AI Engineering and Usage

– Deutsche Telekom professional ethics

KI-Engineering und -Nutzung

– Professionsethik der Deutschen Telekom

LIFE IS FOR SHARING.

Content

Deutsche Telekom professional ethics

Inhaltsverzeichnis

Professionsethik der Deutschen Telekom

Introduction

As the opportunities and risks of new technologies are not always easy to foresee, guidelines are required so that we can remain confident during development and use. To us at DT, having the trust of our customers and of society is of huge importance. And we foster this trust when we handle technological developments responsibly.

This document serves as a self-imposed guideline for employees of DT. It provides best practices, methods, and tips for transferring the Digital Ethics Guidelines in AI[1] to DT's development processes. Similar regulations already exist for classic software development and engineering.[2]

What is the concrete function of these guidelines?

- They provide help in ethical questions. With clearly defined principles, we avoid risking the reputation of DT.
- They offer employees a corporate reference framework that helps them to minimize their personal liability by adhering to these guidelines.
- They support the creation of the best customer experience.
- They avoid unnecessary delays in development by creating clarity and spreading knowledge.

These guidelines are based on two documents: the aforementioned DT Digital Ethics Guidelines on AI and the Employee Code of Conduct.[3] They address three target groups: Those with main responsibility for a project (Project Owners), the Technical Development Team and the entire project team (Team).

The AI guidelines from the DT *Digital Ethics Guideline on AI* are:

1. We are responsible
2. We care
3. We put customers first
4. We are transparent
5. We are secure
6. We set the grounds
7. We keep control
8. We foster the cooperative model
9. We share and enlighten

1 Published by DT in May 2018
see https://www.telekom.com/en/company/digitalresponsibility/details/artificial-intelligence-ai-guideline-524366
2 VDI, Ethical Principles for Engineers, see https://www.vdi.de/ IEEE, Software Engineering Code of Ethics and Professional Practices
3 DT's Code of Conduct, see https://yam.telekom.de/groups/code-of-conduct

Einleitung

Da die Chancen und Risiken neuer Technologien nicht jederzeit klar abzusehen sind, braucht es Leitlinien, damit wir bei ihrer Entwicklung und ihrem Einsatz souverän bleiben. Für die Deutsche Telekom ist das Vertrauen der Kundinnen und Kunden sowie der Gesellschaft von zentraler Bedeutung. Wir fördern dieses Vertrauen, indem wir mit technologischen Entwicklungen verantwortungsvoll umgehen.

Dieses Dokument ist ein selbstverpflichtender Leitfaden für die Mitarbeiterinnen und Mitarbeiter der Deutschen Telekom. Es liefert Best Practices, Methoden und Tipps, die dabei helfen sollen, die Digital Ethics Guidelines für KI[1] auf Entwicklungsprozesse zu übertragen. Für klassische Softwareentwicklung und Engineering gibt es ähnliche Regelungen bereits.[2]

Was ist die konkrete Funktion dieses Leitfadens?

- Er leistet Hilfestellung bei ethischen Fragen. Mit klar festgelegten Grundsätzen vermeiden wir Reputationsrisiken für die Deutsche Telekom.
- Er bietet Mitarbeiterinnen und Mitarbeitern einen betrieblichen Bezugsrahmen, um persönliche Haftungsrisiken zu minimieren, wenn sie sich an diese Empfehlungen halten.
- Er unterstützt bei der Gestaltung der besten Kundenerlebnisse.
- Er vermeidet durch die Schaffung von Klarheit und Wissen unnötige Verzögerungen in der Entwicklung.

Dieser Leitfaden basiert auf zwei Dokumenten: den bereits erwähnten *Digital Ethics Guidelines on AI* der Deutschen Telekom ebenso wie auf dem Code of Conduct für Mitarbeiter*innen.[3] Dabei werden drei Personengruppen adressiert. Die Hauptverantwortlichen im Projekt (Project Owner), das technische Entwicklungsteam (Development Team) und das gesamte Projektteam (Team).

Die KI-Leitlinien aus der *Digital Ethics Guideline on AI* der Deutschen Telekom lauten:

1. Wir übernehmen Verantwortung
2. Wir gehen sorgsam mit KI um
3. Wir stellen unsere Kund*innen in den Mittelpunkt
4. Wir stehen für Transparenz
5. Wir bieten Sicherheit
6. Wir legen das Fundament
7. Wir behalten den Überblick
8. Wir leben das Kooperationsmodell
9. Wir teilen und erklären

1 Deutsche Telekom im Mai 2018,
https://www.telekom.com/en/company/digitalresponsibility/details/artificial-intelligence-ai-guideline-524366
2 VDI: Ethische Prinzipien für Ingenieure; IEEE: Software Engineering Code of Ethics and Professional Practices
3 Code of Conduct der Deutschen Telekom, https://yam.telekom.de/groups/code-of-conduct

This document further elaborates upon these guidelines. In doing so, questions are addressed such as "What results from the guidelines?" or "How can we achieve or maintain the desired state?" The recommendations for action in this document can already be taken now.

Furthermore, all develops involved with AI should adhere to the "Guideline for the Design of AI Supported Business Models, Services and Products at Deutsche Telekom"[4] as well as with the Data Privacy Regulations in which, for example, the quality-ensuring Privacy and Security Assessment process (PSA)[5] is detailed. Furthermore, if in doubt when evaluating a project/product in terms of digital ethics, the expert team "Digital Ethics" can and should be contacted for further advice and support via Digital-Ethics@Telekom.de.

4 further information in the YaM-group Privacy@AI, see https://yam.telekom.de/docs/DOC-552202 (registration is necessary); Guideline: "For the design of ai-supported business models, services and products at deutsche Telekom in compliance with data privacy regulations" Group Privacy Version 1.0 Stand 10.01.2019

5 further information on YaM, see https://yam.telekom.de/groups/psa-en

Die Leitlinien werden in diesem Dokument weiter konkretisiert. Damit werden Fragen beantwortet wie „Was bedeutet das für mich?“ oder: „Wie können wir das Zielbild erreichen und aufrechterhalten?“ Die Handlungsempfehlungen in diesem Dokument können bereits jetzt berücksichtigt und umgesetzt werden.

Bestehende rechtliche Verpflichtungen und Policys werden hier nur erwähnt, aber nicht vollständig dargestellt; ihre Gültigkeit wird vermutet. Darüber hinaus sollen sich alle mit KI befassten Entwickler*innen an die *Richtlinie für die Gestaltung KI-gestützter Geschäftsmodelle, Dienste und Produkte bei der Deutschen Telekom*[4] wie auch an die Datenschutzbestimmungen halten, in die z. B. der qualitätssichernde *Privacy and Security Assessment Prozess (PSA)*[5] einbezogen ist. Bei allen Zweifeln sollte das Ethik-Experten-Team zur Bewertung eines Projektes oder Produktes im Hinblick auf die digitale Ethik einbezogen werden: Digital-Ethics@Telekom.de.

4 Mehr Informationen in der YaM-Gruppe Privacy@AI, https://yam.telekom.de/docs/DOC-552202 (Registrierung notwendig); Leitfaden „Zur Datenschutzkonformen Gestaltung von KI-Gestützten Geschäftsmodellen, Diensten und Produkten bei der Deutschen Telekom“, Version 1.0; Stand 10.01.2019

5 Mehr Informationen in YaM, https://yam.telekom.de/groups/psa-en

Definition of AI – Evolutional Model

In general, artificial intelligence describes the imitation and emulation of human intelligence and its underlying processes with the help of machines, especially computer systems. The aforementioned processes include the fields of learning, concluding, self-correction and self-reflection.

As artificial intelligence is an evolutional technology that iterates itself, its use and its integration into the lives of humans must also follow an evolutional model. And evolutional model for the use of AI systems could consist of three levels:

On the first level, AI algorithms and features are only a marginal part of the software are used by users to support their decisions. It is purely a supportive system that does not make any decisions that have an impact on the user.

The second level describes the phase in which the AI system is an integral part of the overall IT system and is capable of autonomously proposing decisions to users. The user can accept or reject the proposals and is actively supported by the AI system. On the highest level, level three, the system consisting of majority AI algorithms performs clearly defined tasks autonomously. The execution is supervised by at least one user, who is superior to the AI system and has the authority to terminate the execution.

All three application levels exist for different tasks and market environments. Challenges include creating transitions between the different automation levels or deciding which level to use for which task and system.

The following steps apply for new projects and AI solutions. For AI solutions already developed and in use, they should also be taken into account in the case of adaptations, new releases and in operation. The aim of the authors was to gather the latest information and latest input on the ethical development of AI. Due to the speed of the technology‘s advancement, this should be regarded as a living document – we look forward receiving your suggestions, references to best practices and tips at Digital-Ethics@Telekom.de.

The following best practices are structured according to the *Digital Ethics Guidelines on AI*: They derive a concrete target picture, define several fields of action for each of the guidelines, and provide a to-do list for the different phases of a project, organized by addressee.

Definition von KI – evolutionäres Modell

Künstliche Intelligenz beschreibt im Allgemeinen die Nachahmung und Simulation des menschlichen Denkvermögens und der diesem zugrunde liegenden Prozesse mit Hilfe von Maschinen, insbesondere Computersystemen. Zu den genannten Prozessen gehören die Bereiche des Lernens, des Schlussfolgerns, der Selbstkorrektur und der Selbstreflexion.

Da es sich bei der künstlichen Intelligenz um eine evolutionäre Technologie handelt, die sich selbst iterieren kann, müssen auch ihre Nutzung und ihre Integration in das Leben der Menschen einem evolutionären Modell folgen. Ein Evolutionsmodell für die Nutzung von KI-Systemen könnte drei Ebenen beinhalten: Auf der ersten Ebene stellen KI-Algorithmen und -Funktionen nur einen marginalen Teil der Software dar und werden von Nutzer*innen zur Unterstützung ihrer Entscheidungen verwendet. Es ist ein rein unterstützendes System, das keine eigenständigen Entscheidungen trifft.

Die zweite Ebene beschreibt die Phase, in der das KI-System einen integralen Teil des gesamten IT-Systems darstellt und in der Lage ist, den Menschen eigenständig Entscheidungen vorzuschlagen. Diese können die Vorschläge annehmen oder ablehnen und werden dabei vom KI-System aktiv unterstützt.

Auf der höchsten Ebene, der dritten Ebene, erledigt das aus Majoritäts-KI-Algorithmen bestehende System klar definierte Aufgaben selbständig. Die Ausführung wird von mindestens einer Person überwacht, die dem KI-System übergeordnet und befugt ist, die Ausführung zu unterbrechen.

Alle drei Anwendungsebenen existieren für verschiedenen Aufgaben und Marktumfelder. Herausforderungen liegen unter anderem darin, Übergänge zwischen den verschiedenen Automatisierungsebenen zu schaffen oder zu entscheiden, welche Ebene für welche Aufgabe und welches System verwendet werden soll.

Die folgenden Schritte gelten für neue Projekte und KI-Lösungen. Für bereits entwickelte und im Einsatz befindliche KI-Lösungen sollten sie bei Anpassungen, neuen Releases und im Betrieb ebenfalls berücksichtigt werden. Es war unser Bestreben, neueste Informationen und neuesten Input zur ethischen Entwicklung der KI zu sammeln. Aufgrund der Geschwindigkeit, mit der die Technologie voranschreitet, soll dies ein lebendiges Dokument sein – wir freuen uns auf Vorschläge, Hinweise auf Best Practices und Tipps an Digital-Ethics@Telekom.de.

Die folgenden Best Practices sind entsprechend den *Digital Ethics Guidelines on AI* strukturiert: Sie leiten ein konkretes Zielbild ab, definieren und beschreiben für jede der Leitlinien einige Handlungsfelder und bieten eine nach Adressaten geordnete To-do-Liste für die verschiedenen Phasen eines Projekts.

01

We are responsible

01

Wir übernehmen Verantwortung

01
We are responsible

Responsibility – Documentation – Use of external AI

The human always remains responsible. Our solutions come with a clear definition of who is responsible for which AI system or feature. We are responsible for our products and services. And, we know who is responsible AI systems on the part of our partners and service providers.

With AI technology in its infancy, we are aware of our responsibility in development. We clarify which product or project owner has which responsibilities. As a prerequisite for working with partners and third-party vendors, we define clear guidelines on who we work with and keep a record of what responsibilities are associated with each AI function.

Target picture

For each and every part of an AI system, there is a dedicated person responsible for the system's conformity with the existing laws and policies. Additionally, the person responsible ensures the lawful data processing.

To-dos

Prior to development

- Responsible persons must be documented.
- Product Owner advises external suppliers in the field of AI of the validity of the current Supplier Code of Conduct as part of the order.
- Product Owner takes responsibility for open source components meeting the DT standards for KI.
- Team is familiar with the Digital Ethics Guideline on AI, is therefore aware of their own responsibility and uses this document as an opportunity for more in-depth information.

During development

- Product Owner ensures compliance with all laws and rules.

After launch

- Product Owner defines a person responsible for the entire development and operating period for the AI within DT.
- Product Owner defines a new responsible person in the case that they leave the project.

01
Wir übernehmen Verantwortung

Verantwortlichkeiten – Dokumentation – Verwendung externer KI

Der Mensch bleibt immer in der Verantwortung. Für unsere Lösungen ist klar definiert, wer für welches KI-System und welche KI-Funktion verantwortlich ist. Wir tragen die Verantwortung für unsere Produkte und Dienste – und wir wissen, wer seitens unserer Partner und Dienstleister die Verantwortung für die KI-Systeme trägt.

*Die KI-Technologie befindet sich noch in ihren Anfängen und wir sind uns unserer Verantwortung für die weitere Entwicklung bewusst. Wir sorgen für Klarheit, welche*r Produkt- oder Projektverantwortliche welche Zuständigkeiten hat. Als Voraussetzung für die Zusammenarbeit mit Partnern und Drittanbietern definieren wir eindeutige Vorgaben, mit wem wir zusammenarbeiten, und halten fest, welche Verantwortlichkeiten mit den jeweiligen KI-Funktionen einhergehen.*

Zielbild

Für jeden einzelnen Teil eines KI-Systems gibt es eine bestimmte Person, die für die Konformität des Systems mit den bestehenden Gesetzen und Richtlinien verantwortlich ist. Zusätzlich stellt die verantwortliche Person die rechtmäßige Datenverarbeitung sicher.

To-dos

Vor der Entwicklung

- Verantwortliche Personen werden dokumentiert.
- Product Owner weist externe Lieferanten im Bereich KI auf die Gültigkeit des aktuellen Supplier Code of Conduct als Teil des Auftrags hin.
- Product Owner übernimmt Verantwortung dafür, dass Open-Source-Komponenten den Telekom-Standards für KI entsprechen.
- Team kennt die Digital Ethics Guideline on AI, ist sich dadurch der eigenen Verantwortung bewusst und nutzt dieses Dokument als mögliche Vertiefung.

Während der Entwicklung

- Product Owner gewährleistet die Einhaltung aller Gesetze und Regeln.

Nach dem Launch

- Product Owner definiert eine verantwortliche Person für die gesamte Entwicklungs- und Betriebszeit der KI innerhalb der Deutschen Telekom.
- Product Owner definiert vor dem eventuellen Verlassen des Projekts eine neue verantwortliche Person.

Responsibility

Before starting a new project, the overall person responsible should be defined and their name, work e-mail address and work telephone number documented. By default, the Product Owner is the person responsible, but the responsibility can also be assumed by another team member. In the case that another person is defined as being in charge, this should also be documented. This can be done as part of the PSA process. The project manager should make sure that responsibilities for each part of the project are defined and that the responsibilities are transparent for those involved. The same applies for external contractors or partners.
Responsibility in the sense of this guideline must not conflict with the project function of a person or party. The responsible person must be informed about possible risks as well as about measures that can be initiated in the event of an incident in order to avert damage.

Documentation

There is no formal requirement on how to document this information. We recommend using existing project management tools used within the project (e.g. Confluence, Wiki, etc.) or to integrate this into existing project organization documents (e.g. project charta, project plans, etc.). The documentation should be easy accessible by the project team.

Use of external AI

The Supplier Code of Conduct[6] is binding when external AI is used. Additionally, a contact person should be documented for the partner.

6 Supplier Code of Conduct: https://www.telekom.com/en/company/global-procurement/topics/how-to-become-a-supplier-526024

Verantwortlichkeiten

Vor Beginn eines neuen Projekts sollte die gesamtverantwortliche Person festgelegt und deren Name sowie Arbeits-E-Mail-Adresse und Arbeits-Telefonnummer dokumentiert werden. Standardmäßig ist der*die Product Owner die verantwortliche Person, es kann aber auch ein anderes Team-Mitglied sein. Wenn eine andere Person als verantwortlich definiert wird, sollte dies ebenfalls dokumentiert werden. Dies kann auch im Rahmen des PSA-Prozesses geschehen. Die projektleitende Person stellt sicher, dass die Verantwortlichkeiten für jeden Teil des Projekts festgelegt und dass die jeweiligen Zuständigkeiten für die beteiligten Personen transparent sind. Dasselbe gilt für externe Auftragnehmer oder Partner.
Die Verantwortlichkeit im Sinne dieses Leitfadens darf nicht mit der Projektfunktion einer Person oder Partei im Widerspruch stehen. Die verantwortliche Person muss über mögliche Risiken informiert werden sowie über Maßnahmen, die im Falle eines Ereignisses eingeleitet werden können, um Schaden abzuwenden.

Dokumentation

Für die Dokumentation der Zuständigkeiten gibt es keine formalen Anforderungen. Es wird empfohlen, bestehende Projektmanagementtools, die in Ihrem Projekt eingesetzt werden (z. B. Confluence, Wiki, etc.), zu verwenden oder diese in bestehende Dokumente der Projektorganisation (z. B. Projektcharta, Projektpläne, etc.) zu integrieren. Die Dokumentation sollte für das Projektteam leicht zugänglich sein.

Die Verwendung externer KI

Der Supplier Code of Conduct[6] ist bei Verwendung externer KI verbindlich, darüber hinaus sollte eine Kontaktperson des Partners dokumentiert werden.

6 Supplier Code of Conduct, https://www.telekom.com/en/company/global-procurement/topics/how-to-become-a-supplier-526024

02

We care

02

Wir gehen sorgsam mit KI um

02

We care

Compliance with official rules, laws and guidelines – Rules for the application of the system – PSA process

We act in line with our company values. Our systems and solutions must be compliant with human-defined rules, law and legislation. In addition to our high technical requirements, our AI systems are therefore subject to the rules, law and legislation that apply to our employees – and to humans in general.

AI systems must meet the same high technical requirements as any other of our IT systems, for example in the areas of security and stability. As artificial intelligence is becoming (and already is) part of our everyday lives and will even advise and guide us, we must ensure that AI systems and their use conform to our corporate values (such as DT's Guiding Principles and Code of Conduct), fundamental ethical values as well as social conventions.

Target picture

All digital and AI solutions comply with the IT standards. Furthermore, ethical fundamental values are converted into clearly defined rules, e.g. no misuse of software for self-beneficial reasons. The person responsible ensures the conformity of the project/product with these rules and all applicable laws.

To-dos

Prior to development

- Product Owner is familiar with all relevant rules, laws, and regulations and seeks assistance as necessary.

During development

- Product Owner ensures that the team is familiar with the relevant laws, rules, and DT regulations and acts in accordance with these.
- Team goes through the PSA process.
- Team defines clear rules for users of the AI system and documents these.

After launch

- Product Owner ensures that the product or service complies with the latest rules, laws and guidelines.

02

Wir gehen sorgsam mit KI um

Einhaltung der offiziellen Regeln, Gesetze und Richtlinien – Regeln für die Anwendung des Systems – PSA-Prozess

Wir handeln im Einklang mit unseren Unternehmenswerten. Unsere Systeme und Lösungen müssen mit den vom Menschen definierten Regeln, Recht und Gesetzen konform sein. Unsere KI-Systeme unterliegen deshalb — zusätzlich zu unseren hohen technischen Anforderungen — den Regeln, dem Recht und den Gesetzen, die für unsere Beschäftigten – und die Menschen an sich – gelten.

KI-Systeme müssen dieselben hohen technischen Anforderungen – etwa in Bereichen der Sicherheit und der Stabilität – erfüllen wie alle unsere IT-Systeme. Da künstliche Intelligenz ein Bestandteil unseres täglichen Lebens wird (und bereits ist) und uns sogar beraten und leiten wird, müssen wir dafür sorgen, dass die KI-Systeme und ihre Nutzung mit unseren Unternehmenswerten (wie zum Beispiel den Leitlinien und dem Code of Conduct der Deutschen Telekom), ethischen Grundwerten sowie gesellschaftlichen Konventionen konform sind.

Zielbild

Für alle digitalen und KI-Lösungen werden die IT-Standards eingehalten. Darüber hinaus sind ethische Grundwerte in klar definierte Regeln überführt wie: *Kein Missbrauch von Software zum Selbstzweck.* Die verantwortliche Person stellt die Konformität des Projekts/Produkts mit diesen Regeln und allen geltenden Gesetzen sicher.

To-dos

Vor der Entwicklung

- Product Owner kennt die relevanten Regeln, Gesetze und Vorschriften und nimmt ggf. Hilfestellung in Anspruch.

