Universität Witten/Herdecke
Institute for Digital Transformation in Healthcare

DERA

Ansätze für ein digital-ethisches Risikoassessment für Data Sharing Ökosysteme

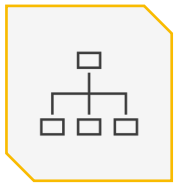
28.04.2023

Inhaltsverzeichnis

Über den Bericht

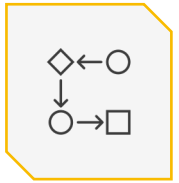
Dieser Bericht zeigt die Herleitung eines Konzepts zur Entwicklung eines digital-ethischen Risikoassessments für Data Sharing Ökosysteme auf. Er wurde im Rahmen des Förderprojekts “Health-X dataLOFT” im Auftrag der IDSA erstellt und dient als Impuls für eine zukünftige Ausarbeitung und Nutzung des Konzepts bei der Entwicklung von Data Sharing Ökosystemen.

Nach einer kurzen Einleitung in die Zusammenhänge von Data Sharing und digitaler Ethik wird das methodische Vorgehen vorgestellt. Vier Schritten folgend, werden zunächst die verwendete Literatur sowie die Zwischenergebnisse des vergleichenden Screenings vorgestellt. Im anschließenden Analyseschritt werden dann drei Analyseebenen aufgezeigt, die jeweils mit Deep Dives in einzelne Dokumente exemplifiziert werden. Basierend auf den Analyseergebnissen werden im letzten Schritt die Kernelemente eines digital-ethischen Risikoassessments für Data Sharing Ökosysteme zusammengeführt und konzeptualisiert.



Dokumentstruktur

1. Methodik
2. Analyse
    a) Kontextbestimmung
    b) Werte & Prinzipien
    c) Anforderungen
3. DERA für Data Sharing Ökosysteme



Arbeitsweise

- Vertiefende Literaturrecherche in Fachportalen und im Internet
- Verwendung von thematischen Datenbanken
- Rückgriff auf internes Expertenwissen

HEALTH-X dataLOFT

HEALTH-X dataLOFT ist ein im Rahmen des GAIA-X Förderwettbewerbs des Bundesministeriums für Wirtschaft und Klimaschutz (BMWK) bewilligtes Fördervorhaben. 14 Partner aus Forschung und Industrie arbeiten gemeinsam unter der Konsortialleitung der „Charité – Universitätsmedizin Berlin“ an einem Datenraum im Gesundheitswesen (Laufzeit: 11/2021-10/2024). Auf Basis der Entscheidungen von Bürger:innen sollen die Gesundheitsdaten in einer **l**egitimierten, **o**ffenen und **f**öderierten **dataLOFT** Plattform implementiert und gemäß Gaia-X Standards zugänglich gemacht werden.[1]

Der Gesundheitsbereich ist hochreguliert und Gesundheitsdaten werden als besonders schützenswert betrachtet. Die meisten Daten aus dem ersten und zweiten Gesundheitsbereich sind in proprietären Systemen eingeschlossen, die eine organisationsübergreifende und patient:innen-zentrierte Nutzung von Gesundheitsdaten erschweren. Über den aufzubauenden **Datenraum** und in Use Cases instanziierte Dienste sollen die individuen-zentrierte Datennutzung ermöglicht sowie innovative Geschäftsmodelle für die deutsche und europäische Gesundheitsindustrie erarbeitet werden.[2]

International Data Spaces Association (IDSA)

Die International Data Spaces Association ist Konsortialpartner im Förderprojekt HEALTH-X dataLOFT. Die Arbeiten des IDSA orientieren sich an dem Aufbau des Gaia-X Technologie Stacks zur datensouveränen Erschließung und Verknüpfung von Daten. Die IDSA hat eine Referenzarchitektur entwickelt, die die Grundlage für Datenökosysteme und Marktplätze auf der Grundlage europäischer Werte, d.h. Datenschutz und Sicherheit, bildet.[3]

Quellen: [1]https://www.health-x.org/plattform | [2]https://www.isi.fraunhofer.de/de/themen/technikfolgenabschaetzung/methodenentwicklung.html | [3]https://www.health-x.org/partner

Data Sharing & digitale Ethik

Der Zugriff auf Daten ist eine tragende Säule des digitalen Wandels und digitaler Geschäftsmodelle. Mit der Analyse dieser Daten können Zusammenhänge hergestellt, Vorhersagen getroffen und Entscheidungen gefällt werden. Daten werden daher zu einem wertvollen Asset für Unternehmen und das Interesse am branchenübergreifenden Austausch von Daten sowie dessen Monetarisierung steigt.

Auch im Kontext immer leistungsstärkerer algorithmischer Systeme („künstliche Intelligenz“) gewinnt das Teilen von Daten und das Bilden von Datenpools zunehmend an Relevanz: Das **Training von Algorithmen** auf Basis von Deep oder Reinforcement Learning Methoden erfordert riesige Datenmengen. Nicht selten ergeben sich aus der Nutzung KI-basierter Systeme jedoch nichtbeabsichtigte Implikationen.

Das seit Ende 2022 allgegenwärtige ChatGPT des Unternehmens OpenAI wirft bspw. **soziale und digital-ethische Fragen** auf, die einen möglichen Bias in den Trainingsdaten, das Missbrauchspotenzial oder die Auswirkungen auf Arbeitsplätze und den Verlust menschlicher Interaktion adressieren. Diese soziotechnologischen Implikationen zu erkennen, einzuordnen und zu minimieren, ist Teil einer digital-ethischen Betrachtungsweise. Bei digitaler Ethik geht es darum, die Grundpfeiler des digitalen Wandels – die exponentielle Steigerung der Rechenleistung, Daten und Algorithmen – sowie deren Organisation über verschiedene Praktiken ethisch zu hinterfragen und Lösungskonzepte für den Umgang mit diesen zu entwickeln. Insbesondere stehen dabei digitale Geschäftsmodelle im Fokus der Betrachtung.

Data Sharing Ökosysteme wie „HEALTH-X dataLOFT“ als **Enabler der Datenökonomie** im Allgemeinen und als mögliche Grundlage für das Training algorithmischer Systeme im Speziellen rücken daher ebenfalls in den Fokus digital-ethischer Betrachtungen. Digital-ethische Dilemma und Risiken aufzudecken und die Resilienz dieser Systeme zu beurteilen sind hier zwei entscheidende Faktoren für die nachhaltige Entwicklung und Nutzung von Data Sharing Ökosystemen, insbesondere wenn diese in kritischen Kontexten wie dem Gesundheits-, Finanz- oder Rechtswesen eingesetzt werden.

Ziele digitaler Ethik

Mögliche negative Auswirkungen auf die Gesellschaft und Einzelpersonen minimieren

Verantwortung übernehmen in Bereichen, wo Regulierungsvorhaben hinterher hinken

Vertrauen in neue und komplexe digitale Technologien schaffen

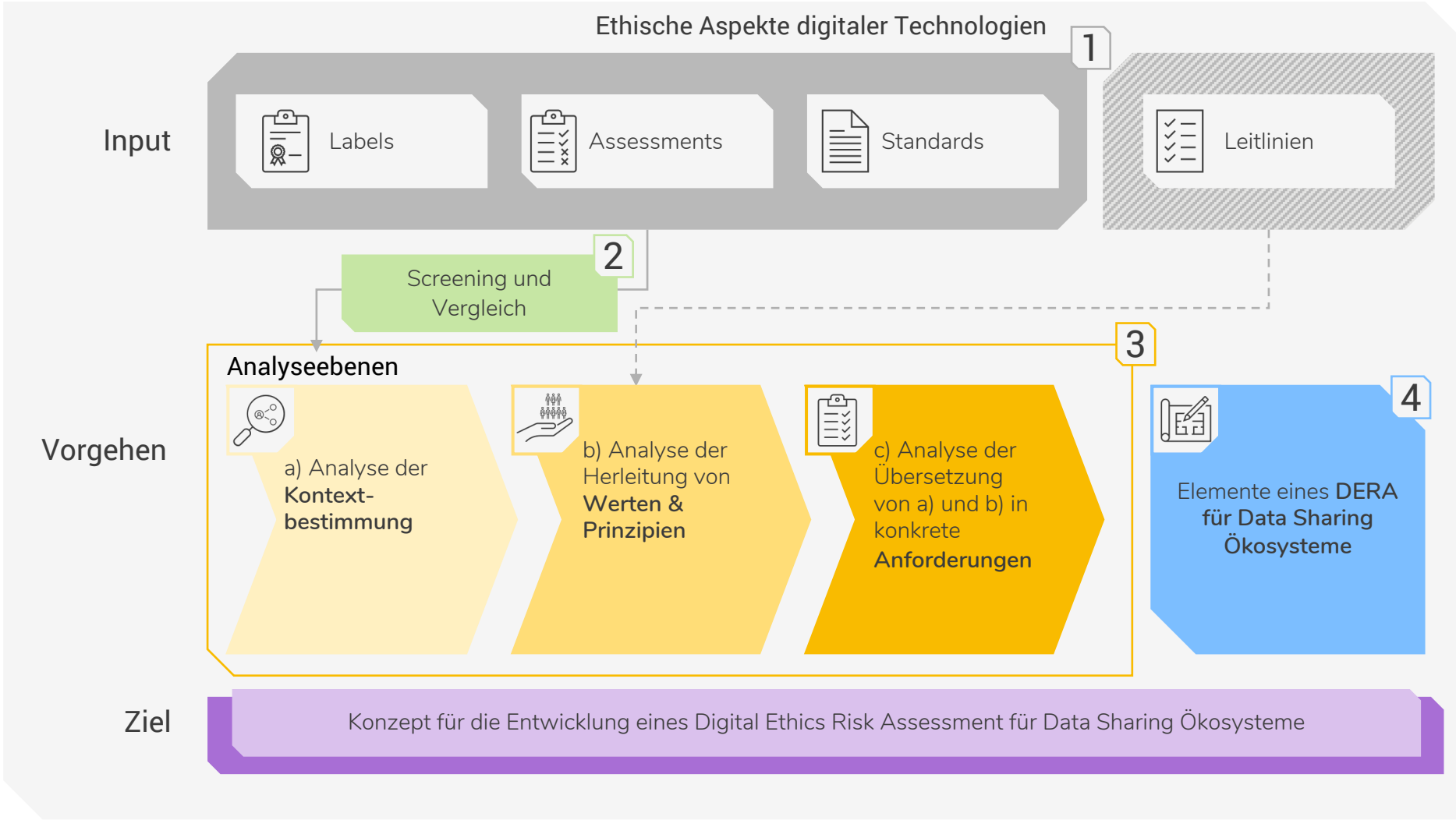
Zu diesem Zweck sollen die Voraussetzungen für ein **digital-ethisches Risikoassessment (DERA)** geprüft werden, das unterschiedliche Data Sharing Ökosysteme systematisch untersuchen und mögliche digital-ethische Risiken sowie Maßnahmen zu deren Minimierung aufzeigen kann. Funktion eines solchen DERAs ist es, nichtbeabsichtigte Implikationen erkennen und Ansatzpunkte für ein menschenzentriertes und gesellschaftsorientiertes Design auf den Ebenen der Technologie sowie der Governance finden zu können. Die Bewertung des etwaigen digital-ethischen Risikos soll zu verschiedenen Zeitpunkten der Entwicklung durchführbar sein und ermöglichen, digital-ethische Aspekte zu dokumentieren und über die Zeit zu vergleichen.

Um den potenziellen Aufbau und die Komponenten eines DERA für Data Sharing Ökosysteme wie HEALTH-X dataLOFT zu ermitteln, wurde eine Umfassende Literaturrecherche und -analyse durchgeführt, wie auf der folgenden Seite dargestellt wird.

Methodik

Übersicht

Ein digital-ethisches Risikoassessment für Data Sharing Ökosysteme sollte auf möglichst etablierte Prozesse zurückgreifen, um anschlussfähig zu sein. Um diese Prozesse zu identifizieren wurden zunächst bestehende Labels, Assessments und Standards zu digitaler Ethik und verwandter Bereiche gesammelt [1] und im Hinblick auf Kriterien und Methodik verglichen [2]. So konnten drei Analyseebenen [3] identifiziert werden, zu denen jeweils Deep Dives durchgeführt wurden. Während für die Analyse der Art und Weise der Kontextexploration und der Anforderungserstellung auf den Korpus aus Labeln, Assessments und Standards zurückgegriffen werden konnte, bedurfte das Thema „Werte & Prinzipien" weiterer Quellen in Form von selbstverpflichtenden Leitlinien. Die Zusammenführung der Ergebnisse ergab ein Bild davon, wie ein digital-ethisches Risikoassessment für Data Sharing Ökosysteme aussehen könnte [4].



Datenmaterial (Input)

Das digital-ethische Design von digitalen Produkten, Ökosystemen und algorithmischen Systemen wird unter unterschiedlichen Begriffe behandelt, die im Kern auf dieselben Aspekte abzielen: „Digital Ethics“, “Ethical AI” „CDR“, „Value-based Design“; „Digital Trust“ uvm. Im Kontext dieser Begriffe gibt es zahlreiche Labels, Assessments und Standards, die sowohl vor, während und nach dem Entwicklungsprozess digitaler Innovationen zum Einsatz kommen sollen. Ein Großteil dieser Dokumente fokussiert auf künstliche Intelligenz, denn nicht erst seit ChatGPT werden die gesellschaftlichen Implikationen des Einsatzes von (teil)automatisierten Systemen auch in der breiten Öffentlichkeit diskutiert. Entsprechend groß ist das Interesse am angemessenen Umgang mit dieser Technologie. Labels, Assessments und Standards, die explizit auf den digital-ethischen Umgang mit Daten und Data Sharing abzielen, sind hingegen seltener zu finden.

