



WHITEPAPER

SKALIERBARE PRÜFUNG VON KI-SYSTEMEN MIT SMART STANDARDS: SZENARIEN AUS ANWENDERSICHT

Autoren

Dr. Philipp Heß, Raymond Puppen (DKE)

Ansprechpartner

Janos Koschwitz (DKE)

Zielgruppe & Hintergrund

Dieses Whitepaper richtet sich an Anwenderinnen und Anwender von Normen im Bereich KI, Beteiligte im Bereich KI-Qualität und -Prüfung, Interessierte am Thema SMART Standards, sowie sonstige Entscheidungstragende, die eine digitale Qualitäts- bzw. Prüfinfrastruktur für KI-Systeme mitgestalten wollen.

Es ist das Ergebnis von Interviews und dem Austausch in der Gruppe „Skalierbare KI-Prüfung“ (SKIP) der Initiative Digitale Standards (IDiS) von DIN und DKE. SKIP wurde von 2024 bis 2025 im Kontext des Projekts MISSION KI (acatech, BMDV/BMDS) durchgeführt.

Danksagung

Wir danken allen Expertinnen und Experten, die durch Interviews und fachlichen Austausch zur Entstehung dieses Whitepapers beigetragen haben.

Ein besonderer Dank gilt den Teilnehmenden der Gruppe „Skalierbare KI-Prüfung“ (SKIP) und des Auftaktworkshops im November 2024.

Für ihre Beiträge im Rahmen der Konzeption und des inhaltlichen Reviews danken wir insbesondere:

Martina Paul (Huawei)

Dr.-Ing. Marc P. Hauer (TÜV AI.Lab)

Achim Pascal Meyer (sphin-X e.V.)

Taras Holoyad (BNetzA)

sowie weiteren Beteiligten von DKE, DIN, QI-Digital und dem AI Quality & Testing Hub (AIQ).

HERAUSGEBER



DIN e. V.

Burggrafenstraße 6
10787 Berlin
Tel.: +49 30 2601-0
E-Mail: presse@din.de
Internet: www.din.de



VDE e. V.

DKE Deutsche Kommission Elektrotechnik
Elektronik Informationstechnik

Merianstraße 28
63069 Offenbach am Main
Telefon: +49 69 6308-0
Telefax: +49 69 6308-9863
E-Mail: standardisierung@vde.com
Internet: www.dke.de

Fotonachweis: Mit generativer KI (DALL-E) erstellt (Bereitgestellt von Philipp Heß)

Stand: August 2026

MANAGEMENT SUMMARY	4
IDIS - DIE INITIATIVE DIGITALE STANDARDS VON DIN UND DKE	5
1 EINLEITUNG	8
2 HINTERGRUND	12
2.1 Überblick: Prüfung von Künstlicher Intelligenz auf Basis von Normen und Standards	12
2.2 Skalierbare Prüfung von KI-Systemen	17
3 SZENARIEN FÜR DEN EINSATZ VON SMART STANDARDS BEI DER PRÜFUNG VON KI-SYSTEMEN	24
3.1 API-basierte Auslieferung von maschinenles- und interpretierbaren Norminhalten	26
3.2 Zuverlässig KI-generierte Prüferfakte	29
3.3 Einbindung in ein Ökosystem für KI-Qualität	31
3.4 Skalierbare Generierung von SMART Content und Operationalisierungsinhalten	38
4 ANWENDUNGSBEISPIEL	41
4.1 Szenario 1: Einsatz von maschinenlesbaren Norminhalten per APIs	42
4.2 Szenario 2: Generierung von KI-Prüfungen	45
4.3 Szenario 3: Einbindung in ein Ökosystem für KI-Qualität	48
5 DISKUSSION	52
5.1 Bewertung der Szenarien	52
5.2 Einordnung in zentrale Konzepte der Digitalisierung der Normung	54
5.3 Übertragbarkeit	56
6 FAZIT	59
ABKÜRZUNGEN	61

MANAGEMENT SUMMARY

Bedarf: Prüfungen von KI-Systemen werden zur Daueraufgabe

Mit der zunehmenden Bedeutung von KI-basierten Produkten und Dienstleistungen wachsen die Anforderungen an Sicherheit, Zuverlässigkeit und Transparenz, die in der Praxis über systematische Prüfungen auf Basis von Normen wirksam werden. Auch gerade die europäische KI-Verordnung und weitere horizontale und sektorale Rechtsakte machen normbasierte Prüfungen von KI-Systemen zu einem dauerhaft mitlaufenden Bestandteil von Entwicklung und Betrieb.

Herausforderungen: häufige Prüfung, häufige Updates, komplexe Normenlandschaft

Drei Faktoren treiben den Gesamtaufwand: die Anzahl zu prüfender Systeme, die Häufigkeit ihrer Veränderungen durch Modellaktualisierungen und Datenänderungen sowie die Komplexität einer sich entwickelnden Gesetzes- und Normenlandschaft. Damit Innovation und Qualitätssicherung in Einklang bleiben, müssen Prüfungen so weit wie möglich automatisiert und verbleibende manuelle Tätigkeiten durch passgenau bereitgestellte Informationen beschleunigt werden.

SMART Standards machen Norminhalte automatisiert nutzbar

SMART Standards stellen Norminhalte als granulare, maschinenlesbare und maschineninterpretierbare Ressourcen bereit. Erst dadurch lässt sich normbasierte Prüfung von KI-Systemen skalierbar gestalten. In Verbindung mit standardisierten Prüfwerkzeugen, Referenzdatensätzen, Prüfplattformen und einer datengetriebenen Qualitätsinfrastruktur können wiederkehrende Prüfkaktivitäten in die laufenden Entwicklungs- und Betriebsabläufe von KI-Systemen integriert, teilweise automatisiert und reproduzierbar dokumentiert werden.

Vom Norm-Bezug per API zum Ökosystem für KI-Qualität

Drei Szenarien beschreiben einen stufenweisen Weg von der heutigen manuellen Praxis zu einer zunehmend automatisierten Prüfumgebung. Die beschriebenen Szenarien adressieren die technische Unterstützung und Teilautomatisierung von Prüfkaktivitäten. Sie ersetzen weder die fachliche Bewertung noch die Verantwortung von Prüfenden, Audi-

toren oder Konformitätsbewertungsstellen. Anforderungen an Unabhängigkeit, Vertraulichkeit, risikobasierte Entscheidungsfindung und weitere organisatorische Rahmenbedingungen bleiben vom jeweiligen Prüfkontext abhängig und werden durch die dargestellten technischen Konzepte nicht ersetzt.

1. Szenario: API-basierte Auslieferung von Norminhalten.

SMART Standards werden über Software-Schnittstellen in Anforderungsverwaltung und Entwicklungswerkzeuge integriert, sodass Norminhalte mit eindeutigen Verweisen direkt in Entwicklung und Test einfließen. Aktualisierungen werden automatisch erkannt, betroffene Anforderungen markiert und Konformitätsberichte automatisiert aus den dokumentierten Prüfschritten abgeleitet. Dieses Szenario könnte sich mit heute verfügbaren Komponenten grundsätzlich bereits umsetzen lassen.

2. Szenario: Zuverlässig KI-generierte Prüfartefakte.

Aufbauend auf Szenario 1 übernimmt ein KI-Assistent die Steuerung: Er ermittelt einschlägige Anforderungen, schlägt Prüfpläne vor und generiert Prüfcode. Norminhalte wie Grenzwerte und Schwellen werden dabei unverändert aus Normen übernommen und durchlaufen das Sprachmodell nicht, sodass sie auf ihren Normabschnitt zurückführbar bleiben. Voraussetzung für die vollständige Umsetzung ist die Verfügbarkeit neuer Arten von SMART Content.

3. Szenario: Einbindung in ein Ökosystem für KI-Qualität.

Norminhalte, Daten und Prüfwerkzeuge und -services sind als souveräne Ressourcen in einem gemeinsamen Datenraum verfügbar. Konformitätsbescheinigungen werden als digitale Nachweise mit dem KI-System mitgeführt. Dieses Szenario ist langfristig angelegt und setzt umfangreiche Verständigungen über die Regeln eines solchen Datenraums voraus.

IDIS - DIE INITIATIVE DIGITALE STANDARDS VON DIN UND DKE

Hintergrund und Ziel

Das Deutsche Institut für Normung (DIN) und die Deutsche Kommission Elektrotechnik Elektronik Informationstechnik (DKE) treiben seit 2020 mit der Initiative Digitale Standards (IDiS) die digitale Transformation der Normung voran. IDiS fungiert als Bindeglied zwischen Normerstellung und digitaler Normanwendung in Form einer offenen Community-Plattform, die Unternehmen, Verbände, Behörden und Wissenschaftseinrichtungen vereint und konkrete Anforderungen an die digitale Norm erarbeitet. Initiiert wurde IDiS als direkte Antwort auf eine grundlegende Lücke: Während die Industrie ihre Prozesse zunehmend automatisiert und digitale Requirements-Engineering-Systeme einsetzt, lagen Normen weitgehend in dokumentenbasierten Formaten vor, die für eine maschinengestützte Verarbeitung nicht geeignet waren.

SMART Standards: Kern von IDiS

Im Zentrum der Aktivitäten stehen SMART Standards: Normen, deren Inhalte sowohl für Menschen als auch für Maschinen lesbar, anwendbar und digital bereitstellbar sind. IDiS verfolgt dabei ein stufenweises Reifegradmodell (Utility-Stufen Level 2 bis 5), das von erster digitaler Bereitstellung von Norminhalten bis hin zur vollständig automatisierten, maschineninterpretierbaren Normanwendung reicht.

Beteiligte und Arbeitsstruktur

IDiS vereint über 130 Teilnehmer aus Industrie, Wirtschaft, Verbänden und der Wissenschaft. Inhaltlich arbeitet die Initiative in Ad-hoc-Gruppen und Pilotprojekten, deren Ergebnisse in Whitepapers und Spezifikationen einfließen. Zweimal jährlich trifft sich das IDiS-Plenum zum übergreifenden Austausch. International anschlussfähig wird die Arbeit von IDiS durch die Koordination mit CEN/CENELEC und ISO/IEC, sowie nationalen Normungsorganisationen wie AFNOR.

Digitale Services und Werkzeuge von DIN und DKE

DIN und DKE treiben die Digitalisierung ihrer eigenen Prozesse und Produkte voran. Mit dem Online Standards Development Tool (OSD) steht bei ISO/IEC seit 2020 eine gemeinsame Plattform zur Verfügung, deren breite Einführung seit 2023 läuft und die ab 2025 der Default für neue Normungsprojekte ist. OSD ermöglicht die strukturierte, XML-basierte

Erstellung von Normen und schafft damit die Grundlage für die schrittweise Anreicherung der Inhalte zu SMART Content.

Ergänzend dazu bauen DIN und DKE ein wachsendes Portfolio digitaler Services entlang des gesamten Lebenszyklus einer Norm auf, von der Erstellung über die Qualitätssicherung bis zur Anwendung. Dazu gehören KI-gestützte Werkzeuge zur Unterstützung von Normungsexpertinnen und -experten in Recherche und Redaktion, Services zur Sicherung terminologischer Konsistenz über Normengrenzen hinweg sowie API-basierte Zugänge, die es Unternehmen ermöglichen, Normen und normungsbezogene Prozesse direkt in ihre eigenen Systemlandschaften zu integrieren. Gemeinsam bilden diese Bausteine die Brücke von klassischen Normungsprozessen hin zu einem durchgängig digitalen, maschinenlesbaren und nutzungsorientierten Normenwerk.

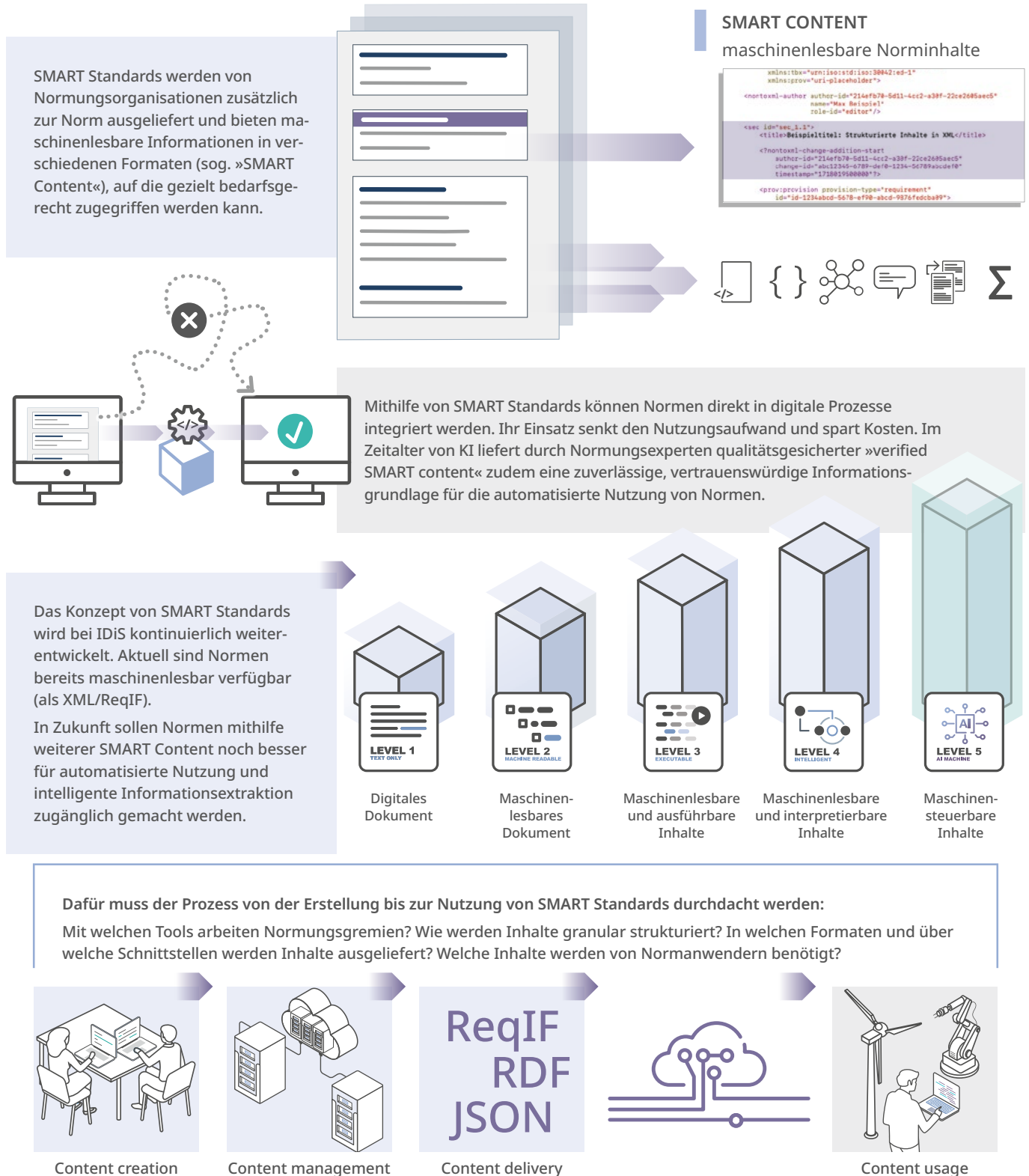
Wesentliche Outputs

Mehrere konkrete Arbeitsergebnisse sind bereits verfügbar. Die DIN DKE SPEC 99200 legt eine einheitliche ReqIF-Interpretation für öffentliche Normen fest und etabliert damit ein erstes standardisiertes Austauschformat für maschinenlesbare Norminhalte, das Unternehmen ihre Systeme SMART-Standards-ready machen lässt, einschließlich Änderungsmanagement auf Anforderungsebene. Abgeschlossene Pilotprojekte demonstrierten u.a. automatisierte Konformitätsprüfung und die Integration von Norminhalten in die Verwaltungsschale (AAS) für digitale Vorzertifizierungsprozesse.

Einbettung in die digitale Qualitätsinfrastruktur

SMART Standards sind in der QI-Digital-Roadmap als zentraler Baustein einer digitalen Qualitätsinfrastruktur positioniert. Sie ermöglichen die strukturierte und maschinenverarbeitbare Bereitstellung normativer Anforderungen und schaffen damit eine wichtige Voraussetzung für automatisierte Konformitätsprüfung und -dokumentation, digitale Zertifikate sowie die Verknüpfung normativer Anforderungen mit digitalen Produktinformationen. Ihre Nutzung setzt dabei geeignete Datenmodelle, Austauschformate und semantische Beschreibungen voraus. In diesem Sinne können SMART Standards auch eine wichtige Rolle in digitalen Produktdatenökosystemen wie dem Digitalen Produktpass (DPP) spielen, indem sie die normative Referenz für maschinenlesbar verknüpfte Produktinformationen und Compliance-Nachweise bereitstellen.

Was sind SMART Standards?



Weitere Informationen:

IDiS Whitepaper 1. Szenarien zur Digitalisierung der Normung und Normen
IDiS Whitepaper 2. Anwendungsfälle von SMART Standards
IDiS Whitepaper 3. Der betriebswirtschaftliche Nutzen von SMART Standards

 dke.de/idis

Abbildung 1

1

EINLEITUNG

1 EINLEITUNG

KI-basierte Produkte und Dienstleistungen zählen mit prognostizierten Wachstumsraten von über 30 Prozent pro Jahr zu den am schnellsten wachsenden Segmenten der Weltwirtschaft. Ihr Einsatz erstreckt sich von Medizin und Mobilität über Bildung bis hin zu Verwaltung und Industrie. Mit der zunehmenden gesellschaftlichen und wirtschaftlichen Bedeutung solcher Systeme wächst auch die Verantwortung, ihre Sicherheit und Zuverlässigkeit sicherzustellen.

Normen und Standards gewinnen bei der Gestaltung und Prüfung von KI-Systemen zunehmend an Bedeutung. Sie kodifizieren wissenschaftlich und industriell erarbeitetes Wissen, etablieren eine gemeinsame technische Sprache und schaffen Rahmenbedingungen, die die Qualität, Sicherheit und Verlässlichkeit von KI-Systemen sichern sollen.

Viele KI-Normen bilden dabei nicht nur technische, sondern auch gesellschaftliche Anforderungen ab, etwa hinsichtlich Nicht-Diskriminierung, Transparenz, menschlicher Aufsicht sowie dem Schutz von Sicherheit, Gesundheit und Umwelt. Diese Aspekte spiegeln politische Zielsetzungen und rechtliche Leitprinzipien wider, die über rein technische Qualitätsmerkmale hinausgehen.

Als Ergebnis umfangreicher Konsensfindungsprozesse werden Normen international zunehmend als Instrument staatlicher KI-Regulierung eingesetzt, etwa durch ihre Referenzierung im Rahmen der europäischen KI-Verordnung (KI-VO). Für viele Unternehmen dürfte die KI-VO und somit auch die Anwendung einschlägiger KI-Normen höchst relevant werden.

Auch die Prüfung von KI-Systemen auf Konformität mit Normen wird damit immer wichtiger. Dabei ist absehbar, dass die Prüfung von KI die betroffenen Entwickler und Betreiber vor besondere Herausforderungen stellen wird, insbesondere aufgrund nichtdeterministischer KI-Systeme, der starken Abhängigkeit von Datenqualität, der begrenzten Erklärbarkeit komplexer Modelle sowie der Dynamik lernender Systeme über ihren Lebenszyklus hinweg.

Auch gerade für das Anforderungsmanagement von KI-Systemen bzw. Produkten mit KI-Komponenten ergeben sich neue Herausforderungen. Dazu zählen insbesondere die Übersetzung abstrakter normativer Anforderungen in überprüfbare technische Kriterien, die kontinuierliche Überprüfung der Erfüllung von Anforderungen bei Modell- oder Datenänderungen, sowie die durchgängige Nachvollziehbarkeit, welche Norminhalte zu welchen technischen Anforderungen, welcher Implementierung und welchen Prüfnachweisen geführt haben (Traceability).

Dieses Whitepaper richtet sich an (potenzielle) Anwender von Normen im Bereich KI, Akteure im Bereich KI-Qualität und -Prüfung, sowie generell Interessierte zum Thema SMART Standards. Es soll dabei helfen, den bereits möglichen und den potenziellen Beitrag von SMART Standards zur Verbesserung der Skalierbarkeit von normbasierten Prüfungen von KI¹ besser verständlich zu machen. Gleichzeitig soll es als Diskussionsgrundlage für die Weiterentwicklung von SMART Standards im Kontext solcher Prüfungen dienen.

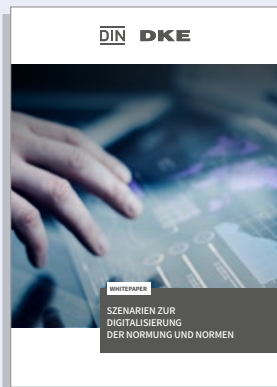
Folgende Fragen stehen im Mittelpunkt:

- Welchen Bedarf gibt es für die zukünftige Skalierbarkeit von normbasierter Prüfung von KI?
- Welchen Beitrag können SMART Standards potenziell in welchen Anwendungsfällen leisten?
- Welche Funktionen müssen SMART Standards dafür in Zukunft erfüllen?
- Wie kann das erreicht werden? Was muss dafür technisch umgesetzt werden, z.B. auch bei der Normerstellung, und in welchem Ökosystem sollen SMART Standards dafür eingesetzt werden?
- Was kann aus den Szenarien für den Einsatz von SMART Standards bei der Prüfung von KI für die Weiterentwicklung von SMART Standards insgesamt geschlossen werden?

Inhaltlich und strukturell baut dieses Whitepaper auf eine Reihe durch IDiS herausgegebene Texte und Pilotprojekte² auf, welche die konzeptuelle und praktische Weiterentwick-

1 Abgrenzung: Die genaue inhaltliche Ausgestaltung von KI- (Prüf-) Normen, oder die konkrete Anwendung und Operationalisierung einschlägiger KI-Normen, bzw. deren Umwandlung in SMART Standards wird in diesem Whitepaper dagegen nicht im Einzelfall diskutiert.

2 <https://www.dke.de/idis/pilotprojekte/roadmap>



Whitepaper 1



Whitepaper 2



Whitepaper 3



Whitepaper 4

lung von SMART Standards in den letzten Jahren vorange-
trieben haben.

Insbesondere das Whitepaper 1 „Szenarien zur Digitalisie-
rung der Normung und Normen“ (WP1), das Whitepaper
2 „Anwendungsfälle von SMART Standards“ (WP2) und
das Whitepaper 3 „Der Betriebswirtschaftliche Nutzen von
SMART Standards in Normenanwendungsprozessen“ (WP3)
bilden eine wichtige Grundlage.

Whitepaper 1 befasst sich mit Szenarien zur Digitalisierung
der Normung und Normen. Die Szenarien wurden anhand
der vier Wertschöpfungs-Prozessphasen Content Creation,
Content Management, Content Delivery und Content Usage
beschrieben und erläutern, welche Prozesse bei einer „Di-
gitalen Norm“ angepasst werden müssen. Das erweiterte
Utility-Modell beschreibt fünf Ausbaustufen digitalisierter
Normen, angefangen bei digitalen, aber unstrukturierten
Textdokumenten (Word, PDF), über vollständig maschinen-
interpretierbare Inhalte, bis hin zu einer Vision automati-
siert geprüft und durch eine KI überarbeiteter Normen.
Die Szenarien dieses Whitepapers lassen sich direkt darauf
abbilden und konkretisieren damit für die Prüfung von KI,
welche Beiträge zu welchen Utility-Stufen welchen prakti-
schen Nutzen bringen.

Whitepaper 2 befasst sich mit den Fragestellungen der
Anwendbarkeit von SMART Standards im Rahmen von An-
wendungsfällen. Diese beschreiben grundlegend, was im
Umgang mit SMART Standards bei Erstellung, Management,
Bereitstellung und Nutzung von Normen möglich sein wird.
Eine Vielzahl an Use Cases wurde in 11 Generic User Stories
(GUS) zusammengefasst, um generische Anwendungssze-
narien von SMART Standards zu beschreiben. Verschiedene

dieser GUS wurden in diesem Whitepaper auf den Anwen-
dungsfall „KI-Prüfung“ übertragen.

Whitepaper 3 befasst sich mit dem wirtschaftlichen Nutzen
von SMART Standards in der Anwendung von Normen. Es
analysiert, wie digitalisierte und maschinenlesbare Normen
in betriebliche Prozesse eingebettet werden können und
welche Effekte sich daraus für Unternehmen ergeben. An-
hand eines nutzenorientierten Modells werden Zeit-, Kosten-
und Qualitätsvorteile entlang typischer Normanwendungs-
prozesse aufgezeigt. Das Whitepaper zeigt, dass SMART
Standards insbesondere durch Automatisierung, reduzierte
manuelle Aufwände und geringere Fehlerquoten zu Effi-
zienzsteigerungen beitragen und damit einen messbaren
betriebswirtschaftlichen Mehrwert ermöglichen.

Ein entscheidender Effekt liegt im Aufbrechen der Sequen-
zialität von Prozessen. Im PDF-basierten Status quo müssen
Normanforderungen zunächst recherchiert, dann manuell
extrahiert und anschließend in nachgelagerte Systeme
übertragen werden, bevor eigentliche Entwicklungsarbeit
beginnen kann. Jeder Schritt wartet auf den vorherigen.
SMART Standards entkoppeln diese Abhängigkeiten. Da
Norminhalte als granulare, stabil adressierbare Einheiten
vorliegen, können verschiedene Teams und Prozessschrit-
te gleichzeitig auf denselben normativen Inhalt zugreifen,
ohne eine zentrale Extraktions- und Übertragungsschicht zu
durchlaufen.

Das Ergebnis ist ein Wechsel von seriellen zu parallelen Pro-
zessen, den Whitepaper 3 anhand eines Mehrwertmodells
quantifiziert: Mit gleichem Personalaufwand können laut
Modell 1,5- bis 2-mal mehr Aufträge bearbeitet werden; die
Kosten für normbezogene Tätigkeiten sinken um 48 bis 64 %.

Dieses Whitepaper widmet sich der normbasierten Prüfung von Künstlicher Intelligenz mithilfe von SMART Standards, und damit im Kontrast zu den bisher veröffentlichten Whitepapers einem einzelnen, sehr spezifischen Technologiebereich.

können zudem besonders weitgehende Integrationsszenarien für SMART Standards entworfen werden.

