TRUSTWORTHY-RICHTLINIEN DER SOFTWARE

storywise

Version 1.1 vom 18.07.2023

Ein Produkt der

IREO

ireo GmbH
Griesgasse 1
8020 Graz

INHALT

TRUSTWORTHY-RICHTLINIEN VON STORYWI.SE

Zentraler Bestandteil der Software storywi.se ist neben der spezialisierten Benutzeroberfläche für die Eingabe und für den Vergleich von Software-Requirements die intelligente Anbindung von KI-Systemen.

Der ireo GmbH ist bewusst, dass die Anwendung von KI-Systemen nicht unbedacht erfolgen darf und einen gezielten Einsatz sowie Kontrolle benötigt.

Es existieren verschiedene Leitlinien und Frameworks, die als Referenz für eine ethische Datenverarbeitung dienen, wie die Datenschutz-Grundverordnung (DSGVO) der Europäischen Union[1], die Ethical Guidelines for Trustworthy AI der Europäischen Kommission[2], die Fair Information Practice Principles (FIPPs) des National Institute of Standards and Technology (NIST)[3] und die Datenschutz- und Ethik-Standards des Institute of Electrical and Electronics Engineers (IEEE)[4].

Für die Entwicklung unserer KI-Systeme für die Software storywi.se sowie für die Verarbeitung der Daten im Trainings- sowie Produktiveinsatz hat sich IREO deshalb folgenden 10 Punkte aus den oben genannten Quellen als interne Richtlinie abgeleitet und definiert:

1. Transparenz

Die Datenverarbeitung muss transparent sein und die Kunden werden über die Art der gesammelten Daten, den Zweck und die Verwendung informiert.

Um diesen Punkt einzuhalten, werden die Benutzer bei der Verwendung der Free-Version darauf hingewiesen, dass diese Daten für das Training von KI-Daten hinzugezogen werden dürfen. Sobald ein Kunde einen bezahlten Plan verwendet, werden die Daten nicht mehr für das Training verwendet. Falls einzelne Kunden der Verwendung der Daten aktiv zustimmen, ist das in Ausnahmefällen möglich.

[1] Europäische Datenschutz-Grundverordnung (DSGVO): Die DSGVO ist ein umfassendes Datenschutzgesetz, das in der gesamten Europäischen Union gilt. Sie stellt strenge Anforderungen an den Schutz von personenbezogenen Daten und die Transparenz von KI-Systemen, die solche Daten verarbeiten.
2 EU-Leitlinien für vertrauenswürdige KI: Die Europäische Kommission hat Leitlinien für vertrauenswürdige KI entwickelt, die ethische Grundsätze und rechtliche Vorschriften umfassen. Diese Leitlinien betonen die Bedeutung von Transparenz, Fairness, Sicherheit, Datenschutz und Verantwortlichkeit.
3 Die Fair Information Practice Principles (FIPPs) des National Institute of Standards and Technology (NIST) sind eine Reihe von Datenschutzprinzipien, die als Leitfaden für den Umgang mit personenbezogenen Daten dienen. Diese Prinzipien wurden entwickelt, um einen angemessenen Schutz der Privatsphäre und der Vertraulichkeit von personenbezogenen Daten zu gewährleisten.
[4] Das Institute of Electrical and Electronics Engineers (IEEE) hat eine Reihe von Datenschutz- und Ethikstandards entwickelt, um den verantwortungsvollen Umgang mit Technologie und personenbezogenen Daten zu fördern.

2. Datensparsamkeit

Es werden nur Daten gesammelt, die für den Zweck der Verarbeitung notwendig sind. Es ist wichtig, dass keine unnötigen oder irrelevanten Daten gesammelt werden.

Das System speichert die Vorschläge und Anmerkungen vom KI System in einer Datenbank, damit nicht jede Interaktion mit dem System eine neue Anfrage bedingt. Diese Daten werden nur dem jeweiligen Kunden zur Verfügung gestellt und nicht für analytische Zwecke o.Ä. weiterverwendet.