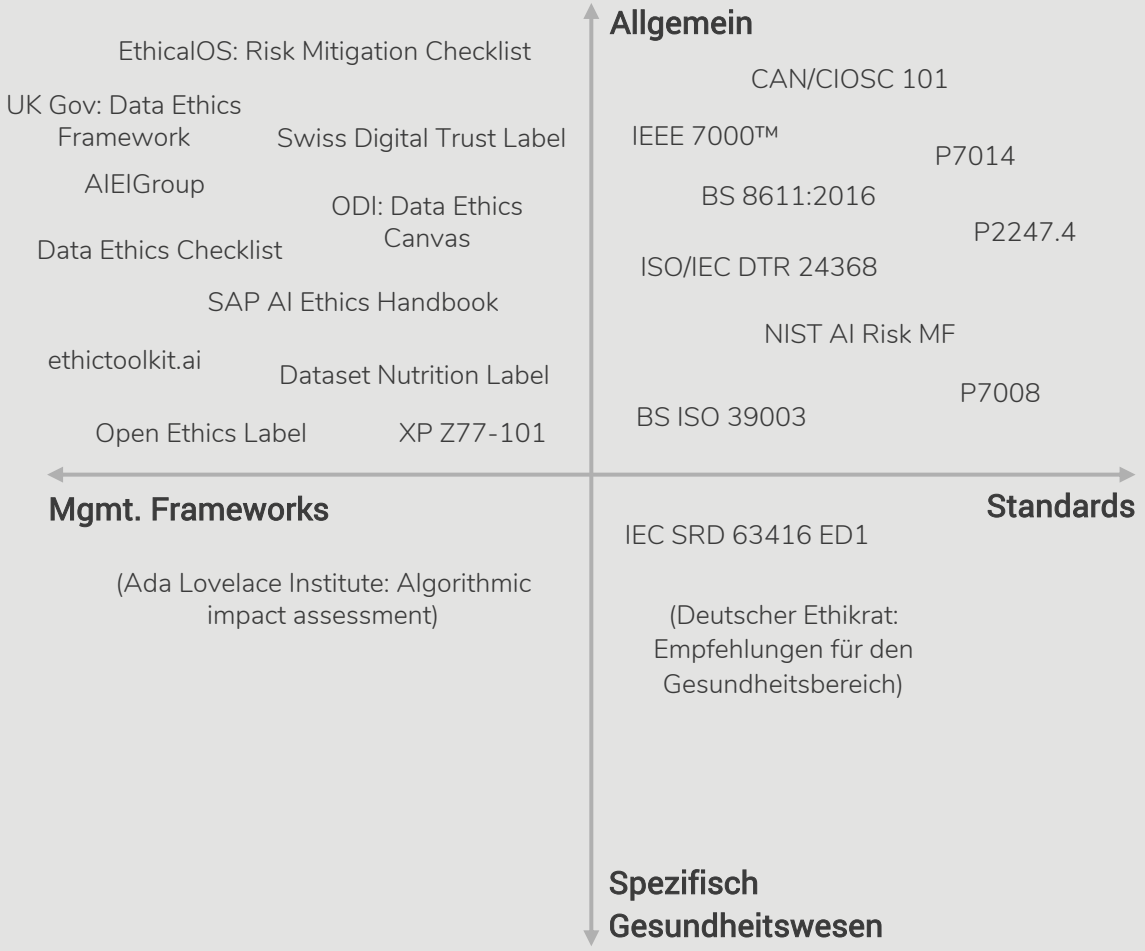
Die Matrix rechts zeigt einen Ausschnitt der gesammelten Dokumente (N=65), die sich auf funktionaler Ebene (X-Achse) zwischen Management Frameworks und Standards bewegen. Aufgrund der Verortung von HEALTH-X dataLOFT im Gesundheitswesen wurde neben sektorenunabhängigen Labels, Assessments und Standards auch nach gesundheitsspezifischen Dokumenten gesucht.

Für den Gesundheitssektor konnten keine spezifischen Management Frameworks identifiziert werden. Allerdings lassen sich die meisten allgemeinen Frameworks auf Use Cases im Gesundheitssektor anwenden. Ein konkretes Beispiel ist der Report “Algorithmic impact assessment: a case study in healthcare” des britischen Ada Lovelace Institutes, weshalb er in die Matrix aufgenommen wurde.

Auch für Standards konnten kaum gesundheitsspezifische Beispiele gefunden werden. Der hier aufgenommene IEC SRD 63416 ED1 („Ethical considerations of Artificial Intelligence (AI) when applied in the Active Assisted Living (AAL) context”) ist mit seinem Fokus auf “Active Assisted Living“ nur im weitesten Sinne dem Gesundheitssektor zuzuordnen. Die Empfehlungen des deutschen Ethikrats zu Big Data sind zwar gesundheitsbezogen, bewegen sich aber auf einer Metaebene ohne direkte Anwendbarkeit in der Praxis.

Verteilung der untersuchten Dokumente

Auszug (n=23)



Labels, Assessments und Standards (Screening und Vergleich)

Nach der Identifizierung relevanter Labels, Assessments und Standards wurden die jeweiligen Dokumente gescreent und miteinander verglichen. Der Vergleich ergab drei zentrale Erkenntnisse, die für ein digital-ethisches Risikoassessment von Bedeutung sind.

Die vorhandenen Labels, Assessments und Standards zeigen die Komplexität der betreffenden Systeme auf, die eine **Berücksichtigung des Kontextes** erfordert. Je nach dem, aus welcher Perspektive bspw. auf einen Datensatz geschaut wird, sind andere Analyseergebnisse denkbar. Die Daten aus einer App zur Corona-Kontaktdatenverfolgung könnten so z. B. auch für polizeiliche Ermittlungen genutzt werden. Während das eine ein sinnvoller Mechanismus während einer Pandemie darstellt, ist letzteres je nach politischem System des jeweiligen Landes kritisch zu betrachten. Bei einem digital-ethischen Risikoassessment sollte daher die Kontextbestimmung mitgedacht werden.

Zudem haben diese Ansätze gemeinsam, dass sie auf ausgewählte **Werte & Prinzipien** zurückgreifen, um die digital-ethische Umsetzung zu leiten oder zu bewerten. Diese Werte & Prinzipien können sich aufgrund soziokultureller Differenzen je nach Ursprungsland unterscheiden. Jedoch spricht vieles dafür, dass zumindest für Europa eine einheitliche Wertebasis gefunden werden kann, auf der sich auch ein digital-ethisches Risikoassessment aufbauen lässt. Die idigiT Studie „Zwischen Unternehmenswerten und Operationalisierung“[1] brachte bereits Erkenntnisse zu Werten & Prinzipien zutage, weshalb im weiteren Verlauf der Analyse auf diese Ergebnisse zurückgegriffen wurde.

Werte & Prinzipien dienen häufig vor allem als abstrakte Orientierungspunkte. Um sie im digital-ethischen Design von digitalen Produkten, Ökosystemen und algorithmischen Systemen praktisch umzusetzen, müssen sie in konkrete **Anforderungen** übersetzt werden. Der oben angeführte Kontext hat hierbei Einfluss auf diese Anforderungen. Dies ist bei der Evaluation mit Hilfe eines digital-ethischen Risikoassessments zu berücksichtigen, weshalb entsprechende Anforderungen zu formulieren sind.

Quelle: [1]https://www.transforming-healthcare.com/insights/

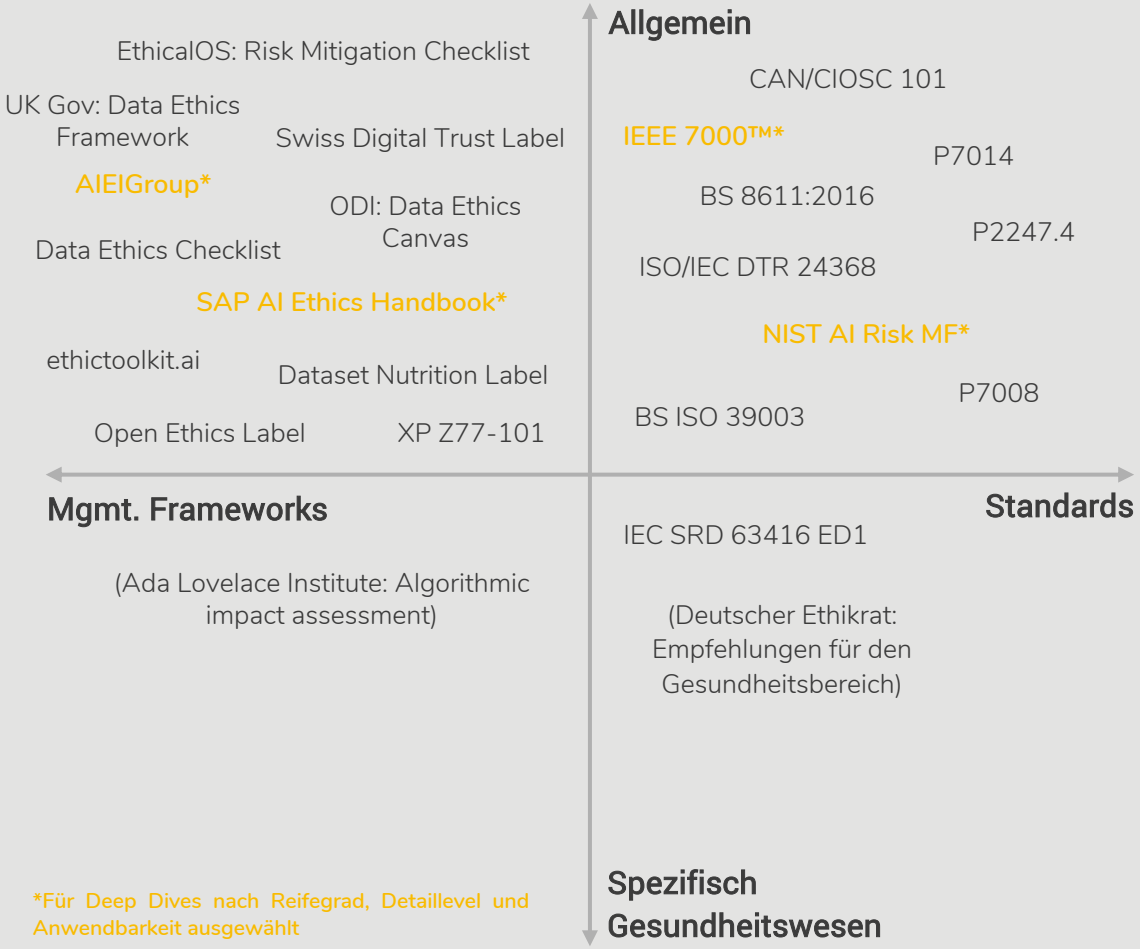
Ausgewählte Beispiele für Labels, Assessments und Standards

Zentrale Erkenntnisse

- Der Kontext spielt eine entscheidende Rolle
- Werte & Prinzipien dienen als Maßstab für die digital-ethische Evaluation
- Werte & Prinzipien müssen in konkrete Anforderungen übersetzt werden

Auswahl der Dokumente für die Deep Dive Analyse

Auszug (n=23)



*Für Deep Dives nach Reifegrad, Detaillevel und Anwendbarkeit ausgewählt

Analyseebenen

Für die Entwicklung eines digital-ethischen Risikoassessments, welches auch auf Data Sharing Ökosysteme anwendbar ist, zeigte der Vergleich von bestehenden Labels, Assessments und Standards drei Analyseebenen auf. Zunächst stellt sich die Frage, wie der Kontext eines Data Sharing Ökosystems erfasst werden kann und ab wann ein Kontext als „kritisch“ einzustufen ist und ein digital-ethisches Assessment notwendig wird. Darüber hinaus ist zu ermitteln, wie und welche Werte & Prinzipien herangezogen werden sollten und woran deren Einhaltung festgemacht werden kann. Dazu werden im Folgenden mehrere Deep Dives zu einzelnen Dokumenten auf den drei Analyseebenen vorgestellt.

Analyse: Kontextbestimmung

Der Auswahlprozess von Datensätzen beginnt im Allgemeinen mit einer Geschäfts- oder Forschungsfrage, und die Bewertung des Datensatzes erfolgt im Kontext dieser Frage. Entsprechend muss zur Evaluation des digital-ethischen Risikos der Kontext systematisch erfasst werden. Im Rahmen von KI-Systemen lassen sich unterschiedliche Methoden zur Bestimmung des Kontextes finden. Einige dieser Ansätze werden im Folgenden analysiert.

Analyse: Werte & Prinzipien

Werte & Prinzipien werden vorallem in selbstverpflichtenden Leitlinien in unterschiedlichen Kontexten beschrieben. Um relevante Werte & Prinzipien für Data Sharing Ökosysteme wie HEALTH-X dataLOFT bestimmen zu können, wird auf unterschiedliche vergleichende Studien zu diesem Thema zurückgegriffen. Insbesondere die idigiT Studie „Zwischen Unternehmenswerten und Operationalisierung“ ermöglicht aufschlussreich Einblicke und wird daher vorgestellt.

Analyse: Anforderungen

Welche Anforderungen sich aus Werten & Prinzipien im jeweiligen Kontext ergeben, ist ausschlaggebend für die praktische Anwendbarkeit des digital-ethischen Risikoassessments. Entsprechend werden zwei ausgewählte Ansätze zur Anforderungsentwicklung vorgestellt.

a) Analyse: Kontextbestimmung 3

Executive Summary

Die gefundenen Ansätze zur Ermittlung des Kontextes nehmen alle Bezug auf die Entwicklung und den Betrieb von KI-Systemen. Das Spektrum ist dabei groß und reicht von Entscheidungsbäumen (siehe SAP) über die Einteilung in Risikoklassen (siehe AIEIGroup) bis hin zu qualitativen Stakeholderansätzen (siehe IEEE oder NIST). Im Hinblick auf die Methodik zur Kontextermittlung lassen sich einige Erkenntnisse ableiten, die auf Data Sharing Ökosysteme übertragen werden könnten.

Zunächst haben die Konzepte gemein, dass sie sich stark auf den **Anwendungsfall** und weniger auf die technische Funktionsweise von KI-Systemen beziehen. Mit Kontext ist also derjenige Kontext gemeint, den der geplante Anwendungsfall definiert. Dazu werden u. a. folgende Fragen gestellt: Welchen Zweck erfüllt das System? Welcher Lebensbereich ist betroffen? Handelt es sich um besonders schützenswerte Anwendungsfälle, bspw. in Gesundheit oder Politik? Welche Absichten haben die Entwickler bzw. die Betreiber des Systems?

