Fraunhofer CML

Fraunhofer-Center für Maritime Logistik und Dienstleistungen CML

BSH
BUNDESAMT FÜR SEESCHIFFFAHRT UND HYDROGRAPHIE

VerifAI

Studie zur zielbasierten Standardisierung in der Prüfung und Zulassung intelligenter Entscheidungseinrichtungen von teilautonomen Überwasserfahrzeugen

Titelbild: Fraunhofer CML/SHansche – istockphoto.com

Fraunhofer
CML

BUNDESAMT FÜR
SEESCHIFFFAHRT
UND
HYDROGRAPHIE

VerifAI

Studie zur zielbasierten Standardisierung in der Prüfung und Zulassung intelligenter Entscheidungseinrichtungen von teilautonomen Überwasserfahrzeugen

Fraunhofer-Center für Maritime Logistik und Dienstleistungen

Paul Koch, M.Sc.
Thomas Stach, M.Sc.
Manfred Constapel, M.Sc.
Hans-Christoph Burmeister, Dipl.-Wirtsch.-Ing. Univ.

Inhalt

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AI	Artificial Intelligence
AIS	Automatisches Identifikationssystem
ANN	Artificial Neural Network
AP	Arbeitspaket
AS	Arbeitsschritt
BSH	Bundesamt für Schifffahrt und Hydrographie
CI	Computational Intelligence
CNN	Convolutional Neural Network
COLREGs	Convention on the International Regulations for Preventing Collisions at Sea, 1972
DIN	Deutsches Institut für Normung
DL	Deep Learning
DVO	Durchführungsverordnung
EG	Europäische Gemeinschaft
EU	Europäische Union
EVA	Eingabe, Verarbeitung und Ausgabe
GNSS	Global Navigation Satellite System
IMO	International Maritime Organization
IMU	Inertial Measurement Unit
KI	Künstliche Intelligenz
LSTM	Long Short-term Memory
MED	Marine Equipment Directive
MASS	Maritime Autonomous Surface Ships
ML	Machine Learning
MMSI	Maritime Mobile Service Identity
MRU	Motion Reference Unit
MS	Meilenstein
NMEA	National Marine Electronics Association
RADAR	Radio Detection and Ranging
RGB	Rot, Grün und Blau
sAI	Symbolic Artificial Intelligence
SOLAS	International Convention for the Safety of Life at Sea, 1974

1. Einleitung

1.1. Einordnung der Studie

Der maritime Sektor ist durch eine zunehmende Digitalisierung geprägt. Die Automatisierung und Teilautomatisierung von Prozessen an Bord haben in den letzten Jahren deutlich zugenommen. Neue Technologien setzen dabei neben klassischen Ansätzen zur regelbasierten Steuerung auf den Einsatz von Künstlicher Intelligenz (KI), um frühzeitig und proaktiv Situationen in der Navigation oder Betriebssicherheit eines Seeschiffes zu erkennen und kontextgerecht zu reagieren. Der Einsatz von KI könnte den wachsenden Druck auf Nautiker ausgelöst durch zunehmenden komplexen Schiffsverkehr und fehlendem Personal (Brooks & Greenberg, 2022; Minter, 2021) reduzieren und die Betriebssicherheit erhöhen (Daranda & Dzemyda, 2020; Yoshida et al., 2021). Während die Vorteile der Einführung KI-gestützter Systeme in der Industrie auf wachsendes Interesse stoßen, ist vor allem mit Blick auf ihrer Einführung eine Reihe von Herausforderungen zu bewältigen. Im Rahmen dieser Studie wird ein Konzept zur Prüffähigkeit, Technologiefolgenabschätzung und Zulassung dieser Systeme erarbeitet.

KI-basierte Systeme können Entscheidungsprozesse imitieren und ohne die explizite Beschreibung von Einzelschritten rationale oder regelbasierte Entscheidungen treffen. Sie nutzen hierbei verschiedene Mechanismen, um große Datenmengen im Betriebsablauf ohne menschlichen Eingriff zu interpretieren, Rückschlüsse zu ziehen und selbständig Handlungen durchzuführen (Norvig & Russell, 2021).

Die vorliegende Studie zielt auf die Prüfung von KI-Systemen, welche auch mit Methoden des maschinellen Lernens erstellt wurden, ab. Solche Modelle bringen eine Reihe an Herausforderungen mit sich, die bei der Entwicklung von Prüf- und Zulassungsprozessen eine zentrale Rolle spielen:

- Generalisierung der Problemdomäne
- Datenqualitätsmanagement in Entwicklungs- und Validierungsprozessen
- Komplexe und neue Modellarchitekturen

Für die Anpassung der Prüf- und Zulassungsprozesse bedeutet dies zum einen, dass teilweise Systeme betrachtet werden, deren Entscheidungen nicht ohne weiteres nachvollziehbar sind. Zum anderen müssen die Systeme aufgrund ihrer unvorhersehbar wachsenden Vielfalt und Vielzahl nicht individuell, sondern generisch betrachtet werden, um den Aufwand für Prüf- und Zulassungsprozesse in einem umsetzbaren und zukunftsfähigen Rahmen zu halten.

1.2. Zielvorgaben und Anforderungen

Zielstellung dieser Studie ist die Ausarbeitung eines Prüfkonzeptes für maritime KI-basierte, maschinelle Entscheidungssysteme sowie eines Sicherheitskonzeptes für die Hersteller von KI-basierten Schiffstechnologien. Das Prüfkonzept zeigt auf, wie das Bundesamt für Seeschifffahrt und Hydrographie (BSH) auf ordnungsgemäße Sicherheit und Funktion prüfen kann. Das Sicherheitskonzept unterstützt die Hersteller darin, relevante sicherheitstechnische Aspekte bei der Entwicklung zu berücksichtigen sowie ein KI-System prüffähig zu gestalten. Prüfkonzept und Sicherheitskonzept sind in ihrer Ausarbeitung aufeinander abgestimmt.

Die Studie gliedert sich in vier wesentliche Aspekte, welche im Verlauf des Projektes bearbeitet wurden:

- Sichtung und Bewertung des bestehenden Zulassungswesens (Kapitel 3) mit Blick auf die Eignung zur Prüfung von KI-Systemen unterschiedlicher Art. Teilergebnis ist die Ermittlung von Unzulänglichkeiten im aktuellen Prüf- und Zulassungswesen. Damit wird der notwendige Rahmen der nachfolgenden Arbeitspakete abgeschätzt und Anknüpfungspunkte mit bestehenden Prozessen werden entwickelt.
- Abschätzung der Informationsbedarfe autonomer Systeme durch eine gezielte Marktanalyse von KI-Systemen (Kapitel 4). Mit den Ergebnissen soll die Entwicklung des Prüfkonzeptes an Systeme der Industrie angepasst werden, um Systeme zeitnah in einen Prüfprozess überführen zu können.
- Erarbeitung eines Prüfkonzeptes (Kapitel 5 und 6), welches die verschiedenen Schritte aufzeigt, die für eine Prüfung von einem KI-System erforderlich sind. Das Prüfkonzept zeigt Empfehlung für Verfahrensweisen auf und beschreibt Kompetenzen und Prozesse, die für eine Zulassung notwendig sind.
- Entwicklung eines Sicherheitskonzeptes (Kapitel 7) zur Gewährleistung der Prüffähigkeit eines KI-Systems. Das Sicherheitskonzept enthält darüber hinaus Hinweise, welche bei der Entwicklung eines KI-Systems Beachtung finden müssen, um die Betriebssicherheit zu gewährleisten.

1.3. Planung und Ablauf der Studie

Für eine erfolgreiche und effiziente Durchführung der Studie wurde ein auf 12 Monate ausgelegter Zeitplan festgelegt. Unterteilt wurde die Durchführung der Studie in zwei Arbeitspakete (AP). Die AP sind wiederum in Arbeitsschritte (AS) untergliedert. Die AP und untergeordneten AS mit den jeweils geplanten Bearbeitungsdauern sind in ihrem Zeitplan in Abbildung 1 aufgeführt. Die Meilensteine (MS) tragen eine besondere Bedeutung, da der Projektfortschritt an diesen gemessen und auch gesondert vorgestellt worden ist.

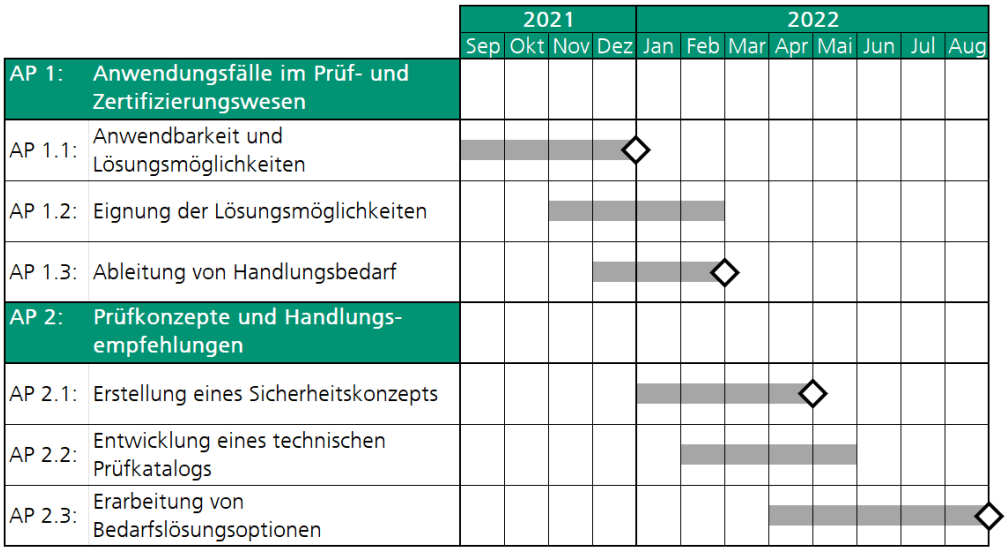

Abbildung 1: Zeitplan mit eingeordneten Arbeitspaketen (AP) zur Durchführung der Studie. Die Karos in dieser Abbildung stellen Meilensteine (MS) dar.

In AP 1 (vgl. „AP 1: Anwendungsfälle im Prüf- und Zertifizierungswesen“ in Abbildung 1) wurden Anwendungsfälle formuliert, mithilfe welcher die Eignung des bestehenden Prüf- und Zertifizierungswesen abgeleitet sowie der Handlungsbedarf abgeschätzt worden ist. In AP 2 (vgl. „AP 2: Prüfkonzepte und Handlungsempfehlungen“ in Abbildung 1) wurde ein Sicherheitskonzept und Prüfkonzept als Handlungsempfehlung erarbeitet. Jedem Arbeitspaket sind jeweils zwei Meilensteine zugeordnet. Die insgesamt vier Meilensteine sind in Tabelle 1 aufgeführt und als Karos im Zeitplan in Abbildung 1 dargestellt.

Tabelle 1: Meilensteine in der Projektumsetzung.

#	Meilenstein-Beschreibung	Monat
MS 1	Die Grenzen der Anwendbarkeit des Prüf- und Zertifizierungswesen für die Anwendungsfälle sind erhoben und dokumentiert.	4
MS 2	Die Eignung der von alternativen Prüf- und Zertifizierungsszenarien für die Gewährleistung der Betriebssicherheit ist analysiert.	6
MS 3	Ein Sicherheitskonzept zur Gewährleistung der Betriebssicherheit mit den Erkenntnissen aus AP1 ist erstellt.	8
MS 4	Bedarfslösungsoptionen und Handlungsempfehlungen nach aktuellem Stand der Wissenschaft und Technik sind etabliert.	12

Über diesen organisatorischen Rahmen hinaus fand ein regelmäßiger Austausch zwischen dem BSH und dem Fraunhofer CML statt. Für den Austausch wurde zunächst ein 2-Wochen-Rhythmus angesetzt, der nach Vorstellung von MS 2 zu einem 3-Wochen-Rhythmus ausgeweitet wurde. Die Abschlusspräsentation wurde am 15.09.2022 gehalten und die Studie im darauffolgenden Monat an das BSH übergeben.

1.4. Aufbau der Studie

Im Rahmen der Studie werden die Themenpunkte „Intelligente Entscheidungseinrichtungen“ mit der Spezifikation „Teilautonome Überwasserschiffe“ im Kontext der Prüfung und Zulassungsfähigkeit betrachtet. Die Bearbeitung der Aufgaben vor dem Hintergrund der definierten AP und AS verlief gemäß der Darstellung in Abbildung 2.

Die Studie ist wie folgt aufgebaut. Beginnend mit Kapitel 2 findet eine Einführung in wesentliche Begriffe und Definitionen statt. Kapitel 3 beginnt mit einer Aufarbeitung des bestehenden Zulassungswesens und einer Betrachtung aktuell vorhandener Verfahren, die bei der Konformitätsbewertung von maritimen Ausrüstungsgegenständen Anwendung finden. Aufbauend aus den vorgestellten Bestandteilen eines solchen Konformitätsbewertungsverfahren werden detailliert die Grenzen der Verfahren im Kontext der Zulassung von KI-Systemen aufgezeigt und erläutert.

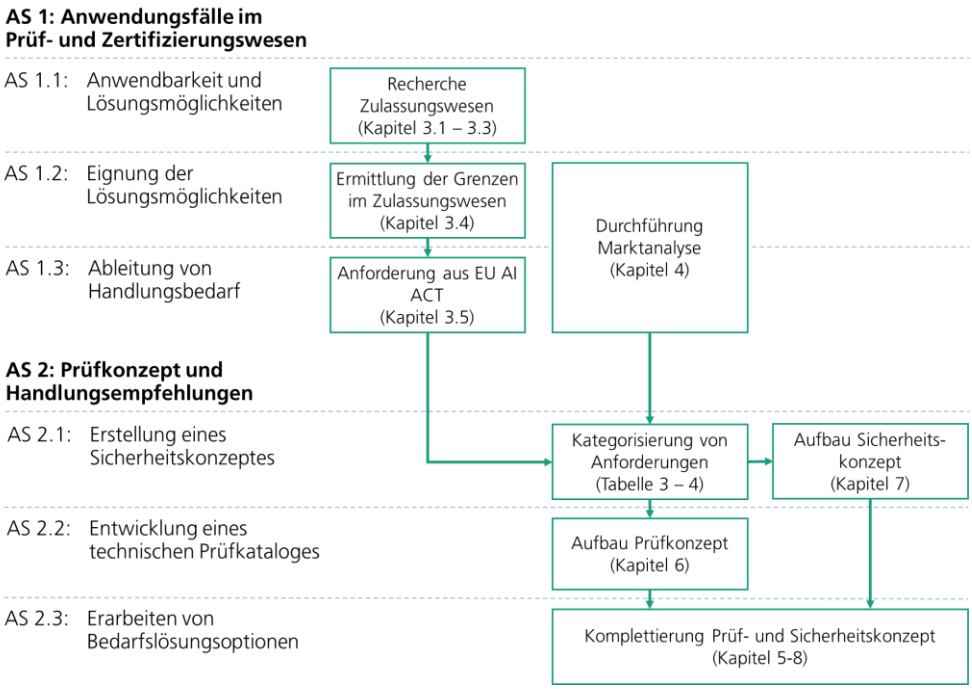

Abbildung 2: Arbeitsplan der Studie und Sequenz durchgeführten Aufgaben.

In Kapitel 4 werden in Kombination mit den ermittelten Anforderungen und einer im Rahmen der Studie durchgeführten Marktanalyse entsprechende Informationsbedarfe identifiziert und nach Datenquellen kategorisiert. Hierfür werden in einer Marktanalyse bereits verfügbare, aber noch nicht prüfbare Schiffstechnologien mit KI-Ansatz gesichtet und zusammengefasst. Die Ergebnisse dieser Marktanalyse sind ein zentraler Bestandteil der Studie, weil sie die Grundlage für die Ausarbeitung des Prüf- und Sicherheitskonzeptes bilden.

Basierend auf den technischen Grundlagen sowie dem identifizierten Handlungsbedarf werden in Kapitel 5 das Prüfkonzept und Sicherheitskonzept, sowie ihre Integration eingeführt. Anschließend wird in Kapitel 6 ein an das BSH gerichtetes Prüfkonzept vorgestellt. Das Prüfkonzept besteht aus einer Sequenz aus Prüfschritten, mit denen das BSH ein KI-basiertes System modellagnostisch prüfen kann. In Kapitel 7 wird ein an die Hersteller gerichtetes Sicherheitskonzept vorgeschlagen. Darin wird beschrieben, wie der Hersteller sein System auf die Prüfung vorbereiten kann, damit es prüffähig ist und eine gute Aussicht auf eine Zertifizierung hat.

In den Kapiteln 5 bis 7 sind Beispiele zur Anwendung der einzelnen Schritte des Prüf- und Sicherheitskonzeptes anhand eines fiktiven KI-basierten Produktes aufgeführt (s. Einführung in Kapitel 4.3). Diese Beispiele befinden sich in solchen türkisen Kästen.

Die Studie schließt in Kapitel 8 mit einer Zusammenfassung und den für das BSH identifizierten Handlungsempfehlungen ab.

2. Einführung von Begriffen und Definitionen

Die Thematik Künstliche Intelligenz hält Einzug in viele verschiedene Branchen – so auch in der maritimen Industrie. Dabei werden die Begriffe Künstliche Intelligenz und Autonome Schifffahrt sowie im Zusammenhang stehende Begriffe oftmals unterschiedlich definiert oder verstanden. Für das Grundverständnis ist es notwendig ein gemeinsames Verständnis über Begriffe und Konzepte zu schaffen und ihre Zusammenhänge zu erklären.

2.1. Künstliche Intelligenz als Entscheidungseinrichtung

Mit der Entwicklung der ersten Computer und ihrer stetig steigenden Rechenleistung rückte das Thema Künstliche Intelligenz nicht nur in die Aufmerksamkeit der Wissenschaft, sondern zunehmend in vielen Bereichen der Industrie. Die vorliegende Studie betrachtet künstlich intelligente Systeme als Entscheidungseinrichtungen zur Durchführung sicherheitskritischer Entscheidungen an Bord von teilautonom ausgerüsteten Überwasserschiffen.

Zur Fragestellung, ab wann man bei einem System von einer KI sprechen kann, existieren in der Fachwelt verschiedene Antworten (Norvig & Russell, 2021). Zur Definitionsauslegung einer KI kann zum Beispiel bewertet werden, wie rational und richtig oder wie menschlich die Entscheidungen von einem System getroffen werden. Diese Betrachtung des Systems kann sowohl intern, also hinsichtlich der sukzessiven internen Schlussfolgerungen, als auch extern, also hinsichtlich des äußerlichen Verhaltens oder Ergebnisses, unternommen werden.

In diesem Unterkapitel werden im weiteren Verlauf die Begriffe Machine Learning und Deep Learning als Teil von Künstlicher Intelligenz, wie in Abbildung 3 vereinfacht dargestellt, eingeordnet.

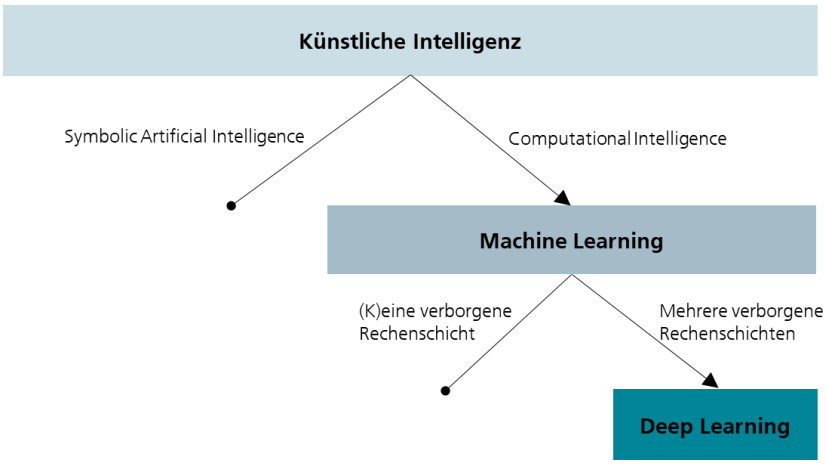

Abbildung 3: Zusammenhang zwischen Künstliche Intelligenz (KI), Machine Learning (ML) und Deep Learning (DL).

Grundlegend können KI-Ansätze einer der folgenden Methoden zugeordnet werden: Symbolic Artificial Intelligence (sAI) oder Computational Intelligence (CI) (Flasiński, 2016). Ein sAI-Modell zeichnet sich dadurch aus, dass es explizit definiert sein kann. Das heißt das Wissen ist symbolisch repräsentiert und „gedankliche" Prozesse im Modell sind als formale Operationen beschrieben. Als Beispiele können hier explizit

formulierte Entscheidungsbäume oder Expertensysteme aufgeführt werden. In einem CI-Modell hingegen sind Informationen grundsätzlich numerisch repräsentiert. Das heißt „gedankliche" Prozesse werden hauptsächlich in Form von numerischen Berechnungen durchgeführt und Wissen ist nicht zwangsläufig in expliziter Darstellung hinterlegt. Als Beispiel kann ein neuronales Netz aufgeführt werden, bei welchem das Wissen in Form eines Netzwerkes mit numerisch gewichteten Knotenpunkten gespeichert ist. Bei einem CI-Modell handelt es sich um ein Modell, welches auf Grundlage eines vorhandenen Datensatzes algorithmisch erstellt wurde. Der Prozess der Erstellung eines solchen Modells im Allgemeinen ist bekannt als Machine Learning (dt.: Maschinelles Lernen, ML) und die Erstellung und sukzessive Verbesserung des Modells im Spezifischen wird als Training bezeichnet.

2.1.1. Machine Learning

Zwei wesentliche Elemente im ML sind der Datensatz und der algorithmische Ansatz. Der Datensatz bildet die „Erfahrung" ab, mit welcher das KI-Modell trainiert wird. Das Training erfolgt dabei nach dem gewählten algorithmischen Ansatz. Datensatz, algorithmischer Ansatz und die verfolgte Anwendung befinden sich in einem Wechselspiel und können nicht unabhängig voneinander gewählt werden.

Die Datensätze und Daten, die beim ML zum Einsatz kommen, können sehr verschieden sein. Entsprechend relevant ist es, diese zu charakterisieren. Datensätze können sich hinsichtlich ihrer Modalität unterscheiden und Datentypen wie z.B. Bild-, Text-, Audio-, sequenzielle oder tabellarische Daten beinhalten. Neben dieser und weiterer möglichen Datentypen spielt auch die statistische Verteilung der Daten eine zentrale Rolle. Denn damit ein KI-Modell in der Realwelt zielführend funktioniert, muss die statistische Verteilung des für das Training verwendeten Datensatzes die zu erwartende statistische Verteilung aus der realen Anwendung widerspiegeln. Die statistische Verteilung der Daten lässt sich mit Methoden der deskriptiven Statistik beschreiben (Navlani et al., 2021).

Je nach angestrebter Anwendung und verfügbarem Datensatz eignen sich unterschiedliche ML-Ansätze. Ein ML-Ansatz lässt sich meist einem der folgenden drei grundlegenden Lernansätze zuordnen (Burkov, 2019):

- Supervised Learning (dt.: Überwachtes Lernen): Training mit einem Datensatz, in welchem die Zielwerte für korrespondierende Eingangswerte hinterlegt sind.
- Unsupervised Learning (dt.: Unüberwachtes Lernen): Training mit einem Datensatz ohne Verwendung von Zielwerten.
- Reinforcement Learning (dt.: Bestärkendes Lernen): Training, welches durch Bestärkung und Bestrafung für die Auswirkung von Entscheidungen, durchgeführt wird.

