorhersage gleich bleibt?**	
	Welcher Wertebereich ist für ein bestimmtes Merkmal zulässig, damit die Vorhersage gleich bleibt?	
	Welche(s) Merkmal(e)/Merkmalswert(e) ist/sind erforderlich, um diese Vorhersage zu gewährleisten?	
	Welche Art von Instanz erhält die gleiche Vorhersage?	
Was, wenn	**Was würde das System vorhersagen, wenn diese Instanz sich ändert in...?**	
	Was würde das System vorhersagen, wenn ein bestimmtes Merkmal sich zu... ändert?	

	Was würde das System für [eine andere Instanz] vorhersagen?	
Andere	Wie/Warum wird sich das System im Laufe der Zeit ändern/anpassen/verbessern/abdriften? (Veränderung)	
	Kann ich, und wenn ja, wie kann ich das System verbessern? (Verbesserung)	
	Warum verwendet das System einen bestimmten Algorithmus/ein bestimmtes Merkmal/eine bestimmte Regel/einen bestimmten Datensatz oder nicht? (Nachbereitung)	
	Was bedeutet [eine Terminologie des maschinellen Lernens]? (terminologisch)	
	Was sind die Ergebnisse anderer Personen, die das System verwenden? (sozial)	

Fragenanalyse

In diesem Schritt werden die Fragen anhand von Fragekategorien, relevanten Themen und dem Grad der Selbsterklärung analysiert.
Tragen Sie die Ergebnisse der Analyse bitte übersichtlich auf den Anforderungsblättern ein, die Sie am Ende der Fragenanalyse finden. Die Granularität von der Fragenerhebung sollte bei der Fragenanalyse beibehalten werden und die Anforderungsblätter dementsprechend kopiert werden.

Fragekategorien identifizieren

Für diesen Teil werden ähnliche Fragen gruppiert und in Kategorien eingeordnet. Anschließend werden die Kategorien (z.B. nach Anzahl der enthaltenden Fragen) priorisiert. Für die Zuordnung kann der Fragenkatalog des vorherigen Abschnittes verwendet werden.
Die folgende Tabelle zeigt eine mögliche Zuordnung der Fragekategorien zu einem inhaltlichen Ansatz für die Gestaltung der Erklärungen.

Fragekategorie	**Allgemeine Erklärungsansätze**
Daten	- Dokumentieren Sie umfassende Informationen über die Trainingsdaten, einschließlich der Quelle, Herkunft, Typ, Größe, Abdeckung der Population, potenzielle Verzerrungen usw.
Ausgabe	- Beschreiben Sie den Umfang der Ausgabe oder der Systemfunktionen. - Schlagen Sie vor, wie die Ausgabe für nachgelagerte Aufgaben oder den Arbeitsablauf der Nutzenden verwendet werden sollte.
Performanz	- Bereitstellung von Leistungskennzahlen des Modells. - Angabe der Unsicherheitsfaktoren für jede Vorhersage. - Beschreiben Sie mögliche Stärken und Grenzen des Modells.
Wie (globale modellweite Erklärung)	- Beschreiben Sie die allgemeine Modelllogik in Form von Merkmalsauswirkungen, Regeln oder Entscheidungsbäumen (manchmal muss ein einfaches Ersatzmodell verwendet werden). - Wenn Nutzende nur an einer Übersicht interessiert sind, beschreiben Sie die wichtigsten Merkmale oder Regeln, die berücksichtigt wurden.
Warum	- Beschreiben Sie, welche Schlüsselmerkmale der Instanz die Vorhersage des Modells bestimmen. - Beschreiben Sie die Regeln, die der Fall erfüllt, um die Vorhersage zu gewährleisten. - Zeigen Sie ähnliche Beispiele mit demselben vorhergesagten Ergebnis, um die Vorhersage des Modells zu rechtfertigen.
Warum nicht	- Beschreiben Sie, welche Änderungen für die Instanz erforderlich sind, um die alternative Vorhersage zu erhalten und/oder, welche Merkmale der Instanz die aktuelle Vorhersage garantieren. - Zeigen Sie prototypische Beispiele, die das alternative Ergebnis liefern.
Wie wird das (eine andere Vorhersage)	- Hervorhebung von Merkmalen, deren Änderung (Erhöhung, Verringerung, Fehlen oder Vorhandensein), die Vorhersage verändern könnte. - Zeigen Sie Beispiele mit minimalen Unterschieden, die ein anderes Ergebnis als die Vorhersage liefern.