Angesichts eines sich noch entwickelnden Felds ist das mit einigen Herausforderungen verbunden. So müssen teilweise Annahmen zur zukünftigen Relevanz bestimmter Technologien und Konzepte getroffen werden. Etwa zur Nutzung von standardisierten KI-Prüfwerkzeugen, KI-Prüfplattformen, Referenzdatensätzen, oder Datenräumen. Auch relevante Praktiken wie das Anforderungsmanagement für KI befinden sich aktuell noch in Anpassung bzw. Entwicklung.³ Zudem ist die KI-Normenlandschaft bei Veröffentlichung in einem frühen Entwicklungsstadium, insbesondere da die Veröffentlichung einiger harmonisierter KI-Normen für den AI Act noch aussteht und gerade die Ausgestaltung vertikaler Normen mit KI-Relevanz noch nicht klar ist. Welche Norminhalte also in welcher Form per SMART Standard bereitgestellt werden können („Wie konkret werden die Anforderungen an KI in Normen?“) unterliegt einer gewissen Unsicherheit.

Diese Faktoren machen es herausfordernd, KI-Prüfprozesse in technischem Detail und „zukunftssicher“ abzubilden, und den Bedarf für bestimmte Leistungsmerkmale von SMART Standards richtig einzuschätzen. Aber gerade Szenarien und Vorschläge eröffnen die Chance, eine Gesprächsgrundlage zu erzeugen, die das Potenzial von SMART Standards frühzeitig in die Debatte zur Entwicklung des Qualitätsökosystems (bzw. der Qualitätsinfrastruktur) für KI einführt.

Zudem kann die Weiterentwicklung des Konzepts für SMART Standards von der Verbindung mit dem Bereich der Künstlichen Intelligenz profitieren. Unternehmen, die KI-Normen anwenden und KI prüfen, weisen typischerweise einen hohen Digitalisierungsgrad auf und arbeiten häufig in agilen Entwicklungsprozessen⁴. Der Bedarf an SMART-Standards-Funktionalitäten ist in diesem Umfeld entsprechend hoch. Angesichts vorliegender digitaler Infrastruktur und Prozesse, die häufig automatisierte Bestandteile aufweisen,

3 Siehe z.B.: Habiba, U., Haug, M., Bogner, J., & Wagner, S. (2024). How mature is requirements engineering for AI-based systems? A systematic mapping study on practices, challenges, and future research directions. *Requirements Engineering*, 29(4), 567–600. <https://doi.org/10.1007/s00766-024-00432-3>

4 Siehe Whitepaper 3 bzgl. einer Diskussion des Digitalisierungsgrads von Unternehmen, sowie agiler Entwicklungsprozesse (SCRUM) und der Nutzung von SMART Standards

2

HINTERGRUND

2 HINTERGRUND

Mithilfe von Prüfungen wird bestimmt, ob oder inwieweit eine Künstliche Intelligenz den Anforderungen gerecht wird, die an sie gestellt werden. Teil einer solchen Prüfung können Aktivitäten wie das (automatisierte) Testen von Daten und KI-Modellen oder das Überprüfen von Prozessen und organisatorischen Strukturen auf Basis von Dokumentationen sein. „KI-Prüfungen“⁵ können beispielsweise regelmäßiger Bestandteil von Entwicklungs- und Validierungsprozessen sein, oder während eines Audits bzw. einer Konformitätsbewertung durchgeführt werden.

Gerade für Anbieter oder Betreiber von künstlicher Intelligenz, deren Anwendung mit einem hohen Risiko verbunden ist, sind solche Prüfungen angesichts klarer regulatorischer Vorgaben essenziell. Gleichzeitig zeigt sich, dass auch für niedrigere Risikostufen verbindliche Vorgaben entstehen, und dass die Nachfrage nach unabhängigen Prüfungen allgemein deutlich zunimmt. Freiwillige Prüfungen tragen nicht nur dazu bei, ein Mindestmaß an Qualität zu gewährleisten, sondern ermöglichen es Anbietern, ihre technische und ethische Verantwortung sichtbar zu machen und so Vertrauen gegenüber Nutzern, Kunden und Aufsichtsbehörden zu stärken.

Zunehmend entstehen daher Ansätze, die eine systematische, transparente und effiziente Bewertung von KI-Systemen über alle Risikoklassen hinweg ermöglichen sollen. Hier setzen Konzepte skalierbarer Prüfung mit dem Ziel an, Qualitätssicherung und Innovationsdynamik miteinander zu vereinen und die Prüfung in zukünftigen KI-Ökosystemen praktikabel zu gestalten.

In den folgenden Abschnitten wird eine Einordnung verschiedener, auf Normen und Standards basierender Prüfarten vorgenommen. Ziel ist es nicht, einen einzelnen regulatorischen Kontext, etwa die Konformitätsbewertung nach der europäischen KI-Verordnung, vollständig darzustellen. Vielmehr werden unterschiedliche Anwendungs- und Prüfungskontexte betrachtet, in denen KI-Systeme anhand normativer Anforderungen bewertet werden. Die dargestellten Prüfarten dienen als analytisches Raster, um mögliche Einsatzszenarien

skalierbarer Prüfungen mit SMART Standards zu strukturieren. In der Praxis überschneiden und kombinieren sich diese Kontexte häufig.

2.1 Überblick: Prüfung von Künstlicher Intelligenz auf Basis von Normen und Standards

Während Anbieter und Betreiber von KI-Systemen eine Vielzahl eigener Anforderungen aufstellen, findet eine öffentliche, systematische und oft rechtlich relevante Aufstellung von operationalisierbaren Anforderungen an die Eigenschaften von KI in Normen und Standards statt. Aktuelle Normen und Standards regeln für KI etwa Aspekte wie das Risiko-, oder Qualitätsmanagement, Daten-Governance, Sicherheit, oder, zusammengefasst unter „Vertrauenswürdigkeit“: Transparenz, Menschliche Aufsicht und Kontrolle, Verlässlichkeit, etc.⁶

Eine normenbasierte Prüfung von KI kann auf unterschiedliche Weise erfolgen, abhängig davon, was geprüft wird, wer die Prüfung in welchem Kontext durchführt, wann sie stattfindet und welche Ziele sie verfolgt.

Warum (in welchem Kontext) wird geprüft?

Die Ziele von KI-Prüfungen unterscheiden sich je nach Anwendungskontext. Sie reichen von der internen Qualitätssicherung und Risikominimierung über Anforderungen entlang der Wertschöpfungskette bis hin zur Darlegung rechtlicher Compliance. Insbesondere in regulierten Anwendungsbereichen spielen Konformitätsbewertungen eine wichtige Rolle. Im europäischen Rechtsrahmen erfolgt der Nachweis der Konformität häufig unter Rückgriff auf Normen und Standards. Für Hochrisiko-KI-Systeme sieht die KI-Verordnung hierzu spezifische Konformitätsbewertungsverfahren nach Artikel 43 sowie den Anhängen VI und VII vor. Die Anwendung harmonisierter europäischer Normen kann dabei eine Konformitätsvermutung hinsichtlich der adressierten gesetzlichen Anforderungen begründen. Auch für die Einhaltung anderer, etwa sektoraler Gesetze mit KI-Relevanz,

5 Im Folgenden wird KI-Prüfung als ein Bestandteil von KI-Bewertungstätigkeiten verstanden. Der Begriff bezieht sich insbesondere auf die Prüfung von KI-Systemen (und nicht notwendigerweise auf Prüfungen mithilfe von KI).

6 Siehe z.B.: Normungsroadmap Künstliche Intelligenz 2. Ausgabe, S.136 ff; Normungsthemen des JTC21, <https://jtc21.eu/working-groups/MISSION-KI-Qualitätsstandard>, <https://github.com/mission-ki/standard/>; Fraunhofer IAIS KI-Prüfkatalog, <https://www.iais.fraunhofer.de/de/publikationen/studien/2021/ki-pruefkatalog.html>; VDE SPEC 90012

Prüfung von Künstlicher Intelligenz

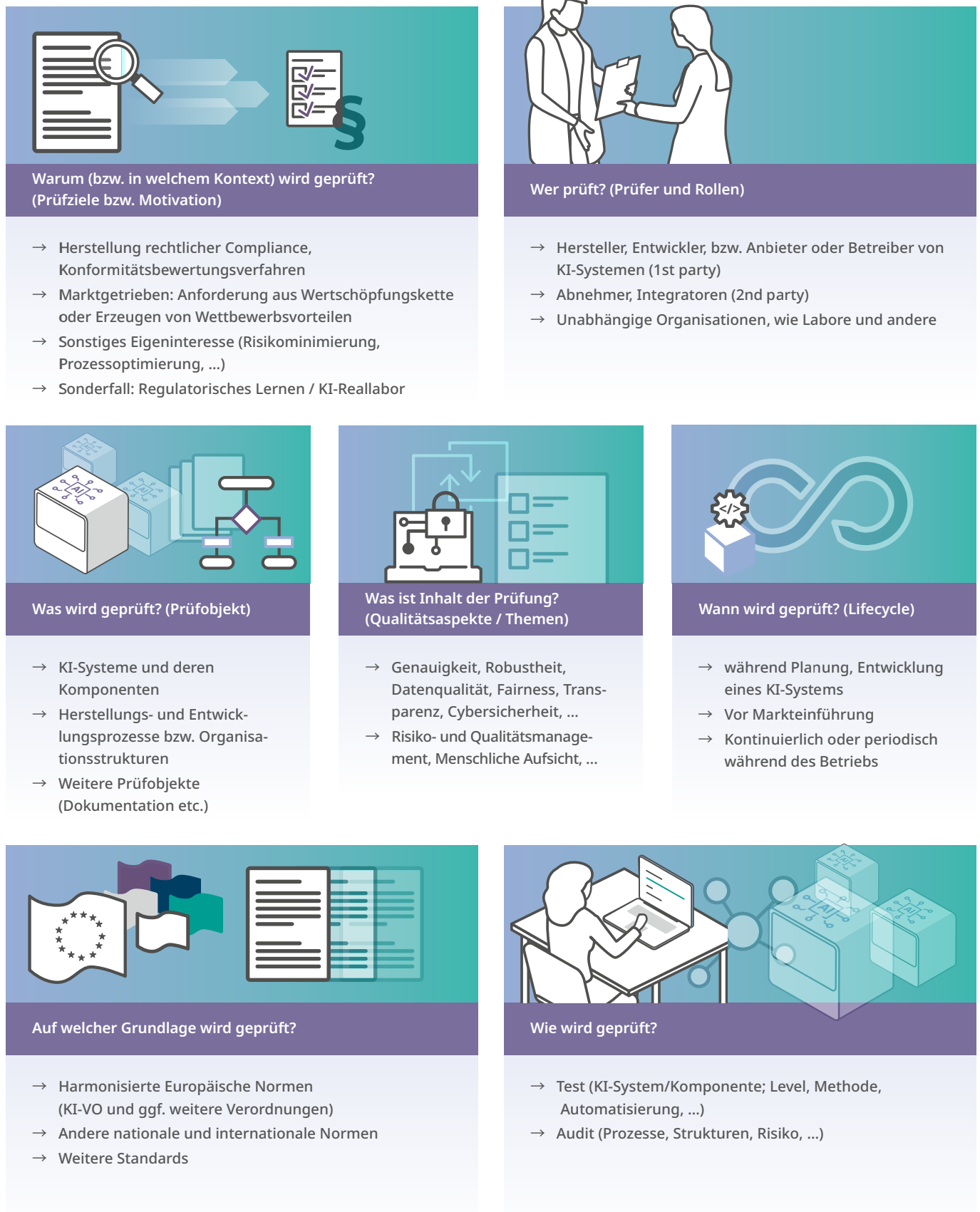


Abbildung 2

wie die Medizinprodukteverordnung (MDR), oder etwa die neue Maschinenverordnung (MVO) kann die Prüfung eines KI-Systems im Rahmen eines Konformitätsbewertungsverfahrens eine Rolle spielen.⁷

Werden KI-Systeme sowohl unter der KI-VO als auch unter anderen sektoralen Rechtsvorschriften, etwa der MDR, geprüft, kann es zu Überschneidungen zwischen den jeweiligen Konformitätsbewertungsverfahren kommen. Derzeit gibt es noch keine vollständige Harmonisierung zwischen diesen Verfahren. Die genaue Ausgestaltung dieser Schnittstellen zwischen KI-VO und sektoralen Rechtsakten bleibt ein offenes Thema der regulatorischen Weiterentwicklung.

Die Reproduzierbarkeit standardisierter Prüfungen durch SMART Standards könnte hier einen wichtigen Beitrag leisten: Einmal implementierte und dokumentierte Prüfaktivitäten könnten in verschiedenen regulatorischen Kontexten wiederverwendet werden, ohne jedes Verfahren vollständig neu aufsetzen zu müssen.

Neben der Durchführung im Rahmen gesetzlich vorgeschriebener Konformitätsbewertungen kann es auch verschiedene marktbedingte Gründe für KI-Prüfungen, bzw. KI-Audits, und KI-Zertifizierungen geben, die teilweise andere Prüfausprägungen beinhalten und auch für KI-Systeme mit niedrigerem Risiko nach KI-VO relevant sein können. So können etwa Anforderungen aus der Wertschöpfungskette Prüfungen erfordern. Geschäftspartner, Integratoren oder Abnehmer von KI-Komponenten verlangen hier zunehmend Nachweise über Qualität und Sicherheit. Gleichzeitig kann ein geprüftes und gegebenenfalls zertifiziertes KI-System Wettbewerbsvorteile bieten, indem es Qualitätsmerkmale wie Transparenz, Zuverlässigkeit und Verantwortungsbewusstsein signalisiert und die Positionierung im Markt stärkt.⁸ Schließlich spielen auch andere Eigeninteressen der Hersteller bzw. Entwickler, Anbieter und Betreiber eine Rolle. Sie prüfen KI mithilfe von Normen und Standards auch, um Risiken zu minimieren, eigene Qualitätsstandards zu sichern und interne Entwicklungsprozesse zu verbessern.

Einen Sonderfall stellen Prüfungen von KI-Systemen in KI-Reallaboren dar.⁹ Hier können Prüfungen unter Anleitung und Aufsicht relevanter Behörden und unter eingeschränktem (rechtlichem) Risiko durchgeführt werden¹⁰. Für Behörden kann das Ziel der Prüfungen hier sein, die Regulierung innovativer KI-Lösungen zu erproben und weiterzuentwickeln (regulatorisches Lernen), während Unternehmen Konformitätsbewertungen durch Prüfungen in KI-Reallaboren gegebenenfalls einfacher bzw. risikoärmer umsetzen können.

Wann wird geprüft?

KI-Prüfungen können in unterschiedlichen Phasen des Systemlebenszyklus stattfinden, zum Beispiel während der (Neu-) Planung bzw. des (Re-) Designs eines KI-Systems, oder während der (Weiter-) Entwicklung zur Qualitätssicherung und Fehlervermeidung. Relevant ist auch die Phase vor bzw. während des Deployments, also etwa vor der Markteinführung (im Rahmen eines Konformitätsbewertungsverfahrens oder einer Zertifizierung) und während des Betriebs in Form kontinuierlicher oder periodischer Überwachung, etwa bei Modellupdates, z.B. im Rahmen von Post-Market Monitoring (siehe auch Kapitel 2.2).

Auf welcher Grundlage wird geprüft?

Je nach Prüfkontext und Zielsetzung kommen unterschiedliche Prüfgrundlagen zum Einsatz. Eine grundlegende Unterscheidung lässt sich anhand des regulatorischen Verbindlichkeitsgrads treffen:¹¹

Im Regelfall werden Konformitätsbewertungen auf Grundlage der durch das Gesetz referenzierten Normen durchgeführt. Diese harmonisierten Europäischen Normen (hEN) werden auf Grundlage von Normungsaufträgen der Europäischen Kommission durch CEN, CENELEC oder ETSI entwickelt. Ihre Anwendung ist freiwillig und Rechtskonformität kann grundsätzlich auch mithilfe von KI-Prüfungen nachgewiesen werden, die nicht auf (diesen) Normen basieren, sondern die gesetzlichen Anforderungen anders adressieren¹². Da Konformitätsbewertungen auf Grundlage harmonisierter Nor-

7 <https://www.vde.com/topics-de/health/beratung/zulassung-von-ki-basierten-medizinprodukten-in-europa>

8 Siehe z.B. MISSION KI Qualitätsstandard, <https://github.com/mission-ki/standard/>

9 Z.B.: <https://epthintank.eu/2026/04/01/ai-regulatory-sandboxes-state-of-play-and-implementation-challenges/>

10 vgl. Art. 57 KI-VO

11 Vgl. Normungsrroadmap (2), S. 128 ff / Abb. 27

12 Anhang VII KI-VO

KI-NORMEN ALS PRÜFGRUNDLAGE (BEISPIELE)

Harmonisierte Europäische Normen (CEN/CENELEC):

- prEN 18286 - KI-Qualitätsmanagementsystem
- prEN 18228 - KI-Risikomanagement: unterstützt die Anforderungen an das Risikomanagementsystem nach Artikel 9 KI-VO und ist über Artikel 17 eng mit dem Qualitätsmanagementsystem verzahnt
- prEN 18282 - Cybersicherheit für KI-Systeme: ergänzt KI-spezifische Sicherheitsanforderungen und baut auf bestehenden Normen wie der ISO/IEC-27000-Reihe und ISO/IEC 15408 (Common Criteria) auf

Andere internationale Normen (ISO/IEC):

- ISO/IEC 42001 - Informationstechnik - Künstliche Intelligenz - Managementsystem
- ISO/IEC 23894 - Informationstechnik - Künstliche Intelligenz - Leitlinien für Risikomanagement
- ISO/IEC 25059 - Software-Engineering - Qualitätskriterien und Bewertung von Systemen und Softwareprodukten (SQuaRE) - Qualitätsmodell für KI-Systeme

men jedoch eine spezielle „Konformitätsvermutung“ (zum Gesetz) verleihen,¹³ dürften normbasierte KI-Prüfungen in diesem Kontext die in den meisten Fällen relevanteste, da mit höherer Rechtssicherheit und mit am wenigsten Aufwand verbundene Form der Prüfung sein.

Aktuell entstehen die zentralen harmonisierten Normen (hEN) zur KI-VO. Die erste davon (prEN 18286, siehe Infobox) durchlief Ende 2025 die öffentliche Konsultation und soll im Juli 2026 veröffentlicht werden. Es entstehen weitere Kernnormen, die einzelne Anforderungen der KI-VO adressieren. CEN und CENELEC haben im Oktober 2025 Sondermaßnahmen beschlossen, um die Verfügbarkeit der Kernnormen bis Q4 2026 sicherzustellen.

Daneben existieren weitere nationale und internationale Normen. Diese Normen können auch ohne formale Harmonisierung als Prüfgrundlage dienen. Sie sind häufig bereits verfügbar und werden in der Praxis zum Beispiel auch dann genutzt, solange harmonisierte Normen noch ausstehen. Dabei adressieren Managementsystem-Normen wie ISO/IEC 42001 die Prozessebene; für die produktbezogene Qualitätsprüfung werden ISO/IEC 25059 und darauf aufbauende Konformitätsbewertungsverfahren maßgeblich.

Schließlich spielen weitere Standards wie Konsortialstandards, Frameworks und Guidelines (z.B. NIST AI RMF, MIS-SION KI Qualitätsstandard, Fraunhofer IAIS KI-Prüfkatalog) eine ergänzende Rolle, insbesondere dann, wenn formale Normen noch nicht ausreichend operationalisierbare Anforderungen liefern.

Wer prüft?

Abhängig von Prüfzielen und -kontexten gibt es bei KI-Prüfungen verschiedene Rollen und Verantwortlichkeiten. Diese lassen sich insbesondere für Prüfungen im Rahmen von formellen Konformitätsbewertungen in drei Ebenen gliedern. Im Fall einer Selbstprüfung (first party, Erststellenprüfung) erfolgt die Prüfung durch den Hersteller, Entwickler bzw. den Anbieter oder den Betreiber selbst, etwa während der Entwicklungsphase oder im laufenden Betrieb, bzw. im Rahmen einer gesetzlich vorgeschriebenen Erstbewertung (z.B. bei einer „internen Kontrolle“ von Hochrisiko KI-Systemen¹⁴). Die hier betrachtete Selbstprüfung bezieht sich auf die Durchführung von Prüftätigkeiten durch den Anbieter bzw. Betreiber selbst. Sie stellt damit lediglich einen möglichen Bestandteil einer umfassenderen Konformitätsbewertung oder Selbsterklärung dar.

Externe Organisationen, wie Abnehmer oder Integratoren, prüfen die Qualität von zugekauften KI-Komponenten oder -Systemen, etwa um deren Eignung für eigene Anwendungen sicherzustellen (second party, Zweitstellenprüfung). Externe, unabhängige Stellen (third party, Drittstellenprüfung) führen formelle Prüfungen bzw. Zertifizierungen durch. Hierunter fallen gesetzlich vorgeschriebene Konformitätsbewertungsverfahren, Produktzertifizierungen oder Systemaudits.

Welche Ebene zum Einsatz kommt, ergibt sich nicht primär aus der Eigenschaft des KI-Systems, sondern aus dem regulatorischen Kontext. Sowohl Erststellen- als auch Dritt-

¹³ Art. 40 KI-VO

¹⁴ Art. 43 Abs. 1 (a) KI-VO

stellen-Prüfungen können gesetzlich vorgeschrieben sein. Die KI-VO erlaubt für die meisten Hochrisiko-Systeme eine interne Kontrolle durch den Anbieter (Art. 43 i.V.m. Anhang VI), schreibt aber für bestimmte biometrische Systeme die Einbindung einer benannten Stelle vor (Anhang VII). Zweitstellen-Prüfungen sind regulatorisch seltener gefordert, spielen aber praktisch eine wachsende Rolle bei der Absicherung von KI-Komponenten in Lieferketten.

An KI-Prüfungen sind in der Regel mehrere Rollen beteiligt, deren genaue Ausgestaltung je nach Organisationsform und Prüfkontext variiert. Auf Anbieter- und Betreiberseite sind typischerweise Rollen wie Product Owner, ML Engineer, Quality Engineer, Compliance Manager oder Data Scientists involviert. Sie tragen Verantwortung für die Operationalisierung von Anforderungen, die Implementierung und Ausführung von Tests sowie die Dokumentation von Prüfergebnissen. Auf Seiten externer Prüfstellen agieren akkreditierte Konformitätsbewertungsstellen (KBS), Prüflabore oder spezialisierte Auditoren. Marktüberwachungsbehörden übernehmen schließlich die nachgelagerte Kontrolle im Betrieb. Welche dieser Rollen konkret für KI-Prüfungen verantwortlich zeichnen und wie sie in Organisationen verankert werden, ist ein noch aktiv diskutiertes Thema, unter anderem im Kontext von ETSI TR 104 119, das Empfehlungen zur Dokumentation und Rollenverantwortung bei KI-Systemen gibt.

Was wird wie geprüft?

Gegenstand von KI-Prüfungen können verschiedene Prüfobjekte sein, für die in KI-Normen und -Standards Regeln festgelegt wurden. Das zentrale Prüfobjekt ist das eingesetzte KI-System, womit meist ein System gemeint wird, das auf Machine Learning Modellen basiert. Allgemeine Definitionen, wie die der KI-VO¹⁵, schließen aber auch andere Formen von KI ein.

Die Prüfung eines KI-Systems beinhaltet etwa das Testen des Systems als Ganzes (End-to-End-testing), oder das einzelne Testen von Komponenten wie (trainierte) Model-

le, eingesetzte Daten (Trainings-, Test- und Validierungsdaten), sowie weitere Systemkomponenten und Schnittstellen (Code, APIs, Infrastruktur). Tests unterscheiden sich neben Testobjekt und -level auch etwa hinsichtlich des Grads der Automatisierbarkeit, dem notwendigen Zugang zu KI-Modell und (Trainings-) Daten und den eingesetzten Arten von Testmethoden (z.B.: Black-, White-, oder Grey-Box Testing).¹⁶ Prüfbare Artefakt eines KI-Systems ist zudem auch die begleitende Dokumentation¹⁷, die etwa technische Beschreibungen oder Risikoanalysen umfasst.

Neben dem KI-System selbst können auch die zugrundeliegenden Entwicklungs- und Betriebsprozesse Gegenstand einer Prüfung sein (vgl. Prozessaudits nach ISO/IEC 17021). Geprüft werden die Methoden, Verfahren und organisatorischen Strukturen, die sicherstellen, dass Qualität, Sicherheit und Nachvollziehbarkeit während der Entwicklung gewährleistet sind. Dazu zählen etwa Entwicklungs- und Validierungsprozesse, eingesetzte Datenmanagementpraktiken, Maßnahmen zum Risikomanagement sowie interne Audit- und Freigabeprozesse. Eine besondere Rolle spielen in diesem Zusammenhang Managementsysteme, die diese Abläufe institutionalisieren. Relevante Beispiele sind das KI-Managementsystem nach ISO/IEC 42001, das Qualitätsmanagementsystem nach ISO 9001 oder das Informationssicherheitsmanagementsystem nach ISO/IEC 27001. Sie schaffen die organisationale Grundlage für eine reproduzierbare, dokumentierte und überprüfbare KI-Entwicklung.

Relevant im Kontext von KI-Prüfungen kann nach ISO/IEC 17024 auch die Prüfung bzw. Zertifizierung der beteiligten Akteure sein: Personen, die Prüfungen durchführen oder für KI-Entwicklung und -Überwachung zuständig sind. Nach ISO/IEC 17025 ist zudem die Prüfung und Akkreditierung von Prüf- und Kalibrierlaboren (Konformitätsbewertungsstellen) geregelt, um deren technische Kompetenz und Unabhängigkeit sicherzustellen. Im Rahmen dieses Whitepapers werden diese Prüfarten jedoch nicht betrachtet.¹⁸

¹⁵ Art. 3 KI-VO

¹⁶ siehe DIN SPEC 92006

¹⁷ Die Prüfung von Managementsystemen erfolgt in der Regel zu einem großen Teil ebenfalls anhand von Dokumentationen. Insofern ist die Unterteilung je nach Interpretation nicht trennscharf.