Wenn das Training eines Modells abgeschlossen ist, dann kann das Modell als eingefroren betrachtet werden. In diesem Fall ist das Verhalten des Modells deterministisch und damit reproduzierbar. Denn das zu erwartende Verhalten verändert sich nicht mit der Zeit.

Neben der Einordnung in Lernansätze, lassen sich ML-Ansätze auch hinsichtlich ihrer zu lösenden Aufgabe einordnen. Dabei besteht grundlegend ein Zusammenhang zwischen dem gewählten Lernansatz und lösbarer Aufgabentypen. Dies wird im weiteren Verlauf anhand von Abbildung 4 verdeutlicht. Grundlegend existieren folgende drei Aufgabentypen (Burkov, 2019):

- Klassifikation: Bei der Klassifikation werden eingehenden Daten Ausgabewerte zugeordnet, die bei einer erfolgreichen Ausgabe den erwarteten Zielwerten[1] entsprechen. Ein Zielwert ist Element einer Menge aus verschiedenen Klassen. Die Klassen repräsentieren diskrete Werte und können z.B. Zeichenketten, also Texte, als Datentyp darstellen.
- Regression: Wie bei der Klassifikation, werden bei der Regression den eingehenden Daten Zielwerte zugeordnet. Der Unterschied zur Klassifikation ist, dass es sich hier um Zielwerte aus einem kontinuierlichen Raum, also um z.B. reellwertige Zahlen, handelt.
- Clustern: Grundlegender Unterschied zur Klassifikation und Regression ist beim Clustern, dass es zu den Eingangswerten keine korrespondieren Zielwerte gibt. Beim Clustern werden basierend auf einem Datensatz Gruppen aus zueinander ähnlichen Daten erzeugt. Neue Eingangswerte können entsprechend der Clustering-Kriterien eingruppiert werden.

Künstliche Neuronale Netze (engl. Artificial Neural Network, kurz: ANN) sind ein spezieller Fall von Machine-Learning-Methoden. Ihr Name rührt daher, dass ihr Aufbau an den Neuronalen Netzen des menschlichen Gehirns angelehnt ist (Flasiński, 2016). ANNs besitzen grundlegend eine Eingangsschicht, eine oder mehrere verborgene Rechenschichten und eine Ausgabeschicht. Die Neuronen zwischen benachbarten Schichten können dabei netzwerkartig miteinander verbunden sein. ANNs lassen sich nicht eindeutig einem Lernansatz oder Aufgabentyp zuordnen. Ein ANN kann je nach Methode und Aufbau unterschiedliche Anwendungen, wie z.B. Objekterkennung oder Zeitreihen-Vorhersage, erfüllen.

Prominente Beispiele für ML-Methoden sind in Abbildung 4, hinsichtlich ihrer zugrundeliegenden Lernansätze und Aufgabentypen hierarchisch sortiert, aufgeführt. Die Abbildung stellt nur einen Ausschnitt an möglichen Lernansätzen, Aufgabentypen und Methoden dar. Für eine größere Übersicht und weiteren Beispielen wird auf (Sarker, 2021) verwiesen.

Eine geläufige Klassifikationsmethode ist der Decision Tree (Entscheidungsbaum) (Burkov, 2019; Sarker, 2021). Zur Erzeugung eines Decision Tree können verschiedene Algorithmen herangezogen werden. Schlussendlich besteht das Modell aus einem Wurzelknoten und mehreren Entscheidungsknoten. Letztere stellen Verästelungen dar und an welchen mittels Berechnungen der weitere Pfadverlauf evaluiert wird, der dann am Ende zu einem von vielen Blattknoten führt. An den Blattknoten werden die Ausgabewerte ermittelt und ausgegeben.

Eine Support Vector Machine („Stützvektormethode", kurz: SVM) ist ein Machine-Learning-Ansatz, der sich wohl für Klassifikations- als auch Regressionsprobleme eignet. Bei diesem Ansatz werden Datenpunkte durch eine oder mehrere Trennebenen (bzw. im linearen zweidimensionalen Fall durch Trenngeraden) in Gruppen getrennt. Es wird dabei die Trennebene gewählt, welche zu den Datenpunkten den größten Abstand hat. Die Klassifikation wird folglich durch die Aufteilung des Datenraumes durch die Trennebenen ermöglicht. Modifikationen des SVM-Ansatzes ermöglichen es, die Trennebene als Regressionsebene zu verwenden (Burkov, 2019; Sarker, 2021).

Eine geläufige Regressionsmethode ist die Lineare Regression (Burkov, 2019; Sarker, 2021). Hierbei wird eine Ausgleichsgerade entlang der Eingangswerte gelegt. Die Ausgleichsgerade kann beispielsweise durch die Methode der kleinsten Fehlerquadrate

[1] Diskrete Zielwerte, also insbesondere solche in Klassifizierungsproblemen, werden in der Fachliteratur i.d.R. als „Label" bezeichnet. In dieser Studie wird der allgemeinere Begriff Zielwerte verwendet.

optimiert werden. Anhand der Ausgleichsgerade lassen sich für Eingangswerte genäherte Ausgangswerte ablesen.

Ein einfaches Beispiel für eine Clustering-Methode ist die Methode k-means (Burkov, 2019; Sarker, 2021). Mit dieser Methode wird eine Anzahl von k Clustern, also Datenpunkt-Gruppen, gebildet. Jede Gruppe besitzt einen geometrischen Schwerpunkt. Ein neuer Datenpunkt wird einer Gruppe zugeordnet, indem über ein Abstandsmaß berechnet wird, welchem geometrischen Schwerpunkt aus welcher Gruppe er am nächsten ist.

Bei DBSCAN handelt es sich um eine Clustering-Methode, die bei Schiffsverkehrsdaten zum Einsatz kommt (Riveiro et al., 2018). Im Vergleich zu k-means wird bei DBSCAN nicht die Anzahl der zu erzeugenden Clustern im Vorfeld festgelegt. Stattdessen wird definiert, wie nah Datenpunkte benachbart sein müssen und wie viele hinreichend nah benachbarte Datenpunkte notwendig sind, um sie als eigenständiges Cluster zu definieren (Burkov, 2019). Aus dieser Vorgehensweise ergibt sich, dass nicht alle Datenpunkte einem Cluster zugehörig sein können oder müssen.

Bei Methoden des Reinforcement Learning (RL) werden Aktionen, z.B. Entscheidungen, sequentiell von einem Agenten umgesetzt (Sarker, 2021). Jede Aktion bringt eine Belohnung oder Bestrafung mit sich, worauf basierend der Zustand der Umgebungswahrnehmung des Agenten sich anpasst. Idealerweise passt sich der Agent langfristig der Umgebung so an, dass optimale Entscheidungen getroffen werden. Bei Reinforcement Learning wird zwischen modellbasierten und modellfreien Ansätzen unterscheiden. Bei einem modellbasierten Ansatz versucht der Agent ein Modell von seiner Umwelt aufzubauen indem es seine Entscheidung und die damit verbundene Belohnung mit seiner Vorhersage abzugleichen. Entscheidung trifft er damit unter Berücksichtigung seiner Wahrnehmung (Vorhersage) der Umwelt. Ein Agent in einem modellfreien Ansatz trifft eine Entscheidung basierend allein auf seiner Erfahrung, ohne eine Vorhersage a priori durchzuführen.

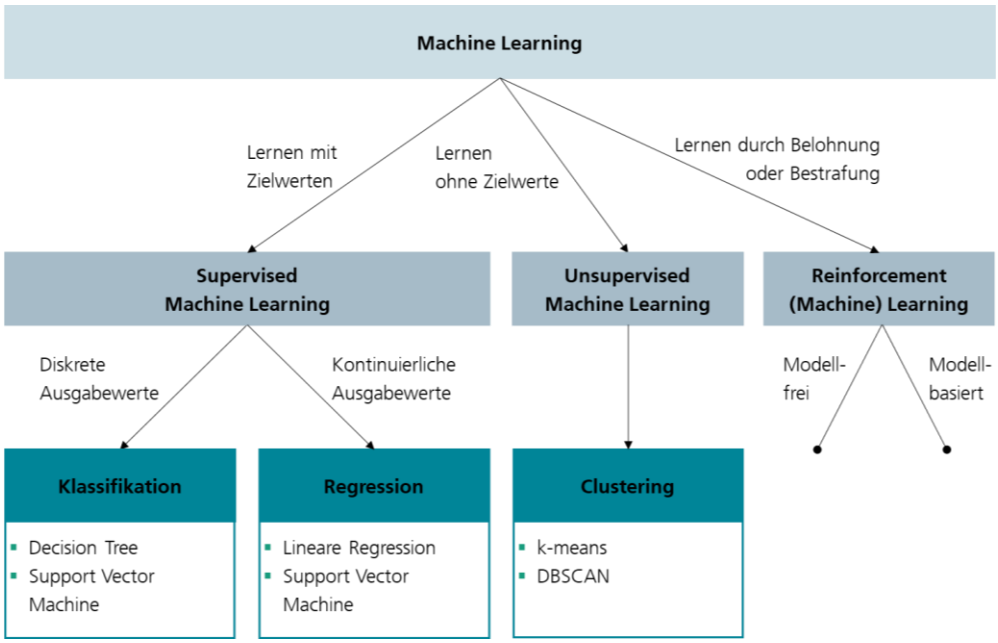

Abbildung 4: Beispielhafte Machine-Learning-Methoden sortiert nach Lernansatz und Aufgabentyp.

Werden bei einem Machine-Learning-Ansatz mehrere Rechenschichten verwendet, so spricht man von Deep Learning (dt.: tiefes Lernen, DL) (Norvig & Russell, 2021). Deep-Learning-Methoden kommen zum Beispiel bei visueller Objekterkennung zur Anwendung, wo in der Regel ein Convolutional Neural Network (CNN, dt.: faltendes neuronales Netz) als Modell verwendet wird. Sofern es sich bei dem mit DL trainiertem Modell um ein neuronales Netz handelt, wird das Modell auch als ein tiefes neuronales Netz bezeichnet.

Sowohl in der Entwicklungsphase beim Training eines Machine-Learning-basierten Modells als auch während der Anwendungsphase kann es dazu kommen, dass das Modell weniger zuverlässige Ausgaben liefert. Ein bekanntes, aber lösbares, Problem stellt das Under- oder Overfitting dar (Burkov, 2019; Norvig & Russell, 2021). Wenn ein Modell bei seinen Trainingsdaten zu viele Fehler, also stark abweichende Ausgabewerte, liefert, dann spricht man von Underfitting. Underfitting kann aus für das Modell wenig informativen Daten oder der Wahl eines zu einfachen Modells folgen. Abbildung 5 illustriert dies am Beispiel einer linearen Regression, die an Datenpunkten genähert wird, die näherungsweise einer quadratischen Funktion folgen. Beim Overfitting hingegen ist das Modell zu stark auf die Trainingsdaten ausgerichtet und versagt bei der richtigen Ausgabe für Eingabewerte, die nicht in den Trainingsdaten enthalten waren. In Abbildung 5 ist dies daran erkennbar, dass der Fit alle Datenpunkte trifft, aber nicht mehr dem näherungsweisen quadratischen Verlauf der zugrundeliegenden Funktion folgt. Overfitting kann aus einem zu komplexen Modell (im Beispiel einer Näherungsfunktion eines zu hohen Grades) oder zu wenig Trainingsdaten folgen. Bei Overfitting spricht man auch davon, dass das Modell eine hohe Varianz besitzt, da das Modell sehr stark streut.

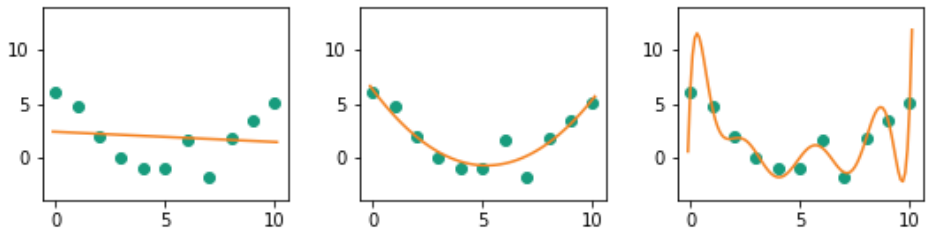

Abbildung 5: Eine quadratische Verteilung von Punkten mit Ausreißern (grün) und zugehörige Ausgleichsfunktionen (orange). Underfitting (links), guter Fit (mitte) und Overfitting (rechts).

Die Performanz von KI-Systemen, die während der Entwicklungsphase erfolgreich funktioniert haben, kann sich in der Anwendungsphase mit der Zeit verschlechtern. Die Ursache hierfür ist weniger eine intrinsische Verschlechterung am KI-System selbst (welches im Verhalten als eingefroren betrachtet wird), sondern Änderungen an der Umwelt, in der das KI-System agiert. Diese Veränderung von der Beziehung zwischen Modell und Realität wird als Konzept-Drift bezeichnet (Žliobaitė, 2010). Ein Drift kann dabei unterschiedliche Formen annehmen:

- Statistische Verteilung der Daten verändert sich mit der Zeit.
- Die Beziehung zwischen Modell-Ein- und -Ausgabe verändert sich mit der Zeit.
- Die ursprünglich angenommene Ground Truth, beispielsweise Bezeichnung oder Auswahl von Zielwerten, ändert sich.

Aus diesem Grund spielt es eine große Rolle die Zuverlässigkeit von (eingefrorenen) KI-Systemen auch nach ihrer Entwicklung zu prüfen.

2.1.2. Nachvollziehbarkeit von Entscheidungen

Die Nachvollziehbarkeit von Entscheidungen von KI-basierten Modellen ist stark vom gewählten ML-Ansatz des Modells abhängig. Der ML-Ansatz bestimmt darüber, ob eine künstliche Entscheidungseinrichtung aus einem tiefen neuronalen Netz oder Entscheidungsbaum besteht und damit, ob das Modell weniger oder mehr nachvollziehbar ist. Die Nachvollziehbarkeit eines Systems ist anhand seiner Interpretierbarkeit oder Erklärbarkeit beschreibbar. Interpretierbarkeit und Erklärbarkeit sind dabei unterschiedliche Begriffe im Kontext von Künstlicher Intelligenz (Norvig & Russell, 2021).

Modelle mit hoher Interpretierbarkeit sind solche, anhand deren inneren Aufbaus intuitiv erkennbar ist, warum das Modell zu jenen Entscheidungen gelangt. Ein Beispiel hierfür ist ein Entscheidungsbaum, weil dieser durch seine einfache Wenn-Dann-Struktur den Weg seiner Entscheidung menschlich verständlich wiedergibt. Die Entscheidungsfindung ist also ohne ein externes System nachvollziehbar. Dies ist schematisch in Abbildung 6 links dargestellt.

Konträr dazu ist ein tiefes neuronales Netz nicht ohne Weiteres interpretierbar und damit nachvollziehbar (s. Abbildung 6, rechts). Denn beim Einblick in seine Struktur ist eine Vielzahl Neuronen erkennbar (s. hellblaue Punkte in Abbildung 6, rechts), die parallel in hintereinandergeschalteten Rechenschichten (s. Reihen aus hellblauen Punkten in Abbildung 6, rechts) aufgebaut sind. Jedes Neuron besteht darüber hinaus aus einer Aktivierungsfunktion und einer Gewichtung zwischen 0 und 1 (anstelle einer Wenn-Dann-Struktur). Das Modell und seine Entscheidungsfindung ist somit für das menschliche Verständnis unzugänglich. In diesem Fall wird das Modell als Black-Box bezeichnet. Mit Hilfe eines externen Moduls lässt sich Erklärbarkeit einführen. Zum Beispiel könnte bei einer visuellen Schiffserkennung ein Modul herangezogen werden, welches Merkmale markiert, die es typisch für ein Objekt von der Art „Schiff" hält. Das Forschungsfeld, das sich mit der Erklärbarkeit von KI beschäftigt wird als Explainable AI bezeichnet (Samek & Müller, 2019).

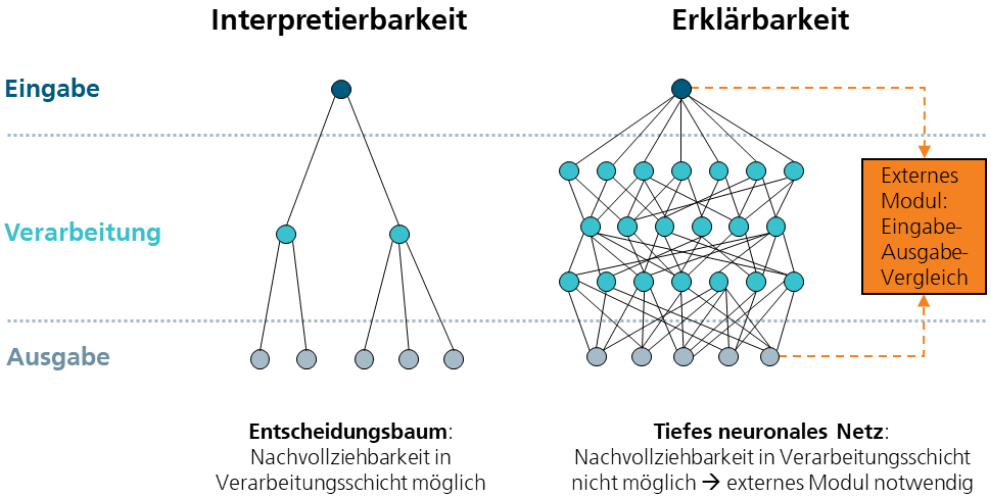

Abbildung 6: Unterscheidung zwischen Interpretierbarkeit und Erklärbarkeit bei KI-Modellen. Das Beispiel zeigt einen Entscheidungsbaum (links) und ein tiefes neuronales Netzes (rechts). Zur Nachvollziehbarkeit des neuronalen Netztes wird ein externes Zusatzmodul für einen Eingabe-Ausgabe-Vergleich benötigt.

Weller führt den Begriff der Transparenz ein, dessen Bedeutung davon abhängt, welcher Personenkreis mit dem betrachteten KI-System in Berührung kommt (Weller, 2019). So bedeutet Transparenz für einen Hersteller etwas anderes als für eine Prüfstelle. Ersterer interessiert sich z.B. dafür, wie die Neuronen innerhalb eines CNN zu der Klassifizierung von bestimmten Objekten beitragen. Das Interesse von Letzterem

hingegen sollte insbesondere darin liegen, die Richtigkeit der Entscheidungen von einem KI-System zu prüfen. Vereinfacht ausgedrückt versucht der Hersteller die Frage nach dem „wie“ und die Prüfstelle nach dem „ob“ zu beantworten. Damit wird Explainable AI, welches sich mit der Frage nach dem „wie“ beschäftigt, in dieser Studie nicht als Lösungsansatz gewertet und verfolgt.

2.2. Teilautonome Überwasserfahrzeuge

Maritime autonome Überwasserfahrzeuge, bekannt als MASS (engl.: maritime autonomous surface ships), sind Schiffe, die in unterschiedlichem Maße unabhängig von menschlicher Beteiligung betrieben werden können. Betrieb umfasst zum einem Teilaufgaben des Schiffbetriebes wie Maschinenüberwachung, Lagebilderfassung oder Navigation als auch die vollumfängliche Aufgabe der sicheren Schiffsführung.

Der Begriff autonom beschreibt hierbei nicht die Selbstbestimmung des Systems im eigentlichen Sinne, sondern die Automatisierung von Betriebsabläufen (Etzkorn, 2022).

Nach Untersuchungen der International Maritime Organization (IMO) können teil- und vollautonome Überwasserfahrzeuge in vier Autonomiegrade unterteilt werden (IMO, 2022), welche in Tabelle 2 dargestellt sind. Zweck dieser Unterteilung ist die Kategorisierung von Entwicklungen im MASS-Sektor und die Identifikation von Herausforderungen bei der Zulassung und dem Betrieb von MASS.

Tabelle 2: Autonomiegrade nach Untersuchungen der IMO im Rahmen der IMO Scoping Excercise.

Grad der Autonomie	Bedeutung
1	*Schiff mit automatisierten Prozessen und Entscheidungshilfen:* Seeleute sind an Bord, um Systeme und Funktionen an Bord zu bedienen und zu steuern. Einige Vorgänge sind automatisiert und unbeaufsichtigt, können aber jederzeit durch den Menschen unterbrochen werden.
2	*Ferngesteuertes Schiff mit Seeleuten an Bord:* Das Schiff wird von einem anderen Ort aus gesteuert und bedient. Seeleute sind an Bord, um die Kontrolle zu übernehmen und die Systeme und Funktionen an Bord zu überwachen.
3	*Ferngesteuertes Schiff ohne Seeleute an Bord:* Das Schiff wird von einem anderen Ort aus gesteuert und betrieben. Es sind keine Seeleute an Bord.
4	*Völlig autonomes Schiff:* Das Betriebssystem des Schiffes ist in der Lage Entscheidungen zu treffen und Aktionen selbst zu bestimmen.

Die vorliegende Studie fokussiert sich auf teilautonome Überwasserfahrzeuge, betrachtet also Schiffe, welche unter anderem mit autonomen Systemen ausgestattet oder bei denen bestimmte Prozesse durch maschinelle Entscheidungseinrichtungen ersetzt sind. Die Fernsteuerung von Systemen wurde in dieser Studie nicht betrachtet. Entsprechend fokussiert sich diese Studie auf Schiffe und Schiffstechnologien der Autonomiegrade 1 und 4 nach IMO.

Die Studie betrachtet Einzelsysteme mit domänenspezifischen Aufgaben und vereinfacht damit die Entwicklung von Prüfprozessen. Komplexe KI-Systeme werden im

Rahmen der Studie auf ihre Komponenten heruntergebrochen und einzeln geprüft. Das Prüf- und Sicherheitskonzept zielt weiterführend darauf, unabhängig vom Grad der Autonomie, allgemeingültige Prozesse zu entwickeln, welche sich auf einzelne KI-Systeme als auch auf eine große Sammlung von KI-Systemen übertragen lassen.

3. Prüf- und Zertifizierungswesen

Die Markteinführung neuartiger Bauteile für ein Schiff (Schiffsaurüstung) erfordert nach der International Convention for the Safety of Life at Sea (SOLAS) die Zertifizierung und Prüfung der Funktionsweise des Gerätes sowie den Herstellungsprozess und Betrieb des Ausrüstungsgegenstandes an Bord des Schiffes. Insbesondere bei der Zulassung von Einrichtungen, mit dem Ziel Prozesse an Bord eines Schiffes zu autonomisieren, ist eine umfassende Prüfung notwendig, um die Betriebssicherheit dieser Systeme zu gewährleisten. Dieses Kapitel erläutert die Prozesse, die für eine Zulassung und Prüfung eines Ausrüstungsgegenstandes notwendig sind und zeigt die Grenzen bestehender Verfahren auf.

3.1. EU-Konformitätsbewertungsverfahren

Die Überprüfung der Sicherheit von Schiffsausrüstung in der Europäischen Union (EU) wird gemäß der Marine Equipment Directive (MED) mittels eines Konformitäts-bewertungsverfahrens durch notifizierte Stellen durchgeführt. Notifizierte Stellen sind von nationalen Behörden akkreditierte Institutionen, die den Auftrag dazu bekommen, Überprüfungsverfahren durchzuführen. Die Sicherstellung der Konformität der Produkte, die auf dem europäischen Markt eingeführt werden sollen, finden hinsichtlich ihres Entwurfs, dem Bau und ihrer Leistung statt. Die EU skizziert den Prozess der Konformitätsbewertung mit seinen möglichen Prüfmodulen und -optionen im Rahmen der Schiffsausrüstungsverordnung (Europäisches Parlament und Rat der Europäischen Union, 2014).

