# Vertrauenswürdige KI-WebAPI Spezifikationen

Hochschule für Wirtschaft und Recht Berlin
Berlin School of Economics and Law

OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG

Sandro Hartenstein, M.Sc.
sandro.hartenstein@hwr-berlin.de
www.hwr-berlin.de 15.11.2022

## Motivation

Die Vertrauenswürdigkeit von Onlinediensten ist essentiell für eine nachhaltige Digitalisierung, da die Akzeptanz beim Konsumenten die Nutzungsquoten maßgeblich bestimmen. Aus diesem Grund wird von allen Stakeholdern eine hohe Vertrauenswürdigkeit der Angebote angestrebt. Eine adäquate Kommunikation von vertrauensschaffenden Maßnahmen und Implementierungen von Best Practices zum Dienstnutzer sind wichtige Faktoren. Diese Untersuchung stellt einen Ansatz vor, der vertrauenswürdige Aspekte in den Fachspezifikationen der technischen Umsetzung von Onlinediensten (WebAPIs) verorten möchte. Insbesondere KI-Algorithmen sind im Fokus, da Attribute wie Fairness und Transparenz besondere Relevanz haben.

**Frage**: Kann die OpenAPI Spezifikation Anforderungen an vertrauenswürdige KI-WebAPIs abbilden?

**Hypothese**: OpenAPI Spezifikationen bilden derzeit nur technische Attribute der Vertrauenswürdigkeit ab.

**Ziel**: Eine Erweiterung von OpenAPI für vertrauensschaffende Attribute ist das Ziel.

## Methode

Vertrauenswürdige Attribute können mithilfe von Indikatoren und Metriken gemessen und bewertet werden [1]. Diese quantitative Analyse gibt ein schnelles objektives Bild des einzelnen Attributes, allerdings erfordern einige weichere Attribute qualitative Analysen. Es wird eine **GAP-Analyse** verwendet, um die Möglichkeiten der OpenAPI Spezifikationen im Kontext der Beschreibung von vertrauenswürdigen WebAPIs zu identifizieren:

1. Hierzu werden die idealen **Soll-Anforderungen** aus der Literatur in Form der fachlichen Aspekte zusammengetragen. Konkret sind die Attribute vertrauenswürdiger WebAPIs, siehe Abbildung 1, mithilfe der Spezifikationen messbar bereitzustellen.
2. Die derzeit aus den Spezifikationen ermittelbaren vertrauenswürdigen Attribute stellen den **Ist- Stand** dar und werden aus den technischen Aspekten der OpenAPI Spezifikation extrahiert [2].
3. Ein anschließender **Soll - Ist Abgleich** liefert die Lücken, die für eine vertrauenswürdige OpenAPI Spezifikation geschlossen werden sollten.
4. Durch die Identifikation der notwendigen Metriken und Indikatoren können konkrete Vorschläge für die benötigten Daten bereitgestellt werden.