3. Datenschutz

Es werden geeignete Maßnahmen ergriffen, um den Datenschutz hinsichtlich unbefugten Zugriffs, Verlust oder Diebstahlt sicherzustellen. Die Umsetzung erfolgt basierend auf gängigen Industrie-Standards.

Nach dem Zero-Trust Prinzip haben nur die dafür absolut notwendigen Personen Zugriff auf die Server und auf die Protokolldaten. Entwickler bekommen beispielsweise keinen Zugriff auf das Produktivsystem.

4. Rechtmäßigkeit

Die Datenverarbeitung erfolgt in Übereinstimmung mit geltenden Gesetzen und Vorschriften.

Um dies zu gewährleisten, wurden basierend auf existierenden Guidelines für sichere Verarbeitung von Daten wie der DSGVO, FIPPs, IEEE Datenschutz & Ethik Standard, firmeninterne und softwarespezifische Regeln abgeleitet und umgesetzt.

5. Einwilligung

Falls personenbezogene Daten verarbeitet werden, müssen die betroffenen Personen der Verarbeitung ihrer Daten zustimmen. Diese Einwilligung muss freiwillig und informiert erfolgen.

Durch die manuelle Überprüfung der Trainingsdaten wird es nicht zur Verarbeitung von personenbezogenen Daten kommen. Das System soll in weiterer Folge dabei helfen, Entwickler auf mögliche Probleme bei der korrekten Datenschutz-Umsetzung hinzuweisen.

6. Fairness

Die Datenverarbeitung soll fair und ohne Diskriminierung erfolgen.
Der Ansatz, User Stories vorzuschlagen, soll helfen, um die Fairness zwischen Auftraggeber und Auftragnehmer zu erhöhen, da es seltener zu der Situation kommt, dass Anforderungen erst während der Umsetzung auftauchen, die einer Partei aber vielleicht sogar schon bewusst waren.
Unser System stellt eine Unterstützung beim Entwurf von Softwaresystemen dar. Dabei werden keine personenbezogenen Daten verwendet, und die Gefahr einer Diskriminierung, Sozialer Bias o.Ä. existiert nicht.

7. Transfers von Daten

Bei der Übertragung von Daten werden geeignete Sicherheitsmaßnahmen getroffen, um sicherzustellen, dass die Daten sicher und geschützt sind. Dies trifft vor allem auf den Transfer von Daten über Ländergrenzen hinweg zu (z.B. bei Verwendung von Cloud-Storages, externen API's o.ä.).

Der Industrie-Standard bezüglich der Speicherung und Übertragung von Daten wird eingehalten. Es werden keine Testdaten aus dem Produktivsystem verwendet oder Entwicklern zur Verfügung gestellt, ohne dafür geeigneten Anonymisierungs- oder sonstige Schutzmaßnahmen zu ergreifen. Zusätzlich sind sämtliche Entwickler-Maschinen verschlüsselt und nur mit Passwort zugreifbar. Build- und Versionierungs-Server sind nur im VPN erreichbar und alle Dev- und Log-Umgebungen sind über einen Login mittels 2FA gesichert.

8. Aktualität

Die Daten sollen aktualisiert und korrekt gehalten werden.

Diese Richtlinie verhindert die unterschiedliche Behandlung von Personen bzw. dient der Umsetzung der DSGVO spezifischen Richtlinien, um Personendaten immer aktuell zu halten.
Im System selbst ist die Pflege der personenbezogenen Daten (Vorname, Nachname) Aufgabe der Benutzer und kann per se nicht von der ireo GmbH übernommen werden.

9. Verwendungszweck

Die Daten dürfen ausschließlich für den angegebenen Verwendungszweck genutzt werden und dürfen ohne vorherige Absprache mit den betroffenen Akteuren nicht für andere Zwecke verwendet werden.

Die Daten werden nur beim „Free tier“ oder mit ausdrücklichem Einverständnis und nach manueller Kontrolle für das Training verwendet.