Allgemein kann beim Konformitätsbewertungsverfahren in Abhängigkeit von der Produktionsart zwischen zwei Prüfoptionen unterschieden werden, welche teilweise in weiterführende Module unterteilt werden können. Die in Abbildung 7 dargestellten Module werden in den folgenden Abschnitten erläutert und mit Blick auf ihre Zielsetzung betrachtet.

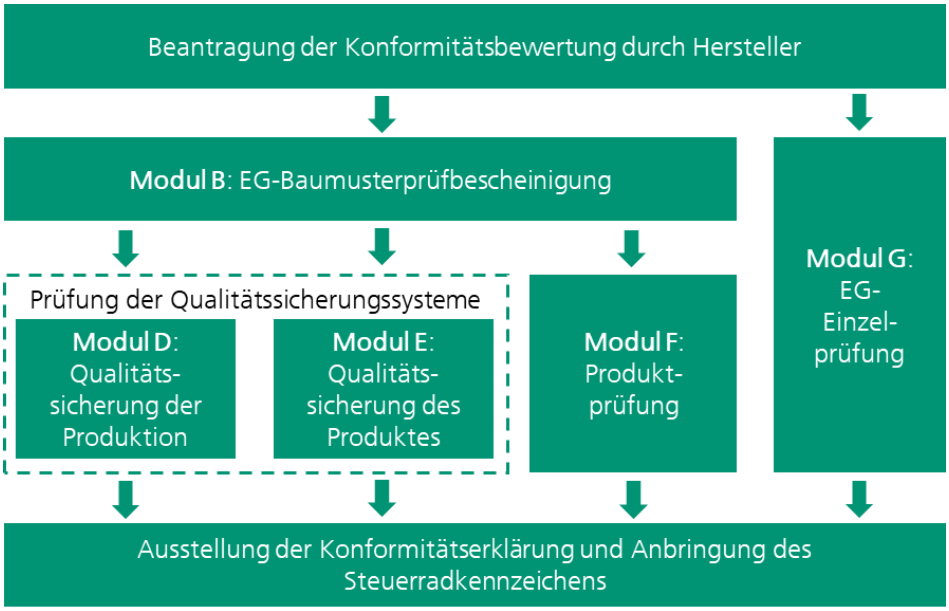

Abbildung 7: Konformitätsbewertungsverfahren nach MED.

Wird ein Produkt mittels einer Massen- oder Serienfertigung produziert, so ist eine EG-Baumusterprüfung (Modul B) (EG steht für Europäische Gemeinschaft) sowie nach Wahl des Herstellers und Art des Produktes eines der folgenden Prüfmodule vorgeschrieben (Europäisches Parlament und Rat der Europäischen Union, 2014):

- Qualitätssicherung der Produktion (Modul D)
- Qualitätssicherung des Produktes (Modul E)
- Prüfung der Produkte (Modul F)

Wird ein Produkt nicht durch Serien- oder Massenfertigung, sondern entweder einzeln oder in kleinen Mengen produziert, so wird dieses durch eine EG-Einzelprüfung (Modul G) auf seine Sicherheit und Qualität überprüft (Europäisches Parlament und Rat der Europäischen Union, 2014).

3.1.1. Modul B – EG-Baumusterprüfung

Die Baumusterprüfung untersucht den technischen Entwurf der Schiffsausrüstung auf Sicherheit und Konformität. Ziel dieser Überprüfung ist es, die Angemessenheit des technischen Entwurfs, die Übereinstimmung des Musters mit den technischen Unterlagen sowie die Konformität mit bestehenden Normen und Richtlinien sicherzustellen.

Die Prüfung kann entweder anhand repräsentativer Muster des vollständigen Produktes oder eines bzw. mehrerer wichtiger Teile des Produkts durchgeführt werden. Die zweite Option setzt zusätzlich eine Bewertung der Eignung des technischen Entwurfs anhand technischer Unterlagen und Nachweise voraus (Europäisches Parlament und Rat der Europäischen Union, 2014).

3.1.2. Modul D – Qualitätssicherung der Produktion

Modul D untersucht das Qualitätssicherungssystem des Produktionsprozesses auf seine Konformität mit internationalen Normen und Standards. Das zu überprüfende System gewährleistet sowohl die Übereinstimmung der Produkte mit der zuvor in Modul B überprüften Bauart als auch die Qualitätsziele der hergestellten Produkte über den vollständigen Produktionsprozess. Um dies sicherzustellen wird die Produktqualität vor, während und nach der Herstellung überprüft (Europäisches Parlament und Rat der Europäischen Union, 2014).

3.1.3. Modul E – Qualitätssicherung des Produktes

Im Prüfmodul E wird die Konformität anhand einer Überprüfung des Qualitäts-sicherungssystems des Produktes untersucht. Wie im Modul D gewährleistet das zu überprüfende System die Übereinstimmung der Produkte mit der in Modul B überprüften Bauart. Darüber hinaus stellt es sicher, dass die Produkte am Ende des Produktionsprozesses die anvisierte Produktqualität erzielen. Um dies sicherzustellen, wird die Produktqualität bei Endabnahme der Produktion überprüft (Europäisches Parlament und Rat der Europäischen Union, 2014).

3.1.4. Modul F – Prüfung der Produkte

Im Prüfmodul F werden einzelne Produkte nach dem Produktionsprozess durch die notifizierten Stellen auf ihre Konformität kontrolliert. Dazu muss der Hersteller zunächst die Übereinstimmung der Produkte mit der in Modul B überprüften Bauart und die Übereinstimmung der Anforderungen an den Fertigungsprozess gewährleisten. Der Hersteller kann in einem weiteren Schritt entweder ein Prüfungsverfahren anhand einer Überprüfung jedes einzelnen Produkts oder anhand einer Überprüfung durch statistische Mittel wählen. Sollte zweiteres gewählt werden, muss die Einheitlichkeit aller produzierten Lose sichergestellt sein (Europäisches Parlament und Rat der Europäischen Union, 2014).

3.1.5. Modul G – Einzelprüfung

Wird ein Produkt einzeln oder in kleinen Stückmengen gefertigt, wird dieses im Rahmen des Konformitätsbewertungsverfahrens durch Modul G einzeln geprüft. Dazu müssen in einem ersten Schritt die zugrunde liegenden technischen Unterlagen auf ihre Konformität mit existierenden Normen und Richtlinien untersucht sowie Konstruktionsberechnungen und -prüfungen nachvollzogen werden. In einem weiteren Schritt werden die einzelnen Produkte auf ihre zuvor definierten Anforderungen überprüft (Europäisches Parlament und Rat der Europäischen Union, 2014).

3.2. Durchführungsverordnung

Zusätzlich zur MED, in der das Verfahren zur Konformitätsbewertung definiert wird, veröffentlicht die Europäische Kommission die Durchführungsverordnung (DVO) (Europäische Kommission, 2022). In dieser werden die verschiedenen Entwurfs-, Bau- und Leistungsanforderungen sowie die Prüfnormen für Schiffsausrüstungen festgelegt. Die unterschiedlichen Ausrüstungen werden in den folgenden neun Kategorien eingeordnet und veröffentlicht (Europäische Kommission, 2022):

1. Rettungsmittel
2. Verhütung der Meeresverschmutzung
3. Brandschutzausrüstung
4. Navigationsausrüstung
5. Funkausrüstung
6. Ausrüstung nach der Kollisionsverhütungsregulierung von 1972 (COLREGs)
7. Sonstige Sicherheitsausrüstung
8. Ausrüstung nach SOLAS Kapitel II-1
9. Ausrüstung, für die der Normensatz für die MED-Zertifizierung nicht vollständig ist

Lassen sich Schiffsausrüstungen thematisch in die Kategorien 1 bis 8 eingliedern und sind keine nach Bestimmungen der IMO notwendigen Normen und Anforderungen vorhanden oder angemessen, so sind diese in Kategorie 9 eingegliedert (Europäische Kommission, 2022).

3.3. Nationale Zulassung

Für Schiffsausrüstungen, die nach deutschem Recht zulassungspflichtig sind und für die keine international harmonisierten Anforderungen bestehen, kann eine nationale Zulassung erteilt werden. Diese Zulassung muss von anderen europäischen Mitgliedsstaaten anerkannt werden, falls das garantierte Sicherheitsniveau mit den

Regeln des jeweiligen Landes übereinstimmt (Bundesamt für Schifffahrt und Hydrographie, 2022).

3.4. Unzulänglichkeit der ermittelten Normen und Richtlinien

Um KI-Systeme auf Schiffen mittels des EU-Konformitätsbewertungsverfahrens zertifizieren zu können, muss vorher sichergestellt werden, dass die verschiedenen Prüfelemente auf diese Systeme anwendbar sind. Um dies zu überprüfen, werden im Folgenden die aktuelle DVO und die Prüfverfahren auf ihre Konformität mit den Anforderungen an KI-Systemen untersucht.

Die 2021 verfasste, aktuelle Fassung der DVO (Europäische Kommission, 2022) enthält keine Verfahren zur Testung von KI-Systemen. Es gibt weder eine mögliche Kategorie zur Einordnung von solchen Systemen noch ein Verfahren, nach welchem KI-Systeme eingegliedert werden können. Eine Eintragung in die DVO benötigt ausreichende Normen und Spezifikationen, welches die nötigen Anforderungen an ein solches System beschreibt. Im Allgemeinen können Prüfnormen für ein EU-Konformitäts-bewertungsverfahren von unterschiedlichen Organisationen festgelegt werden (Europäisches Parlament und Rat der Europäischen Union, 2014). Hierin ist insbesondere die IMO dafür zuständig, eine international einheitliche Regulierung von autonomen und teilautonomen Schiffen zu entwickeln (Danish Maritime Authority, 2018).

Um das Fundament zur Regulierung von KI-Systemen zu legen, hat die IMO eine Roadmap zur Entwicklung eines solchen festgelegt. Demnach soll eine international verpflichtende Regulierung von MASS im Jahr 2028 eingeführt werden (IMO, 2022). Auf nationaler Ebene ergibt eine erste Überprüfung der aktuell veröffentlichten DIN-Normen, dass es zwar vereinzelt Normen, Spezifikationen und Normen im Entwurf (DIN-SPEC) gibt, diese jedoch nicht die Anforderungen an ein KI-System modellagnostisch beschreiben. Nach Sichtung aller im Rahmen der Studie ausgewerteten Dokumente wurde festgestellt, dass existierende DIN-Normen spezielle Einzelfälle standardisieren und keine generalisierte Zertifizierung zum gegenwärtigen Zeitpunkt ermöglichen.

Neben den Normen und Spezifikationen werden zur Zertifizierung geeignete Werkzeuge innerhalb eines Prüfverfahrens benötigt. Eine Untersuchung des EU-Konformitätsbewertungsverfahrens ergibt, dass die zuvor erläuterten Prüfmodule nicht für die Zertifizierung genutzt werden können, da sich die darin genutzten Prüfmethoden und -werkzeuge als ungeeignet herausstellen. Weder in einer Baumusterprüfung (Modul B) noch innerhalb der Produktqualitäts- und Qualitätssicherungsverfahren (Module D, E, F) oder der aktuell definierten Methoden zur Einzelprüfung (Modul G) werden KI-Systeme betrachtet (Europäisches Parlament und Rat der Europäischen Union, 2014). Dies bezieht sich auf alle existierenden Module und entsprechenden Verordnungen. Somit sind auf CI-Modellen (s. Kapitel 2.1) basierende KI-Systeme mit bisherigen Prüfprozessen nicht abbildbar.

Es ergibt sich also ein dringender Handlungsbedarf dedizierte Prozesse zu entwickeln, die das Verhalten von KI-Systemen auf ihre korrekte Funktionsweise prüfen. Zur Eingliederung neuer Prüfprozesse sollte das bestehende Konformitätsbewertungs-verfahren um ein zusätzliches, speziell für KI-Systeme ausgelegtes Modul erweitert werden, um KI-Systeme unter Nutzung bestehender und neuartiger Prozesse prüfen zu können. Eine mögliche Erweiterung des MED-Konformitätsbewertungsverfahren ist in Abbildung 8 dargestellt.

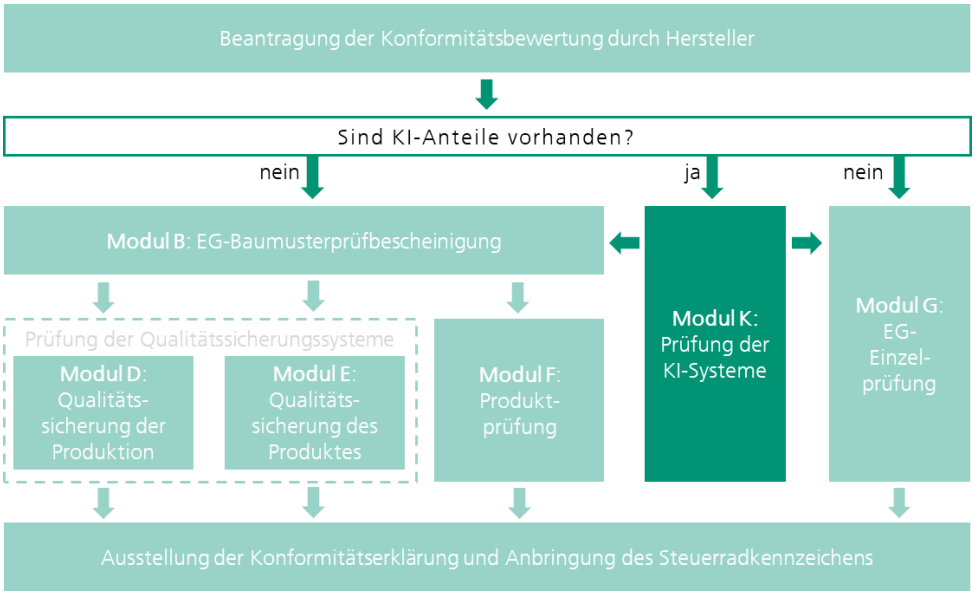

Abbildung 8: Einführung Modul K im Konformitätsbewertungsverfahren nach MED.

Das vorgeschlagene Prüfmodul K schließt die aktuell existierende Regulierungslücke zur Zertifizierung von KI-Systemen spezifisch in der maritimen Industrie. Der empfohlene Inhalt dieses Prüfmoduls wurde im Rahmen dieser Studie erarbeitet und ist im Prüfkonzept in Kapitel 6 festgehalten.

Neben diesem Ansatz, wird aktuell von der Europäischen Union ein generalisierter Vorschlag zur Zertifizierung von KI-Systemen erarbeitet. Um die mögliche Veränderung durch die Ratifizierung dieses Gesetzesvorschlages zu antizipieren, wird dieser im folgenden Unterkapitel näher betrachtet.

3.5. EU-Gesetzesvorschlag über künstliche Intelligenz

Der Vorschlag für ein generalisiertes Verfahren zur Regulierung von KI-Systemen, auch EU AI Act genannt, wurde von der Europäischen Union im April 2021 eingebracht (Europäische Kommission, 2021). In dem vierstufigen risikobasierten Ansatz wird zwischen Systemen mit unannehmbarem, hohem, geringem und minimalem Risiko unterschieden. Die vorgestellten Prozesse dienen nach aktuellem Stand nur als Entwurfsprozess für eine zukünftige Verordnung und spiegeln möglicherweise nicht alle zukünftigen Anforderungen wider. Es ist davon auszugehen, dass bereits ausgeführte Anforderungen im Kern jedoch ähnlich bleiben und für eine erste Skizzierung von Prüfprozessen für teilautonome Überwasserfahrzeuge Berücksichtigung finden sollten.

3.5.1. Risikobasiertes Stufenmodell

Gemäß des Gesetzesvorschlags könnten KI-Systeme mit unannehmbarem Risiko grundsätzlich im Raum der Europäischen Union (EU) verboten werden. Diese Risikokategorie würde Praktiken enthalten, in denen Grundrechte und -werte der EU verletzt werden oder das Potenzial dazu haben manipulativ und ausbeuterisch zu agieren.

Eine Stufe darunter sind Hochrisiko-Systeme anzusiedeln. Diese müssten bei einer Ratifizierung des Gesetzes ein Konformitätsbewertungsverfahren durchlaufen, in dem

die Systeme auf die Einhaltung unterschiedlicher Kriterien untersucht werden. Die vorgeschlagenen Regelungen für Systeme mit hohem Risiko sowie ihre Anwendungsfälle werden im folgenden Unterkapitel ferner erläutert. KI-Systeme mit minimalem oder geringem Risiko würden nur einer minimalen Transparenzpflicht unterliegen (Europäische Kommission, 2021).

3.5.2. Hochrisiko-KI-Systeme

Einer der Kernvorschläge im EU AI Act ist die Reglementierung der Hochrisikosysteme. Betroffen sind KI-Systeme, die entweder als Sicherheitskomponenten von Produkten dienen, die eine Vorab-Konformitätsbewertung benötigen, oder sich auf andere Art und Weise auf die europäischen Grundrechte auswirken. Da die MED bei Schiffsausrüstungen eine Konformitätsbewertung vorschreibt, ist davon auszugehen, dass diese als Hochrisikosystem klassifiziert werden (Europäische Kommission, 2021).

Die im EU AI Act vorgeschlagene Konformitätsbewertungen von Schiffsausrüstungen mit integrierten KI-Systemen würde nach aktuellem Stand anhand von zwei Bewertungen stattfinden: der Bewertung des Qualitätsmanagements und der Bewertung der technischen Dokumentation (Europäische Kommission, 2021).

Anbieter von Hochrisiko-KI-Systemen wären dazu verpflichtet, ein einschlägiges Qualitätsmanagementsystem einzurichten. Dieses müsste im Bewertungsverfahren nachweisen können, dass es die Qualität und Sicherheit des Systems über den gesamten Lebenszyklus gewährleistet. Zur Sicherstellung der Qualität müssten bestimmte Mindestanforderungen an technischen Systemen und Dokumentationen erfüllt sein. Diese beinhalten z.B. Verfahren für das Datenmanagement, die Qualitäts-sicherung, das Risikomanagement oder zur Meldung von Fehlfunktionen. Um die richtige Pflege und Anwendung des Qualitätsmanagementsystems zu prüfen, könnten notifizierte Stellen regelmäßige Audits durchführen (Europäische Kommission, 2021).

Neben dem Qualitätsmanagementsystem müssten Anbieter für eine Konformität in der EU ebenfalls eine vollständige technische Dokumentation nachweisen. In dieser wäre eine detaillierte Beschreibung aller technischen Systeme sowie aller Hardware- und Software-Elementen, die im Laufe des Lebenszyklus genutzt werden oder unterstützend wirken. Außerdem benötigt die technische Dokumentation eine Auflistung aller angewandten harmonisierten Normen und technischen Spezifikationen (Europäische Kommission, 2021).

Im Allgemeinen müsste durch die beiden Bewertungen sichergestellt sein, dass folgende Anforderungen erfüllt werden (Europäische Kommission, 2021):

- Ein einschlägiges Risikomanagementsystem, welches Risiken systematisch ermittelt und analysiert.
- Ein Daten-Governance- und Daten-Verwaltungssystem, das eine hohe und konstante Datenqualität sicherstellt.
- Eine vollständige technische Dokumentation.
- Die verpflichtende Aufzeichnung aller Vorgänge und Ereignisse.
- Die transparente Bereitstellung von Informationen für die Nutzer.
- Eine kontinuierliche, menschliche Aufsicht.
- Ein angemessenes Maß an Genauigkeit, Robustheit und Cybersicherheit.

Da sich der EU AI Act erst in einem Entwurfszustand befindet, ist es möglich, dass einzelne Elemente über den Gesetzgebungsprozess verändert werden. Um auf die umfangreichen Auswirkungen dieses Gesetzes auf die KI-Zertifizierung vorbereitet zu sein, wird in der Studie davon ausgegangen, dass eine Reglementierung innerhalb der nächsten Jahre ratifiziert wird.

4. Marktanalyse von KI-Systemen im maritimen Kontext

Wie bereits in den vorigen Kapiteln aufgezeigt, gibt es aktuell kein geeignetes Verfahren zur Prüfung und Zertifizierung von KI-Systemen. Nach aktuellem Stand obliegt daher die Sicherheit der KI-Systeme in der Hand der Unternehmen und internen, nicht öffentlich einsehbaren, Prozessen. Diese existierende Regulierungslücke gewinnt durch das Fortschreiten der Digitalisierung und Autonomisierung der Schifffahrt an Bedeutung. Durch die zunehmende Anzahl an Technologien, die auf den Markt gebracht werden, wächst der Druck auf Prüfstellen und beauftragende Behörden. Die Zunahme der Autonomisierung kann anhand des Verlaufs der internationalen Patentanmeldungen im MASS-Sektor aufgezeigt werden. Abbildung 9 zeigt das zunehmende Wachstum der Patentanmeldungen, die 1990 bis 2021 jährlich eingereicht wurden. Hierin werden Patentanmeldung gezählt, die sich unter der gleichzeitigen Verwendung der beiden Suchworte „autonomous“ und „ship“ auffinden lassen. Der Verlauf der Kurve kann eine Indikation dafür sein, dass sich die Anzahl der jährlich auf den Markt hinzukommenden KI-unterstützten Produkte weiter erhöhen wird.

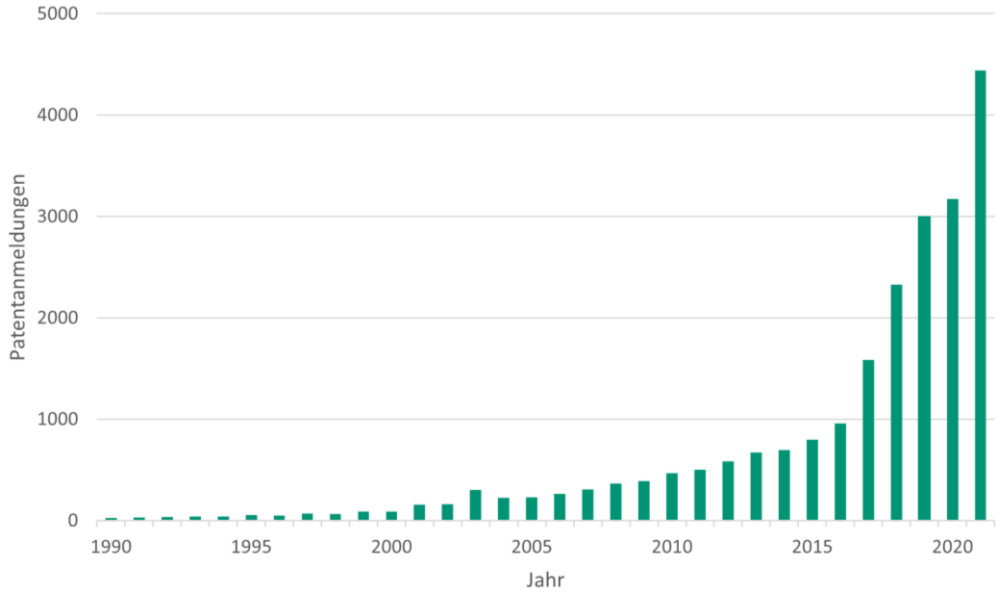

Abbildung 9: Patentanmeldungen mit MASS-Bezug für 1990 bis 2021. Auszug aus „Google Patents“ (Alphabet Inc., o. D.).