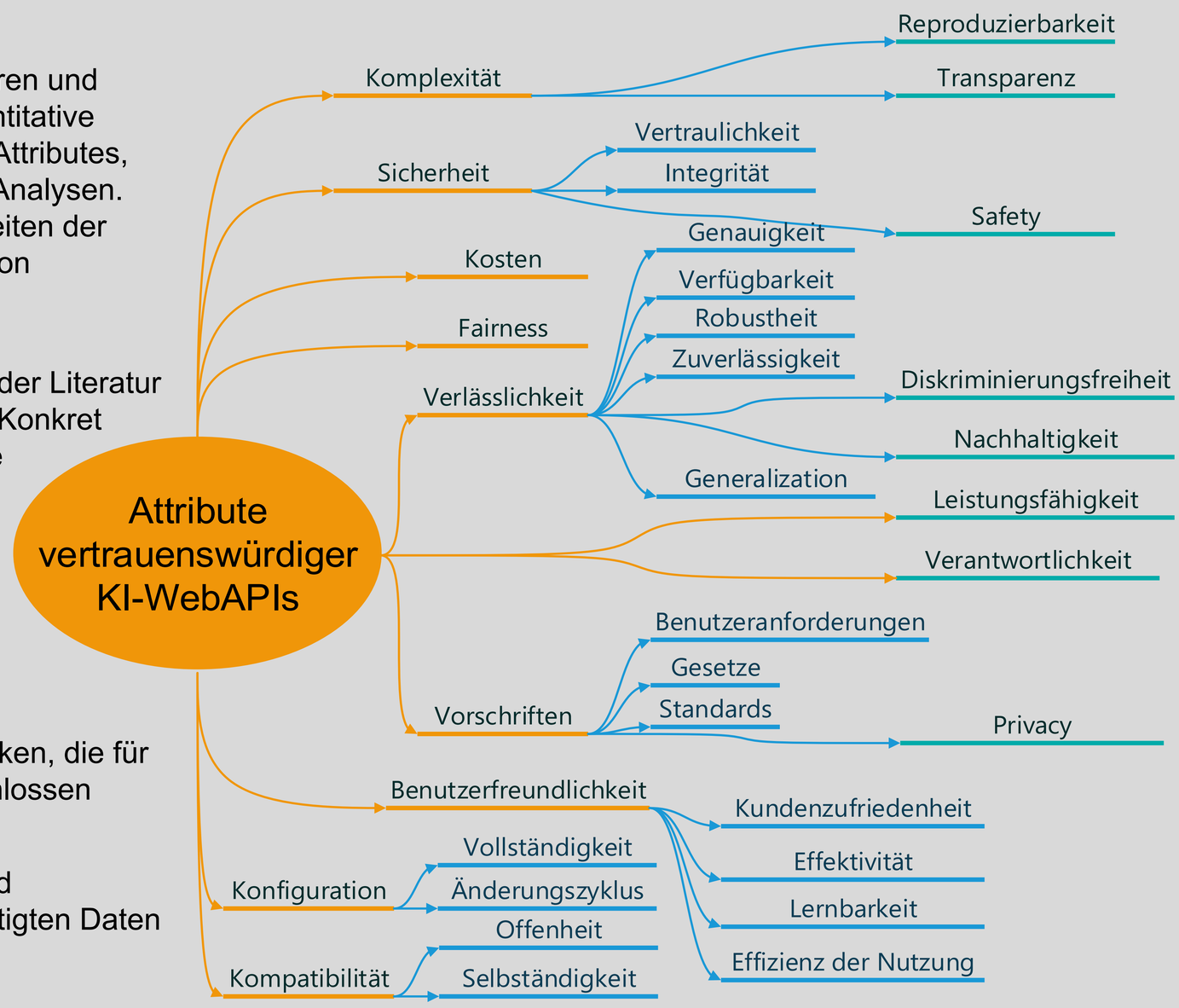


**Abbildung 1 Attribute vertrauenswürdiger KI-WebAPIs**

## Ergebnisse

Mithilfe der GAP-Analyse konnten die relevanten Lücken identifiziert werden. Ausgewählte Attribute und Indikatoren zur Bestimmung der Vertrauenswürdigkeit von KI-WebAPIs , sowie Ansätze zur Integration sind in Tabelle 1 aufgeführt.

**Tabelle 1 Ausgewählte Lücken der OpenAPI -Spezifikation für vertrauenswürdige KI-WebAPIs**

| Attribut | Fehlende Indikatoren | Ansätze zur Integration |
|---|---|---|
| **Safety** | Kritische Funktionalitäten im Kontext, z.B. Bilderkennung in Automotive (kritisch) vs. Bilderkennung zur Authentifikation (nicht kritisch) | Safety kritische Funktionen sollten vom Anbieter markiert werden |
| **Transparenz [3] S. 8** | Externe Verarbeitung, Qualität der KI-Dokumentation | Informationen über die weitere Verarbeitung sollten durch den Anbieter bereitgestellt werden. |
| **Reproduzierbarkeit [3] S. 8–9** | Grad der Reproduzierbarkeit (mit gleicher Ausgangsparametern) | Angabe des Anbieters über den Einfluss von externen Daten auf das Ergebnis. |
| **Leistungsfähigkeit** | Antwortzeiten, Durchsatz, Laufzeit | Angaben der breitgestellten Leistungen sollten vom Anbieter pro Funktion bereitgestellt werden |
| **Verantwortlichkeit** | Transparenz- und Testindikatoren (Beispiel werte, Erwartbare Ergebnisse, Zwischenergebnisse, externe Datenquellen) | Indikatoren zu Transparenz und Rückverfolgbarkeit sollte vom Anbieter angegeben werden |
| **Privacy** | Rechtsgrundlagen, technische Umsetzungen (Privacy by Design), Dokumentation der Verarbeitung | Informationen über die weitere Verarbeitung sollten durch den Anbieter bereitgestellt werden. |
| **Diskriminierungs-freiheit [4] S. 10** | sensible Variablen (z.B. Geschlecht, Rasse), erwartete Systemreaktionen (vorurteilsfreie Ergebnismenge) wird regelmäßig auditiert | Angaben über Filterung von sensiblen Variablen vor der Verarbeitung oder erwarteten Systemreaktion sollten vom Anbieter angegeben werden. |

## Fazit & Ausblick

Die Betrachtung der Architektur von OpenAPI hat ergeben: Mit der OpenAPI-Spezifikation ist es möglich, die vertrauenswürdigen Anforderungen an KI-WebAPIs abzubilden.

Aus den Ergebnissen der Untersuchung geht hervor, dass die OpenAPI Spezifikation derzeit vorrangig die technischen Aspekte der Vertrauenswürdigkeit adressieren.

Für das Ziel sind die folgenden weitere Schritte geplant:

1. Formal korrekte Formulierungen und Verortungen in der Spezifikation
2. Einreichung als Feature Request bei der OpenAPI Community.

Die Anbieter sollten diese Chance der Attraktivität durch Transparenz der vertrauensschaffenden Aspekte nutzen, um die Akzeptanz ihrer Services zu verbessern. Insbesondere KI-Algorithmen werden in vielen Anwendungsbereichen sehr kritisch betrachtet. Eine Erweiterung der OpenAPI-Spezifikation unterstützt dies.

### Quellen

[1] HARTENSTEIN, Sandro: Toolunterstütztes Messen der Vertrauenswürdigkeit von Webapplikationen, Bd. 15. In: SCHMIETENDORF, Andreas; KUNISCH, Matthias (Hrsg.): BSOA : 10. Workshop Bewertungsaspekte service- und cloudbasierter Architekturen, 03. November 2015, Leipzig. 1. Aufl. Herzogenrath : Shaker, 2015 (Berliner Schriften zu modernen In-tegrationsarchitekturen, 15), S. 85–99
[2] OPENAPI INITIATIVE: OpenAPI Specification. URL https://spec.openapis.org/oas/v3.1.0.
[3] LI, Bo ; QI, Peng ; LIU, Bo ; DI SHUAI ; LIU, Jingen ; PEI, Jiquan ; YI, Jinfeng ; ZHOU, Bowen: Trustworthy AI: From Principles to Practices.
[4] GILLESPIE, Nicole ; CURTIS, Caitlin ; BIANCHI, Rossana ; AKBARI, Ali ; VAN FENTENER VLISSINGEN, Rita: Achieving Trustworthy AI: A Model for Trustworthy Artificial Intelligence. Australia, 2020