¹⁸ Siehe zu diesem Thema bspw.: Kuch, S., & Kirmes, R. (2025). AI Certification: An Accreditation Perspective (Practitioner Track). In R. Gorge, E. Haedecke, M. Poretschkin, & A. Schmitz (Hrsg.), OASICS, Volume 126, SAIA 2024 (Nr. 14; Bd. 126, S. 14:1-14:7). Schloss Dagstuhl – Leibniz-Zentrum für Informatik. <https://doi.org/10.4230/OASICS.SAIA.2024.14>

IM DETAIL: RECHTLICHE VORGABEN FÜR REGELMÄSSIGE ODER FORTLAUFENDE PRÜFUNGEN VON KI

Die KI-VO etabliert für Hochrisiko-KI-Systeme eine Reihe von Pflichten, die Prüfaktivitäten systematisch über den Lebenszyklus verteilen, statt sie auf den Zeitpunkt der Markteinführung zu konzentrieren. Das Risikomanagementsystem (Art. 9) ist ausdrücklich als kontinuierlicher, iterativer Prozess definiert, das Qualitätsmanagementsystem (Art. 17) schreibt Test-, Validierungs- und Änderungsverfahren über den gesamten Lebenszyklus vor, und die geforderten Eigenschaften Genauigkeit, Robustheit und Cybersicherheit (Art. 15) sind ebenfalls lebenszyklusbezogen formuliert.

Operativ wird diese Daueraufgabe vor allem über die Beobachtung nach dem Inverkehrbringen konkretisiert (Art. 72). Anbieter müssen Performance-Daten aktiv und systematisch erheben und gegen die Anforderungen aus Kapitel III Abschnitt 2 KI-VO abgleichen, um die fortwährende Konformität zu bewerten. Die engen Meldefristen aus Art. 73 für schwerwiegende Vorfälle und die Beobachtungspflichten der Betreiber aus Art. 26 Abs. 5 setzen eine ebenso kontinuierliche Detektionsinfrastruktur voraus.

Veränderungen am System unterliegen mit Art. 43 Abs. 4 einer eigenen Prüflogik: Wesentliche Änderungen lösen eine

erneute Konformitätsbewertung aus, und die Ausnahme für lernende Systeme verlagert den Aufwand lediglich in die kontinuierliche Überwachung eines vorab definierten Änderungsraums. Hinzu kommt die quantitative Dimension, da die KI-VO acht Hochrisiko-Anwendungsfelder umfasst. Unternehmen mit mehreren Produkten oder Modellvarianten müssen daher möglicherweise nicht nur ein, sondern viele Systeme prüfen.

Die KI-VO ist zudem nicht der einzige Rechtsakt, der wiederkehrende Prüfungen vorschreibt. Ist ein KI-System Sicherheitskomponente eines bereits regulierten Produkts, addieren sich die Lifecycle-Pflichten der sektoralen Rahmen (Medizinprodukte- und IVD-Verordnung, Maschinenverordnung, Cyber Resilience Act). Querschnittsregelungen wie DORA, NIS2 und die DSGVO verstärken das Bild durch Anforderungen an kontinuierliches IKT-Risikomanagement, laufende Cybersicherheits-Maßnahmen und die jederzeitige Nachweisbarkeit der Compliance. Managementsystem-Standards wie ISO/IEC 42001 institutionalisieren diese Daueraufgabe organisatorisch durch interne Audits, Management Reviews und Überwachungsaudits.

2.2 Skalierbare Prüfung von KI-Systemen

Die Vorbereitung und Durchführung von Prüfungen von KI-Systemen im Kontext von Gesetzen, Normen und anderen Standards bestehen aus einer Reihe komplexer Aktivitäten, wie

- der Identifikation von relevanten Qualitätsdimensionen, Gesetzen und Normen,
- der Durchführung von Risikobewertungen,
- der Operationalisierung von Anforderungen,
- inklusive der Untersuchung, Festlegung und Erhebung von konkreten Messgrößen,
- sowie der Auswahl von Softwarebibliotheken bzw. der Implementierung (und Verifizierung) von Code.

Diese Aktivitäten sind nicht notwendigerweise spezifisch für KI, doch ihre Durchführung unterscheidet sich von klassischen Softwaretests in mehreren Punkten. KI-Systeme arbeiten meist nicht-deterministisch, und ihr Verhalten kann

nicht vollständig durch feste Regeln beschrieben werden. Ergebnisse hängen wesentlich von Daten ab, die selbst Fehlerquellen darstellen können, etwa durch Verzerrungen, Qualitätsmängel oder Drift im laufenden Betrieb.

Anders als bei klassischen Softwareprodukten fehlen häufig vollständige und eindeutige Spezifikationen, sodass Eigenschaften wie Genauigkeit, Fairness oder Robustheit nur über Metriken und Näherungen geprüft werden können und Bewertungen mit Unsicherheiten verbunden sind. Hinzu kommt, dass KI-Systeme sich über die Zeit verändern, etwa durch Retraining oder veränderte Nutzungskontexte, wodurch einmalige Tests schnell an Aussagekraft verlieren. Die oft eingeschränkte Erklärbarkeit vieler Modelle erschwert zudem die Ursachenanalyse von Fehlverhalten.

Zahlreiche gesetzliche Regelungen sowie Anforderungen in Normen und Standards sehen die regelmäßige oder kontinuierliche Prüfung und Berichterstattung zur Konformität von KI-Systemen vor (siehe Infobox). Daher müssen Prüfun-

gen von KI-Systemen regelmäßig und in großem Umfang durchgeführt werden.¹⁹ Der Prüfbedarf wächst dabei mit drei Faktoren: der Anzahl der zu prüfenden Systeme, der Häufigkeit ihrer Veränderungen durch Retraining, Datenupdates oder Anpassungen abhängiger Komponenten, sowie der Komplexität der Normenlandschaft mit ihrer Vielfalt an Informationsquellen und Prüfaspekten.

Damit Innovation und Qualitätssicherung in Einklang bleiben, muss der Aufwand pro Prüfung in einem zur Anzahl der Prüfanlässe handhabbaren Verhältnis stehen. Skalierbar heißt in diesem Kontext, dass Prüfungen so weit wie möglich automatisiert werden und verbleibende manuelle Aktivitäten durch eine vereinfachte Bereitstellung geeigneter Informationen und Hilfsmittel beschleunigt werden.^{20,21}

Verschiedene technische Konzepte können zur Skalierbarkeit der Prüfung von KI-Systemen beitragen, indem sie Abläufe automatisieren, reproduzierbar machen, und Zusatzaufwände an Schnittstellen reduzieren. Hierzu zählt beispielsweise die Einbettung kontinuierlicher, automatisierter Prüfungen in MLOps-Frameworks.

Während die MLOps-Integration vor allem die zeitliche Skalierung sowie die parallele Verwaltung mehrerer Systemvarianten adressiert, betrifft eine zweite Klasse von Maßnahmen die Vergleichbarkeit und Wiederverwendbarkeit von Prüfungen über Systeme und Prüfanlässe hinweg:

Bei der Operationalisierung von Anforderungen aus Normen gibt es Interpretationsspielräume. Es müssen Entscheidungen gefällt werden, die etwa die genaue Auswahl von Algorithmen und Metriken, das Setzen von (Hyper-)Parametern oder die Implementierung bzw. Auswahl von Bibliotheken und Validierungsdatensätzen betreffen.

Je mehr solche Entscheidungen standardisiert werden, je mehr Regeln also für die Operationalisierung vorliegen, desto einfacher ist es, Prüfungen vergleichbar und wiederverwendbar zu machen. Dies erleichtert es, Prüfungen zu verifizieren, Prüfergebnisse zu vergleichen und sie in ver-

schiedenen Prüfkontexten wiederzuverwenden, etwa um implementierte automatische Prüfungen für eine Konformitätsbewertung im Rahmen einer Markteinführung eines KI-Systems zu verwenden.

Verschiedene Komponenten einer Infrastruktur für KI-Prüfungen können zur Standardisierung und Reproduzierbarkeit von KI-Prüfungen beitragen, wie etwa KI-Prüfwerkzeuge, spezialisierte Prüfservices auf entsprechenden technischen Plattformen, standardisierte Referenz- bzw. Test-Daten, oder semantische Modelle von Regelwerken. Eine wichtige Rolle spielt auch die Integration in die digitale Qualitätsinfrastruktur, wo etwa durch den Einsatz digitaler, standardisierter Prüf- bzw. Konformitätsberichte Aufwände reduziert, und Qualität erhöht werden kann.

19 Hiervon gibt es Ausnahmen, etwa im Fall von „Locked AI“ – Systemen, die nach dem Inverkehrbringen nicht weiterlernen und daher in einem unveränderlichen Zustand bleiben. Ihre nach der Bewertung gleichbleibenden Eigenschaften passen besser zu Verfahren, die für statische Produkte konzipiert sind.

20 Haedecke, E. G., Mock, M., Pintz, M. A., & Poretschkin, M. (2023). KI-Anwendungen systematisch prüfen und absichern. Prüfwerkzeuge und Prüfplattform zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz.

21 Symposium on Scaling AI Assessments (SAIA 2024). Open Access Series in Informatics (OASICs), Volume 126, pp. 0:i-0:xii, Schloss Dagstuhl – Leibniz-Zentrum für Informatik (2025) <https://doi.org/10.4230/OASICs.SAIA.2024.0>

BEISPIELE FÜR KONZEPTE UND KOMPONENTEN SKALIERBARER PRÜFUNG VON KI:

- 2.2.1 CABCA / MLOps
- 2.2.2 KI-Prüfwerkzeuge
- 2.2.3 Referenzdatensätze
- 2.2.4. KI-Prüfplattformen
- 2.2.5. Semantische Modellierung
- 2.2.6 Standardisierte Prüfberichte
- 2.2.7 SMART Standards

2.2.1 Kontinuierliche Konformitätsprüfung in MLOps-Framework

Beim Konzept sogenannter kontinuierlicher Konformitätsbewertungen von KI (CABCA)^{22,23,24} werden punktuelle Audits durch eine laufende, automatisierte Überprüfung der Konformität ersetzt. Kern des Ansatzes ist die Operationalisierung abstrakter Anforderungen, die systematisch in konkrete, messbare und maschinenlesbare Kriterien übersetzt werden. Darauf basierend erfolgt eine automatisierte und kontinuierliche Erhebung von Nachweisen aus dem KI-System, etwa aus Logs, Leistungskennzahlen oder Daten- und Modellartefakten. Die gesammelten Messwerte werden fortlaufend analysiert und mit definierten Schwellenwerten verglichen. Abweichungen, neue Risiken oder Nichtkonformitäten werden zeitnah erkannt, dokumentiert und in einem aktuellen Konformitätsstatus zusammengeführt. Prüfung und Überwachung erstrecken sich von Entwicklung und Training über den Betrieb bis hin zum Post-Market-Monitoring, wobei Tiefe und Intensität der Prüfung vom Risikoprofil des KI-Systems abhängen.

Technisch und organisatorisch ist ein solches Konzept idealerweise in ein MLOps (Machine Learning Operations) Framework eingebettet, welches Praktiken, Prozesse und Werkzeuge umfasst, mit denen der gesamte Lebenszyklus von KI-Modellen, von Datenaufbereitung und Training über Deployment bis hin zu Betrieb und Wartung, systematisch gesteuert wird.

Im Kontext von KI-Prüfungen stellt MLOps die operative Infrastruktur für kontinuierliche Messung und Überwachung bereit. Typische MLOps-Mechanismen wie Versionierung von Modellen und Daten, automatisierte Tests, Monitoring von Modellperformance, Drift-Erkennung und Logging liefern jene Artefakte und Evidenzen, die für eine kontinuierliche Konformitätsprüfung benötigt werden. Prüftätigkeiten, wie die Berechnung und der Abgleich von Metriken auf Basis von Daten oder KI-Modellen, lassen sich direkt in bestehende MLOps-Pipelines integrieren, sodass Prüfungen nicht als zusätzlicher manueller Schritt, sondern als integrierter Bestandteil des laufenden Betriebs erfolgen.

2.2.2 KI-Prüfwerkzeuge

(Standardisierte) KI-Prüfwerkzeuge²⁵ sind darauf ausgerichtet, transparente, nachvollziehbare und wiederholbare Prüfprozesse zu unterstützen. Dies setzt voraus, dass ihre Funktionsweise sowie die verwendeten Metriken, Eingaben, Parameter und Ausgaben nachvollziehbar dokumentiert sind. Testziele und Testkriterien werden so definiert, dass erkennbar ist, welche Eigenschaften eines KI-Systems unter welchen Bedingungen geprüft werden.

Das zentrale Element des Konzepts ist die fachliche Validität der eingesetzten Prüfmethode, bei dem die Methodik und technische Umsetzung der Werkzeuge sowie ihre Schnittstellen dokumentiert und fachlich begründet werden, sodass die Prüfungen inhaltlich nachvollziehbar bleiben. Für

22 Knoblauch, D., & Großmann, J. (2023). Towards a risk-based continuous auditing-based certification for Machine Learning. *The Review of Socionetwork Strategies*, 17(2), 255–273. <https://doi.org/10.1007/s12626-023-00148-w>

23 Minkinen, M., Laine, J., & Mäntymäki, M. (2022). Continuous Auditing of Artificial Intelligence: a Conceptualization and Assessment of Tools and Frameworks. In *Digital Society* (Vol. 1, Issue 3). Springer Science and Business Media LLC. <https://doi.org/10.1007/s44206-022-00022-2>

24 ETSI TS 104 008 V1.1.1 (2026-01)

25 DIN SPEC 92006

den vorgesehenen Einsatzbereich wird angestrebt, dass die Werkzeuge konsistent arbeiten, vorhersehbare Laufzeiten aufweisen und Fehlerzustände transparent handhaben.

Ein weiterer Aspekt ist der Umgang mit Messunsicherheit und Reproduzierbarkeit. Relevante Unsicherheitsquellen werden identifiziert und, soweit möglich, quantifiziert. Nicht-deterministische Einflüsse und ihr Beitrag zur Ergebnisvarianz werden dokumentiert, um die Vergleichbarkeit von Prüfergebnissen zu unterstützen.

Schließlich umfasst das Konzept eine systematische Protokollierung relevanter Parameter, Konfigurationen und Ergebnisse, sodass einzelne Testläufe rekonstruierbar bleiben und eine Einbindung in wiederkehrende oder automatisierte Prüfprozesse ermöglicht wird.

2.2.3 Referenzdatensätze

Standardisierte Referenzdatensätze²⁶ sind ein weiteres zentrales Element einer reproduzierbaren KI-Prüfinfrastruktur. Sie ermöglichen es, die Performance von KI-Modellen unter kontrollierten, vergleichbaren Bedingungen zu bewerten, unabhängig davon, wer die Prüfung durchführt. Referenzdatensätze müssen dabei klar dokumentiert sein hinsichtlich Herkunft, Zusammensetzung, bekannter Biases und Gültigkeitsbereich. Im Kontext physikalisch-technischer Prüfungen sind solche Datensätze vergleichbar mit zertifizierten Referenzmaterialien in der klassischen Messtechnik.

In Deutschland wird der Aufbau einer solchen Infrastruktur u.a. durch Akteure wie die Physikalisch-Technische Bundesanstalt (PTB) und im Rahmen von QI-Digital vorangetrieben. Ein konkretes Beispiel ist das Pilotprojekt „KI in der Medizin“, das standardisierte Testdatensätze für medizinische Bildgebungs-KI erprobt. SMART Standards können in diesem Kontext eine wichtige Rolle spielen, indem sie Referenzdatensätze mit persistenten Identifiern referenzieren und so eine automatisierte Integration in Prüfworkflows ermöglichen. Damit Werkzeuge, Referenzdaten und semantische Modelle in einem konsistenten Prüfablauf zusammenwirken, braucht es eine integrierende technische und semantische Schicht.

2.2.4 KI-Prüfplattformen

Als KI-Prüfplattformen werden technische Frameworks beschrieben, die eine strukturierte Definition, Ausführung und Dokumentation von Prüfprozessen für KI-Systeme ermöglichen. Sie stellen eine Infrastruktur bereit, in der Prüfmethoden unabhängig von ihrem inhaltlichen Fokus spezifiziert und reproduzierbar ausgeführt werden können. Im Zentrum steht dabei die technische Operationalisierung von Prüfabläufen: (standardisierte) Prüfwerkzeuge, Modelle, Datensätze und Konfigurationen werden als klar abgegrenzte, maschinenlesbar beschriebene Komponenten behandelt, deren Zusammenspiel explizit definiert ist. Damit werden Mechanismen bereitgestellt, um beliebige Prüfmethoden konsistent zu integrieren und auszuführen.

Darüber hinaus können durch eine Prüfplattform Mechanismen zur Versionierung, Protokollierung und Archivierung bereitgestellt werden. Jede Ausführung eines Prüfworkflows kann mit eindeutigen Referenzen auf eingesetzte Komponenten, Parameter und Artefakte dokumentiert werden. Dadurch entsteht eine nachvollziehbare und wiederholbare Prüfbeschreibung, die unabhängig von der inhaltlichen Ausrichtung der jeweiligen Tests ist. Die Plattform wirkt somit als generisches, inhaltlich agnostisches Framework, das die formale Strukturierung und technische Konsistenz von KI-Prüfprozessen sicherstellt.

Entwickler bzw. Betreiber von Prüfplattformen können KI-Prüfungen als Service anbieten, etwa zentralisiert (off-premise), wobei Prüflinge ihr KI-System oder einzelne Komponenten für die Prüfung bereitstellen, oder on-premise, auf der Infrastruktur des Prüflings. Das Konzept einer KI-Prüfplattform kann durch Berücksichtigung der zentralen Konzepte auch im Rahmen der Implementierung einer MLOps Infrastruktur unternehmensintern selbst umgesetzt werden.

Als Betreiber solcher Plattformen kommen unterschiedliche Akteure in Frage. Akkreditierte Prüfstellen und Konformitätsbewertungsstellen können sie als Teil ihres Dienstleistungsangebots betreiben. Forschungseinrichtungen und nationale Metrologieinstitute (etwa die PTB, u.a. im Rahmen von QI-Digital) entwickeln Plattformen mit Referenzcharakter. Schließlich können Plattformbausteine als Open-Source-Komponenten verfügbar sein und intern, etwa innerhalb eines unternehmenseigenen MLOps-Stacks, betrieben werden. In

26 DIN SPEC 92007

der Praxis dürften sich Mischformen herausbilden, in denen spezialisierte Plattformen für bestimmte Einsatzbereiche (z.B. medizinische Bildgebung, Credit-Scoring) mit weiterer Prüfinfrastruktur integriert werden.

2.2.5 Semantische Modellierung

Eine grundlegende Methode, KI-Prüfungen auch bei wachsender Anzahl und Komplexität skalierbar zu machen, besteht im Einsatz von Ontologien und Knowledge Graphs (Wissensgraphen).

Ontologien schaffen eine formale, maschinenlesbare Struktur für das, was bei Prüfungen von KI-Systemen relevant ist. Dazu gehören Datentypen, Modellklassen, Trainingsverfahren, Metriken, Risiken, Annahmen und insbesondere Anforderungen aus Normen und Gesetzen.

Durch diese Formalisierung lassen sich prüfrelevante Objekte standardisieren und vergleichbar machen. Anstatt jede Prüfung individuell zu definieren, können Prüfende auf wiederverwendbare ontologische Bausteine zurückgreifen, die festlegen, was geprüft werden muss, zum Beispiel Datenherkunft, Fairnesskriterien oder Robustheit, und wie diese Aspekte zueinander in Beziehung stehen. Zum Beispiel lassen sich so Konzepte und Anforderungen aus Gesetzen auf Normen und Standards abbilden.²⁷

Konkrete, mit spezifischem Wissen angereicherte Knowledge Graphs bauen auf Ontologien auf und ermöglichen es, konkrete Informationen über Daten, Modelle, Pipelines und deren Lebenszyklus miteinander zu verknüpfen. Statt isolierte Checklisten abzuarbeiten, können Prüfsysteme so gezielt Abhängigkeiten analysieren, etwa welche Trainingsdaten ein Modell beeinflusst haben, welche Vorverarbeitungen angewendet wurden oder welche Modelle von denselben Datenquellen abhängen. Anforderungen können formal modelliert und direkt mit Modelleigenschaften und Daten verknüpft werden, etwa mit Risikoanalysen oder der Dokumentation von KI-Systemen.²⁸ Auf dieser Basis lassen sich automatisierte Abfragen formulieren, die Inkonsistenzen oder Regelverstöße identifizieren können. Neue Modelle oder Datenflüsse

werden dann automatisch in den bestehenden Wissensgraphen integriert und gegen bekannte Prüfregeln evaluiert.

2.2.6 Standardisierte Prüfberichte

Ein weiterer Baustein skalierbarer KI-Prüfungen ist die automatisierte Generierung von Ergebnisdokumentation in standardisierten Formaten. Prüfergebnisse, Nachweise und Konformitätsstatus werden nicht mehr manuell in Berichte übertragen, sondern direkt aus den Prüfsystemen exportiert. Relevante Formate hierfür sind etwa digitale Konformitätsnachweise (Digital Certificates of Conformity, D-CoCs), Model Cards oder ein sogenannter „KI-Pass“, der strukturierte, maschinenlesbare Beschreibungen eines KI-Systems und seiner Prüfergebnisse enthält. Solche Nachweise können perspektivisch auch über digitale Produktdatenökosysteme, etwa Digitale Produktpässe (DPP), bereitgestellt oder referenziert werden.²⁹ Standardisierte Berichtsformate ermöglichen die direkte Weitergabe von Prüfergebnissen an Aufsichtsbehörden, benannte Stellen oder nachgelagerte Akteure in der Wertschöpfungskette und reduzieren so den Aufwand für Berichterstattung und Wiederverwendung von Nachweisen.

2.2.7 SMART Standards

Während die bisher beschriebenen technischen Elemente die Durchführung und Dokumentation von Prüfungen adressieren, betrifft eine zweite Aufwandsdimension die Erschließung der normativen Grundlage selbst. Der Weg von einer Anforderung aus einer Norm zu einer durchführbaren KI-Prüfung erfordert bislang teils erheblichen manuellen Aufwand und ist fehleranfällig. Relevante Normen müssen identifiziert und interpretiert, Anforderungen operationalisiert, passende Prüfmethoden ausgewählt und Code geschrieben werden (siehe Abbildung auf nächster Seite). Dieser Aufwand wiederholt sich potenziell bei jeder Normaktualisierung und bei jeder Änderung am KI-System.

Die Szenarien im nächsten Kapitel beschreiben, wie SMART Standards zu einer Vereinfachung dieses Aufwands, und somit zur Skalierbarkeit von KI-Prüfungen beitragen könnten.

27 Hernandez, J., Golpayegani, D., & Lewis, D. (2025). An open knowledge graph-based approach for mapping concepts and requirements between the EU AI act and international standards. *AI and Ethics*, 5(5), 4463–4474. <https://doi.org/10.1007/s43681-025-00708-6>

28 Golpayegani, D., Pandit, H. J., & Lewis, D. (2022). AIRO: An Ontology for Representing AI Risks Based on the Proposed EU AI Act and ISO Risk Management Standards. In *Studies on the Semantic Web*. IOS Press. <https://doi.org/10.3233/ssw220008>

29 <https://www.qi-digital.de/en/digital-certificates>

Aktivitäten mit Normenbezug bei der Prüfung von KI

Normen identifizieren



- Um die Rahmenbedingungen für eine normkonforme Prüfung festzulegen, müssen die einschlägigen Qualitätsdimensionen, gesetzlichen Anforderungen und relevanten Norminhalte zunächst identifiziert werden.
- Etwa durch Analyse des Zusammenhangs zwischen gesetzlichen Grundlagen und harmonisierten Normen, das Tracing zwischen Anforderungen über verschiedene Normen hinweg, oder der Identifikation von konkretisierenden Inhalten in anderen Dokumenten (z.B. Technical Specifications, Technical Reports, Guidelines).
- Manuelle Rechercheergebnisse müssen ggf. angesichts von Neuerscheinungen und Norm-Aktualisierungen regelmäßig überarbeitet werden.

Inhalte beziehen und integrieren



- Identifizierte Normen und einzelne Norminhalte müssen bezogen werden.
- Da viele Normen nicht frei verfügbar sind, müssen dafür zunächst Zugangsrechte erworben bzw. nachgewiesen werden.
- Manuelle Rechercheergebnisse müssen ggf. angesichts von Neuerscheinungen und Norm-Aktualisierungen regelmäßig überarbeitet werden.
- Einzelne Norminhalte, wie etwa Anforderungen, Definitionen, Formeln etc., die als prüfrelevant identifiziert wurden, müssen extrahiert und in eigene Systeme (z.B. Anforderungsmanagementsysteme) übertragen und integriert werden.
- Die Inhalte müssen hinsichtlich Änderungen in den zugrundeliegenden Dokumenten aktuell gehalten werden.

Prüfungen operationalisieren



- Aus extrahierten Norminhalten müssen funktionale und nicht-funktionale Anforderungen abgeleitet werden und messbare Kriterien und Kennzahlen identifiziert oder ausgewählt werden.
- Für automatisierbare Prüfungen müssen Prüfservices oder Codebibliotheken ausgewählt, Parameter gesetzt, Algorithmen ausgewählt und ggf. Prüfcode implementiert werden.
- Relevante Artefakte, wie Testdaten, müssen bezogen und integriert werden.

Konformität dokumentieren



- Prüfergebnisse, eingesetzte Methoden und referenzierte Normfassungen müssen so dokumentiert werden, dass sie gegenüber Auditoren, benannten Stellen oder Aufsichtsbehörden nachvollziehbar sind.
- Bislang erfolgt diese Dokumentation überwiegend manuell, oft in proprietären Formaten und mit redundanter Erfassung über Berichte, Zertifikate und Anforderungsmanagementsysteme hinweg.
- Bei jeder Aktualisierung des KI-Systems oder der zugrundeliegenden Normen muss diese Dokumentation überarbeitet und in den verschiedenen Zielsystemen synchronisiert werden.



Aufwändige Arbeitsschritte mit Normbezug können durch den Einsatz von **SMART Standards** vereinfacht werden.