Ein wichtiges Element der Kontextermittlung sind die betroffenen **Stakeholder**. Hier geht es um Fragen wie: Auf wen hat der Einsatz des Systems direkte und indirekte Auswirkungen? In welchem Ausmaß bestehen Risiken für diese Stakeholder? Dabei stehen nicht nur individuelle Stakeholdergruppen, sondern oft auch die Implikationen für die Gesellschaft als Ganzes im Fokus. Die systematische Bestimmung der Stakeholder eines Anwendungsfalls hilft entsprechend dabei, zentrale Elemente des Kontextes zu erfassen.

Häufig zu beobachten ist die Einteilung in **Risikoklassen**. Für jede Klasse werden Kriterien definiert, nach denen der Kontext des Anwendungsfalls eingestuft wird – mit unterschiedlichen Folgen. So kann es vorkommen, dass der Kontext so unkritisch ist, dass ein anschließendes detailliertes digital-ethisches Risikoassessment nicht notwendig ist oder aber dass in als besonders kritisch festgelegten Fällen ein Stopp der Entwicklung notwendig wird. Solche „Red-Line-Cases“ sind von vornherein ausgeschlossen und erübrigen ebenfalls eine detailliertes Evaluation. Die anfänglichen Einteilung eines Anwendungskontextes in festgelegte Risikoklassen ist somit ein hilfreicher Filter für ein DERA.

Zentrale Erkenntnisse

- Fokussierung auf den Anwendungsfall
- Betroffene Stakeholder sind ein wichtiges Element der Kontextermittlung
- Einteilung in Risikoklassen

AIEIGroup | From Principles to Practice (1/2) - Risikomatrix

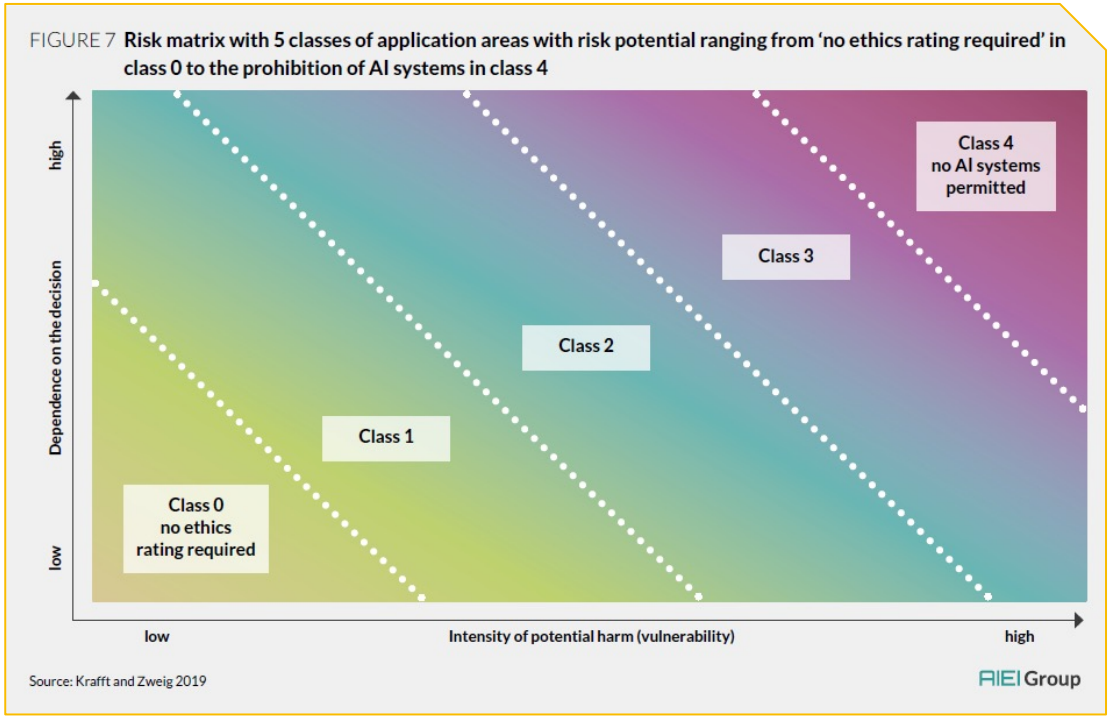
Mit der Publikation „From Principles to Practice“ hat die „AI Ethics Impact Group (AIEIGroup)“ 2020 ein Konzept für ein Framework vorgestellt, das speziell darauf ausgerichtet ist, ethische Grundsätze bei der Entwicklung, Implementierung und Evaluierung von KI-Systemen in der Praxis umzusetzen.

Wichtiger Teil der Evaluierung ist die Berücksichtigung des Kontextes, in dem KI-Systeme verwendet werden. Wie Werte wie Gerechtigkeit und Transparenz in der Praxis umgesetzt und priorisiert werden, hängt bis zu einem gewissen Grad vom Anwendungsbereich und dem kulturellen Kontext ab, in dem ein KI-System operiert. Ein System, das im Justizbereich eingesetzt wird, muss zwangsläufig ein höheres Maß an Privatsphäre und Fairness aufweisen als ein System, das in der Organisation der industriellen Produktion eingesetzt wird. Für die Anwendung im medizinischen Bereich könnte die Zuverlässigkeit als wichtigster Wert angesehen werden.

Die Einstufung des digital-ethischen Risikos orientiert sich an dem gesamten Schadenspotenzial, das ein KI-System in seinem jeweiligen gesellschaftlichen Prozess verursachen kann. Da der Fokus auf KI-Systemen liegt, sind die Bewertungskriterien dieses Schadenspotenzials thematisch eng gefasst und können nicht ohne weiteres auf Data Sharing Ökosysteme übertragen werden: Entscheidend für die Bewertung dieses Potenzials sind die Intensität des potenziellen Schadens des KI-Systems und die Abhängigkeit der betroffenen Person(en) von der jeweiligen Entscheidung. Beide Kriterien dienen als Dimensionen einer Risikomatrix nach Krafft/Zweig[2], anhand derer das jeweilige KI-System bewertet werden kann.

Für die x-Achse ist der entscheidende Aspekt der potenzielle Schaden, der die Bewertung der Intensität betrifft, mit der ein KI-System potenziell Menschen, Organisationen und der Gesellschaft schaden könnte. Dazu müssen die Auswirkungen auf eine bestimmte Anzahl von Menschen oder den Zugang zu Ressourcen betrachten und geprüft werden, ob die Gesellschaft als Ganzes gefährdet ist. Hier stellen sich Fragen nach dem Einfluss des Systems auf Grundrechte, Gleichheit und soziale Gerechtigkeit.

Risikomatrix mit unterschiedlichen Risikoklassen[1]



Die y-Achse zeigt die Abhängigkeit der potenziell betroffenen Parteien von der algorithmischen Entscheidung und damit die Möglichkeiten zur Vermeidung des auf der x-Achse angegebenen potenziellen Schadens. Bei der Bewertung der Abhängigkeit von der Entscheidung sind Kontrolle, Austauschbarkeit und Korrekturmöglichkeiten zu berücksichtigen. Sind Menschen an der Entscheidung beteiligt oder agiert das KI-System autonom? Lassen sich KI-Systeme gegen andere austauschen oder entsteht eine Abhängigkeit? Können Entscheidungen im Nachgang korrigiert werden?

Quellen: [1]https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/WKIO_2020_final.pdf | [2]19-01-22_zweig_krafft_transparenz_adm-neu.pdf (vzbv.de)

AIEIGroup | From Principles to Practice (2/2) - Risikoklassen[1]

Je nach Ausprägung der Schadensintensität und der Abhängigkeit von Entscheidungen schlagen die Autoren verschiedene Risikoklassen für KI-Systeme vor.

Klasse 0 (keine ethische Bewertung erforderlich)

KI-Systeme mit geringem Gesamtschadenspotential wie bspw. bei personalisierter Auswahl von Werbung für Kleidung.

Klasse 1

KI-Systeme, deren potenzieller Schaden und die Abhängigkeit von der Entscheidung eine bestimmte Schwelle überschreiten, z. B. bei personalisierter Auswahl von Suchergebnissen. Erste ethische Bewertungen etwa in Hinblick auf Transparenz sollten erfolgen.

Klasse 2

KI-Systeme, deren Intensität des potenziellen Schadens und der Abhängigkeit von der Entscheidung hoch sind, z. B. bei personalisierter Auswahl von Werbung für Stellenangebote. Eingabedaten sollten vollständig offengelegt, und Informationen über die Qualität des Systems überprüfbar sein.

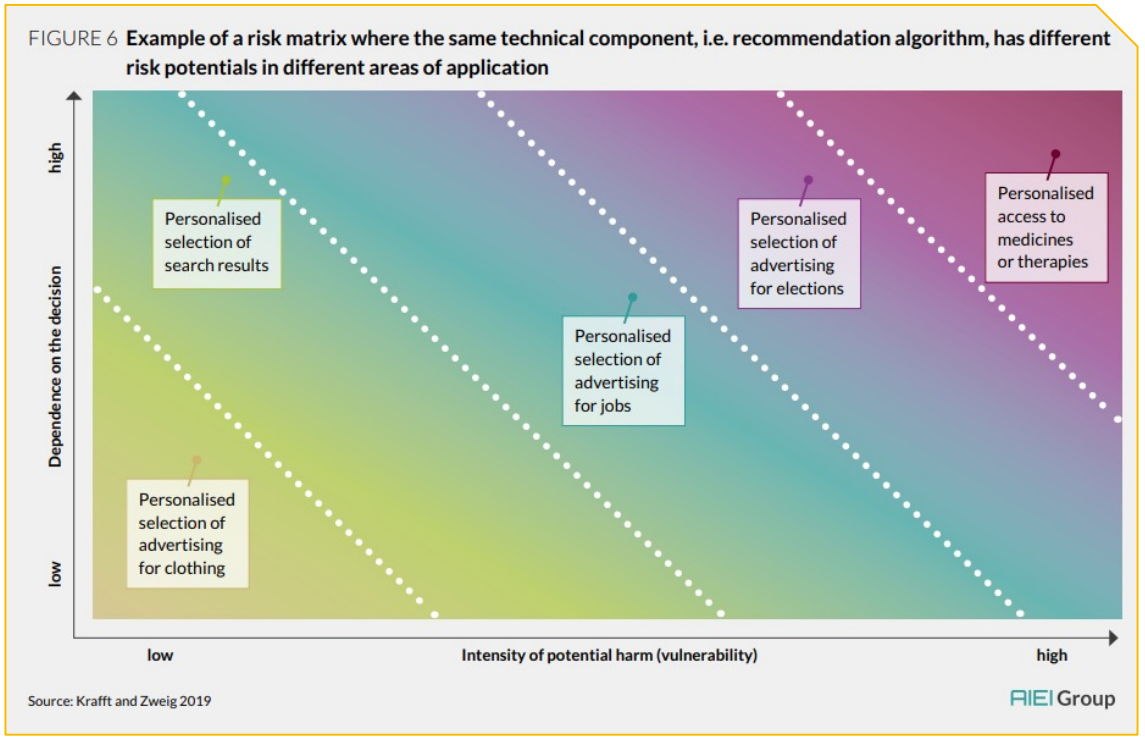
Klasse 3

KI-Systeme, in denen entweder der potenzielle individuelle/gesellschaftliche Schaden von Entscheidungen des KI-Systems sehr hoch ist, das System ohne das Wissen der Betroffenen eingesetzt wird, oder in denen es gegen deren Erwartungen an das System operieren kann, z. B. personalisierte Auswahl von Wahlwerbung.

Klasse 4

Einige KI-Systeme haben ein so hohes Gesamtschadenspotenzial, dass sie überhaupt nicht mit einer maschinellen Lernkomponente eingesetzt werden sollten, z. B. autonome Waffensysteme oder Entscheidungen über den Zugang zu Medikamenten und Therapien.

Beispiel einer Risikomatrix aus dem Bereich „Personalisierung“[1]



Quelle: [1]https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/WKIO_2020_final.pdf

SAP AI Ethics Handbook

Das deutsche Softwareunternehmen SAP verfügt über eine globale KI-Ethikrichtlinie, die sicherstellen soll, dass die Entwicklung und der Einsatz von KI-Systemen der SAP im Einklang mit den etablierten Leitprinzipien und zentralen Unternehmenswerten erfolgt. Ein KI-Ethikhandbuch[1] soll SAP-Mitarbeiter bei der Umsetzung der KI-Ethikrichtlinie in allen Phasen des KI-Entwicklungszyklus unterstützen. Die Anforderungen der Richtlinie sind den jeweiligen Phasen der SAP AI Factory Entwicklungsprozesse zugeordnet.

Zu Beginn der Prozesse werden zu entwickelnde Use Cases in drei Kategorien eingeteilt:

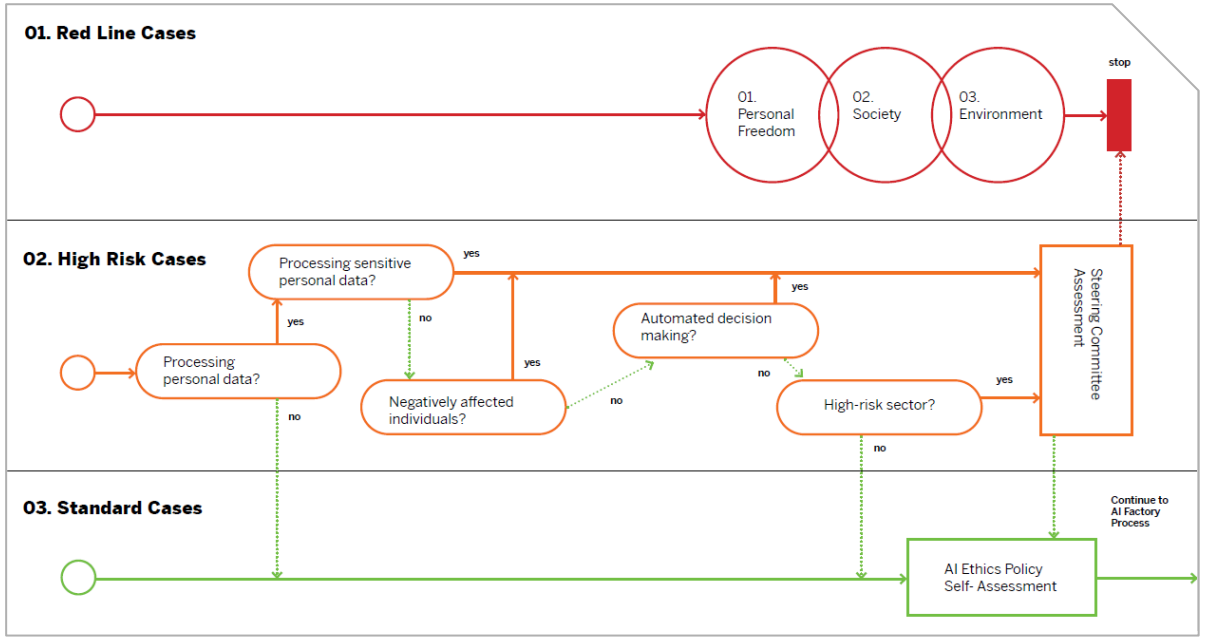
- **Red Line Cases**
- **High Risk Cases**
- **Standard Cases**

Diese Einteilung betrachtet den Kontext des jeweiligen Use Cases. Unter **Red Line Cases** fallen diejenigen Anwendungsfälle, die gegen die persönliche Freiheit (Menschliche Überwachung, Diskriminierung, Deanonymisierung), die Gesellschaft (Manipulation, Untergraben von Debatten) oder die Umwelt (Umweltschädigung) gerichtet sind. Wenn ein Anwendungsfall für diese Zwecke entwickelt wurde, besteht die Pflicht, die Entwicklung, den Einsatz und den Verkauf dieses Anwendungsfalls einzustellen.