Während der Entwicklung

- Product Owner stellt sicher, dass das Team die relevanten Gesetze, Regeln und Vorschriften der Telekom kennt und danach handelt.
- Team durchläuft den PSA-Prozess.
- Team legt klare Regeln für Nutzer*innen des KI-Systems fest und dokumentiert diese.

Nach dem Launch

- Product Owner stellt sicher, dass das Produkt oder die Dienstleistung mit den aktuellsten Regeln, Gesetzen und Richtlinien konform ist.

Compliance with official rules, laws and guidelines

The development and use of AI systems and solutions should be conducted on the basis of a thorough evaluation of laws and regulations, DT's IT standards, DT's Code of Conduct[7], our Digital Ethics Guidelines on AI and Guiding Principles[8] and the societal as well as cultural conventions and values affected by the systems in question. Cultural conventions and values deviate depending on the region of deployment. If in doubt, European values should be taken as standard. The Product Owner is responsible for ensuring that all relevant laws and regulations are complied with – at national and international level. These include, among other: EU primary law EU directives, the Constitution of the European Union and its fundamental rights, EU secondary law (such as the General Data Protection Regulation, the Product Liability Directive, the Regulation on the Free Flow of Non-Personal Data, anti-discrimination directives, consumer law, and directives on health and safety at work), as well as numerous laws of the EU member states, depending on the market on which the product is used. For use in a specific area or topic, specific rules must be taken into account, such as medical device regulation, regulations for the financial market, critical infrastructure or autonomous driving.

Rules for usage

Manufacturers generally cannot exclude the possibility that a product may be misused. Nevertheless, measures can be taken so that the manufacturer's responsibilities can be clearly separated from those of the user. The more critical the use case, the more clearly the customer should be informed and made aware of his or her responsibilities when operating the product or service.

PSA process

Successful completion of the PSA process[9] is mandatory.

7 Deutsche Telekom Code of Conduct: https://www.telekom.com/en/company/compliance/code-of-conduct
8 Corporate Guiding Principles: https://www.telekom.com/en/company/company-profile/corporate-values
9 PSA process: https://psa-portal.telekom.de

Einhaltung der offiziellen Regeln, Gesetze und Richtlinien

Die Entwicklung und Nutzung von KI-Systemen und -Lösungen sollte auf der Grundlage einer gründlichen Evaluierung der Gesetze und Vorschriften, der IT-Standards der DT, des Code of Conduct der DT[7], der *Digital Ethics Guidelines on AI* der Deutschen Telekom und der *Guiding Principles*[8] sowie der gesellschaftlichen und kulturellen Konventionen und Werte erfolgen, innerhalb derer die entsprechenden Systeme eingesetzt werden. Kulturelle Konventionen und Werte können je nach Einsatzregion abweichen. Im Zweifelsfall sind die europäischen Werte als Maßstab heranzuziehen. Der*die Product Owner ist dafür verantwortlich, dass alle relevanten Gesetze und Regeln eingehalten werden – auf nationaler und internationaler Ebene. Dazu gehören unter anderem: das EU-Primärrecht, die EU-Richtlinien, die Verfassung der Europäischen Union und ihre Grundrechte, das EU-Sekundärrecht (wie die Allgemeine Datenschutzverordnung, die Produkthaftungsrichtlinie, die Verordnung über den freien Verkehr nicht-persönlicher Daten, Antidiskriminierungsrichtlinien, Verbraucherrecht und Richtlinien über Sicherheit und Gesundheitsschutz am Arbeitsplatz) sowie zahlreiche Gesetze der EU-Mitgliedstaaten, je nachdem, auf welchem Markt das Produkt verwendet wird. Für den Einsatz in bestimmten Bereichen oder zu bestimmten Themen kann es wiederum spezifische Regeln geben, die berücksichtigt werden müssen, wie z. B. die Verordnung über Medizinprodukte, Vorschriften für den Finanzmarkt, kritische Infrastrukturen oder autonomes Fahren.

Regeln für die Anwendung des Systems

Hersteller können im Allgemeinen die Möglichkeit nicht ausschließen, dass ein Produkt missbräuchlich verwendet wird. Dennoch können Maßnahmen ergriffen werden, damit die Verantwortlichkeiten des Herstellers klar von denen der Nutzerinnen und Nutzer getrennt werden können. Je kritischer der Anwendungsfall, desto klarer sollten Kund*innen informiert und auf ihre Verantwortung beim Betrieb des Produkts oder der Dienstleistung hingewiesen werden.

PSA-Prozess

Der erfolgreiche Abschluss des PSA-Prozesses[9] ist obligatorisch.

7 Deutsche Telekom Code of Conduct, https://www.telekom.com/en/company/compliance/code-of-conduct
8 Corporate Guiding Principles, https://www.telekom.com/en/company/company-profile/corporate-values
9 PSA-Prozess, https://psa-portal.telekom.de

03

We put customers first

03

Wir stellen unsere Kund*innen in den Mittelpunkt

03
We put customers first

Human-centered KI – Universal design – The use of data – Machine learning and data privacy

We enrich and simplify our customers' lives. If an AI system or the usage of customer-related data helps us to benefit our customers, we embrace this opportunity to meet their demands and expectations.

The collection and use of customer data – especially in AI systems – should always serve a useful purpose towards our customers. This applies to systems and processes that support in the background as well as services that interact with our customers directly.

Target picture

Developing and using AI solutions is not an end in itself. It should be purposeful. If the AI solution provides a practical effect for the customer, it should be used. The AI solutions should simplify customers' lives and support them, whether internally or externally. This trust is becoming evident by users delegating decisions in certain areas to AI-based programs.

To-dos

Prior to development

- Product Owner plans which user data should be collected and if anonymization of the data is required.
- Product Owner ensures compliance with the regulations for data use.

During development

- Product Owner ensures an easy-to-use, transparent interface and tests whether people of all ages and people with disabilities are able to use the product or service.
- Product Owner ensures that customer or testing data is protected in all phases of the product or service.
- Developer implements functionalities to give customers options regarding the handling of their data.

03
Wir stellen unsere Kund*innen in den Mittelpunkt

Menschenzentrierte KI – Universelles Design – Die Verwendung von Daten – Maschinelles Lernen und Datenschutz

*Wir vereinfachen und bereichern das Leben unserer Kund*innen. Wenn künstliche Intelligenz und die Nutzung kundenbezogener Daten uns dabei helfen, Lösungen im Sinne unserer Kund*innen zu entwickeln, begrüßen wir dies als Chance, die Bedürfnisse und Erwartungen unserer Kund*innen zu erfüllen.*

*Die Sammlung und Weiterverarbeitung kundenbezogener Daten – insbesondere in KI-Systemen – soll unseren Kund*innen nutzen. Dies betrifft Systeme und Prozesse, die im Hintergrund unterstützen, genauso wie solche, die unmittelbar mit unseren Kund*innen interagieren.*

Zielbild

Die Entwicklung und Nutzung von KI-Lösungen ist kein Selbstzweck. Sie erfolgt zielgerichtet. Wenn die KI-Lösung einen anwendungspraktischen Nutzen hat, ohne an anderer Stelle zu riskieren, Schaden zu erzeugen, soll sie eingesetzt werden. Die KI-Lösungen sollen das Leben der Menschen vereinfachen und sie unterstützen, sei es intern oder extern. Dieses Vertrauen erfüllt sich in der Delegation von Entscheidungen von Nutzerinnen und Nutzern an KI-basierte Programme in bestimmten Bereichen.

To-dos

Vor der Entwicklung

- Product Owner plant, welche Nutzerdaten gesammelt werden sollen und ob eine Anonymisierung der Daten erforderlich ist.
- Product Owner stellt die Einhaltung der Vorschriften für die Verwendung von Daten sicher.

Während der Entwicklung

- Product Owner sorgt für ein gut benutzbares, klares Interface und testet, ob Menschen aller Altersgruppen und Menschen mit Behinderungen das Produkt oder die Dienstleistung nutzen können.
- Product Owner stellt sicher, dass Kunden- oder Testdaten in allen Phasen des Produkts oder der Dienstleistung geschützt sind.
- Developer implementiert Funktionen, die Kund*innen Entscheidungen über den Umgang mit ihren Daten ermöglichen.

After launch
- Product Owner ensures that the AI meets all guidelines and regulations.
- Product Owner implements an overview for customers on how their data will be used.

Human-centered AI

Putting the customer first means ensuring that the motivations for the project are based on customer needs even before development begins. Design Thinking Methods[10] and Telekom Personas (both employees and customers) can be used for this purpose, for example. During the development process and prior to rollout, the match between customer needs and the envisioned solutions should be continuously checked.

This customer focus is also important for avoiding bias, for the fairness of the system, and for ensuring that the desired function is performed in the individual case. Always keep possible influential cognitive biases (see definition) of the team in mind during development.

Universal Design

Especially in consumer services, the systems should be designed in a way that allows all possible customers to use the service, regardless of their age, gender, abilities, or characteristics. In addition, neither race, color, religion, political or other opinion, national or social origin, membership in a national minority, property, birth or other status shall make a difference in the operation of the system.[11] To do so, consider the principles of universal design and the relevant accessibility standards.[12] In addition, DT's design templates and interfaces can be found in the Telekom Brand and Design Guide.[13]

The use of data

It is obligatory to provide information about the processing purposes of the personal data (follows Art. 6 DSGVO Rechtmäßigkeit der Verarbeitung[14]). It must be clear to the user which data must be collected in order to provide the service, e.g. full name and billing address. The information required for each individual possible function must be documented so that it can be integrated as soon as it is available. For this, please use a way to provide this information in a concise, comprehensible, and easy-to-understand manner and in clear, simple language. This can be done in the service itself or within an external service like a website. It is also necessary that customers have the opportunity to object to the use of their personal data or to have their personal data subsequently deleted. To incorporate the customer's potential decision not to share their data into the process, a mechanism to stop the collection of the data must be implemented in the functionalities from the very beginning. If the purpose of the data processing changes during the life of the product, the permission to use the available data must be reviewed by the person responsible in accordance

10 Telekom New Work Akademy: https://yam.telekom.de/groups/newworkacademy/projects/new-workacademy-english/pages/methods
11 General Equal Treatment Act; Art. 8 G v. 3.4.2013 I 610
12 For instance EN 301 549.
13 Telekom Brand and Design, see https://www++.brand-design.telekom.com
14 See also: Regulation of the European parliament: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

Nach dem Launch
- Product Owner stellt sicher, dass die AI jederzeit allen aktuellen Richtlinien und Vorschriften entspricht.
- Product Owner implementiert einen Überblick für Kund*innen zur Verwendung ihrer Daten.

Menschenzentrierte KI

Den Menschen an die erste Stelle setzen, heißt, bereits vor Beginn der Entwicklung sicherzustellen, dass die Beweggründe für das Projekt auf den Kundenbedürfnissen basieren. Dazu können zum Beispiel Methoden des Design Thinking[10] und die Telekom Personas eingesetzt werden. Während des Entwicklungsprozesses und vor dem Rollout sollte die Übereinstimmung von Kundenbedürfnissen mit den angestrebten Lösungen laufend überprüft werden.

Dieser Kundenfokus ist wichtig zum Vermeiden von Verzerrungen, für die Fairness des Systems und zur Sicherstellung, dass die gewünschte Funktion auch im Einzelfall ausgeführt wird. Haben Sie mögliche einflussreiche kognitive Verzerrungen des Teams während der Entwicklung immer im Blick (siehe Glossar im Anhang → Bias).

Universelles Design

Insbesondere im Privatkundengeschäft sollten die Systeme so gestaltet sein, dass alle potenziellen Kundinnen und Kunden den Dienst nutzen können, unabhängig von Alter, Geschlecht, Fähigkeiten oder Eigenschaften. Außerdem dürfen weder Rasse, Hautfarbe, Religion, politische oder sonstige Anschauung, nationale oder soziale Herkunft, Zugehörigkeit zu einer nationalen Minderheit, Vermögen, Geburt noch der sonstige Status einen Unterschied für das Funktionieren des Systems machen.[11] Berücksichtigen Sie dafür die Grundsätze des Universal Design und die geltenden Normen für die Zugänglichkeit.[12] Darüber hinaus sind die Designvorlagen und Oberflächen der Telekom im Telekom Brand and Design Guide zu finden.[13]

Die Verwendung von Daten

Es besteht eine Informationspflicht über die Zwecke der Verarbeitung der personenbezogenen Daten (in Anlehnung an Art. 6 DSGVO, Rechtmäßigkeit der Verarbeitung[14]). Bei der Nutzung muss klar sein, welche Daten zur Erbringung der Dienstleistung erhoben werden müssen, z. B. vollständiger Name und Rechnungsanschrift. Die benötigten Informationen für jede einzelne mögliche Funktion müssen dokumentiert werden, damit sie integriert werden können, sobald diese zur Verfügung steht. Verwenden Sie dazu bitte eine Möglichkeit, diese Informationen in einer präzisen, verständlichen und leicht verständlichen Art und Weise und in klarer und einfacher Sprache zur Verfügung zu stellen. Dies kann im Dienst selbst oder innerhalb eines externen Dienstes wie einer Website geschehen. Es muss die Möglichkeit geben, der Verwendung der eigenen Daten

10 Telekom New Work Akademy, https://yam.telekom.de/groups/newworkacademy/projects/new-workacademy-english/pages/methods
11 Allgemeines Gleichbehandlungsgesetz; Art. 8 G v. 3.4.2013 I 610
12 Zum Beispiel EN 301 549
13 Telekom Brand and Design, https://www.brand-design.telekom.com
14 Siehe auch *Regulation of the European parliament*, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

with ethics guidelines No. 1 "We are responsible" and, if necessary, new permission must be obtained from the customer. The lawfulness of the data processing must be guaranteed at all times and communicated to the customer.

Machine learning and data privacy

The "Guideline for data privacy-compliant design of AI-supported business models, services and products at Deutsche Telekom"[15] provides guidance on evaluation, the creation of transparency and data privacy impact assessment from a data privacy perspective and highlights difficulties that may arise when using AI in compliance with the law.

In the following, methods for dealing with these challenges are presented.
The purpose of an AI can be to achieve other objectives or other benefits by intelligently combining different input data. Concrete information on as yet unknown purposes is logically not possible in practice. However, further processing of the data is only possible with the original purpose under the conditions of Art. 6 (4) DSGVO.

In any case, an attempt should be made to formulate the purposes – including future purposes – for which the product or service could be used, as specifically as possible. Conceivable future purposes can be found, for example, through creative thinking based on practical examples. It is advisable to formulate the purposes in this agreement in a transparent but sufficiently flexible manner to be able to achieve compatibility with the new purposes later. A Bitkom publication states: "The performance claim of the processing of personal data for ML (machine learning) is not determined exclusively by the assessment of whether or not a service can be provided at all – the decisive factor is rather whether it can be provided in the contractually agreed form and quality".[16]

Example: An AI is used to recognize bird species. For this purpose, the user is asked for consent for the camera function, the location and the microphone function of the smartphone, as the birds are recognized via the image and sound at the moment of the camera recording. In the course of processing, it turns out that forest populations are shrinking at the location of some users, on all of whose sound recordings the AI detected a certain noise which comes from an insect which damages trees.

At the beginning of the project, it was not clear that the tree inventory data would be integrated into the AI. However, the intent was to provide additional value to the service through extended data. Therefore, for such a project, it should be communicated before the first use that the anonymized results of the recordings will be combined with geological and environmental developments.
Various technologies[17] can be used to make working with personal data more secure, also in connection with machine learning processes, or even to exempt it completely from further transparency requirements:

15 Version 1.0; 10.01.2019
16 Bitkom: Machine Learning und die Transparenzanforderungen der DS-GVO, 2018, 30–41
17 Bitkom: Machine Learning und die Transparenzanforderungen der DS-GVO, 2018, 30–41

zu widersprechen oder persönliche Daten nachträglich löschen zu lassen. Um die potentielle Entscheidung, die eigenen Daten nicht weiterzugeben, in den Prozess einzubeziehen, muss von Anfang an ein Mechanismus zum Stoppen der Erfassung der Daten implementiert werden. Ändert sich der Zweck der Datenverarbeitung während der Laufzeit des Produkts, muss die Erlaubnis zur Nutzung der vorliegenden Daten von der verantwortlichen Person gemäß der Leitlinie 01 – *Wir übernehmen Verantwortung* überprüft und bei Bedarf eine erneute Erlaubnis eingeholt werden. Die Rechtmäßigkeit der Datenverarbeitung muss zu jeder Zeit gewährleistet und kommuniziert sein.

Maschinelles Lernen und Datenschutz

Der *Leitfaden zur datenschutzkonformen Gestaltung von KI-gestützten Geschäftsmodellen, Diensten und Produkten bei der Deutschen Telekom*[15] gibt aus datenschutzrechtlicher Sicht Hinweise zur Bewertung, zur Schaffung von Transparenz und zur Datenschutz-Folgenabschätzung. Außerdem zeigt er Schwierigkeiten auf, die bei der gesetzeskonformen Nutzung von KI auftauchen können.

In folgendem werden Methoden zum Umgang mit diesen Herausforderungen aufgezeigt.
Sinn und Zweck einer KI kann sein, dass durch das intelligente Verbinden von verschiedenen Eingabedaten auch andere Verwendungszwecke mit neuem Nutzen erzielt werden.
Konkrete Hinweise zu noch unbekannten Verwendungszwecken sind in der Praxis nicht möglich; eine Weiterverarbeitung der Daten ist jedoch nur mit dem ursprünglichen Zweck unter den Bedingungen des Art. 6 Abs. 4 DSGVO zulässig.

In jedem Fall sollte versucht werden, die – auch zukünftigen – Zwecke, für welche das Produkt oder der Service eingesetzt werden könnte, so konkret wie möglich zu formulieren. Zukünftig denkbare Zwecke können z. B. durch kreatives Nachdenken anhand von Beispielen aus der Praxis gefunden werden. Es empfiehlt sich, die Zwecke in dieser Vereinbarung transparent, aber ausreichend flexibel zu formulieren, um später die Kompatibilität mit den neuen Verwendungszwecken erreichen zu können. In einer Veröffentlichung der Bitkom heißt es: „Der Leistungsanspruch der Bearbeitung von Personendaten für ML (Machine Learning) wird nicht ausschließlich durch die Beurteilung bestimmt, ob eine Leistung überhaupt erbracht werden kann oder nicht – entscheidend ist vielmehr, ob sie in der vertraglich vereinbarten Form und Qualität erbracht werden kann".[16]

Beispiel: Eine KI wird verwendet, um Vogelarten zu erkennen. Hierzu wird vor der ersten Nutzung die Erlaubnis zur Verwendung der Kamerafunktion, des Standorts und der Mikrofonfunktion des Smartphones eingeholt, denn ein Vogel soll anhand des Bildes und des Tones im Moment der Kameraaufnahme erkannt werden. Im Verlauf der Verarbeitung stellt sich heraus, dass Waldbestände am Standort einiger Nutzer*innen schrumpfen, deren Tonaufnahmen alle ein bestimmtes von der KI erkanntes Geräusch enthalten, das von einer die Bäume schädigenden Insektenart stammt.

Zu Beginn des Projekts stand nicht fest, dass die Daten zum Baumbestand in die KI eingebunden werden. Allerdings war beabsichtigt, dem Service durch erweiterte Daten einen zusätzlichen

15 Version 1.0; Stand 10.01.2019
16 Bitkom, Machine Learning und die Transparenzanforderungen der DS-GVO, 2018, 30–41

1. Pseudonymization: Pseudonymization attempts to protect data by replacing the values of direct identifiers (ID card number, name, etc.) of a data set with pseudonyms. These pseudonyms are either generated from the original value or assigned completely new. A pseudonym can have the same format as the original data type, e.g. an artificial name or random ID card number, or it can be given a completely new type of designation. It is only important that the assignment is unique, i.e. that the same pseudonym is always generated for two identical input values. For many applications, pseudonymization must also be reversible, i.e. it must be possible to derive the original data value from the pseudonym, if necessary with additional information such as a key or a table.

2. K-Anonymization: K-anonymization is a characteristic of a database that represents a certain degree of anonymity. The model states that every entry about a person in the database should be impossible to distinguish from K-1 others.[18] If data are combined in such a way that K persons look the same, the individual privacy of each data subject is protected. In K anonymity, the value of K is the unit of measurement for anonymity. If K=1, then the original data is unchanged and there is no anonymity – the greater the value of K, the better the anonymity.

3. Differential Privacy: Differential privacy gives a measure of the risk that a data set runs of being assigned to a person within a statistical database. The principle applies that personal data must not lead to a difference in the results of an evaluation.[19]

18 Latanya Sweeney. 2002. »k-anonymity: A model for protecting privacy«, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10, 05 (2002), 557–570.

19 Petrlic, Ronald und Sorge, Christoph. Datenschutz: Einführung in technischen Datenschutz, Datenschutzrecht und angewandte Kryptographie. Springer Vieweg, 2017

Nutzen zu ermöglichen. Daher ist für ein solches Projekt vor der ersten Nutzung zu kommunizieren, dass die anonymisierten Ergebnisse der Aufnahmen mit geologischen und umweltbasierten Entwicklungen kombiniert werden.