Bezugnehmend auf die in Kapitel 3.4 ermittelte Unzulänglichkeit bestehender Standardisierungen und Zulassungsverfahren sowie fehlender nationaler und internationaler Prozesse und Standards ergibt sich die Notwendigkeit, entsprechende Prüf- und Zertifizierungsprozesse zu etablieren. Insbesondere mit der Annahme, dass eine solche Prüfung die Sicherheitsprotokolle und -mechanismen bei der Implementierung von KI-Algorithmen gewährleistet, wächst der Druck kurzfristig nutzbare Konzepte zur Prüfung von KI-Systemen zu entwickeln.

Um ein Verständnis dafür zu erlangen, welche Produkte eine Zertifizierung benötigen und wie eine solche aussehen kann, werden im Folgenden existierende Produkte für MASS mit KI-Unterstützung identifiziert und nach verschiedenen Kriterien analysiert.

Die Auswahl der Kriterien resultiert aus der Bandbreite an Sensoren, welche im Kontext von MASS-bezogenen Entwicklungen und Veröffentlichungen aufgeführt wird und für eine erfolgreiche Umsetzung notwendig erscheint. Die Marktanalyse wurde mit Fokus auf teilautonome Überwassersysteme durchgeführt, mit dem Ziel die Informationsbedarfe entsprechender Systeme zu identifizieren und eine Kategorisierung durchzuführen. Ermittelte Informationsbedarfe dienen zur Recherche bestehender Standardisierungen der genutzten Daten mit dem Ziel diese im Rahmen des Prüf- und Sicherheitskonzeptes zu vereinheitlichen. Das Kapitel schließt mit der Einführung eines fiktiven Anwendungsbeispiels ab, welches zur Erklärung des Prüf- und Sicherheitskonzeptes verwendet wird.

4.1. Sichtung und Analyse von KI-gestützten Produkten

Die folgende Marktanalyse untersucht verschiedene aktuell auf dem Markt befindliche, KI-unterstützte Produkte anhand von zwei Leitfragen:

- Welche **Datenquellen** nutzen die Produkte?
- Welchen **Anwendungsfällen** dienen die Produkte?

Im Rahmen der Marktanalyse wurden 18 Produkte von 16 verschiedenen Unternehmen gesichtet. Von diesen zur Autonomisierung genutzten Systeme werden 17 auf den Schiffen selbst und eines an Land installiert. Abhängig von ihrer Zielsetzung und Anwendung unterscheiden sich die gesichteten Produkte durch ihren Grad der autonomen Komponenten. Es befinden sich daher sowohl einfache, kamerabasierte Anlegeunterstützungen als auch ganzheitlich autonome Containerschiffe unter den gesichteten Produkten. Hergestellt werden diese von weltweit verteilten kleinen, mittleren oder Großunternehmen. Die gesammelten Informationen basieren auf den Produkterläuterungen der Unternehmen, welche auf die genannten Leitfragen untersucht worden sind.

Welche Datenquellen von den Produkten genutzt werden und welche Anwendungsfälle sie bedienen wird nachfolgend erläutert. Eine detaillierte Übersicht der untersuchten Systeme ist in Anhang A.1. in Tabelle 5 und Tabelle 6 hinterlegt.

4.1.1. Kategorisierung von Datenquellen

Die identifizierten Datenquellen der untersuchten Systeme lassen sich anhand der genutzten Sensorik einordnen. Die Kategorisierung in Abhängigkeit der Sensorik dient der Fokussierung auf die Art und Form der Daten, welche durch das System erfasst und verarbeitet werden. Nachfolgend fasst Tabelle 3 alle identifizierten Sensoriken und, sofern vorhanden, Referenzen zu ihren Standardisierungen zusammen. Um die Nutzbarkeit für die Entwicklung maritimer KI-Systeme einschätzen zu können wird die Betrachtung mit Fokus auf eine vorliegende Standardisierung seitens der IMO durchgeführt. Diese fokussierte Betrachtung erlaubt eine Einschätzung, inwiefern vorliegende Sensoren bereits grundlegend für maritime Einsatzzwecke nutzbar sind und ob eine Zulassung in KI-Systemen nur noch mit Blick auf die KI-Komponenten durchgeführt werden muss.

Tabelle 3: In Marktanalyse identifizierte Sensoriken für Datenquellen aus KI-Systemen und ihre Kommunikationsstandardisierungen. Die Bedeutungen der Abkürzungen der Sensoriken sind im Abkürzungsverzeichnis aufgeführt.

Sensorik für Datenquelle	Leistungsanforderungen durch IMO
AIS	(IMO, 2015; ITU, 2014)
GNSS	(IMO, 1995, 2001)
IMU / MRU	(IMO, 2017)
Infrarot- und RGB[1]-Kamerasysteme	nein
LIDAR	nein
RADAR	(IMO, 2004)
Tiefenmessgeräte	(IMO, 1971, 1998)
Wettersensoren	nein

Von Sensoren mit vorhandenen Kommunikationsstandards werden insbesondere RADAR und AIS von den Herstellern häufig als Bestandteil der Produkte genannt. RGB-Kamerasysteme werden in allen gesichteten Produkten verwendet, wobei diese keiner Standardisierung folgen. Insbesondere die Nutzung von Kamerasystemen und daraus resultierenden bildgestützten KI-Systemen stellt durch eine fehlende Standardisierung im maritimen Kontext Hürden dar. KI-Systeme, die auf solche Sensoriken setzen, müssen im Vergleich zu Sensoriken mit standardisiertem Informationsaustausch mit zusätzlichem Aufwand geprüft und zertifiziert werden. Der Hintergrund ist der, dass die Standardisierung eines Informationsaustausches bereits Aufschlüsse über die zu erwarteten Eingangs- und Ausgangsdaten geben kann.

4.1.2. Analyse der Anwendungsfälle

Die Auswertung der Anwendungsfälle hat in einem ersten Schritt eine Kategorisierung in vier Gruppen hervorgebracht in welchen KI-Systeme eingeordnet werden können: die Identifikation von Objekten, die Verhaltensvorhersagen von Verkehrsteilnehmenden, die Routenplanung und die Lagebilderstellung. Die Kategorisierung hilft Produktgruppen hinsichtlich ihrer Datengrundlage einordnen zu können und entsprechende Informationsbedarfe zu extrahieren.

Die Objekterkennung mittels kamerabasierter Daten ist eine Anwendung, die in ihren unterschiedlichen Ausprägungen in allen gesichteten Produkten angewandt wird. Dabei können die Objekte Hindernisse, andere Schiffe oder die Küste sein. Diese Objekte werden teilweise unter Verwendung von AIS-Daten identifiziert und zur Entscheidungsunterstützung ausgegeben. Außerdem kann die Objektidentifikation zusätzlich als System zur An- oder Ablege-Unterstützung dienen, indem der Schiffsführung gesammelte, relevante Zusatzinformationen bereitgestellt werden.

In einigen Produkten dienen die gesammelten Daten und die Identifizierung der erkannten Objekte zur Verhaltensvorhersage anderer Schiffe. Diese Vorhersage dient der Kollisionsvermeidung und der Evaluation der COLREGs.

[1] Als RGB-Kamerasysteme (RGB kurz für Rot, Grün und Blau) werden in dieser Studie Kamerasysteme bezeichnet, die das sichtbare Farbspektrum bedienen.

In der Routenplanung werden die Position und Geschwindigkeit des eigenen Schiffes, anderer Schiffe sowie Wetterdaten genutzt um die Routenplanung, mit Blick auf einen Zeit- oder Verbrauchsrahmen, zu optimieren.

Im Rahmen der Lagebilderstellung werden Daten des Schiffes gesammelt, um diese näher zu untersuchen. Dazu werden andere Verkehrsteilnehmer oder lokale Wetterphänomene mittels einer großen Bandbreite an Sensoren erkannt, kontextualisiert und für ein gemeinsames Verständnis fusioniert.

Wichtig im Kontext dieser Anwendungsfälle ist eine harmonisierte Beschreibung der Ergebnisdatensätze der einzelnen Anwendungsfälle, wie beispielsweise ein Datensatz der *Maritime Perceived Environment* (dt.: Maritime Umgebungswahrnehmung) als Ergebnis einer Objekterkennung. Erst durch eine solche Harmonisierung kann eine Modularisierung eines KI-Systems, aber auch Prüfbarkeit der Systeme realisiert werden (Burmeister et al., 2020). Darüber hinaus würde eine allgemeingültige Harmonisierung die Integration mehrerer KI-Systeme und die Prüfung solcher integrierten Systeme vereinfachen.

4.2. Zusammenfassung der Marktanalyse und abgeleiteter Handlungsbedarf

Mit Blick auf den Entwicklungsstand zeigen sich deutliche Unterschiede in Umfang, Beschreibung und Nutzung von KI-Anteilen für die Funktion der Produkte. Die Marktanalyse hilft im Kontext der Studie grobe Informationsbedarfe abzuschätzen und entsprechende Empfehlungen zu geben. Die einfachen Produktbeschreibungen erlauben hier jedoch nur eine erste Abschätzung, inwiefern die gesichteten Produkte Gebrauch von KI-Technologien machen.

Aus den jährlich zunehmenden Patentanmeldungen innerhalb der letzten zwei Jahrzehnte lässt sich ableiten, dass die Entwicklung eines Prüfkonzeptes zum aktuellen Zeitpunkt von hoher Bedeutung ist und für viele Hersteller der anstehende Schritt zur Marktreife abbildet. Insbesondere die Vielfalt an KI-Systemen, die im Rahmen der Marktanalyse identifiziert wurden, deutet auf eine große Menge verschiedener zu prüfender Technologien hin.

Um die hohe Anzahl verschiedenartiger Systeme abbilden zu können, wird empfohlen Prüfprozesse modellagnostisch auszulegen. Daraus folgt, dass eine generische Betrachtung der KI-Systeme und ihrer Anwendungen und Datenquellen von hoher Bedeutung ist. Weiterführend ist nicht vorhersehbar, welche zukünftigen KI-Systeme in MASS Anwendung finden werden. Diese noch unbekannten Technologien könnten ebenfalls durch einen modellagnostischen Ansatz abgedeckt werden. Hierfür wird in den folgenden Kapiteln eine Prüfungsprozedur mit einem evidenzbasierten Ansatz vorgeschlagen, welche sich auf die Verarbeitung von Eingabedaten und den resultierenden Ausgabedaten des KI-Systems fokussiert. Damit soll in erster Linie die Frage beantwortet werden können, „ob“ das KI-System funktioniert und nicht „wie“.

4.3. Fiktives KI-System als Anwendungsbeispiel

Im Folgenden wird ein fiktives Anwendungsbeispiel eingeführt, anhand welchem das vorgeschlagene Prüf- und Sicherheitskonzept erläutert wird. Bei dem Beispiel handelt es sich um einen bildgestützten Peilungssensor, welcher mithilfe des nachfolgend aufgeführten Produktdatenblatts beschrieben wird. Die Erläuterungen anhand des

Beispiels sind in türkisen Kästen eingerahmt und enthalten unterstützende Erläuterungen, wie sich Ergebnisse der Studie auf die Prüfung eines KI-Systems übertragen lassen.

Produktbeschreibung:
Der bildgestützte Peilungssensor wertet ein Kamerabild mit Fokus auf die Identifikation und Bestimmung des Winkels von Schiffen in der Umgebung aus. Peilungen werden relativ zur Schiffsmitte grafisch auf einem Bildschirm über den Kameradaten angezeigt und geben eine Indikation an, in welchem Winkel sich dieses Schiff nach Auswertung der Kameradaten befindet.

Leistungsmerkmale Gesamtsystem:
Erkennung von 95% aller Schiffe bei Erfüllung aller Betriebsvoraussetzungen. Maximaler Anteil in der fehlerhaften Detektion von nicht Schiffsobjekten von weniger als 5%.

Leistungsmerkmale zur Kamera:
- Farbgebung: RGB
- Sichtfeld: 122°
- Bildausrichtung zur Schiffsmitte: 000°
- Brennweite: 10 mm
- Bildauflösung: 2592 px × 1944 px
- Bildrate: 15 Hz

Betriebsvoraussetzungen:
- Helligkeit: Tageslicht
- Maximaler Rollwinkel: 5°
- Maximaler Pitchwinkel: 3°
- Wetterbedingungen: klare Sicht (kein Niederschlag, Nebel, Gischt, Staub)
- Erkennbare Schiffstypen: Frachtschiffe

Installation:
- 1 Bildschirm auf der Brücke.
- 1 Kamera auf dem Mast des Vorschiffes ohne sichtbare Verdeckung der Linse.

Eingabe:
- Videostream der installierten Kamera.

Ausgabe:
- Kamera-Stream mit Overlay.
- Peilungen der Schiffe als Zahl im Overlay über den Schiffen augmentiert.
- Proprietärer Datenstrom nach dem Standard NMEA 0183 (DIN, 2011), ähnlich Radar Target Message.

5. Integration von Prüf- und Sicherheitskonzept

In der Studie wird ein Prüfkonzept (s. Kapitel 6) für das BSH und ein Sicherheitskonzept (s. Kapitel 7) für die Hersteller vorgeschlagen. Beide Konzepte sind integrativ miteinander verknüpft und können nicht ohne jeweils das andere betrachtet werden. Durch diese Integration sind im Sicherheitskonzept bereits Prozesse verankert, die ein KI-System auf eine möglichst erfolgreiche Prüfung nach dem Prüfkonzept vorbereiten. Auf diese Weise kann die Entwicklung und Prüfung des KI-Systems zielführend und ressourcenschonend verfolgt worden.

Das Sicherheitskonzept richtet sich an den Hersteller eines KI-Systems und hat folgende zwei Ziele:

- Die Gewährleistung der Prüfbarkeit des Systems.
- Die Entwicklung eines hinreichend sicheren Systems mit Aussicht auf erfolgreiche Prüfung.

Es wird vorausgesetzt, dass der Hersteller die Inhalte des Sicherheitskonzeptes nicht unmittelbar vor der Prüfung, sondern spätestens während der Entwicklung des KI-Systems zur Kenntnis nimmt. Damit können frühzeitig Grundsteine für ein prüfbares und hinreichend sicheres KI-System gelegt werden.

Das Prüfkonzept ist an das BSH gerichtet und es werden im Hinblick auf das betrachtete KI-System folgende Ziele verfolgt:

- Die Prüfung der ordnungsgemäßen informations- und sicherheits-technischen Funktion.
- Zuverlässige Zulassung von KI-Systemen.

Die Zusammenarbeit zwischen Hersteller und Prüfer beim Prüfprozess äußert sich in erster Linie an einer prüfungsvorbereitenden Kommunikation vom Hersteller zum Prüfer. Die prüfungsvorbereitende Kommunikation ist das Bindeglied zwischen Sicherheitskonzept und Prüfkonzept. Dieser Zusammenhang ist in Abbildung 10 vereinfacht dargestellt.

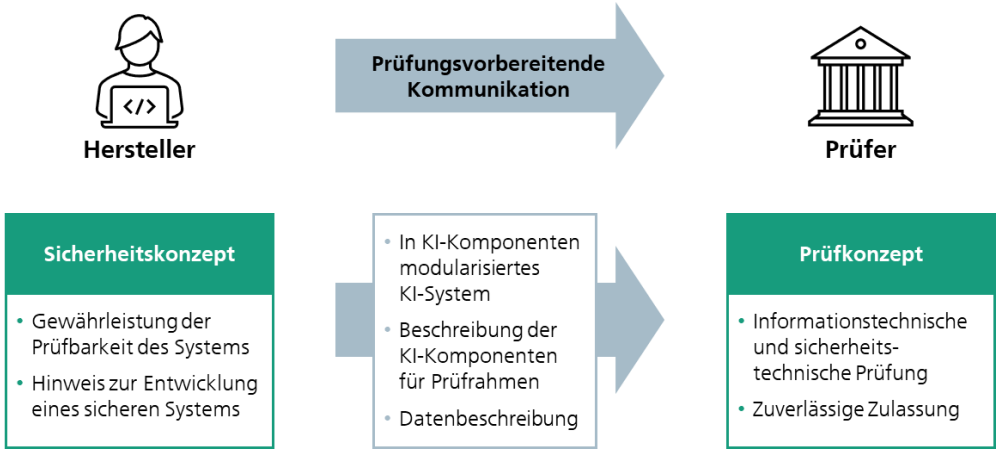

Abbildung 10: Prüfungsvorbereitende Kommunikation von Hersteller zu Prüfer.

Der Inhalt dieser prüfungsvorbereitenden Kommunikation geht für den Hersteller aus dem Sicherheitskonzept hervor. Im Wesentlichen besteht die prüfungsvorbereitende Kommunikation aus den folgenden Bestandteilen:

- Das in KI-Komponenten[1] modularisierte KI-System, welches geprüft wird (s. Kapitel 7.1 Unterabschnitt F1).
- Eine spezifizierte Beschreibung der KI-Komponenten, um den Prüfrahmen zu definieren (s. Kapitel 7.1 Unterabschnitt F2).
- Eine Datenbeschreibung, zur Beschaffung von geeigneten Testdaten (s. Kapitel 7.3 Unterabschnitt D2).

Die prüfungsvorbereitende Kommunikation ist zur Durchführung einer Prüfung und Zulassung durch das BSH absolut notwendig.

In dieser Studie werden nur solche KI-Systeme betrachtet, die über eingefrorene KI-Modelle verfügen. Der Grund hierfür ist der, dass sich das Verhalten von nicht eingefrorenen Modellen nach der Prüfung und Zertifizierung ändern kann und somit eine Prüfung und Zertifizierung sich als bedeutungslos erweisen würde.

Im Folgenden wird die Verwendung des Sicherheits- und Prüfkonzeptes und die Kommunikation zwischen Hersteller und Prüfer einleitend grob erläutert. Tiefergehende Erläuterungen finden sich am Ende der Kapitel zum Sicherheitskonzept (Kapitel 7) und Prüfkonzept (Kapitel 6).

Der Hersteller beabsichtigt die Entwicklung eines Produktes, welches ein KI-System beinhaltet. Bei dem Produkt handelt es sich um einen bildgestützten Peilungssensor. Durch die Berücksichtigung des Sicherheitskonzeptes wird der Hersteller an zwei Stellen unterstützt.

Verwendung des Sicherheitskonzeptes zur Entwicklung eines sicheren KI-Systems

Erstens werden Empfehlungen ausgesprochen, welche bei der Entwicklung des KI-Systems unterstützend wirken. Bei den Empfehlungen handelt es sich um bewährte Entwicklungspraktiken, die ein KI-System möglichst sicher gestalten. Im Sicherheitskonzept wird hierbei der Fokus auf die Datenqualität gesetzt. Im Beispiel des bildgestützten Peilungssensors werden im Wesentlichen RGB-Bilddaten und AIS-Daten verwendet. Entsprechend beziehen sich die Datenqualitätshinweise auf diese Datensätze. Die Befolgung dieser Empfehlung erhöht die Chancen, dass das KI-System mit Erfolg geprüft und zertifiziert wird.

Kommunikation an den Prüfer

Zweitens werden notwendige, prüfungsvorbereitende Maßnahmen, die durch den Hersteller sicherzustellen sind, erläutert. Die Befolgung dieser Maßnahmen, gewährleistet, dass das KI-System des Herstellers geprüft werden kann. Die erste Maßnahme ist die Modularisierung des KI-Systems. Der Hersteller würde dieses in prüfbare KI-Komponenten zerlegen. Eine KI-Komponente könnte in dem Anwendungsbeispiel die Schiffserkennung in Bilddaten sein. Als zweite Maßnahme würde der Hersteller alle KI-Komponenten hinsichtlich ihres Anwendungsbereiches

[1] Bei KI-Komponenten handelt es sich um individuell prüfbare Untereinheiten des zu prüfenden KI-Systems (s. Kapitel 7.1 Unterabschnitt F1).

beschreiben, also u.a. unter welchen Bedingungen die Schiffserkennung funktionieren muss. Als letzte Maßnahme müsste der Hersteller die Datensätze beschreiben, mit welchen die KI-Komponenten trainiert worden sind. Im Anwendungsbeispiel würden der RGB-Bilddatensatz und der AIS-Datensatz beschrieben werden müssen.

Verwendung des Prüfkonzeptes zur zuverlässigen Prüfung des KI-Systems

Dem Prüfer wird ein Prüfprozess vorgeschlagen, welcher eine zuverlässige sicherheits- und informationstechnische Prüfung ermöglicht. Um die Prüfung des bildgestützten Peilungssensor durchzuführen, muss der Hersteller die prüfungsvorbereitende Kommunikation befolgen.

Diese besteht, erstens, darin, dass das KI-System modularisiert in KI-Komponenten an den Prüfer übergeben wird. Der Prüfer ist dann im Stande die KI-Komponenten, z.B. die Schiffserkennung, einzeln zu prüfen.

Zweitens erhält der Prüfer von dem Hersteller Beschreibungen zu den KI-Komponenten, aus denen der Prüfrahmen folgt. Der Prüfrahmen umfasst den Anwendungsbereich der KI-Komponenten und messbare Kriterien zur Bewertung einer erfolgreichen Prüfung.

Drittens erhält der Prüfer eine Beschreibung zu den verwendeten Datensätzen. Daraus soll für den Prüfer hervorgehen, mit welchen Daten er die KI-Komponenten testen kann.

6. Prüfkonzept

Zielstellung dieses Prüfkonzeptes ist es mittels einer Sequenz von Prüfschritten KI-Systeme hinsichtlich ihrer ordnungsgemäßen informations- und sicherheitstechnischen Funktion modellagnostisch zu prüfen und zuverlässig zulassen zu können. Der Fokus liegt dabei festzustellen „ob“ und nicht „wie“ ein KI-System funktioniert.

Die sequenzielle Struktur des Prüfkonzepts ist in Abbildung 11 dargestellt und zeigt die drei Prüfabschnitte der Prüfung (Vorprüfung, Hauptprüfung und Nachprüfung) und die untergegliederten Prüfschritte. Ein Prüfschritt repräsentiert Aufgaben, die im Rahmen des Prüfkonzeptes seitens des Prüfers durchgeführt werden müssen. Die sequenzielle Struktur wurde im Rahmen der Studie gewählt, um einzelne Abschnitte des Prüfkonzeptes fachlich zu trennen und dem Prüfer einen Leitfaden zu geben, welcher sich leicht befolgen lässt und die Struktur des Prüfprozesses inhärent vorgibt.

Abbildung 11: Prüfkonzept mit Prüfabschnitten und untergliederten Prüfschritten.

Die Prüfabschnitte (Vor-, Haupt- und Nachprüfung) bilden einen thematischen Überbau, um bestimmte Aspekte der Prüfung von KI-Systemen zu bündeln und in dedizierte Prüfschritte zu kombinieren. Ein Prüfschritt ist die kleinste Einheit der Prüfung und bildet eine dedizierte fachliche Aufgabe ab, die im Rahmen der Prüfung und Zulassung eines KI-Systems erfolgen muss. Prüfschritte werden dabei nacheinander durchgeführt und müssen mit Blick auf ihre Aufgabe und Zweck jeweils „bestanden“ werden. Kann ein Prüfschritt aufgrund des KI-Systems und etwaiger Ergebnisse der Untersuchung innerhalb eines Prüfschrittes nicht vollständig durchgeführt werden, gilt die Prüfung als fehlgeschlagen und eine Nacharbeit seitens des Herstellers ist notwendig.

6.1. Vorprüfung

Die Vorprüfung beinhaltet Prüfschritte zur Abschätzung der Notwendigkeit dieser spezifischen Prüfung von KI-Systemen. Mithilfe einer expliziten Differenzierung sollen nur Systeme geprüft werden, die unter die Definition eines KI-Systems fallen. Weiterführend wird im Rahmen der Vorprüfung die grundsätzliche Prüffähigkeit des KI-Systems festgestellt. Die Vorbereitung der Prüffähigkeit ist notwendig, damit der modellagnostische Prüfprozess auf das KI-System angewendet werden kann.