High Risk Cases sind hingegen innerhalb von SAP nicht verboten; müssen aber zunächst einen Bewertungsprozess des AI Ethics Steering Committee durchlaufen, bevor sie weiterentwickelt, eingesetzt und verkauft werden können. Ob es sich um einen solchen Anwendungsfall handelt, wird anhand eines Entscheidungsbaumes festgestellt (siehe Grafik). Relevante Kriterien sind die Verarbeitung (sensibler) personenbezogener Daten, automatisierte Entscheidungsfindung, negativer Einfluss auf Personen und die Zugehörigkeit zu einem risikoreichen Anwendungsgebiet wie bspw. Gesundheitswesen, Personalwesen, Verwaltung und Betrieb kritischer Infrastrukturen oder Strafverfolgung. Werden keine personenbezogenen Daten verarbeitet oder ist der Einsatzbereich kein Hochrisikosektor, so handelt es sich um **Standard Cases**, die lediglich einem KI-Ethikrichtlinien-Self-Assessment unterzogen werden müssen. In allen anderen Fällen ist ein Bewertungsprozess des AI Ethics Steering Committee von Nöten, um über die Weiterentwicklung, den Einsatz und den Verkauf entsprechender Use Cases zu entscheiden.

Ein Entscheidungsbaum zur Kategorisierung des Kontextes von Use Cases, hier im Bezug auf KI-Systeme, ist auch für die Entwicklung eines digital-ethischen Assessments für Data Sharing Ökosysteme wie HEALTH-X dataLOFT denkbar. Die Herausforderung besteht darin, relevante Kriterien für zu definieren um sinnvolle Unterscheidungen treffen zu können.

SAP Risiko-Klassifizierung & Bewertungsprozess[1]



Quelle: [1]https://www.sap.com/documents/2023/03/7211ee96-647e-0010-bca6-c68f7e60039b.html

IEEE 7000™ (1/2) – Übersicht Kontextexplorationsprozess

Der technische Standard IEEE 7000™ der Vereinigung “Institute of Electrical and Electronics Engineers” ist ein Ethikstandard für intelligente und autonome Systeme. Er soll dabei helfen, dass bei der Entwicklung technologischer Systeme Ethik von Anfang an mitgedacht und beachtet wird, indem er menschliche und soziale Werte in die traditionelle Systemtechnik und -gestaltung integriert. Er richtet sich sowohl an große als auch an kleine und mittlere Unternehmen, die bei der Konzeption, der Entwicklung oder dem Betrieb von KI- oder anderen technischen Systemen umfassendere ethische Wertkriterien und Belange berücksichtigen wollen. Die Limitationen des IEEE 7000™ liegen darin, dass er nicht vorschreibt, was ethisch und was unethisch ist, und auch dem einzelnen Entwickler keine ethische Anleitung für seine persönliche ethische Beurteilung bietet. Genauso schreibt er Organisationen keine spezifischen ethischen Richtlinien vor.

Für die ethische Entwicklung von intelligenten und autonomen Systemen schlägt der Standard eine Reihe von Prozessen vor, die wiederum in mehrere Teilschritte unterteilt sind. Ein grundlegendes Element des ethischen Entwicklungsprozesses ist die Erfassung des Kontextes des jeweiligen „System of Interest“ (SOI), auf das hier genauer eingegangen wird.

Betriebskonzept (ConOps) und Prozess der Kontextexploration

Der Zweck des ConOps- und Kontextexplorationsprozesses besteht darin, zu definieren, wie ein System aus der Sicht der Nutzenden und des Nutzungskontextes, der Stakeholdergruppen und des Potenzials für ethischen Nutzen oder Schaden funktionieren soll. Der Prozess der Kontextexploration entwickelt ein Verständnis für das ethische Umfeld, in dem das SOI und sein Betrieb Auswirkungen auf die Stakeholder haben. Wenn der Kontext erforscht und für die Zukunft eines Systems ins Auge gefasst wird, sollte dies unter der Annahme geschehen, dass das System in großem Maßstab umgesetzt wird, d. h. dass es erhebliche Auswirkungen auf die Ziel-Stakeholder und -Märkte hat. Dabei sollten Beschreibungen von Anwendungsfällen oder ConOps einen langen Zeithorizont (d. h. 10 bis 20 Jahre) berücksichtigen. Folgende Schritte werden zur Klärung des Kontextes empfohlen:

Die 7 Hauptschritte des Kontextexplorationsprozesses

1. **Verständnis für den beabsichtigten Nutzungskontext des Systems bzw. der Systeme.**
2. **Identifizierung von Stakeholdern, die an dem System interessiert oder davon betroffen sein könnten.**
3. **Beschreibung und Analyse der technischen und organisatorischen Kontrolle.**
4. **Zugang zu den Grundsystemen erhalten.**
5. **Sammlung verfügbarer sozialer, rechtlicher und ökologischer Informationen über die Machbarkeit des System of Interest (SOI).**
6. **Lücken und Diskrepanzen zwischen den Annahmen und den Ergebnissen ermitteln und beseitigen.**
7. **Vollständige Konzept- und Kontextanalysen.**

Zu jedem der Schritte gibt es Teilaufgaben, die durchgeführt werden sollten. So wird der Schritt „**Verständnis für den beabsichtigten Nutzungskontext des Systems / der Systeme”** durch die Teilaufgaben “1) Beschreiben Sie den Kontext der derzeitigen Vorgänge, die durch das künftige System ersetzt oder verändert werden sollen.“ und „2) Identifizieren Sie einen oder mehrere tatsächliche oder mögliche Systemnutzungskontexte und stellen Sie diese in geeigneter Weise dar.“ beschrieben.

Quelle: IEEE-7000-2021 S. 35-38

IEEE 7000™ (2/2) - Prozess der Kontextexploration

Der IEEE 7000™ legt besonderen Wert auf die Fokussierung beteiligter Stakeholder und Stakeholder-Gruppen. Wesentlicher Teil der Erfassung des Kontextes ist die „**Identifizierung von Stakeholdern, die an dem System interessiert oder davon betroffen sein könnten**“. Die entsprechenden Teilaufgaben sehen folgendes vor:

1) Identifizierung der relevanten Stakeholder, einschließlich

   I. organisatorische Vertreter, die die Innovationsbemühungen vorantreiben
   II. ein vielfältiges Spektrum von Stakeholdern, die sowohl kritisch sind als auch eine breite Streuung von technischen Fähigkeiten und ethischer Wertorientierung aufweisen
   III. Interessenvertreter für indirekte Stakeholder
   IV. Fachleute, die den sozialen Kontext des SOI verstehen
   V. Fachleute, die die technischen Möglichkeiten des SOI verstehen
   VI. Interessenvertreter, die auf transparente Weise ausgewählt werden
   VII. Potenzielle Nutzende des Systems, die gegebenenfalls an den Prozessen teilnehmen; insbesondere Endnutzende aus dem Markt oder den Weltregionen, in denen das System eingesetzt wird oder werden soll
   VIII. Institutionen, die von dem SOI betroffen sind, oder deren Vertreter, je nachdem
   IX. Gegebenenfalls Vertreter der Zivilgesellschaft und Rechtsbeistände

2) Identifizierung von Stakeholdergruppen.

Einzelne Vertreter der identifizierten Stakeholder sollten ausgewählt und möglichst auch für die Zusammenarbeit in den weiteren Prozessen des IEEE 7000™ herangezogen werden.

Der IEEE7000™ zeigt, dass das Spektrum relevanter Stakeholder, die entscheidenden Einfluss auf den Kontext eines intelligenten oder autonomen Systems haben, groß ist. Entsprechend sind ausreichende Kapazitäten für eine digital-ethisches Assessment einzuplanen.

Ergebnisse der erfolgreichen Implementierung

- ✓ Der geplante Nutzungskontext des SOI ist beschrieben.
- ✓ Die Stakeholder, die während des gesamten Lebenszyklus mit dem geplanten System zu tun haben, sind identifiziert und ihre Vertreter sind ausgewählt.
- ✓ Konzepte der Kontrolle über das SOI sind identifiziert und analysiert.
- ✓ Relevante Informationen über die soziale, rechtliche und ökologische Machbarkeit des SOI werden gesammelt.
- ✓ Die Aktivitäten und Aufgaben dieses Prozesses werden mit anderen Aufgaben integriert, die den Kontext und die anfänglichen ConOps für das SOI definieren.
- ✓ Es wird festgestellt, ob es notwendig ist, die potenziellen Schäden und Vorteile des Systemkonzepts für ethische Werte weiter zu untersuchen.

Quelle: IEEE-7000-2021 S. 35-38

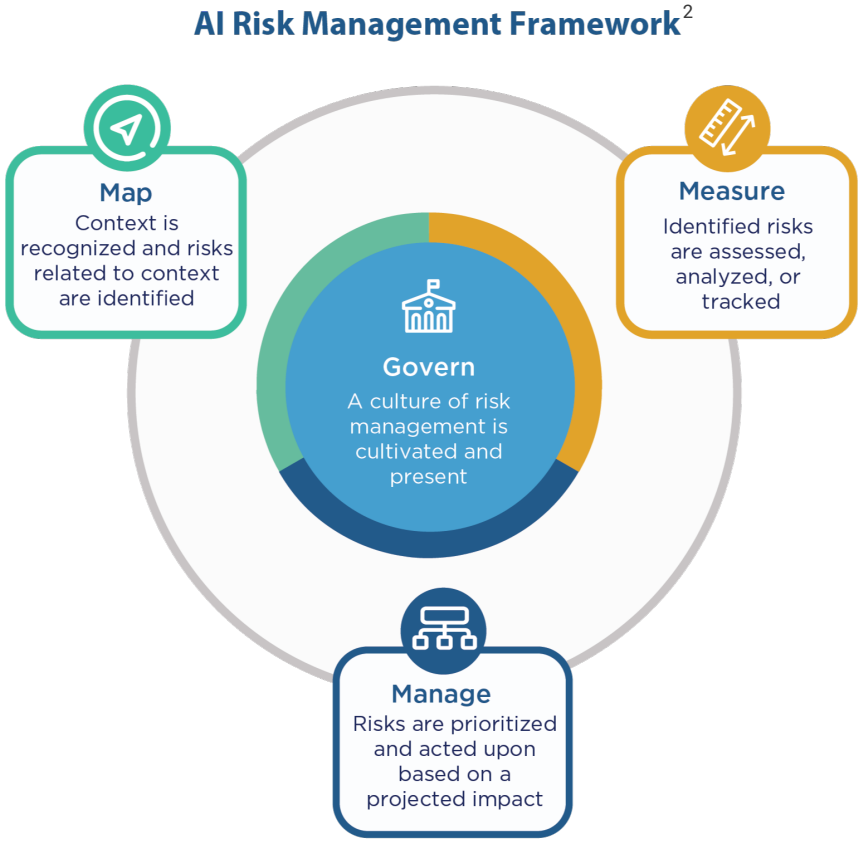
NIST AI Risk Management Framework (1/2) - Übersicht

Das 2023 veröffentlichte "AI Risk Management Framework"[1] (AI RMF) des US-amerikanischen National Institute of Standards and Technology (NIST) ist für die freiwillige Nutzung gedacht und soll die Fähigkeit verbessern, Überlegungen zur Vertrauenswürdigkeit in die Konzeption, Entwicklung, Nutzung und Bewertung von KI-Produkten, -Dienstleistungen und -Systemen einzubeziehen. Das AI RMF wurde in einem offenen, transparenten, multidisziplinären und von mehreren Stakeholdern getragenen Verfahren über einen Zeitraum von 18 Monaten und in Zusammenarbeit mit mehr als 240 beitragenden Organisationen aus der Privatwirtschaft, der Wissenschaft, der Zivilgesellschaft und der Regierung entwickelt.

Das Framework ist in die vier Aufgabenbereiche "Govern", "Map", "Measure" und "Manage" aufgeteilt. Die einzelnen Bereiche umfassen jeweils mehrere empfohlene Schritte. Die Ergebnisse des Map-Aufgabenbereichs bilden die Grundlage für die Bereiche „Measure" und „Manage", weshalb er am Anfang des Frameworks steht. Auch die Autoren des AI RMF sind davon überzeugt, dass ohne Kontextwissen und Bewusstsein für Risiken innerhalb der identifizierten Kontexte ein Risikomanagement nur schwer möglich ist. Daher wird im Folgenden ein Blick auf die im "Playbook" definierten ersten sechs Schritte des Map-Aufgabenbereichs geworfen.

NIST AI RMF Playbook[2]

Das Playbook enthält Vorschläge für Maßnahmen zur Erreichung der im AI Risk Management Framework (AI RMF) Core dargelegten Ergebnisse. Die Vorschläge sind auf jede Unterkategorie innerhalb der vier AI RMF-Bereiche (Govern, Map, Measure, Manage) ausgerichtet. Das Playbook ist weder eine Checkliste noch eine Abfolge von Schritten, die in ihrer Gesamtheit befolgt werden müssen. Die Vorschläge im Playbook sind freiwillig. Unternehmen können diese Informationen nutzen, indem sie so viele – oder so wenige – Vorschläge übernehmen, wie für ihren Anwendungsfall oder ihre Interessen zutreffen.