Wie bleibt das (die aktuelle Vorhersage)	- Beschreiben Sie Merkmale/Merkmalsbereiche oder Regeln, die die gleiche Vorhersage garantieren könnten. - Zeigen Sie Beispiele, die sich von dem speziellen Fall unterscheiden, aber dennoch das gleiche Ergebnis liefern.
Was, wenn	- Zeigen Sie, wie sich die Vorhersage entsprechend der angefragten Änderung abwandelt.

Relevante Themen identifizieren

Im zweiten Teil werden relevante Themen aus den Absichten und Erwartungen aufgedeckt.
Um relevante Themen zu identifizieren, markieren Sie bitte Themen in den Tabellen mit den Absichten und Erwartungen. Wenn Themen wiederholt vorkommen, sind sie als relevant zu betrachten.

Grad der Selbsterklärung einschätzen

Menschen besitzen die Eigenschaft, dass sie sich Möglichkeiten suchen, Erklärungen zu erschließen. Diese Eigenschaft sollte sich bei dem Design von Erklärungen zu Nutze gemacht werden. Durch die Analyse der Fragen kann bestimmt werden, welche Bedürfnisse die Nutzenden an die Tiefe der Erklärungen haben. Es können oberflächliche erklärende Hilfestellungen, wie Heatmaps verwendet werden, bis hin zu anspruchsvollen Diagnosen von Fehlschlägen des Systems. Die Fragen der Nutzenden sollten dahingehend untersucht werden, zu welchem Grad eine Selbsterklärung sinnvoll ist. Die folgende Tabelle zeigt die verschiedenen Stufen der Selbsterklärung.

Grad	**Beschreibung**
1. Null	Keine Erklärungen
2. Oberflächlich	Es werden beispielsweise durch Heatmaps oder linguistische Merkmale einige der, von der KI durchgeführten Analysen, visualisiert. Sie reichen nicht aus, um die KI zu verstehen, können aber nützlich sein. Die Nutzenden erklären sich die Funktionsweise, in Verbindung mit weiteren Hilfsmitteln, selbst.
3. Erfolge	Darstellung von Instanzen oder Demonstrationen, in denen die KI korrekte Vorhersagen oder Empfehlungen generiert.
4. Mechanismen	Globale Beschreibungen über die Funktionsweise der KI.
5. KI-Reasoning	Einblick in die Funktionsweise, wie die KI Entscheidungen trifft. Dazu gehören beispielsweise Logiken oder Regeln und die Gewichtung von Informationen.
6. Fehlschläge	Darstellung von Misserfolgen der KI. Diese zeigen die Grenzen der KI auf und zeigen wie sie funktioniert bzw. nicht funktioniert.
7. Vergleiche	Durch Vergleiche können Entscheidungslogiken oder Gewichtungen verdeutlichen, indem verschiedene Bedingungen gegenübergestellt werden.
8. Diagnosen von Fehlschlägen	Wenn die Misserfolge nicht nur dargestellt, sondern zusätzlich diagnostiziert werden, können die Nutzenden mehr Informationen daraus ziehen. Die Nutzenden können durch verschiedene Eingaben die Auswirkungen auf die KI-Ausgaben erkennen und ihre eigenen Schlüsse zu Diagnosen ziehen.