V1 | Erfüllung der Definition einer KI

Durch die Einführung des Prüfmoduls K (s. Abbildung 8) innerhalb des Prüf- und Zertifizierungswesens wurde die getrennte Prüfung der KI-Anteile eines KI-basierten Systems vorgeschlagen. Um bestehende Prozesse effektiv nutzen zu können, muss zu Beginn der Prüfung eines KI-basierten Systems geprüft werden, ob das System der Definition eines KI-Systems im Rahmen der Studie entspricht. Kapitel 2.1 gibt einen Überblick in die Bandbreite dieser Technologien und entsprechender Definitionen. KI-Systeme, die Gebrauch von den eingeführten Technologien machen, erfüllen somit die Definition einer KI und werden in nachfolgend eingeführten Prozessen betrachtet. Es sei zu beachten, dass sich die Definition eines KI-Systems mit der Zeit aufgrund von fortschreitenden Entwicklungen verändern kann.

Der Funktionsumfang im Anwendungsbeispiel bildgestützter Peilungssensor lässt auf Grundlage der beschriebenen Funktionalität auf ein System schließen, das im Sinne dieser Studie als KI-System bezeichnet werden kann. In Anbetracht aktueller Entwicklungen ist anzunehmen, dass eine Erkennung auf Basis eines CI-Ansatzes (s. Kapitel 2.1) erfolgt, um generisch Formen zu erkennen, die der eines Schiffes entsprechen. Vergleichbare Ansätze nutzen Modelle, welche durch eine große Anzahl von Schiffsbildern auf die Erkennung dieser trainiert wurden. Das Wissen des Systems besteht somit nicht in der symbolischen Beschreibung von Schiffen, sondern in mathematischen Bildauswertungen, um bekannte Muster zu erkennen und diese bekannten Klassen von Objekten zuordnen zu können.

V2 | Modularisierung in KI-basierte Systemkomponenten

Bedingt durch die Vielfalt von Architekturen von KI-Systemen und der möglichen Komplexität dieser Systeme ist eine ganzheitliche Prüfung nur in vereinzelten Fällen möglich. Ebenfalls ist es in der Praxis nicht umsetzbar KI-System nach architektur-spezifischen Prozessen zu prüfen. Eine Modularisierung der KI-Systeme in dedizierte Komponenten, welche spezifische Aufgaben lösen, kann die Prüfung überhaupt erst ermöglichen. Denn dadurch reduziert sich die Komplexität in dem Maße, dass die Beschreibungen der Funktionsweise und prüfungsrelevante Prozesse komponentenweise durchgeführt werden kann.

Ein System gilt im Rahmen der Studie als modularisiert, wenn jeder Menge an Eingaben die korrespondierende Menge an Ausgaben eindeutig zugeordnet werden kann. Die Zuordnung von Ein- und Ausgabewerten bildet die Grundlage für die evidenzbasierte Prüfung von KI-Systemen. Abbildung 12 zeigt schematisch, wie mithilfe einer eindeutigen Zuordnung von Ein- und Ausgabedaten eine Modularisierung erfolgen kann. Die Kürzel E1 bis E7 sowie A1 bis A7 stehen hierbei beispielhaft für Ein- bzw. Ausgabewerte und K1 bis K3 für die Systemkomponenten. Die Modularisierung erfolgt dabei in zwei Schritten. Im ersten Schritt werden korrespondierende Eingabe- und

Ausgabedatenströme des KI-Systems eingeteilt und im zweiten Schritt wird das KI-System aufbauend auf der Datenstromeinteilung in KI-Komponenten modularisiert.

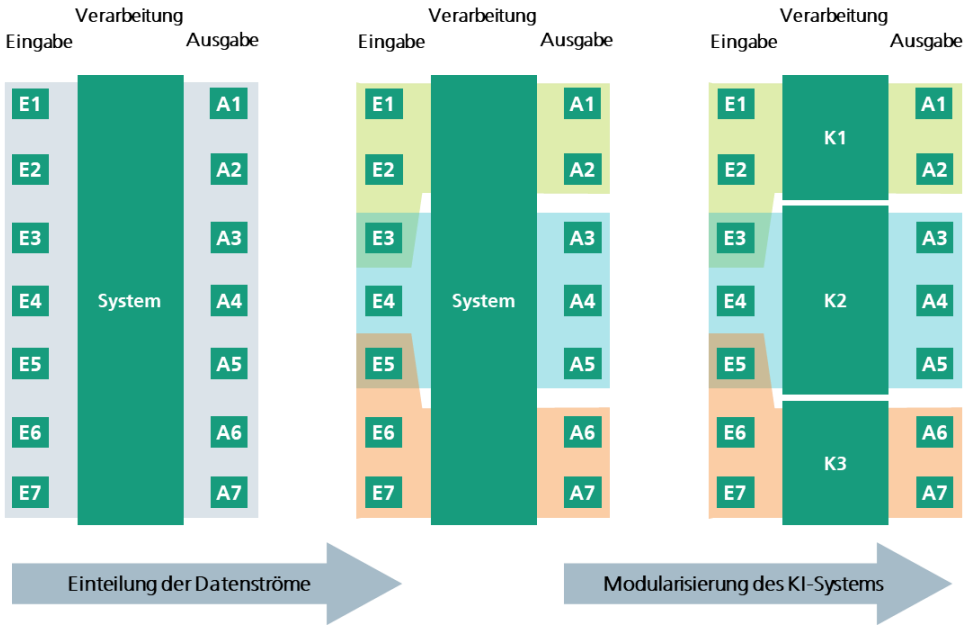

Abbildung 12: Modularisierungsprozess von KI-Systemen. E steht jeweils für Eingabe, A für Ausgabe und K für Systemkomponente.

Der gesamte Prüfprozess bedient sich zur Durchführung der evidenzbasierten Prüfung dem Eingabe-, Verarbeitung- und Ausgabe-Prinzip (EVA-Prinzip). Evidenzbasiert bedeutet hier, dass das System auf Grundlage der Beobachtung und Bewertung seines (reproduzierbaren) Verhaltens untersucht wird. Dem Prüfer sind nur die Ein- und Ausgabedaten bekannt. Die verarbeitende Einheit (KI-Komponente) wird im Rahmen der Studie modellagnostisch als Blackbox angesehen und eine Einsicht ist nicht angestrebt. Die Betrachtung als Blackbox liegt sowohl in dem Interesse der Hersteller, neuartige Technologien unter Verschluss zu halten, als auch im Interesse der Prüfstelle, die hohe Anzahl von bestehenden und potenziellen Architekturen nicht individuell, sondern einheitlich, also modellagnostisch, zu betrachten. Eine Betrachtung als Whitebox würde die Prüfung und Zertifizierung verkomplizieren und ohnehin nur bei sAI-Modellen realisierbar sein. Bei CI-Modellen ist die Verarbeitung nicht ohne Weiteres nachvollziehbar. Durch den Fokus auf eine Prüfung eines KI-Systems nach dem EVA-Prinzip bleibt diese hingegen realisierbar und skalierbar. Im weiteren Verlauf wird die zu prüfende Einheit als KI-Komponente bezeichnet.

Die Modularisierung des bildgestützten Peilungssensors hängt vom Grad der Integration der einzelnen mathematischen Prozesse ab (Erkennung und Peilungsbestimmung).

Zur Veranschaulichung der Modularisierung wird davon ausgegangen, dass die Schiffserkennung und die Peilungsschätzung zwei getrennte Komponenten sind und einzeln geprüft werden können. Der Hersteller stellt hierfür in einer ausführbaren Umgebung beide Komponenten mit ihren Eingabe- und Ausgabeschnittstellen zur Verfügung.

Die folgende Skizze (Abbildung 13) stellt das modularisierte KI-System in Anlehnung an Abbildung 12 dar.

Abbildung 13: Mögliche Modularisierung des Anwendungsbeispiels.

Die Schiffserkennungskomponente spiegelt die Technologie wider, die für die Erkennung von Schiffen in einem fortlaufenden Datenstrom von Bildern mit einer Auflösung von 2592 px × 1944 px und einer Bildrate von 15 Hz verantwortlich ist. Ausgabe ist das RGB-Kamerabild sowie eine annotierte Liste an Begrenzungsrahmen, welche durch die Komponente innerhalb des Bildes erkannt wurden.

Die Peilungskomponente erhält die Kamerabilder und eine Liste an Begrenzungsrahmen, welche in eine Liste mit Peilungen mit Winkelangaben in Grad überführt werden und gibt ferner eine Komposition der Bilder und Peilungsangaben über eine grafische Schnittstelle aus. Weiterhin erzeugt die Komponente einen NMEA-konformen Datenstrom, um die Ergebnisse des Sensors für andere Systeme zugänglich zu machen.

Es ist grundlegend vom Prüfer sicherzustellen, ob die Komponenten für die nachfolgenden Prüfschritte nutzbar sind und eine Auswertung von Ergebnissen durch die Eingabe von Daten gemäß der Datenbeschreibung möglich ist.

V3 | Formalisierung der Anwendungsdomäne

Um die Funktionsweise einer KI-Komponente abgrenzen zu können und die Rahmenbedingungen für die Hauptprüfung zu legen, muss der Hersteller eindeutig spezifizieren, unter welchen Rahmenbedingungen die KI-Komponente funktionieren soll. Diese als Anwendungsdomäne definierte Abgrenzung gibt Größen vor, durch die sich das Verhalten der KI-Komponente eingrenzen lässt. Als Verhaltensgrenze wird der Wertebereich definiert, innerhalb welchem die Komponente vom Hersteller definierte Eingaben erhält und korrespondierende Ausgaben erzeugt. Entsprechend kann die Komponente innerhalb dieses Wertebereiches Anwendung finden. Außerhalb dieses Wertebereiches hat sie keine Funktion und sollte keine Ausgabe erzeugen, um unvorhersehbare Entscheidungen oder Handlungen zu vermeiden.

Jede KI-Komponente muss innerhalb einer definierten Anwendungsdomäne liegen, welche sowohl die Kapazitäten der KI-Komponente abbildet als auch im Rahmen der Anwendung der KI-Komponente sinnvolle Grenzwerte bildet. Die Anwendungsdomäne liefert somit die Grundlage für die Datenbeschreibung des Herstellers, die für die Hauptprüfung notwendigen Größen beinhaltet, um Daten zur Prüfung der KI-Komponente beschaffen zu können.

Die Anwendungsdomäne kann mithilfe anerkannter Methodiken formalisiert werden und sollte sich an allgemein bekannte Formen ausrichten. Das Werkzeug der Operational Design Domain stammt aus der Automobilbranche und dient zur Beschreibung des möglichen Einsatzbereiches von (teil-)autonomen Fahrzeugen (Gyllenhammar et al., 2020). Für den maritimen Bereich, perspektivisch auch MASS, wird das Konzept Operational Envelope vorgeschlagen. Dieses baut auf der Operational Design Domain auf. Zudem berücksichtigt es die Verantwortlichkeiten und Schnittstellen sowohl innerhalb als auch zwischen den Bereichen der Autonomisierung und den menschlichen Operateuren (Rødseth et al., 2022). Unabhängig von der gewählten Methode zur Formalisierung sollte die Anwendungsdomäne die folgenden Kriterien erfüllen, um die Prüffähigkeit einer KI-Komponente zu gewährleisten.

- Bei einer modularisierten KI-Komponente muss die Anwendungsdomäne für jede KI-Komponente definiert werden. Dies bedeutet das bei einer erforderlichen Modularisierung jede KI-Komponente mit einer eigenen Anwendungsdomäne zur Verfügung gestellt werden muss.
- Die Anwendungsdomäne beschreibt und grenzt alle im Rahmen der KI-Komponente genutzten Eingabewerte eindeutig ein und definiert feste Wertebereiche.
- Für jeden Eingabewert muss klar definiert sein, ob und inwiefern dieser von der KI-Komponente genutzt wird und welche Eingabewerte bei der Entwicklung der Komponente angenommen wurden. Im Umkehrschluss muss fest definiert sein bei welchen Werten keine Verarbeitung erfolgen darf und wie die Komponente darauf reagiert.
- Die Ausgabe der KI-Komponente sollte analog zu den Eingabewerten beschrieben werden. Hier ist ein Ansatz nach dem *Maritime Perceived Environment* (Burmeister et al., 2020) denkbar.

Für den Aufbau eines Frameworks zur Formalisierung und Prüfung von Anwendungsdomänen ist der Aufbau einer zentralen Datenbank für bereits geprüfte Anwendungsdomänen empfehlenswert. Ein durch die Association for Standardization of Automation and Measuring Systems eingeführte offene und an das Operational Design Domain angelehnte Methode (ASAM, 2021) zur Beschreibung und Generalisierung von Anwendungsdomänen ist auch für die Prüfung von KI-Systemen im maritimen Sektor adaptierbar. Die zentrale Sammlung von geprüften Anwendungsdomänen ermöglicht es dem Prüfer Gemeinsamkeiten zu identifizieren und als Grundlage für zukünftige Prüfungen zu berücksichtigen. Der offene Zugang zu

einer solchen Datenbank kann durch Hersteller von KI-Systemen als Referenz zur Formalisierung der Anwendungsdomäne genutzt werden.

Durch die Aufteilung des bildgestützten Peilungssensors in zwei Komponenten müssen ebenfalls zwei Anwendungsdomänen definiert werden.

Die Anwendungsdomäne der Schiffserkennungskomponente zeichnet sich hauptsächlich durch die technischen Betriebsvoraussetzungen aus und ist durch diese bereits zu einem Großteil definiert.

Die Anwendungsdomäne der Peilungskomponente ergibt sich durch die Abhängigkeit von der Schiffserkennungskomponente und ihren Ausgabedaten. Somit muss die Anwendungsdomäne alle Domänengrenzen der Ausgabedaten, welche durch die Schiffserkennungskomponente bereitgestellt werden, berücksichtigen.

V4 | Definition der Prüfmetriken und Erfolgskriterien

Ausgehend von einer erfolgreichen Prüfung der Anwendungsdomäne muss der Prüfer die definierten Prüfmetriken der KI-Komponente bewerten und mit Blick auf die geplante Funktion der KI-Komponente prüfen. Prüfmetriken beschreiben in diesem Kontext direkt messbare oder indirekt bestimmbare Größen, die zur Bewertung der Ausgaben einer KI-Komponente genutzt werden und die Ergebnisse quantifizierbar und vergleichbar machen.

Die Prüfmetrik muss eine Bewertung der Ausgabewerte einer KI-Komponente hinsichtlich der Güte der Ausgabewerte ermöglichen. Dabei muss die vorliegende Funktionalität der KI-Komponente und damit die Form der entsprechenden Ausgabewerte berücksichtigt werden. Bei den Ausgabewerten kann es sich im einfachsten Fall um quantitative Werte (z.B. indiskrete Zahlenbereiche, kategoriale Zielwerte oder zweiwertige Logikaussagen) oder solche mit qualitativer Ausprägung (z.B. navigatorische Handlungsempfehlungen) handeln. Quantitative Prüfmetriken können z.B. mithilfe verschiedener Methoden wie einer Wahrheitsmatrix oder dem euklidischen Distanzmaß für die Ähnlichkeit zwischen zwei Zahlen bestimmt werden.

Der Hersteller muss bei der Definition der Prüfmetriken zudem Erfolgskriterien für die Funktion der KI-Komponente im Entwicklungsprozess festlegen. Die Definition der Erfolgskriterien steht mit der Anwendungsdomäne in Verbindung und gibt dem Prüfer Verhaltensgrenzen, die von der KI-Komponente eingehalten werden müssen.

Folgende Anforderung müssen bei der Definition der Erfolgskriterien erfüllt werden:

- Die Prüfmetriken geben dem Prüfer die Möglichkeit unter Betrachtung der Ein- und Ausgabewerte bewerten zu können, ob die KI-Komponente so (gut) funktioniert, wie es vom Hersteller erwartet wird.
- Sofern bereits Standards oder Normen zu Prüfmetriken und Erfolgskriterien für vergleichbare KI-Komponenten bestehen, dürfen die vom Hersteller genannten Erfolgskriterien gleich kritisch oder kritischer sein.

Es wird empfohlen, dass der Prüfer im Verbund mit anderen Prüfeinrichtungen internationale Standards oder Normen zu geeigneten Prüfmetriken und Erfolgskriterien für KI-Komponenten erstellt. Dies kann den Aufwand für Hersteller und Prüfer reduzieren. Der Hersteller könnte so auf bestehende Standards oder Normen zurückgreifen und der Prüfer könnte skalierbare Prüfprozesse einführen.

Die Prüfmetriken und zugehörige Erfolgskriterien teilen sich beim Anwendungsbeispiel auf beide KI-Komponenten auf.

Die Schiffserkennungskomponente sollte mittels einer Prüfmetrik bewertbar sein, die den Anteil richtig und fehlerhaft klassifizierter Schiffe wiedergibt. Dies kann z.B. über eine Wahrheitsmatrix, wie in Tabelle 4 dargestellt, und daraus ableitbaren Größen erfolgen (Navlani et al., 2021).

Tabelle 4: Beispielhafte Wahrheitsmatrix als Grundlage für Prüfmetrik im Anwendungsbeispiel.

	Schiff erkannt: JA	**Schiff erkannt: NEIN**
Schiff vorhanden: JA	**Richtig positiv: 980**	**Falsch negativ: 4**
Schiff vorhanden: NEIN	**Falsch positiv: 1**	**Richtig negativ: 265**

Die Prüfmetriken für die Peilungskomponente definieren sich aus der Genauigkeit der geschätzten Peilung durch die Bildauswertung und einer Referenz aus AIS- oder Radardaten. Die benötigte Genauigkeit sollte in Abgleich mit Vorgaben des Herstellers geprüft werden oder anderweitig festgelegt werden. Sollte der Hersteller im Falle des Beispiels keine Kriterien vorgegeben haben, wie hoch die Genauigkeit aus seiner Sicht sein muss, sollte ein sinnvoller Wert angenommen werden. In diesem Fall macht es Sinn die Genauigkeit dadurch zu definieren das eine Überprüfung mit einem anderen Gerät zur Peilungsermittlung eine Peilung mit hoher Übereinstimmung zurückgibt.

6.2. Hauptprüfung

Alle im Rahmen der Hauptprüfung durchgeführten Prüfschritte konzentrieren sich auf die eigentliche Untersuchung der KI-Komponenten des KI-Systems und der Feststellung der korrekten Funktionsweise. Dieser Abschnitt beinhaltet die Untersuchung der rechtlichen Konformität, die Beschaffung einer Testdatengrundlage für die Prüfung sowie die Prüfung und Bewertung gemäß definierter Prüfmetriken und Erfolgskriterien. Abschließend werden Rahmenbedingungen für die Gültigkeit der durchgeführten Prüfung und Bedingungen für eine Nachprüfung festgelegt.

H1 | Einhaltung geltender Verordnungen

Durch die hohe Bandbreite der in der Entwicklung befindlicher Richtlinien, Standards und Gesetzen, die sich mit der Regulierung von KI-Systemen auseinandersetzen, ist es schwierig ein festes Regelwerk zu definieren, an dem sich ein Prüfkonzept orientieren kann. Eine Sichtung ausformulierter Regularien (s. Kapitel 3.5) zeigt, wie hoch der Umfang und die Dynamik der in Arbeit befindlichen Dokumente ist und wie unterschiedlich sie sich auf die Zulassung und den Betrieb von KI-Systemen auswirken können.

Die Prüfung von existierenden Verordnungen, also Dokumente, die KI-(Teil-)Systeme definieren, Verhalten festlegen und Betriebsgrößen standardisieren, kann im Rahmen eines Prüfkonzeptes nur konzeptuell festgehalten werden. Normen, die sich auf die Architektur und den Aufbau von KI-Systemen beziehen, welche unter die gewählte Definition in dieser Studie fallen, könnten abweichende Aussagen und Empfehlungen zur Zulassung und Entwicklung dieser Systeme anführen. Dies kann insbesondere bei KI-Technologien der Fall sein, die bereits stark fortgeschritten sind und in anderen Industriesektoren experimentell schon zugelassen worden sind. Darunter fallen Bilderkennungssysteme, die mithilfe eines CNN trainiert wurden (DIN, 2020).

Bestehende Anforderungen sollten somit durch den Prüfer in die Prozesse integriert werden können und gegebenenfalls abweichende Empfehlungen von den hier aufgeführten Prüfschritten festlegen. Die genaue Abtrennung bedarf einer Einzelfallentscheidung und sollte somit individuell für jede KI-Komponente abgestimmt werden, die dadurch betroffen ist.

Aus Sicht des Prüfers ist es entscheidend, kontinuierlich neue Verordnungen zu sichten und den Prüfprozess auf diese Änderungen anzupassen. Durch die hohe Dynamik auf dem Gebiet muss dieser Prozess regelmäßig durchgeführt werden.

> Die Betrachtung bestehender Verordnungen trifft bei der Betrachtung des Anwendungsbeispiels vor allem auf die Ausgabe der Daten in einem Format nach dem NMEA-0183-Standard (DIN, 2011) sowie die Visualisierung der grafischen Komponenten auf der Brücke (IHO, 2014) zu.

H2 | Daten-Beschaffungsprozess

Kernstück der evidenzbasierten Prüfung von KI-Komponenten ist der Vergleich von Ein- und Ausgabedaten in Abgleich mit der Anwendungsdomäne. Durch die Eingabe von Daten, die die Komponente nicht kennt, soll die erfolgreiche Funktion der KI-Komponente durch den Prüfer nachgewiesen werden können.

Mit einer Beschreibung der Eingabe- und Ausgabedaten durch den Hersteller muss der Prüfer in die Lage versetzt werden, Daten in die Komponente eingeben zu können, die

einer Form entsprechen, die von der Komponente erwartet wird. Diese als Datenbeschreibung im Rahmen der Studie eingeführte Beschreibung der Daten dient als Mechanismus, den Austausch zwischen Hersteller und Prüfer zu formalisieren. Zielstellung ist es anschließend die Daten unabhängig zu beschaffen. Mögliche Methoden zur Erstellung einer Datenbeschreibung durch den Hersteller werden in Kapitel 7.3 unter Unterabschnitt D2 aufgeführt.

Bei Vorlage einer Datenbeschreibung durch den Hersteller ist zunächst eine Prüfung auf Vollständigkeit im Rahmen des Zulassungsprozesses zu prüfen. Im ersten Schritt muss der Prüfer sicherstellen, dass in der Datenbeschreibung und in der formalisierten Anwendungsdomäne die gleichen Größen verwendet werden. Dies stellt sicher, dass mit Referenz zur formalisierten Anwendungsdomäne festgelegt werden kann, in welchem Datenraum neue Daten beschafft werden müssen. Stellt der Prüfer zu diesem Zeitpunkt fest, dass die Datenbeschreibung unvollständig ist oder sich nicht mit den Größen aus der Anwendungsdomäne deckt, muss der Prüfprozess unterbrochen werden und seitens des Herstellers eine erneute Formalisierung der Datenbeschreibung erfolgen. Für den Prozess der Datenbeschaffung wurden zwei mögliche Szenarien identifiziert, die unterschiedlichen Handlungen auf Seite des Prüfers erfordern:

- Daten gemäß Datenbeschreibung existieren bereits: Das erste Szenario geht davon aus, dass Daten zur Prüfung bereits existieren und der Datenbeschreibung des Herstellers entsprechen. Dies ist insbesondere dann der Fall, wenn eine Komponente mit gleicher oder ähnlicher Funktionsweise bereits zugelassen wurde und entsprechende Daten durch den Prüfer beschafft wurden. Dies bedingt eine Datenbank mit möglichen Eingabedaten und entsprechenden Anwendungsdomänen. Es erleichtert die Vergleichbarkeit von Komponenten mit gleicher Funktion aber unterschiedlicher Referenzimplementierung. Grundsätzlich wird geraten, dass Daten anwendungsspezifisch, z.B. unter Verwendung von Operational Envelope, gespeichert werden. Es wird von der ziellosen und unstrukturierten Erzeugung sogenannter „Datenhalden“ abgeraten.
- Daten gemäß Datenbeschreibung existieren noch nicht: Dieses Szenario erfordert die Beschaffung von Daten durch Methoden der Datengeneration oder der Verwendung von öffentlich verfügbaren oder erwerbbaren Daten. Prozesse zur Beschaffung der Daten können durch den Prüfer intern durchgeführt oder an externe Dienstleister übertragen werden. Bei der Beschaffung der Daten muss sichergestellt werden, dass diese nicht durch den Hersteller bei der Entwicklung der KI-Komponenten eingesetzt worden sind.