Quellen: [1]https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF | [2]https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook

NIST AI RMF Playbook (2/2) - Map-Aufgabenbereiche 1.1 - 1.6[1]

Map 1.1

Ziel: Der beabsichtigte Zweck, die potenziell nutzbringenden Anwendungsfälle, die kontextspezifischen Gesetze, Normen und Erwartungen sowie die voraussichtlichen Rahmenbedingungen, unter denen das KI-System eingesetzt wird, werden verstanden und dokumentiert. Fragen, die gestellt werden, sind u. a.:

- *Inwieweit ist der Output der einzelnen Komponenten für den operativen Kontext geeignet?*
- *Welche KI-Akteure sind für die Entscheidungen der KI verantwortlich und ist sich diese Person der beabsichtigten Verwendungszwecke und Grenzen der Analyse bewusst?*

Map 1.2

Ziel: Interdisziplinäre KI-Akteure, Kompetenzen, Fähigkeiten und Kapazitäten für die Erstellung von Kontexten spiegeln die demografische Vielfalt und ein breites Spektrum an Fachwissen und Nutzererfahrungen wider, und ihre Beteiligung wird dokumentiert. Gelegenheiten zur interdisziplinären Zusammenarbeit werden vorrangig behandelt. Fragen, die gestellt werden, sind u. a.:

- *Inwieweit spiegeln die für die Entwicklung und Wartung des KI-Systems zuständigen Teams unterschiedliche Meinungen, Hintergründe, Erfahrungen und Perspektiven wider?*
- *Inwieweit hat sich das Unternehmen mit den Ansichten der Beteiligten zu den möglichen negativen Auswirkungen des KI-Systems auf die Endnutzer und die betroffenen Bevölkerungsgruppen auseinandergesetzt?*

Map 1.3

Ziel: Die Absicht der Organisation und die entsprechenden Ziele für die KI-Technologie sind bekannt und dokumentiert. Fragen, die gestellt werden, sind u. a.:

- *Wie kann das KI-System dem Unternehmen helfen, seine Ziele zu erreichen?*
- *Wie stimmen die technischen Spezifikationen und Anforderungen mit den Zielen des KI-Systems überein?*

Map 1.4

Ziel: Der Geschäftswert oder der Kontext der geschäftlichen Nutzung wurde klar definiert oder – im Falle der Bewertung bestehender KI Systeme – neu bewertet. Fragen, die gestellt werden, sind u. a.:

- *Welche Ziele will die Einrichtung durch die Konzeption, Entwicklung und/oder den Einsatz des KI-Systems erreichen?*
- *Inwieweit stehen die Ergebnisse des Systems im Einklang mit den Werten und Grundsätzen der Einrichtung zur Förderung des öffentlichen Vertrauens und der Gerechtigkeit?*

Map 1.5

Ziel: Organisatorische Risikotoleranzen werden festgelegt und dokumentiert. Fragen, die gestellt werden, sind u. a.:

- *Welche bestehenden Vorschriften und Leitlinien gelten und wurden von der Einrichtung bei der Entwicklung von Systemrisikotoleranzen befolgt?*
- *Welche Kriterien und Annahmen hat die Einrichtung bei der Entwicklung von Systemrisikotoleranzen zugrunde gelegt?*

Map 1.6

Ziel: Systemanforderungen (z. B. "das System muss die Privatsphäre seiner Nutzer respektieren") werden von den relevanten KI-Akteuren erfragt und verstanden. Bei Designentscheidungen werden soziotechnische Auswirkungen berücksichtigt, um KI-Risiken anzugehen. Fragen, die gestellt werden, sind u. a.:

- *Welche Messgrößen hat das Unternehmen entwickelt, um die Leistung des KI-Systems zu messen?*
- *Welche Begründungen hat das Unternehmen ggf. für die Annahmen, Grenzen und Beschränkungen des KI-Systems geliefert?*

Quelle: [1] https://airc.nist.gov/AI_RMF_Knowledge_Base/Playbook

b) Analyse: Werte & Prinzipien 3

Executive Summary

Für die Betrachtung möglicher Werte & Prinzipien als Bewertungsmaßstab eines digital-ethischen Risikoassessments für Data Sharing Ökosysteme wurden zwei Analyseschwerpunkte verfolgt.

Zum einen wurde der Frage nachgegangen, **welche Werte & Prinzipien** im Kontext von Data Sharing im allgemeinen von Bedeutung sein könnten. Dafür wurden die Ergebnisse vergleichender Studien zu digital-ethischen Leitlinien in Europa herangezogen (siehe Leitlinienanalyse). Diese Leitlinien von unterschiedlichen Akteuren aus Wirtschaft, Politik und Gesellschaft haben entweder einen Schwerpunkt in algorithmischen Systemen oder in Daten. Hier offenbart die Studie eine Vielzahl an Werten & Prinzipien, die sich je nach Schwerpunkt unterscheiden. Dabei wird deutlich, dass in Data Sharing Ökosystemen möglicherweise andere Werte & Prinzipien zu berücksichtigen sind, als dies bei KI-Systemen der Fall ist. „Typische" Werte & Prinzipien mit reinem KI-Bezug wie Erklärbarkeit sollten z. B. zugunsten datenbezogener Werte & Prinzipien wie Rückverfolgbarkeit in den Hintergrund rücken. Eine Weitere Erkenntnis ist, dass es eine geteilte Wertebasis in Europa gibt, deren Berücksichtigung die Anschlussfähigkeit eines darauf basierenden digital-ethischen Risikoassessments sicherstellt.

Zum anderen wurde nach einem Prozess gesucht, der dabei unterstützt, **wie die Auswahl** der Werten & Prinzipien erfolgen kann. Dieser wurde u. a. im Standard IEEE 7000™ gefunden. Eine zentrale Erkenntnis des dort beschriebenen Prozesses zur Ermittlung und Priorisierung ethischer Werte ist die Involvierung der möglichen Stakeholder des betreffenden Systems. Im iterativen Austausch sollen Werte gesammelt, geclustert und priorisiert werden. Neben den Stakeholdern, die Einfluss auf die Auswahl der Werte haben, werden auch potenzielle technische und organisatorische Risiken und Chancen betrachtet, die Auswirkungen auf diese Werte haben. So wird sichergestellt, dass das jeweilige KI-System auf denjenigen Werten aufgebaut wird, die dem Verwendungskontext angemessen sind (siehe IEEE).

Zentrale Erkenntnisse

Berücksichtigung kultureller Gegebenheiten

Abwägen der dem Assessment zugrundeliegenden Werte & Prinzipien und Anpassung auf den Betrachtungsgegenstand

Involvierung relevanter Stakeholder

Leitlinienanalysen

Wie die Analyse unterschiedlicher Labels, Assessments und Standards zeigt, werden häufig Werte & Prinzipien herangezogen, die als Bewertungsmaßstab genutzt werden. Eine große Relevanz von Werten & Prinzipien ist auch in Leitlinien zum digital-ethischen Umgang mit Daten und Algorithmen zu beobachten. Unterschiedliche Studien (Field et al. 2020[1]; Hagendorff 2020[2]; idigiT 2022[3]; Jobin & Vayena 2019[4]) zu diesem Thema zeigen, dass Leitlinien nach Begriffen wie Transparenz, Gerechtigkeit bzw. Fairness oder Verantwortung strukturiert werden, unter denen bestimmte Anforderungen an den Umgang mit Daten und Algorithmen gefasst werden. Im Folgenden sollen exemplarisch einzelne Ergebnisse der idigiT Studie „Zwischen Unternehmenswerten und Operationalisierung“[3] vorgestellt werden.

idigiT Studie: Leitlinien zu digitaler Ethik in Europa

Gegenstand

Leitlinien zu digitaler Ethik in Europa

Die idigiT Studie untersucht **Leitlinien aus Europa**, die den verantwortungsvollen Umgang mit digitalen Anwendungen thematisieren.

Methodik

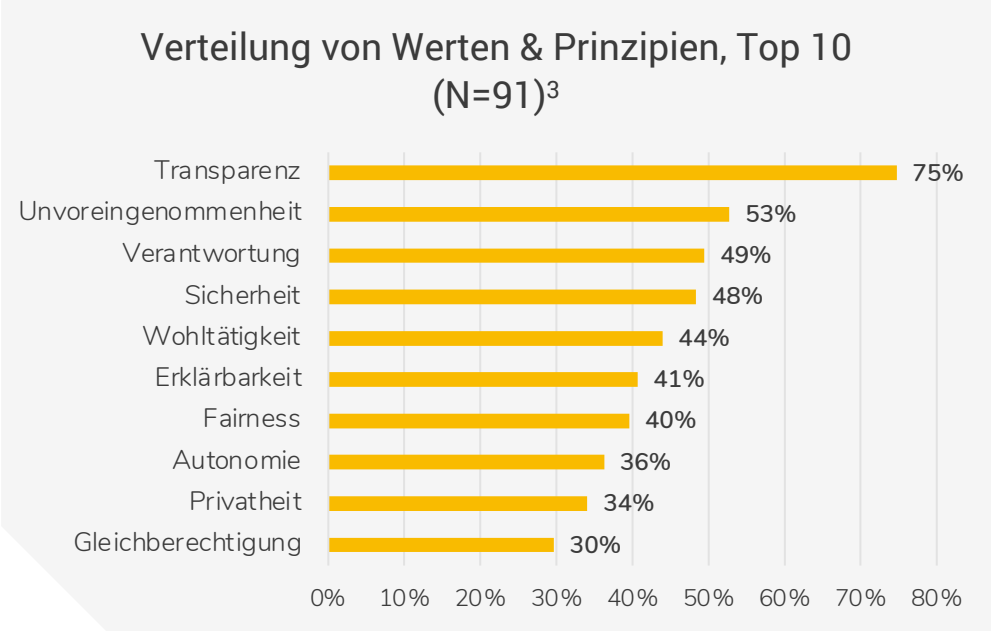
Qualitative Analyse (Grundgesamtheit N=91)

In einer **qualitativen Analyse** wurden 91 Leitlinien in Form von öffentlich zugänglichen Dokumenten empirisch nach Werten und Themen ausgewertet.

Zeitraum

2014 - 2021

Die Analyse berücksichtigt Leitlinien, die zwischen den **Jahren 2014 bis 2021** in Europa veröffentlicht wurden.



In den durch idigiT untersuchten Veröffentlichungen zeigt sich eine Vielzahl an **Werten**. Die 10 häufigsten Werte sind als „Top 10“ abgebildet. **Transparenz** kommt am häufigsten vor. Darauf folgen Werte wie **Unvoreingenommenheit** und **Verantwortung**. Diese Werte werden immer wieder in den Veröffentlichungen genannt. Sie lassen sich keiner besonderen **Herausgebergruppe** zuordnen, sondern werden universal aufgegriffen.

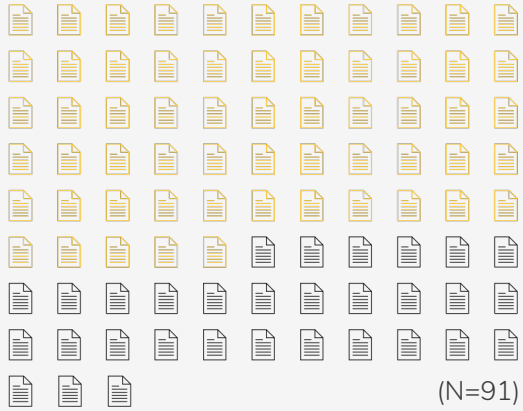
Der Wert **Transparenz** bezieht sich zumeist darauf, wie viele und welche Informationen im Zusammenhang mit Daten und KI geteilt werden. Bei **Unvoreingenommenheit** geht es vor allem darum, Diskriminierungen durch „verzerrte“ Datenlagen und fehlgeleitete statistische Annahmen, den sogenannten „Bias“ in Daten, zu verhindern. Unter dem Wert **Verantwortung** wird verstanden, dass Unternehmen und ihre Mitarbeiter:innen Verantwortung bei der Datennutzung und Entwicklung algorithmischer Systeme übernehmen sollen.

Quellen: [1]http://dx.doi.org/10.2139/ssrn.3518482 | [2]https://doi.org/10.1007/s11023-020-09517-8 | [3]https://www.transforming-healthcare.com/insights/ | [4]https://doi.org/10.1038/s42256-019-0088-2

idigiT Leitlinienanalyse: KI als Treiber

Die idigiT Leitlinienanalyse ergab ebenfalls eine Unterscheidung in zwei Arten von digital-ethischen Leitlinien. Während 66% der Leitlinien auf das Thema KI, vor allem in Form von algorithmischen Systemen, ausgerichtet sind, gibt es einen Teil von 34% der allgemeiner formuliert ist und den Umgang mit Daten fokussiert. Diese Einteilung hat auch Implikationen auf die zu beobachtenden Werte & Prinzipien, da letztere eher dem einen oder dem anderen Thema zugeordnet werden. **Der Wert Transparenz wird häufiger in Verbindung mit einem KI-Schwerpunkt genannt.** Ein Grund hierfür ist, dass, wenn es um die Anwendung von KI geht, oft von einer Black Box gesprochen wird. Darunter wird verstanden, dass bei einigen Formen von KI nicht nachzuvollziehen ist, wie ein Endergebnis zustande kommt, z. B. bei selbst-lernenden Systemen. Insbesondere dann wird Transparenz gefordert. Transparenz ist auch im Umgang mit Daten ein wichtiger Wert. Jedoch ist es grundsätzlich einfacher, die Nutzungsart von Daten, etwa in einem Datenraum, transparent zu gestalten.