10. Verantwortlichkeit

Es gibt eine Verantwortlichkeit für die Datenverarbeitung, insbesondere in Bezug auf die Einhaltung der oben genannten Punkte.
Die Geschäftsleitung ist selbst für die Einhaltung der Richtlinien und der Einhaltung verantwortlich und verpflichtet sich ausdrücklich zur Einhaltung dieser.

Zusätzlich zu diesen Richtlinien haben wir den Data Canvas von Datentreiber[5] herangezogen, um die Datenstrategie holistisch während des gesamten Life-Cycles zu untersuchen. Für die Sicherstellung der ethischen Verarbeitung der Daten haben wir zusätzlich den Data Ethics Canvas des Open Data Institutes[6] herangezogen. Dadurch wollen wir nicht nur die effektive Konzeptionierung der Datenpipeline gewährleisten, sondern auch die rechtmäßige Entwicklung und Verwendung dieser.

[5] Der Datenstrategie-Canvas ist ein Werkzeug, das Unternehmen und Organisationen dabei unterstützt, eine umfassende und klar definierte Datenstrategie zu entwickeln. Es ist ein visuelles Framework, das verschiedene Aspekte der Datenstrategie abdeckt und es ermöglicht, diese auf einer einzigen Seite zu erfassen.
[6] Der Data Ethics Canvas des Open Data Institutes ist ein Werkzeug, das entwickelt wurde, um Organisationen bei der ethischen Bewertung und Reflexion über Datenprojekte zu unterstützen. Es soll dabei helfen, mögliche ethische Herausforderungen zu identifizieren und Maßnahmen zu entwickeln, um diese zu adressieren.

ERWEITERTE ANSÄTZE ZUR GEWÄHRLEISTUNG DER KONTROLLE ÜBER DIE FUNKTIONSWEISE VON KI-SYSTEMEN

Vorrang menschlichen Handelns und menschlicher Aufsicht

Der Benutzer des KI-Systems wird darauf aufmerksam gemacht, dass es eine automatisiert-erstellte Analyse durch KI in diesem Berechnungsschritt geben wird. Dies wird mithilfe eines Pop-up-Fensters realisiert. Das AI System wird von Human-in-the-loop bzw. human-in-command überwacht. Wenn die AI sich in einem Punkt nicht sicher ist, wird keine Aussage getroffen. Es muss eine manuelle Verifikation/Erweiterung des Systems durchgeführt werden.

Das KI-System wird anhand der Leitlinien aus dem “Cybersecurity Act in Europe”[7] entwickelt. Im Allgemeinen ist ein sehr geringes Risiko durch Cyber-Attacken aufgrund einer Closed Solution Entwicklung gegeben[8].

Sensitivity Detection System

Das Risiko von Alpha- und Betafehlern in der Detektion von Data Governance Verletzungen besteht. Im Bereich der Künstlichen Intelligenz (KI) und der Detektion von Data Governance beziehen sich Alpha- und Betafehler auf Fehler, die bei der Klassifizierung oder Erkennung von Daten auftreten können.

Der Alphafehler, auch als Typ-I-Fehler bezeichnet, tritt auf, wenn ein KI-System fälschlicherweise ein Ereignis oder eine Datenklasse erkennt, obwohl es nicht existiert. Mit anderen Worten, das System gibt ein positives Ergebnis, obwohl es nicht sein sollte. Ein Alphafehler kann dazu führen, dass irrtümlich Daten als relevant oder problematisch eingestuft werden, wenn sie es nicht sind.

Der Betafehler, auch als Typ-II-Fehler bezeichnet, tritt auf, wenn ein KI-System ein Ereignis oder eine Datenklasse fälschlicherweise nicht erkennt, obwohl es vorhanden ist. Mit anderen Worten, das System gibt ein negatives Ergebnis, obwohl es nicht sein sollte. Ein Betafehler kann dazu führen, dass relevante oder problematische Daten übersehen oder nicht erkannt werden.