Als vielversprechender Prozess zur Datenbeschaffung wird die Datengeneration eingestuft. Die Vorteile sind folgende: Erstens, werden hiermit Daten beschafft, die vom Hersteller in der Entwicklung nicht verwendet worden sind, zweitens, übersteigt die mögliche Menge und Varietät von theoretisch generierbaren Daten, die der anderweitig verfügbaren Daten und, drittens, können generierte Daten bereits mit Zielwerten hinterlegt werden (Nikolenko, 2021a). Bei der Datengenerierung sollte grundlegend zwischen zwei Paradigmen unterschieden werden: Datenaugmentation und Datensynthese. Bei der Datenaugmentation werden auf Grundlage von vorhandenen Daten neue Daten generiert (Nikolenko, 2021b). Das gelingt z.B. bei Bilddaten durch Transformationsprozesse (symmetrische Operationen wie Spiegelungen oder Drehungen) oder durch Bereicherungsprozesse (Ergänzungen von Objekten). Bei der Datensynthese werden hingegen vollständig neue, also künstliche, Daten generiert. Sowohl Datenaugmentation und Datensynthese finden bereits Anwendung. Datenaugmentation kommt z.B. bei Bilddaten zum Einsatz, welche durch künstliche Objekte ergänzt werden (Ekbatani et al., 2017; Frid-Adar et al., 2018). Bei der Datensynthese werden Bild- und Videodaten unter Verwendung von Spiele-Engines

(Korakakis et al., 2018; Tsirikoglou et al., 2017) oder auch andere Datenformen wie Zeitreihen (Zhang et al., 2018) synthetisiert. Datensynthese kann ferner unter Verwendung von Deklarierungssprachen umgesetzt werden, welche simulativ beliebige Szenarien erstellen können. Auf diese Weise ließen sich jederzeit zahlreiche Szenarien mit vorgegebenen Abweichungen generieren. Der Vorteil dieser Herangehensweise wäre das Vorbeugen einer Datenhalde, denn es müssten nur die Befehlssätze zur Szenariengenerierung, nicht aber der dadurch erzeugte Datensatz, abgespeichert werden. Die Entwicklung und Umsetzung dieses Ansatzes ist derzeit Forschungsgegenstand am Fraunhofer CML.

Beim Training von Modellen zeigt sich, dass die Verwendung eines durch Datenaugmentation erweiterten Datensatzes das Modellverhalten mehr verbessert als ein rein synthetischer Datensatz, weil ein augmentierter Datensatz über eine höhere Varianz verfügt (Seib et al., 2020). Es wird davon ausgegangen, dass analog die Prüfung mit einem augmentierten Datensatz wirkungsvoller ist als mit einem rein synthetischen oder ausschließlich realem Datensatz. Bei der Datengeneration können etablierte Standards (s. Kapitel 4.1.1) der anvisierten Datenquellen unterstützend wirken, da diese die möglichen Formate und Werte der zu generierenden Daten vorgeben können.

Die Prüfstelle ist dafür verantwortlich, nur solche Daten bei der Prüfung zu verwenden, die nicht vom Hersteller verwendet worden sind. Sofern der Hersteller angezeigt hat, dass er beim Training auf öffentlich verfügbare oder erwerbbare Datensätze zurückgegriffen hat, muss die Prüfstelle gewährleisten, dass sie auf diese nicht zurückgreift. Die Prüfstelle kann dies gewährleisten, indem sie auf Methoden der Datensynthese oder -augmentation oder grundsätzlich nicht auf öffentlich zugängliche Datensätze zurückgreift.

Nach Beschaffung der Daten ist im Austausch mit dem Hersteller zu prüfen, ob diese den Erwartungen des Herstellers in Abgleich mit seiner eingereichten Datenbeschreibung entsprechen. Der in Abbildung 14 dargestellte Prozess zeigt zusammenfassend wie die Abstimmung zwischen Prüfer und Hersteller zur Datenbeschreibung gestaltet werden kann.

Der vorgeschlagene Prozess sieht an zwei Punkten den Austausch zwischen Prüfer und Hersteller vor. Zum einen kann die Vollständigkeit der Datenbeschreibung durch den Hersteller nachgeschärft werden, sollten hier durch den Prüfer Unvollständigkeiten in Abgleich mit der Anwendungsdomäne festgestellt worden sein. Eine zweite Prüfung erfolgt nach der Beschaffung von Beispieldaten, die der Datenbeschreibung des Herstellers entsprechen. Dieser Schritt gibt dem Hersteller die Möglichkeit, durch den Prüfer erzeugte Daten mit internen Erwartungen zu vergleichen und mögliche Abweichungen zu kommunizieren, um gegebenenfalls Anpassungen an der Datenbeschreibung vorzunehmen. Abweichungen in den durch den Prüfer oder Hersteller beschafften Testdaten können hierbei beispielsweise in der statistischen Verteilung, Rauschen oder allgemein einer Falschannahme (Cognitive Bias) liegen.

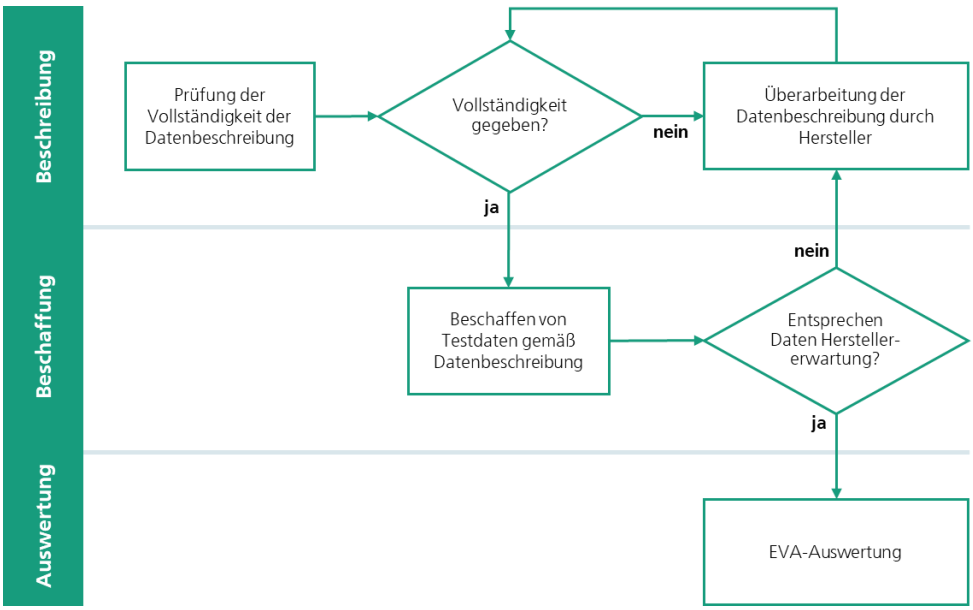

Abbildung 14: Iterative Abstimmung zur Datenbeschaffung zwischen Prüfer und Hersteller.

Die Kommunikation zwischen Hersteller und Prüfer zur Beschaffung der Testdaten auf Basis einer Datenbeschreibung bietet die Möglichkeit, Fehler im Verhalten der KI-Komponente frühzeitig zu erkennen. Durch die detaillierte Beschreibung der zu erwartenden Ein- und Ausgabedaten durch den Hersteller steigt die Erwartung, dass etwaige Testdaten im Entwicklungsprozess tiefgreifend exploriert wurden. Weiterführend hat eine streng formalisierte Datenbeschreibung zwischen Hersteller und Prüfer den Vorteil, dass Ambiguitäten oder Missverständnisse in der Beschreibung der Daten oder Funktionsweise der Komponente minimiert werden, da Doppeldeutigkeiten und Interpretationsspielräume vermieden werden.

Der Datenbeschaffungsprozess im Falle des bildgestützten Peilungssensors konzentriert sich auf Beschaffung von Bilddaten, die den Kennwerten der Schiffserkennungskomponente entsprechen und reale Aufnahmen von Schiff-zu-Schiff Situationen widerspiegeln.

Da ein solches System möglicherweise zum ersten Mal geprüft wird liegen keine Daten zum Prüfzeitpunkt vor und müssen durch den Prüfer generiert werden. Dieser Prozess kann entweder durch fähige Unternehmen mit Spezialisierung in der digitalen Erzeugung von Bilddaten realisiert werden oder durch den Aufbau interner Kompetenzen.

Durch die aktuellen Forschungsergebnisse auf dem Gebiet der KI-gestützten Bildgenerierung können weiterhin spezialisierte Unterstützungssysteme entwickelt werden, die eine Erzeugung notwendiger Daten ermöglichen. Systeme wie das KI-Model „DALL-E 2“ zeigen derzeit große Erfolge in der programmatischen Erzeugung realitätsnaher Bilddaten, welche grundlegend auch in der Erzeugung von Datenprodukten zur Zulassung des Anwendungsbeispiels genutzt werden könnten. Im Folgenden sind beispielhafte synthetische Bilderserien (Abbildung 15, Abbildung 16 und Abbildung 17) aufgeführt, die mit DALL-E 2 erzeugt worden sind (Ramesh et al., 2022).

Abbildung 15: Synthetische Bilder für den Ausdruck "fleet of ships on the horizon".

Abbildung 16: Synthetische Bilder für den Ausdruck "fleet of ships on the horizon during storm".

Abbildung 17: Synthetische Bilder für den Ausdruck "ships on the horizon looking towards the camera".

Die Beschaffung der notwendigen Daten für die Peilungskomponente hängt grundlegend von den Daten ab, die für die Schiffserkennungskomponente genutzt werden. Es ist bei der Beschaffung der Daten für die Prüfung der Schiffserkennungskomponente darauf zu achten die Anwendungsdomäne der Peilungskomponente zu berücksichtigen. Die Systeme sollten grundlegend getrennt prüfbar sein. Eine kombinierte Prüfung kann als Schritt gesehen werden den Prozess zu beschleunigen und die Effizienz des Prüfkonzeptes zu erhöhen.

H3 | Erfüllung der Erfolgskriterien

Hauptziel des Prüfprozesses ist die Validierung der Erfolgskriterien, welche durch den Hersteller im Rahmen der Vorprüfung formalisiert wurden. Diese Validierung erfolgt entlang der gesamten Anwendungsdomäne. Außerhalb der Anwendungsdomäne soll geprüft werden, ob die KI-Komponente wie erwartet inaktiv ist. Dieser Schritt spiegelt die eigentliche informations- und sicherheitstechnische Prüfung der KI-Komponente wider.

Die Daten, die im Zuge des Datenbeschaffungsprozesses erzeugt wurden, können in die Komponenten in der ausführbaren Umgebung eingegeben werden und die Ausgaben protokolliert werden. Die Erfüllung der Erfolgskriterien ergibt sich dabei im Grad der Übereinstimmung der Prüfergebnisse mit den Erfolgskriterien des Herstellers.

Beim Anwendungsbeispiel muss die Anzahl korrekt erkannter Schiffe im Bereich von 95% der Testdaten liegen. Weiterführend dürfen nicht mehr als 5% von Nicht-Schiffen fälschlicherweise als solche klassifiziert werden. Dies muss für alle Testdaten, die mit Blick auf die Anwendungsdomäne generiert worden sind, gewährleistet sein.

H4 | Notwendigkeitskriterien einer Nachprüfung

Bei erfolgreicher Prüfung einer KI-Komponente sind durch den Prüfer die Bedingungen für eine Nachprüfung (siehe Unterabschnitt N1) festzulegen. Die Notwendigkeitskriterien einer Nachprüfung können zeit- oder ereignisbasiert sein und gelten komponentenweise. Entsprechend kann die Notwendigkeit einer Nachprüfung komponentenweise festgestellt werden.

Beispiele für eine zeitbasierte Nachprüfung sind gesetzlich vorgeschriebene Zeiträume, in denen die Funktion der KI-Komponente erneut geprüft werden muss. Der Grund hierfür ist, dass in dem Prüfkonzept eingefrorene KI-Systeme, genau genommen ihre Modelle, und die Anwendungsdomäne als statisch betrachtet werden. Daraus kann mit der Zeit ein Drift zwischen Modell und Realität entstehen (s. Kapitel 2.1.1). Um die Aktualität der Komponenten zu gewährleisten, ist eine regelmäßige Prüfung der ordnungsgemäßen Funktion unabdingbar. Der in Kapitel 3.5 vorgestellte EU AI Act sieht zum Beispiel eine periodische Konformitätsbewertung für zugelassene KI-Systeme vor.

Die ereignisbasierte Nachprüfung orientiert sich hingegen an direkten Änderungen der KI-Komponente und möglichen Auswirkungen auf die Betriebssicherheit. Insbesondere bei Softwareupdates oder Veränderungen der Hardware ist durch den Prüfer festzustellen, ob die während der Prüfung validierte Funktion der KI-Komponente davon betroffen ist.

Die Festlegung der Nachprüfungskriterien wird am Ende des Prüfprozesses durchgeführt und mit Blick auf die Anwendungsdomäne festgelegt. Die Funktion des Anwendungsbeispiels hängt grundlegend von den Kennwerten der RGB-Kamera ab. Eine Änderung dieser Hardwarekomponente würde entsprechend zu einer Nachprüfung führen.

Die Peilungskomponente hängt inhärent von den Kennwerten der Schiffserkennungskomponente sowie den Installationsdaten zur Schätzung der Peilung ab. Bauliche Veränderungen der Kamera müssen auch auf Seiten der Peilungskomponenten berücksichtigt werden. Dies kann als Leistungsmerkmal im Produktdatenblatt stehen, führt anderweitig jedoch zu einer Nachprüfung.

Sollte die Schiffserkennungskomponente auf die Erkennung von Objekten erweitert werden, die nicht im ursprünglichen Umfang enthalten waren (z.B. Seezeichen oder Hindernisse), kann das eingefrorene Modell nicht mehr angewandt werden und muss durch ein neues Modell ersetzt werden. Das neue

Modell muss wiederum erneut geprüft werden, um nachzuweisen, dass die Funktionsfähigkeit für die neuen Klassifizierungen sowie alle bisherigen Leistungsumfänge weiterhin ihre Gültigkeit haben (siehe Kapitel 2.1.1).

6.3. Nachprüfung

Die Notwendigkeit einer Nachprüfung kann aufgrund von internen Änderungen am KI-System, externen Veränderungen in der Anwendungsdomäne oder dem zeitlichen Ablauf der Zertifizierung resultieren. Die Bedarfsuntersuchung und Umfangsbestimmung werden in folgenden Unterabschnitten aufgeführt.

N1 | Bedarfsuntersuchung einer Nachprüfung

Der Prüfer muss eine Übersicht aller zugelassenen KI-Komponenten führen, um den Bedarf einer Nachprüfung ermitteln zu können. Dies ist insbesondere dann notwendig, wenn die Betriebssicherheit einer Komponente, durch Eintreten eines externen Einflusses, nicht mehr gegeben ist (s. Kapitel 6.2 Unterabschnitt H4). Es ist im Einzelnen zu prüfen, ob Veränderungen der Software die erneute Prüfung der KI-Komponente erfordern oder durch vereinfachte Prozesse vor dem Aufspielen auf die Komponente validiert werden können.

N2 | Umfangsbestimmung der Nachprüfung

Wurde der Bedarf einer Nachprüfung ermittelt, ist vom Prüfer der Umfang zu bestimmen. Durch die bei der Zulassung festgelegten Bedingungen für eine Nachprüfung können so je nach Umfang eine einzelne KI-Komponenten, mehrere (zusammenhängende) KI-Komponenten oder das gesamte KI-System erneut durch den Prüfprozess geführt werden.

Eine Nachprüfung gilt als abgeschlossen, wenn analog zu den in Kapitel 6.1 festgelegten Erfolgskriterien die geforderten Ausgaben des Systems erfüllt werden. Bei jeder Nachprüfung oder Änderung der Funktionsweise der Komponente sind auch die Notwendigkeitskriterien zu prüfen und ggf. anzupassen.

7. Sicherheitskonzept

Das Sicherheitskonzept richtet sich an die Hersteller der zu prüfenden KI-Systeme und dient den beiden wesentlichen Zielen:

- Gewährleistung der Prüfbarkeit des Systems.
- Vermittlung von Hinweisen zur Entwicklung eines hinreichend sicheren Systems mit Aussicht auf erfolgreiche Prüfung.

Zur Erreichung dieser beiden Ziele wurde das Sicherheitskonzept in enger Abstimmung mit dem empfohlenen Prüfkonzept entwickelt. Das Sicherheitskonzept besteht aus drei Hauptabschnitten mit jeweils einzelnen Schritten, wobei die Schritte nicht zwangsläufig in der festgehaltenen Reihenfolge befolgt werden müssen. Die Abschnitte und Schritte sind nachfolgend in Abbildung 18 dargestellt.

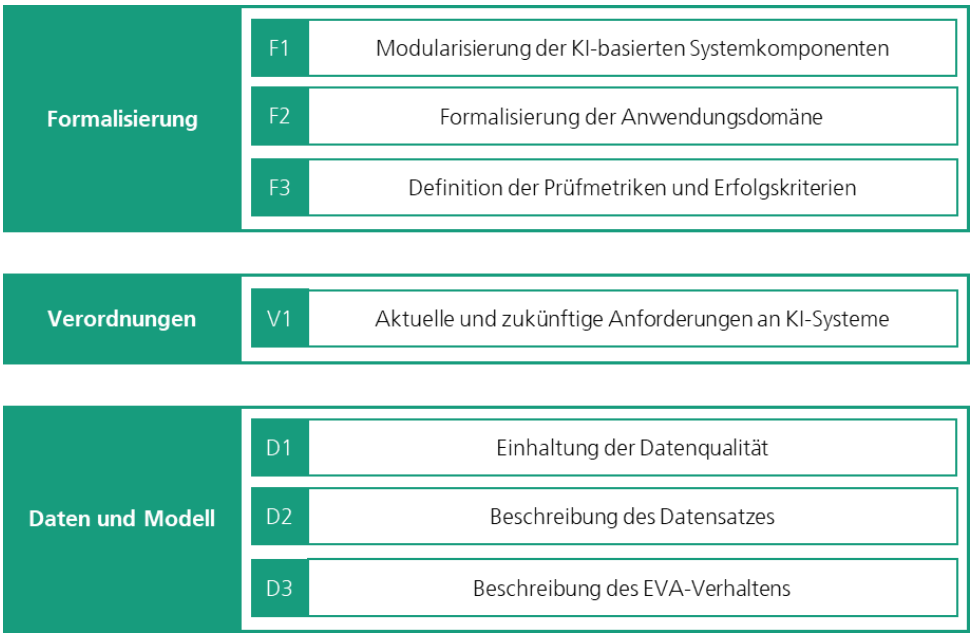

Abbildung 18: Sicherheitskonzept mit Hauptabschnitten und untergeordneten Schritten.

Mit dem ersten Abschnitt Formalisierung wird das Ziel verfolgt, ein KI-basiertes System zur Prüfung vorzubereiten. Hierin sind Schritte verortet, mit denen ein KI-System hinreichend beschrieben werden kann, damit es von der Prüfstelle geprüft werden kann. Hierfür wird das gesamtheitliche System komponentenweise hinsichtlich seiner Anwendungsdomäne und der zu erwarteten Performanz beschrieben.

Im Abschnitt Verordnungen wird dem Hersteller der Umgang mit aktuell und potentiell geltenden Verordnung aufgezeigt.

Im letzten Abschnitt Daten und Modell werden Schritte erläutert, die zum einen auf die Datenqualität und zum anderen auf die Beschreibung und Reproduzierbarkeit des Systemverhaltens abzielen. Für Ersteres werden Hinweise für einen qualitativ ordnungsgemäßen Umgang mit Daten beim Entwicklungsprozess aufgeführt. Für Letzteres werden gängige mathematische und technische Methoden aufgeführt, mittels welcher es möglich ist, den in der Entwicklung verwendeten Datensatz und das Verhalten des erzeugten KI-Modells verständlich zu beschreiben.

7.1. Formalisierung

F1 | Modularisierung in KI-basierte Systemkomponenten

Das KI-System des Herstellers wird modularisiert geprüft. Dies impliziert für den Hersteller, dass ein KI-System als prüfungsvorbereitende Maßnahme hinreichend in einzelne KI-Komponenten unterteilt werden muss (s. Kapitel 6.1). Der Hersteller kann hierfür früh die Grundsteine legen, indem er eine zur Prüfung geeignete Systemarchitektur wählt, bei der die KI-Komponenten modularisiert vorliegen.

Für den Hersteller hat die modularisierte Prüfung den Vorteil, dass er die Prüfungsergebnisse zu seinem System komponentenweise erfährt. Im Falle einer erforderlichen Nachbesserung des Systems, können diese gezielt an den verbesserungsbedürftigen Komponenten durchgeführt werden. Darüber hinaus kann der Hersteller Verbesserungen aus eigenem Interesse komponentenweise durchführen, welche die Prüfstelle dann ebenfalls gezielt prüft.

Im Folgenden wird im Sicherheitskonzept die zu prüfende Einheit als KI-Komponente betrachtet.

Bei dem bildgestützten Peilsensor handelt es sich um ein KI-System, dass sich hinsichtlich seiner Funktionen in einzelne KI-Komponenten modularisieren lässt. Der Prüfer würde dann die KI-Komponenten einzeln prüfen.

Der Hersteller könnte das KI-System in folgende beide KI-Komponenten zerlegen:

- Schiffserkennungskomponente: Schiffserkennung anhand von RGB-Bilddaten.
- Peilungskomponente: Schätzung der Peilung der Schiffe.

Das in KI-Komponenten modularisierte KI-System lässt der Hersteller dem Prüfer im Rahmen der prüfungsvorbereitenden Kommunikation (s. Kapitel 5) zukommen. Ein möglicher Weg der Modularisierung ist in Abbildung 13 skizziert.

F2 | Formalisierung der Anwendungsdomäne

Bei der Anwendungsdomäne handelt es sich aus der Sicht des Herstellers um den Anwendungsbereich, in welchem die KI-Komponente funktionieren soll. Dem Hersteller ist die Anwendungsdomäne jeder modularisierten KI-Komponente aus der Entwicklung bekannt. Konkret betrachtet handelt es sich um den Raum der Eingabe- bzw. Ausgabedaten, welche die KI-Komponente verarbeiten bzw. als Ergebnis ausgeben kann.

Die formalisierte Anwendungsdomäne dient der Prüfstelle dazu, ihren Prüfrahmen für die KI-Komponenten einzugrenzen. Damit die Prüfstelle dies möglichst richtig und genau umsetzen kann, bedarf es einer Formalisierung der Anwendungsdomäne.