→ Szenarien: Kapitel 3

3

SZENARIEN FÜR DEN EINSATZ VON SMART STANDARDS BEI DER PRÜFUNG VON KI-SYSTEMEN

3 SZENARIEN FÜR DEN EINSATZ VON SMART STANDARDS BEI DER PRÜFUNG VON KI-SYSTEMEN

SMART Standards können die Prüfung von KI-Systemen skalierbarer machen, indem sie prüfrelevante Inhalte maschinenlesbar und -interpretierbar machen, sie mit relevanten Ressourcen verknüpfen, und ihre Integration in relevante Abläufe vereinfachen. Ein durch SMART Standards angereichertes System kann insbesondere folgende Beiträge zur Vereinfachung von Aktivitäten mit Normbezug leisten:

1. Die Identifikation relevanter Normen und Inhalte vereinfachen,
2. Informationen zu inhaltlichen Änderungen und der Publikation von neuen relevanten Inhalten bereitstellen,
3. Norminhalte passgenau für die Integration in digitale Prozesse liefern,
4. Relevante Inhalte zuverlässig automatisiert nutzbar machen (etwa für die Verarbeitung durch KI),
5. den Bezug zu weiteren Hilfsmitteln für die Operationalisierung von Prüfungen herstellen, sowie,
6. Prüfergebnisse automatisiert reproduzierbar und interoperabel dokumentierbar machen.

Das kann insbesondere durch die Integration mit den in Kapitel 2 dargestellten Konzepten und Ressourcen erreicht werden:

- Norminhalte können besser für eine automatisierte Verarbeitung bereitstellen, um sie gut in MLOps Prozessen und für die Logik einer kontinuierlichen Konformitätsprüfung zugänglich zu machen.
- Wo möglich, kann einen Bezug zwischen SMART Standards und qualitätsgesicherten Ressourcen wie Prüfwerkzeugen, Referenzdaten, oder gesamten Prüfservices (-plattformen) hergestellt werden, um die Lücke zwischen Normen und deren Operationalisierung zu verkleinern.

Wie dieser Beitrag erzeugt und aus Anwenderperspektive eingesetzt werden könnte, zeigen die folgenden technisch-organisatorischen Szenarien. In den aufeinander aufbauenden Szenarien entwickelt sich der Einsatz von Normen von einer rechercheintensiven, dokumentbasierten Nutzung über die Integration von Norminhalten per API zu einem KI-unterstützten Prüfmanagement innerhalb eines Ökosystems für KI-Qualität.

ANWENDUNGSSZENARIEN:

3.1. API-basierte Auslieferung von maschinenles- und interpretierbaren Norminhalten

Zusätzlich zur Norm als PDF werden maschinenlesbare SMART Contents per API bereitgestellt. Semantische Inhalte ermöglichen die intelligente Identifikation relevanter Normen und die Verknüpfung mit anderen Datenquellen. Norminhalte können direkt in Informationssysteme übernommen und aktuell gehalten werden. Inhalte stehen zuverlässig für die Operationalisierung von Prüfungen und die Generierung von Prüfberichten zur Verfügung.

3.2. Zuverlässig KI-generierte Prüfartefakte

Die Durchführung von KI-Prüfungen wird durch einen KI-Assistenten unterstützt, der auf Grundlage des Prüfkontextes und relevanter Normen einschlägige Anforderungen identifiziert, Prüfpläne vorschlägt, Prüfberichte erstellt und, soweit möglich, ausführbaren Testcode generiert. Die Zuverlässigkeit bei der Verwendung von Normen durch die KI kann durch die Einbindung von Prüfwerkzeugen und -services, und die Verwendung von eindeutigem SMART Content erhöht werden.

3.3. Einbindung in ein Ökosystem für KI-Qualität

SMART Standards rücken in einer geteilten Infrastruktur näher an Ressourcen, die zu einer einheitlicheren Operationalisierung von KI-Prüfungen beitragen. Prüfungen, Prüfaufträge und benötigte Werkzeuge, Daten und Services werden vertrauenswürdig durch eine digitale KI-Qualitätsinfrastruktur integriert.

ERSTELLUNGSSZENARIO:

3.4 Skalierbare Generierung von SMART Content und Operationalisierungsinhalten

Um dem steigenden Bedarf an vielfältigem SMART Content zu begegnen wird dieser (teilweise) mit KI-Unterstützung aus natürlichsprachlichem Normtext gewonnen und validiert. (Referenzierte) Inhalte, die die Operationalisierung der Norm unterstützen, können aus dem Ökosystem für KI-Qualität beigetragen werden.

Skalierbare KI-Prüfung mit SMART Standards

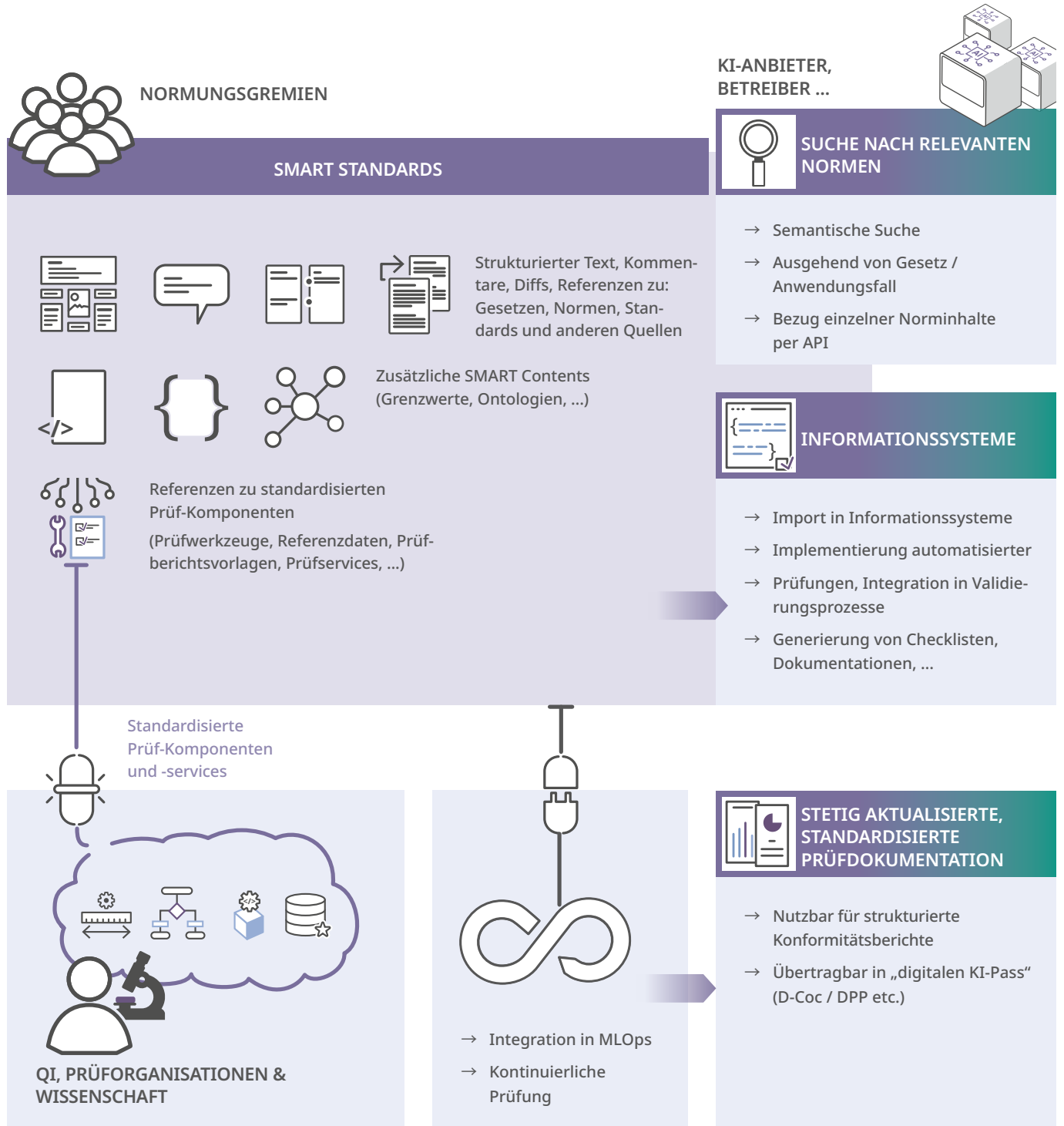


Abbildung 3

3.1 API-basierte Auslieferung von maschinenles- und interpretierbaren Norminhalten

Im ersten, grundlegenden Skalierungsszenario wird vor allem der Zugriff auf einzelne Norminhalte und deren Verarbeitung adressiert: Statt Normen als unstrukturierte Textdokumente manuell über Webportale zu beziehen, werden sie durch die Normungsorganisationen und ihre Verlage als SMART Content strukturiert über APIs bereitgestellt.

Einzelne Norminhalte, wie etwa Anforderungen, sind granular identifizier- und extrahierbar, und können ohne manuelle Vorverarbeitung in digitale Prozesse integriert werden. Zusammenhängende Inhalte werden über die Auflösung der Referenzen zwischen Normen integriert.

Inhalte werden versioniert zur Verfügung gestellt; aktualisierte Inhalte können automatisch neu bezogen werden. Es gibt einen „Normen-Changelog“ und es können „Diffs“ für Normen verschiedener Versionen angezeigt werden. Bei Aktualisierungen und Neuerscheinungen werden Bedarfsanalysen und Bezug vereinfacht.

Einzelne Bestandteile der Norm können von den Normerstellern mit zusätzlichen Metadaten, wie etwa thematischer Gruppierung, oder Kommentaren, angereicht werden. Komplexe Inhalte werden in spezialisierten maschinenlesbaren Formaten zur Verfügung gestellt, z.B. mathematische Formeln im MathML Format (siehe Tabelle 1). Wichtige, bislang nur natürlichsprachlich bereitgestellte Inhalte können mithilfe von Zusatzinhalten maschinenlesbar gemacht werden, etwa durch die explizite Auflistung von Werten oder Wertebereichen (XML/JSON).

Semantische Inhalte wie Ontologien, Vokabulare und Constraint-Modelle (z.B. in RDF, OWL, SHACL, SKOS) werden zusätzlich zu den eigentlichen Norminhalten ausgeliefert. Sie machen Beziehungen zwischen Begriffen, Anforderungen und Prüfobjekten explizit und erlauben es, Norminhalte mit anderen Datenquellen (Anforderungsmanagement, Wissensgraphen, Prüfartefakte) automatisiert zu verknüpfen. Konkrete Anwendungsfälle dafür ergeben sich vor allem in den folgenden Szenarien.³⁰

Damit lassen sich KI-Prüfungen einfacher und zuverlässiger umsetzen:

Relevante Normen intelligent identifizieren

Auf Basis der Angaben zum System werden systemgestützt einschlägige gesetzliche Vorgaben und harmonisierte Normen ermittelt, inklusive der Referenzkette von Gesetzen zu harmonisierten Normen und weiter zu verwiesenen Normen. Relevante Norminhalte (einzelne Anforderungen, Definitionen, Referenzen) werden gezielt bezogen. Mithilfe semantischer Inhalte lassen sich zudem komplexere Beziehungen zwischen Normen und dem eigenen Prüfkontext abbilden:

- Es werden komplexe Reports für die Identifikation relevanter Normen in eigenen Informationssystemen angelegt und regelmäßig automatisch aktualisiert. Etwa: „zeige mir alle Norminhalte bestimmter Normungsorganisationen die direkt oder indirekt vom AI Act referenziert werden und mit dem Tag ‚Datenqualität‘ versehen sind“.
- Änderungen an Normen oder neue Fassungen werden automatisiert erkannt und angezeigt, was die manuelle Recherche nach neuen Publikationen ersetzt.

Norminhalte automatisiert integrieren

Als relevant identifizierte Norminhalte werden automatisiert als granulare, maschinenlesbare Einheiten in lokale Informationssysteme übernommen. Sie können zum Beispiel ins Anforderungsmanagement integriert werden, etwa als übergeordnete Compliance-Anforderungen, als Constraints für Epics oder als Akzeptanzkriterien für User Stories. Jede integrierte Anforderung bleibt dabei eindeutig mit ihrer Quelle verknüpft (Normbezug, Abschnitt, Version) und kann automatisch aktualisiert werden.

- Alle Anforderungen an ein KI-System aus ausgewählten Normen werden per API ins Backlog übernommen (Jira, Confluence, Github, Gitlab, ...), inkl. Anforderungstext & -typ (muss, kann, soll), sowie Referenzen zu Normen und Gesetzestexten (Traceability). Alle Anforderungen und sonstige aus Normen extrahierten Texte können bei Veröffentlichung einer neuen Norm-Version aktualisiert und Änderungen angezeigt werden.

30 VDE SPEC 90010 V1.0, Anleitung zur Erstellung ausführbarer semantischer Standards (SemNorm)

Angebotene Inhalte (SMART Content)		
SMART Content	Beschreibung	Format-Beispiele
Normtext als granulare Einheiten	Strukturierte, klassifizierte Textbestandteile einer Norm, z.B. einzeln adressierbare Anforderungen mit eindeutiger ID, Normabschnitt, Version	XML (ReqIF)
Querverweise zwischen Normen und Gesetzen	Explizite Relationen, z. B. KI-VO → harmonisierte Norm → weitere referenzierte Normen und Standards	XML (ReqIF), RDF
Zusätzliche Metadaten	Tags, Kommentare, ...	XML (ReqIF), RDF
Grafiken		PNG, SVG, ...
Direkt verwendbare Werte	Numerische Grenzwerte, Wertebereiche, Schwellen, Klassifikationsgrenzen, Referenzbereiche, Einheiten, kategoriale Festlegungen, ...	XML, JSON, CSV, RDF
Mappings/Übersetzungstabellen	Zuordnungen zwischen Klassifikationssystemen, zwischen Einheiten oder Messmethoden, zwischen Versionsständen, ...	XML, JSON, SKOS, RDF
Einfache algorithmische und operative Inhalte	Formeln, Prüfalgorithmen, Entscheidungsbäume, Zustandsautomaten, ...	MathML / LaTeX, DMN, BPMN, ...
Semantische Informationen	Deklarative Semantik über Ontologien und strukturierte Anforderungen, operationale Semantik über ausführbare Modelle	RDF, OWL, SHACL, MathML, DMN, BPMN, ...

Schnittstellen	
Schnittstelle	Beschreibung
Such-API	Suche in Volltext & Metadaten, semantische Suche
Content-API (Content as a Service)	Abruf aller Inhalte einer Norm, bzw. gezielter Abruf eines Normabschnitts bzw. eines Zusatzinhalts per Identifikator, versioniert
Änderungs-API	Benachrichtigung bei Normänderungen; strukturiertes Delta zwischen Versionen.
RMS-Integration	Import von Anforderungen und anderen Textabschnitten in relevante Anforderungsmanagementsysteme (Jira, Confluence, GitLab u. a.) über ReqIF (Converter/Wrapper), oder ggf. direkt über CaaS API

Tabelle 1: Digitale Auslieferung von SMART Content für KI-Prüfungen (technisches Umsetzungsszenario)

Zuverlässig operationalisieren und strukturiert Prüfungen dokumentieren

Prüfrelevante Norminhalte werden operationalisiert und als automatisierte Tests und Monitoring-Checks in CI/CD- bzw. MLOps-Pipelines integriert. Ergebnisse fließen kontinuierlich in eine automatisch generierte Konformitätsdokumentation ein.

- Norminhalte werden direkt in Arbeitsumgebungen bereitgestellt, etwa in Notebooks oder der IDE, wo beim Schreiben von Validierungscode inline auf relevante Normabschnitte hingewiesen wird. Normbasierte Anforderungen sind damit frühzeitig und kontinuierlich sichtbar, statt erst spät im Projektverlauf wieder aufgegriffen zu werden.
- Gültige Wertebereiche, Parametereinstellungen, etc. für Tests werden automatisiert aus Normen in Implementierungen integriert, sodass Übertragungsfehler bei manueller Übernahme vermieden werden.
- Durch die Verknüpfung zwischen SMART Content und Prüfimplementierungen ist jederzeit nachvollziehbar, welche Tests sich auf welchen Norminhalt beziehen. Prüf-Trigger können bei Abweichungen automatisiert Workflows wie Retraining oder Alarmierung auslösen.
- Da Prüfungen und normbasierte Anforderungen eindeutig miteinander verknüpft sind, können Nachweise systemgestützt aggregiert, versioniert und ausgewertet werden, als Grundlage für first-party Konformitätsbewertungen oder die Vorbereitung externer Audits.

3.2 Zuverlässig KI-generierte Prüfarbefakte

Die Fähigkeit generativer KI-Systeme, aus strukturierten Vorgaben funktionalen Softwarecode zu erzeugen, verändert die Welt der Softwareentwicklung aktuell maßgeblich. Sie eröffnet auch für normenbasierte KI-Prüfungen neue Möglichkeiten (und Risiken).

Generative KI so zu nutzen, dass sie möglichst zuverlässig bei der Umsetzung von KI-Prüfungen auf Basis von Normen unterstützt, ist ein logischer Schritt Richtung Skalierbarkeit.³¹ In diesem Szenario sind maschinenlesbare Normen mit eindeutigen, qualitätsgesicherten Inhalten die Eingabe für einen LLM-Agenten, der daraus automatisiert Checklisten, Prüfpläne und -berichte oder direkt ausführbaren Prüfcode generiert, aber auch selbstständig Prüfungen durchführt.

Bereits jetzt lassen sich Teile dieses Szenarios mit gängigen Sprachmodellen für populäre KI-Normen umsetzen.³² Die große Herausforderung dabei ist, dass LLMs als statistische Sprachmodelle nicht automatisch den hohen Grad der Zuverlässigkeit bieten, der für die Anwendung von Normen Voraussetzung ist. Sie generieren plausible Formulierungen auf Basis von Trainingsdaten, verfügen aber weder über eine verifizierbare Wissensbasis noch über einen definierten Verwendungskontext der von ihnen reproduzierten Aussagen.

Ein LLM könnte also Anforderungen paraphrasieren und verfälschen, es könnte Begriffe verwenden, ohne deren standardisierte Definition zu referenzieren, und es erfindet im Zweifel plausibel klingende Textpassagen oder sogar Werte und Formeln („Halluzination“).

Dennoch ist der Einsatz von KI für die Automatisierung von Konformitätsprüfungen auf Basis von Regelwerken ein in

verschiedenen regulierten Domänen verfolgtes Konzept. So gibt es etwa im Bereich der DSGVO, der Finanzregulierung, oder für das Baurecht entsprechende Forschungsprojekte.³³ Auch im Bereich der KI-Regulierung entsteht semantische Modellierung, die zu einer Automatisierung von „Compliance-Checks“ beitragen soll.³⁴

Um die Qualität von Prüfarbefakten zu erhöhen, die aus Normen generiert werden, bietet sich der Einsatz von neurosymbolischer KI an, also einer Kombination aus einem statistischen und einem regelbasierten Ansatz.³⁵ Eine solche KI verbindet die Flexibilität und Generalisierungsfähigkeit neuronaler Netze mit der Verifizierbarkeit und formalen Garantie symbolischer Systeme und könnte mithilfe zusätzlicher semantischer Informationen, und weiterer, deterministisch verwendeter strukturierter Daten aus SMART Standards bessere Ergebnisse liefern.

Semantische Informationen aus SMART Standards könnten dabei auf verschiedene Art und Weise in die Generierung von Prüfarbefakten einfließen. Entsprechend eines „knowledge infusion“ Konzepts³⁶ wäre zum Beispiel der Einsatz eines spezialisierten „Norm-LLMs“³⁷ denkbar, das zusätzlich zu natürlichsprachlichen Norminhalten mit den semantischen Informationen aller (relevanten) Normen trainiert wird („knowledge infused learning“). Zum anderen ist ein Ansatz denkbar, bei dem die relevanten semantischen Informationen bei Bedarf in den Generierungsprozess eines generischen (Foundation-) Sprachmodells integriert werden („knowledge infused inference“ / „inference stage enhancement“).

31 Einen verwandten Ansatz außerhalb der KI-Prüfung verfolgt das Projekt FAICE (Framework for AI-Enabled Collaborative Engineering, [prostep ivip](https://www.prostep.org/projekte/framework-for-ai-enabled-collaborative-engineering-faice)). Dort wird ein „Standards-Chatbot“ entwickelt, der einen KI-gestützten, granularen Zugriff auf Norminhalte ermöglichen soll. Siehe <https://www.prostep.org/projekte/framework-for-ai-enabled-collaborative-engineering-faice>

32 Beispiel: Generieren einer Checkliste gemäß ISO 42001 und Überprüfung auf Basis lokaler Dokumentation mithilfe von Claude / ChatGPT etc.

33 Z.B.: Using Large Language Models for the Interpretation of Building Regulations. (2024). Journal of Engineering, Project, and Production Management. <https://doi.org/10.32738/jepm-2024-0035>

34 Z.B.: AIRO (AI Risk Ontology), <https://delaramglp.github.io/airo/>

35 Garcez, A. d'Avila, & Lamb, L. C. (2023). Neurosymbolic AI: the 3rd wave. Artificial Intelligence Review, 56(11), 12387–12406. <https://doi.org/10.1007/s10462-023-10448-w>

36 Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2024). Unifying Large Language Models and Knowledge Graphs: A Roadmap. IEEE Transactions on Knowledge and Data Engineering, 36(7), 3580–3599. <https://doi.org/10.1109/tkde.2024.3352100>

37 Hierzu gab es bei IdIS ein entsprechendes Pilotprojekt, bei dem jedoch zunächst keine semantischen Informationen und keine neurosymbolische KI zum Einsatz kamen.

Grundsätzlich lassen sich zwei Einsatzszenarien ableiten:

a) Zentrales, auf SMART Standards trainiertes „Normen-LLM“

Ein auf allem verfügbaren SMART Content trainiertes neurosymbolisches KI-Modell wird zentralisiert angeboten (z.B. durch eine oder mehrere Normungsorganisationen), beispielsweise als Service (z.B. als Chatbot und/oder per API). Zu prüfende und andere kontextrelevante Inhalte der Anwenderorganisation werden der KI zur Laufzeit für die Inferenz zur Verfügung gestellt.

Und/oder:

b) Generisches LLM + SMART Standards RAG

Eine Anwenderorganisation verwendet ein generisches (ggf. neurosymbolisches) Sprachmodell, das mit den lokalen Kontextdaten des zu prüfenden Systems ausgestattet ist. Das Modell wird zur Laufzeit über RAG/MCP-Mechanismen mit von Normungsorganisationen bereitgestellten relevanten SMART Standards angereichert und kann semantische Inhalte für die Inferenz verwenden.

Neben semantischen Informationen kommen bei der Generierung von Prüfartefakten in beiden Fällen die weiteren durch SMART Standards bereitgestellten, strukturierten Zusatzinhalte zum Einsatz:

Einbindung strukturierter Zusatzinhalte als „ground truth“

Fest definierte Werte, Mappings, und klar definierte algorithmische und andere operative Inhalte (siehe Tabelle 1), werden von der KI als strukturierte Daten aus SMART Standards übernommen. Die Notwendigkeit, sie aus dem Text zu extrahieren, oder aus Trainingswissen zu rekonstruieren entfällt. Das Modell übernimmt die Werte direkt und verwendet sie für die Generierung von Prüfartefakten. Zum Beispiel: Direktes Einsetzen von Grenzwerten, Implementierung von Metriken aus mathematischen Formeln. Werte sind unverändert aus der Norm übernommen und bleiben auf den jeweiligen Normabschnitt rückverfolgbar. So werden „normative Wahrheit“ und Generierungslogik getrennt, sodass „Halluzinationen“ an dieser Stelle reduziert werden können.

Referenzierung weiterer Prüffressourcen und -tools

Es ist absehbar, dass Normen und entsprechende Zusatzinhalte (wie operationale semantische Informationen) nicht immer einen Informationsgehalt anbieten können, der für

die vollständige Operationalisierung einer Prüfung ausreichend ist. Beispielsweise könnten Normen Prüfkonzepte beschreiben, die für sehr spezifische Prüfkontexte auf verschiedene Arten umgesetzt werden können. Genaue Implementierungsanweisungen (z.B. die Auswahl der Code-Bibliothek für implementierte Standardalgorithmen) übersteigen den Detailgrad einer Norm, können aber für letztendliche Prüfergebnisse entscheidend sein und Auswirkungen auf Vergleich- und Reproduzierbarkeit und Korrektheit haben.

Solche Unsicherheiten können und sollten durch die KI entsprechend für die Qualitätsbewertung durch einen Menschen markiert werden. Eine automatisierte Qualitätsbewertung könnte aber auch implizit über den Einsatz extern qualitätsgesicherter Ressourcen einfließen, die bei besonders komplexen, uneindeutigen, bzw. unterspezifizierten Anforderungen fertige Implementierungen anbieten:

Für spezialisierte Prüfungen, wie etwa statistische Tests auf Fairness, Robustheitstests oder Sicherheitsanalysen, können standardisierte und ggf. zertifizierte Prüfwerkzeuge (vgl. DIN SPEC 92006) als referenzierte Services direkt im SMART Content hinterlegt werden (siehe auch nächstes Szenario).

Anwendungsszenario: KI-generierte Prüfinhalte

Das dargestellte Szenario ermöglicht die Umsetzung verschiedener, auf SMART Standards basierender Anwendungsfälle. Im Fokus steht dabei stets die Verwendung einer KI dafür, im Kontext von KI-Prüfungen Informationen aus Normen zu verarbeiten, Prüfartefakte zu generieren (Code, Checklisten, Prüfpläne) und Prüfungen auf deren Basis durchzuführen.

Die Verantwortung für die Qualität der Prüfungen bleibt dabei beim Menschen. Generierte Artefakte werden vor dem produktiven Einsatz durch qualifizierte Fachpersonen überprüft und freigegeben. KI beschleunigt die Implementierung, ohne das fachliche Urteil zu ersetzen.

- Durch Interaktion im Stil eines Chatbots geben Prüferinnen und Prüfer an, welches KI-System geprüft werden soll und nach welchen Normen. Eine KI identifiziert alle prüfbaren Anforderungen und generiert einen strukturierten Prüfplan, in dem für jede Prüfkategorie Verweise auf die zugrundeliegenden Normabschnitte angegeben werden.
- Anforderungen, die automatisiert prüfbar sind (z. B. durch definierte Schwellenwerte, Metriken, Pflichtfelder

in Dokumentationen), werden von solchen unterschieden, die manuelles Urteil erfordern.

