KI Bundesverband e.V.

KI Gütesiegel

Berlin, 03.03.2019

Zusammenfassung

Das KI Gütesiegel des KI Bundesverband e.V. verfolgt das Ziel einen menschen-zentrierten und menschen-dienlichen Einsatz von Künstlicher Intelligenz zu sichern. Durch das Definieren und Einhalten von einem übergeordneten Werte- und Prozessverständnis stellt das Gütesiegel eine ethisch verträgliche Service- und Produktentwicklung sicher. Im Zentrum stehen die Gütekriterien Ethik, Unvoreingenommenheit, Transparenz sowie Sicherheit und Datenschutz. Zu jedem Gütekriterium sind notwendige Maßnahmen festgehalten. Zum Zeitpunkt der Einführung umfasst das Gütesiegel eine Selbstverpflichtungserklärung.

Inhaltsverzeichnis

Präambel

Der KI Bundesverband hat sich zum Ziel gesetzt, einen menschen-zentrierten und menschen-dienlichen Einsatz von Künstlicher Intelligenz (KI) zu fördern.

Im Positionspapier „KI Situation und Maßnahmenkatalog“ vom 25.06.2018 definierte der KI Bundesverband über hundert einzelne Maßnahmen um den Einsatz von Künstlicher Intelligenz in den Bereichen **Bildung & Forschung**, **Wirtschaft & Infrastruktur** sowie **Arbeit & Gesellschaft & Recht** nachhaltig zu fördern.

Zum Aufbau exzellenter Rahmenbedingungen sehen die Mitglieder des KI Bundesverbandes sich auch selbst in der Pflicht, ein übergeordnetes Werte- und Prozessverständnis, das in der Entwicklung von KI berücksichtigt wird, zu schaffen. Eine erste Maßnahme ist das KI-Gütesiegel.

Motivation

Der Einsatz von Künstlicher Intelligenz (KI) prägt den technologischen Fortschritt wesentlich und wird dafür sorgen, dass gesellschaftliche und wirtschaftliche Strukturen grundlegend verändert werden. Viele Dimensionen des gesellschaftlichen Zusammenlebens sind vom Einsatz künstlicher Intelligenz betroffen, was gleichsam zu ethischen als auch zu ökonomischen sowie sicherheitsrelevanten Fragestellungen führt.
Mit dem KI-Gütesiegel haben vor allem deutsche Unternehmen die Möglichkeit, sich auf die Einhaltung grundlegender Qualitätsparameter berufen zu können.

Ziele

Das Ziel ist die Etablierung eines Gütesiegels für eine ethisch verträgliche Service- und Produktentwicklung, um Vertrauen in der Gesellschaft aufzubauen und die Wettbewerbsfähigkeit der beteiligten Unternehmen international zu stärken.
Das Gütesiegel arbeitet mit den Gütekriterien Ethik, Bias, Transparenz sowie Sicherheit und Datenschutz.

Kontext und Abgrenzung

Das Gütesiegel des KI-Verbands ist ein Indikator dafür, dass die Produkte und Dienstleistungen eines Unternehmens branchenübliche Qualitätsanforderungen erfüllen und nach anerkannten Regeln der Technik entwickelt werden.
Das Gütesiegel ersetzt keine anerkannten Vorgehensmodelle oder Zertifizierungen wie ITIL, COBIT oder ISO-20000, sondern beschränkt sich weitest möglich auf KI-spezifische Probleme und Lösungen, wobei für die meisten KI-Anwendungen zugleich die Anforderungen an den Betrieb herkömmlicher IT-Systeme gelten und beispielsweise in der Regel ein Mindestmaß an IT-Sicherheit (z.B. IT-Grundschutz) gewährleistet sein muss, um das Gütesiegel zu erlangen.

Gütekriterien

Die vier Grundsätze des Gütesiegels bilden die Gütekriterien Ethik, Unvoreingenommenheit, Transparenz sowie Sicherheit und Datenschutz.