Methoden zur Formalisierung der Anwendungsdomäne werden in Kapitel 6.1 Unterabschnitt V3 eingeführt.

Beide zuvor eingeführten KI-Komponenten müssen vom Hersteller hinsichtlich der Anwendungsdomäne formalisiert werden. Am Beispiel der Schiffserkennungskomponente muss der Hersteller die zur Entwicklung verwendeten und in der Anwendung vorkommenden Bilddaten zum einen aus technischer Perspektive und zum andern aus der inhaltlichen Perspektive beschreiben.

Mit der technischen Perspektive sind Größen des RGB-Kamerasystems wie Bildauflösung, Brennweite, Sichtfeld und ähnliche Größen gemeint. Diese sind notwendig, damit bei der Prüfung der KI-Komponenten möglichst dieselbe Bildtechnik verwendet wird, wie sie in der Anwendung vorkommt, um Bilddaten zu reproduzieren, die von der KI-Komponente verarbeitbar sind.

Darüber hinaus muss der in Anwendung vorkommende Bildinhalt beschrieben werden. Hierzu zählt die Beschaffenheit, also z.B. die Wetterbedingungen bei Aufnahme des Bildes. Aus dem Produktdatenblatt (s. Kapitel 4.3) ließe sich entnehmen, dass der bildgestützte Peilsensor nur bei Tageslicht unter klarer Sicht agiert.

Aus der Formalisierung der Anwendungsdomäne ist für den Hersteller ersichtlich, unter Verwendung welcher Bilder der bildgestützte Peilungssensor funktionieren muss.

Die formalisierte Anwendungsdomäne lässt der Hersteller dem Prüfer im Rahmen der prüfungsvorbereitenden Kommunikation (s. Kapitel 5) zukommen.

F3 | Definition der Prüfmetriken und Erfolgskriterien

Als letzten Schritt der Formalisierung muss der Hersteller der Prüfstelle Prüfmetriken und Erfolgskriterien vorlegen, mit denen die Güte der Ausgabe der KI-Komponenten bewertet werden kann. In diesem Zuge muss der Hersteller eine Methode zur Messung (Prüfmetrik) sowie die zu erfüllende Güte (Erfolgskriterien) bereitstellen. In Abstimmung mit der Prüfeinrichtung wird untersucht, ob diese die Prüfanforderungen der Prüfeinrichtung erfüllen. Sofern bereits Standards oder Normen zu Erfolgskriterien vergleichbarer KI-Komponenten bestehen, dürfen die vom Hersteller genannten Erfolgskriterien gleich kritisch oder kritischer sein. Die Definition der Prüfmetriken und Erfolgskriterien muss separat für jede KI-Komponente durchgeführt werden.

Bereits bei der Entwicklung eines KI-Modells sollten vom Hersteller Messmethoden angewendet werden, um die Performanz des Modells zu messen und zu optimieren. Die Wahl einer geeigneten Messmethode hängt in erster Linie von der Art der Ausgabe ab. Beispielsweise eignet sich im Falle von binären oder mehrklassigen Klassifikationsproblemen die Verwendung einer Wahrheitsmatrix (Navlani et al., 2021).

Damit der Prüfer die Performanz der KI-Komponenten bewerten und bei hinreichendem Ergebnis zulassen kann, muss der Hersteller komponentenweise eine Prüfmetrik und ein Erfolgskriterium mitgeben.

Im Rahmen der Schiffserkennungskomponente kann der Hersteller eine Wahrheitsmatrix verwenden, da es sich um ein binäres Klassifikationsproblem handelt (Navlani et al., 2021). Denn, sofern ein hinreichend erkennbares Schiff auf einem Bild vorhanden ist, muss dieses von der Schiffserkennungskomponente

erfolgreich detektiert werden. Der Hersteller muss dabei Erfolgskriterien angeben, welche erfüllt werden müssen, damit die KI-Komponente als sicherheits- und informationstechnisch funktionstüchtig gilt. Diese Messgrößen würde der Hersteller gewöhnlicherweise schon während der Entwicklung dieser KI-Komponente formulieren, um Verbesserungen messbar zu machen.

Diese Definition der Prüfmetriken und Erfolgskriterien lässt der Hersteller dem Prüfer im Rahmen der prüfungsvorbereitenden Kommunikation (s. Kapitel 5) zukommen.

7.2. Verordnungen

V1 | Aktuelle und zukünftige Anforderungen an KI-Systeme

Der Hersteller ist im Zuge des Entwicklungsprozesses für die Sichtung und Auswertung bestehender Anforderungen für die gewählte Architektur und Anwendungsdomäne verantwortlich. Wurde die Nutzung einer KI-Technologie für die Lösung eines Problems ausgewählt können verschiedene Anforderungen existieren, die in unterschiedlichem Umfang Einfluss auf die Entwicklung eines KI-Systems haben können.

Wegen der fehlenden Standardisierung von KI-Systemen und zugrunde liegenden Architekturen ist ein detaillierter Überblick im Rahmen der Studie nur bedingt möglich. Grundlegend ist bei der Entwicklung mit dem Prüfer Rücksprache zu halten welche Regularien aktuell für den Prüfungsprozess Anwendung finden.

Der Hersteller prüft schon während der Entwicklung und auch bei der Prüfungsvorbereitung sein KI-System darauf, ob und welche Regularien es erfüllen muss.

Weil die Ausgabe der Peilungskomponente über den Standard NMEA 0183 erfolgt, muss der Hersteller dieses erfüllen (DIN, 2011). Darüber hinaus beobachtet der Hersteller die Entwicklung der Umsetzung des EU AI Acts (Europäische Kommission, 2021), da dieses Auswirkungen auf sein Produkt haben könnte (s. Kapitel 3.5.2) und nimmt den Leitfaden zur Entwicklung von Deep-Learning-Bilderkennungssystemen (DIN, 2020) zur Kenntnis.

7.3. Daten und Modell

D1 | Einhaltung der Datenqualität

Mangelnde Datenqualität äußert sich im Allgemeinen anhand von fehlenden, unvollständigen, inkonsistenten, ungenauen oder doppelten Daten (DIN, 2020; Gudivada et al., 2017). Speziell bei Machine Learning sind häufig weitere Erscheinungsformen mangelnder Datenqualität zu beobachten: die Verwendung von zu vielen Variablen, stark miteinander korrelierte Variablen oder Ausreißer im Datensatz.

Datenqualität im Kontext von Machine Learning ist von zentraler Bedeutung. Denn die verwendbaren Daten bilden das Wissen ab, basierend auf welchem das KI-Modell trainiert wird (s. Kapitel 2.1.1). Bei geringer Datenqualität kann es u.a. zu Erscheinungsformen wie hohem Bias oder hoher Varianz kommen, welche sich in geringer Richtigkeit bzw. Präzision manifestieren. In Abbildung 19 sind diese Auswirkungen anhand von Würfen auf Dartscheiben vereinfacht dargestellt. Mangelnde Datenqualität kann zur Folge haben, dass ein KI-System nicht zuverlässig funktioniert, auch wenn ihre gewählten Machine-Learning-Methoden überaus geeignet sind und die sonstige Entwicklung hohen Qualitätsstandards folgte.

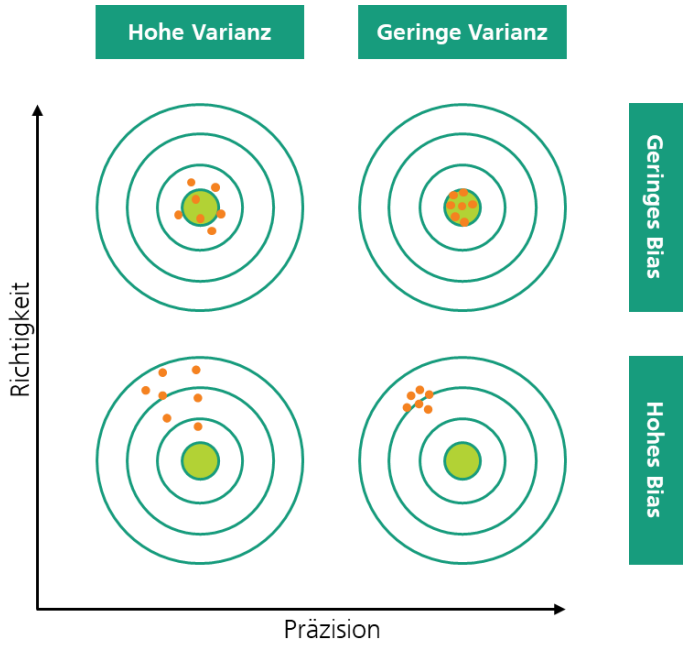

Abbildung 19: Zusammenhang zwischen Varianz und Präzision sowie Bias und Richtigkeit als Folgen mangelnder Datenqualität.

Die Einhaltung und Verbesserung von Datenqualität mittels geeigneter Methoden und Kontrollen wird als Datenqualitätsmanagement bezeichnet (Gudivada et al., 2017). Die Umsetzung von Datenqualitätsmanagement erhöht die Chancen, dass ein funktionstüchtiges KI-System entwickelt wird, welches folglich eine Prüfung erfolgreich besteht. Datenqualitätsmanagement kann z.B. mittels unternehmens- und branchenweiten Datenstandards oder einheitlichen Prozessen zur Beseitigung von Ungereimtheiten wie Inkonsistenzen, Ungenauigkeiten oder Ausreißern umgesetzt werden. Methoden zur Umsetzung von Datenqualitätsmanagement werden in (Burkov, 2020; Gudivada et al., 2017) gründlich behandelt.

Bei der Entwicklung der Schiffserkennungskomponente muss der Hersteller auf verschiedene Datensätze für das Training des KI-Modells zurückgreifen, um einen hinreichend großen, gesamtheitlichen Datensatz zu erhalten. Den zusammengeführten Datensatz prüft der Hersteller auf unterschiedliche Aspekte und formuliert folgende Fragestellungen:

- Ist die statistische Verteilung der im Datensatz vorkommenden Bilddaten zu peilender Objekte repräsentativ für die Realanwendung?
- Sind nur solche Bilddaten enthalten, die auch von der Anwendungsdomäne abgedeckt sind (u.a. klare Sichtverhältnisse im Tageslicht)?
- Sind die Bilddaten richtig klassifiziert?

Unter der Betrachtung der obigen Fragestellungen passt der Hersteller gegebenenfalls den Datensatz an.

D2 | Beschreibung des Datensatzes

Bei der Entwicklung eines KI-Systems verwendet der Hersteller einen oder mehrere Datensätze. Bereits durch die Ausübung von Datenqualitätsmanagement und insbesondere beim Training von ML-basierten KI-Modellen sollte sich der Hersteller mit den verwendeten Datensätzen auseinandergesetzt haben. Denn, wie zuvor erwähnt, haben die Datenqualität, aber auch schon der verwendete Trainingsdatensatz einen erheblichen Einfluss auf das Verhalten der KI-Modelle.

Damit die Prüfstelle das Verhalten der KI-Modelle untersuchen kann, muss sie über hinreichende Kenntnisse zu den Eigenschaften der Trainingsdatensätze verfügen, damit sie sich eigenständig geeignete Testdatensätze beschaffen kann. Diese Eigenschaften der Datensätze sind von dem Hersteller im Rahmen der Datenbeschreibung an die Prüfstelle frühzeitig zu kommunizieren. Insbesondere muss durch den Hersteller angezeigt werden, wenn das Training auf öffentlich verfügbaren oder erwerbbaren Datensätzen erfolgt.

Die Datenbeschreibung liefert komponentenweise eine Beschreibung der bei der Entwicklung verwendeten Datensätze. Eine Möglichkeit zur Beschreibung eines Datensatzes wird im Folgenden tabellarisch entlang zweier Achsen (s. Abbildung 20) veranschaulicht und anhand eines AIS-Datensatzes kurz beispielhaft erläutert. Daten in einem Datensatz bestehen aus einer oder mehreren Dimensionen (horizontale Achse in Abbildung 20). Die Dimensionen repräsentieren in einem AIS-Datensatz statische und dynamische Variablen, wie z.B. die MMSI-Nummer der Schiffe, ihre Positionen und die Zeit der Schiffspositionen. Jede Dimension lässt sich beispielsweise anhand ihres Dimensionstyps (nominal, ordinal oder numerisch) unterscheiden und bei numerischen Dimensionstypen ferner zwischen diskreten und kontinuierlichen Dimensionen unterscheiden (Navlani et al., 2021). Die möglichen Werte jeder Dimension lassen sich mit der Angabe des möglichen Wertebereich einschränken. Wenn vorhanden, müssen auch die Beziehungen zwischen den Dimensionen angegeben werden.

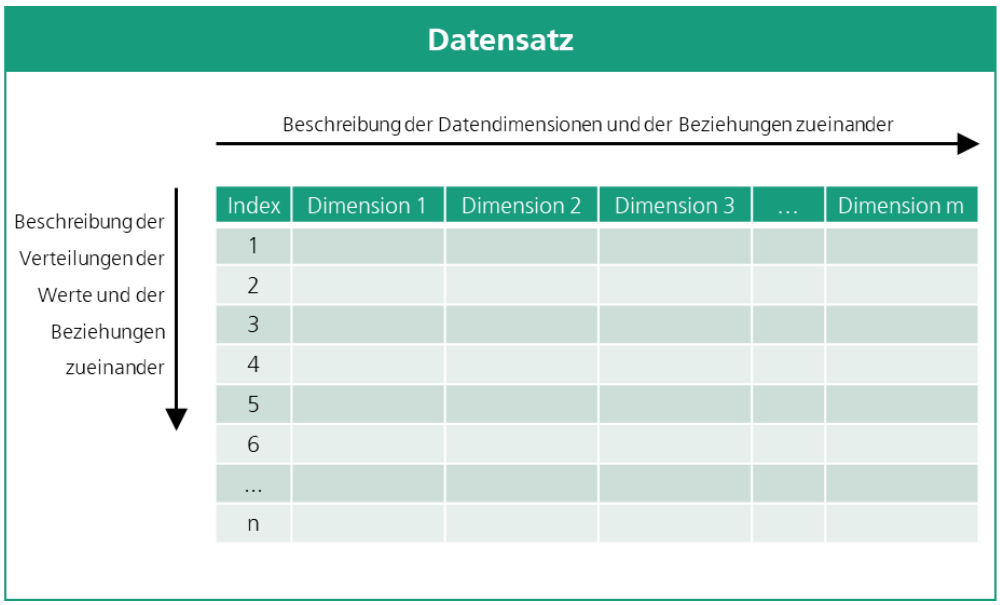

Abbildung 20: Beispielhafte Veranschaulichung einer Datenbeschreibung mit Hilfe einer Matrix.

Die verschiedenen Daten eines Datensatzes werden entlang der vertikalen Achsen aufgeführt. Im Beispiel des AIS-Datensatz würden AIS-Nachrichten untereinander aufgeführt werden. Zur Beschreibung der auftretenden Werte jeder Dimension kann eine geeignete statistische Verteilung angegeben oder ein Stichprobensatz zur Schätzung bereitgestellt werden. Darüber hinaus sollte, wenn vorhanden, die Beziehung der Werte zueinander beschrieben werden, z.B. bei Zeitreihen.

Für weitere Methoden, mit der sich ein Datensatz beschreiben lässt, sei auf die deskriptive Statistik verwiesen (Navlani et al., 2021).

Die Datenbeschreibung sollte gemeinsam mit Beispieldaten der Prüfstelle übergeben werden. Der Umgang der Prüfstelle mit der Datenbeschreibung und wann diese hinreichend ausführlich ist, wird in Kapitel 6.1 im Unterabschnitt V2 erläutert.

Jeden verwendeten Datensatz, welchen der Prüfer reproduzieren muss, beschreibt der Hersteller. Im Falle der Bilddaten für die Schiffserkennung besitzt der Hersteller einen Bilddatensatz, welcher in Anlehnung an Abbildung 20 wie folgt aufgebaut ist.

Eine Dimension repräsentiert ein Bild an sich. In der gleichen Zeile wird der Bildinhalt anhand weiterer Dimensionen und gemäß einer Semantik beschrieben. So könnten weitere Dimensionen Wetter-, Wasser-, und Lichtverhältnisse beinhalten. Eine weitere Dimension beinhaltet die Bildpositionsdaten für vorkommende Schiffe und eine hierzu korrespondierende Dimension die Schiffstypenbezeichnungen. Damit repräsentiert jede Zeile ein Bild und die zugehörige Beschreibung.

Die Rückschlüsse zur statistischen Verteilung der Bildinhalte zieht der Hersteller, indem er versucht die statistische Verteilung der einzelnen Dimension zu beschreiben. Damit versucht er u.a. folgende Fragestellung zu beantworten:

- Welche Schiffstypen tauchen wie oft auf?

- Wie viele Bilder sind gänzlich ohne Schiffe?
- Tauchen unter bestimmten Voraussetzungen manche Schiffstypen öfter oder seltener auf?
- Welche Wetter-, Wasser-, und Lichtverhältnisse kommen wie häufig vor?

Analog geht der Hersteller für alle weiteren verwendeten Datensätze der übrigen KI-Komponenten vor. Die Beschreibungen der Datensätze lässt er dem Prüfer im Rahmen der prüfungsvorbereitenden Kommunikation zukommen.

D3 | Beschreibung des Modellverhaltens

Damit die Prüfstelle der Fragestellung nachgehen kann, ob das vorliegende KI-System nach der Vorgabe des Herstellers (hinreichend) richtig funktioniert, muss der Hersteller eine Beschreibung zum Modellverhalten der KI-Komponenten bereitstellen. Die Beschreibung des Modellverhaltens kann dem zuvor eingeführten EVA-Prinzip folgen (s. Kapitel 6.1 Unterabschnitt V2).

Aus dieser Beschreibung soll hervorgehen, welche Eingabewerte zu welchen Ausgabewerten führen sollen. Diese Beschreibung muss dahingehend hinreichend genau sein, dass die Prüfstelle das beschriebene Modellverhalten in der Anwendungsdomäne prüfen und die Richtigkeit entsprechend der definierten Prüfmetriken komponentenweise messen und bewerten kann.

Der Hersteller beschreibt das Modellverhalten seiner KI-Komponenten. Dabei bedient er sich dem EVA-Prinzip und erklärt das Modellverhalten anhand der erwarteten Ausgaben für verschiedene Eingaben.

Im Falle der Schiffserkennungskomponente muss betrachtet werden, unter welcher Eingabe die Schiffserkennung angestoßen wird und ein Schiff identifiziert werden soll. Aus dem Produktdatenblatt (s. Kapitel 4.3) folgt z.B., dass der bildgestützte Peilungssensor und damit auch die Schiffserkennung nur bei klarer Sicht unter Tageslicht funktioniert. Diese Umgebung beschreibt der Hersteller als Grundvoraussetzungen für eine erfolgreiche Funktion des Modells.

Unter der Voraussetzung der Erfüllung der Grundvoraussetzungen existieren zwei Fälle, welche das Modellverhalten hinreichend beschreiben. Im ersten Fall der Eingabe ist kein Schiff hinreichend groß auf einem Bild zu erkennen. Hier darf die Schiffserkennung in der Ausgabe an keiner Stelle ein Schiff identifizieren. Im zweiten Fall, sind ein oder mehrere hinreichend groß abgebildete Schiffe erkennbar. Hier muss in der Ausgabe jedes einzelne Schiff angemerkt sein.

Aus dieser Beschreibung des Modellverhaltens der Schiffserkennungskomponente ist für den Prüfer ersichtlich, welche Ausgabe bei welcher Eingabe zu erwarten ist. Darauf aufbauend können die KI-Komponenten auf Funktion geprüft werden.

Die Beschreibungen der Modellverhalten aller KI-Komponenten lässt der Hersteller dem Prüfer im Rahmen der prüfungsvorbereitenden Kommunikation zukommen.

8. Zusammenfassung und Handlungsempfehlungen

Das Ergebnis der Studie ist ein Konzept zur Prüfung und Zulassung von KI-Systemen. In diesem Kapitel werden die wichtigsten Teilergebnisse zusammengefasst und resultierende Handlungsempfehlungen (blau hinterlegt) für das BSH abgeleitet.

Verortung der Prüf- und Zertifizierungsprozesse in einem separaten Modul K

Die Untersuchung des bestehenden Konformitätsbewertungsverfahren von maritimen Ausrüstungsgegenständen ergibt, dass keine anwendbaren Prozesse zur Prüfung und Zulassung von KI-Systemen (s. Kapitel 3) existieren. Zwar werden Anstrengungen von der Europäischen Kommission bei der Zulassung von KI-Systemen betrieben (s. EU AI Act in Kapitel 3.5), doch rechtlich bindende oder allgemein nutzbare Verfahren wurden bisher noch nicht vorgelegt.

Unter Betrachtung bestehender Veröffentlichungen der Europäischen Kommission sowie der Verfahren in der MED wird zur Einführung eines separaten Moduls in den bisherigen Prüf- und Zertifizierungsprozess geraten. Es liegt nahe, in diesem separaten Modul (siehe Modul K in Kapitel 3.4) die in dieser Studie herausgearbeiteten Prüf- und Zertifizierungsprozesse für KI-Systeme zu verorten. Dies ermöglicht die Integration der vorgeschlagenen Prüfprozesse, ohne die bestehenden Module anpassen zu müssen.

Standardisierung des Informationsaustausches bei KI-Systemen

Aus der Marktanalyse bestehender und sich in Entwicklung befindender KI-basierter Produkte im maritimen Umfeld (s. Kapitel 4) resultiert eine Übersicht zu Datenquellen (s. Tabelle 5 in Anhang A.1.) und Anwendungsfällen (s. Tabelle 6 in Anhang A.1.) der Produkte. Hierbei wird eine intensive Nutzung von Kamerasystemen festgestellt (s. Kapitel 4.1.1). Diese bedürfen eine detailliertere Betrachtung, weil für ihren Informationsaustausch bisher keine Standardisierung im maritimen Kontext vorliegt.

Es wird geraten bei der Prüfung und Zertifizierung von KI-Systemen die Standardisierungen des Informationsaustausches und ihrer Datenquellen voranzutreiben. Solche Standardisierungen können den Prüfprozess maßgeblich vereinfachen und skalieren.

Einführung eines modellagnostischen Prüfprozesses

Zur Integration des Modul K in den bestehenden Prüfprozess wird in dieser Studie eine Möglichkeit zur Organisation des Austausches zwischen Hersteller und Prüfer vorgestellt, welche als Grundlage für das Sicherheits- und Prüfkonzept dient (s. Kapitel 5). Dieser Austauschprozess beschreibt den Rahmen der Kommunikation zwischen Hersteller (Sicherheitskonzept) und Prüfer (Prüfkonzept).

Der gesamte Prüfprozess versteht sich als iterativer Prozess zur schrittweisen Prüfung eines KI-Systems. Die Prüfung basiert auf der evidenzbasierten Betrachtung von Ein- und Ausgabedaten, angelehnt an das EVA-Prinzip (s. Kapitel 6.1 Unterabschnitt V2). Dieses Verfahren kann dabei sowohl für sAI- als auch für CI-Systeme einheitlich verwendet werden. Das Prüfkonzept ist also modellagnostisch ausgelegt und

ermöglicht es ohne Einblick in die Funktionsweise oder Architektur des KI-Systems die ordnungsgemäße Funktion zu prüfen. Die Aufteilung des Prüfprozesses erlaubt es ferner ein KI-System schrittweise zu prüfen und somit neuartige Prozesse in das bestehende Prüf- und Zulassungswesen zu integrieren.