- Für automatisiert prüfbare Anforderungen werden Prüfergebnisse vorgeschlagen (falls aus dem Kontext / lokaler Dokumentation und Code ableitbar), oder es wird Prüfcode generiert (und ggf. ausgeführt).
- Nicht automatisiert prüfbare Anforderungen werden durch Frage & Antwort, bzw. Einbinden weiterer Dokumente im Chat Kontext überprüft.
- Der Prüfplan kann zur regelmäßigen Durchführung in die lokale Infrastruktur eingebunden und mit einem

KI-Agenten verbunden werden: Automatisierte Prüfungen werden in das MLOps-Framework integriert. Dokumentationen zu manuell durchzuführenden Prüfungen werden in einem Informationssystem abgelegt. Ein KI-Agent erinnert Prüfer per Mail / Ticket-System zu relevanten Zeitpunkten an die Durchführung der Prüfung.

3.3 Einbindung in ein Ökosystem für KI-Qualität

In den Szenarien 1 und 2 greifen Unternehmen auf eine Vielzahl an Ressourcen unterschiedlicher Anbieter zu: SMART Standards von Normungsorganisationen, Prüfwerkzeuge und Referenzdatensätze von Prüfinstituten und Forschungseinrichtungen, Akkreditierungsnachweise von zuständigen Stellen sowie weitere Inhalte aus der Wertschöpfungskette der KI-Entwicklung. Bereits heute existiert eine große Zahl an Open-Source-Werkzeugen und -Datensätzen für die Prüfung von KI-Systemen. Allein der OECD Catalogue of Tools & Metrics for Trustworthy AI listet mehrere Hundert Einträge auf.³⁸

Werden solche Ressourcen für die Prüfung auf Normkonformität eingesetzt, kann eine zentrale Qualitätssicherung durch qualifizierte Stellen mehrere Vorteile bieten:

- **Operationalisierung und Bezug zu Normen.** Bei der Operationalisierung von Normen bleiben Auswahl von Algorithmen, Bibliotheken, Parametern, Schwellenwerten und Validierungsdatensätzen als Entscheidungen beim Anwender, ebenso der Bezug zu konkreten Normabschnitten. Eine qualifizierte Stelle könnte Open-Source-Bibliotheken und andere Implementierungen für die Prüfung gegen eine bestimmte Norm verifizieren und das Ergebnis als signiertes, maschinenlesbares Eignungsattest an die Ressource binden. Anwender lösen damit direkt aus dem SMART Standard heraus auf, welche Werkzeuge in welcher Version für eine Anforderung geeignet sind, statt diese Bewertung selbst zu führen.

- **Förderung des Ökosystems.** Mit einer solchen zentralen Verifizierung erhalten Open-Source-Projekte und Tool-Anbieter einen Anreiz, ihre Implementierungen an etablierten Qualitätsanforderungen auszurichten und sich um eine entsprechende Bewertung zu bewerben. Das stärkt die Community und vermeidet, dass Prüfwerkzeuge parallel in verschiedenen Organisationen neu entwickelt werden müssen.
- **Vergleichbarkeit von Prüfergebnissen.** Ohne abgestimmte Werkzeugversionen, Default-Parameter und Referenzdaten können bei identischer Eingabe abweichende Ergebnisse entstehen. Konformitätsbewertungen und gegenseitige Anerkennung setzen jedoch zuverlässige Vergleichbarkeit voraus.
- **Verlässlichkeit der Werkzeuge selbst.** Prüfwerkzeuge sind Software und können fehlerhaft sein. Eine fehlerhafte Fairness-Metrik oder ein numerisch instabiler Robustheitstest produziert systematisch falsche Aussagen. In sicherheitskritischen Anwendungsfeldern ist ihre Verlässlichkeit nicht weniger wichtig als die der geprüften Systeme selbst.
- **Schutz vor Datenkontamination.** Werden Referenz- und Testdatensätze öffentlich verfügbar gemacht, fließen sie über Trainingsdaten in spätere Modelle ein und verlieren als unabhängiges Prüfmittel ihre Aussagekraft. Eine zentrale Verwaltung erlaubt es, Testdaten gezielt vorzuhalten und nur über kontrollierte Mechanismen zugänglich zu machen.

38 <https://oecd.ai/en/catalogue/tools>

→ **Effizienz auf Anwenderseite.** Ohne verlässliche Empfehlungen muss jedes Unternehmen eigene Kapazitäten für Recherche, Bewertung und Validierung externer Werkzeuge aufbauen, was insbesondere kleinere Anbieter überfordert. Eine Bewertung durch qualifizierte Stellen macht aus dieser individuellen Last eine kollektive Aufgabe.

→ **Akkreditierungs- und Aufsichtskonformität.** Akkreditierte Konformitätsbewertungsstellen müssen die Eignung ihrer Prüfmittel nachweisen; ohne zentrale Qualitätssicherung trägt jede Stelle diese Last allein. Eine geteilte Bewertung erleichtert zugleich die gegenseitige Anerkennung über Sektor- und Ländergrenzen hinweg.

Auf diese Lage reagiert die digitale Qualitätsinfrastruktur bereits mit konkreten Vorhaben. Die QI-Digital-Roadmap formuliert die Vision einer „Vertrauenswürdigen KI mithilfe von Prüfplattformen und KI-Pass“, in der KI-Systeme in abgesicherten Umgebungen einer Prüfplattform ausgeführt oder über definierte Schnittstellen integriert werden, mit automatisiertem Abgleich gegen Referenzwerte. Erste Vorarbeiten zu Plattform, Referenzdaten und methodischer Grundlage liegen bereits vor.³⁹ Diese Vorhaben adressieren die zentrale Qualitätssicherung von Werkzeugen, Daten und Diensten als Teil der laufenden Transformation der Qualitätsinfrastruktur.

An der Stelle setzt dieses Szenario an und erweitert die zentralen Bausteine der QI-Digital-Vision um den Aspekt der kontrollierten, vertrauenswürdigen Vermittlung dieser Ressourcen zwischen vielen Anbietern und Konsumenten.

Die konkrete Ausgestaltung wirft eine Reihe offener Fragen auf, deren Beantwortung ein abgestimmtes Vorgehen der Beteiligten voraussetzt:

- Wie werden Werkzeuge, Datensätze und Services über Anbietergrenzen hinweg auffindbar und vergleichbar?
- Wer übernimmt die Qualitäts- und Eignungsbewertung, und wie werden Empfehlungen qualifizierter Stellen sichtbar und überprüfbar gemacht?
- Wie wird der Bezug zwischen Werkzeug und Norm explizit hergestellt und bei Norm-Updates nachgeführt?

- Unter welchen Zugangsbedingungen, Lizenzen und Abrechnungsmodellen werden Ressourcen bereitgestellt, auch wenn ihre Nutzung kostenpflichtig ist?
- Wie werden Referenz- und Testdatensätze so bereitgestellt, dass ihre Integrität trotz wiederholter Nutzung über die Zeit erhalten bleibt?

Datenraum als geteilte Infrastruktur

Da Einzelvereinbarungen zwischen verschiedenen Akteuren sehr aufwändig wären, macht hier der Einsatz eines Datenraums als geteilte Infrastruktur Sinn, in der Auffindbarkeit, Vertrauen, Zugangskontrolle und Abrechnung zwischen allen Teilnehmern in einheitlicher Form geregelt sind. **Tabelle 2** gibt eine strukturierte Übersicht zu möglichen beteiligten Rollen. **Tabelle 3** listet verschiedene im Ökosystem relevante Ressourcen auf.

Drei zentrale Prinzipien dieses Szenarios sind:

- **Datensouveränität.** Jeder beteiligte Akteur der QI behält die Kontrolle über die von ihm bereitgestellten Ressourcen, ohne sie in eine zentrale Plattform übergeben zu müssen. Normungsorganisationen können Lizenzen und Zugriffsmodelle für Norminhalte selbst durchsetzen. Anbieter von KI-Prüfwerkzeugen können kostenpflichtige Lizenzen anbieten oder den Zugriff auf akkreditierte Stellen beschränken. KI-Hersteller können Modellgewichte und proprietäre Implementierungen schützen. Anbieter von Referenzdatensätzen können diese gegen Kontamination absichern.
- **Föderierte Auffindbarkeit.** SMART Standards der Normungsorganisationen, KI-Prüfwerkzeuge und Referenzdatensätze, oder Prüfservices von Prüfinstituten verbleiben bei ihren jeweiligen Anbietern und werden über einen verteilten Katalog auffindbar gemacht. Der Bezug zu konkreten Normabschnitten und KI-Anwendungsdomänen kann dabei maschinenlesbar mitgeführt werden.
- **Vertrauen durch verifizierbare Identitäten.** Akkreditierungen, Zertifizierungen durch Konformitätsbewer-

³⁹ https://www.qi-digital.de/fileadmin/user_upload/website/Roadmap/QID_Roadmap_V1.0_20251008.pdf. Die Roadmap führt unter anderem die laufende Entwicklung einer Plattform zur objektiven Prüfung und Qualitätsbewertung von KI-Algorithmen durch die PTB an, sowie das PTB-Pilotprojekt "Künstliche Intelligenz in der Medizin" mit Arbeiten an synthetischen Referenzdaten und anwendungsfallspezifischen Referenzdatensätzen. Als methodische Grundlage referenziert sie das METRIC-Framework, vgl. Schwabe, D., Becker, K., Seyferth, M., Klaub, A., Schäffter, T. (2024): The METRIC-framework for assessing data quality for trustworthy AI in medicine: a systematic review. npj Digital Medicine 7, 203, <https://doi.org/10.1038/s41746-024-01196-4>.

tungsstellen und Eignungsbewertungen für KI-Prüfwerkzeuge durch qualifizierte Stellen werden als signierte Verifiable Credentials geführt. Anwender können damit nachvollziehen, ob eine Werkzeug-Empfehlung tatsäch-

lich von der angegebenen Stelle für die angegebene Norm stammt, und Zugriffe auf zertifizierte Prüfservices lassen sich auf akkreditierte Stellen begrenzen.

Rolle	Akteure (Deutsche Beispiele)	Beitrag im Ökosystem
Normungsorganisationen	DIN, DKE	Bereitstellung von SMART Standards; Quelle normativ verankerter Vokabulare für die semantische Beschreibung von Werkzeugen, Datensätzen und Attestierungen
Anbieter von Prüfwerkzeugen und Referenzdaten	Forschungseinrichtungen, Open-Source-Communities, kommerzielle Anbieter	Entwicklung und Bereitstellung von KI-Prüfwerkzeugen und Referenzdatensätzen; Verknüpfung mit Normabschnitten über Katalog-Metadaten
Akkreditierungsstellen	DAkkS	Akkreditierung von Konformitätsbewertungsstellen; Ausstellung verifizierbarer Akkreditierungsnachweise
Qualifizierte Stellen	Stellen mit nachgewiesener fachlicher Kompetenz für die Bewertung von KI-Prüfwerkzeugen	Eignungsattestierung von Werkzeugen für bestimmte Normabschnitte; Ausstellung signierter Verifiable Credentials
Konformitätsbewertungsstellen	Akkreditierte und benannte Stellen, sowie übergreifende Governance Projekte	Bereitstellung von Compliance-Katalogen und Prüfungs-Workflows, Durchführung von Konformitätsbewertungen unter Nutzung qualitätsgesicherter Werkzeuge; Betrieb zertifizierter Prüfservices, gegebenenfalls im Compute-to-Data-Modus, Harmonisierung der nötigen Dienstleistungen durch ein domänenspezifisches „Rulebook“
KI-Hersteller und Inverkehrbringer	Unternehmen, die KI-Systeme entwickeln und in Verkehr bringen	Bereitstellung der zu prüfenden Modelle (lokal oder im Compute-to-Data-Modus); Konsum von Werkzeugen, Referenzdaten und Attestierungen aus dem Datenraum
KI-Anwender	Organisationen, die KI-Systeme einsetzen	Auswahl und Beschaffung konformer KI-Komponenten anhand der im Datenraum aufgelösten Bezüge
Aufsichtsbehörden	Marktüberwachungsbehörden, Datenschutzaufsicht	Stichprobenartige Überprüfung; Verifikation von Attestierungen und Konformitätsbescheinigungen
Plattformbetreiber	Konsortium eines Datenraums entlang des Konzept-Frameworks Quality-X	Betrieb der föderierten Infrastruktur einschließlich Konnektoren, Katalog, Trust- und Vokabular-Diensten mit Beteiligung von QI-Institutionen

Tabelle 2: Mögliche Akteure und Rollen in einem Ökosystem für KI-Qualität

Bausteine und Erweiterungen für die Prüfung von KI im Quality-X Konzept

Mit Quality-X entsteht bereits ein Datenraum-Konzept speziell für die Qualitätsinfrastruktur.⁴⁰ Quality-X betrachtet dabei zentrale Bausteine, an die ein Ökosystem für die qualitätsgesicherte KI-Prüfung anknüpfen könnte, wie etwa Konnektoren für den souveränen Austausch (etwa Eclipse EDC), einen föderierten Katalog für die Auffindbarkeit verteilter Ressourcen (etwa auf Basis von DCAT-AP), geteilte Vokabulare für deren semantische Beschreibung, sowie Identity- und Trust-Services für die Verifikation von Akteuren und Bescheinigungen.

Die konkrete Ausgestaltung dieser Bausteine für die KI-Prüfung liegt jenseits des Fokus dieses Whitepapers. Einige Anknüpfungspunkte für SMART Standards lassen sich aber schon benennen:

- Konnektoren ermöglichen Berechnungen, ohne dass sensible Ressourcen geteilt werden müssen. Im Compute-to-Data-Modus verbleiben Modellgewichte beim Hersteller und Black-Box-Testdatensätze bei der qualifizierten Stelle. Dabei werden nur die Prüfanforderung und das signierte Ergebnis ausgetauscht. Eine in einem SMART Standard verankerte Anforderung führt so unmittelbar zu einer Prüfausführung, deren Provenienz, Werkzeugversion und Eingangsdaten am Konnektor protokolliert sind.
- Föderierter Katalog: Für KI-Prüfwerkzeuge und Referenzdatensätze sind maschinenlesbare Bezüge zu Normabschnitten und KI-Anwendungsdomänen zu hinterlegen (zusätzliche Metadaten im DCAT-AP-Vokabular), damit sich aus einer konkreten Anforderung heraus passende Ressourcen auflösen lassen.⁴¹ Konkret könnte ein Anwender ausgehend von einer Anforderung an Robustheit aus einem KI-bezogenen SMART Standard sowohl geeignete Prüfwerkzeuge als auch passende Referenzdatensätze auflösen, jeweils mit Hinweis auf die zugehörige Eignungsbewertung und die abgedeckte Anwendungsdomäne.
- Bei den geteilten Vokabularen könnten SMART Standards sowohl für Inhalte verwendet werden, die im Datenraum referenziert werden, als auch als Quelle normativ verankerter Begriffe (etwa für KI-Testverfahren, Qualitätsdimensionen, Anwendungsdomänen), mit denen Werkzeuge, Datensätze und Attestierungen einheitlich beschrieben werden. Auf diese Weise würde derselbe Begriff, der in einer Norm definiert ist, auch im Katalogeintrag eines Prüfwerkzeugs und im Credential einer Eignungsbewertung verwendet, ohne dass eine separate Übersetzungsschicht zwischen Norm und Datenraum gepflegt werden müsste.
- Die organisatorischen und rechtlichen Anforderungen aus SMART Standards und Datenraum-Prozessen, sowie Beschränkungen in Confidential-Computing, Compute-to-data oder Secure Processing Environments (SPE) könnten beispielsweise mithilfe der Open Digital Rights Language⁴² vereinheitlicht werden.
- Bei den Trust-Services käme zur bekannten Verifikation von Akkreditierungen und Konformitätsbescheinigungen, für die mit dem Digital Certificate of Conformity (D-CoC) im Rahmen von QI-Digital bereits ein maschinenlesbares Format entwickelt wird, der Bedarf hinzu, auch Eignungsattestierungen für KI-Prüfwerkzeuge maschinenlesbar mitzuführen (als signierte Verifiable Credentials, am Werkzeug im Katalog hinterlegt). Ein Konsument könnte vor der Nutzung automatisiert prüfen, ob ein Werkzeug für die einschlägige Norm in der vorliegenden Version durch eine qualifizierte Stelle bewertet wurde, ohne diesen Nachweis manuell aus separaten Quellen zusammentragen zu müssen.
- Über die heutige Funktionalität hinaus müssten Werkzeuge selbst auffindbar, beziehbar und ausführbar sein, zunehmend auch in agentenbasierter Form (mit Schnittstellen wie dem Model Context Protocol). Ein agentenbasiertes Prüfwerkzeug könnte aus einem SMART Standard heraus die einschlägigen Anforderungen lesen, im Katalog passende Subwerkzeuge und Datensätze auflösen und die Prüfung am Ort der Daten anstoßen, ohne dass diese Schritte manuell verkettet

40 Quality-X: föderierte Plattform für die Qualitätsinfrastruktur, QI-Digital-Roadmap, Kapitel 2.4.2; ergänzende Informationen unter <https://www.qi-digital.de/digitale-qi/quality-x>.

41 Mit Daseen liegt aus dem Projekt MISSION KI bereits ein Proof-of-Concept für einen föderierten Datensatzkatalog vor, an den eine Erweiterung um Werkzeuge, Eignungsattestierungen und Norm-Bezüge anknüpfen könnte: <https://github.com/Mission-KI/Dataset-Search-Engine>

42 <https://www.w3.org/TR/odrl-model/>

KI-unterstützte Generierung von SMART Content

MANUELLE NORMERSTELLUNG & QUALITÄTSSICHERUNG VON SMART CONTENTS

AUTOMATISCHE GENERIERUNG VON SMART CONTENT ENTWÜRFEN

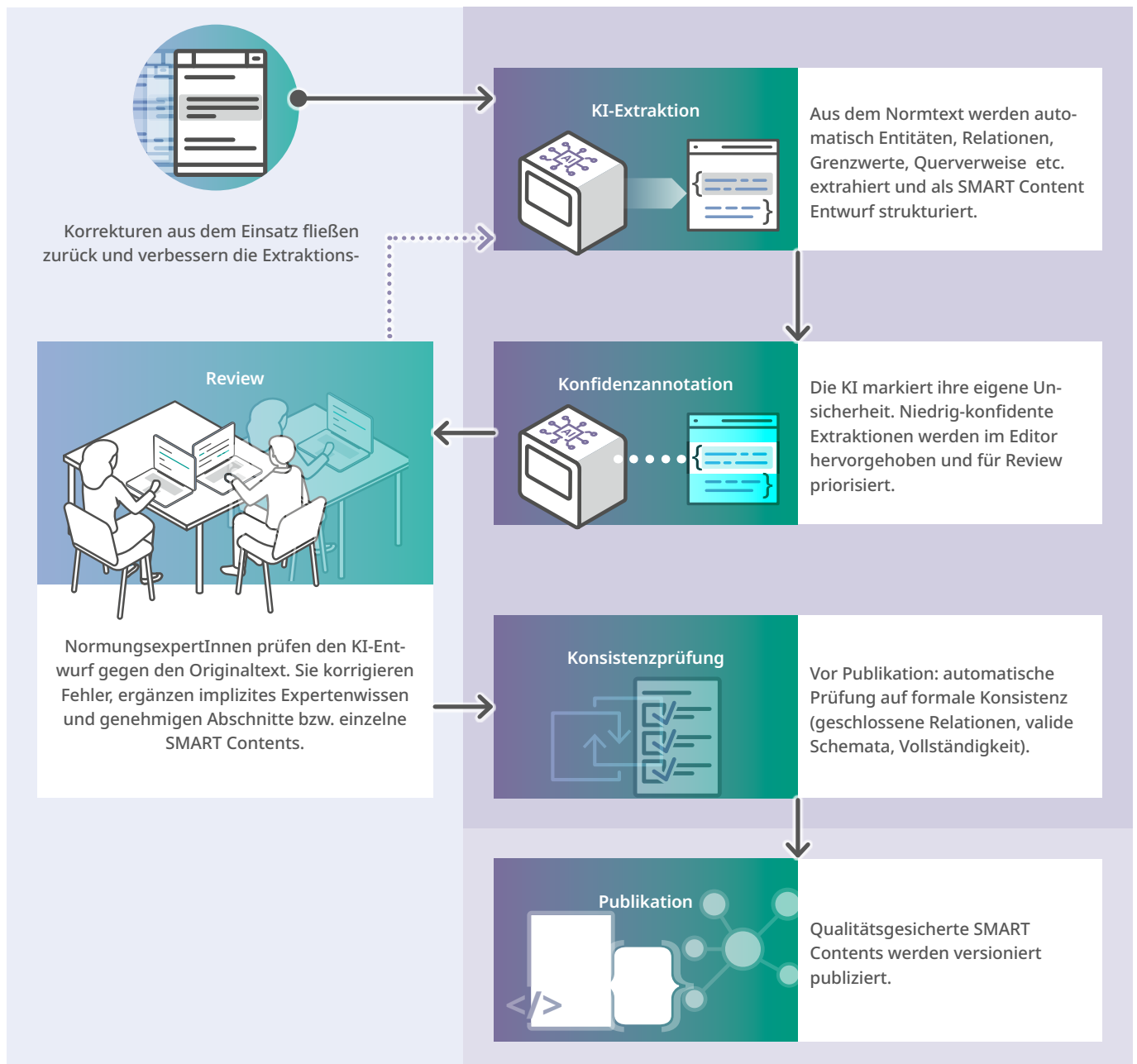


Abbildung 4

werden müssten. Der in Quality-X vorgesehene Marktplatz wäre ein naheliegender Anknüpfungspunkt. Die technische Umsetzung müsste im Detail geklärt werden.

- Ein Datenraum der digitalen Qualitätsinfrastruktur für KI-Prüfungen könnte in Kombination mit domänenspezifischen Datenräumen⁴³ auf die jeweiligen Anforderungen der Domäne angepasst werden. Ein Nachweis, der in einem Datenraum ausgestellt wurde, könnte in einem anderen direkt verifiziert und anerkannt werden, ohne dass die ausstellende Stelle erneut eingebunden werden muss. Compliance-Freigaben und Prüfnachweise würden datenraumübergreifend als verifizierte Credentials in einen gemeinsamen Workflow einfließen, ohne Medienbrüche an den Grenzen der beteiligten Infrastrukturen zu verursachen. So würde ein gemeinsames Ökosystem aus spezialisierten Datenräumen entstehen. Hierfür könnte etwa das Dataspace Protocol⁴⁴ für die bilaterale Aushandlung von Nutzungsrechten und Compliance-Anforderungen zwischen dezentralen Teilnehmern verwendet werden. Verifiable Credentials⁴⁵ und das Decentralized Claims Protocol⁴⁶ könnten die Grundlage für organisatorische Interoperabilität bilden.

KI-Prüfung lokal oder im Datenraum

Für die Ausführung von KI-Prüfwerkzeugen im Ökosystem sind zwei komplementäre Szenarien denkbar, die sich in der Frage unterscheiden, wo die Berechnung stattfindet, am Ort des Anwenders oder am Ort der zu schützenden Ressource. Beide Szenarien haben ihre Berechtigung und hängen von den Schutzbedürfnissen der beteiligten Ressourcen ab.

Im lokalen Szenario würde ein Prüfwerkzeug aus dem Datenraum bezogen und in der Infrastruktur des Anwenders ausgeführt. Dieses Szenario wäre naheliegend, wenn weder die geprüften Modelle noch die Testdaten besonders schutzbedürftig sind, etwa wenn ein Hersteller seine eigenen Modelle gegen öffentlich verfügbare Referenzdaten prüft oder

einen Konformitätsnachweis in einer kontrollierten internen Umgebung erbringt. Auch die frühe Phase der Konformitätsprüfung, in der ein Werkzeug zunächst mit eigenem Material erprobt wird, eignete sich für die lokale Ausführung.

Im Compute-to-Data Szenario verbleiben Daten und Modell an ihrem Ursprungsort, das Prüfwerkzeug wird dort ausgeführt. Nur die Prüfanforderung und das signierte Ergebnis werden ausgetauscht. Dieses Szenario erlaubt eine in der KI-Prüfung besonders relevante symmetrische Anwendung der Datensouveränität. Modellgewichte verbleiben beim Hersteller, sodass die geprüfte KI ihre Infrastruktur nicht verlassen muss. Black-Box-Testdatensätze verbleiben bei der qualifizierten Stelle und behalten so ihre Integrität gegen wiederholte Nutzung und Kontamination. Zertifizierte Prüf-services akkreditierter Stellen könnten sich entsprechend so betreiben lassen, dass ihre interne Logik aus dem Datenraum nicht ausgelesen werden kann. Der Anwender würde am Ende ein nachvollziehbares Ergebnis erhalten, ohne Zugang zum Inneren des Werkzeugs oder zu den Referenzdaten zu bekommen.

Beide Szenarien könnten sich in dedizierten Prüfplattformen bündeln lassen, die von qualifizierten Stellen betrieben oder Open Source als hostbare Software bereitgestellt werden. Aus Sicht des Datenraums wären solche Plattformen genauso über dieselben Konnektoren, Katalog-Einträge und Eigenschaftsattestierungen ansprechbar wie einzelne Werkzeuge.

In bestehenden Datenraum-Initiativen finden sich für den Compute-to-Data-Modus bereits konkrete Vorarbeiten. Der Extended Dataset Profile Service (EDPS) ist ein Dienst, der am Ort der Daten ausgeführt wird, dort standardisierte Datenqualitätsprofile erzeugt und diese über die föderierte Suchmaschine Daseen auffindbar macht, ohne dass die Originaldaten ihre Umgebung verlassen.⁴⁷ Es gibt außerdem Open-Source-Prototypen für on-premise ausgeführte KI-Prüfwerkzeuge, die Datenaustauschvorgänge gegen Vorgaben aus umfangreichen Regelwerken prüft.⁴⁸

43 wie der European Health Data Space (EHDS) oder der sphin-X Datenraum; Sphin-X: one data space for health <https://sphin-x.de/>

44 ISO/IEC DIS 26450

45 <https://www.w3.org/TR/vc-overview/>

46 <https://github.com/eclipse-dataspace-dcp/decentralized-claims-protocol>

47 Extended Dataset Profile Service (EDPS) und föderierte Suchmaschine Daseen, entwickelt im Rahmen der Initiative MISSION KI als Compute-to-Data-Dienst. Open Source verfügbar unter <https://github.com/Mission-KI/Dataset-Search-Engine>; ergänzende Beschreibung unter <https://mission-ki.de/de/datenraeume-vernetzen/data-set-search-engine>.