Verschiedene Techniken[17] können eingesetzt werden, um den Umgang mit personenbezogenen Daten auch im Zusammenhang mit maschinellen Lernverfahren sicherer zu machen oder sogar ganz von zusätzlichen Transparenzanforderungen zu befreien:

1. Pseudonomisierung: Bei der Pseudonymisierung wird versucht, Daten zu schützen, indem die Werte der direkten Identifikatoren (Ausweisnummer, Name usw.) eines Datensatzes durch Pseudonyme ersetzt werden. Diese Pseudonyme werden entweder aus dem ursprünglichen Wert generiert oder völlig neu vergeben. Ein Pseudonym kann dasselbe Format wie der ursprüngliche Datentyp haben, z. B. ein Kunstname oder eine zufällige Personalausweis-Nummer oder eine ganz neue Art der Bezeichung bekommen. Wichtig ist nur, dass die Zuordnung eindeutig ist, d.h. dass für zwei identische Eingabewerte immer das gleiche Pseudonym generiert wird. Für viele Anwendungen muss eine Pseudonymisierung zusätzlich reversibel sein, d.h. es muss möglich sein, aus dem Pseudonym den ursprünglichen Datenwert abzuleiten, ggf. mit zusätzlichen Informationen wie mit einem Schlüssel oder einer Tabelle.

2. K-Anonymisierung: K-Anonymisierung ist eine Eigenschaft einer Datenbank, die einen gewissen Grad an Anonymität aufweist. Das Modell besagt, dass jeder Eintrag über eine Person in der Datenbank unmöglich von K-1 anderen zu unterscheiden sein sollte.[18] Wenn Daten so kombiniert werden, dass K Personen gleich aussehen, wird die individuelle Privatsphäre jeder betroffenen Person geschützt. In der K-Anonymität ist der Wert von K die Maßeinheit für die Anonymität. Wenn K=1, dann sind die Originaldaten unverändert und es gibt keine Anonymität – je grösser K, desto höher die Anonymität.

3. Differential Privacy: Die Differential Privacy soll verhindern, dass ein Datensatz Gefahr läuft, innerhalb einer statistischen Datenbank eindeutig einer Person zugeordnet werden zu können. Durch zufällige Datenzeilen ist in Datensätzen, die mit Diffential Privacy geschützt wurden, jedes Ergebnis gleich wahrscheinlich. Es gilt der Grundsatz, dass personenbezogene Daten nicht zu einem Unterschied in den Ergebnissen einer Auswertung führen dürfen.[19]

17 ebd., 40ff

18 Latanya Sweeney, k-anonymity: A model for protecting privacy, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10, 05 (2002), 557–570.

19 Petrlic, Ronald und Sorge, Christoph, Datenschutz: Einführung in technischen Datenschutz, Datenschutzrecht und angewandte Kryptographie, Springer Vieweg, 2017

04

We are transparent

04

Wir stehen für Transparenz

04
We are transparent

Honesty – communicating limitations and purpose

We make it clear to our customers when they are communicating with an AI system. We are also transparent about how we use customer data. As Deutsche Telekom, we always have the customer's trust in mind – trust is what we stand for.

We act transparently towards our customers. It is obvious to our customers that they are interacting with an AI when they do. In addition, we make it clear to our customers how and to which extent they can decide how their data is processed.

Target picture

In order to prevent misunderstanding and confusion, an AI solution must never be disguised as a human being. Furthermore, Deutsche Telekom always has an overview of AI systems and features developed and used within and by Deutsche Telekom. Limitations and purpose should be as well communicated to the customer.

To-dos

Prior to development

- Product Owner analyzes the type of customer relationship the AI will have and plans the communication accordingly.

During development

- Product Owner tests the appearance of the AI with testers who are not yet familiar with the service.

Honesty

AI systems should never present themselves to users as humans; humans have the right to be informed that they are interacting with an AI system. By default, this should be stated at least in the GTCs, but there are various options which should be considered in order to comply with ethical values avoid misunderstandings. Include the following examples in your service if the service could give the customer the impression of a real human:

04
Wir stehen für Transparenz

Ehrlichkeit – Grenzen und Zweck kommunizieren

*Wir machen es für unsere Kund*innen klar ersichtlich, wenn sie mit einem KI-System interagieren. Zudem legen wir offen, wie wir Kundendaten nutzen. Das Vertrauen unserer Kund*innen bestimmt unser Handeln - denn wir als Deutsche Telekom stehen für Vertrauenswürdigkeit*

*Wir handeln transparent für unsere Kund*innen. Interagieren sie mit einem KI-System, dann ist es für sie erkennbar. Darüber hinaus machen wir für sie deutlich, wie und in welchem Maße sie über die Art und Weise der Weiterverarbeitung ihrer Daten entscheiden können.*

Zielbild

Um Missverständnisse und Verwirrung zu vermeiden, versucht eine KI-Lösung niemals, wie ein Mensch zu wirken. Darüber hinaus behält die Deutsche Telekom jederzeit einen Überblick über KI-Systeme und Funktionen, die innerhalb und von der Deutschen Telekom entwickelt und eingesetzt werden. Grenzen und Zweck der KI sollten immer klar kommuniziert werden.

To-dos

Vor der Entwicklung

- Product Owner analysiert die Art der Kundenbeziehung, die die KI haben wird und plant die Kommunikation entsprechend.

Während der Entwicklung

- Product Owner testet das Erscheinungsbild der KI mit Personen, für die der Dienst noch unbekannt ist.

Ehrlichkeit

KI-Systeme sollten sich gegenüber Nutzerinnen und Nutzern nie als Menschen ausgeben; Menschen haben das Recht, darüber informiert zu werden, dass sie mit einem KI-System interagieren. Standardmäßig muss dies mindestens in den AGBs festgehalten werden, aber Sie sollten weitere Optionen in Betracht ziehen, um den ethischen Werten gerecht zu werden und Missverständnisse zu vermeiden. Nehmen Sie folgende Beispiele in Ihren Service auf, falls der Service Nutzerinnen und Nutzern unter Umständen fälschlicherweise den Eindruck eines echten Menschen vermitteln könnte:

- Context based: implement statements such as "if I were a human ..." or refer to the identity "digital assistant".
- Via the name: use names not commonly used for humans (examples below).
- Graphics-based: implement a graphical representation of a robot or other species to clarify an artificial identity.

Examples:

(1) Karl Klammer Tinka Woebot

Communicate limitations and purpose

The AI system's capabilities and limitations should be communicated to AI practitioners or end-users depending on the use case at hand. This could encompass communication of the AI system's level of accuracy, as well as its limitations. In order to prevent miscommunication and confusion, but also to avoid a fear of technology, it may be necessary to communicate a brief overview of how the AI reaches its results. Furthermore, a decision taken solely by an AI system should be identified as such.

Figure 1:
Karl Klammer: https://apps.derstandard.de/privacywall/story/2000121096252/clippy-microsofts-virtuelle-bueroklammer-die-nutzer-in-den-wahnsinn-trieb
Tinka: https://newsroom.magenta.at/2017/02/27/t-mobile-chatbot-tinka-im-rampenlicht-des-mobile-world-congress/
Woebot: https://chatbotslife.com/meet-woebot-the-mental-health-chatbot-changing-the-face-of-therapy-44e8c6ff4fc2

- Kontextbasiert: Aussagen wie „Wenn ich ein Mensch wäre ...“ oder Verweise auf die Identität „Digitaler Assistent“ einbauen.
- Durch den Namen: Verwendung nicht gebräuchlicher menschlicher Namen (siehe unten).
- Grafisch: Mit der Darstellung eines Roboters oder einer anderen Spezies verdeutlichen, dass das Gegenüber nur eine künstliche Identität hat.

Beispiele:

(1) Karl Klammer Tinka Woebot

Grenzen und Zweck kommunizieren

Die Möglichkeiten und Grenzen des KI-Systems sollten je nach Anwendungsfall umfassend kommuniziert werden. Dies könnte sowohl die Mitteilung über den Genauigkeitsgrad des KI-Systems als auch über dessen Schwächen umfassen. Um Fehlkommunikation, Verwirrung, aber auch Angst vor der Technologie zu vermeiden, kann ein kurzer Überblick darüber, wie die KI zu ihren Ergebnissen gelangt, notwendig sein. Darüber hinaus ist eine ausschließlich von einem KI-System getroffene Entscheidung entsprechend zu kennzeichnen.

Abbildung 1:
Karl Klammer: https://apps.derstandard.de/privacywall/story/2000121096252/clippy-microsofts-virtuelle-bueroklammer-die-nutzer-in-den-wahnsinn-trieb
Tinka: https://newsroom.magenta.at/2017/02/27/t-mobile-chatbot-tinka-im-rampenlicht-des-mobile-world-congress/
Woebot: https://chatbotslife.com/meet-woebot-the-mental-health-chatbot-changing-the-face-of-therapy-44e8c6ff4fc2

05

We are secure

05

Wir bieten Sicherheit

05
We are secure

Damage security – Security measures – Robust AI – Data privacy – External threat – Emergency button

Data security and data privacy are part of DT's self-image. We not only ensure that our security measures are in line with the latest state of development, but also keep track of how customer-related data is used – and who is allowed to access what data.

We do not process any data relevant to data privacy without a legal basis. This applies to our AI systems just as it does to all our other activities. We limit the use of customer-related data to permitted use cases. At the same time, we protect our systems from unauthorized external access to ensure data security and data privacy.

Target picture

Data privacy-relevant data is never processed without legal permission and is stored in a secure way, with end-to-end decryption. Access to encryption keys is protected. Furthermore, neither employees of DT nor of third parties will ever have access to privacy-relevant data without customer or legal permission. A list of components and their functionalities (e.g. server without storage) is maintained for AI systems and data that are distributed throughout several network components.

To-dos

Prior to development
- Product Owner assesses the necessary security measures.
- Product Owner communicates risks to superiors.
- Product Owner ensures that those involved in development keep the code, state of development and other information confidential.

During development
- Development Team keeps the systems and features updated according the latest technology and security standards.
- Development Team assesses the robustness and implements measures to increase the robustness depending on the use case.
- Development Team implements an emergency mode for shut-down.

05
Wir bieten Sicherheit

Schadenssicherheit – Sicherheitsmaßnahmen – Robust AI – Datensicherheit – Externe Bedrohung – Not-Aus-Schalter

Datensicherheit und Datenschutz gehören zum Selbstverständnis der Deutschen Telekom. Wir sorgen nicht nur dafür, dass unsere Sicherheitsmechanismen dem aktuellen Entwicklungsstand entsprechen, sondern behalten auch den Überblick darüber, wie kundenbezogene Daten genutzt werden – und wer auf welche Daten zugreifen darf.

Wir verarbeiten keine datenschutzrelevanten Daten ohne Rechtsgrundlage. Das gilt für unsere KI-Systeme genauso wie für alle anderen unserer Aktivitäten. Die Verwendung kundenbezogener Daten begrenzen wir auf erlaubte Anwendungsfälle. Gleichzeitig schützen wir unsere Systeme vor unerlaubtem externen Zugriff, um die Datensicherheit und den Datenschutz sicherzustellen.

Zielbild

Datenschutzrelevante Daten werden niemals ohne gesetzliche Erlaubnis verarbeitet und werden in einer sicheren, durchgehend verschlüsselten Weise gespeichert, wobei der Zugang zu Verschlüsselungscodes geschützt ist. Darüber hinaus können weder Mitarbeiter*innen der Deutschen Telekom noch Dritte jemals ohne Zustimmung von Kund*innen oder Gesetzgeber auf datenschutzrelevante Daten zugreifen. Für KI-Systeme und Daten, die über mehrere Netzkomponenten verteilt sind, wird eine Liste von Komponenten und deren Funktionalitäten (z. B. „Server ohne Speicherung“) geführt.

To-dos

Vor der Entwicklung
- Product Owner beurteilt notwendige Sicherheitsmaßnahmen.
- Product Owner kommuniziert Risiken an Vorgesetzte.
- Product Owner stellt sicher, dass die an der Entwicklung beteiligten Personen den Code, den Entwicklungsstand und andere Informationen vertraulich behandeln.

Während der Entwicklung
- Development Team hält Systeme und Funktionen auf dem neuesten Stand von Technik und Sicherheitsstandards.
- Development Team bewertet die Robustheit und führt Maßnahmen zur Steigerung der Robustheit je nach Anwendungsfall durch.
- Development Team implementiert einen Notfallmodus zum Abschalten.

After launch
- Product Owner keeps the security measures up-to-date through regular updates.
- Product Owner informs people working with the AI about the necessity of updates the emergency mode and other security measures.

Damage security

Security in AI systems not only means data protection, but also:

1. That it does not cause nor exacerbate damage. Damage can be of an individual or collective nature, and can include non-material damage to social, cultural and political environments.

2. That it does not otherwise adversely affect humans. This also includes the lifestyle of individuals and social groups.
Particularly vulnerable persons should receive greater attention and be included in the development and tests of the AI systems. Especially in situations where AI systems can cause or exacerbate adverse impacts due to asymmetries of power or information, such as between employers and employees, businesses and consumers or governments and citizens, the use case and the model must be analyzed carefully and the level of robustness must be accordingly high.[20]

AI – Security measures

Security measures for AI systems and the combination of AI functionalities as known for existing IT systems should be set up in a so-called meta system. The security measures are thus superior to the AI functionalities, which means that the security measures are executed even if an AI functionality is no longer or only partially functional in the process. The security system must be continuously executed and improved.

Robust AI

Robust AI solutions are solutions that are impervious to influence and perform as expected in uncertain and unpredictable environments without reproducing any human bias. When developing AI systems, these capabilities must be thoughtfully designed and implemented. The robustness of an AI system depends mainly on three factors: (1) the robustness of the (business) process in which the system is integrated, (2) the structure and type of the AI model used, and (3) the database that forms the basis for training the model. AI-based systems are known to provide accurate and high-quality predictions, especially if a good database is available. However, the more complex the data models an AI represents, the more difficult it becomes to understand why the system chooses a certain outcome and whether the decision is actually justified. Therefore, robust AI is a prerequisite for trustworthy AI. This is particularly necessary given that AI technology will increasingly be used

20 European Comission: Ethics Guidelines for Trustworthy AI; 2019

Nach dem Launch
- Product Owner hält die Sicherheitsmaßnahmen durch regelmäßige Updates auf dem neuesten Stand.
- Product Owner informiert Personen, die mit der KI arbeiten sollen, über die Notwendigkeit von Updates, den Notfallmodus und andere Sicherheitsmaßnahmen.

Schadenssicherheit

Sicherheit in KI-Systemen bedeutet nicht nur Datenschutz, sondern auch:

1. dass sie keinen Schaden verursachen. Schaden kann individueller oder kollektiver Natur sein und schließt auch immateriellen Schaden für das soziale, kulturelle und politische Umfeld ein.

2. dass sie Menschen nicht negativ beeinträchtigen. Dies umfasst die Lebensweise von Individuen und sozialen Gruppen.[20]
Besonders schutzbedürftige Personen sollten größere Aufmerksamkeit erhalten und in die Entwicklung und Tests der KI-Systeme einbezogen werden. Vor allem in Situationen, in denen KI-Systeme aufgrund von Macht- oder Informationsasymmetrien – Arbeitgebende/Arbeitnehmende, Unternehmen/Verbrauchende oder Regierung/Bevölkerung – negative Auswirkungen verursachen oder verschärfen können, müssen der Anwendungsfall und das Modell sorgfältig analysiert werden; der Grad der Robustheit muss entsprechend hoch sein.

KI - Sicherheitsmaßnahmen

Sicherheitsmaßnahmen für KI-Systeme und die Kombination von KI-Funktionalitäten, wie sie für bestehende IT-Systeme bekannt sind, sollten in einem so genannten Meta-System eingerichtet werden: Die Sicherheitsmaßnahmen sind damit den KI-Funktionalitäten überlegen, was bedeutet, dass die Sicherheitsmaßnahmen ausgeführt werden, auch wenn eine KI-Funktionalität dabei nicht mehr oder nur noch teilweise funktional ist. Das Sicherheitssystem muss laufend ausgeführt und verbessert werden.

Robust KI

Robuste-KI-Lösungen (auch englisch: *robust AI*) sind Lösungen, die gegenüber Einflüssen unempfindlich sind und in unsicheren und unvorhersehbaren Umgebungen die erwartete Leistung erbringen, ohne menschliche Vorurteile zu reproduzieren. Bei der Entwicklung von KI-Systemen müssen diese Fähigkeiten durchdacht konstruiert und implementiert werden. Die Robustheit eines KI-Systems hängt hauptsächlich von drei Faktoren ab: 1. der Robustheit des (Geschäfts-) Prozesses, in den das System integriert ist, 2. der Struktur und Art des verwendeten KI-Modells und 3. der Datenbank, die die Grundlage für das Training des Modells bildet. KI-basierte Systeme sind

20 Europäische Kommission, Ethics Guidelines for Trustworthy AI, 2019

for decisions with high risk potential; it must be possible to understand how a decision is made. It is especially important to check whether a decision is resilient or based on a misunderstanding of AI in order to fulfill our responsibility to act fairly and ethically. To learn how to make AI more robust and assess the robustness of the AI developed, please see the whitepaper on "Robust AI Assessment".

2 DIN norms (Status November 2020) currently define measures for robust AI:

- DIN SPEC 92001–1: Artificial Intelligence - Life Cycle Processes and Quality Requirements – Part 1: Quality Meta Model;
- DIN SPEC 92001–2: Artificial Intelligence –Life Cycle Processes and Quality Requirements – Part 2: Robustness

Data safety

AI systems and features must be updated continuously to meet safety standards.
Every collected data shall be categorized according to the data privacy classes (Datenschutzklasse[21]) provided by Group Privacy. For privacy-relevant data, a high level of security encryption is envisioned if suitable.

External threat

At DT, all software systems should be protected against vulnerabilities that can be exploited by hackers.

Mathematically speaking, all deep neural networks are trained to optimize their behavior in relation to a specific task, such as language translation or image classification. During training, this required behavior is usually expressed as an optimization problem that minimizes a loss value according to a specific formula that quantifies the differences from the required behavior. Attacks can produce inputs that have the opposite effect (adversarial inputs). They maximize the loss value and consequently maximize the differences from the desired behavior. This requires either knowledge of the inner workings of the deep neural network or the attacker estimates the behavior of the model and designs adversarial examples based on this estimate.

Therefore, not only should the model itself be protected from public access, but repeated queries (reverse engineering) should also be prevented from reproducing the functionality.

An attack can be conducted via malicious data designed to reveal information about the training process, it can target the infrastructure as well as hardware and software. This can end in the following scenarios: The systems make different decisions than intended, the data is corrupted, which, in turn, results in wrong decisions and in the worst case harm to people.

21 see https://yam.telekom.de/docs/DOC-556657

dafür bekannt, genaue und qualitativ hochwertige Vorhersagen zu liefern, insbesondere wenn eine gute Datenbank zur Verfügung steht. Je komplexer die Datenmodelle, die eine KI darstellt, desto schwieriger wird es jedoch zu verstehen, warum das System ein bestimmtes Ergebnis wählt und ob die Entscheidung tatsächlich gerechtfertigt ist. Daher ist eine robuste KI eine Voraussetzung für eine vertrauenswürdige KI. Dies ist insbesondere vor dem Hintergrund notwendig, dass die KI-Technologie zunehmend für Entscheidungen mit hohem Risikopotenzial eingesetzt wird; es muss nachvollziehbar sein, wie eine Entscheidung getroffen wird. Es ist besonders wichtig zu prüfen, ob eine Entscheidung belastbar ist oder ob sie auf einem Missverständnis der künstlichen Intelligenz beruht. Um zu erfahren, wie man die KI robuster machen und die Robustheit der entwickelten KI beurteilen kann, lesen Sie bitte das Dokument *Whitepaper zum Robust AI Assessment*.

Zwei DIN-Normen (Stand Nov. 2020) definieren derzeit Maßnahmen für eine robuste AI:

- DIN SPEC 92001–1: Artificial Intelligence - Life Cycle Processes and Quality Requirements – Part 1: Quality Meta Model;
- DIN SPEC 92001–2: Artificial Intelligence –Life Cycle Processes and Quality Requirements – Part 2: Robustness

Datensicherheit

KI-Systeme und Funktionen müssen ständig aktualisiert werden, um den Sicherheitsstandards zu entsprechen.
Alle gesammelten Daten sind nach den von Group Privacy bereitgestellten Datenschutzklassen[21] zu kategorisieren. Für datenschutzrelevante Daten ist teilweise ein sehr hohes Sicherheitsniveau bei der Verschlüsselung vorgesehen.

Externe Bedrohung

Bei der Deutschen Telekom sollten alle Softwaresysteme vor Schwachstellen geschützt werden, die von Hackern ausgenutzt werden könnten.

Mathematisch gesehen werden alle tiefen neuronalen Netze darauf trainiert, ihr Verhalten in Bezug auf eine bestimmte Aufgabe, z. B. Sprachübersetzung oder Bildklassifikation, zu optimieren. Während des Trainings wird dieses erforderliche Verhalten normalerweise als Optimierungsproblem ausgedrückt, das einen Verlustwert gemäß einer bestimmten Formel, die die Unterschiede zum erforderlichen Verhalten quantifiziert, auf ein Minimum reduziert. Angriffe können Eingaben erzeugen, die den gegenteiligen Effekt haben (kontradiktorische Eingaben). Sie maximieren den Verlustwert und damit die Unterschiede zum gewünschten Verhalten. Dazu ist entweder die Kenntnis der inneren Funktionsweise des tiefen neuronalen Netzes erforderlich oder der Angreifer schätzt das Verhalten des Modells und entwirft gegnerische Beispiele auf der Grundlage dieser Schätzung.