1. Ethik

Die Entwicklung und Anwendung von Künstlicher Intelligenz erfolgt menschenzentriert und nach den europäischen Grundwerten: Menschenwürde, Freiheit, Demokratie, Gleichheit, Rechtsstaatlichkeit. Die Werte zeichnen sich durch Pluralismus, Nichtdiskriminierung, Toleranz, Gerechtigkeit, Solidarität sowie der Gleichheit der Geschlechter aus.

Menschenzentrierte KI bedeutet, dass wirtschaftlicher Nutzen unter Wahrung der europäischen Grundwerte erfolgt und in Mensch-Maschine-Interaktionen der Mensch immer die Möglichkeit hat zu intervenieren oder ein System anhalten oder unterbrechen kann.

Künstliche Intelligenz soll zu mehr Freiheit, Gleichheit, Gerechtigkeit, Solidarität, Toleranz und Pluralismus führen und es ist sicher zu stellen, dass sie nicht zur Diskriminierung oder gegen Demokratie oder Menschenrechte genutzt wird.

In der Qualifizierung im Rahmen des KI-Gütesiegels ist sicherzustellen, dass:

- hinsichtlich der Implementierung in ein Produkt die o.g. gesellschaftlichen, rechtlichen und ethischen Grundlagen berücksichtigt werden.
- das einsetzende Unternehmen sich zu den genannten Werten im Zusammenhang des Grundgesetzes und der UN-Menschenrechtskonvention bekennt.
- in der Interaktion Mensch-Maschine die Entscheidungsprozesse durch menschliche Intervention so beeinflussbar sind, dass maschinelle Abläufe jederzeit angehalten und im notwendigen Falle abgeschaltet werden können.

2. Unvoreingenommenheit (Bias)

Bias bezeichnet eine verzerrte Wahrnehmung, z.B. aufgrund eigener Vorurteile, Überzeugungen und Lebenserfahrungen. Diese Verzerrung der eigenen Wahrnehmung kann sich in technischen Systemen widerspiegeln. Die transparente Auseinandersetzung mit Bias ist ein Gütekriterium.

Bias wird hauptsächlich durch zwei Faktoren verursacht:

- durch ungeeignete Datenauswahlprozesse
- durch fehlerhafte Verfahren und Algorithmen

Die Datenauswahlprozesse beziehen sich auf das Sammeln von Daten, die verwendet werden, um KI-Systeme zu trainieren. Dabei ist die Auswahl der Daten eine der wesentlichen Komponenten. Wird diese falsch getroffen, werden alte Stereotypen und Vorurteile, die bereits in den Daten vorhanden sind, durch das KI System gelernt und dabei verstärkt.
Es ist sicherzustellen, dass mit geschultem Personal gearbeitet wird und bekannte Datenanalyseverfahren zur Erkennung von Bias angewandt werden.
Die Ergebnisse der Bias-Erkennung sind in regelmäßigen Abständen zu dokumentieren, unabhängig davon, wie ein Unternehmen sich entscheidet mit dem Bias umzugehen.

Im Rahmen der Erteilung des Gütesiegels ist deshalb eine Verfahrensbeschreibung empfohlen, durch welche eine erhöhte Transparenz ermöglicht wird und so letztlich festgestellt werden kann, ob Verfahren und Zweck einander bedingen. Es ist zudem angeraten zu prüfen, ob die Auswahl der Verfahren und Algorithmen hinsichtlich des Einsatzzweckes und des Nutzens sinnvoll und zielführend sind.

3. Transparenz

Im Rahmen eines transparenten Entwicklungsprozesses wird das folgende generische Vorgehensmodell empfohlen:

- Datenvorverarbeitung
  Die Datenvorverarbeitung beinhaltet je nach Problemstellung und Datentyp verschiedenste und teilweise mehrstufige Aufbereitungsarten. Jeder Schritt in Bezug auf die Datenvorverarbeitung ist im Rahmen eines transparenten Vorgehens zu dokumentieren. Dies geschieht in erster Linie zur Nachvollziehbarkeit aber auch zur Transparenz über etwaige nicht verwendete Daten (Erkennen und Entfernen von Ausreißern).
  Eine gute Dokumentation der Datenvorverarbeitung ist weiter zur KI-Modellbereitstellung ratsam, um neue Daten ebenso für das erstelle KI-Modell aufzubereiten.