Themen der Leitlinien: KI als Treiber[1]

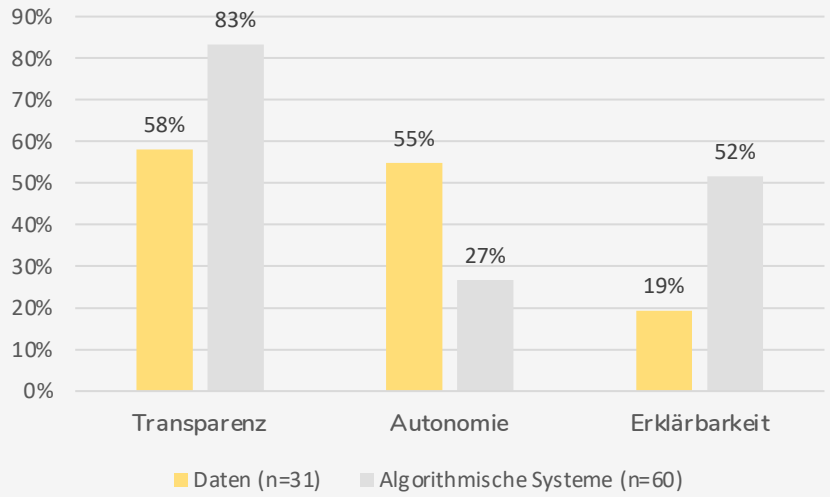


(N=91)

66 %
der Leitlinien konzentrieren sich auf algorithmische Systeme

34 %
der Leitlinien konzentrieren sich auf den Umgang mit Daten

Werteverteilung der Leitlinien: Fokusthemen je Werte[1]



Der Wert Autonomie kommt häufiger bei Leitlinien mit einem Datenbezug vor. Dies ist nachvollziehbar, da oft betont wird, dass Personen ihre Daten selbstbestimmt teilen können sollten. Dennoch ist Autonomie auch im Umgang mit KI ein wichtiges Thema. Dieses wird allerdings spezifischer über den Wert Erklärbarkeit zum Ausdruck gebracht. Erklärbarkeit kommt zu 52 % in Leitlinien mit einem KI-Bezug – durch die Forderung nach Erklärbarkeit von algorithmischen Systemen – und zu 19 % mit einem Datenbezug vor.

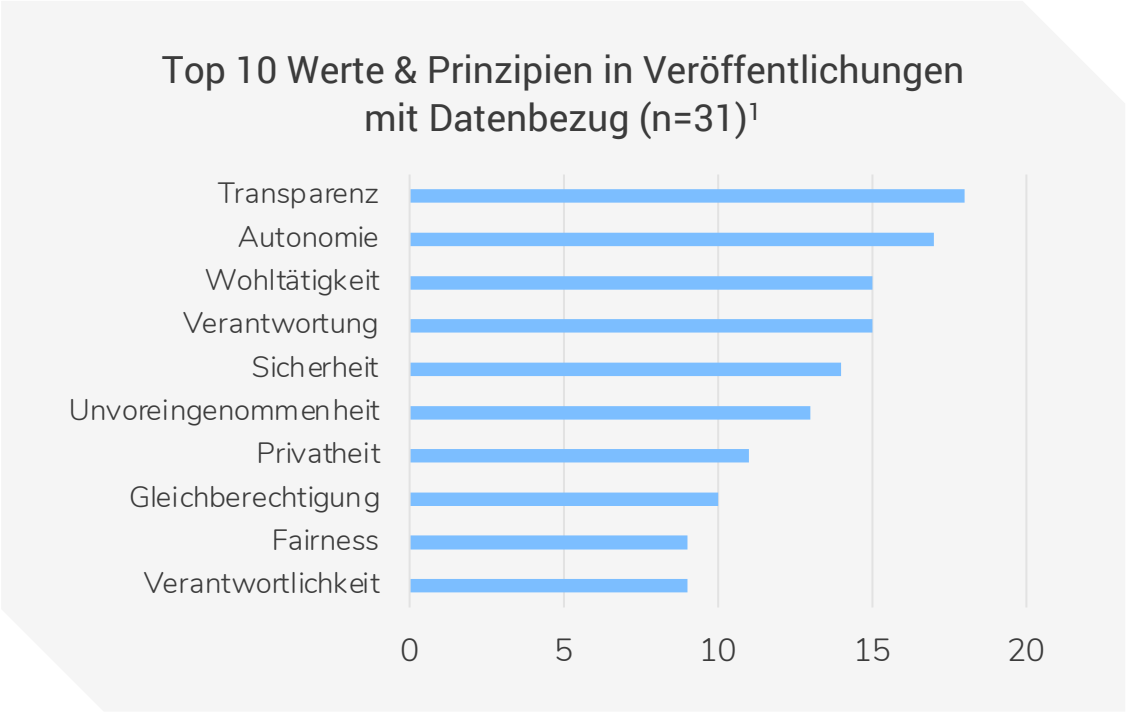
Diese Erkenntnisse über die Zuordenbarkeit von Werten & Prinzipien zu einzelnen Schwerpunkten spielt ebenfalls eine Rolle bei der Konzeption eines digital-ethischen Risikoassessments für Data Sharing Ökosysteme. Mit dem Fokus auf Daten verändert sich die Auswahl relevanter Werte & Prinzipien für die digital-ethische Bewertung. Dies ist zu berücksichtigen, um eine aussagekräftige Risikobewertung zu erhalten. Die in der idigiT Leitlinienanalyse gefundenen Werte & Prinzipien in Veröffentlichungen mit Datenschwerpunkt könnten daher als Basis für die weitere Entwicklung herangezogen werden. Diese werden im Folgenden vorgestellt.

Quelle: [1]https://www.transforming-healthcare.com/insights/

idigiT Leitlinienanalyse: Werte & Prinzipien mit Datenbezug

Eine detaillierte Betrachtung der Top 10 Werte & Prinzipien, die sich in Veröffentlichungen mit Datenbezug zeigen, ergibt im Vergleich zu der Top 10 mit Bezug auf Veröffentlichungen zu Daten & Algorithmischen Systemen ein leicht verändertes Bild. Wie bereits erwähnt, rückt etwa Autonomie hinter Transparenz auf den zweiten Platz. Ungeachtet der Verteilung stellt sich in Bezug auf die Entwicklung eines digital-ethischen Risikoassessments für Data Sharing Ökosysteme die Frage, was die einzelnen Werte & Prinzipien bedeuten können:

- **Transparenz:** Es ist jederzeit nachvollziehbar, wie und vom wem die Daten genutzt werden.
- **Autonomie:** Personen können selbstbestimmt entscheiden, was mit ihren Daten geschieht.
- **Wohltätigkeit:** Daten sollen zum Wohl der Allgemeinheit verwendet werden.
- **Verantwortung:** Beteiligte an Data Sharing Prozessen übernehmen Verantwortung für ihr Handeln.
- **Sicherheit:** Die Daten sind vor unberechtigtem Zugriff geschützt.
- **Unvoreingenommenheit:** Es wird sichergestellt, dass Daten möglichst wenig Verzerrungen aufweisen.
- **Privatheit:** Die Privatsphäre von Personen wird respektiert und personenbezogene Daten bspw. als besonders Schützenswert behandelt.
- **Gleichberechtigung:** Ergebnisse des Data Sharings kommen können prinzipiell allen gleichermaßen zu gute.
- **Fairness:** Der Zugang zu Data Sharing Ökosystemen steht allen (natürlichen und juristischen Personen) offen.
- **Verantwortlichkeit:** Es ist eindeutig, wer im Fall von Zwischenfällen die Verantwortung übernimmt (es gibt entsprechende Strukturen).



Nicht alle dieser gefundenen Werte & Prinzipien lassen sich auf den ersten Blick klar von einander trennen. Zudem sind sie nicht alle gleichermaßen sinnvoll auf Data Sharing Ökosysteme bzw. Datenräume übertragbar. Für die Anwendung als Maßstab für das digital-ethische Risikoassessment müssen daher zunächst einzelne Werte & Prinzipien zusammengefasst oder ganz verworfen werden. Ein möglicher Prozess, um die relevanten Werte & Prinzipien zu bestimmen, ist im bereits vorgestellten technischen Standard IEEE 7000™ enthalten. Die dort beschriebene Ermittlung und Priorisierung wird im Folgenden vorgestellt.

Quelle: [1]https://www.transforming-healthcare.com/insights/

IEEE 7000™: Ethische Werte ermitteln und priorisieren

Der in der Analyse bzgl. des Kontextes bereits vorgestellte Standard IEEE 7000™ enthält ebenfalls einen Prozessschritt zur Ermittlung und Priorisierung ethischer Werte. Der Zweck des Prozesses besteht darin, Werte und Prinzipien nicht nur zu ermitteln, sondern auch in eine Rangfolge zu bringen. Zur Auswahl der Werte wird auf die in der Kontextanalyse identifizierten Stakeholder zurückgegriffen. Dazu werden in einem ersten Schritt diejenigen Stakeholder ausgewählt, deren Werte eruiert werden sollen. Der IEEE 7000™ sieht vor, die ethischen Fragen, Werte und Potenziale der Stakeholder über Utilitarismus, Tugendethik und Pflichtethik zu ermitteln.

Die 7 Hauptschritte zur Ermittlung und Priorisierung ethischer Werte[1]

1. **Auswahl der Stakeholder in der SOI, deren Werte eruiert werden sollen.**
2. **Ermitteln und Erfassen der für die ConOps relevanten Werte der Stakeholder.**
3. **Analyse und Einordnung der ermittelten Werte.**
4. **Priorisierung der Grundwerte für die SOI.**
5. **Identifizierung und Erfassung potenzieller technischer und organisatorischer Risiken und Chancen mit Auswirkungen auf die Werte.**
6. **Durchführung einer konzeptionellen Wertanalyse und Verfeinerung der priorisierten Wertcluster.**
7. **Einholung der Genehmigung für die priorisierten Werte.**

Ergebnisse der erfolgreichen Implementierung[1]

- ✓ Es werden die Werte der Stakeholder, ethische Fragen sowie potenzielle Schäden und Vorteile in Bezug auf die SOI eruiert.
- ✓ Mithilfe einer konzeptionellen Analyse werden die Werte und Wertdemonstratoren verfeinert und in Wertclustern organisiert.
- ✓ Die Wertecluster werden nach Prioritäten geordnet.
- ✓ Die Zustimmung des Managements zu den priorisierten Werten wird eingeholt.
- ✓ Die Aktivitäten und Aufgaben dieses Prozesses werden mit den anderen Aufgaben zur Entwicklung der SOI integriert.

Aus den gesammelten Werten, Problemen und Potenzialen werden Kernwerte identifiziert, die dann in Form von Werteclustern beschrieben werden, einschließlich der ethischen Probleme, Werte und Potenziale, die angesprochen werden. Die Wertecluster werden von den Stakeholdern validiert. Die Kernwerte werden priorisiert und mit Wertprioritäten verglichen, die von maßgeblichen externen Quellen vorgeschlagen werden. Bei Unvereinbarkeiten zwischen den Wertprioritäten werden die Prioritäten und Wertcluster angepasst. Die sich daraus ergebenden Wertecluster können durch den Werteverantwortlichen konzeptionell präzisiert werden. Die Wertcluster werden von ausgewählten Stakeholdern und dem Management freigegeben.

Dieser Prozess der Werteermittlung ist entsprechend iterativ aufgebaut und erfordert den Austausch mit relevanten Stakeholdern. Als Grundlage wird empfohlen, auf Literatur zur angewandten Ethik mit konzeptionellen Rahmen für die Taxonomie der einzelnen Werte, Menschenrechts-Frameworks oder andere Listen mit Werten zurückzugreifen. Die identifizierten Werte der oben besprochenen idigiT Leitlinienanalyse könnten beispielsweise ein solcher Input sein.

Quelle: [1]IEEE-7000-2021 S. 39-42

c) Analyse: Anforderungen 3

Executive Summary

Die Übersetzung von Werten & Prinzipien in messbare Anforderungen ist die größte Herausforderung bei der Konzeption eines digital-ethischen Risikoassessments für Data Sharing Ökosysteme. Die analysierten Dokumente bieten in dieser Hinsicht nur wenige Anhaltspunkte. Zwar definieren die meisten Labels, Assessments und Standards Kriterien, die erfüllt sein müssen, beschreiben aber nur selten, wie diese Kriterien entstanden sind. Trotzdem ließen sich einige zentrale Erkenntnisse gewinnen.

Wie sich Werte & Prinzipien in der Praxis äußern, hängt vom **Kontext** ab. Transparenz in Bezug auf Data Sharing betont die Nachvollziehbarkeit der Datennutzung – also wie und vom wem die Daten genutzt werden. Transparenz von KI-Systemen betrifft hingegen neben der verwendeten Datenbasis eher die Art und Weise des Zustandekommens von automatisierten Entscheidungen. Genauso kann Transparenz im Kontext von KI-Systemen auch bedeuten, dass Nutzende darüber informiert werden, dass ein eben solches System überhaupt zum Einsatz kommt (z. B. bei Chatbots). Sollen also Anforderungen überprüft werden, sollten diese in Relation zum Betrachtungsgegenstand definiert werden.

Die Definition von Anforderungen reicht für eine anwendbare Bewertung noch nicht aus. Hier ist es hilfreich, die Anforderungen in Indikatoren aufzuteilen, anhand derer die Erfüllung festgemacht werden kann. Diese wiederum sollten in Observables unterteilt werden, die etwas über den Grad der Erfüllung aussagen. So entsteht ein **mehrstufiges Verfahren**, dass die Überprüfbarkeit der Anforderungserfüllung ermöglicht (siehe VCIO-Model).

Für die Erarbeitung der Anforderungen spielt erneut die Berücksichtigung von **Stakeholdern** eine große Rolle. Stakeholder verfügen in der Regel über eine Idee davon, was entsprechende Werte & Prinzipien im jeweils zu bewertenden System bedeuten. Da diese im Idealfall bereits Einfluss auf die Ermittlung und Priorisierung von Werten & Prinzipien haben, ist es naheliegend, auch die Entwicklung entsprechende Anforderungen an ihnen auszurichten (siehe IEEE).

Zentrale Erkenntnisse

- Anpassung der Anforderungen an den jeweiligen Kontext
- Mehrstufige Bewertung der Anforderungen, die sich auch Werten & Prinzipien ergeben
- Abgleich der Anforderungen mit den Stakeholdern

VCIO-Model (1/2)

Die bereits in der Analyse zu Methoden der Kontextexploration aufgegriffene Publikation „From Principles to Practice“ der AIEIGroup enthält auch ein Model zur Operationalisierung von Werten & Prinzipien: **V**alues, **C**riteria, **I**ndicators, **O**bservables (VCIO).