48 Compliance Monitor, Proof-of-Concept eines KI-basierten on-premise-Prüfsystems für Datentransaktionen unter Vorgaben des Data Act, Data Governance Act und DORA, entwickelt von Noxtua (ehem. Xayn) im Rahmen der Initiative MISSION KI. Sprachmodell und Dokumentation verfügbar unter <https://huggingface.co/ACATECH/ncos>; weitere Informationen unter <https://mission-ki.de/de/datenraeume-vernetzen/compliance-monitor>.

Ressource	Beschreibung	Format/Standard (Beispiele)
SMART Standards	Maschinenlesbare Norminhalte mit Anforderungen, Prüfvorgaben und Definitionen	XML (ReqIF), perspektivisch weitere Formate
Domänenspezifische Vokabulare und Ontologien	Begriffssysteme für KI-Testverfahren, Qualitätsdimensionen und Anwendungsdomänen, abgeleitet aus SMART Standards oder ergänzend dazu	RDF, OWL, SKOS
KI-Prüfwerkzeuge	Software zur automatisierten Prüfung von KI-Systemen gegen Normvorgaben; Listen an Compliance-Anforderungen für konkrete Nutzungsszenarien;	Open-Source-Repositories, Container-Images (OCI); Katalog-Beschreibung in DCAT-AP / JSON-LD
Referenzdatensätze	Kuratierte Datensätze für Training, Validierung und Test, einschließlich standardisierter Datenqualitätsprofile	Domänenübliche Datenformate (CSV, Parquet u.a.); Datensteckbriefe nach EDPS-Schema
Black-Box-Prüfservices	Prüfdienste, die geschützte Testdaten oder geschützte Werkzeuglogik enthalten und im Compute-to-Data-Modus angesprochen werden (Prüfplattformen)	API-Endpunkte mit OpenAPI-Spezifikation; Aufruf über Datenraum-Konnektoren
Eignungsattestierungen	Bewertungen qualifizierter Stellen über die Eignung eines Werkzeugs für bestimmte Normabschnitte	Verifiable Credentials in JSON-LD
Akkreditierungsnachweise	Bescheinigungen über die Akkreditierung von Konformitätsbewertungsstellen	Verifiable Credentials in JWT & JSON-LD
Konformitätsbescheinigungen	Ergebnisse durchgeführter Konformitätsbewertungen; in DPP bzw. „KI-Pass“ entlang des Produktlebenszyklus mitgeführt	Digital Certificate of Conformity (D-CoC), signierte JSON-LD-Dokumente, DPP-Datenmodelle
Asset Administration Shell	Beschreibung physisch eingebetteter KI-Komponenten mit Verweis auf Norm- und Klassifikationsvokabulare über semanticIds	AAS-Submodelle in JSON, XML oder AASX; semanticIds nach IEC CDD und ECLASS

Tabelle 3: Ressourcen und Formate in einem Datenraum für KI-Qualität (Formatangaben sind als Beispiele und nicht als abschließende Festlegung zu verstehen)

3.4 Skalierbare Generierung von SMART Content und Operationalisierungsinhalten

Die zuvor beschriebenen Szenarien setzen qualitativ hochwertigen SMART Content und ergänzende Operationalisierungsinhalte voraus, deren Bandbreite erheblich ist. Eine rein manuelle Erstellung des gesamten Bestands wäre nicht skalierbar. Um die Bandbreite handhabbar zu machen, könnte eine KI-gestützte Generierung mit menschlicher Validierung mit einer arbeitsteiligen Erstellung durch mehrere Akteure mit unterschiedlichen Kompetenzschwerpunkten kombiniert werden.

KI-gestützte Generierung mit menschlicher Validierung

Der vorgeschlagene Ansatz folgt dem Prinzip Human-in-the-Loop. KI übernimmt die aufwändige Erstextraktion, Normungsexperten sichern die Qualität (siehe Abbildung 5). Die Grundlage bildet ein Editor wie das Online Standards Development Tool (OSD)⁴⁹, ergänzt um KI-Unterstützung und Validierung. Der von der DKE entwickelte TermAnalyzer zur Qualitätsprüfung und semantischen Harmonisierung von Normbegriffen kann dabei ebenfalls eingesetzt werden.

KI-Modelle eignen sich dabei vor allem für Aufgaben mit hoher Strukturiertheit und großem Volumen, etwa die Extraktion von Entitäten und Termen aus Normtext, die Identifikation expliziter Querverweise, die Ableitung von Constraints aus Anforderungssätzen, das Aufzeigen inhaltlicher Änderungen zwischen Normversionen und die Konsistenzprüfung über mehrere Normen hinweg. Menschliches Urteil bleibt dort erforderlich, wo fachliche Bewertung im Vordergrund steht, insbesondere beim Auflösen normativer Ambiguität, beim Ergänzen impliziten Expertenwissens, bei Entscheidungen in Grenzfällen, bei der Bewertung der Auswirkung von Normänderungen sowie bei der Priorisierung im Fall konkurrierender Normen.

Dasselbe Human-in-the-Loop-Prinzip lässt sich womöglich auf die Generierung von Ontologien und Knowledge Graphs anwenden. LLMs können aus Normtext Konzeptionshierarchien, Relationen und formale Constraints extrahieren und als strukturierten Graphentwurf formulieren, typischerweise

in hybriden Pipelines, die LLMs mit domänenspezifischen Embeddings, Retrieval-Augmented Generation und Fine-Tuning kombinieren.⁵⁰

Sinnvoll ist eine stufenweise Vorgehensweise mit steigendem Formalisierungsgrad. Auf der ersten Stufe entstehen Begriffe, Definitionen und Synonyme mit geringem Validierungsaufwand, vergleichbar mit bekannter Terminologiearbeit. Die zweite Stufe bringt Relationen zwischen Anforderungen wie „setzt_voraus“, „konkretisiert“ oder „verweist_auf“ als RDF/OWL, deren Relationstypen fachlich bewertet werden müssen. Auf der dritten Stufe werden formale Prüfregelelemente generiert, deren Validierung Prüffachwissen erfordert, da stille Fehler große Auswirkungen auf alle nachgelagerten Szenarien haben können. Auf jeder Stufe liefert die KI einen Entwurf einschließlich Konfidenz-Annotation, auf deren Grundlage Normungsexpertinnen und -experten Entscheidungen treffen.

Wer erstellt welche Inhalte?

Der vorherige Abschnitt skizziert umfangreichen neuen SMART Content. Hinsichtlich Kapazität und Kompetenzschwerpunkten dürfte es für Normungsgremien schwierig sein, diese Inhalte eigenständig beizutragen. Bei einigen dieser Inhalte handelt es sich ohnehin nicht um Norminhalte im klassischen Sinne, sondern um Operationalisierungen und Interpretationen, für die andere Akteure bessere Voraussetzungen mitbringen. Wie eine arbeitsteilige Erstellung dieser Inhalte konkret organisiert werden könnte, ist eine offene Frage, deren Beantwortung umfangreiche Verständigungen voraussetzt.

Ein mögliches Schichtenmodell unterscheidet vier Ebenen, mit dem „normativen Kern“ auf der ersten Ebene. Normungsgremien stellen zunächst granulare Norminhalte, Begriffsdefinitionen und Querverweise als verbindliche Grundlage bereit. Auf einer zweiten Ebene könnten Forschungs- und Metrologieeinrichtungen Beiträge zur Operationalisierung liefern, etwa wissenschaftlich fundierte Prüfkataloge, SHACL-Shapes, oder Referenzdatensätze.

49 <https://www.iec.ch/online-standards-development>

50 Vgl. Babaei Giglou, H., D'Souza, J., Mihindukulasooriya, N., & Auer, S. (2025), "LLMs4OL 2025 Overview: The 2nd Large Language Models for Ontology Learning Challenge", Open Conference Proceedings 6, <https://doi.org/10.52825/ocp.v6i.2913>. Pilotprojekte zur Maschineninterpretierbarkeit von Industriestandards wie Gjerver et al. (2025), "Machine-interpretable Engineering Design Standards for Valve Specification", IEEE Access, <https://doi.org/10.1109/ACCESS.2025.3643528>, demonstrieren die konkrete Anwendung auf technische Standards.

Ihr Vertrauensbeitrag würde sich dabei auf wissenschaftliche Methodik, Peer-Review und institutionelle Verantwortung stützen. Auf einer dritten Ebene erfolgt die sektorspezifische Konkretisierung durch Branchenverbände, Industriekonsortien und Forschungsprojekte, die Interpretationsleitfäden oder Norm-Mappings beisteuern könnten. Eine vierte Werkzeug- und Anwendungsebene fasst Beiträge von (KI-Prüf-) Werkzeugherstellern, Open-Source-Communities und Anwendern zusammen, die Werkzeugadapter, Referenzimplementierungen und Bibliotheken beitragen.

Prüforganisationen wären in einem solchen Modell weniger primäre Erzeuger von Operationalisierungsinhalten, sondern Anwender, die die Inhalte aus den oberen Ebenen in akkreditierten Prüfdurchführungen nutzen und eigene Beiträge etwa als Konformitätsbescheinigungen (D-CoC) und akkreditierungsbezogene Verifiable Credentials zurückgeben würden.

Die Infrastruktur für ein solches arbeitsteiliges Modell wäre der in Szenario 3 beschriebene Datenraum. Jeder Inhalt würde die Identität seines Herausgebers tragen, eine maschinenlesbar attestierte Vertrauensklasse als Verifiable Credential und eine explizite Referenz auf die Norm-Version, auf die er sich bezieht. Bei einer neuen Normversion könnten abhängige Inhalte automatisch als möglicherweise veraltet markiert werden, vergleichbar mit Abhängigkeiten in einem Paketmanager.

Ein wesentliches Risiko des Modells wäre eine Fragmentierung, wenn mehrere Akteure konkurrierende Operationalisierungen für dieselbe Norm veröffentlichen. Eine kuratierende Rolle könnte etwa von Forschungs- und Metrologieeinrichtungen mit etabliertem Vertrauensstatus übernommen werden, die qualitätsgesicherte „Schicht-2-Inhalte“ transparent empfehlen (ohne sie „normativ“ zu machen). Die praktische Umsetzung eines solchen Modells würde umfangreiche Verständigungen über Rollen, Schnittstellen und Vertrauensbegründungen voraussetzen und stellt damit ein langfristiges Entwicklungsziel dar.

4

ANWENDUNGS- BEISPIEL

4 ANWENDUNGSBEISPIEL

Die zuvor beschriebenen Szenarien sind am konkretesten anhand eines Unternehmens nachvollziehbar, das sie schrittweise umsetzt. Dieses Kapitel verfolgt den Weg eines KI-Unternehmens von den ersten Schritten mit SMART Standards bis zu einer in eine digitale Qualitätsinfrastruktur eingebettete Prüfumgebung und macht damit greifbar, was die konzeptionellen Szenarien in der Praxis bedeuten können.⁵¹



DAS UNTERNEHMEN

Ein europäisches Unternehmen entwickelt und betreibt ein KI-basiertes Credit-Scoring-System, das Banken und Konsumentenkreditgebern als SaaS angeboten wird. Das System ordnet Kreditanträge anhand soziodemografischer und finanzieller Merkmale Bonitätsklassen zu und unterstützt damit Entscheidungen über Kreditvergabe und -konditionen. Damit fällt es nach Anhang III der KI-VO unter die Hochrisiko-Kategorie. Das Unternehmen muss die Anforderungen aus Kapitel III Abschnitt 2 der KI-VO erfüllen, insbesondere ein angemessenes Risikomanagement (Art. 9) und ein Mindestmaß an Genauigkeit, Robustheit und Cybersicherheit (Art. 15).

Im Unternehmen wird eine agile, SCRUM-artige Softwareentwicklung praktiziert. Anforderungen werden in Gitlab gepflegt und mit Epics und User Stories operationalisiert. Die Entwicklung des KI-Systems besteht zu großen Teilen aus Softwareentwicklung und Data-Science-Prozessen. Die MLOps-Infrastruktur ist als GitLab-CI/CD-Pipeline umgesetzt, die für ML-spezifische Aufgaben ein MLOps-Framework (etwa MLflow) als Experiment-Tracking- und Model-Registry-Backend einbindet. Prüfsingenieurinnen und

Prüfsingenieure hinterlegen Prüfcode als CI-Jobs. Trainingsläufe, Modellversionen und Prüfergebnisse werden im MLOps-Framework geloggt und versioniert. Neue Modellversionen werden nach automatisiertem Testen in Release-Zyklen ausgeliefert. Die Verantwortung für die Analyse und Umsetzung von Normanforderungen trägt ein Product Owner gemeinsam mit den Prüfsingenieurinnen und Prüfsingenieuren.

Eine zentrale Qualitätsdimension für das System ist Datendrift, also die Veränderung der statistischen Eigenschaften produktiver Eingangsdaten gegenüber dem Trainings- bzw. Referenzdatensatz. Bei einem Credit-Scoring-System entsteht Datendrift etwa durch konjunkturelle Verschiebungen, neue Antragstellergruppen oder veränderte Konsumgewohnheiten. Wird sie nicht überwacht, drohen schleichende Verzerrungen der Bonitätsentscheidungen, die sowohl die Robustheit (KI-VO Art. 15) als auch das Risikomanagement (Art. 9) beeinträchtigen.

⁵¹ Bei diesem Beispiel handelt es sich um eine adaptierte Version des Beispiels aus ETSI TS 104 008 V1.1, A.3

DIE (FIKTIVEN) NORMEN

Für diese Qualitätsdimension greifen drei fiktive Normen ineinander (siehe auch Abbildung 6):

- **EN 123456:2026** „Datenqualität von KI-Systemen“, harmonisierte Norm zur KI-VO mit horizontalem, abstraktem Anforderungscharakter. Sie fordert: „The data-drift should be constantly monitored and should not exceed a moderate level.“
- **ISO/IEC 98761:2026** „Datenqualitätsmetriken“, internationale Norm, die den Population Stability Index

(PSI) als Indikator für Datendrift definiert, dessen Berechnungsformel angibt und auf eine registrierte Referenzimplementierung verweist.

- **ISO/IEC 98768:2026** „Qualität von KI-Anwendungen für Credit-Scoring“, internationale Norm mit vertikalem, konkretem Charakter. Sie konkretisiert die Drift-Schwellenwerte für den Anwendungskontext Credit-Scoring: „stable“ bei $PSI < 0.1$, „moderate“ bei $PSI < 0.2$, „significant“ ab $PSI \geq 0.2$.

4.1 Szenario 1: Einsatz von maschinenlesbaren Norminhalten per APIs

Normen identifizieren und beziehen

In der Konzeptionsphase muss das Unternehmen klären, welche Normen und Norminhalte für sein KI-System relevant sind. Statt PDFs manuell über Webshops zu suchen und durchzuarbeiten, nutzt das Unternehmen die SMART-Standards-Schnittstelle eines Normenanbieters (z. B. REST API, vgl. VDE CaaS).

Über eine intelligente Suche mit semantischen Funktionen lässt sich die Recherche an den Systemkontext anbinden: Einsatzzweck (Bonitätsbewertung), Risikoklasse (Hochrisiko nach KI-VO Anhang III), und Branche (Finanzdienstleistung). Ausgangspunkt ist die Referenzkette aus der KI-VO: Art. 9 und Art. 15 verweisen auf harmonisierte Normen, die wiederum auf weitere Normen verweisen. Der Suchdienst liefert die einschlägige harmonisierte Norm EN 123456:2026 als Treffer und folgt automatisch deren Querverweisen zu ISO/IEC 98761:2026. Eine semantische Suche im Katalog vertikaler Normen identifiziert ISO/IEC 98768:2026 als einschlägige Konkretisierung für Credit-Scoring.

Norminhalte ins Anforderungsmanagement integrieren

Die identifizierten Norminhalte werden als granulare, maschinenlesbare Einheiten per ReqIF-Import⁵² in Gitlab übernommen, statt als PDF abgelegt und manuell paraphrasiert zu werden. Für die hier betrachtete Anforderung entstehen drei verknüpfte Inhalte:

- Aus EN 123456:2026 die Anforderung Req-5-2 („The data-drift should be constantly monitored and should not exceed a moderate level.“) als GitLab Issue im Epic „Datenqualität nach EN 123456“, versehen mit dem Label `compliance::EN-123456` und einem Verweis auf die durch sie adressierten KI-VO-Artikel (Art. 9 und Art. 15).
- Aus ISO/IEC 98761:2026 die Definition des PSI: die Berechnungsformel (in MathML, einem standardisierten Auszeichnungsformat für mathematische Formeln) und ein Verweis auf eine registrierte Referenzimplementierung mit persistentem Identifier und semantischer Versionierung. Beide werden im Issue verlinkt.
- Aus ISO/IEC 98768:2026 das kontextspezifische Schwellenwertprofil als JSON-Datei, die im Repository abgelegt wird. Sie verknüpft die numerischen PSI-Schwellen mit

⁵² Hinweis: Gitlab bietet bislang keinen ReqIF-Import an, hat dies jedoch als Work Item dokumentiert: https://gitlab.com/groups/gitlab-org/-/work_items/21272

Fiktive Normen und dazugehöriger SMART Content

AI ACT



Art. 9, Art. 15

STANDARDS

EN 123456:2026 »Datenqualität von KI-Systemen«

Harmonisierte Norm, durch KI-VO referenziert:
horizontal, abstrakt

The data-drift should be constantly monitored and should not exceed a moderate level.



ISO/IEC 98761:2026 »Datenqualitätsmetriken«

The Population Stability Index (PSI)

$$PSI = \sum ((Actual\% - Expected\%) \times \ln(\frac{Actual\%}{Expected\%}))$$



ISO/IEC 98768:2026 »Qualität von KI-Anwendungen
für Credit-Scoring«

Internationale Norm, vertikal, konkret

The following data-drift PSI-thresholds are defined for

[credit scoring test]

stable: $PSI < 0.1$

moderate: $PSI < 0.2$

significant: $\text{PSI} \geq 0.2$



SMART CONTENT

[illegible]

Ontology (OWL)

Compliance Shape (SHACL)

```
3 @prefix en13456: <https://standards.example/en13456/2026#> .
4
5 iso98761 : a ql:InternationalStandard ;
6   rdfs:label "ISO/IEC 98761:2026 - Datenqualitätsmetriken für KI-Systeme"@de ;
7   ql:scope ql:DomainAgnostic .
8
9 iso98761:PSI a ql:Metric ;
10  rdfs:label "Population Stability Index"@en ;
```

Ontology

```
<math xmlns="http://www.w3.o
  <mrow>
  </mrow>
</math>
```

Formula (MathML)

Reference to implementation

```

8  A1=describes CreditScoring""@eq;
9  q1.scope      q1.Ctx-CreditScoring;
10 q1.concretises en123456.., iso89761.;
11
12 q1.Ctx-CreditScoring a1.ApplicationContext;
13   rdfs:label "AnwendungsContext: KI-basiertes Credit-Scoring"@eq..
14
15 iso89768:PSI1.Thresholds-CreditScoring a q1.ThresholdProfile;
16   rdfs:label "PSI1-Schwellenwerte für Credit-Scoring"@eq;
17   q1.appliesMetric      PSI1;
18   q1.appliesInContext   q1.Ctx-CreditScoring;
19   q1.concretises        en123456.., PSI1.en1.;

```

Ontology + thresholds

```
7  "thresholds": {
8    "stable": {
9      "level": "https://standards.example/en123456/2026#Stable",
10     "lower": 0.0,
11     "upper": 0.1
12   },
13   "moderate": {
14     "level": "https://standards.example/en123456/2026#Moderate",
15     "lower": 0.1,
16     "upper": 0.2
17   }
18 }
```

Thresholds (JSON)

Abbildung 5

den abstrakten Drift-Niveaus aus EN 123456 und dient später unmittelbar als Konfiguration für den Test.

Jedes Issue trägt in einem strukturierten Metadatenfeld den persistenten Identifier des SMART Content, etwa einen URN, der Norm, Abschnitt und Version eindeutig adressiert (im Beispiel: `urn:smartstd:en123456:2026:req-5-2`). Damit ist die Quelle dokumentiert, und über die GitLab-API maschinell auswertbar.

Aktualisierungen werden über eine GitLab Scheduled Pipeline erkannt, die täglich für jeden importierten Identifier die SMART-Standards-API abfragt. Liefert die API eine neuere Version eines Norminhalts (etwa weil EN 123456 als Version 1.1 veröffentlicht wurde), öffnet die Pipeline einen Merge Request, der die textuellen und strukturellen Änderungen markiert. Die zuständigen Prüfsingenieurinnen und Prüfsingenieure werden über GitLab automatisch zugeordnet und reviewen das Delta, bevor die aktualisierte Anforderung in die laufende Entwicklung übernommen wird.

In Jupyter-Notebooks im GitLab-Repository (etwa über die GitLab-Web-IDE) und in IDEs wie VS Code wird der SMART Content inline bereitgestellt. Eine SMART-Standards-Extension blendet beim Schreiben von Validierungscode den zugehörigen Normabschnitt, die Formel und die Schwellenwerte direkt im Editor ein.

Operationalisieren in der MLOps-Pipeline

Aus den in GitLab verlinkten SMART Content leiten die Prüfsingenieurinnen und Prüfsingenieure die Prüfümplementierung ab. Die eigentliche Drift-Prüfung läuft in zwei Schritten ab:

- Berechnung und Klassifikation: Der PSI-Wert wird für einen produktiven Datenausschnitt gegen den Referenzdatensatz berechnet und anschließend über das Schwellenwertprofil aus ISO/IEC 98768 einem Drift-Niveau zugeordnet („stable“, „moderate“ oder „significant“).
- Konformitätsprüfung: Das ermittelte Niveau wird gegen die Anforderung aus EN 123456 abgeglichen, also geprüft, ob es nicht über „moderate“ hinausgeht. EN 123456 stellt diese Prüfung als formale, maschinell ausführbare Vorschrift bereit (im SMART Content der Norm als SHACL-Constraint⁵³ hinterlegt).

Die Implementierung übernimmt eine kurze Python-Funktion: Sie ruft die referenzierte PSI-Bibliothek aus ISO/IEC 98761 über deren persistenten Identifier auf, lädt die Schwellenwerte direkt aus dem JSON-Profil und gibt das Drift-Niveau zurück. Die formale Konformitätsprüfung selbst wird mit einer Standardbibliothek (etwa pyshacl) gegen die in der Norm hinterlegte Vorschrift validiert.

Der Test ist als Job in der GitLab-CI/CD-Pipeline angelegt. Er läuft bei jedem Modellrelease und wird zusätzlich über eine Scheduled Pipeline kontinuierlich gegen Produktionsdaten ausgeführt. Der berechnete PSI-Wert, das ermittelte Drift-Niveau und das Ergebnis der Konformitätsprüfung werden im Job über das Tracking-API des MLOps-Frameworks der jeweiligen Modellversion in der Model Registry zugeordnet und damit versioniert nachvollziehbar. Bei Überschreitung des „moderate“-Niveaus erzeugt die Pipeline einen Alarm und triggert einen Retraining-Workflow.

Da Test und Norminhalt über die in GitLab gepflegten Identifier eindeutig miteinander verknüpft sind und Prüfergebnisse über das MLOps-Framework an die Modellversion gebunden sind, ist jederzeit nachvollziehbar, welcher Test welche Anforderung für welche Modellversion abdeckt.

Konformitätsdokumentation generieren

Die Konformitätsdokumentation wird automatisiert aus den vorhandenen, versionierten Artefakten abgeleitet, ein nachgelagerter manueller Schritt entfällt. Belege stammen aus drei Quellen:

- GitLab liefert den normativen Bezug. Für jedes Compliance-Label („compliance::EN-123456“) lassen sich die zugehörigen Issues mit ihren persistenten Identifiern auflisten, einschließlich der Version der referenzierten Norm.
- Das MLOps-Framework liefert die Prüfergebnisse. Jeder Run ist über die Modellversion mit der jeweiligen Norm-Anforderung verknüpft und enthält PSI-Wert, Drift-Niveau und Konformitätsergebnis.
- Git liefert den Code-Stand. Der Commit-Hash der zugrundeliegenden Pipeline-Definition und Prüfümplementierung ist Teil jedes Runs im MLOps-Framework.

53 SHACL ist eine durch W3C standardisierte Sprache zur Beschreibung solcher Konformitätsregeln: <https://www.w3.org/TR/shacl/>

Ein dedizierter CI-Job (im Beispiel „compliance-report“) führt diese Quellen zusammen und erzeugt einen strukturierten Konformitätsbericht. Der Bericht enthält für jede Anforderung die Kette aus Normabschnitt, Anforderungs-Identifizierung, eingesetzter Prüfimplementierung mit Versionsangabe, Datenausschnitt, Prüfergebnis und Konformitätsprüfung. Er wird sowohl als JSON-Datei für die maschinelle Weiterverarbeitung als auch als gerenderte HTML- bzw. PDF-Fassung für menschliche Adressaten erzeugt und als GitLab-Release-Artefakt sowie im MLOps-Framework archiviert.

Dadurch, dass alle Bestandteile des Berichts versioniert und persistent referenziert sind, wird die Reproduzierbarkeit sichergestellt. Die Norm-Inhalte sind über ihre URNs adressiert (etwa `urn:smartstd:en123456:2026:req-5-2`), der Prüfcode über den Git-Commit-Hash, die Modellversion und die Prüfergebnisse über die Run-ID des MLOps-Frameworks. Für jede vergangene Modellversion lässt sich der zugehörige Bericht damit auf Knopfdruck neu erzeugen, etwa wenn ein Auditor oder eine Auditorin zu einem bestimmten Release nachfragt oder rückwirkend nachvollzogen werden muss, gegen welche Normfassung geprüft wurde.

4.2 Szenario 2: Generierung von KI-Prüfungen



DAS UNTERNEHMEN (...FORTSETZUNG)

Das Unternehmen aus Szenario 1 nutzt SMART Standards über spezialisierte Werkzeuge. Im nächsten Schritt übernimmt ein lokaler Chatbot die Steuerung. Eine Prüferingenieurin oder ein Prüferingenieur interagiert mit dem Chatbot in der eigenen IDE (etwa VS Code mit Chatbot-Extension). Der Chatbot greift

über MCP auf die SMART-Standards-API zu und stützt seine Antworten per RAG auf die maschinenlesbaren Norminhalte. Lokale Kontextquellen werden ebenfalls über MCP angebunden, insbesondere das GitLab-Projekt mit seinen Issues, Wikis und dem Repository-Inhalt.