Die Kontrolle und Minimierung von Alpha- und Betafehlern ist ein wichtiger Aspekt der Data Governance im Kontext von KI. Es ist wichtig, dass KI-Systeme so trainiert und optimiert werden, dass sie eine angemessene Balance zwischen der Erkennung von relevanten Daten und der Vermeidung von

[7] Der "Cybersecurity Act in Europe" ist eine Verordnung der Europäischen Union, die darauf abzielt, die Cybersicherheit in der EU zu stärken.
[8] Im Bereich Künstliche Intelligenz (KI) bezieht sich der Begriff "Closed Solution Entwicklung" auf die Entwicklung von KI-Systemen, die auf spezifische Anwendungsfälle oder Probleme zugeschnitten sind und begrenzte Verwendungsmöglichkeiten haben. Im Gegensatz zu einer "Open Solution Entwicklung", bei der ein KI-System so konzipiert wird, dass es flexibel und anpassungsfähig ist, um verschiedene Anwendungsfälle zu unterstützen, ist eine Closed Solution auf eine bestimmte Aufgabe oder einen spezifischen Kontext beschränkt.

Fehlalarmen aufweisen. Dies kann durch die Anwendung geeigneter Algorithmen, die Optimierung der Schwellenwerte für die Klassifizierung und regelmäßige Überprüfungen und Anpassungen des Systems erreicht werden.

Da die Sensitivity-Detection ein Unterstützungssystem darstellt, muss generell ein manueller Entscheidungsprozess durchgeführt werden. Bei beiden Systemen gibt es keine nachteiligen oder schädlichen Folgen aufgrund mangelnder Genauigkeit. Es entsteht maximal ein Mehraufwand durch den Enduser.

Risiko von korrumpierten Daten

Trainingsdaten werden intern erhoben und werden davor verifiziert und nicht automatisch der AI im Trainingsprozess gefüttert. Dabei verwenden wir aktuelle Tools wie JIRA oder storywi.se selbst, aus denen die Daten für das Training bezogen werden. Diese werden in weiterer Folge manuell überprüft und erst danach in den Trainingspool aufgenommen. Wir verwenden keine Strategie des “kontinuierlichen Online-Lernens”.

Die Integrität, Robustheit und die allgemeine Sicherheit des KI-Systems durch potenzielle Angriffe sind durch die oben genannten Punkte gewährleistet. Des Weiteren unterstützen folgende Punkte diese Anforderungen:

- Das Recommender-System muss durch einen bestimmten Trigger aufgerufen werden und läuft nicht in einem kontinuierlichen Status
- Read Only AI: Parameter können während des Betriebs nicht geändert werden.
- Sicherheitsupdates werden über die gesamte Lifetime dieses Produkts gewährleistet.

Nachdem es sich dabei um eine Webanwendung handelt und uns die Sicherheit des Systems bedeutet, wird das System regelmäßig mit den neuesten Betriebssystems- & Sicherheitsupdates versorgt.

Schutz der Privatsphäre und Datenqualitätsmanagement

Das System ist als Closed System konzipiert, sodass keine Datenpunkte oder -quellen zu externen Systemen gelangen können. Es existiert im Zuge der ethischen experimentellen Entwicklung auch keine API-Anbindung nach außen. Diese wird erst im Zuge eines Whitelabeling-Konzepts umgesetzt, und muss dementsprechend in diesem Schritt auch allen ethischen Normen und Prüfungen, wie das vorliegende System, unterzogen werden. Durch unser System soll das Maß an Data Governance für Datenbankmodelle erhöht werden. Die Sensitivity Detection soll sicherstellen, dass sensitive Daten und Datenpunkte erkannt und entsprechend den Richtlinien behandelt werden. Weiters gibt es einen zertifizierten Datenschutzexperten (DI Dr. techn. Andreas Schüppel), der während der gesamten Entwicklungsphase des KI-Systems involviert ist und auf die Datenschutzkonforme Umsetzung achtet.