Zur praktischen Umsetzung von KI-Ethik arbeitet der VCIO-Ansatz mit vier Ebenen:

- **Werte** formulieren ein allgemeines ethisches Anliegen. Sie werden auf der höchsten Ebene definiert (z. B. als Gerechtigkeit oder Transparenz).
- **Kriterien** definieren die Erfüllung oder Verletzung des jeweiligen Wertes.
- Es werden **Indikatoren** benötigt, um zu beobachten, ob ein Kriterium erfüllt ist.
- **Observables** helfen, das aktuelle Qualitätslevel der Indikatoren zu beurteilen.

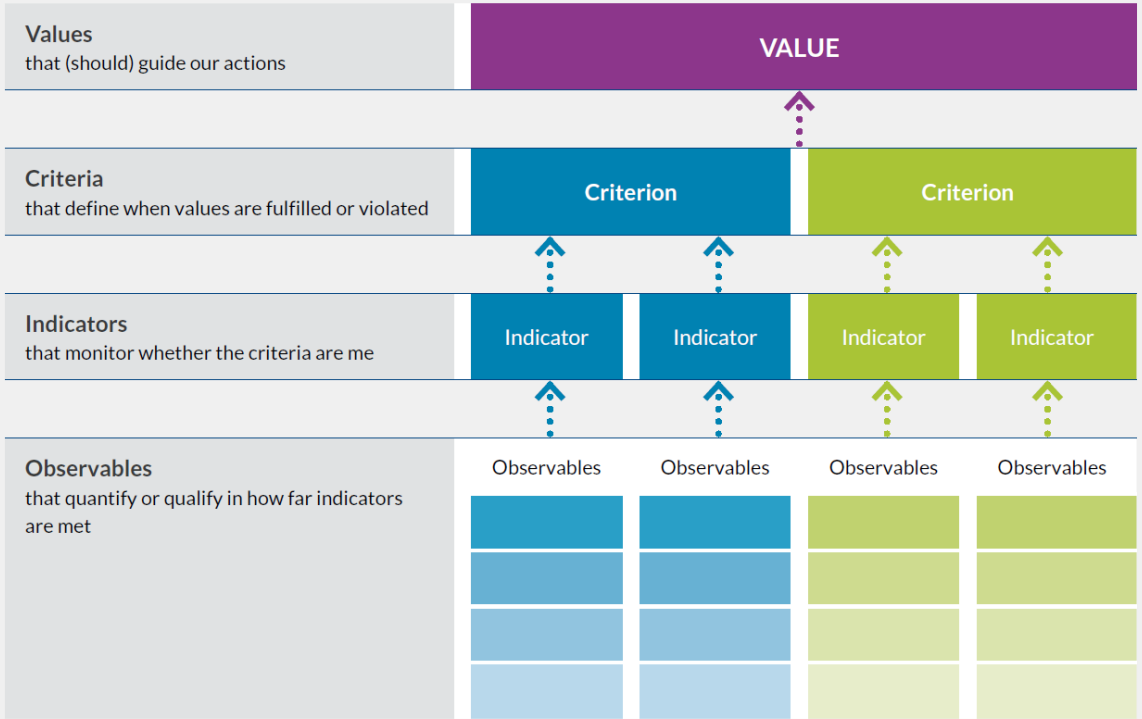
Werte & Prinzipien können so auf beobachtbare Anforderungen heruntergebrochen werden. Das Thema Werte und wie diese auszuwählen sind, wurde bereits im vorherigen Kapitel beschrieben. Es folgt daher ein Blick auf Kriterien, Indikatoren und Observables am Beispiel „Transparenz“.

Beispiel Transparenz

So lässt sich der **Wert** Transparenz, verstanden als Erklärbarkeit und Interpretierbarkeit, in die **Kriterien** „Offenlegung der Herkunft von Datensätzen“, „Offenlegung der Eigenschaften des verwendeten Algorithmus/Modells“ und „Zugänglichkeit“ unterteilen. Die „Offenlegung der Herkunft von Datensätzen“ kann dann an drei **Indikatoren** überprüft werden: „Ist die Herkunft der Daten dokumentiert?“, „Ist es für jeden Zweck plausibel, welche Daten verwendet werden?“ und „Sind die Eigenschaften des Trainingsdatensatzes dokumentiert und offengelegt? Sind die entsprechenden Datenblätter umfassend?“. **Observables** in Bezug auf den Indikator „Ist die Herkunft der Daten dokumentiert?“ sind dann entweder „Ja, umfassende Protokollierung aller Trainings- und Betriebsdaten, Versionskontrolle der Datensätze usw.“, „Ja, Protokollierung und Versionskontrolle durch einen Mittelsmann (z. B. Datenlieferant)“ oder „Keine Protokollierung; verwendete Daten werden nicht kontrolliert oder in irgendeiner Weise dokumentiert.“ Zusammen mit den Ergebnissen der anderen

Struktur des VCIO-Models[1]

FIGURE 2 **The VCIO model**



Indikatoren und Kriterien lässt sich so feststellen, ob Transparenz im jeweiligen Fall verletzt wird. Anzumerken ist, dass es nicht möglich ist, Kriterien oder Indikatoren logisch aus Werten abzuleiten, sondern für diese in Abwägungsprozessen zu argumentieren.

Quelle: [1]https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/WKIO_2020_final.pdf

VCIO-Model (2/2) - Beispiel „Transparenz“[1]

Transparenz bezieht sich auf die Offenlegung der Datenherkunft und der Eigenschaften des verwendeten KI-Modells sowie auf die Zugänglichkeit und Nachvollziehbarkeit der offengelegten Informationen. In diesem Sinne ist Transparenz sowohl beim allgemeinen Funktionsprinzip als auch bei jedem Output des KI-Systems anzustreben. Die Transparenz muss darüber hinaus auf die Bedürfnisse der Zielgruppen wie Nutzer und Betroffene zugeschnitten sein, d. h. das System muss für sie verständlich sein. Wie die Grafik zeigt, lassen sich für jedes Kriterium unterschiedlich viele Indikatoren und davon abhängige Observables bestimmen. Dabei sind die Anforderungen auf KI-Systeme zugeschnitten. Das selbe Prinzip ließe sich jedoch auch auf Data Sharing Ökosysteme wie HEALTH-X dataLOFT übertragen.

2.1.1 Applying the VCIO approach to transparency as a value

Value	TRANSPARENCY						TRANSPARENCY						Value
Criteria	Disclosure of origin of data sets			Disclosure of properties of algorithm/model used			Accessibility						Criteria
Indicators	Is the data's origin documented?	Is it plausible for each purpose, which data is being used?	Are the training data set's characteristics documented and disclosed? Are the corresponding data sheets comprehensive?	Has the model in question been tested and used before?	Is it possible to inspect the model so far that potential weaknesses can be discovered?	Taking into account efficiency and accuracy, has the simplest and most intelligible model been used?[1]	Are the modes of interpretability target-group-specific and have been developed with the target groups?	Who has access to information about data sets and the algorithm/model used?	Is the operating principle comprehensible and interpretable?	Are the modes of interpretability in their target-group-specific form intelligible for the target groups?	Are the hyperparameters (parameters of learning methods) accessible?	Has a mediating authority been established to settle and regulate transparency conflicts?	Indicators
Observables	Yes, comprehensive logging of all training and operating data, version control of data sets etc.[2]	Yes, the use of data and the individual application are intelligible	Yes and the data sheets are comprehensive	Yes, the model is widely used and tested both in theory and practice[3]	Yes, the model can easily be inspected and tested	Yes, the model has been evaluated and the most intelligible model has been used	Yes	Everyone	Yes, the model itself is directly comprehensible	Yes, the modes of interpretability have been tested with target groups for intelligibility	Yes, to everyone	Yes, a competent authority has been established	Observables
									Yes, the modes of interpretability are provided with the model itself				
	Yes, logging and version control through an intermediary (e.g. data supplier)	Yes, it is intelligible on an abstract, not case specific level, which data is being used	Yes, but (some) data sheets contain few or missing information	Yes, the model is known and tested in either theory or practice	Yes, but the model can only be tested by certain people due to non-disclosure	No, but the model wa evaluated regarding interpretability and this evaluation is disclosed to the publi	Yes, but without participation of the target groups	All people directly affected	No, the modes of interpretability can only be used post hoc by experts	Yes, target groups can complain or ask if they do not understand a mode of interpretability	Yes, but only to information and trust intermediaries (regulators, watchdogs, researchers, courts)	Yes, a competent authority has been established but its powers are limited	
		No, but a summary on data usage is available		Yes, the model is known to some experts but has not been tested yet			Yes, but the modes or interpretability are only specific for one target group	Only information and trust intermediaries (regulators, watchdogs, research, courts)	No, the modes of interpretability need to be adjusted to the individual model and use by experts				
	No logging; data used is not controlled or documented in any way	No	No	No, the model has been developed recently	No	No, the model has no been evaluated	No, the modes of interpretability[4] are not target-group-specific	Nobody	No, but the model is theoretically comprehensible	No	No	No	
									No, there are no known modes of interpretability				

[1] This indicator would require further specification regarding the balance between using an efficient and accurate model and using a model which is technically simple and thus naturally easier to comprehend and follow.
[2] This observable could include further levels of logging and documentation of data sets.
[3] This observable could help to determine the levels needed in other observables: If the model has been widely used and tested, it might not require additional testing.
[4] "Modes of interpretability" refers to different methods to ensure or increase interpretability (use of simple model, explanations of data and model used, etc.).

AIEI Group

Quelle: [1]https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/WKIO_2020_final.pdf

IEEE 7000: Definition ethischer Anforderungen

Der bereits in den bisherigen Analyse berücksichtigte Standard IEEE 7000™ enthält ebenfalls einen Prozessschritt zur Definition ethischer Anforderungen. Der Zweck dieses Prozesses ist die Formulierung von sog. Ethical Value Requirements (EVRs) und von wertbasierten Systemanforderungen, die definieren, wie die priorisierten Grundwerte im SOI reflektiert werden. Er gibt keine Werte und Anforderungen vor, zeigt aber die Möglichkeit für eine strukturierte Erarbeitung auf.

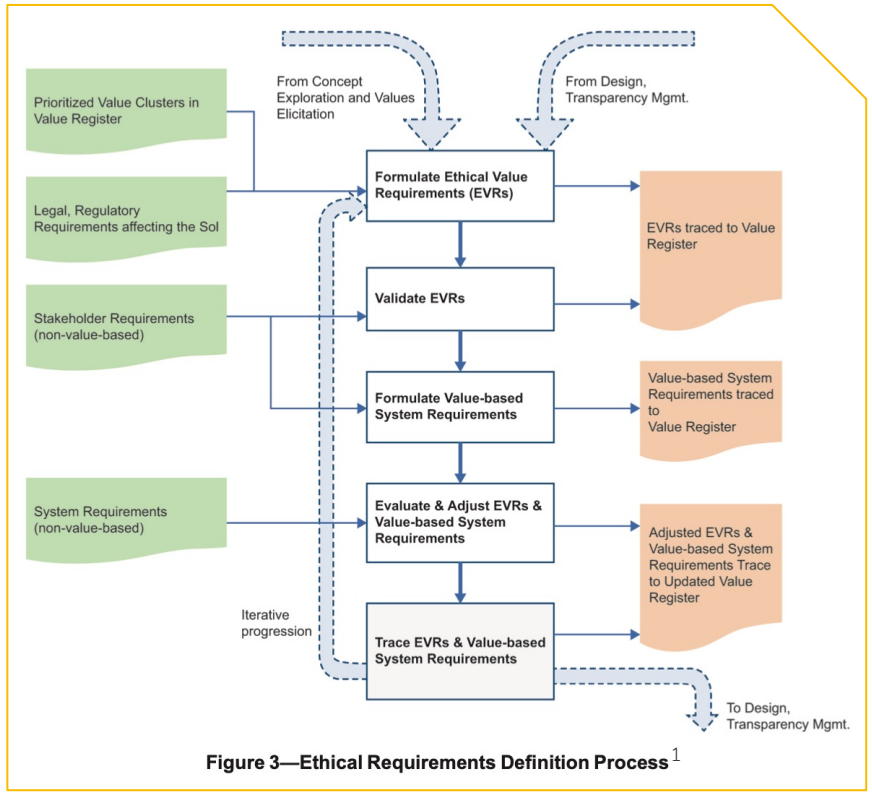


Figure 3—Ethical Requirements Definition Process[1]

Die 5 Hauptschritte zur Definition ethischer Anforderungen[1]

- Ethical Value Requirements (EVRs) formulieren und aufzeichnen.
- Validierung der EVRs zusammen mit anderen Anforderungen der Stakeholder.
- Formulierung und Aufzeichnung der Systemanforderungen, die sich aus jedem EVR ergeben.
- Evaluierung und Anpassung der EVR und der wertorientierten Systemanforderungen.
- Analyse, Verfolgung und Aufzeichnung der weiteren Bearbeitung von wertbezogenen Anforderungen.

Ethische Anforderungen sind Vorschläge zur Risikominderung, um die Grundwerte innerhalb des SOI zu schützen und zu erhalten. Im Rahmen des Prozesses werden die EVRs und die wertebasierten Systemanforderungen auf ethikbezogene Risiken hin analysiert und Abhilfemaßnahmen in Form von Überarbeitungen des Anforderungssatzes identifiziert. Bei diesem Prozess werden die für das SOI verantwortlichen Personen einbezogen und ihr Engagement für wertorientierte Anforderungen durch Validierung dokumentiert. Auch in diesem IEEE 7000™ Prozess spielt der Austausch mit den relevanten Stakeholdern also eine große Rolle.

Quelle: [1]IEEE-7000-2021 S. 43-46

4 DERA für Data Sharing Ökosysteme

Konzeptionelle Merkmale (1/3)

Die ethischen Implikationen des digitalen Fortschritts und ihr Einfluss auf die Gestaltung nachhaltiger Geschäftsmodelle zeigen sich aktuell immer deutlicher. KI-Systeme wie ChatGPT werfen bspw. Fragen nach Verantwortung, Fairness und Diskriminierungsfreiheit auf. Die Identifikation von digital-ethische Risiken digitaler Innovationen ist daher heute unerlässlich für ein menschenzentriertes und gesellschaftsorientiertes Design.