Daher sollte nicht nur das Modell selbst vor öffentlichem Zugang geschützt sein, sondern es sollte auch verhindert werden, dass durch wiederholte Abfragen (Reverse Engineering) die Funktionalität reproduziert wird.

21 siehe https://yam.telekom.de/docs/DOC-556657

The following methods can be used to make attacks more difficult or prevent them:

Avoiding reverse engineering: For this purpose, the frequency of the queries can be limited. If the frequency is chosen correctly, the use of the AI model is not affected – but reverse engineering is no longer possible in a realistic time horizon. To avoid affecting usage, it would be conceivable to gradually increase the delay only after a certain number of requests have been made at high speed. Another way to make reverse engineering more difficult is to increase the complexity of the AI model. This can be achieved, for example, by randomly switching between different models trained for the same task.

Adversarial training[22]: The basic idea of adversarial training is to include adversarial inputs in the training process. However, these are correctly classified so that the application of the trained model allows for correct classification of other adversarial examples. A new method of this training is fast adversarial training.

SafetyNet[23]: SafetyNet consists of the original model (classifier) and a detector that examines the internal state of the activation layers of the original model. If the detector detects that a sample is hostile, the hostile data set (sample) is rejected.

Emergency mode

Furthermore, the possibility of shutting down the system or removing dangerous parts should be provided. For example, since AI systems are often distributed across network nodes, a mechanism such as apoptosis can be applied that ensures a maximum lifetime for network nodes and IT processes, to name just one possibility.

22 Li et al., Towards understanding fast adversarial Training; 2020; https://arxiv.org/abs/2006.03089

23 Lu et al., Safety Net; 2017

Ein Angriff kann über Schaddaten erfolgen, die Informationen zum Trainingsprozess offenbaren sollen, er kann die Infrastruktur wie auch Hard- und Software zum Ziel haben. Dies kann in folgenden Szenarien enden: Die Systeme treffen andere Entscheidungen als beabsichtigt, die Daten werden verfälscht, woraus wiederum Fehlentscheidungen und im schlimmsten Fall Schaden für Menschen entsteht.

Um Angriffe zu erschweren oder zu verhindern, können folgende Methoden eingesetzt werden:

Vermeidung von Reverse Engineering: Dazu kann die Häufigkeit der Anfragen begrenzt werden. Wird die Häufigkeit richtig gewählt, wird die Verwendung des KI-Modells nicht beeinträchtigt – aber Reverse Engineering ist in einem realistischen Zeithorizont nicht mehr möglich. Um die Nutzung nicht zu beinträchtigen, wäre es denkbar, die Verzögerung erst nach einer bestimmten Anzahl von Anfragen mit hoher Geschwindigkeit schrittweise zu erhöhen. Eine weitere Möglichkeit, das Reverse Engineering zu erschweren, besteht darin, die Komplexität des KI-Modells zu erhöhen. Dies kann z. B. durch zufälliges Umschalten zwischen verschiedenen Modellen, die für die gleiche Aufgabe trainiert wurden, erreicht werden.

Kontradiktorisches Training[22]: Grundidee des kontradiktorischen Trainings ist es, kontradiktorische Eingaben in den Trainingsprozess einzubeziehen. Diese werden aber korrekt klassifiziert, so dass die Anwendung des trainierten Modells eine korrekte Klassifizierung anderer feindseliger Beispiele ermöglicht. Eine neue Methode dieses Trainings ist das schnelle kontradiktorische Training (fast adversarial training).

SafetyNet[23]: SafetyNet besteht aus dem ursprünglichen Modell (Klassifikator) und einem Detektor, der den internen Zustand der Aktivierungsschichten des ursprünglichen Modells untersucht. Wenn der Detektor feststellt, dass eine Probe schädlich ist, wird der schädliche Datensatz (Probe) zurückgewiesen.

Notfallmodus (sog. Not-Aus-Schalter)

Darüber hinaus sollte die Möglichkeit vorgesehen werden, das System abzuschalten oder gefährliche Teile zu entfernen. Da KI-Systeme oft über Netzwerkknoten verteilt sind, kann beispielsweise ein Mechanismus wie die Apoptose angewendet werden, der eine maximale Lebensdauer von Netzwerkknoten und IT-Prozessen setzt, um nur eine Möglichkeit zu nennen.

22 Li et al., Towards understanding fast adversarial Training; 2020; https://arxiv.org/abs/2006.03089

23 Lu et al., Safety Net; 2017

06

We set the grounds

06

Wir legen das Fundament

06
We set the grounds

Transparent, comprehensible and fully documented AI systems – Business impact assessment – Contingency scenarios – Explainable AI vs. black box

Thorough analyses and evaluations are the basis for the development and further improvement of our AI systems. These are transparent, comprehensible and fully documented. In this way, we consciously lay the foundation for AI development that leads to the best possible solution.

The starting point for our analyses of the effects of an AI system is the principle of privacy and security by design. This includes, among other things, opportunity and risk considerations as well as the investigation of emergency scenarios. We take great care in the development of the initial algorithm of our own AI solutions to prevent so-called black boxes and to ensure that our systems do not harm users, even by mistake.

Target picture

DT's AI systems are fully documented and transparent. In particular the adaptation of the system to the environment is transparent, i.e. when developing and implementing algorithms for the AI system, an automatic computation trace could be implemented, for example, which can be represented as a (graphical) network. The computational trace should reveal inferences about decisions and how they relate to each other. The ability to shut down the system and remove dangerous/harmful parts is also inherent to all systems that use AI.

To-dos

Prior to development
- Product Owner assesses the business impact.
- Product Owner develops emergency processes.

During development
- Team takes measures to achieve an highly explainable AI.

After launch
- Product Owner assesses new technology for explainable AI and implements new measures if necessary.

06
Wir legen das Fundament

Transparente, nachvollziehbare und vollumfänglich dokumentierte KI-Systeme – Bewertung des Business Impact – Notfallszenarien – Explainable AI vs. Black Box

Gründliche Analysen und Evaluierungen sind die Basis für die Entwicklung und weitere Verbesserung unserer KI-Systeme. Diese sind transparent, nachvollziehbar und vollumfänglich dokumentiert. Damit legen wir bewusst das Fundament für eine KI-Entwicklung, die zur bestmöglichen Lösung führt.

*Ausgangspunkt unserer Analysen zu Auswirkungen eines KI-Systems ist das Prinzip privacy and security by design. Dazu gehören u.a. Chancen- und Risiko-Betrachtungen sowie die Untersuchung von Notfallszenarien. Wir achten sehr sorgfältig auf die Entwicklung des initialen Algorithmus unserer eigenen KI-Lösungen, um sogenannte Black Boxes zu verhindern und sicherzugehen, dass unsere Systeme die Benutzer*innen nicht – auch nicht irrtümlicherweise – schädigen.*

Zielbild

Die KI-Systeme der DT sind vollständig dokumentiert und transparent. Insbesondere die Anpassung des Systems an die Umgebung ist transparent, d. h. bei der Entwicklung und Implementierung von Algorithmen für das KI-System könnte z. B. eine automatische Berechnungsspur implementiert werden, die als (grafisches) Netzwerk dargestellt werden kann. Die Berechnungsspur soll Rückschlüsse zu Entscheidungen und deren Beziehung zueinander zeigen. Auch die Möglichkeit, das System abzuschalten und gefährliche/schädliche Teile zu entfernen, gehört zu allen Systemen, die KI verwenden.

To-dos

Vor der Entwicklung
- Product Owner schätzt den Business Impact ab.
- Product Owner erarbeitet Notfallprozesse.

Während der Entwicklung
- Team ergreift Maßnahmen, um die KI gut erklärbar zu machen.

Nach der Entwicklung
- Product Owner prüft neue Technologie auf Erklärbarkeit und führt ggf. neue Maßnahmen ein.

Transparent, comprehensible and fully documented AI systems

Transparency is not just the comprehensibility of a decision that the AI makes through various derivations (see section "Explainable AI vs. Black Box" in this chapter), but also the transparency of communicating to the customer that he/she is interacting with an AI (see chapter "04 – We are transparent").

Many aspects of AI – how it works, the training processes and codes used, and the security measures taken – cannot always be clearly understood after a certain period of time. This is due to the further development of the technology, changes in team structures and responsibilities, and the increasing number of use cases. Detailed documentation is therefore essential (see "01 – We are responsible").

Assessment of the business impact

In order to control impacts such as breaches of security, privacy or ethical goals, the identification and assessment of business impact is mandatory. This should be done very early on in the development cycle, as subsequent changes often involve a great deal of effort. It is advisable to establish risk management throughout the product lifecycle. It is important to avoid having no worst-case measures available for certain scenarios when they are suddenly needed. The "Structure for risk analysis" in the appendix should serve as a guide for this purpose.[24] The examples presented there briefly introduce some of the more ethically explosive risks that you should consider. In addition, there are numerous other risks (currency risk, commercial risk, etc.) that are not covered in these professional ethics. We speak of fundamental impacts with high risk when one or more worst-case scenarios could lead to damage to life and limb, damage to the environment, financial losses in the millions, or damage to reputation of a particularly serious nature (especially in the case of issues that are essential for DT, such as data privacy and integrity).

How to proceed

1. **Identify:** With the "structure for risk analysis", a detailed risk assessment can be performed. For this assessment, a diverse team is needed, as this helps to avoid blind spots. In order to get a holistic view, it might help to get in contact with experienced colleagues (see "09 – We share and enlighten").

2. **Analyze and measure:** The next step is to quantify the impact in terms of scale of damage and probability of occurrence.

24 C. Bartneck et al., An Introduction to Ethics in Robotics and AI, SpringerBriefs in Ethics; Springer, 2020

Transparente, nachvollziehbare und vollumfänglich dokumentierte KI-Systeme

Unter Transparenz versteht man nicht nur die Nachvollziehbarkeit einer Entscheidung, welche die KI durch verschiedene Herleitungen trifft (s. den Abschnitt „Explainable AI vs. Black Box“ in diesem Kapitel), sondern auch die Transparenz, Nutzer*innen zu kommunizieren, dass sie mit einer KI interagieren (siehe *04 – Wir stehen für Transparenz*).

Viele Aspekte einer KI – Funktionsweise, verwendete Trainingsprozesse und Codes sowie ergriffene Sicherheitsmaßnahmen – lassen sich nach einer gewissen Zeit nicht immer eindeutig nachvollziehen. Das liegt an der Weiterentwicklung der Technologie, an Veränderungen in Teamsstrukturen und Verantwortlichkeiten, sowie an der steigenden Zahl von Anwendungsfällen. Eine ausführliche Dokumentation ist daher unerlässlich (siehe *01 – Wir übernehmen Verantwortung*).

Bewertung des Business Impact

Um Auswirkungen wie Verstöße gegen die Sicherheit, den Schutz der Privatsphäre oder ethische Ziele zu kontrollieren, ist die Ermittlung und Bewertung des Business Impact obligatorisch. Dies sollte sehr früh im Entwicklungszyklus geschehen, da spätere Änderungen oft mit hohem Aufwand verbunden sind. Es ist ratsam, ein Risikomanagement über den gesamten Produktlebenszyklus zu etablieren. Es ist zu vermeiden, dass für bestimmte Szenarien keine Worst-Case-Maßnahmen zur Verfügung stehen, wenn sie plötzlich benötigt werden. Die „Struktur zur Risikoanalyse“ im Anhang soll Ihnen dazu als Orientierungshilfe dienen.[24] Die dort vorgestellten Beispiele stellen kurz einige der ethisch brisanteren Risiken vor, die Sie berücksichtigen sollten. Daneben gibt es zahlreiche andere Risiken (Währungsrisiko, kommerzielles Risiko usw.), die in dieser Professionsethik nicht behandelt werden. Wir sprechen von grundsätzlichen Auswirkungen mit hohem Risiko, wenn ein oder mehrere Worst-Case-Szenarien zu Schäden an Leben und Gesundheit, zu Umweltschäden, zu finanziellen Verlusten in Millionenhöhe oder zu Reputationsschäden besonders schwerwiegender Art führen können (insbesondere bei Themen, die für die Deutsche Telekom von wesentlicher Bedeutung sind, wie Datenschutz und Integrität).

Wie Sie vorgehen können

1. **Identifizieren:** Mit der „Struktur zur Risikoanalyse“ kann eine detaillierte Risikobewertung durchgeführt werden. Für diese Bewertung ist ein diversifiziertes Team erforderlich, denn dies hilft, blinde Flecken zu vermeiden. Um eine ganzheitliche Sichtweise zu erhalten, kann es hilfreich sein, mit erfahrenen Kolleginnen und Kollegen in Kontakt zu treten (siehe *09 – Wir teilen und erklären*).

2. **Analysieren und messen:** Der nächste Schritt ist die Quantifizierung der Auswirkungen in Bezug auf Schadensausmaß und Eintrittswahrscheinlichkeit.

24 Nach C. Bartneck et al., An Introduction to Ethics in Robotics and AI, SpringerBriefs in Ethics; Springer, 2020

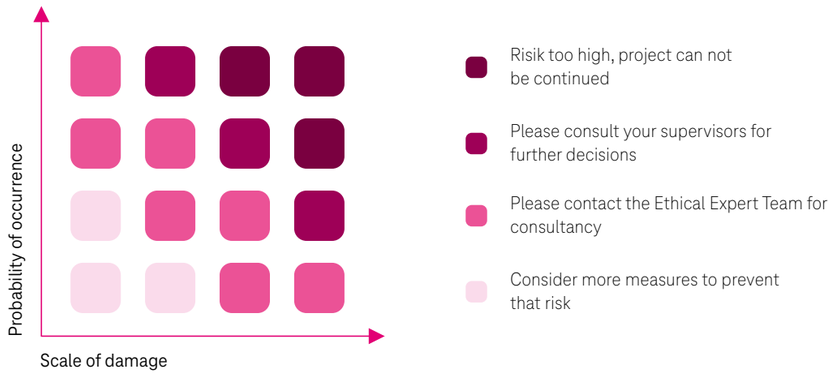
Probability of occurrence

4 = High probability of occurrence	Expected to occur within the next 3 months
3 = Medium probability of occurrence	Expected to occur within 3 years
2 = Low probability of occurrence	Expected to occur within 10 years
1 = Very unlikely	No risk has yet occurred, even at comparable companies. However, risk also cannot be excluded.

Scale of damage

5 = Worst case	The existence/reputation of the company would be endangered in case of the occurrence.
4 = Major impact	The occurrence of the risk forces the company to change its goals or strategy in the short-term. Example: A failure leads to strong impairments of all customers (e.g. through network instability).
3 = Medium impact	The occurrence of the risk necessitates medium-term changes in the company's goals or strategy. Example: An error leads to reputational damage which is publicized in leading media.
2 = Minor impact	The occurrence of the risk forces the change of ways and methods. Example: A system error incorrectly initiates ordering processes for individual customers.
1 = Trivial	No effect on the company value or reputation. Example: Internally processes cannot be carried out at the usual speed.

Position the analyzed impacts in the following matrix:



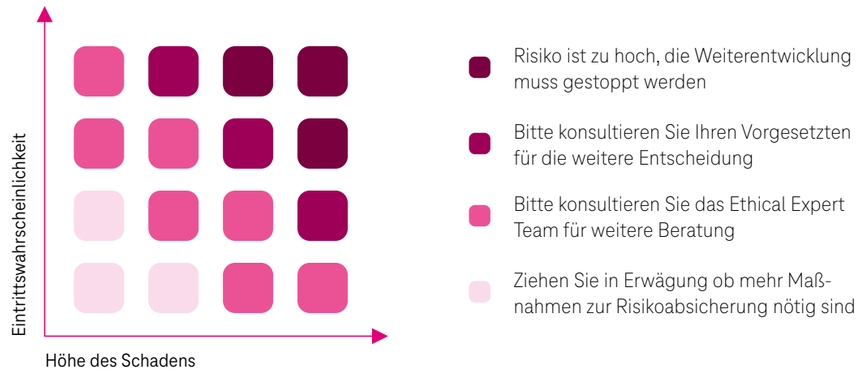
Eintrittswahrscheinlichkeit

4 = Sehr wahrscheinlich	Erwartet innerhalb der nächsten 3 Monate
3 = Wahrscheinlich	Erwartet innerhalb der nächsten 3 Jahre
2 = Unwahrscheinlich	Erwartet innerhalb der nächsten 10 Jahre
1 = Sehr unwahrscheinlich	Auch bei vergleichbaren Unternehmen ist bisher kein Risiko dieser Art aufgetreten. Ein Risiko kann aber nicht ausgeschlossen werden.

Höhe des Schadens

5 = Im schlimmsten Fall	Die Existenz oder der Ruf des Unternehmens sind im Falle des Auftretens gefährdet.
4 = Hoher Einfluss	Das Auftreten des Risikos zwingt das Unternehmen, seine Ziele oder Strategie kurzfristig zu ändern. Beispiel: Ein Ausfall führt zu starken Beeinträchtigungen aller Kund*innen (z. B durch Netzinstabilität).
3 = Mittlerer Einfluss	Das Auftreten des Risikos erfordert mittelfristige Änderungen der Ziele oder der Strategie des Unternehmens. Beispiel: Der Fehler führt zu einem Reputationsschaden, welcher von führenden Medien aufgegriffen wird.
2 = Kleiner Einfluss	Das Auftreten des Risikos zwingt dazu, Wege und Methoden zu ändern. Beispiel: Ein Systemfehler leitet fälschlicherweise Bestellprozesse für einzelne Kund*innen ein.
1 = Trivial	Keine Auswirkung auf Ruf oder Wert des Unternehmens. Beispiel: Interne Prozesse können nicht in gewohnter Geschwindigkeit erfolgen.

Positionieren Sie die analysierten Auswirkungen in der folgenden Matrix:



3. Gather the results according to the scenarios you have prepared and share them with supervisors as needed. If difficult ethical issues arise, please contact the digital ethics experts.[25]

4. Define risk management according to the matrix. For example, develop a step by step plan for the correct communication in the case that an incident happens in the green area.
Develop an emergency mode for the case that an incident happens in the yellow area and make sure that those involved have been instructed accordingly. These measures must be adapted depending on the scenarios and use cases.

Emergency scenarios

As described above, the business impact and risks should be analyzed at the very beginning of the project. This should identify critical situations. These are then generalized and serve as a basis for an emergency concept. The resources required for this are then determined: employees with know-how and capacity, IT applications, possibly buildings or workplaces, as well as information/documents or resources from service providers. Subsequently, time and capacity requirements for eventual recovery, restart and also damage reduction are recorded. This standard procedure can be found in ISO 22301:2012.

Explainable AI vs. Black Box

For a system to be trustworthy, operators and developers must be able to understand why it behaved in a certain way in a given situation and on the basis of which interpretation this was done. Explainability of neural networks in particular is a major challenge, but a critical factor in detecting and avoiding errors and undesirable behavior. Explainable AI (XAI) makes it easier for developers and users to clearly explain the decision-making process of AI systems.

The black box systems, on the other hand, do not provide this transparency, but often achieve more accurate results than more transparent methods. Therefore, a balance must be found between the accuracy of the results and their traceability.[26] In some cases, the benefit for the user may be significantly higher due to better results of the AI system than due to improved traceability. An example here could be the recognition of body movements in games. Therefore, please evaluate the degree of traceability specifically for your model.
From an ethical perspective, the traceability of decision-making processes plays an important role, as it is an essential aspect for assigning responsibility, which is described in the chapter "We are responsible". Transparency of the actions of intelligent systems increases trust in these systems.
There are several approaches to evaluating the results of AI systems that increase the degree of traceability:[27]

25 Digital Ethics Experts: Digital-Ethics@Telekom.de
26 Gleicher, M, A Framework for Considering Comprehensibility in Modeling. Big Data, 75–88
27 Bitkom, Blick in die Blackbox, 2019

3. Sammeln Sie die Ergebnisse gemäß den von Ihnen vorbereiteten Szenarien und teilen Sie sie bei Bedarf den Vorgesetzten mit. Wenn schwierige ethische Fragen auftreten, wenden Sie sich bitte an die Expert*innen für digitale Ethik.[25]

4. Definieren eines Risikomanagements entsprechend der Matrix. Entwickeln Sie z. B. einen Stufenplan für die korrekte Kommunikation für den Fall, dass ein Vorfall im grünen Bereich passiert. Entwickeln Sie einen Notfallmodus, für den Fall, dass ein Vorfall im gelben Bereich passiert, und stellen Sie sicher, dass die beteiligten Personen entsprechend unterwiesen werden. Diese Maßnahmen müssen je nach Szenario und Anwendungsfall angepasst werden.

Notfallszenarien

Wie oben beschrieben, sollten schon zu Anfang des Projekts der Business Impact und Risiken analysiert und kritische Situationen herauskristallisiert werden. Diese werden im Anschluss generalisiert und dienen als Basis für ein Notfallkonzept. Für dieses werden die erforderlichen Ressourcen identifiziert: Mitarbeiterinnen und Mitarbeiter mit Know-How und Kapazität, IT-Anwendungen, ggf. Gebäude, Arbeitsplätze sowie Informationen/Dokumente oder Ressourcen von Dienstleistern. Dann werden zeitliche und kapazitative Anforderungen an eine eventuelle Wiederherstellung, einen Neustarts und auch die Schadensreduzierung festgehalten. Dieses Standard-Vorgehen kann in ISO 22301:2012 nachgelesen werden.