Anforderungen kontextspezifisch ermitteln

Die Prüferingenieurin bzw. der Prüferingenieur kann den Systemkontext auf zwei Wegen einbringen: Eine kurze Beschreibung in natürlicher Sprache reicht für eine schnelle Recherche aus, etwa „KI-basiertes Credit-Scoring-System für Konsumentenkredite, eingesetzt in der EU. Relevant sind die Anforderungen der KI-VO an Datenqualität und Robustheit.“

Für eine fundierte Anforderungsanalyse, etwa im Rahmen einer formalen Konformitätsbewertung, ist es zweckmäßiger, dem Chatbot eine Auswahl einschlägiger Dokumentationsartefakte freizugeben, etwa Systembeschreibung, Architekturdokumentation, Modellkarte und Risikoklassifizierung. Aus den Dokumenten extrahiert der Chatbot die

für die Norm-Identifikation relevanten Systemeigenschaften (Einsatzzweck, Risikoklasse, Branche, Verarbeitungstyp) und nutzt sie als strukturierte Grundlage für die Recherche.

Beide Vorgehen lassen sich kombinieren. Zusätzlich liest der Chatbot den lokalen Kontext aus GitLab (bestehende Issues, Architekturdokumentation in Wikis, Modellkarten und Code im Repository) und gleicht ihn semantisch gegen die im SMART-Standards-Katalog verfügbaren Norminhalte ab.

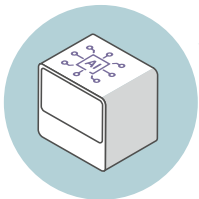
Auf Begriffsebene werden lokale Terme gegen die Norm-Ontologie abgeglichen. „Modellabweichung“ oder „Verteilungsschiebung“ in der internen Dokumentation treffen auf den Normterm „data-drift“ aus EN 123456. Auf struktureller Ebene werden die Systemeigenschaften (Hochrisiko-Klasse,

Unterstützung durch einen Chatbot bei einer normbasierten Prüfung (vereinfachtes Beispiel)



PRÜFINGENIEURIN

Wir entwickeln ein KI-basiertes Credit-Scoring-System für Konsumentenkredite (EU). Welche Anforderungen der KI-VO an Datenqualität und Robustheit sind relevant, und wie testen wir sie?



CHATBOT

```
// Recherchiere harmonisierte Normen über die SMART-Standards-API, lese
lokalen Kontext aus GitLab (Modellkarte, Architektur, Issues), gleiche Be-
griffe gegen die Normontologie ab.
```

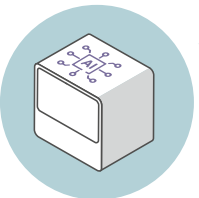
Drei Normen sind einschlägig:

- EN 123456:2026 (harmonisiert): kontinuierliches Drift-Monitoring, Niveau $\leq$ „moderate“
- ISO/IEC 98761:2026: PSI als Drift-Indikator (Formel + Referenzimplementierung)
- ISO/IEC 98768:2026: PSI-Schwellenwerte für Credit-Scoring



PRÜFINGENIEURIN

Bitte erstelle einen Prüfplan und Testcode für die Drift-Anforderung.



CHATBOT

```
// Leite Prüfplan aus den Norminhalten ab, generiere Testcode mit PSI-Bere-
rechnung, Schwellenwert-Klassifikation und SHACL-Konformitätsprüfung.
```

Prüfplan:

- Metrik: PSI auf produktivem Datenausschnitt vs. Referenzdatensatz
- Klassifikation: Schwellenwertprofil aus ISO/IEC 98768
- Konformität: SHACL-Constraint aus EN 123456
- Trigger: bei jedem Release und täglich, Alarm bei „significant“

```
psi    = psi_lib.compute(window, reference)    # ISO/IEC 98761
level  = classify(psi, thresholds_cs)          # ISO/IEC 98768
validate_shacl(level, en123456_constraint)    # EN 123456
mlflow_client.log_metric(„psi“, psi)
```

Soll ich den Test als CI-Job in der GitLab-Pipeline anlegen und das Logging im MLOps Framework konfigurieren?

Abbildung 6

Anwendungskontext Credit-Scoring) gegen die Geltungsbereiche der Normen gematcht. Ergebnis ist eine kommentierte Liste relevanter Anforderungen, wie etwa die Drift-Anforderung aus EN 123456:2026, die zugehörige Metrik aus ISO/IEC 98761:2026 und die kontextspezifischen Schwellenwerte aus ISO/IEC 98768:2026, jeweils mit Verweis auf den Normabschnitt und Begründung der Einschlägigkeit.

Prüfplan vorschlagen

Aus der Anforderungsliste generiert der Chatbot einen strukturierten Prüfplan:

- Was geprüft wird: PSI als Indikator für Datendrift im produktiven Datenausschnitt gegen den Referenzdatensatz.
- Wie geprüft wird: Berechnung gemäß der MathML-Formel aus ISO/IEC 98761:2026, Abgleich gegen das JSON-Schwellenwertprofil aus ISO/IEC 98768:2026.
- Wann geprüft wird: bei jedem Modellrelease und kontinuierlich im Monitoring (gemäß „constantly monitored“ aus EN 123456:2026).
- Welche Aktionen ausgelöst werden: Alarm bei „significant“, Retraining-Trigger bei wiederholtem „moderate“.

Maschinell prüfbare Anforderungen werden dabei von solchen unterschieden, die manuelles Urteil erfordern. Die Prüferin bzw. der Prüfer überprüft den Prüfplan, ergänzt fehlende Aspekte und gibt ihn frei.

Prüfcode generieren

Für die maschinell prüfbaren Aktivitäten generiert der Chatbot ausführbaren Testcode, der im Git-Repository abgelegt und als Job in der GitLab-CI/CD-Pipeline ausgeführt wird. Die Vertrauenswürdigkeit der Generierung beruht auf einer klaren Trennung von SMART Content (Grenzwerte, Berechnungsvorschriften) und Sprachmodell.

- Die PSI-Berechnung wird auf die referenzierte Implementierung aus ISO/IEC 98761:2026 verlinkt; die Bibliothek wird über ihren persistenten Identifier eingebunden, statt vom Modell rekonstruiert zu werden.
- Die Schwellenwerte werden direkt aus dem JSON-Profil aus ISO/IEC 98768:2026 in den Code übernommen.

- Die formale Prüfvorschrift aus EN 123456:2026 (siehe Operationalisierung in Szenario 1) dient als Spezifikation für die Konformitätsprüfung.

Das Sprachmodell liefert die Prüflogik und die Einbettung in die lokale Toolchain, einschließlich des Loggings im MLOps-Framework für die Prüfergebnisse. Die normierten Werte sind fest verankert und auf ihren Normabschnitt zurückführbar. Der generierte Code wird automatisch syntaktisch geprüft und gegen die Prüfvorschrift der Norm validiert, um die Abbildung der Anforderung sicherzustellen. Der Chatbot legt das Ergebnis als Merge Request im GitLab-Repository an, den die Prüferin bzw. der Prüfer reviewt und merged.

Berichte auf Anfrage erzeugen

Der gleiche Chatbot ermöglicht ad-hoc-Reporting aus natürlicher Sprache. Die Prüferin bzw. der Prüfer fragt etwa „Erstelle eine Übersicht der Datenqualitäts-Prüfergebnisse nach EN 123456 für die letzten drei Releases“, und der Chatbot ruft über MCP die entsprechenden Runs aus dem MLOps-Framework und GitLab-Issues ab und erzeugt den Bericht im selben strukturierten Format wie der CI-Job aus Szenario 1.

Zu den standardisierten Berichten zu festen Triggerpunkten (Szenario 1) kommen in Szenario 2 ad-hoc-Auswertungen für spezifische Audits, interne Reviews oder Rückfragen von Konformitätsbewertungsstellen hinzu, ohne dass dafür ein separater Implementierungsaufwand entsteht. Die Evidenzkette bleibt dabei deterministisch, da das Sprachmodell nur die Anfrage interpretiert und das Layout gestaltet, während die belegenden Werte unverändert aus den versionierten Quellsystemen kommen.

Qualitätssicherung und Verantwortung

Die Prüferin bzw. der Prüfer bleibt für die Qualität der Prüfung verantwortlich. Die Rolle verlagert sich vom Erstellen von Prüfcode zum gezielten Review der generierten Artefakte. Für sicherheitskritische Anforderungen ist ein obligatorischer manueller Review-Schritt vorgesehen. Das System protokolliert lückenlos, welche Chatbot-Ausgaben welche normativen Grundlagen hatten, als Voraussetzung für die Nachvollziehbarkeit gemäß KI-VO Art. 12.

4.3 Szenario 3: Einbindung in ein Ökosystem für KI-Qualität



DAS UNTERNEHMEN (...FORTSETZUNG)

Das Unternehmen aus den Szenarien 1 und 2 hat sein Portfolio erweitert. Neben dem bestehenden Credit-Scoring-System entwickelt es inzwischen weitere KI-basierte Dienste, etwa für die Risikobewertung von Geschäftskunden und für die Betrugserkennung im Zahlungsverkehr. Mit dem zusätzlichen Portfolio wachsen Anzahl und Vielfalt der Prüfungen, und es zeigt sich, dass nicht alle benötigten Ressourcen sinnvoll im Unternehmen selbst bereitgestellt werden können. Für bestimmte Prüfungen fehlen entweder die geeigneten Referenzdatensätze, oder die Werkzeuge sind so spezialisiert, dass sich die Eigenentwicklung nicht rechnet, oder die geforderte Unabhängigkeit der Prüfung verlangt nach einer akkreditierten externen Stelle.

Das Unternehmen schließt sich daher an eine föderierte Plattform der Qualitätsinfrastruktur an. Der Anschluss erfolgt schrittweise: Zunächst wird ein Datenraum-Konnektor (Eclipse EDC) als Container-Service in der bestehenden Infrastruktur bereitgestellt und (per REST API, bzw. über MCP) an die GitLab-CI/CD-Pipeline und den Chatbot aus Szenario 2 angebunden. Parallel registriert sich das Unternehmen als Teilnehmer und erhält die zugehörige verifizierbare Identität. Damit kann es im Datenraum Werkzeuge und Datensätze beziehen, Prüfservices beauftragen und eigene Konformitätsnachweise sichtbar machen. Die folgenden Abschnitte beschreiben typische Anwendungen.

Prüfwerkzeuge aus dem Datenraum beziehen

In Szenario 1 wurde angenommen, dass die in ISO/IEC 98761 referenzierte PSI-Bibliothek einfach verfügbar ist. Durch Szenario 3 wird die gesicherte Bereitstellung einer solchen Bibliothek konkretisiert. Mit dem Anschluss an den Datenraum löst das Unternehmen den im SMART Content hinterlegten Identifier im föderierten Katalog auf. Die Suche liefert zwei Implementierungen, die jeweils eine Eignungstestierung einer qualifizierten Stelle für ISO/IEC 98761:2026 tragen. Die Attestierungen liegen als signierte Verifiable Credentials am Werkzeug-Eintrag im Katalog vor und benennen die geprüfte Werkzeugversion, die abgedeckten Normabschnitte und die ausstellende Stelle.

Das Unternehmen wählt anhand der Lizenzbedingungen und der Versionsbindung an die in der GitLab-Pipeline genutzten Modellversionen aus. Über den Konnektor wird das Werkzeug als Container-Image bezogen. Das zugehörige Verifiable Credential wird mit übertragen und in der CI/CD-Pipeline lokal verifiziert, bevor das Image ausgeführt wird. Die GitLab-Pipeline aus Szenario 1 wird so angepasst, dass die Bibliothek nicht mehr aus einem öffentlichen Paket-Re-

pository stammt, sondern aus dem Datenraum, mit verifizierbarem Bezug zur attestierten Eignung.

Modell gegen einen geschützten Referenzdatensatz prüfen

Eine zusätzliche Anforderung an das Credit-Scoring-System ergibt sich aus der branchenweiten Konkretisierung der KI-VO im Finanzsektor: Die Drift-Prüfung soll nicht nur gegen den unternehmenseigenen Referenzdatensatz erfolgen, sondern zusätzlich gegen einen demografisch ausgewogenen Referenzbestand, den eine qualifizierte Stelle für die Branche pflegt. Dieser Referenzbestand wird bewusst nicht öffentlich verfügbar gemacht, damit er nicht in spätere Trainingsdaten einfließt und seine Aussagekraft als unabhängiges Prüfmittel verliert.

Über den Datenraum lässt sich diese Prüfung im Compute-to-Data-Modus realisieren. Das Unternehmen stellt die Inferenz-Funktion seines Modells als Service über den Konnektor zur Verfügung; die qualifizierte Stelle ruft diesen Service mit Datenpunkten aus dem geschützten Referenzbestand auf, berechnet die PSI-Werte und Drift-Niveaus mit der attestierten PSI-Bibliothek aus dem vorigen Abschnitt

und sendet ausschließlich das signierte Ergebnis (PSI-Wert, Drift-Niveau, Konformitätsprüfung) zurück. Modellgewichte verlassen das Unternehmen nicht; der Referenzbestand verlässt die qualifizierte Stelle nicht.

Die Logik der Prüfung selbst bleibt identisch zu Szenario 1, lediglich Ausführungsort und Schutz der beteiligten Ressourcen sind anders aufgeteilt. Das Unternehmen erhält damit einen zusätzlichen, unabhängig erzeugten Beleg für die Drift-Prüfung gegen den unabhängigen Referenzbestand. Das Ergebnis wird im MLOps-Framework an die jeweilige Modellversion gebunden und fließt in den Konformitätsbericht aus Szenario 1 ein. Für die qualifizierte Stelle bleibt der Datensatz vollständig in ihrer Kontrolle, ohne dass eine bilateral verhandelte Datenübergabevereinbarung mit jedem Kunden erforderlich wäre.

Konformitätsbescheinigung als D-CoC im „KI-Pass“

Nachdem das Credit-Scoring-System vor Inverkehrbringen eine externe Konformitätsbewertung durch eine akkreditierte Stelle durchlaufen hat, stellt diese das Ergebnis als Digital Certificate of Conformity (D-CoC) aus. Anders als die heute übliche PDF-Bescheinigung ist das D-CoC ein maschinenlesbares Dokument, das als signiertes Verifiable Credential geführt wird und die geprüften Norm-Versionen, die geprüfte Modellversion, die eingesetzten Prüfverfahren, die ausstellende akkreditierte Stelle und deren Akkreditierungsnachweis maschinenlesbar enthält.

Das D-CoC wird in den „KI-Pass“ des Credit-Scoring-Systems aufgenommen, der das System über seinen Lebenszyklus begleitet und sowohl Modell- als auch Konformitätsangaben bündelt. Die abnehmenden Kunden können vor Vertragsabschluss automatisiert prüfen, ob ein gültiges D-CoC vorliegt, ob es von einer akkreditierten Stelle stammt und ob es sich auf die tatsächlich eingesetzte Modellversion bezieht. Die zugrundeliegende Vertrauenskette ist dabei über den Datenraum durchgängig verifizierbar: Das D-CoC verweist auf die Akkreditierung der ausstellenden Stelle (DAkKS-Akkreditierungsnachweis), auf die Eignungsattestierungen der eingesetzten Prüfwerkzeuge und auf die Versionen der zugrundeliegenden SMART Standards.

Mit dieser Verkettung wird die in der QI-Digital-Roadmap angelegte Vision wirksam: Konformität wird nicht mehr als isolierte Bescheinigung übergeben, sondern als nachvollziehbares Geflecht aus Norm, Werkzeug, Prüfausführung

und ausstellender Stelle entlang der Wertschöpfungskette mitgeführt.

Weitere Anwendungsfälle

Über die ausführlich beschriebenen Anwendungen hinaus entstehen mit dem Anschluss an den Datenraum weitere Möglichkeiten, von denen einige hier kurz skizziert werden:

- **Norm-Update:** Erscheint ISO/IEC 98768 in einer neuen Version, prüft die Pipeline aus Szenario 1 nicht nur den textuellen Diff der Schwellenwerte, sondern auch im Datenraum, ob die genutzte PSI-Bibliothek noch eine gültige Eignungsattestierung für die neue Version trägt. Falls nicht, wird das im Merge Request markiert und das Unternehmen kann frühzeitig auf eine attestierte Variante umsteigen.
- **Beauftragung eines spezialisierten Prüfservice:** Für eine seltene, aber regulatorisch geforderte Bias-Prüfung über geschützte Gruppen beauftragt das Unternehmen über den Marktplatz einen Prüfservice einer akkreditierten Konformitätsbewertungsstelle. Der Service läuft als Compute-to-Data, das signierte Ergebnis fließt als zusätzlicher Beleg in den Konformitätsbericht ein. Das Unternehmen muss den entsprechenden Prüfaufbau nicht selbst betreiben.
- **Audit-Anfrage durch Marktüberwachung:** Eine Aufsichtsbehörde fragt rückwirkend nach Belegen für die Prüfung einer bestimmten Modellversion. Über die Datenraum-Bezüge sind alle Komponenten verifizierbar nachvollziehbar: welche Werkzeugversion mit welcher Eignungsattestierung, welcher Referenzbestand mit welcher Provenienz, welche Konformitätsbescheinigung von welcher akkreditierten Stelle. Die Behörde prüft die Verifiable Credentials direkt, ohne das Unternehmen zur Beibringung zusätzlicher Dokumente auffordern zu müssen.
- **Agentenbasierte End-to-End-Prüfung:** Der Chatbot aus Szenario 2 wird um eine Datenraum-Anbindung erweitert. Aus einer neuen Anforderung heraus löst er nicht nur den SMART Content auf, sondern findet auch passende Werkzeuge im Marktplatz, prüft Eignungsattestierungen und stößt die Prüfung an. Damit verbinden sich die Mechanismen aus Szenario 2 und 3 zu einem durchgehenden Workflow.

- **Eigenes Werkzeug im Marktplatz bereitstellen:** Aus dem laufenden Betrieb hat das Unternehmen eine credit-scoring-spezifische Drift-Variante entwickelt, die für seine Domäne stabilere Ergebnisse liefert als die generische PSI-Bibliothek. Das Unternehmen stellt sie als Open-Source-Werkzeug im Marktplatz bereit und lässt sie durch eine qualifizierte Stelle für ISO/IEC 98768 attestieren. Damit wird das Unternehmen vom reinen Konsumenten zum Beitragenden im Ökosystem.

5

DISKUSSION

5 DISKUSSION

5.1 Bewertung der Szenarien

Die in Kapitel 3 beschriebenen Szenarien stellen unterschiedliche Entwicklungsstufen für den Einsatz von SMART Standards bei der Prüfung von KI-Systemen dar. Gemeinsam verfolgen sie das Ziel, Prüfungen skalierbarer, reproduzierbarer und effizienter zu gestalten. Dabei unterscheiden sie sich hinsichtlich des erforderlichen Reifegrads von SMART Content, der technischen Infrastruktur sowie des organisatorischen Aufwands. Aus Anwendersicht stellt sich dabei weniger die Frage, ob einzelne Prüfschritte vollständig automatisiert werden können, sondern welchen praktischen Nutzen SMART Standards bei der Durchführung von KI-Prüfungen stiften und welche Voraussetzungen hierfür geschaffen werden müssen.

Chancen und Mehrwerte

Ein wesentlicher Mehrwert der dargestellten Szenarien liegt in der Reduktion manueller Tätigkeiten entlang des Prüfprozesses. Die Prüfung von KI-Systemen umfasst typischerweise eine Vielzahl wiederkehrender Aufgaben, darunter die Analyse normativer Anforderungen, die Erstellung von Prüfplänen, die Ableitung von Prüfkriterien, die Dokumentation von Ergebnissen sowie die Nachweisführung gegenüber Kunden, Auditoren oder Aufsichtsbehörden. Diese Tätigkeiten verursachen häufig einen erheblichen Teil des Gesamtaufwands einer Prüfung.

Bereits Szenario 1 ermöglicht es, normative Anforderungen direkt in Anforderungsmanagement-, Entwicklungs- und Prüfsysteme einzubinden. Dadurch können Medienbrüche reduziert, Anforderungen eindeutig referenziert und Änderungen an Normen gezielt nachverfolgt werden. Insbesondere die Nutzung granular referenzierbarer Norminhalte erleichtert die Versionierung und die Rückverfolgbarkeit von Anforderungen über den gesamten Prüfprozess hinweg.

Darüber hinaus fördern SMART Standards die Wiederverwendbarkeit von Prüfartefakten. Prüfpläne, Testfälle, Checklisten oder Konformitätsnachweise müssen nicht für jede Prüfung vollständig neu erstellt werden, sondern können auf bestehenden Artefakten aufbauen. Dies erhöht die Reproduzierbarkeit und Konsistenz von KI-Prüfungen und erleichtert den Vergleich von Prüfergebnissen über Organisationen und Anwendungsfälle hinweg.

Die Prüfung von KI-Systemen kann bereits jetzt mithilfe von SMART Standards skalierbar umgesetzt werden.

Für die Realisierung verschiedener Anwendungsfälle liegen die notwendigen Informationen oft schon im ReqIF-Format vor:

- Import granularer Norminhalte (Anforderungen, Definitionen, Querverweise) ins Anforderungsmanagement, mit eindeutigem Bezug zu Norm, Abschnitt und Version über persistente Identifier.
- Abbildung der Referenzkette zwischen Gesetzen, harmonisierten Normen und referenzierten Normen als explizite Relationen, einschließlich der Verknüpfung importierter Anforderungen mit den durch sie adressierten KI-VO-Artikeln.
- Anreicherung der Inhalte mit Metadaten wie Anforderungstyp (muss, soll, kann), Compliance-Labels oder thematischen Gruppierungen.
- Automatisierte Erkennung neuer Norm-Versionen über Identifier-Abgleich, Diff-Anzeige zwischen Versionsständen und Auslösung von Review-Workflows im Anforderungsmanagement.
- Aggregation von Anforderungen, Quellverweisen und Compliance-Labels für die Konformitätsdokumentation und Anbindung an Prüfergebnisse aus CI/CD- bzw. MLOps-Frameworks über die persistenten Identifier.

Ein weiterer Nutzen liegt in der Entlastung bei der Prüf vorbereitung. Die Identifikation relevanter Anforderungen, die Erstellung von Prüflisten sowie die Zuordnung von Anforderungen zu Prüfschritten kann durch SMART Standards erheblich unterstützt werden. Normänderungen können automatisiert erkannt und auf bestehende Prüfprozesse abgebildet werden, wodurch der Pflegeaufwand sinkt.

Dabei ist Skalierbarkeit nicht mit vollständiger Automatisierung gleichzusetzen. Viele Anforderungen aus KI-Normen erfordern weiterhin fachliche Bewertung, Kontextwissen und menschliche Entscheidungen. Die dargestellten Szenarien verfolgen daher nicht primär das Ziel, Prüfungen vollständig zu automatisieren, sondern wiederkehrende Tätigkeiten bei Vorbereitung, Durchführung und Dokumentation technisch zu unterstützen. Bereits die automatische Bereitstellung relevanter Norminhalte, die Unterstützung bei der Normenanalyse, die Generierung von Prüfartefakten oder die strukturierte Nachweisführung können den Aufwand von KI-Prüfungen erheblich reduzieren, ohne die Verantwortung für das Prüfurteil zu automatisieren.

Insbesondere Szenario 2 eröffnet darüber hinaus neue Möglichkeiten für den Einsatz generativer KI in Prüfprozessen. Bereits heute werden Sprachmodelle zur Analyse von Normen, zur Erstellung von Auditunterlagen oder zur Unterstützung von Konformitätsbewertungen eingesetzt. SMART Contents können dazu beitragen, diese Nutzung zuverlässiger und nachvollziehbarer zu gestalten, indem normative Inhalte strukturiert, versioniert und eindeutig referenzierbar bereitgestellt werden.

Voraussetzungen und Herausforderungen

Die Umsetzung der beschriebenen Szenarien setzt voraus, dass normative Inhalte in geeigneter Form bereitgestellt werden. Normtexte allein reichen hierfür häufig nicht aus. Für die technische Unterstützung von KI-Prüfungen werden zusätzlich strukturierte Anforderungen, Metadaten, Referenzinformationen, Prüfmethodiken, semantische Beschreibungen und geeignete Austauschformate benötigt.

Eine besondere Herausforderung besteht in der Operationalisierung normativer Anforderungen. Viele Anforderungen in KI-Normen sind bewusst prinzipienorientiert formuliert und enthalten nur wenige unmittelbar automatisierbare Schwellenwerte. Für ihre technische Nutzung müssen sie teilweise formalisiert, mit Prüfmethoden verknüpft oder durch zusätzliche Informationen ergänzt werden. SMART Standards

schaffen hierfür eine wichtige Grundlage, ersetzen jedoch nicht die fachliche Interpretation der Anforderungen.

Darüber hinaus gewinnt die semantische Anreicherung normativer Inhalte an Bedeutung. Soll aus Normanforderungen automatisiert Prüfunterstützung entstehen, müssen Begriffe, Zusammenhänge und Anforderungen maschinell interpretierbar beschrieben werden. Dies erfordert zusätzliche Kompetenzen in Bereichen wie Wissensmodellierung, Ontologie-Engineering und formaler Beschreibung normativer Inhalte.

Für eine automatisierte Verarbeitung steigt gleichzeitig die Bedeutung sprachlicher Präzision. Unklare oder mehrdeutige Formulierungen erschweren die technische Nutzbarkeit von Normen und erhöhen den Aufwand bei der Validierung der erzeugten Ergebnisse. Die konsequente Anwendung von Redaktionsrichtlinien wird daher zu einer wichtigen Voraussetzung für die Skalierbarkeit normbasierter KI-Prüfungen.

Eine besondere Rolle spielt dabei die menschliche Validierung. Automatisch erzeugte Prüfartefakte, Bewertungen oder Handlungsempfehlungen müssen weiterhin durch fachkundige Personen überprüft werden. Die Verantwortung für Prüfentscheidungen verbleibt unabhängig vom Automatisierungsgrad bei den jeweiligen Prüfern, Auditoren oder Konformitätsbewertungsstellen.