Transparenz

Es werden Methoden wie SHAP-Values[9] verwendet, um eine Erklärbarkeit und Nachvollziehbarkeit der KI-Entscheidungsfindung zu ermöglichen. Diese werden entweder in Form von konkreten Zahlen oder mit einem Ampelsystem (Rot, Orange, Grün) realisiert. Das Feedback der Benutzer*innen hinsichtlich der Ergebnisse wird implizit gespeichert in dem die Anzahl der angenommenen Ergebnisse gespeichert wird. Des weiteren wird es eine Dokumentation für die Benutzer*innen geben, welche genau erklärt, wie die AI zu verwenden ist, welche Modelle in der Berechnung eingesetzt werden, sowie die Limitationen dieser Modelle.

Vielfalt, Nichtdiskriminierung und Fairness

Der Einfluss eines möglichen Bias hinsichtlich Diskriminierung, Diversität oder Fairness wurde im Vorfeld bereits mit mehreren Handlungsableitungen ausgeschlossen. Es werden keine personenbezogenen Daten verarbeitet, da diese im Zuge der Datenstrategie in mehreren Instanzen ausgefiltert oder ausgeschlossen werden.

Der Input und Output des AI-Systems kann generell als Accessible eingestuft werden. Nachdem IREO sowie unsere Projektpartner selbst Teil der Zielgruppe sind (Entwickler*innen, Systemarchitekt*innen), werden diese dementsprechend intern und extern in den Entwicklungsprozess miteinbezogen. Weiters haben wir Prozesse zur Prüfung und Überwachung potenzieller Verzerrungen während des gesamten Lebenszyklus des KI-Systems bewertet und eingeführt.

Gesellschaftliches und ökologisches Wohlergehen

Das System selbst trägt zu keinen negativen Auswirkungen hinsichtlich des gesellschaftlichen und ökologischen Wohlergehens bei. Das System unterstützt bei der Konzeption des Datenbankschemas und weist auf potenzielle Sicherheitsrisiken hin. Durch diese unterstützende Wirkung besteht kein Risiko in der Dequalifizierung von Arbeitskräften. Zusätzlich werden Schulungsunterlagen für Benutzer*innen zur Verfügung gestellt. Grundsätzlich ermöglicht das System eine Ressourceneinsparung in der Bearbeitung von Datenmodellen, sowie eine Risikominimierung hinsichtlich Data Governance Themen durch die Sensitivity Detection.

Rechenschaftspflicht

Es werden Prozesse eingerichtet, die die Überprüfbarkeit des KI-Systems erleichtern. Darunter zählen unter anderem die Rückverfolgbarkeit des Entwicklungsprozesses, die Beschaffung von Trainingsdaten und Protokollierung der Prozesse und Ergebnisse sowie mögliche positive und negative Auswirkungen des

[9] SHAP-Values (Shapley Additive Explanations) sind eine Technik, die in der Künstlichen Intelligenz (KI) verwendet wird, um die Beiträge einzelner Merkmale bei der Vorhersage von Modellen zu erklären. Sie bieten eine Möglichkeit, die Vorhersagen eines Modells auf eine einzelne Instanz oder Beobachtung herunterzubrechen und den Einfluss jedes Merkmals auf die Vorhersage zu quantifizieren.

KI-Systems. Das KI-System soll während der Entwicklung sowie auch im Produktiveinsatz intern sowie von unabhängigen Dritten evaluiert werden. (Partnerschaften, Zertifizierungsstellen). Durch einen iterativen Entwicklungsprozess werden Aufzeichnung hinsichtlich Versionierung und Funktionalitäten chronologisch abgelegt, sodass zu jedem Zeitpunkt bzw. für jedes Ergebnis erhoben werden kann, mit welcher Modellversion, welchen Daten und in welchem Prozess dieses zustande gekommen ist.