- Einflussgrößenauswahl (Feature Engineering)
  Es ist sicherzustellen, dass die Einflussparameter (Features) in Bezug auf die Aufgabenstellung geeignet gewählt sind. Darin inbegriffen sind Maßnahmen zur Featureauswahl als auch evtl. vorgelagert zur Reduktion von Featuredimensionen.

- KI-Modellerstellung und KI-Modellevaluation
  Die KI-Modellerstellung soll durch geeignete Maßnahmen (Verfahrenspipelines und Kreuzvalidierung) einen generalisierenden Charakter haben. Eine Überanpassung soll grundsätzlich vermieden werden.
  Die Modellevaluation soll je nach Problemstellung und KI-Verfahren geeignete Metriken zur Bewertung der KI-Modellgüte verwenden.

- Nachvollziehbare KI-Modelle
  KI-Verfahren sollen von bekannten und geeigneten Analyseverfahren begleitet werden. Damit soll ein Einblick in die Einflussfaktoren, die von einem KI-Modell gelernt werden, gewährt werden, um jederzeit das Ergebnis eines KI-Modells nachvollziehen zu können. Das Ziel ist, sogenannte Black-Box Systemarchitekturen in vor allem sicherheitsrelevante Anwendungen zu vermeiden.

  Die genannten Punkte sind bei der Erstellung von KI-Produkten und Projekten zu berücksichtigen und, falls nicht explizite Gründe dagegen sprechen, anzuwenden.

4. Sicherheit und Datenschutz

Ein KI-System verarbeitet Daten ebenso wie klassische Datenverarbeitungssysteme. Daher gibt es zunächst die gleichen Anforderungen an die Vertraulichkeit und Integrität der Datenverarbeitung. Darüber hinaus entstehen durch die besondere Natur lernender Maschinen zusätzliche Anforderungen.

Hierzu gehören:

- Sicheres und datenschutzkonformes Sammeln, Verwahren und Verarbeiten großer Mengen von Trainingsdaten
- Robustheit gegenüber feindseligen Eingabedaten (Adversarial Inputs)
- Verhindern der Degeneration und Manipulation selbstlernender Systeme
- Absicherung gegen neue, menschenunübliche Versagensarten (Failure Modes der KI)
- Unberechtigter Zugriff auf personenbezogene Daten durch Ausforschung neuronaler Netze

Im Gegensatz zu “klassischen” IT-Systemen werden KI-Systeme öfter für Zwecke eingesetzt, die bisher menschlicher Tätigkeit vorbehalten waren und oft direkte Schnittstellen zur analogen Welt innehaben. Diese müssen dabei mit unpräzisen Eingabedaten wie Bildern, Geräuschen oder natürlicher Sprache umgehen. Eine derartige Verarbeitung ist prinzipbedingt fehlerbehaftet und offen für unterschiedliche Interpretationen.
Es wird empfohlen zu dokumentieren, welche zusätzliche Anforderungen durch den Einsatz eines spezifischen KI-System für einen spezifischen Zweck entstehen und wie mit ihnen umgegangen wird.

Anhang/Referenzen

http://www.bmc.com/guides/itil-cobit-introduction.html

Autoren

Im Folgenden werden die Autoren in alphabetischer Reihenfolge genannt.

Marc Engenhart (Engenhart ° Design Studio)
Sebastian Eumann (Sebastian Eumann Consulting)
Silviu Homoceanu (BrainPlug)
Tobias Jerzewski (Aspen Institute Germany)
Benedikt Kämpgen (Empolis)
Ansprechpartner: Amadeusz Kargul (kahaura GmbH)
Gero Nagel (Tognos)
Lydia Nemec (privat)
Pavel Mayer (Tognos)
Florian Werner-Jäger (infofeld GmbH)