Explainable AI vs. Black Box

Damit ein System vertrauenswürdig ist, müssen die Menschen, die es betreiben und die, die es entwickeln, in der Lage sein zu verstehen, warum es sich in einer gegebenen Situation auf eine bestimmte Art und Weise verhielt und auf Basis welcher Interpretation dies geschah. Insbesondere die Erklärbarkeit von neuronalen Netzen ist eine große Herausforderung, aber ein kritischer Faktor, um Fehler und unerwünschtes Verhalten zu erkennen und zu vermeiden. Eine Explainable AI (XAI) macht es Entwickler*innen und Anwender*innen einfacher, den Entscheidungsprozess von KI-Systemen klar zu erklären.

Die Black-Box-Systeme auf der anderen Seite leisten diese Transparenz nicht, aber erzielen oftmals genauere Ergebnisse als transparentere Methoden. Daher muss ein Gleichgewicht zwischen der Genauigkeit der Ergebnisse und ihrer Nachvollziehbarkeit gefunden werden.[26] In einigen Fällen kann der Vorteil für Nutzerinnen und Nutzer aufgrund besserer Ergebnisse des KI-Systems deutlich höher sein, als durch verbesserte Nachvollziehbarkeit. Ein Beispiel wäre die Erkennung von Körperbewegungen in Spielen. Bitte bewerten Sie daher den Grad der Nachvollziehbarkeit speziell für Ihr Modell.
Aus ethischer Sicht spielt die Rückverfolgbarkeit der Entscheidungsprozesse eine wichtige Rolle, da sie ein wesentlicher Aspekt für die Zuweisung von Verantwortung ist, die in *01 – Wir übernehmen Verantwortung* beschrieben wird. Die Transparenz der Aktionen intelligenter Systeme erhöht das Vertrauen in diese Systeme. Es gibt verschiedene Ansätze zur Bewertung der Ergebnisse von KI-Systemen, die den Grad der Nachvollziehbarkeit erhöhen:[27]

25 Digital-Ethics@Telekom.de
26 Gleicher, M, A Framework for Considering Comprehensibility in Modeling. Big Data, 75–88
27 Bitkom, Blick in die Blackbox, 2019

1. Description of the result-finding process: With this type of traceability, the AI system provides an explanatory model for local data points in the process of results analysis. Explanatory models can be implemented, for example, with the help of Local Interpretable Model-agnostic Explanations (LIME).

2. LIME[28]: LIME is a frequently used explanatory model for ML models, which was developed to locally explain even complex and strongly nonlinear models (phyton model exists[29]). LIME generates an explanatory pattern for the individual predictions of a model in several steps. First, random data points are generated around the data point to be explained or selected from the training data. Predictions are generated for these data points using the existing ML model. These data points, along with the predictions, are then used to train an explainable model to locally explain the predictions of the actual ML model. This, in turn, can be used to interpret the decision-making process of the actual model.

3. Input-Output Validation[30]: This method is used to examine which parts of the input are of particular importance in obtaining a particular output. Methods such as SHapley Additive exPlanations (SHAP) can be used to evaluate individual data: Just like LIME, SHAP can be applied to generic, table-based data and also provides specific implementations for data formats, such as image data. Similar to LIME, SHAP produces an interpretable local model of an ML procedure that can then be used to explain predictions of the model. The advantage of this method is that explanations are more robust in many cases and do not depend on the user‘s choice of parameters. SHAP also exists as a Python library.

4. Counterface explanations: Starting from a specified data point, a different data point is sought that would significantly change the decision of the model while being as close as possible to the original data point. What exactly represents a significant change depends on the model type: In classification models, for example, a change in the predicted class for the data point would be considered a significant change; in regression models, for example, a certain change in the predicted value might be considered.

5. Accumulated local effects[31]: This method generates more realistic data points than the method of counterface explanations and considers only local differences instead of global dependencies. The main attribute of this method is that only data points are used which are close to the real data, thus providing a more realistic reflection of dependencies between individual attributes.

28 Algoneer – LIME Beispiel – Jupyter Notebook

29 LIME Python-Bibliothek: https://github.com/marcotcr/lime

30 Wachter, Sandra, Brent Mittelstadt, and Chris Russell. »Counterfactual explanations without opening the black box: Automated decisions and the GDPR.« (2017).

31 Zhao, Qingyuan, and Trevor Hastie. »Causal interpretations of black-box models.« Journal of Business & Economic Statistics, to appear. (2017).

1. Darstellung des Ergebnisfindungsprozesses: Mit dieser Art der Rückverfolgbarkeit stellt das KI-System ein Erklärungsmodell für lokale Datenpunkte im Prozess der Ergebnisermittlung zur Verfügung. Erklärungsmodelle können z. B. mit Hilfe von Local Interpretable Modelagnostic Explanations (LIME) implementiert werden.

2. LIME[28]: Lime ist eine häufig verwendete Erklärungsmethode für ML-Modelle, die entwickelt wurde, um auch komplexe und stark nichtlineare Modelle lokal zu erklären (Phyton-Modell vorhanden[29]). LIME erzeugt in mehreren Schritten ein Erklärungsmuster für die einzelnen Vorhersagen eines Modells. Zuerst werden zufällige Datenpunkte um den zu erklärenden Datenpunkt herum generiert oder aus den Trainingsdaten ausgewählt. Für diese Datenpunkte werden Vorhersagen mit dem bestehenden ML-Modell generiert. Diese Datenpunkte werden dann zusammen mit den Vorhersagen dazu verwendet, ein erklärbares Modell zu trainieren, das die Vorhersagen des eigentlichen ML-Modells lokal erklären soll. Dieses kann wiederum dazu verwendet werden, den Entscheidungsfindungsprozess des tatsächlichen Modells zu interpretieren.

3. Input-Output-Validierung[30]: Mit dieser Methode wird untersucht, welche Teile des Inputs für die Gewinnung eines bestimmten Outputs von besonderer Bedeutung sind. Methoden wie SHapley Additive exPlanations (SHAP) können zur Auswertung der einzelnen Daten genutzt werden: Genau wie LIME kann SHAP auf generische, tabellenbasierte Daten angewendet werden und bietet auch spezifische Implementierungen für Datenformate, wie z. B. Bilddaten. Ähnlich wie LIME erzeugt SHAP ein interpretierbares, lokales Modell einer ML-Prozedur, das dann zur Erklärung von Vorhersagen des Modells verwendet werden kann. Der Vorteil dieser Methode besteht darin, dass Erklärungen in vielen Fällen robuster sind und nicht von der Parameterauswahl des*der Nutzer*in abhängen. SHAP existiert auch als Python-Bibliothek

4. Kontrafaktische (Counterface) Erklärungen: Ausgehend von einem spezifizierten Datenpunkt wird ein anderer Datenpunkt gesucht, der die Entscheidung des Modells signifikant verändern würde, während er gleichzeitig so nah wie möglich am ursprünglichen Datenpunkt liegt. Es hängt vom Modelltyp ab, was genau eine signifikante Änderung darstellt: In Klassifikationsmodellen würde z. B. eine Änderung der vorhergesagten Klasse für den Datenpunkt als signifikante Änderung angesehen; in Regressionsmodellen kann eine bestimmte Änderung des vorhergesagten Wertes betrachtet werden.

5. Kumulierte lokale Einflüsse[31]: Diese Methode erzeugt realistischere Datenpunkte als die Methode der Counterface-Erklärung und berücksichtigt nur lokale Unterschiede anstelle von globalen Abhängigkeiten. Ihr Hauptmerkmal ist, dass nur Datenpunkte verwendet werden, die nahe an den realen Daten liegen und somit Abhängigkeiten zwischen einzelnen Attributen realistischer abbilden.

28 Algoneer – LIME Beispiel – Jupyter Notebook

29 LIME Python-Bibliothek, https://github.com/marcotcr/lime

30 Wachter, Sandra, Brent Mittelstadt, and Chris Russell, Counterfactual explanations without opening the black box: Automated decisions and the GDPR; 2017

31 Zhao, Qingyuan, and Trevor Hastie, Causal interpretations of black-box models, Journal of Business & Economic Statistics; to appear; 2017

07

We keep control

07

Wir behalten den Überblick

07

We keep control

Training Data – Bias – Testing and Validation – Quality Enhancement – Product Lifecycle Quality Assessment

We are able to stop or turn off our AI systems at any time ("kill switch"). In addition, we remove data that leads to a bias in the results. We pay attention to both the information that is fed into the system and the decisions that the artificial intelligence makes in order to enhance decision quality.

We assume responsibility for a suitable data basis. However, if discrepancies do occur, we stop the affected AI system rather than proceed with potentially damaging data. We are able to reset our AI systems and delete incorrect or biased data to minimize the number of (even unintentional) unsuitable decisions or actions.

To-dos

Prior to development
- Development Team assesses sources of training data.
- Development Team analyzes the training data in tests in order to prevent bias or non-realistic input.
- Product Owner puts together a diverse team and testing community.

During development
- Development Team implements quality standards and measures.
- Development Team implements a kill switch.
- Team tests early and often.
- Team tests in a realistic environment.

After launch
- Product Owner implements quality assessment over the entire product lifecycle.
- Product Owner assesses the methods and tools used and updates these if necessary.

07

Wir behalten den Überblick

Trainingsdaten – Voreingenommenheit/Bias – Prüfung und Validierung – Steigerung der Qualität – Bewertung der Qualität im Produktlebenszyklus

Wir sind dazu in der Lage, unsere KI-Systeme jederzeit anzuhalten oder abzuschalten („Not-Aus-Schalter“). Zudem entfernen wir Daten, die zu einer Verzerrung der Ergebnisse führen. Wir achten sowohl auf die Informationen, die in das System eingespeist werden, als auch auf die Entscheidungen, die die künstliche Intelligenz trifft, um die Qualität zu verbessern.

Wir übernehmen die Verantwortung für eine geeignete Datengrundlage. Sollten dennoch Unstimmigkeiten auftreten, halten wir das betroffene KI-System eher an, als dass wir mit potentiell schadhaften Daten weiterverfahren. Wir sind in der Lage, unsere KI-Systeme zurückzusetzen und falsche oder vorurteilsbehaftete Daten zu löschen, um die Anzahl (auch unbeabsichtigter) untauglicher Entscheidungen oder Handlungen auf ein Minimum zu reduzieren.

To-dos

Vor der Entwicklung
- Development Team wertet geeignete Quellen für Trainingsdaten aus.
- Development Team analysiert die Trainingsdaten in Tests, um Verzerrungen oder nicht realistische Eingaben zu vermeiden.
- Product Owner stellt eine diverse Team- und Testgemeinschaft zusammen.

Während der Entwicklung
- Development Team setzt Qualitätsstandards und Maßnahmen um.
- Development Team implementiert einen Not-Aus Schalter.
- Team testet früh und oft.
- Team testet in einer realitätsnahen Testumgebung.

Nach dem Launch
- Product Owner implementiert eine Qualitätsbewertung über den gesamten Produktlebenszyklus.
- Product Owner evaluiert die verwendeten Werkzeuge und Methoden und aktualisiert sie, falls erforderlich.

Training data

Training must be carried out in safe environments (a virtual testing area, "sandbox") to prevent possible failures affecting systems already in use.
If any risk occurs during the training or testing process, this must be reported to the person responsible, who must decide whether to continue or to adapt the system or, in the worst case, to shut down.

Proper selection of the source, amount, and variety of training data is of great importance for an ethical AI. To avoid the need to completely shut down an AI, it is recommended to include the ability to reset the AI to its initial state, thereby erasing all learned patterns and decisions. This chapter provides guidance on selecting the right training data and avoiding bias and prejudice.

High-quality training data not only provides a high level of accuracy and precision in decision-making, but is also critical for avoiding bias and unfair decisions. Use cases in which data is heavily influenced by trends, i.e. independent systematic changes such as weather, seasonal user behavior (e.g. vacation shopping), or other phenomena, are more susceptible to the inability to use previously learned patterns to make reliable decisions during the AI product lifecycle. Although AI models are trained to learn only generalized patterns during the learning process, training data can quickly change to not reflect all real-world influences, especially for more complex tasks and limited data. As trends are often not immediately recognized as such and training data are often limited to a narrow time frame, it can be the case that important information is not included in the training data. Especially in the case of cyclical trends, such as the influence of the seasons on outdoor camera images, it is very likely that important data has not been implemented in the training data. Another challenge is to correctly diagnose incorrect decisions, as errors occur periodically and then appear to be corrected. Conversely, continuous trends are more likely to cross a "threshold" after which the quality of the decision noticeably declines, which is usually easier to interpret. However, because the value ranges of both trends change significantly, there is generally the same risk of producing data that leads to incorrect decisions due to these natural changes. The amount of risk depends on the speed and extent to which the trends change. Therefore, it is important that your data set reflects the real-world environment in which the model will be used as closely as possible, from the characteristics of the data to their distribution across classes.

You can find indicators for improving the quality of your dataset below[32]:

32 Lionbridge.ai, training data guide, how can I improve the quality of my data, 2020, https://lionbridge.ai/training-data-guide/#how-can-i-improve-the-quality-of-my-data

Trainingsdaten

Das Training muss in sicheren Umgebungen (einem virtuellen Testgelände, „Sandbox“) durchgeführt werden. Mögliche Ausfälle dürfen bereits in Betrieb befindliche Systeme nicht infizieren.
Tritt während der Schulung oder des Testprozesses ein Risiko auf, muss es der verantwortlichen Person gemeldet werden. Diese Person muss dann entscheiden, ob sie das System weiter betreibt, anpasst oder abschaltet.

Die richtige Wahl der Quelle, der Menge und der Vielfalt der Trainingsdaten ist für eine ethische KI von großer Bedeutung. Um zu verhindern, dass eine KI vollständig abgeschaltet werden muss, wird empfohlen, die Möglichkeit einzubauen, die KI wieder in den Ausgangszustand zurückzusetzen und damit alle gelernten Muster und Entscheidungen zu löschen. In diesem Kapitel finden Sie eine Anleitung zur Auswahl der richtigen Trainingsdaten und zur Vermeidung von Bias und Vorurteilen.

Eine hohe Qualität der Trainingsdaten bietet nicht nur ein hohes Maß an Genauigkeit und Präzision bei der Entscheidungsfindung, sondern ist auch kritisch für die Vermeidung von Bias und unfairen Entscheidungen. Anwendungsfälle, in denen die Daten stark von Trends – d. h. unabhängigen systematischen Veränderungen wie Wetter, saisonalem Nutzerverhalten (z. B. Urlaubseinkäufe) oder anderen Phänomenen – beeinflusst werden, sind anfälliger dafür, dass zuvor gelernte Muster während des Produktlebenszyklus der KI nicht mehr für zuverlässige Entscheidungen genutzt werden können. Obwohl KI-Modelle während des Lernprozesses darauf trainiert werden, nur verallgemeinerte Muster zu lernen, können sich die Trainingsdaten schnell dahingehend ändern, dass sie nicht alle Einflüsse der realen Welt widerspiegeln, insbesondere bei komplexeren Aufgaben und begrenzter Datenmenge. Da Trends oft nicht sofort als solche erkannt werden und Trainingsdaten oft auf einen engen Zeitrahmen beschränkt sind, ist es möglich, dass wichtige Informationen nicht in den Trainingsdaten enthalten sind. Vor allem bei zyklischen Trends, wie z. B. dem Einfluss der Jahreszeiten auf die Bilder von Außenkameras, ist es sehr wahrscheinlich, dass wichtige Daten nicht in die Trainingsdaten implementiert wurden. Eine weitere Herausforderung ist, falsche Entscheidungen richtig zu diagnostizieren, da Fehler periodisch auftreten und dann scheinbar korrigiert werden. Umgekehrt wird bei kontinuierlichen Trends eher eine „Schwelle“ überschritten, nach der die Qualität der Entscheidung merklich abnimmt, was in der Regel leichter zu interpretieren ist. Da sich jedoch die Wertebereiche beider Trends erheblich ändern, besteht im Allgemeinen das gleiche Risiko, Daten zu erzeugen, die aufgrund dieser natürlichen Veränderungen zu Fehlentscheidungen führen. Die Höhe des Risikos hängt von der Geschwindigkeit und dem Ausmaß ab, in dem sich die Trends verändern. Daher ist es wichtig, dass Ihr Datensatz die reale Umgebung, in der das Modell verwendet werden soll, von den Eigenschaften der Daten bis hin zu ihrer Verteilung auf die Klassen möglichst genau widerspiegelt.

Nachfolgend finden Sie Indikatoren zur Verbesserung der Qualität Ihres Datensatzes:[32]

32 vgl. Lionbridge.ai, training data guide, how can I improve the quality of my data; 2020; https://lionbridge.ai/training-data-guide/#how-can-i-improve-the-quality-of-my-data

Desired characteristic	Definition	To-do
Uniformity	All data points were collected under the same framework conditions and come from comparable sources.	Check for irregularities when pulling data from multiple internal or external sources.
Consistency	All data points have the same origin.	Ensure that classes are distributed and separated clearly.
Comprehensiveness	The dataset has enough parameters to cover all of the model's possible use cases, including all edge cases.	Check that enough data is collected; include examples of edge cases.
Relevancy	The dataset contains only parameters which are useful to the model.	Identify important parameters; for complicated data, a data analyst can be brought in (if this has not already been done).
Diversity	The dataset accurately reflects the model's entire user base.	Conduct user analysis to uncover hidden biases; consider pulling data from both internal and external sources; also consider involving an expert.

Bias

Bias (see glossary) is automatically making assumptions based on mental models or memories. Bias can affect the following:

- How data is collected and classified.
- How machine learning systems are designed and developed.

For example, when building a classifier to identify wedding photos, an engineer may use the presence of a white dress in a photo as a feature. However, white dresses have been customary only during certain eras and in certain cultures.[33] Bias/distortions in systems can have serious consequences for DT's reputation, especially if attributes such as gender, race, skin color, language, religion, political or other opinion, national or social origin, membership of a national minority, assets or birth are not given sufficient consideration or even discriminated against.[34]

In general, it is impossible to guarantee complete fairness[35]. This topic is discussed in many dimensions, yet it is important to us that DT's AI is as fair as possible. Therefore, you should deepen your knowledge in this area and include the latest technologies and knowledge to avoid unfairness. In addition, integrity definitions should be defined to ensure certain standards of behavior within specified areas in all cases, e.g. a male customer should not be given preference over a female customer.

33 Glossary for Fairness, Google: https://developers.google.com/machine-learning/glossary/fairness#b
34 General Equal Treatment Act; Art. 8 G v. 3.4.2013 I 610
35 Friedler e.a.: On the (im)possibility of fairness; 2016, https://arxiv.org/pdf/1609.07236.pdf

Angestrebte Charakteristika	Definition	To-do
Einheitlich	Alle Datenpunkte wurden unter gleichen Rahmenbedingen erhoben und stammen aus vergleichbaren Quellen.	Prüfen Sie beim Abrufen von Daten aus mehreren (internen oder externen) Quellen diese auf Unregelmäßigkeiten.
Konsistent	Alle Datenpunkte haben den gleichen Ursprung.	Stellen Sie sicher, dass die Klassen klar verteilt und voneinander getrennt sind.
Ausführlich	Der Datensatz verfügt über genügend Parameter, um alle möglichen Anwendungsfälle des Modells abzudecken, einschließlich aller Randfälle (Edge Cases).	Prüfen Sie, ob Sie genügend Daten haben; fügen Sie Beispiele von Randfällen ein.
Relevant	Der Datensatz enthält nur Parameter, die für das Modell nützlich sind.	Identifizieren Sie wichtige Parameter; bei komplizierten Daten könnte ein Data Analyst hinzugezogen werden (falls nicht bereits geschehen).
Vielfältig	Der Datensatz spiegelt die gesamte Nutzerbasis des Modells genau wider.	Führen Sie eine Benutzeranalyse durch, um versteckte Verzerrungen aufzudecken; ziehen Sie in Betracht, Daten sowohl aus internen als auch aus externen Quellen zu beziehen; ziehen Sie zusätzlich in Betracht, eine*n Expert*in hinzuzuziehen.

Voreingenommenheit / Bias

Voreingenommenheit oder Bias (siehe auch Glossar) ist das automatische Treffen von Annahmen, die auf mentalen Modellen oder Erinnerungen beruhen. Voreingenommenheit kann folgendes beeinflussen:

- Wie Daten gesammelt und klassifiziert werden.
- Wie maschinelle Lernsysteme entworfen und entwickelt werden.

Beispielsweise beim Aufbau eines Klassifizierungsmodells zur Identifizierung von Hochzeitsfotos kann das Vorhandensein eines weißen Kleides auf einem Foto als Merkmal verwendet werden. Allerdings sind weiße Kleider nur zu bestimmten Epochen und in bestimmten Kulturen bei Hochzeiten gebräuchlich.[33] Bias/Verzerrungen in Systemen können weitreichende Folgen für die Reputation der Deutschen Telekom bedeuten, speziell wenn manche Attribute von Kategorien wie Geschlecht, Rasse, Hautfarbe, Sprache, Religion, politische oder sonstige Anschauung, nationale oder soziale Herkunft, Zugehörigkeit zu einer nationalen Minderheit, Vermögen, Geburt[34] nicht ausreichend berücksichtigt oder sogar benachteiligt behandelt werden.