Vorgeschlagener Standard für vertrauenswürdige KI

Als zentraler Standard werden die Ethik-Leitlinien für eine vertrauenswürdige KI der Europäischen Kommission herangezogen. Die Europäische Kommission ist in diesem Bereich ein Vorreiter, weshalb diese Leitlinien als aktueller Standard angesehen werden. Um die sieben Grundsätze von vertrauenswürdiger künstlicher Intelligenz im Zuge der Entwicklung sowie auch im Ergebnis sicherzustellen, haben wir mehrere unterschiedliche Assessments für das System durchgeführt. Dazu zählt die Assessment List for Trustworthy Artificial Intelligence (ALTAI), die Einordnung der Entwicklung in die Risikopyramide und dementsprechender Handlungsanleitungen, sowie der Data und Data Ethics Canvas für eine vollständige und vertrauenswürdige Verarbeitung der Daten. Zusätzlich werden alle geltenden Richtlinien der Datenschutz-Grundverordnung (DSGVO) in die Verarbeitung, sowie in den Output des Systems implementiert. Es wird im Zuge der experimentellen Entwicklung wird auch der Standard “IEEE P7001/D4 Draft Standard for Transparency of Autonomous Systems”[10] angestrebt, da das Produkt einen starken Fokus auf Nachvollziehbarkeit und Transparenz in der Entscheidungsfindung hat. Dies stellt die stabile Basis für ein System dar, welches automatisiert Hinweise und Recommendations für eine ethische und vertrauenswürdige Datenmodellierung ermöglicht.

[10] Der IEEE P7001/D4 Draft Standard for Transparency of Autonomous Systems ist ein Dokument, das von der IEEE (Institute of Electrical and Electronics Engineers) entwickelt wurde. Der Standard zielt darauf ab, die Transparenz von autonomen Systemen zu fördern.

AI REGULIERUNGEN, STANDARDS, NORMEN UND ZERTIFIZIERUNGEN

Beschreibung der Berücksichtigung der Themen EU AI-Act, Standards, Normen und Zertifizierungen bei der Produkt-/Dienstleistungsentwicklung

Als Teilnehmer in der Entwicklung von transformativen Technologien liegt es auch in unserer Verantwortung, die Risiken dieser zu erkennen und dementsprechend zu adressieren. Deswegen hat IREO die vorliegende experimentelle Entwicklung, sowie auch das bestehende System mit mehreren Assessments hinsichtlich ethischer Richtlinien für künstliche Intelligenz untersucht. Die gewonnenen Erkenntnisse in diesem Bereich werden nicht nur im Zuge der F&E-Tätigkeiten einen Einfluss haben, sondern werden im gesamten Unternehmen als Kernkompetenz integriert. Dadurch sollen sich die erarbeiteten Leitlinien durch alle Segmente des Unternehmens durchziehen. Mit dem Ziel der Entwicklung transformativer Technologien zum Vorteil für die Allgemeinheit, und gleichzeitiger Reduzierung jeglicher ethischen Risiken, welche die menschliche Autonomie und den gesellschaftlichen Einfluss betreffen, ist ein neuer Unternehmensleitsatz in IREO entstanden. Mit der vorliegenden experimentellen Entwicklung kann dieser Leitsatz erstmals mit einem praxisrelevanten System umgesetzt werden. Mit der ALTAI-Checkliste haben wir jeden der 7 Kernbereiche der europäischen Ethik-Leitlinien für vertrauenswürdige KI behandelt, und dementsprechende Handlungsschlüsse für die Entwicklung geschlossen. Das System selbst muss einen hohen Anspruch hinsichtlich dieser Faktoren erfüllen, da der Outcome des Systems selbst die Awareness im Daily Business von Entwickler*innen für diese Bereiche erhöhen soll.

Anhand der Risikopyramide haben wir entsprechende Definitionen hinsichtlich der Risikobeurteilung für die Entwicklung abgeleitet. Im Data und Data Ethics Canvas haben wir alle Bereiche in der Datenverarbeitung hinsichtlich Vollständigkeit und rechtmäßiger Verarbeitung überprüft.