Vor allem im Bereich von KI-Systemen wird versucht, dies mit Hilfe von Labeln, Assessments und Standards sicherzustellen. Wie der Vergleich und die Analyse entsprechender Dokumente zeigen, gibt es dafür bisher viele verschiedene Ansätze, aber noch keinen allgemein etablierten Standard, auf den zurückgegriffen werden könnte.

Im Hinblick auf Daten, die als Grundlage dieser Entwicklung eine tragende Rolle spielen, sind spezifische Frameworks ebenfalls selten. Data Sharing Ökosysteme wie „HEALTH-X dataLOFT“ stehen daher vor der Herausforderung, Wege zur Identifikation von digital-ethischen Risiken zu finden, um diese bereits im Entwicklungsprozess zu minimieren. Dies gilt insbesondere für den Gesundheitssektor, für den so gut wie keine spezifischen digital-ethischen Frameworks zu finden sind – weder für KI-Systeme noch für Data Sharing Ökosysteme.

Um diese Lücke zu schließen braucht es daher einen eigenen Ansatz für ein digital-ethisches Risikoassessment für Data Sharing Ökosysteme. Die Analyse bisheriger Labels, Assessments und Standards (N=65) brachte mehrere Komponenten zu Tage, die Bestandteile eines solchen Assessments sein könnten.

Ein anfänglicher Vergleich der entsprechenden Dokumente ergab, dass drei Elemente erforderlich sind, wenn ein digital-ethisches Risikoassessment aufgebaut werden soll:

- **Die Bestimmung des Kontextes**
- **Der Bezug auf Werte & Prinzipien als Maßstab der Evaluation**
- **Die Übersetzung von Werten & Prinzipien in überprüfbare Anforderungen**

Eine vertiefende Analyse brachte zentrale Erkenntnisse über die Art und Weise zutage, wie das jeweilige Element umgesetzt werden kann. Für ein digital-ethisches Risikoassessment für Data Sharing Ökosysteme wie „HEALTH-X dataLOFT“ sind folgende Überlegungen zentral:

Schritt 1: Kontext (1/2)

Bei der Entwicklung eines eigenen Ansatzes, der die spezifischen Bedingungen von Data Sharing Ökosystemen berücksichtigt, ist im Rahmen des Kontextes die Frage nach dem **Bezugsgegenstand** zu stellen. Was soll bewertet werden? Die technologische Basis des Data Sharing Ökosystems (d. h. der Datenraum) oder die Use Cases, die innerhalb des Data Sharing Ökosystems realisiert werden? Die Antwort liegt in der Mitte, da letztlich beides miteinander verbunden ist: Es ist nicht möglich, den Kontext des Data Sharing Ökosystems ohne den jeweiligen Use Case zu bewerten. Ein Data Sharing Ökosystem für (potenziell) personenbezogene Daten muss aufgrund der möglichen Implikationen für Gesellschaft und Individuum anders bewertet werden als bspw. ein System zum Teilen von industriellen Daten (bspw. über die Auslastung von Maschinen). Da der Use Case im Fall von „HEALTH-X dataLOFT“ im Gesundheitssektor angesiedelt ist, hat dies etwa Auswirkungen auf die betroffenen Stakeholder (Patient:innen, Ärzt:innen etc.), die Art der Daten (besonders schützenswerte Gesundheitsdaten) und damit auch auf die mögliche Auswahl und Priorisierung von Werten & Prinzipien (bspw. Autonomie für einen souveränen Umgang mit den eigenen Gesundheitsdaten). Trotzdem sind die Anforderungen, die zu formulieren sind, an die technologische Grundlage des Data Sharing Ökosystems – also den Datenraum – zu stellen. Bewertet wird also nicht der Datenraum *oder* der Use Case, sondern der Datenraum *im Kontext* des Use Cases. Um diesen Kontext systematisch zu erfassen, empfiehlt es sich auf zwei Aspekte zurückzugreifen, die sich in der Analyse zeigten: Klassifizierung und Stakeholder.

Konzeptionelle Merkmale (2/3)

Schritt 1: Kontext (2/2)

Aufgrund des beschriebenen Verhältnisses von technologischer Basis und Use Case ist eine **anfängliche Klassifizierung** sinnvoll, wie sie in den Dokumenten der AIElGroup und SAP vorgenommen wird. Diese Klassifizierung könnte der fünfstufigen Risikomatrix der AIElGroup folgen, aber auch ein drei- oder vierstufiges Klassifizierungssystem, wie es bspw. bei SAP oder im kommenden AI Act[1] der Europäischen Union angelegt ist, ist denkbar. Ziel ist es in jedem Fall, über eine anfängliche grobe Klassifizierung den Bedarf für eine detaillierte digital-ethische Risikoklassifizierung zu ermitteln.

Es ist zu erwarten, dass es immer auch Data Sharing Ökosysteme geben wird, die aufgrund ihrer anfänglichen Einstufung per se nur ein minimales digital-ethisches Risiko bergen und keinen weiteren Bewertungsprozess durchlaufen müssen. Wie aus der Analyse von Labeln, Assessments und Standards hervorgeht, könnten hier folgende Kriterien zur Einstufung herangezogen werden: die Art der Daten, der Sektor inkl. der **involvierten Stakeholder** und das zu erwartenden Ausmaß eines potenziellen Schadens. Andere oder weitere Kriterien sind ebenfalls denkbar und Gegenstand zukünftiger Entwicklungsschritte.

Mit der Ermittlung der **involvierten Stakeholder** im Rahmen der Kontextbestimmung wird auch der Grundstein für weitere Elemente des digital-ethischen Risikoassessments gelegt. Da digitale Ethik den Mensch in den Mittelpunkt technologischer Innovationen stellt, ist der Input von Stakeholdern ein nicht zu vernachlässigender Aspekt. Im Falle von HEALTH-X dataLOFT wären hier etwa vor allem Bürger:innen/Patient:innen, medizinisches Fachpersonal etc. zu nennen.

Anwendungszeitpunkt des DERA

Die untersuchten Assessments und Standards nehmen in der Regel Bezug auf die Entwicklung von KI-Systemen, d. h. die jeweils festgelegten Anforderungen sollen bereits im Entwicklungsprozess greifen. Ein digital-ethisches Risikoassessment für Data Sharing Ökosysteme ist dementsprechend am wirkungsvollsten, wenn es zu verschiedenen Zeitpunkten (mindestens vor Beginn und kurz vor Abschluss) im Entwicklungsprozess durchgeführt wird. Durch eine regelmäßige Reevaluation wird sichergestellt, dass Veränderungen an Aufbau und Zielsetzung berücksichtigt und zuvor identifizierte Risiken im Verlauf der Entwicklung minimiert werden. Dies sieht keine Wiederholung der Kontextermittlung vor, da der Kontext sich in der Regel nicht verändert haben dürfte.

Schritt 2: Werte & Prinzipien

Wie bereits in „Analyse b)“ gezeigt, hat der Fokus auf Daten einen Einfluss auf die Auswahl und Priorisierung von Werten & Prinzipien, die als Bewertungsmaßstab für ein Data Sharing Ökosystem dienen können. Mit Bezug auf **europäische Wertvorstellung** konnten erste Werte wie Transparenz („Es ist jederzeit nachvollziehbar, wie und vom wem die Daten genutzt werden“), Autonomie („Personen können selbstbestimmt entscheiden, was mit ihren Daten geschieht“) und Wohltätigkeit („Daten sollen zum Wohl der Allgemeinheit verwendet werden“) identifiziert werden, die gerade im Gesundheitskontext von HEALTH-X dataLOFT besonders anschlussfähig sind. Mit Hilfe von potenzielle **betroffenen Stakeholdern**, wie sie auch in der Kontextbestimmung identifiziert wurden, sind diese und weitere Werte & Prinzipien festzulegen und zu priorisieren. Dabei ist auch zu definieren, wie diese Werte & Prinzipien verstanden werden, um im anschließenden Schritt entsprechende Anforderungen erstellen zu können.

Quelle: [1]https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

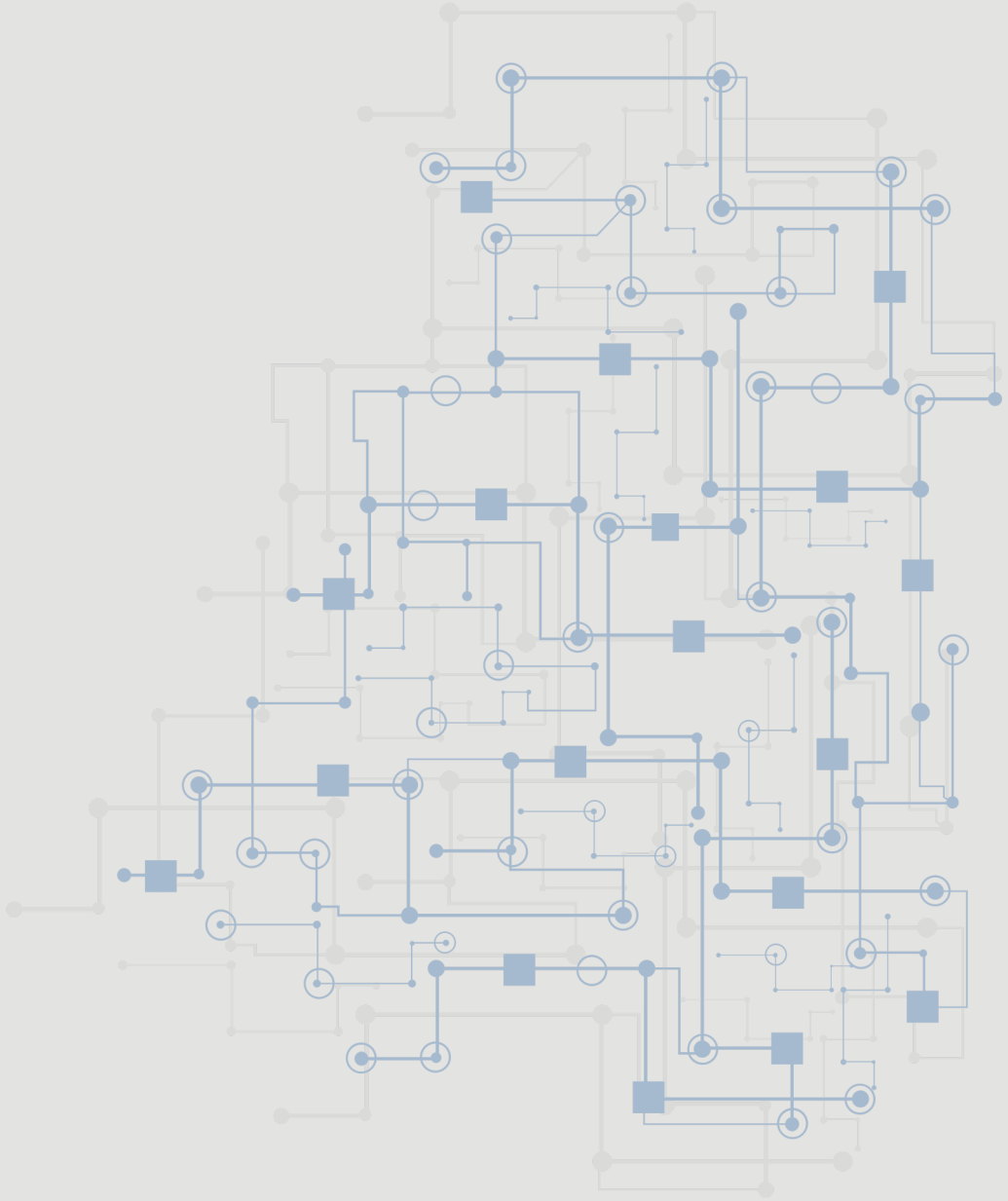
Konzeptionelle Merkmale (3/3)

Schritt 3: Anforderungen

Ebenso wie die Werte & Prinzipien eine gewisse Kontextabhängigkeit besitzen, sind auch die Anforderungen an den Datenaustausch im Kontext des Gesundheitssektors zu überdenken. Wie sehen hier mögliche Kriterien (Erfüllung oder Verletzung des jeweiligen Wertes), Indikatoren (Beobachtung, ob ein Kriterium erfüllt ist) und Observables (Beurteilung des Qualitätslevels der Indikatoren) aus? Bei der Beantwortung dieser Frage können sich Konflikte zwischen verschiedenen Werten und deren Anforderungen zeigen, etwa bei Sicherheit und Transparenz. Wieviel Transparenz kann umgesetzt werden, ohne die Sicherheit des Systems zu gefährden? Für solche und ähnliche Abwägungen sind die betroffenen Stakeholder mit einzubeziehen, wie es etwa im IEEE 7000™ beschrieben ist.

Fazit & Ausblick

Ein digital-ethisches Risikoassessment (DERA) für Data Sharing Ökosysteme erlaubt die strukturierte Identifikation von ethischen Risiken bereits während des Entwicklungsprozesses. Neben den Kernkomponenten, die hier – abgeleitet aus der Literatur zu bestehenden Labeln, Assessments und Standards – beschrieben wurden, ist abschließend noch die Frage nach einer Bewertungsskala zu stellen. So wäre ein vergleichbarer „Riskscore“ hilfreich, da er etwa Veränderungen im Laufe von Entwicklungsprozessen sichtbar machen könnte. Dieser Score ließe sich mit Hilfe der Observables, die das Qualitätsniveau der Indikatoren abbilden, über einfache Punkteverteilungen errechnen. Je höher ein solches Riskscore wäre (d. h. je mehr Observables „schlecht“ ausfallen), desto größer würde auch das digital-ethische Risiko ausfallen. Zur besseren Visualisierung könnte die Skala des Riskscores zudem in die Bereiche „gering“, „mittel“ und hoch geteilt werden. Der Riskscore wäre Teil einer zukünftigen Ausarbeitung des hier skizzierten DERA für Data Sharing Ökosysteme .



Universität Witten/Herdecke

Institute for
Digital Transformation
in Healthcare

Impressum

idigiT – Institute for Digital Transformation in Healthcare GmbH

Alfred-Herrhausen-Straße 45

58455 Witten

Germany

+49 2302 926 874

info@transforming-healthcare.com