Darüber hinaus setzen insbesondere die Szenarien 2 und 3 zusätzliche technische und organisatorische Infrastrukturen voraus. Neben geeigneten Authoring- und Validierungswerkzeugen werden Verfahren zur Qualitätssicherung von SMART Contents, Referenzdaten, Prüfwerkzeugen und semantischen Modellen benötigt. Für viele dieser Aufgaben existieren bislang weder etablierte Prozesse noch klar definierte institutionelle Verantwortlichkeiten.

Risiken

Trotz ihres Potenzials sind mit den dargestellten Ansätzen auch Risiken verbunden. Ein wesentliches Risiko besteht im sogenannten Automation Bias. Automatisch erzeugte Vorschläge, Bewertungen oder Prüfergebnisse könnten ungeprüft übernommen werden, obwohl insbesondere KI-basierte Systeme weiterhin fehlerhafte oder irreführende Ergebnisse erzeugen können. Gleichzeitig besteht die Gefahr, dass der Aufwand für eine ausreichende menschliche Validierung einen Teil der erwarteten Effizienzgewinne wieder reduziert.

Ein weiteres Risiko besteht in einer Verengung des Prüfverständnisses auf diejenigen Anforderungen, die sich besonders gut formalisieren oder automatisieren lassen. Gerade bei KI-Systemen hängen zentrale Eigenschaften wie Robustheit, Fairness, Transparenz, menschliche Aufsicht oder Angemessenheit häufig vom Anwendungskontext ab. Ein starker Fokus auf automatisierbare Prüfkriterien könnte dazu führen, dass schwer operationalisierbare, aber dennoch relevante Anforderungen in den Hintergrund treten.

Mit zunehmender Formalisierung steigt außerdem das Risiko systemischer Fehlerfortpflanzung. Fehler in SMART Contents, Ontologien, Referenzimplementierungen oder Prüfartefakten können sich potenziell auf eine große Zahl von Prüfungen auswirken und dadurch weitreichende Folgen entfalten.

Hinzu kommen Governance-Fragen. Je stärker Prüfungen auf gemeinsam genutzten Inhalten, Werkzeugen und Diensten aufbauen, desto schwieriger kann die Zuordnung von Verantwortung werden. Dies betrifft insbesondere die Zusammenarbeit zwischen Normungsorganisationen, Tool-Anbietern, Prüfstellen und Anwendern. Mit zunehmender Arbeitsteilung besteht die Gefahr einer Verantwortungsdiffusion entlang der gesamten Wertschöpfungskette der Prüfung.

Im Ökosystem-Szenario entsteht darüber hinaus das Risiko einer Fragmentierung in mehrere konkurrierende und nicht interoperable Lösungen. Unterschiedliche Datenmodelle, Schnittstellen oder Governance-Ansätze könnten die angestrebten Skaleneffekte und die Wiederverwendbarkeit von Prüfartefakten einschränken.

Schließlich müssen auch in digital unterstützten Prüfprozessen die etablierten Grundprinzipien vertrauenswürdiger Prüfungen gewahrt bleiben. Dazu zählen insbesondere Vertraulichkeit, Datenschutz, Unabhängigkeit und Unparteilichkeit. Der Zugriff auf sensible Modelle, Datenbestände oder Dokumentationen erfordert daher geeignete technische und organisatorische Schutzmaßnahmen. Digitale Hilfsmittel die bei der Prüfung von KI-Systemen eingesetzt werden, dürfen diese grundlegenden Anforderungen nicht beeinträchtigen.

5.2 Einordnung in zentrale Konzepte der Digitalisierung der Normung

Vom seriellen zum parallelen Prozess

IDiS Whitepaper 3 quantifiziert den wirtschaftlichen Nutzen von SMART Standards in Normenanwendungsprozessen. Sein Mehrwertmodell zeigt, dass SMART Standards Aktivitäten, die heute seriell ablaufen müssen (Norm recherchieren, extrahieren, übertragen, umsetzen, prüfen) entkoppeln und parallelisieren können, mit messbaren Effekten in Zeit, Kosten und Qualität.

Die drei Szenarien konkretisieren diesen Wandel auf drei Ebenen. Szenario 1 löst die zentrale Extraktions- und Übertragungsgengstelle auf: Product Owner, Entwickler und MLOps Engineers greifen gleichzeitig auf dieselben granularen Norminhalte zu. Szenario 2 überträgt das Prinzip in den laufenden MLOps-Zyklus; ein KI-Agent ersetzt sequenzielle Einzelschritte durch nebenläufige Orchestrierung, und Normänderungen lösen eine zielgerichtete Neubewertung aus, ohne dafür einen vollständigen Prüfzyklus neu aufzusetzen. Damit wird das in Kapitel 2 beschriebene CABCA-Prinzip technisch realisierbar. Szenario 3 hebt die Parallelisierung auf die Ebene des Ökosystems: Normungsgremien, Prüforganisationen, Branchenverbände und Werkzeughersteller tragen unabhängig voneinander Inhalte bei und werden über Versionierung und Vertrauensnachweise koordiniert.

Semantische Integration

IDiS Whitepaper 1 beschreibt die Transformation von dokumentbasierten zu maschineninterpretierbaren Normen anhand eines Utility-Modells mit fünf Reifegradstufen. Sein konzeptioneller Kern ist der Wechsel des Bezugsobjekts: Auf unteren Stufen (Level 1 bis 2) steht das Dokument im Vordergrund, das digital verfügbar ist, aber manuell ausgewertet werden muss. Auf den höheren Stufen (Level 3 bis 5) verschiebt sich der Fokus auf die granularen Norminhalte selbst, die eindeutig identifiziert, über Ontologien verknüpft und mit Ausführungsinformationen angereichert sind. Semantische Integration bedeutet dann mehr als Maschinenlesbarkeit: Norminhalte fügen sich nahtlos in vorhandene Verarbeitungsketten ein und werden dort wie native Datenbestände behandelt.

Die drei Szenarien bilden verschiedene Utility-Stufen auf den Anwendungsfall KI-Prüfung ab. Szenario 1 setzt Level 3 bis 4

voraus und nutzt granulare, über einen Katalog adressierbare Inhalte werkzeuggestützt. Szenario 2 erfordert Level 4 mit formal codierten Constraints und Prüfbarkeits-Annotationen, auf deren Basis ein KI-Agent Prüfartefakte deterministisch generiert. Szenario 3 erweitert die Perspektive auf das Ökosystem: Die Norm wird nicht nur maschineninterpretierbar, sondern als souveräne, governance-fähige Ressource in einem Datenraum betrieben. Die Szenarien können damit als anwendungsfallspezifische Konkretisierung des Utility-Modells verstanden werden: Sie zeigen, welche Investitionen in welche Utility-Stufe welchen messbaren Nutzen bringen.

SMART Standards als Single Source of Truth

In der heutigen Praxis existieren Normanforderungen oft in mehreren parallelen Repräsentationen: im Originaldokument, in internen Paraphrasen im Anforderungsmanagement, in Testspezifikationen, in Prüfberichten. Jede dieser Repräsentationen ist manuell erstellt und wird bei Normänderungen selten vollständig nachgezogen, was zu divergierenden Ständen derselben Norm in unterschiedlichen Systemen führt. SMART Standards mit verifizierten Inhalten kehren dieses Muster um: Die Norm selbst ist die einzige maßgebliche Quelle, auf die alle Folgeartefakte über persistente Identifikatoren direkt referenzieren.

Diese Eigenschaft erhält im Zusammenspiel mit generativer KI eine zusätzliche Bedeutung: SMART Standards wirken als verifizierte Wissensbasis (Ground Truth), gegen die KI-Ausgaben geprüft und verankert werden. Statt sich auf Trainingswissen zu verlassen, beziehen Sprachmodelle Werte, Definitionen und Anforderungen direkt aus dem SMART Content. „Halluzinationen“ oder veraltete Trainingsstände werden so an der entscheidenden Stelle ausgeschlossen, und die Generierung von Prüfartefakten bleibt auf den jeweiligen Normabschnitt zurückführbar.

Die Szenarien setzen dieses Prinzip schrittweise um. In Szenario 1 wird die Norm zur Single Source of Truth für das Anforderungsmanagement: Jede importierte Anforderung behält den Bezug auf Normabschnitt und Version, und Normänderungen markieren betroffene Anforderungen automatisch. Szenario 2 erweitert das Prinzip auf von KI generierte Prüfartefakte: Grenzwerte etc. werden deterministisch aus dem SMART Content übernommen und nicht durch LLM-Inferenz erzeugt; die normativen Werte durchlaufen das Sprachmodell nicht und bleiben auf ihren Normabschnitt zurückführbar. Szenario 3 ergänzt diese technische Identität um kryptographisch verifizierbare Vertrauensnach-

weise (Verifiable Credentials), mit denen Herkunft, Akkreditierungsstatus und Vertrauensniveau jedes Norminhalts und Prüfartefakts maschinenlesbar attestiert werden.

Integration in die digitale Qualitätsinfrastruktur

Die QI-Digital-Roadmap beschreibt den Paradigmenwechsel von einer dokumentenbasierten zu einer datengetriebenen Qualitätsinfrastruktur: Normen, Prüfzeugnisse, Akkreditierungsnachweise und Konformitätsbescheinigungen werden nicht mehr als Dokumente versandt, sondern als maschinenlesbare Ressourcen in einem souveränen Datenraum ausgetauscht. SMART Standards bilden in dieser Architektur das normative Fundament, auf dem automatisierte Konformitätsprüfung, Digitale Konformitätsbescheinigung (D-CoC), Digital Calibration Certificates (DCC), der Digitale Produktpass (DPP) oder ein spezieller „KI-Pass“ aufsetzen.

Die Szenarien dieses Whitepapers beschreiben damit kein isoliertes KI-Prüf-Ökosystem, sondern konkrete Anwendungsfälle innerhalb dieser digitalen QI. Die Infrastrukturkomponenten aus Szenario 1 (semantische Suche, Content Delivery, Query-APIs) entsprechen denen, die QI-Digital generell für SMART Standards vorsieht; die in Szenario 2 generierten Prüfberichte und Konformitätsbescheinigungen sind dieselben Bausteine, die auch für DPP und KI-Pass benötigt werden; der Datenraum aus Szenario 3 kann in das Konzept von Quality-X integriert werden.

Das in Szenario 3 skizzierte Ökosystem liefert dabei die Vorprodukte, die in DPP oder einem KI-Pass zusammenfließen: qualitätsgesicherte Prüfwerkzeuge, nachvollziehbare Prüfausführungen und verifizierbare Eignungs- und Konformitätsattestierungen. Wer die Infrastruktur für skalierbare KI-Prüfung aufbaut, baut gleichzeitig an der digitalen QI insgesamt mit und kann die Szenarien als Testbed nutzen: einen Bereich mit klar umrissenen Akteuren, definierten Artefakten und verfügbaren Technologien, in dem sich Übergangskonzepte praktisch erproben lassen. Erste Erprobungsumgebungen bestehen bereits, etwa das Reallabor Additive Fertigung der Initiative QI-Digital, in dem SMART Standards, digitale Zertifikate und Datenraum-Komponenten gemeinsam getestet werden.

5.3 Übertragbarkeit

Physical AI

Die Szenarien fokussieren den Einsatz von Normen für Prüfungen von software-basierten KI-Systemen in MLOps-Umgebungen. Ein wachsendes und zunehmend praxis-relevantes Feld sind jedoch KI-Systeme, die in physische Produkte integriert sind, etwa eingebettete KI in Maschinen, autonome Systeme oder KI-gestützte Sensorik in Medizinprodukten. Die Prüfung solcher Systeme erfordert neben Software-Tests auch physische Prüfungen unter realen oder simulierten Umgebungsbedingungen.

Digitale Zwillinge (Digital Twins) bieten hier einen vielversprechenden Ansatz: Die virtuelle Repräsentation eines physischen Systems erlaubt Tests in simulierten Umgebungen, bevor Hardware verfügbar ist, und ermöglicht kontinuierliche Validierung im Betrieb. Besonders relevant ist in diesem Kontext die Verwaltungsschale (Asset Administration Shell, AAS), die als standardisiertes digitales Abbild physischer Assets auch KI-Komponenten, Trainingsdaten und Prüfergebnisse als Teilmodelle enthalten kann. IDiS hat in einem abgeschlossenen Pilotprojekt bereits die Integration von Norminhalten in AAS-Teilmodelle erprobt.

Für die in diesem Whitepaper beschriebenen Szenarien bedeutet das eine Erweiterung in zwei Dimensionen. Erstens könnten SMART Standards AAS-Teilmodell-Referenzen direkt mitliefern, sodass normative Anforderungen automatisch in das digitale Abbild des physischen Produkts übernommen werden. Zweitens müssten Prüfplattformen (Szenario 2) fähig sein, nicht nur Software-Tests, sondern auch Hardware-in-the-Loop-Prüfungen zu orchestrieren, unter Einbindung physischer Prüfumgebungen über standardisierte Schnittstellen. Szenario 3 würde dann auch physische Prüfstände und Kalibrierdienste als abonnierbare Ressourcen im Datenraum umfassen. Die konzeptionelle Architektur bleibt dabei dieselbe, während die Schnittstellen zur physischen Welt die zusätzliche Integrationsaufgabe darstellen würden.

Weitere Anwendungsgebiete

Wie diskutiert sind die drei Szenarien dieses Whitepapers nur teilweise KI-spezifisch, und schlagen im Kern Umsetzungsmöglichkeiten für zentrale Transformationskonzepte der Normung und Normanwendung vor. Sie beschreiben, wie maschinenlesbare Norminhalte in iterative Entwicklungs- und Betriebsprozesse integriert werden können, um Prüfung skalierbarer, reproduzierbarer und automatisierbarer zu machen. Diese Grundstruktur ist auf eine Vielzahl weiterer Prüfbereiche übertragbar, in denen Normen und Standards eine zentrale Rolle spielen.

Beispiele mit hoher Übertragbarkeit sind die Prüfung von Medizinprodukten (MDR, IEC 62304), funktionale Sicherheit in der Automatisierungstechnik (IEC 61508, ISO 26262), Cybersicherheit (IEC 62443), sowie Energiesysteme und Smart Grid, wo normenbasierte Anforderungen zunehmend in digitale Prozesse eingebettet werden müssen. In all diesen Bereichen wiederholt sich dasselbe Grundproblem: Normen liegen als unstrukturierte Dokumente vor, müssen manuell in Anforderungen übersetzt und in Prüfprozesse eingebettet werden, die aufgrund eines sich häufiger ändernden Prüfobjekts regelmäßig durchgeführt werden müssen.

Die technische Infrastruktur aus den Szenarien, also SMART Content, semantische Suche, maschinenlesbare Constraints, MCP-Schnittstellen, Datenraum-Governance etc., wäre in diesen Bereichen weitgehend identisch. Was sich ändert, sind die domänenspezifischen Normen und Ontologien. Dieses Whitepaper kann damit als Vorlage für ähnliche Szenarien in anderen Normungsbereichen dienen. Umgekehrt wäre eine branchenspezifische Differenzierung, etwa separate Szenarien für Medizinprodukte oder Cybersicherheit, ein wertvoller nächster Schritt, der auf der hier gelegten Grundlage aufbauen könnte.

Entwicklungen auf EU-Ebene

Die hier skizzierten Szenarien sind zudem anschlussfähig an politische Entwicklungen auf EU-Ebene. Der Final Report des Workstream 14 des High-Level Forum on European Standardisation⁵⁴ verfolgt für die Dateninteroperabilität einen

⁵⁴ European Commission, High-Level Forum on European Standardisation, Workstream 14 (2024): Final Report on Data Interoperability. Endorsed Version, April 2024. Online verfügbar unter <https://ec.europa.eu/docsroom/documents/58914>

Ansatz, in dem harmonisierte Europäische Normen die Anforderungen des Data Act technisch ausfüllen und über die Konformitätsvermutung regulatorisch wirksam werden. Empfehlung R.11 des Reports fordert zudem ein europäisches Rahmenwerk für vertrauenswürdige Ontologien, das Auswahl, Pflege und Vertrauenswürdigkeit semantischer Modelle regelt. Eine vergleichbare Governance-Frage stellt sich für SMART Standards, deren Verlässlichkeit dauerhaft von Pflege und institutioneller Verankerung der zugrundeliegenden Schemata abhängt.

Der Report benennt zudem ausdrücklich, dass KI-Modelle durch die Einbindung von Knowledge Graphs und Ontologien an Verlässlichkeit gewinnen und dass „Halluzinationen“ in sicherheitskritischen Domänen wie der Medizin ein zentrales Risiko darstellen. Die in Szenario 2 vorgeschlagene Trennung stochastischer und deterministischer Funktionen, in der das Sprachmodell sprachverarbeitende Aufgaben übernimmt und die SMART-Standard-Infrastruktur die normative Verankerung trägt, lässt sich als domänenspezifische Umsetzung dieses Prinzips für die normbasierte Konformitätsprüfung verstehen.

6

FAZIT

6 FAZIT

Mit der wachsenden Bedeutung Künstlicher Intelligenz wachsen die Anforderungen an Aspekte wie Sicherheit, Zuverlässigkeit, oder Transparenz. Sie werden in der Praxis über systematische Prüfungen auf Basis von Normen wirksam, deren Ausmaß die vorhandenen Prüfkapazitäten zunehmend übersteigen kann. Die richtige Antwort darauf liegt nicht darin, den Anspruch zu reduzieren, gefragt sind vielmehr Werkzeuge und Konzepte, die ihn einfacher und zuverlässiger umsetzbar machen. SMART Standards sind hier ein zentraler Baustein, da sie Norminhalte als maschinenlesbare Ressourcen verfügbar machen und damit das Fundament einer skalierbaren, normbasierten Prüfung von KI bilden.

Die beschriebenen Szenarien zeigen einen realistischen, stufenweisen Weg in Richtung einer digitalen Qualitätsinfrastruktur für KI, in der Normen, Prüfwerkzeuge und Prüfdaten als souveräne, maschinenlesbare Ressourcen verfügbar sind, durch die zuverlässige KI-Prüfungen teilweise automatisiert werden können.

Vier Transformationskonzepte liegen den Szenarien zugrunde und reichen über den Anwendungsfall KI-Prüfung hinaus. Der Übergang von seriellen zu parallelen Prozessen, die semantische Integration entlang des IDIS-Utility-Modells, die Etablierung der Norm als Single Source of Truth, sowie die Einbettung in die digitale Qualitätsinfrastruktur. Diese Konzepte sind nicht KI-spezifisch und lassen sich auf weitere Prüfbereiche übertragen, in denen normbasierte Prüfungen iterativ und in großer Zahl durchgeführt werden müssen.

Die technische Realisierbarkeit der Szenarien stützt sich auf vorhandene Standards, doch ihre vollständige Umsetzung setzt voraus, dass entsprechende Norminhalte (wie etwa Ontologien) schrittweise erstellt werden. Da eine rein manuelle Erstellung für den gesamten relevanten Normenbestand einen sehr hohen Aufwand bedeuten würde, kommt einer KI-unterstützten Generierung von SMART Content nach dem Human-in-the-Loop-Prinzip eine Schlüsselrolle zu. Hierbei übernehmen Sprachmodelle die aufwändige Erstextraktion aus dem Normtext, Normungsexperten validieren die Ergebnisse, sodass am Ende immer Verified Content entsteht, der die für die Szenarien erforderliche Zuverlässigkeit gewährleistet.

Skalierbare Prüfungen von KI-Systemen erfordern neben dem aus Normen generierten SMART Content weitere Ope-

rationalisierungselemente wie Prüfwerkzeuge und Referenzdatensätze, die zuverlässig miteinander verknüpft werden müssen, damit Prüfungen automatisiert ablaufen können. Da diese Bandbreite an Inhalten und Ressourcen nicht von einem einzelnen Akteur bereitgestellt werden kann, beschreibt das Whitepaper eine mögliche arbeitsteilige Verteilung mit verschiedenen Rollen. Der Datenraum aus Szenario 3 verbindet diese Beiträge zu einem konsistenten Prüfablauf und bleibt zugleich anschlussfähig für Erweiterungen wie die Prüfung physisch eingebetteter KI-Systeme über Verwaltungsschalen und Hardware-in-the-Loop-Prüfungen.

Die Szenarien dieses Whitepapers sind als Vorschlag aus der Perspektive (potenzieller) Anwender entstanden und bei IDIS diskutiert worden. Sie spiegeln nicht notwendigerweise die strategischen Pläne der beteiligten Normungsorganisationen wider und lassen sich auch nicht automatisch als Umsetzungsfahrplan lesen. Ihr Zweck ist es, den möglichen Beitrag von SMART Standards zur skalierbaren Prüfung von KI-Systemen frühzeitig konkret zu machen und damit eine Diskussionsgrundlage für die Weiterentwicklung von SMART Standards zu schaffen.

Vor diesem Hintergrund zielen die hier beschriebenen Szenarien nicht darauf, etablierte Prüfakteure zu ersetzen, sondern sie zu entlasten und in ihrer Arbeit zu stärken. Prüforganisationen, Beratungen, Auditoren und Konformitätsbewertungsstellen tragen Kompetenzen bei, die durch keine technische Infrastruktur ersetzbar sind. Sie legen Normen für konkrete Anwendungskontexte aus, bewerten Grenzfälle und tragen die Verantwortung für das Prüfurteil. SMART Standards und die in den Szenarien beschriebene Infrastruktur sind in diesem Sinne ein Angebot an diese Akteure. Sie sind ein Werkzeugkasten, mit dem sich wiederkehrende, formalisierbare Aktivitäten automatisieren lassen, sodass mehr Kapazität für die Aufgaben verbleibt, die fachliches Urteil und menschliche Verantwortung erfordern. Die in den Szenarien skizzierten Dienstleistungen, etwa der Betrieb von Prüfplattformen, die Pflege von Operationalisierungsinhalten oder die Bereitstellung qualitätsgesicherter Prüfdaten, sind dabei zugleich neue Geschäftsfelder, die auf den vorhandenen Kompetenzen dieser Akteure aufbauen.

Mehrere Entwicklungen müssen parallel vorangetrieben werden, damit die beschriebenen Szenarien praktisch tragfähig werden. Die technischen Konzepte und Werkzeuge hinter SMART Standards müssen weiterentwickelt und kon-

kretisiert werden. Für zentrale (KI-) Normen muss qualitäts-gesicherter SMART Content aufgebaut werden. Dabei geht es um die Etablierung KI-unterstützter Generierungsprozesse mit menschlicher Validierung, aber auch um andere technisch-organisatorische Projekte, wie der Pilotierung von Datenräumen, in denen Norminhalte, Prüfwerkzeuge und Referenzdaten als souveräne Ressourcen verfügbar gemacht werden. Organisatorisch müssen die Rollen der beteiligten Akteure entlang der vorgeschlagenen Aufgabenverteilung konsolidiert und Mechanismen für Kuratierung und Vertrauensbildung etabliert werden. Parallel sollten die Konzepte auf weitere Prüfbereiche übertragen werden, in denen sich vergleichbare Skalierungsprobleme stellen.

NÄCHSTE SCHRITTE

Dieses Whitepaper beschreibt kein abgeschlossenes Konzept, sondern ist eine Einladung zur Mitgestaltung. Die Weiterentwicklung von SMART Standards für die Prüfung von KI-Systemen ist eine Gemeinschaftsaufgabe, die technisches Know-how aus KI-Entwicklung ebenso erfordert, wie Expertise in Normung, Prüfung und Qualitätsinfrastruktur.

IDiS ist die zentrale Plattform für alle, die SMART Standards aktiv mitgestalten wollen. Die Mitarbeit ist entgeltfrei und steht Unternehmen, Verbänden, Behörden und wissenschaftlichen Einrichtungen offen. Inhaltlich arbeitet IDiS in thematischen Ad-hoc-Gruppen und in Pilotprojekten, in denen SMART Standards in konkreten Anwendungskontexten erprobt werden; zweimal jährlich tagt das IDiS-Plenum zum übergreifenden Austausch.

Wer die in diesem Whitepaper beschriebenen Szenarien, insbesondere die laufenden und geplanten Arbeiten zu SMART Content, Ontologien und Datenräumen, aktiv mitprägen möchte, ist bei IDiS richtig. Neben den Adhoc-Gruppen initiiert und begleitet IDiS auch Pilotprojekte, in denen SMART Standards in konkreten Anwendungskontexten erprobt werden.

Wer eigene Anwendungsfälle, Konzepte oder Ideen einbringen möchte, kann sich mit einem Projektvorschlag an IDiS wenden.

JETZT MITMACHEN:

dke.de/idis/mitwirken

Aktuelle Projekte und die IDiS-Roadmap:

dke.de/idis/pilotprojekte

ABKÜRZUNGEN

Benennung	
AAS	Asset Administration Shell (Verwaltungsschale)
API	Application Programming Interface
CaaS	Content as a Service
CABCA	Continuous Auditing Based Conformity Assessment
CI/CD	Continuous Integration / Continuous Delivery
DCAT-AP	Data Catalogue Vocabulary, Application Profile
DCC	Digital Calibration Certificate
D-CoC	Digital Certificate of Conformity
DevOps	Development Operations
DORA	Digital Operational Resilience Act
DPP	Digital Product Passport
EDC	Eclipse Dataspace Components
EDPS	Extended Dataset Profile Service
IDE	Integrated Development Environment
KI-VO	EU KI-Verordnung (AI Act)
LLM	Large Language Model
MDR	Medical Devices Regulation
ML	Machine Learning
MCP	Model Context Protocol
MLOps	Machine Learning Operations
OSD	Online Standards Development
OWL	Web Ontology Language

Benennung	
QI	Qualitätsinfrastruktur
RAG	Retrieval-Augmented Generation
RDF	Resource Description Framework
ReqIF	Requirements Interchange Format
RMS	Requirement Management System
SaaS	Software as a Service
SHACL	Shapes Constraint Language
VC	Verifiable Credentials



DIN e. V.

Burggrafenstraße 6
10787 Berlin
Telefon: +49 30 2601-0
E-Mail: presse@din.de
Internet: www.din.de



VDE e. V.

**DKE Deutsche Kommission Elektrotechnik
Elektronik Informationstechnik**

Merianstraße 28
63069 Offenbach am Main
Telefon: +49 69 6308-0
Telefax: +49 69 6308-9863
E-Mail: standardisierung@vde.com
Internet: www.dke.de

Stand: August 2026