Anforderungen an die Erklärungen BLANKO

Gruppe von Nutzenden: **Das betroffene KI-Modell**:

Anwendungsszenario:

Fragen und Ansätze zur Lösung (lokaler (L) oder globaler (G) Ansatz und Nummer der Selbsterklärung)

Frage-kategorie	Frage - Erwartung an die Lösung	L / 2
	Frage - Erwartung an die Lösung	L / 2
	Frage - Erwartung an die Lösung	L / 2
	Allgemeine Erklärungsansätze - Ansätze für die einzelnen Fragekategorien aus der Tabelle unter „Fragekategorien“	L / 3,4

Frage-kategorie	Frage - Erwartung an die Lösung	G / 4
	Allgemeine Erklärungsansätze - Ansätze für die einzelnen Fragekategorien aus der Tabelle unter „Fragekategorien“	G / 5

Wichtige Themen

1.
2.
3.

Fragenzuordnung

Bei der Zuordnung ist es das Ziel den priorisierten Fragekategorien mögliche Erklärungstechniken zuzuordnen. Dazu müssen Erklärungstechniken recherchiert werden, welche die gesammelten Anforderungen abdecken. Neben den Anforderungen, die bereits gesammelt wurden, muss die Komptabilität zu den KI-Komponenten beachtet werden.

Wenn ein Modell des maschinellen Lernens verwendet wird, können die folgenden Erklärungstechniken einen Ansatz bieten. Ansonsten müssen passende Erklärungstechniken auf Basis der Anforderungen recherchiert werden.

Fragekategorie	**Erklärungstechniken für maschinelles Lernen**
Daten	FactSheets, Model Cards
Ausgabe	FactSheets, DataSheets
Performanz	Precision, Recall, Accuracy, F1, AUC Uncertainty Quantification 360 FactSheets, Model Cards
Wie (globale modellweite Erklärung)	ProfWeight, Global Feature Importance, PDP, BRCG, GLRM, Rule List, DT Surrogate
Warum	LIME, SHAP, LOCO, Anchors, ProtoDash
Warum nicht	CEM, Prototype counterfactual, ProtoDash (on alternative class)
Wie wird das (eine andere Vorhersage)	CEM, Counterfactuals, DiCE
Wie bleibt das (die aktuelle Vorhersage)	CEM, Anchors
Was, wenn	PDP, ALE, What-if Tool

Nach dem Abgleich zwischen Erklärungstechniken und den Anforderungen können, je nach Bearbeitung, für jedes Anwendungsszenario und/ oder jede Gruppe von Nutzenden, eine oder mehrere Erklärungstechniken festgehalten werden. Kopieren Sie bitte je nach Granularität die folgende Tabelle und tragen Sie die gewünschten Erklärungstechniken ein.
Dabei ist zu beachten, dass verschiedene Techniken (z.B. lokale und globale Erklärungen) miteinander kombiniert werden sollen, um die Effektivität zu verstärken.

HINWEIS: Sollten die bevorzugten Erklärungstechniken nicht mit der verwendeten KI kompatibel sein, sollte die Wahl der KI überdacht werden. Es kann es im Anfangsstadium der Entwicklung sinnvoll sein eine andere KI zu verwenden oder eine andere Art der Problemlösung einzusetzen.

Gruppe von Nutzenden:	
Anwendungsszenario	**Erklärungstechniken**

Design der Erklärungen

Mit diesem Schritt startet der iterative Design- und Evaluationsprozess. Durch die Iterativität kann eine kontinuierliche Verbesserung der Erklärungsqualität erreicht werden.

Wenn dieser Prozess das erste Mal durchlaufen wird, ist es das Ziel die Erklärungen mindestens skizzenhaft umzusetzen. Im weiteren Verlauf können Prototypen eingesetzt werden, bis die Erklärungen schließlich im System integriert werden.
Für den ersten Entwurf ist es notwendig, die, in diesem Leitfaden gesammelten Informationen einzubinden. Dazu sollten die ausgewählten Erklärungstechniken sowie die Anforderungsblätter bereitgelegt werden.
Anschließend können die Erklärungstechniken Schritt für Schritt, mit Fokus auf die Anforderungen, umgesetzt werden.

Wenn bereits ein Evaluationsprozess durchlaufen wurde, müssen die, aus diesem Schritt entstandenen, Anforderungen umgesetzt werden. Hierbei ist es sinnvoll auch die anderen Anforderungsblätter zu aktualisieren und bereitzulegen.