Im Allgemeinen ist es unmöglich, vollständige Fairness zu garantieren[35]. Dieses Thema wird in vielen Dimensionen diskutiert, trotzdem ist es uns ein Anliegen, dass KI der Deutschen Telekom so fair wie möglich ist. Deshalb sollten Sie Ihr Wissen in diesem Bereich vertiefen und die neuesten Technologien und Kenntnisse einbeziehen, um Ungerechtigkeiten zu vermeiden. Zusätzlich sollten Integritätsdefinitionen definiert werden, um gewisse Verhaltensstandards innerhalb vorgegebener

33 Siehe auch Glossary for Fairness, Google: https://developers.google.com/machine-learning/glossary/fairness#b
34 Allgemeines Gleichbehandlungsgesetz; Art. 8 G v. 3.4.2013 I 610
35 Friedler et al., On the (im)possibility of fairness; 2016, https://arxiv.org/pdf/1609.07236.pdf

When does bias occur[36]

It is important to understand when bias occurs in a model. Biases usually occur unintentionally and unnoticed by the development team. In most cases, they occur in the base data rather than in the algorithm itself. For example, models may be trained with data that include human decisions or contain second-order effects of societal or historical inequalities. Biases may also arise from the way in which data are collected or selected for use. Likewise, user-generated data can create a feedback loop which, in turn, then leads to bias.

An AI could also pick up on statistical correlations that are socially unacceptable or even illegal. For example, if a mortgage lending model determines that older individuals are more likely to default and subsequently reduces lending based on age, society and legal institutions may view this as illegal age discrimination.

Tools and methods to avoid biases in the model

Bias in the training data can occur when the data collected does not reflect the conditions under which the model will operate. There are several methods and tools to quantify the level of bias in the training data. Especially in critical use cases it is recommended to consult an expert if a bias cannot be excluded.

Analyze the results during:

1. Pre-processing: To avoid a bias in the training data, the comprehensive processing of such is recommended to ensure that the data are as accurate as possible, while reducing any relationship between the results and socio-demographic characteristics (age, gender, race ...). Alternatively, these sensitive attributes can be omitted in the generated database. The idea is to always make the same decision in the counterfactual world where sensitive attributes (according to which, from an ethical point of view, a decision should not be made) are modified.[37]

2. Post-processing[38]: Here, some of the model's predictions can be transformed after they have been made to satisfy a particular fairness requirement.

The following measures are recommended to detect and correct biases if they occur:

1. Collect feedback from the user: Users should always have the possibility to give feedback to the project team if they detect any bias or have any other issue with the model or service. The contact possibility should be provided as an e-mail address or phone number, reference to a direct message service or any other channel that is easy to find.

36 McKinsey: Tackling bias in artificial intelligence; 2019
37 Matt J. Kusner, Joshua R. Loftus, Chris Russell, Ricardo Silva; Counterfactual fairness; 2017
38 Morith Hardt, Eric Price, Nathan Srebro; Equality of opportunity in supervised learning; 2016

Bereiche in jedem Fall zu gewährleisten, z. B. darf ein männlicher Kunde nicht gegenüber einer Kundin bevorzugt werden.

Wann tritt Verzerrung/Bias auf[36]

Es ist wichtig zu verstehen, wann eine Verzerrung in einem Modell auftritt. Verzerrungen treten in der Regel unbeabsichtigt und vom Entwicklungsteam unbemerkt auf. In den meisten Fällen kommen sie eher in den Basisdaten als im Algorithmus selbst vor. Modelle können z. B. mit Daten trainiert werden, die menschliche Entscheidungen enthalten oder Auswirkungen gesellschaftlicher oder historischer Ungleichheiten zweiter Ordnung enthalten. Verzerrungen können auch durch die Art und Weise, wie die Daten gesammelt oder zur Verwendung ausgewählt werden, entstehen. Ebenfalls können von Nutzer*innen erzeugte Daten einen Feedback-Loop erzeugen, der dann wiederum zu Verzerrungen führt.

Eine KI könnte auch statistische Korrelationen aufgreifen, die gesellschaftlich nicht akzeptabel oder sogar illegal sind. Wenn z. B. ein Hypotheken-Kredit-Modell feststellt, dass ältere Personen eine höhere Wahrscheinlichkeit von Zahlungsausfällen haben, und daraufhin die Kreditvergabe aufgrund des Alters reduziert, können Gesellschaft und rechtliche Institutionen dies als illegale Altersdiskriminierung ansehen.

Werkzeuge und Methoden zur Vermeidung von Verzerrungen im Modell

Verzerrungen in den Trainingsdaten können auftreten, wenn die von Ihnen gesammelten Daten nicht die Bedingungen widerspiegeln, unter denen Ihr Modell arbeiten wird. Es gibt einige Methoden und Werkzeuge, um den Grad der Verzerrung in den Trainingsdaten zu quantifizieren. Insbesondere in kritischen Anwendungsfällen, wenn eine Verzerrung nicht ausgeschlossen werden kann, empfiehlt es sich, Personen mit Expertenkompetenz zu Rate zu ziehen.

Analysieren Sie die Ergebnisse im:

1. Pre-Processing: Um eine Verzerrung der Trainingsdaten zu vermeiden, wird ihre umfassende Aufbereitung empfohlen, so dass die Daten so genau wie möglich sind und gleichzeitig jede Beziehung zwischen den Ergebnissen und soziodemografischen Merkmalen (Alter, Geschlecht, Rasse ...) reduziert wird. Alternativ kann in der erzeugten Datenbank auf diese sensiblen Attribute verzichtet werden. Die Idee besteht darin, in der kontrafaktischen Welt, in der sensible Attribute (nach denen aus ethischer Sicht eine Entscheidung nicht zu treffen ist) geändert werden, immer die gleiche Entscheidung zu treffen.[37]

2. Post-Processing[38]: Hierbei können einige der Prognosen des Modells umgewandelt werden, nachdem sie getroffen wurden, um einer bestimmten Fairnessanforderung zu genügen.

36 McKinsey, Tackling bias in artificial intelligence; 2019
37 Matt J. Kusner, Joshua R. Loftus, Chris Russell, Ricardo Silva, Counterfactual fairness; 2017
38 Morith Hardt, Eric Price, Nathan Srebro, Equality of opportunity in supervised learning; 2016

2. Implement fairness metric[39]: Implementation of a mathematical definition of "fairness" that is measurable. There are, for example, libraries with several algorithms to reduce software bias and increase its fairness. Commonly used metrics are equalized odds, predictive parity, counterfactual fairness or demographic parity.

3. Introduce a fairness constraint: Applying a constraint to the developed AI ensures that one or more previously defined definitions of fairness are always satisfied.

These fairness conditions can be applied as follows:

- For post-processing the model's output when checking for that constraint.
- Modification of the loss function for correction in the case of violation of a fairness metric. (For more on the loss function, see the "Improving quality" section later in the chapter.)
- Directly adding a mathematical constraint to an optimization problem in the model.

Further recommendations to prevent bias:

1. Install a feedback mechanism or open dialog with users to get feedback on user-identified biases or issues.
2. Consider a diverse team to help represent a wider variation of experiences to minimize bias. Embrace team members of different ages, ethnicities, genders, educational disciplines, and cultural perspectives.
3. Based on the type of data it ingests, the AI may be susceptible to different types of bias. Monitor training and results to quickly respond to issues. Test early and often.

Testing and validation:

To test properly, plan sufficient time for the testing period, as well as review, evaluation, troubleshooting and feedback loops. From a professional ethics perspective, this may look like this: For each learning task, identify scenarios that are safety or hazard-prone and/or occur frequently during deployment. For these scenarios, capture relevant data sets, simulation environments, and potential real-world test environments, and formulate an argument as to why the given data, simulations, and real-world test environments are representative of the critical scenarios. This will include consideration of potential limitations of the model. Additionally or alternatively, certain regions of the data set may need to be prioritized. This will give the critical scenarios a higher priority (regardless of whether they pose risks in terms of safety)and allow the development team to develop informed strategies for the model. Below are several methods for testing the model that should be considered and possibly run depending on the use case, as well as the testing period

39 Fairness Metrics: https://developers.google.com/machine-learning/glossary/fairness#fairness_metric

Falls Verzerrungen auftreten, werden folgende Maßnahmen zur Erkennung und Korrektur empfohlen:

1. Feedback vom Nutzer einholen: Nutzerinnen und Nutzer sollten immer eine Möglichkeit haben, dem Projektteam eine Rückmeldung zu geben, falls eine Voreingenommenheit festgestellt wurde oder ein anderes Problem mit dem Modell oder der Dienstleistung besteht. Diese Kontaktmöglichkeit sollte in Form einer E-Mail-Adresse oder Telefonnummer, dem Hinweis auf einen Direktnachrichten-Dienst oder über einen anderen leicht zu findenden Kanal zur Verfügung gestellt werden.

2. Implementierung von Fairness-Metriken[39]: Implementierung einer mathematischen Definition von „Fairness", die messbar ist. Es gibt z. B. Bibliotheken mit mehreren Algorithmen, um die Verzerrung von Software zu reduzieren und ihre Fairness zu erhöhen. Häufig verwendete Metriken sind ausgeglichene Zufallszahlen, prognostische Fairness, kontrafaktische Fairness oder demographische Gleichheit.

3. Eine Auflage zur Fairness einführen: Durch die Anwendung einer Randbedingung auf die entwickelte KI wird sichergestellt, dass eine oder mehrere Definitionen von Fairness, die Sie zuvor definiert haben, immer erfüllt werden.

Diese Fairnessbedingungen können folgendermaßen angewendet werden:

- Für die Nachbearbeitung des Outputs des Modells, bei der Überprüfung auf jene Randbedingung.
- Veränderung der Loss-Function für die Korrektur bei einer Verletzung einer Fairness-Metrik. (Zur Loss-Function siehe den Abschnitt „Steigerung der Qualität" weiter unten im Kapitel).
- Direktes Hinzufügen dieser Randbedingung als mathematische Nebenbedingung zu einem Optimierungsproblem im Modell.

Weitere Empfehlungen zur Vermeidung von Bias/Verzerrungen:

1. Richten Sie einen Feedback-Mechanismus oder einen offenen Dialog mit Nutzerinnen und Nutzern ein, um Rückmeldungen über wahrgenommene Verzerrungen oder Probleme zu erhalten.
2. Wählen Sie ein diverses Team, das eine größere Bandbreite an Erfahrungen repräsentieren kann, um Verzerrungen zu minimieren. Beziehen Sie verschiedene Altersgruppen, Ethnien, Geschlechter, Bildungsdisziplinen und kulturelle Perspektiven in das Team ein.
3. Je nach Art der Daten, die die KI erfasst, kann sie anfällig für verschiedene Arten der Verzerrung sein. Überwachen Sie Training und Ergebnisse, um schnell auf Probleme reagieren zu können. Testen Sie früh und oft.

39 Fairness Metrics, https://developers.google.com/machine-learning/glossary/fairness#fairness_metric

that should be chosen according to the criticality level of the product or service (highly critical use cases with customer contact require a longer testing period).

Accuracy method

The aim of this method is that the developers of the model already know prior to launch in which situations the model works better or worse. The situations that are known to be particularly challenging should be given special attention. This can be done either in the pre- or post-processing phase. The advantage of implementing the accuracy method during the pre-processing phase is both to ensure that the AI model is not overloaded and to allow the input data to be modified before it is entered into the AI model to avoid the edge case and still allow processing.

Runtime error detection[40]

Runtime error detection is a software-based verification to analyze if the model reports and executes errors while running. It can be applied during different test phases.

Improving quality

Quality standards are listed in DIN SPEC 92001–2[41].
Below are possible methods for reducing incorrect decisions by the AI and increasing trust in general:

1. Critical decisions are only given as recommendation: In general, it must be assumed that an AI will also make wrong decisions. It is therefore recommended that the AI should only make a recommendation in particularly critical use cases or for sensitive decisions. This recommendation should be understandable and well justified for the user.

2. Limitation of input data: In use cases where it is easy to predict the range of input data, a restriction or filter process can be implemented to limit the possible input data. The limitation of the input data leads not only to a reduction of possible incorrect decisions, it also lowers the risk of an adversarial attack.

3. Limitation of the output data: A similar method to the one above is the limitation of output data. By limiting the output data to a limited, possible output number, a plausibility check can also be performed.

4. Increasing the confidence score:
The Confidence Score, or Classification Threshold, indicates how certain the model is that the relevant intent was correctly assigned. The score can

40 Seshia, Verified Artificial Intelligence: A Runtime Verification Perspective; 2019
41 DIN SPEC 92001–2: Artificial Intelligence –Life Cycle Processes and Quality Requirements – Part 2: Robustness

Prüfung und Validierung

Planen Sie ausreichend Zeit für die Testperiode, sowie Überprüfung, Bewertung, Fehlerbehebung und Feedback-Schleifen ein. Aus berufsethischer Sicht kann dies so aussehen: Identifizieren Sie für die jeweiligen Lernaufgaben Szenarien, die sicherheits- oder gefahrenträchtig sind und/oder während des Einsatzes häufig auftreten. Erfassen Sie für diese Szenarien relevante Datensätze, Simulationsumgebungen sowie potenzielle reale Testumgebungen und formulieren Sie eine Argumentation, warum die gegebenen Daten, Simulationen und realen Testumgebungen für die kritischen Szenarien repräsentativ sind. Hierbei werden auch mögliche Einschränkungen des Modells berücksichtigt. Zusätzlich oder alternativ dazu müssen bestimmte Regionen des Datensatzes möglicherweise priorisiert werden. Damit geben Sie den kritischen Szenarien eine höhere Priorität (unabhängig davon, ob sie Risiken in Bezug auf die Sicherheit darstellen), und das Team kann fundierte Strategien für das Modell entwickeln. Nachfolgend finden Sie verschiedene Methoden zum Testen des Modells, die je nach Anwendungsfall erwogen und ggf. durchlaufen werden sollten, sowie die Testperiode, die entsprechend der Kritikalitätsstufe des Produkts oder der Dienstleistung gewählt werden sollte (hochkritische Anwendungsfälle mit Kundenkontakt benötigen eine längere Testperiode).

Genauigkeits-Methode

Ziel dieser Methode ist es, dass die Entwickler*innen des Modells schon vor der Einführung wissen, in welchen Situationen das Modell besser oder schlechter funktioniert. Die Situationen, von denen bekannt ist, dass sie eine besondere Herausforderung darstellen, sollten besonders betrachtet werden. Dies kann entweder in der Pre- oder der Post-Processing-Phase geschehen. Der Vorteil, die Genauigkeits-Methode während der Pre-Processing-Phase durchzuführen, ist zum einen, das sichergestellt wird, dass das KI-Modell nicht überlastet wird, und zum anderen darin, dass die Eingabedaten modifiziert werden können, bevor sie in das KI-Modell eingegeben werden, um den Edge-Case zu vermeiden und trotzdem eine Verarbeitung zu ermöglichen.

Erkennung von Laufzeitfehlern[40]

Bei der Laufzeitfehlererkennung handelt es sich um eine softwarebasierte Verifikation zur Analyse, ob das Modell beim Auftreten von Fehlern diese Fehler auch meldet oder ein fehlerhaftes Verhalten ausführt. Sie kann in verschiedenen Testphasen angewendet werden.

Steigerung der Qualität

Qualitätsstandards sind in DIN SPEC 92001–2[41] aufgeführt.
Nachfolgend finden Sie mögliche Methoden, um Fehlentscheidungen der KI zu reduzieren und die Vertrauenswürdigkeit insgesamt zu erhöhen:

40 Seshia, Verified Artificial Intelligence: A Runtime Verification Perspective; 2019
41 DIN SPEC 92001–2: Artificial Intelligence –Life Cycle Processes and Quality Requirements – Part 2: Robustness

have a value between 0 and 1, depending on the model. By increasing the confidence score for a output delivery, the possibility for a wrong decision can be reduced.

5. Implementation of the loss function: The loss function is used to calculate deviations and wrong decisions between the real and the received answers from the model. The loss functions provide a value for "how good" the AI is at dealing with real data compared to the extent with training data. The goal is to reduce the loss function value by evaluating the incorrect or inaccurate responses and improving the process or training data.

6. Continuous monitoring: This method describes a continuous feedback mechanism that all processes run as intended and problems are detected and resolved at early stage. This can be implemented with a software or even another AI, which hardens the use case of the monitored AI.

Lifecycle quality assessment

The quality requirements for AI modules must be linked with service life. Certain requirements, such as avoiding harmful bias, are necessary both for development and during deployment. If an AI system or feature is no longer needed, is redundant, or if an updated version is available such that the previous one is obsolete, the AI system or feature must be deactivated.

1. Kritische Entscheidungen nur als Empfehlung aussprechen: Im Allgemeinen muss davon ausgegangen werden, dass eine KI auch falsche Entscheidungen trifft. Es wird daher empfohlen, die KI in besonders kritischen Anwendungsfällen oder bei sensiblen Entscheidungen nur eine Empfehlung aussprechen zu lassen. Diese Empfehlung sollte leicht verständlich und gut begründet sein.

2. Beschränkung der Eingabedaten: In Anwendungsfällen, in denen es einfach ist, den Bereich der Eingabedaten vorherzusagen, kann ein Einschränkungs- oder Filterprozess implementiert werden, um die möglichen Eingabedaten einzuschränken. Die Begrenzung der Eingabedaten führt nicht nur zu einer Reduzierung möglicher Fehlentscheidungen, sondern senkt auch das Risiko eines gegnerischen Angriffs.

3. Beschränkung der Ausgabedaten: Eine ähnliche Methode wie die oben beschriebene ist die Beschränkung der Ausgabedaten. Durch die Beschränkung der Ausgabedaten auf eine begrenzte, mögliche Ausgabezahl kann zudem eine Plausibilitätsprüfung durchgeführt werden.

4. Erhöhung des Confidens-Scores: Der Confidens-Score oder Klassifizierungsschwellenwert gibt an, wie wahrscheinlich es für das Modell ist, dass die relevante Absicht richtig zugeordnet wurde. Der Score kann je nach Modell einen Wert zwischen 0 und 1 haben. Durch die Erhöhung des Confidens-Scores für eine Output-Übermittlung kann die Möglichkeit einer Fehlentscheidung verringert werden.

5. Implementierung der Verlustfunktion (Loss Function): Die Verlustfunktion wird verwendet, um Abweichungen und Fehlentscheidungen zwischen den realen und den erhaltenen Antworten des Modells zu berechnen. Die Verlustfunktionen liefern einen Wert dafür, „wie gut“ die KI im Umgang mit realen Daten im Vergleich zum Umfang mit Trainingsdaten ist. Ziel ist es, den Wert der Verlustfunktion zu verringern, indem die falschen oder ungenauen Antworten bewertet und der Prozess oder die Trainingsdaten verbessert werden.

6. Kontinuierliche Überwachung: Diese Methode beschreibt einen kontinuierlichen Feedback-Mechanismus, der sicherstellt, dass alle Prozesse wie vorgesehen ablaufen und Probleme frühzeitig erkannt und gelöst werden. Dies kann mit einer Software oder sogar mit einer anderen KI implementiert werden, was den Anwendungsfall der überwachten KI härtet.

Bewertung der Qualität im Produktlebenszyklus

Die Qualitätsanforderungen für KI-Module müssen mit der Lebensdauer verzahnt werden. Bestimmte Anforderungen, die z. B. die Vermeidung von schädlichem Bias betreffen, sind sowohl für die Entwicklung als auch während des Einsatzes erforderlich. Wird ein KI-System oder Feature nicht mehr benötigt, ist es redundant oder liegt eine aktualisierte Version vor, sodass die vorherige veraltet ist, so ist das KI-System oder Feature zu deaktivieren.

08

We foster the cooperative model

08

Wir leben das Kooperationsmodell

08

We foster the cooperative model

Autonomy – Building trust – Designing human-centered AI

We believe that human and machine intelligence are complementary, with each bringing its own strength to the table. While we believe in a people-first approach to human-machine collaboration, we recognize that humans can benefit from the strength of AI to unleash a potential that neither human nor machine can unleash on their own.

We recognize the widespread fear, that AI-enabled machines could outsmart human intelligence. At DT, we think differently. We know and believe in human strengths such as inspiration, intuition, sensory perception and empathy. But we also recognize the strengths of AI such as data transfer, processing speed and analysis resources. Through the cooperation of humans and machines, AI systems can help people make better decisions and achieve goals more effectively and efficiently.

Target picture

AI systems are intelligent support systems and recommendation systems for internal and external users. At a later stage, when people understand how decisions are made by the AI system (evolution model) – and if they approve of this kind of decision-making – this may change to humans delegating decisions to AI systems; even in more crucial areas, so that systems and humans can unleash their combined potential.

To-dos

Prior to development

- Product Owner ensures that the team has matching values and mindset and follows the vision of the product/service.

During development

- Development Team implements measures for human decision-making, for securing autonomy.

After launch

- Product Owner gets in contact with customers to collect feedback in order to improve the AI.