Mit der “Trust Your AI” Zertifizierung des Know Centers wollen wir die Erfüllung dieser Kompetenzen in der Entwicklung und im Unternehmen von einer externen Expertenstelle bestätigen (Ongoing).

Einordnung des Projektes entlang der Risikopyramide des EU AI-Acts

Die experimentelle Entwicklung und das dadurch entstehende KI-Produkt ist in der Risikopyramide des EU AI-Acts unter dem Sektor “begrenztes Risiko” einzuordnen. Es wird kein System entwickelt, dass menschliches Verhalten beeinflusst oder kritische Institutionen und Bereiche während der Anwendung gefährdet.

Die Entwicklung und der daraus resultierende Output des KI-Systems muss spezifische Transparenzverpflichtungen erfüllen. Dies wird gewährleistet, indem das System Methoden wie SHAPValues zur transparenten Interpretierung der Entscheidungen in der Darstellung des Ergebnisses

inkludiert. Zusätzlich wird im Tool, in welchem das vorgeschlagene Datenmodell weiterbearbeitet werden kann, darauf hingewiesen, dass es sich um KI-generierte Inhalte handelt. Die Hinweise in Kombination mit den präsentierten Nachvollziehbarkeits-Werten dienen als Grundlage, damit Benutzer*innen eine fundierte Entscheidung in der Verwendung der Ergebnisse treffen können. Die Ergebnisse können zusätzlich jederzeit revidiert und somit gelöscht werden. Das System kann jederzeit während der Berechnung gestoppt und somit unterbrochen werden. Dabei werden alle hochgeladenen Daten gelöscht und es kommt zu keiner weiteren Verwendung, bis die entsprechende Aktion neu gestartet wird.

Darstellung der projektrelevanten AI Standards oder Normen

Datenschutz: Die DSGVO stellt sicher, dass die Daten unter den gegebenen Richtlinien geschützt sind.

Das bedeutet, dass Daten nur für die Zwecke verwendet werden, für die sie erhoben wurden, und dass sie nur mit Zustimmung der betroffenen Person verarbeitet werden. Zusätzlich werden Methoden zur Anonymisierung und Pseudonymisierung angewandt.

Ethik-Standards und -Richtlinien: Gemäß den Leitlinien für vertrauenswürdige KI, soll das System

• rechtmäßig - unter Einhaltung aller geltenden Gesetze und Vorschriften

• ethisch - unter Beachtung ethischer Grundsätze und Werte

• robust - sowohl aus technischer Sicht als auch unter Berücksichtigung des sozialen Umfelds sein.

Dies wird durch die Implementierung der 7 Grundsätze aus den Ethischen Richtlinien für vertrauenswürdige KI der europäischen Kommission durchgeführt. Zur methodischen Evaluierung wird/wurde hierbei die Assessment List for Trustworthy Artificial Intelligence (ALTAI) eingesetzt.

Risikoevaluierung: Die Risikoevaluierung wurde anhand der Risikopyramide durchgeführt. Konkret wurde das Assessment von Deloitte zur Definition des Risikosegments herangezogen. Aus den Erkenntnissen dieser Evaluierung wurden entsprechende Handlungskonsequenzen für die Entwicklung und den Output abgeleitet und definiert.

Datenstrategie: Zur Sicherstellung einer nachvollziehbaren und rechtmäßigen Datenstrategie wurde der Datenstrategie Canvas von www.datentreiber.de sowie der Data Ethics Canvas des Open Data Instituts herangezogen. Anhand der erhobenen Themen wurden entsprechende Ableitungen hinsichtlich Data Governance holistisch in das Konzept integriert.

AI Zertifizierung

Wir sind momentan mit der Zertifizierungsstelle Know Center GmbH in Kontakt:

Die “Trust Your AI”-Zertifizierung vom Know Center ist ein 360 Grad Audit für Systeme, welche künstliche Intelligenz verwenden. Dabei wird angefangen mit der Datenstrategie bis hin zur Inference der Modelle der holistische Prozess der Verarbeitung analysiert und überprüft.

https://trustyour.ai/