Bei dem Design der Erklärungen sollten folgende Hinweise beachtet werden:

1. KI-Systeme, die sich selbst durch Lernen weiterentwickeln oder in einem dynamischen Umfeld eingesetzt werden, verändern ihre Algorithmen und müssen daher die Nutzenden bei Bedarf über Änderungen informieren. Ebenso wichtig ist, dass Erklärungen nicht erneut gegeben werden müssen, wenn die Umstände gleichgeblieben sind.
2. Erklärungssysteme statt Erklärungen: Es reicht nicht aus, die Erklärungen, die von den Erklärungstechniken generiert werden, nur an die Nutzenden weiterzugeben. Die Erklärungen müssen in das System integriert und durch, z.B. Anleitungen oder Erkundungsschnittstellen, begleitet werden, damit sie von den Nutzenden effektiv verwendet werden können.

Evaluation

In dieser Phase werden die Erklärungen evaluiert. Dafür werden die Erklärungen aus drei Blickwinkeln betrachtet. Anschließend werden die Mängel gesammelt, damit diese in der nächsten Designphase verbessert werden können.

Intrinsische Evaluation

In diesem Schritt wird untersucht, ob das System in der Lage ist, die gewünschten Erklärungen zu generieren. Dazu werden die Anforderungen an die Erklärungen mit der Ausgabe der Erklärungen verglichen. Dies wird ohne die Nutzenden durchgeführt. Bitte passen Sie die folgende Tabelle, je nach Granularität der Erklärungen an und tragen Sie die Unterschiede zwischen den gewünschten Erklärungen und den generierten Erklärungen ein (z.B. anhand Struktur, Modalität und semantischen Eigenschaften).

Gruppe von Nutzenden:	
Anwendungsszenario	**Bewertung**

Dialogische Evaluation

Bei der dialogischen Evaluation wird untersucht, ob der Output des Systems für die Nutzenden als Erklärung wirkt.
Für die Bewertung können beispielsweise Studien mit Nutzenden, Reaktionsstudien, experimentelle Studien und qualitative und quantitative Methoden im Allgemeinen eingesetzt werden.
Bitte passen Sie die folgende Tabelle, je nach Granularität der Erklärungen an und tragen Sie identifizierten Mängel ein.

Gruppe von Nutzenden:	
Anwendungsszenario	**Bewertung**

Evaluation der Wirkung

In dem letzten Evaluationsschritt wird die Wirkung der Erklärungen untersucht. Dazu wird bewertet, ob die gewünschten Auswirkungen erzielt werden. Hierfür werden Zielerreichung der Erklärungen, die Gebrauchstauglichkeitskriterien sowie die Kontextfaktoren herangezogen. Für die Bewertung können beispielsweise Studien mit Nutzenden, Reaktionsstudien, experimentelle Studien und qualitative und quantitative Methoden im Allgemeinen eingesetzt werden. Bitte passen Sie die folgende Tabelle, je nach Granularität der Erklärungen, an und tragen Sie Bewertungen ein.

Ziele der Erklärungen

Gruppe von Nutzenden:	
Ziel der Erklärung	**Bewertung**

Gebrauchstauglichkeitskriterien

Gruppe von Nutzenden:	
Kriterium	**Bewertung**

Kontextfaktoren

Gruppe von Nutzenden:	
Kontextfaktor	**Bewertung**

Zusammenfassung der Evaluation

In dem letzten Schritt werden die identifizierten Mängel gesammelt, damit sie in der nächsten Designphase gebündelt beseitigt werden können. Je nach Granularität der Betrachtung können die Mängel gruppiert werden.

Gruppe von Nutzenden:			
Anwendungs-szenario	**Intrinsische Evaluation**	**Dialogische Evaluation**	**Evaluation der Wirkung**

Anschließend folgt erneut der Designschritt, in dem die identifizierten Mängel der Erklärungen beseitigt werden. Auch ohne identifizierte Mängel, sollte die Qualität der Erklärungen weiterhin überprüft werden.