08

Wir leben das Kooperationsmodell

Autonomie – Vertrauensbildung – Menschenzentrierte KI gestalten

Wir glauben, dass menschliche und maschinelle Intelligenz sich ergänzen, weil jeder seine eigenen Stärken mitbringt. Da der Mensch an erster Stelle steht, sind wir überzeugt, dass er von der Stärke der KI profitieren kann, um ein Potential zu entfalten, das weder Mensch noch Maschine alleine freisetzen können.

Wir sind uns der Befürchtungen bewusst, dass computergesteuerte Maschinen die menschliche Intelligenz überlisten könnten. Wir als Deutsche Telekom denken anders. Wir kennen und glauben an die menschlichen Stärken wie Inspiration, Intuition, Sinneswahrnehmung und Empathie. Aber wir erkennen auch die Stärken von KI wie Datentransfer, Verarbeitungsgeschwindigkeit und Analyseressourcen. Durch die Kooperation von Mensch und Maschine können KI-Systeme den Menschen helfen, bessere Entscheidungen zu treffen und Ziele effektiver und effizienter zu erreichen.

Zielbild

KI-Systeme sind intelligente Unterstützungs- und Empfehlungssysteme für interne und externe Nutzerinnen und Nutzer. In einer späteren Phase, wenn Menschen verstehen, wie Entscheidungen durch das KI-System (Evolutionsmodell) getroffen werden – und wenn sie diese Art der Entscheidungsfindung billigen – kann sich dies dahingehend ändern, dass Menschen die Entscheidungen an KI-Systeme delegieren; sogar in entscheidenderen Bereichen, damit Systeme und Menschen ihr kombiniertes Potenzial entfalten können.

To-dos

Vor der Entwicklung

- Product Owner stellt sicher, dass das Team übereinstimmende Werte und Denkweisen hat und der Vision des Produkts / der Dienstleistung folgt.

Während der Entwicklung

- Development Team implementiert Maßnahmen zur menschlichen Selbstentscheidung, zur Sicherung der Souveränität.

Nach dem Launch

- Product Owner nimmt Kontakt mit den Kund*innen auf, um Feedback zur Verbesserung der KI zu sammeln.

Autonomy

A DT AI must always respect human autonomy, which means that a human's decision-making ability must take precedence. The user must be given the opportunity to take risks, provided these have been communicated. Furthermore, DT systems must not support people in pursuing illegal or immoral goals or actions. The user should actively choose how and whether to delegate decisions to AI systems.

It must also be possible to reverse any delegation to the AI. In practical terms, this means that, depending on the use case, there is a special mode for overriding and, if necessary, reversing a decision made by the AI.[42] This mode should be documented, and the user working with the AI must be informed about how this mode can be activated and used correctly.

Building trust

For the team, it is crucial to understand that even simple AI systems can lead people to trust machines more. In particular, people who are unfamiliar with how systems work and how they are built might attribute human characteristics to the corresponding machines (anthropomorphism) and give up too much of their own judgment. This gives rise to various ethical risks, as AI might be able to use large amounts of data to manipulate people into doing things that are not in their best interests.[43]

This effect can never be completely ruled out, so it is important to ensure that users are guided and their trust is not abused. Below you will find a guideline for maintaining this trust and also strengthening trust in DT.[44] Keep in mind:

1. Technology as a tool influences users and societies: As a development team, you influence the user of your service with content and interface. It is impossible to present all available choices with equal priority, values and assumptions are always involved. Make sure that the options provided and the form they take are consistent with DT values.

2. Dynamics of hierarchy, value creation, and social dynamics shape the impact of new technology, regardless of the intentions of the inventors. Economic pressures (e.g. pressure to increase sales for shareholders) or power politics (e.g. an ethnic group using a powerful technology against a marginalized ethnic group) can have profound consequences. Most often, the result is a worsening of inequalities in the world.

3. The human brain is inherently sensitive, malleable and subjective. To reflect the design and service in terms of ethical value and human interaction, not only cognitive biases but also cognitive weaknesses need to be taken into account – on the part of the user and the development team.

42 Floridi, Cowls: A Unified Framework of Five Principles for AI and Society; 2019, https://hdsr.mitpress.mit.edu/pub/l0jsh9d1/release/6

43 Barneck e.a.: an introduction to ethics in Robotics and AI; 2020

44 Center for human technology: Principles of human design; 2020: https://www.humanetech.com/technologists#principles

Autonomie

Eine KI der Deutschen Telekom muss immer die menschliche Autonomie achten, das heißt die Entscheidungsfähigkeit eines Menschen muss vorrangig sein. Allen, die die KI verwenden, muss die Möglichkeit gegeben werden, Risiken einzugehen, sofern diese kommuniziert wurden. Darüber hinaus dürfen Systeme der Deutschen Telekom den Menschen nicht bei der Ausübung illegaler oder unmoralischer Ziele oder Handlungen unterstützen. Nutzerinnen und Nutzer sollten aktiv wählen, wie und ob sie Entscheidungen an KI-Systeme delegieren.

Jede Delegation an die KI muss auch wieder zurückgenommen werden können. Praktisch bedeutet dies, dass es je nach Use Case einen besonderen Modus gibt, um eine Entscheidung der KI zu überwinden und gegebenenfalls rückgängig zu machen.[42] Dieser Modus sollte dokumentiert werden – wer mit der KI arbeitet, muss darüber informiert werden, wie dieser Modus aktiviert und korrekt verwendet werden kann.

Vertrauensbildung

Für das Team ist es entscheidend, zu verstehen, dass selbst einfache KI-Systeme Menschen dazu bringen können, Maschinen mehr Vertrauen entgegenzubringen. Vor allem Menschen, die mit der Funktionsweise und dem Aufbau von Systemen nicht vertraut sind, könnten den entsprechenden Maschinen menschliche Eigenschaften zuschreiben (Anthropomorphismus) und die eigene Urteilskraft zu sehr aufgeben. Daraus ergeben sich verschiedene ethische Risiken, da die KI in der Lage sein könnte, mit Hilfe großer Datenmengen Menschen dahingehend zu manipulieren, dass diese Dinge tun, die nicht in ihrem Sinne sind.[43]

Dieser Effekt lässt sich nie ganz ausschließen, deshalb muss sichergestellt werden, dass Nutzerinnen und Nutzer angeleitet werden und ihr Vertrauen nicht missbraucht wird. Nachfolgend finden Sie einen Leitfaden, um dieses Vertrauen zu erhalten und auch das Vertrauen in die Deutsche Telekom zu stärken.[44] Bedenken Sie:

1. Technologie als Werkzeug beeinflusst Menschen und Gesellschaften: Als Entwicklungsteam beeinflussen Sie Nutzerinnen und Nutzer Ihres Dienstes mit Inhalt und Interface. Es ist unmöglich, alle verfügbaren Auswahlmöglichkeiten mit gleicher Priorität zu präsentieren, Werte und Annahmen fließen immer mit ein. Achten Sie darauf, dass die zur Verfügung gestellten Möglichkeiten und deren Form mit den Werten der Deutschen Telekom übereinstimmen.

2. Dynamiken von Hierarchie, Wertschöpfung und soziale Dynamiken prägen die Auswirkungen von neuer Technologie, unabhängig von den Absichten der Erfinder*innen. Wirtschaftlicher Druck (z. B. der Druck, den Umsatz für Aktionäre zu steigern) oder Machtpolitiken (z. B. von Seiten einer ethnische Gruppe, die eine mächtige Technologie gegen eine marginalisierte ethnische Gruppe einsetzt) können tiefgreifende Folgen haben. Meistens ist das Ergebnis eine Verschärfung der Ungleichheiten in der Welt.

42 Floridi, Cowls, A Unified Framework of Five Principles for AI and Society; 2019, https://hdsr.mitpress.mit.edu/pub/l0jsh9d1/release/6

43 Barneck e.a., an introduction to ethics in Robotics and AI; 2020

44 Vgl. Center for human technology, Principles of human design; 2020: https://www.humanetech.com/technologists#principles

Designing human-centered AI

Here are a few more specific tips on how you can meet the above challenges[45]:

1. Choose values instead of engagement metrics to measure success[46]: The problem with engagement metrics is that they assume that customer value has been met if a category such as screen time optimization has been met. Instead, you can use values to guide you in implementing your KPIs and measuring the success of your product. Select values you want to support, such as health, well-being, connection, productivity, fun, creativity ...

2. Enable wise choices, instead of assuming that more choice is always better: As the world becomes increasingly complex and unpredictable, the ability to understand the new reality and make meaningful decisions can quickly become overwhelming. As a technology employee at DT, you can help people make choices in ways that are informed, thoughtful, and aligned with their values. For example, when you contextualize information with meaningful data, appropriate presentation can help people make good decisions: Hearing that Covid-19 has a 1 percent mortality rate may not mean much to you. But hearing that Covid-19 is several times deadlier than the flu helps anyone immediately understand it in the context of something they already know. When people are presented with information in an intuitive way, they are empowered to make wise choices.

3. Nurture mindfulness instead of excessive attention: Mindfulness allows us to act with intention and to avoid a life that becomes a series of automatic actions and reactions, often based on fear. You can help your users regain and increase their capacity for awareness, rather than racing to win more of their attention. Question signals and interactions you build. Also, question whether the signals you use to vie for attention are consistent with the value and importance of the service.
For example, a mail application that by default makes a sound and displays a notification when mail is received might invite the user to turn on the silent notification. Another example is the Apple Watch Breathe app, which helps people take a moment at regular intervals to simply focus on their breathing.

4. Growth comes with responsibility:
AI can have asymmetric power over different user types. Machine learning, micro-targeting, recommendation engines and deep fakes are all examples of technologies that could be used for good and bad, but at the same time increase the likelihood of being used with greater harm once the technology is scaled. To prevent this effect, consider the cognitive, social,

45 Center for Humane Technology: technologist principles; 2020: https://www.humanetech.com/technologists#principles
46 Center for Humane Technology: technologist principles; 2020: https://www.humanetech.com/technologists#principles

3. Das menschliche Gehirn ist von Natur aus sensibel, formbar und subjektiv. Um das Design und den Service in Bezug auf den ethischen Wert und die menschliche Interaktion zu reflektieren, müssen nicht nur kognitive Voreingenommenheit, sondern auch kognitive Schwächen berücksichtigt werden – bei denen, die sie nutzen und beim Entwicklungsteam.

Menschenzentrierte KI gestalten

Hier noch ein paar konkrete Hinweis dafür, wie Sie den oben genannten Herausforderungen begegnen können[45]:

1. Entscheiden Sie sich für Werte anstatt Engagement-Metriken zur Erfolgsmessung[46]: Bei Engagement-Metriken besteht das Problem, dass angenommen wird, dass der Kundennutzen erfüllt wurde, wenn eine Kategorie wie Bildschirmzeitoptimierung erfüllt wurde. Stattdessen können Sie sich bei der Implementierung Ihrer KPIs und der Messung des Erfolgs Ihres Produkts von Werten leiten lassen. Wählen Sie Werte aus, die Sie unterstützen möchten, z. B. Gesundheit, Wohlbefinden, Verbindung, Produktivität, Spaß, Kreativität ...

2. Ermöglichen Sie kluge Entscheidungen anstatt davon auszugehen, dass mehr Auswahl immer besser ist: Da die Welt immer komplexer und unberechenbarer wird, kann die Fähigkeit, die neue Realität zu verstehen und sinnvolle Entscheidungen zu treffen, schnell überfordert werden. Als Technolog*in bei der Deutschen Telekom können Sie den Menschen helfen, Entscheidungen auf eine Art und Weise zu treffen, die sachkundig und überlegt ist und mit ihren Werten übereinstimmt. Wenn Sie zum Beispiel Informationen mit aussagekräftigen Daten in einen Kontext setzen, kann eine angemessene Darstellung den Menschen helfen, gute Entscheidungen zu treffen: Wenn Sie hören, dass Covid-19 eine Sterblichkeitsrate von 1 Prozent aufweist, bedeutet das für Sie vielleicht nicht viel. Aber zu hören, dass Covid-19 um ein Vielfaches tödlicher ist als eine Grippe, hilft jedem, es sofort im Zusammenhang mit etwas zu verstehen, das er bereits kennt. Wenn Menschen Informationen auf intuitive Weise vermittelt bekommen, sind sie in der Lage, kluge Entscheidungen zu treffen.

3. Fördern Sie Achtsamkeit statt übermäßiger Aufmerksamkeit: Achtsamkeit ermöglicht es uns, absichtsvoll zu handeln und konditionierte Reaktionen zu vermeiden, die oft auf Angst beruhen. Sie können Ihren Nutzerinnen und Nutzern helfen, die Fähigkeit zur Achtsamkeit wiederzuerlangen und zu steigern, anstatt um mehr Aufmerksamkeit zu buhlen. Hinterfragen Sie Signale und Interaktionen, die Sie bauen. Hinterfragen Sie auch, ob die Signale, mit denen Sie um Aufmerksamkeit buhlen, mit dem Wert und der Wichtigkeit des Dienstes übereinstimmen. Beispielsweise könnte eine E-Mail-Anwendung, die standardmäßig einen Ton erzeugt und eine Benachrichtigung anzeigt, wenn eine

45 Center for Humane Technology, technologist principles; 2020: https://www.humanetech.com/technologists#principles
46 Vgl. ebd

economic and environmental impacts of your technology as it is scaled and look for ways to mitigate these impacts.

For example, a product originally intended only for adults will almost inevitably be used by children as it scales; and a product that mostly benefits people in a developed country may cause harm in the global South.

In order to implement these principles of human-centered technology, you can use the canvas for a team workshop, which is provided in the appendix.[47]

47 Center for Humane Technology: technologist principles; 2020: https://www.humanetech.com/technologists#principles

Mail eingetroffen ist, den Nutzer dazu einladen, die stumme Benachrichtigung einzuschalten. Ein weiteres Beispiel ist die Apple Watch App *Breathe*, die Menschen dabei unterstützt, sich in regelmäßigen Abständen einen Moment Zeit zu nehmen, um sich einfach auf ihren Atem zu konzentrieren.

4. Wachstum kommt mit Verantwortung: KI kann asymmetrische Macht über verschiedene Benutzertypen haben. Maschinelle Lernverfahren, Micro-Targeting und Empfehlungsmaschinen können für Gutes und Schlechtes (z. B. Deep Fakes) eingesetzt werden – wenn diese Technologien skaliert werden, erhöht sich das Risiko größerer Schäden. Um diesen Effekt zu verhindern, sollten Sie die kognitiven, sozialen, wirtschaftlichen und ökologischen Auswirkungen Ihrer Technologie bei der Skalierung berücksichtigen und nach Möglichkeiten suchen, diese Auswirkungen abzuschwächen.

So wird zum Beispiel ein Produkt, das ursprünglich nur für Erwachsene bestimmt war, fast zwangsläufig bei seiner Skalierung von Kindern genutzt; und ein Produkt, das den Menschen in einem Industrieland am meisten nützt, kann im globalen Süden Schaden anrichten.

Um diese Prinzipien der menschenzentrierten Technologie umzusetzen, können Sie einen Canvas für einen Workshop mit Ihrem Team verwenden, der im Anhang zu finden ist.[47]

47 Center for Humane Technology, technologist principles; 2020: https://www.humanetech.com/technologists#principles

09

We share and enlighten

09

Wir teilen und erklären

09

We share and enlighten

Mindset – Teaching – Sharing best practices – Responsibility – Sustainability

We are aware of the transformative power of AI for our society. We want to help people and society prepare for this future world. We live our digital responsibility by sharing our knowledge and highlighting the opportunities of the new technology without neglecting its risks. We want to work with our customers, other companies, policymakers, educational institutions and all other stakeholders to understand their concerns and needs and take the right protective measures. We want to engage in education on AI and ethics. In this way, we are preparing ourselves, our colleagues and our fellow humans for the new tasks ahead.

Many tasks performed by humans will be further automated in the future. This will lead to a shift in skill requirements. While this development seems predictable, few know what AI technologies will be capable of. Prejudice and half-knowledge lead to either demonization of progress or blind faith. Both approaches require educational work. We at DT feel a responsibility to educate people and help society deal with the digital transformation so that new adequate skills can be developed and new jobs created. And we start from within – by empowering our colleagues and employees accordingly. We are aware that this task cannot be solved by one company alone. That is why we want to enter into partnerships with other companies and make our expertise available to policymakers and education providers so that we can tackle the challenges ahead together.

Target picture

Knowledge fails if it is not spread. A lack of knowledge leads to fear, neglect and/or regression. To be able to act appropriately, we help people gain a better understanding of the available options with their opportunities and risks. DT employees either know about the risks and opportunities of the technology themselves or know someone they can ask. Experts give presentations within DT and to the outside world to share their expertise.

09

Wir teilen und erklären

Mindset – Lehren – Austausch von Best Practices – Verantwortung – Nachhaltigkeit

*Die transformative Kraft der KI für unsere Gesellschaft ist uns bewusst. Wir wollen Menschen und die Gesellschaft bei der Vorbereitung auf diese zukünftige Welt unterstützen. Wir leben unsere digitale Verantwortung, indem wir unser Wissen teilen und die Möglichkeiten der neuen Technologie aufzeigen, ohne ihre Risiken zu vernachlässigen. Wir wollen mit unseren Kund*innen, anderen Unternehmen, politischen Entscheidungsträger*innen, Bildungseinrichtungen und allen anderen Interessengruppen zusammenarbeiten, um ihre Anliegen und Bedürfnisse zu verstehen und die richtigen Schutzmaßnahmen treffen zu können. Wir wollen uns in der Bildung zu KI und Ethik engagieren. Hierdurch bereiten wir uns, unsere Kolleg*innen und unsere Mitmenschen in diesem Sinne auf die neuen Anforderungen vor.*

*Viele Aufgaben, die von Menschen ausgeführt werden, werden in Zukunft weiter automatisiert. Dieses führt zu einer Verschiebung des Qualifikationsbedarfs. Während diese Entwicklung vorhersehbar erscheint, wissen nur Wenige, wozu KI-Technologien in der Lage sein werden. Vorurteile und Halbwissen führen entweder zu einer Dämonisierung des Fortschritts oder zu blindem Vertrauen. Beide Ansätze erfordern Aufklärungsarbeit. Wir als Deutsche Telekom fühlen uns verantwortlich, Menschen aufzuklären und der Gesellschaft zu helfen, mit dem digitalen Wandel umzugehen, damit neue adäquate Fähigkeiten entwickelt und neue Arbeitsplätze geschaffen werden können. Und wir beginnen bei uns selbst – indem wir unsere Kolleg*innen und Mitarbeiter*innen entsprechend befähigen. Wir sind uns bewusst, dass diese Aufgabe nicht von einem Unternehmen allein gelöst werden kann. Daher wollen wir Partnerschaften mit anderen Unternehmen eingehen und unser Know-how politischen Entscheidungsträger*innen und Bildungsanbietern zur Verfügung stellen, um die anstehenden Herausforderungen gemeinsam zu bewältigen.*

Zielbild

Wissen scheitert, wenn es nicht verbreitet wird. Nicht-Wissen führt zu Angst, Vernachlässigung und/oder Rückschritt. Um angemessen handeln zu können, verhelfen wir den Menschen zu einem besseren Verständnis der vorhandenen Möglichkeiten mit ihren Chancen und Risiken. Die Mitarbeitenden der Telekom wissen entweder selbst über die Risiken und Möglichkeiten der Technologie Bescheid oder kennen jemanden, den sie fragen können. Expert*innen halten Vorträge innerhalb der DT und in der Außenwelt, um ihr Fachwissen weiterzugeben.

To-dos

Prior to development
- Product Owner chooses a team with the right skills and uses the possibility of re- and upskilling.
- Product Owner informs the team about training opportunities and initiatives.

During development
- Team gets in contact with initiatives within Telekom in order to collect best practices.

After development
- Team participates in initiatives.
- Team shares a positive mindset and enlightens about the topic in the environment.

Mindset

DT employees should share the available knowledge about the risks and opportunities of technology (not only in relation to AI) with colleagues, but also with people outside of DT, e.g. in schools, in dialog. Of course, this does not include insider knowledge, as well as knowledge relevant to security and business or information about parties external to DT. Mechanisms for detecting harmful and unsuitable data should also not be published to avoid increasing the vulnerability of AI systems.

Teaching

The respective departments should work together on formats such as eLearning methods and training for colleagues and keep their knowledge up-to-date. Knowledge and expertise as well as cross-functional thinking and collaboration form the basis for sustainable solutions. Throughout the development process and beyond, employees, especially developers, should keep themselves informed about advances in AI technology and the application of such. In this context, exchange with research institutes and the like is crucial in order to evaluate current and future use cases and opportunities.

Sharing best practices

Additionally, for every AI system or feature, it should be clearly documented which kinds of algorithms and models are used for which outcome. Code should also be commented to make it easier for other developers to understand the selected approach and the reasons for this selection.

Responsibility

As a knowledge provider that implements new technologies for society, DT has a responsibility to inform society about AI and share its knowledge. If you wish to participate in this responsibility, please use DT‘s official channels.

To-dos

Vor der Entwicklung
- Product Owner wählt ein Team mit den richtigen Skills aus und nutzt Möglichkeit des Re- und Upskilling.
- Product Owner informiert das Team über Schulungsmöglichkeiten und Initiativen.

Während der Entwicklung
- Team nimmt Kontakt zu Initiativen innerhalb der Telekom auf, um Best Practices zu sammeln.

Nach der Entwicklung
- Team nimmt an Initiativen teil.
- Team teilt eine positive Denkweise und klärt im Umfeld über das Thema auf.

Mindset

Mitarbeiterinnen und Mitarbeiter der Deutschen Telekom sollten das zur Verfügung stehende Wissen über die Risiken und Chancen der Technik (nicht nur in Bezug auf KI) mit den Kolleg*innen, aber auch mit Menschen außerhalb der Telekom, z. B. in Schulen im Dialog teilen. Dazu gehören natürlich nicht Insiderwissen, sowie sicherheits- und geschäftsrelevantes Wissen oder Informationen über Dritte. Auch Mechanismen zur Erkennung schädlicher und ungeeigneter Daten sollten nicht veröffentlicht werden, um die Verwundbarkeit von KI-Systemen nicht zu erhöhen.

Lehren

Die jeweiligen Abteilungen sollen gemeinsam an Formaten wie eLearning und Schulungen für Kolleg*innen arbeiten und ihr Wissen auf dem neuesten Stand halten. Wissen und Expertise sowie funktionsübergreifendes Denken und Zusammenarbeit bilden die Grundlage für nachhaltige Lösungen. Während des gesamten Entwicklungsprozesses und darüber hinaus sollen sich die Mitarbeiter*innen, insbesondere die Entwickler*innen, über Fortschritte auf dem Gebiet der KI-Technologie und deren Anwendung informieren. Dabei ist der Austausch mit Forschungsinstituten und dergleichen entscheidend, um gegenwärtige und zukünftige Anwendungsfälle und Chancen zu bewerten.

Austausch von Best Practices

Zusätzlich sollte für jedes KI-System oder Merkmal klar dokumentiert werden, welche Arten von Algorithmen und Modellen für welches Ergebnis verwendet werden. Der Code sollte auch kommentiert werden, um es anderen Personen in der Entwicklung leichter zu machen, den gewählten Ansatz und die Gründe für die Auswahl zu verstehen.

Verantwortung

Die Deutsche Telekom als Wissensträger in der Gesellschaft, der neue Technologien umsetzt, hat die Verantwortung, die Gesellschaft über KI zu informieren und das Wissen zu teilen. Nutzen Sie daher gegebenenfalls die offiziellen Kanäle der Telekom, falls Sie sich an dieser Verantwortung beteiligen möchten.

Sustainability

Sustainability means that humanity has the ability to shape development in a sustainable way – ensuring that it meets current needs without reducing the chances of future generations to meet their own needs.[48] Products and Services of Deutsche Telekom should be developed according to this principle. For this reason, if possible, align your developments with the Sustainable Development Goals[49].

48 Bundtland-Bericht: Unsere gemeinsame Zukunft; 1987

49 Sustainable Development Goals: https://www.undp.org/content/undp/en/home/sustainable-development-goals.html

Nachhaltigkeit

Nachhaltigkeit bedeutet, dass die Menschheit die Fähigkeit hat, Entwicklung nachhaltig zu gestalten – sicherzustellen, dass die Anforderungen der Gegenwart erfüllt werden, ohne die Chancen künftiger Generationen zu schmälern, ihre eigenen Bedürfnisse zu befriedigen.[48] Nach diesem Prinzip sollten Produkte und Services der Deutschen Telekom entwickelt werden. Daher sollten sich Ihre Entwicklungen auch nach den Sustainable Development Goals[49] ausrichten.

48 Bundtland-Bericht: Unsere gemeinsame Zukunft; 1987

49 Sustainable Development Goals, https://www.undp.org/content/undp/en/home/sustainable-development-goals.html

Appendix

Anhang

Appendix

Many AI projects have been launched within DT, with communities and initiatives sharing their knowledge. It is recommended to get in touch with colleagues not only after the launch but also prior to development in order to to learn and improve. Other initiatives and groups will follow, recommendations are welcome:
Write to Digital-Ethics@Telekom.de

- Digital Ethics Center in Berlin: https://www.telekom.com/en/company/details/deutsche-telekom-launches-forum-for-digital-ethics-566438
- IT@School: https://yam.telekom.de/groups/itschools

Knowledge is what keeps us moving. Knowledge is what eliminates fear. Let's spread the word for a better future that respects both people and the environment – with the help of technology.

Structure for the risk assessment

Functional risk: The risk of the functionality of the system failing, as well as other software which might interact with the model, when released to the public or in other unanticipated situations.

Systemic risk: Systemic risks are risks that affect an entire system, which can have a greater impact than just the system of a single company, for example: The Global Financial Crisis of 2007–2008 was caused by widespread loan defaults in the US subprime market.

Subprime loans are given to people with weak credit ratings. Defaults stressed the major US mortgage lenders Fannie Mae and Freddie Mac and many homeowners abandoned their houses. This led to a collapse in confidence that exposed inadequate risk models.

Risk of fraud: This risk addresses the situation when software has been used to perpetrate fraud. For example: Volkswagen deliberately designed its emissions reduction system to only function during laboratory tests. As a result of this criminal deception, the cars passed tests in labs but emitted up to forty times these volumes on the road.

Nevertheless, Volkswagen promoted its fraudulently obtained "green" credentials. The result of this fraud is billions of dollars in fines[50] and societal damage.

50 Contag et al.; 2017

Anhang

Innerhalb der Deutschen Telekom wurden viele KI-Projekte gestartet; Communities und Initiativen teilen ihr Wissen. Es wird empfohlen, nicht erst nach dem Start, sondern auch schon vor der Entwicklung mit Kolleginnen und Kollegen in Kontakt zu treten, um zu lernen und sich zu verbessern. Andere Initiativen und Gruppen werden folgen, Empfehlungen sind willkommen:
Schreiben Sie an Digital-Ethics@Telekom.de

- Forum für digitale Ethik in Berlin: https://www.telekom.com/en/company/details/deutsche-telekom-launches-forum-for-digital-ethics-566438
- IT@School: https://yam.telekom.de/groups/itschools

Wissen ist das, was uns in Bewegung hält. Wissen ist, was Furcht benötigt. Lassen Sie uns das Wort für eine bessere Zukunft verbreiten, die sowohl den Menschen als auch die Umwelt respektiert – mit Hilfe der Technologie.

Struktur zur Risikoanalyse

Funktionales Risiko: Das Risiko eines Ausfalls der Funktionalität des Systems sowie anderer Software, die mit dem Modell interagieren könnte, wenn es für die Öffentlichkeit oder in anderen vorhergesehenen Situationen veröffentlicht wird.

Systemisches Risiko: Systemische Risiken sind Risiken, die Systeme betreffen, die über ein einzelnes Unternehmen hinausreichen: Die globale Finanzkrise von 2007–08 wurde durch weit verbreitete Kreditausfälle auf dem US-Subprime-Markt verursacht.

Subprime-Kredite wurden an Personen mit schwacher Bonität vergeben. Ausfälle belasteten die großen US-Hypothekenfinanzierer Fannie Mae und Freddie Mac, und Menschen verloren ihre Häuser. Dies führte zu einem Zusammenbruch des Vertrauenskapitals, wobei unzulängliche Risikomodelle aufgedeckt wurden.

Betrugsrisiko: Dieses Risiko betrifft die Situation, in der Software zur Ausführung von Betrug verwendet wird. Zum Beispiel: Volkswagen hat sein Emissionsminderungssystem bewusst so konzipiert, dass es nur während der Labortests funktioniert. Infolgedessen bestanden ihre Autos zwar Tests im Labor, emittierten aber bis zu vierzigmal höhere Mengen auf der Straße. Trotzdem warb Volkswagen für ihre betrügerisch erlangten „grünen“ Zeugnisse. Die Folge dieses Betrugs sind Strafen in Milliardenhöhe[50] und gesellschaftlicher Schaden.

50 Contag et al.; 2017

Safety risk: A safety risk to life or health can occur via a system failure, among other things. An example here is the robots that control industrial production which present physical risks to the people that work near them. People may be injured through collisions or by objects a robot is carrying or moving.

Risk to reputation: Systems that appear biased or prejudiced can cause great reputational damage. Depending on the scenario, a reputational risk is often also contained in other risks; a security risk would also damage DT‘s reputation. An example of a reputational risk: One of the first major scandals in the still young social media economy is the Cambridge Analytica scandal, which severely damaged Facebook‘s reputation. Facebook supplied data obtained by Cambridge Analytica for targeted political advertising.

Legal risk: Legal risk encompasses situations where a system becomes too successful and is viewed as causing an anti-competitive environment.

For example: In 2004, Microsoft was fined 497 million euros by the European Union for anti-competitive behavior. In 2018, the European Commission imposed a record fine of USD 5 billion on Google for antitrust violations involving Android technology.

Environmental risk: System failures can cause environmental disasters or damage. These include not only general failure, but also false-positive or false-negative decisions and mistaken decisions after the model has been attacked. As an example, the Bhopal disaster in India was the result of a gas leak at a Union Carbide factory in Bhopal, India. The leak resulted in an explosion and release of methyl isocyanate. The immediate explosion caused the death of nearly 4,000 people.

Social Risk: Social risks include actions that could involve the people, society or communities around the company. These risks may include increased traffic, noise pollution, and issues related to employee morale. Social risks associated with AI include technology-related social isolation, increased inequality, and local community issues related to the acceptable use of technology and AI.

Sicherheitsrisiko: Ein Sicherheitsrisiko für Leben oder Gesundheit kann unter anderem durch einen Systemausfall entstehen. Beispiele hierfür sind Roboter in der industriellen Produktion, die physische Risiken für die Menschen darstellen, die in ihrer Umgebung arbeiten. Menschen könnten durch Kollisionen oder durch Gegenstände, die ein Roboter trägt oder bewegt, verletzt werden.

Reputationsrisiko: Systeme, die befangen oder voreingenommen erscheinen, können großen Reputationsschaden anrichten. Je nach Szenario ist ein Reputationsrisiko oft auch in anderen Risiken enthalten, ein Sicherheitsrisiko würde auch die Reputation der Deutschen Telekom schädigen. Ein Beispiel für ein Reputationsrisiko: Einer der ersten großen Skandale in der noch jungen Social-Media-Ökonomie ist der Cambridge-Analytica-Skandal, der die Reputation von Facebook stark beschädigt hat. Facebook lieferte Daten, die von Cambridge Analytica gewonnen wurden zur gezielten politischen Werbung.

Rechtliches Risiko: Das rechtliche Risiko umfasst unter anderem Situationen, in denen ein System zu erfolgreich wird und als Ursache für ein wettbewerbsfeindliches Umfeld angesehen wird. Zum Beispiel: Im Jahr 2004 wurde Microsoft von der Europäischen Union wegen wettbewerbswidrigen Verhaltens mit einer Geldstrafe von 497 Millionen Euro belegt. Im Jahr 2018 verhängte die Europäische Kommission eine Rekordstrafe von 5 Milliarden US-Dollar gegen Google wegen kartellrechtlicher Verstöße mit der Android-Technologie.

Risiko von Umweltschäden: Systemausfälle können Umweltkatastrophen oder -schäden verursachen. Dazu gehören nicht nur das generelle Versagen, sondern neben anderen Möglichkeiten auch falsch-positive oder falsch-negative Entscheidungen und Fehlentscheidungen, nachdem das Modell angegriffen wurde. Ein Beispiel: Die Katastrophe von Bhopal in Indien war die Folge eines nicht erkannten Gaslecks in einer Fabrik von Union Carbide in Bhopal, Indien. Das Leck führte zu einer Explosion und zur Freisetzung von Methylisocyanat. Die unmittelbare Explosion verursachte den Tod von fast 4.000 Menschen.

Soziales Risiko: Zu den sozialen Risiken gehören Handlungen, die die Menschen, die Gesellschaft oder die Gemeinschaften um das Unternehmen herum einschließen könnten. Zu diesen Risiken können erhöhter Verkehr, Lärmbelästigung und Probleme im Zusammenhang mit der Moral der Mitarbeiter*innen gehören. Zu den sozialen Risiken im Zusammenhang mit KI zählen technologiebedingte soziale Isolation, zunehmende Ungleichheit und Fragen der lokalen Gemeinschaft im Zusammenhang mit der akzeptablen Nutzung von Technologie und KI.

(2)

Humane Design Guide (Alpha Version)

Use this worksheet to identify opportunities for Humane Technology.

Product of feature:

Value proposition:

Measure of success

Human Sensitivity	We are inhibited when	What inhibits	We are supported when	Opportunity to improve
Emotional What we feel in our body and in our physical health.	We are stressed, low on sleep, afraid or emotionally exhausted.	· Artificial scarcity · Urgency signalling · Constant monitoring · Optimizing for screentime	Design engenders calm, balance, safety, pauses and supports circadian rhythms.	High / Low
Attention How and where we focus our attention.	Attention is physicalogically drawn, overwhelmed or fragmented.	· Constant context switching · Many undifferentiated choices · Fearful information · No stopping cues (e.g. infinite scroll) · Unnecessary movement	Enabled to consider, learn, express and feel grounded.	High / Low
Sensemaking How we integrate what we sense with what we know.	Information is fear-based, out of context, confusing, or manipulative.	· Facts out of context · Over-personalized filters · Equating virality with credibility · Deceptive authority (ads vs. content)	Enabled to consider, learn, express and feel grounded.	High / Low
Decisionmaking How we align our actions with our intentions.	Intentions and agency are not solicited nor supported.	· Avatars to convey authority · Stalking ads and messages · Push content models · Serving preference over intent	Enabled to gain agency, purpose, and mobilization of intent.	High / Low
Social Reasoning How we understand and navigate our personal relationships.	Status, relationships or self-image are manipulated.	· Quantified social status · Viral sharing · Implied obligation · Enabling impersonation	Enabled to connect more safely and authentically with others.	High / Low
Group Dynamics How we navigate larger groups, status, and shared understanding.	Excluded, divided or mobilized through fear.	· Suppressing views and nuance · Enabling ad hominem or hate speech · Enabling viral outrage · Lack of agreed-upon norms	Enabled to develop a sense of belonging and cooperation.	High / Low

Figure 2: Center for Humane Technology, www.humanetech.com

What are Human Sensitivities ?

Human Sensitivites are instincts that are often vulnerable to new technologies.

Now rank the sensitivities 1–6 based on what you now see as the largest opportunities for Humane Design. Then use the second sheet to develop an action statement.

Glossary

A

accuracy
The fraction of predictions that a classification model got right. In multi-class classification, accuracy is defined as follows:
Accuracy=Correct Predictions/Total Number Of Examples

In binary classification, accuracy has the following definition:
Accuracy=True Positives+True Negatives/Total Number Of Examples

B

bias · Bias (ethics fairness)
1. Stereotyping, prejudice or favoritism towards some things, people, or groups over others. These biases can affect collection and interpretation of data, the design of a system, and how users interact with a system. Forms of this type of bias include:
automation bias, confirmation bias, experimenter's bias, group attribution bias, implicit bias,in-group bias, out-group homogeneity bias

2. Systematic error introduced by a sampling or reporting procedure. Forms of this type of bias include:
coverage bias, non-response bias, participation bias, reporting bias, sampling bias, selection bias, prediction bias (in machine learning)

3. A value indicating how far apart the average of predictions is from the average of labels in the dataset.

C

classification model
A type of machine learning model for distinguishing among two or more discrete classes. For example, a natural language processing classification model could determine whether an input sentence was in French, Spanish, or Italian.

counterfactual fairness
fairness metric that checks whether a classifier produces the same result for one individual as it does for another individual who is identical to the first, except with respect to one or more sensitive attributes. Evaluating a classifier for counterfactual fairness is one method for surfacing potential sources of bias in a model.

Glossar

B

Bias · Voreingenommenheit
1. Stereotypisierung, Vorurteile oder Voreingenommenheit gegenüber bestimmten Dingen, Personen oder Gruppen. Diese Voreingenommenheit kann die Sammlung und Interpretation von Daten, das Design eines Systems und die Art und Weise, wie Nutzer*innen mit einem System interagieren, beeinflussen. Formen dieser Art von Voreingenommenheit umfassen: Automatisierungs-Bias, Bestätigungs-Bias, Experimentator's Bias, Gruppenzuweisungs-Bias, impliziter Bias, In-Group-Bias, Out-Group Homogenitäts-Bias ...

2. Systematischer Fehler, der durch ein Stichproben- oder Berichtsverfahren eingeführt wurde. Formen dieser Art der Verzerrung umfassen: Erfassungsbias, Antwortausfall-Bias, Teilnahme-Bias, Bias bei der Berichterstattung, Bias bei der Probenahme, Bias bei der Auswahl, Vorhersage-Abweichung (beim maschinellen Lernen).

3. Ein Wert, der angibt, wie weit der Durchschnitt der Vorhersagen vom Durchschnitt der Kennzeichnungen im Datensatz entfernt ist.

E

Equalized odds *(Ausgeglichene Wahrscheinlichkeiten)*
Ein Maß für die Fairness im Bereich des maschinellen Lernens, das die Frage beantwortet: Werden alle Werte für eine bestimmte Kategorie oder ein bestimmtes Attribut gleichermaßen gut vorhergesagt?

F

Falsch-negative · Fehler
Tritt auf, wenn das Modell fälschlicherweise die negative Klasse vorhergesagt hat. Beispielsweise schloss das Modell daraus, dass eine bestimmte E-Mail-Nachricht kein Spam (die negative Klasse) war, obwohl diese E-Mail-Nachricht tatsächlich Spam war.

Falsch-positive · Fehler
Tritt auf, wenn das Modell fälschlicherweise die positive Klasse vorhersagt. Beispielsweise schloss das Modell daraus, dass eine bestimmte E-Mail-Nachricht Spam (die positive Klasse) sei, diese E-Mail-Nachricht war aber kein Spam.

E

equalized odds
A fairness metric that checks if, for any particular label and attribute, a classifier predicts that label equally well for all values of that attribute.

F

false-negative error
An example in which the model mistakenly predicted the negative class. For example, the model inferred that a particular email message was not spam (the negative class), but that email message actually was spam.

false-positive error
An example in which the model mistakenly predicted the positive class. For example, the model inferred that a particular email message was spam (the positive class), but that email message was actually not spam.

P

post-processing
Processing the output of a model after the model has been run. Post-processing can be used to enforce fairness constraints without modifying models themselves.

predictive parity
A fairness metric that checks whether, for a given classifier, the precision rates are equivalent for subgroups under consideration.
Predictive parity is sometime also called predictive rate parity.

S

sensitive attributes
A human attribute that may be given special consideration for legal, ethical, social, or personal reasons.

G

Genauigkeitsgrad · Genauigkeit
Der Anteil der Vorhersagen, bei denen ein Klassifizierungsmodell richtig lag. Bei der Mehrklassenklassifizierung wird die Genauigkeit wie folgt definiert:
Genauigkeit = Korrekte Prognosen / Gesamtzahl der Beispiele

In der binären Klassifikation hat die Genauigkeit folgende Definition:
Genauigkeit = wahr positiv + wahr negativ / Gesamtzahl der Beispiele

K

Klassifizierungsmodell
Eine Art des maschinellen Lernmodells zur Unterscheidung zwischen zwei oder mehr einzelnen Klassen. Ein Klassifizierungsmodell für die Verarbeitung natürlicher Sprache könnte zum Beispiel bestimmen, ob ein Eingabesatz in Französisch, Spanisch oder Italienisch gesprochen wurde.

Kontrafaktische Fairness
Fairness-Metrik, die prüft, ob ein Klassifikator für ein Individuum das gleiche Ergebnis liefert wie für ein anderes Individuum, das mit dem ersten identisch ist, außer in Bezug auf ein oder mehrere sensible/schützenswerte Attribute. Die Bewertung eines Klassifikators auf Kontrafaktische Fairness ist eine Methode, um potenzielle Quellen von Voreingenommenheit in einem Modell aufzudecken.

P

Post-processing
Verarbeitung des Ergebnisses eines Modells, nachdem das Modell ausgeführt wurde. Die Nachbearbeitung kann zur Durchsetzung von Fairnessbeschränkungen verwendet werden, ohne das Modell selbst zu modifizieren.

Prognostische Fairness
Eine Fairness-Metrik, mit der überprüft wird, ob für einen bestimmten Klassifikator die Präzisionsraten für die betrachteten Untergruppen gleichwertig sind.

S

Sensible Merkmale
Menschliche Beschreibungsmerkmale, mit denen aus rechtlichen, ethischen, sozialen oder persönlichen Gründen besonders sensibel umgegangen werden muss. Zu diesen gehören: Rasse, Geschlecht, Alter, sexuelle Orientierung, Behinderung und andere.

Imprint

Impressum

Further information and contacts: /
Weitere Informationen und Kontakte:

Deutsche Telekom AG
Group Compliance Management
Friedrich-Ebert-Allee 140
53113 Bonn

Digital.Ethics@Telekom.de

https://www.telekom.com/de/konzern/digitale-verantwortung/ethische-ki-leitlinien-der-telekom

Bonn, 31.03.2021

Authors: /
Autor*innen:

Manuela Mackert
Manuela.Mackert@Telekom.de

Maike Scholz
Maike.Scholz01@Telekom.de

Manuel Mikoleit
M.Mikoleit@Telekom.de