Um die Vielzahl verschiedenartiger KI-Systeme einheitlich prüfen zu können sowie die Zukunftsfähigkeit des Prüfprozesses zu gewährleisten, wird die Etablierung eines modellagnostischen Prüfprozesses nahegelegt. Der Fokus der Prüfung sollte darin liegen, festzustellen, „ob" und nicht „wie" ein KI-System funktioniert.

Formalisierung der Anwendungsdomänen von KI-Systemen

Das vorgeschlagene Konzept betrachtet die Prüfung der ordnungsgemäßen Funktion eines KI-Systems auf einer formalisierten Anwendungsdomäne. Im Rahmen dessen wird die Formalisierung der Eingabe- und Ausgabedaten vorgestellt (s. Kapitel 6.1 Unterabschnitte V2 und V3). Die kann einheitlich mit einer Methode zur standardisierten Domänenbeschreibung, wie das Operational Envelope, gelingen. Basierend darauf kann die Prüfstelle stets ermitteln, welche Testdaten für den vorliegenden Anwendungsfall bei einer Prüfung benötigt werden und ob diese schon vorliegen. Darüber hinaus werden im vorliegenden Konzept Möglichkeiten zur Messung (s. Prüfmetrik in Kapitel 6.1 Unterabschnitt V4) und Bewertung (s. Erfolgskriterium in Kapitel 6.1 Unterabschnitt V4) der ordnungsgemäßen Funktion eines KI-Systems erläutert (s. Kapitel 6.1 Unterabschnitt V4). Diese ermöglichen auch die Kommunikation einer Erwartungshaltung der Funktionstüchtigkeit eines KI-Systems von Prüfer zum Hersteller sowie die Vergleichbarkeit zwischen ähnlichen KI-Systemen.

Um eine einheitliche Prüfung, die Vergleichbarkeit ähnlicher Produkte sowie die Skalierbarkeit des Prüfprozesses zu ermöglichen, wird geraten, sowohl bei der Beschreibung der Anwendungsdomäne als auch bei der Messung und Bewertung der Funktionstüchtigkeit auf standardisierte Methoden zur Formalisierung zu setzen.

Aufbau einer automatisierbaren Datenverarbeitungsinfrastruktur

Zur Skalierung und zügigen Reproduzierbarkeit des Prüfprozesses wird die Entwicklung einer automatisierten Datenverarbeitungsinfrastruktur geraten. Diese Infrastruktur sollte im Stande sein für (standardisierte) Anwendungsdomänen Testdaten zu beschaffen, sie vom KI-Modell verarbeiten zu lassen und schließlich ein gemäß Prüfmetrik und Erfolgskriterium gemessenes und bewertetes Ergebnis zurückzugeben. Die Beschaffung der Testdaten, welches den ersten Schritt einer solchen Datenverarbeitungsinfrastruktur darstellt, kann immens von der Verwendung formalisierter Anwendungsdomänen und standardisierter Domänenbeschreibung (s. Operational Envelope in Kapitel 6.1 Unterabschnitt V3) profitieren. Darüber hinaus wird gezeigt, dass Datengenerierung (sowohl durch Datenaugmentation als auch durch Datensynthese) ein vielversprechender Weg zur zielgerichteten Datenbeschaffung ist (s. Kapitel 6.1 Unterabschnitt H2). Die Verwendung standardisierter Domänenbeschreibungen des KI-Systems und eine jederzeit mögliche Datengenerierung beugen darüber hinaus die Bildung von unerwünschten Datenhalden vor.

Die technische Umsetzung der Prüfprozesse sollte anhand einer automatisierbaren Datenverarbeitungsinfrastruktur erfolgen, um Skalierbarkeit und Reproduzierbarkeit zu gewährleisten. Hierfür ist es fundamental auf standardisierte Methoden der Domänenbeschreibung der KI-Produkte zu setzen, um den Datenbeschaffungsprozess zu automatisieren. Des Weiteren ist die Verwendung von synthetischen oder augmentierten Daten ein vielversprechender Weg, um unabhängig jederzeit die notwendigen Testdaten zu beschaffen, ohne dabei langfristig Datenhalden aufzubauen. Ein weiterer Vorteil in der Verwendung von synthetischen (oder augmentierten) Testdaten liegt darin, dass die Prüfeinrichtung Daten erzeugen kann, die vom Hersteller bei der Entwicklung nicht verwendet worden sind.

Berücksichtigung der Veränderlichkeit der KI-Systeme und ihrer Anwendungsdomänen

In dieser Studie werden nur eingefrorene KI-Modelle (s. Kapitel 2.1.1) betrachtet, also jene, die sich durch Interaktion mit ihrer Umwelt in ihrem Modellverhalten nicht verändern. Unerwartetes und nicht geprüftes Modellverhalten am KI-System ist mit der Zeit dennoch möglich (siehe Drift in Kapitel 2.1.1), und zwar zum einen durch die einfache Veränderlichkeit von KI-Systemen (z.B. durch Updates) sowie durch Veränderungen der Anwendungsdomäne der KI-Systeme (Schiffserkennungslösungen betreffend z.B. durch verändertes Aussehen von Schiffen). Daher wird die Notwendigkeit und Funktionsweise einer Nachprüfung erläutert und eingeführt (s. Kapitel 6.3). In Zusammenhang mit der beschriebenen Modularisierung eines KI-Systems sind die Bedingungen gegeben, KI-Systeme komponentenweise weiterzuentwickeln und diese Veränderungen in ein bereits geprüftes KI-System zu integrieren, ohne den gesamten Prüfprozess am gesamten System erneut durchlaufen zu müssen.

Abschließend wird dem BSH empfohlen, das vorgeschlagene modellagnostische Prüf- und Sicherheitskonzept initial für eine einfache Anwendungsdomäne in Prüfprozesse umzusetzen. Um frühzeitig die Prüfprozesse skalierbar und umsetzbar zu gestalten wird empfohlen, frühestmöglich Mess- und Leistungsstandards einzuführen. Dies vermittelt den Herstellern eine Erwartungshaltung und es bietet innerhalb einer Anwendungsdomäne eine Vergleichbarkeit zwischen den KI-Systemen. Zu guter Letzt muss betont werden, dass zu erwarten ist, dass insbesondere eine geeignete Datenverarbeitungsinfrastruktur die Umsetzbarkeit, Skalierbarkeit und Zukunftsfähigkeit der Prüfprozesse maßgeblich erhöht.

9. Würdigung und abschließende Anmerkung

Die vorliegende Studie im Rahmen des BMDV-Expertennetzwerks „Wissen – Können – Handeln“ wurde durch das Bundesministerium für Digitales und Verkehr (BMDV) finanziert.

Die Inhalte der vorliegenden Studie repräsentieren den aktuellen Wissenstand der Autoren des Fraunhofer CML und keine subjektive Einschätzung des BSH.

Die Erkenntnisse und Betrachtungen basieren auf dem Recherchestand vom August 2022.

10. Literaturverzeichnis

ABB. (2018). *ABB Ability TM Marine Pilot Vision Modular Situational Awareness Platform*. ABB. Abgerufen am 09/09/2022, von https://library.e.abb.com/public/12ae485d68b0428884dcd453baf0c296/3AFV6116339_A_en_Pilot_Vision Leaflet.pdf

Alphabet Inc. (o. D.). *Google Patents*. Google. Abgerufen am 22/06/2022, von https://patents.google.com/?q=(%22Autonomous%22+AND+%22Ship%22)&before=priority:20220101&after=priority:19900101

ASAM. (2021). *ASAM OpenODD: Concept Paper Version 1.0*. ASAM. Abgerufen am 09/09/2022, von https://www.asam.net/index.php?eID=dumpFile&t=f&f=4544&token=1260ce1c4f0afdbe18261f7137c689b1d9c27576

Avikus. (o. D.-a). *Avikus AiBOAT*. Avikus. Abgerufen am 09/09/2022, von https://www.avikus.ai/eng/product/aiboat

Avikus. (o. D.-b). *Avikus HiNAS*. Avikus. Abgerufen am 09/09/2022, von https://www.avikus.ai/eng/product/hinas

Brooks, S. K., & Greenberg, N. (2022). Mental health and psychological wellbeing of maritime personnel: a systematic review. *BMC Psychology*, *10*(1), 1–26. https://doi.org/10.1186/s40359-022-00850-4

BSB AI. (2022). *Oscar Navigation Products*. BSB AI. Abgerufen am 09/09/2022, von https://www.oscar-navigation.com/#products

Bundesamt für Schifffahrt und Hydrographie. (2022). *Nationale Zulassung*. BSH. Abgerufen am 09/09/2022, von https://www.bsh.de/DE/THEMEN/Schifffahrt/Schiffsausruestung_Marktueberwachung/Nationale_Zulassung/nationale_zulassung_node.html

Burkov, A. (2019). *The Hundred-Page Machine Learning Book*. Andriy Burkov.

Burkov, A. (2020). *Machine Learning Engineering*. True Positive Inc.

Burmeister, H. C., Constapel, M., Ugé, C., & Jahn, C. (2020). From Sensors to MASS: Digital Representation of the Perceived Environment enabling Ship Navigation. *IOP Conference Series: Materials Science and Engineering*, *929*(1). https://doi.org/10.1088/1757-899X/929/1/012028

Captain AI. (2022). *Technology – Captain AI*. Captain AI. Abgerufen am 09/09/2022, von https://www.captainai.com/technology/

Danish Maritime Authority. (2018). *Analysis of regulatory barriers to the use of autonomous ships. Final Report*. https://dma.dk/Media/637745499808186153/Analysis of Regulatory Barriers to the Use of Autonomous Ships.pdf

Daranda, A., & Dzemyda, G. (2020). Navigation decision support: Discover of vessel traffic anomaly according to the historic marine data. *International Journal of Computers, Communications and Control*, *15*(3), 1–9. https://doi.org/10.15837/IJCCC.2020.3.3864

DIN. (2011). *DIN EN 61162-1:2011-09 Navigations- und Funkkommunikationsgeräte und -systeme für die Seeschifffahrt - Digitale Schnittstellen - Teil 1: Ein Datensender und mehrere Datenempfänger*. https://www.beuth.de/de/norm/din-en-61162-1/143482391

DIN. (2020). DIN SPEC 13266:2020-04 Leitfaden für die Entwicklung von Deep-Learning-Bilderkennungssystemen. In *DIN SPEC (PAS)* (Ausgabe April 2020). https://doi.org/https://dx.doi.org/10.31030/3134557

Ekbatani, H. K., Pujol, O., & Segui, S. (2017). Synthetic Data Generation for Deep Learning in Counting Pedestrians. *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods*, *2017-Januar*, 318–323. https://doi.org/10.5220/0006119203180323

Etzkorn, P. (2022). Der Schiffszusammenstoß unter Beteiligung autonom fahrender Schiffe. In *Der Schiffszusammenstoß unter Beteiligung autonom fahrender Schiffe*. Nomos Verlagsgesellschaft mbH & Co. KG. https://doi.org/10.5771/9783748934073

Europäische Kommission. (2021). *Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für Künstliche Intelligenz (Gesetz über Künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union*. Europäische Kommission. https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX%3A52021PC0206

Europäische Kommission. (2022). *Durchführungsverordnung (EU) 2022/1157 der Kommission vom 4. Juli 2022 mit Vorschriften für die Anwendung der Richtlinie 2014/90/EU des Europäischen Parlaments und des Rates hinsichtlich der Entwurfs-, Bau- und Leistungsanforderungen sowie der Prüfnormen* (s. 1–243). Amtsblatt der Europäischen Union. http://data.europa.eu/eli/reg_impl/2022/1157/oj

Europäisches Parlament und Rat der Europäischen Union. (2014). *Richtlinie 2014/90/EU des europäischen Parlaments und des Rates vom 23. Juli 2014 über Schiffsausrüstung und zur Aufhebung der Richtlinie 96/98/EG des Rates (2014/90/EU)* (s. 146–185). Amtsblatt der Europäischen Union. http://data.europa.eu/eli/dir/2014/90/oj

Flasiński, M. (2016). *Introduction to Artificial Intelligence*. Springer International Publishing. https://doi.org/10.1007/978-3-319-40022-8

Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H. (2018). Synthetic data augmentation using GAN for improved liver lesion classification. *Proceedings - International Symposium on Biomedical Imaging*, *April 2018*, 289–293. https://doi.org/10.1109/ISBI.2018.8363576

Gudivada, V. N., Ding, J., & Apon, A. (2017). Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations. *International Journal on Advances in Software*, *10.1*, 1–20. https://www.researchgate.net/publication/318432363

Gyllenhammar, M., Johansson, R., Warg, F., Chen, D., Heyn, H.-M., Sanfridson, M., Söderberg, J., Thorsén, A., Ursing, S., Ab, Z., & Com, M. G. (2020). Towards an Operational Design Domain That Supports the Safety Argumentation of an Automated Driving System. *10th European Congress on Embedded Real Time Systems*, 1–10. https://www.diva-portal.org/smash/get/diva2:1390550/FULLTEXT01.pdf

IHO. (2014). *Specifications for Chart Content and Display Aspects of ECDIS*. Abgerufen am 10/04/2022, von www.iho.int

IMO. (1971). *Resolution A.224(VII), Performance standards for Echo-Sounding equipment*. https://wwwcdn.imo.org/localresources/en/KnowledgeCentre/IndexofIMOResolutions/AssemblyDocuments/A.224(7).pdf

IMO. (1995). *Resolution A.819(19), Recommendation on Performance Standards for*

Shipborne Global Positioning System (GPS) Receiver. https://wwwcdn.imo.org/localresources/en/KnowledgeCentre/IndexofIMOResolutions/AssemblyDocuments/A.819(19).pdf

IMO. (1998). *Resolution MSC.74(69), Adoption of New and Amended Performance Standards*. https://wwwcdn.imo.org/localresources/en/OurWork/Safety/Documents/AIS/Resolution MSC.74(69).pdf

IMO. (2001). *Resolution A.915(22), Revised maritime policy and requirements for a future Global Navigation Satellite System (GNSS)*. https://wwwcdn.imo.org/localresources/en/KnowledgeCentre/IndexofIMOResolutions/AssemblyDocuments/A.915(22).pdf

IMO. (2004). *Resolution MSC.192(79), Adoption of the Revised Performance Standards for Radar Equipment*. https://wwwcdn.imo.org/localresources/en/KnowledgeCentre/IndexofIMOResolutions/MSCResolutions/MSC.192(79).pdf

IMO. (2015). *Resolution A.1106(29), Revised guidelines for the onboard operational use of shipborne automatic identification systems (AIS)*. https://wwwcdn.imo.org/localresources/en/OurWork/Safety/Documents/AIS/Resolution A.1106(29).pdf

IMO. (2017). *MSC.1/Circular.1575, Guidelines for Shipborne Position, Navigation And Timing (PNT) Data Processing*. https://www.imorules.com/MSCCIRC_1575.html

IMO. (2022). *Maritime Safety Committee (MSC 105), 20-29 April 2022*. International Maritime Organization. Abgerufen am 24/04/2022, von https://www.imo.org/en/MediaCentre/MeetingSummaries/Pages/MSC-105th-session.aspx

ITU. (2014). *Technical characteristics for an automatic identification system using time division multiple access in the VHF maritime mobile frequency band (Recommendation ITU-R M.1371-5)*. https://www.itu.int/dms_pubrec/itu-r/rec/m/R-REC-M.1371-5-201402-I!!PDF-E.pdf

Kongsberg. (2017). *Autonomy is here - Powered by Kongsberg*. Kongsberg. Abgerufen am 09/09/2022, von https://www.kongsberg.com/maritime/about-us/news-and-media/our-stories/autonomy-is-here--powered-by-kongsberg/

Kongsberg. (2019). *Kongsberg Autonomous Shipping*. Kongsberg. Abgerufen am 09/09/2022, von https://www.kongsberg.com/maritime/support/themes/autonomous-shipping/?_t_id=MJquBrAbUl9fFaQl-KpGfQ%3D%3D&_t_uuid=fbm8o_i1TbajRZ0iocOo4w&_t_q=Autonomous+shipping&_t_tags=language%3Aen%2Csiteid%3A24c9be7d-c7a0-47ff-9aff-d09ef8b15bbc%2Candquerymatch&_t_hit.

Kongsberg. (2022). *Kongsberg Situation Awareness*. Kongsberg. Abgerufen am 09/09/2022, von https://www.kongsberg.com/maritime/products/situational-awareness/?_t_id=I_tj8xlY0kjRMfKEZQECPA%3D%3D&_t_uuid=sGRclj3wQHygY_d6z1awwg&_t_q=Intelligent+Awareness&_t_tags=language%3Aen%2Csiteid%3A24c9be7d-c7a0-47ff-9aff-d09ef8b15bbc%2Candquerymatch&_t_hit.id

Korakakis, M., Mylonas, P., & Spyrou, E. (2018). A short survey on modern virtual environments that utilize AI and synthetic data. *MCIS 2018 Proceedings*, 34. https://aisel.aisnet.org/mcis2018/34

Marine AI Ltd. (2022). *Guardian by Marine AI*. Abgerufen am 12/09/2022, von https://marineai.co.uk/products/guardian/#autonomy

Mayflower Autonomous Ship. (2022). *Mayflower Autonomous Ship - Technology*. Abgerufen am 09/09/2022, von https://mas400.com/technology

Minter, A. (2021). *The Next Shipping Crisis: A Maritime Labor Shortage*. Bloomberg.Com. Abgerufen am 26/09/2022, von https://www.bloomberg.com/opinion/articles/2021-11-06/the-next-shipping-crisis-a-maritime-labor-shortage

Navlani, A., Fandango, A., & Idris, I. (2021). *Python Data Analysis: Perform data collection, data processing, wrangling, visualization, and model building using Python* (3rd ed.). Packt Publishing.

Nikolenko, S. I. (2021a). *Synthetic Data for Deep Learning* (Vol. 174). Springer International Publishing. https://doi.org/10.1007/978-3-030-75178-4

Nikolenko, S. I. (2021b). Introduction: The Data Problem. In *Springer Optimization and Its Applications* (Vol. 174, pp. 1–17). Springer International Publishing. https://doi.org/10.1007/978-3-030-75178-4_1

Norvig, P., & Russell, S. J. (2021). *Artificial Intelligence: a modern approach* (4th ed.). Pearson.

Orca AI. (2022). *Orca AI - Solutions*. Orca AI. Abgerufen am 09/09/2022, von https://www.orca-ai.io/solutions

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). *Hierarchical Text-Conditional Image Generation with CLIP Latents*. https://doi.org/10.48550/arXiv.2204.06125

Riveiro, M., Pallotta, G., & Vespe, M. (2018). Maritime anomaly detection: A review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(5), e1266. https://doi.org/10.1002/widm.1266

Rødseth, Ø. J., Lien Wennersberg, L. A., & Nordahl, H. (2022). Towards approval of autonomous ship systems by their operational envelope. *Journal of Marine Science and Technology*, *27*(1), S. 67-76. https://doi.org/10.1007/s00773-021-00815-z

Saildrone Inc. (2022). *Saildrone Voyager*. Abgerufen am 09/09/2022, von https://assets.website-files.com/5beaf972d32c0c1ce1fa1863/629a8318e699b0b1981d06c4_SD_Voyager-Bathymetry_Product_Card_r8-web final-2206.pdf

Samek, W., & Müller, K. (2019). Towards Explainable Artifical Intelligence. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (pp. 5–22). Springer, Cham. https://doi.org/10.1007/978-3-030-28954-6_1

Samsung Heavy Industries. (o. D.). *SHI SVessel Onboard Solution*. Abgerufen am 12/09/2022, von https://shi.svessel.com/?page_id=298

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Sea Machines Robotics. (2022). *Sea Machines 300*. Abgerufen am 09/09/2022, von https://sea-machines.com/wp-content/uploads/SM-300-Insert-Sheet_2022_final-web.pdf

Seadronix. (2022). *Seadronix - Our Products*. Abgerufen am 09/09/2022, von https://www.seadronix.com/products

Seafar NV. (2022). *Seafar Services*. Abgerufen am 09/09/2022, von https://seafar.eu/services/

Seib, V., Lange, B., & Wirtz, S. (2020). *Mixing Real and Synthetic Data to Enhance Neural Network Training -- A Review of Current Approaches*. https://doi.org/10.48550/arXiv.2007.08781

Skredderberget, A. (2018). *Yara Birkeland - The first ever zero emission, autonomous ship*. Abgerufen am 09/09/2022, von https://www.yara.com/knowledge-grows/game-changer-for-the-environment/

Tsirikoglou, A., Kronander, J., Wrenninge, M., & Unger, J. (2017). *Procedural Modeling and Physically Based Rendering for Synthetic Data Generation in Automotive Applications*. http://arxiv.org/abs/1710.06270

Wang, W., Shan, T., Leoni, P., Fernandez-Gutierrez, D., Meyers, D., Ratti, C., & Rus, D. (2020). Roboat II: A Novel Autonomous Surface Vessel for Urban Environments. *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1740–1747. https://doi.org/10.1109/IROS45743.2020.9340712

Wärtsilä. (2022). *Wärtsilä Advanced Assistance Systems*. Abgerufen am 09/09/2022, von https://www.wartsila.com/voyage/autonomy-solutions/advanced-assistance-systems

Weller, A. (2019). Transparency: Motivations and Challenges. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 11700, Issue Section 2, pp. 23–40). Springer Cham. https://doi.org/10.1007/978-3-030-28954-6_2

Yoshida, M., Shimizu, E., Sugomori, M., & Umeda, A. (2021). Identification of the Relationship between Maritime Autonomous Surface Ships and the Operator's Mental Workload. *Applied Sciences*, *11*(5), 2331. https://doi.org/10.3390/app11052331

Zhang, C., Kuppannagari, S. R., Kannan, R., & Prasanna, V. K. (2018). Generative Adversarial Network for Synthetic Time Series Data Generation in Smart Grids. *2018 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, 1–6. https://doi.org/10.1109/SmartGridComm.2018.8587464

Žliobaitė, I. (2010). *Learning under Concept Drift: an Overview*. https://doi.org/10.48550/arXiv.1010.4784

A Anhang

A.1. Ergebnisse der Marktanalyse

Tabelle 5: In Marktanalyse gesichtete Unternehmen oder Produkte und ihre Datenquellen.

Unternehmen / Produkt	Quellen	RGB-Kamera	Infrarot-Kamera	LIDAR / RADAR	IMU / MRU	GNSS	AIS	Wetter-daten / -sensorik	Tiefen-messung
ABB Ability Marine Pilot Vision	(ABB, 2018)	x	x	x	x	x	x		
Avikus AiBOAT	(Avikus, o. D.-a)	x		x		x	x		
Avikus HiNAS	(Avikus, o. D.-b)	x	x	x		x	x		
Captain AI	(Captain AI, 2022)	x		x		x	x		
Kongsberg Situation Awareness	(Kongsberg, 2022)	x		x	x	x	x		
Kongsberg Maritime Autonomous Shipping	(Kongsberg, 2019)	x		x	x	x	x		
Marine AI Guardian Autonomy	(Marine AI Ltd, 2022)	x		x			x	x	
Mayflower Autonomous Ship[1]	(Mayflower Autonomous Ship, 2022)	x		x	x	x	x	x	x
Orca AI	(Orca AI, 2022)	x	x						
Oscar	(BSB AI, 2022)	x	x		x	x			
Roboat	(Wang et al., 2020)	x		x	x	x			
Saildrone	(Saildrone Inc., 2022)	x			x	x	x	x	x
Sea Machines SM300	(Sea Machines Robotics, 2022)	x		x	x	x	x		x
Seadronix AVISS	(Seadronix, 2022)	x							
Seafar	(Seafar NV, 2022)	x		x	x	x	x		
SVessel Samsung Heavy Industries	(Samsung Heavy Industries, o. D.)	x					x		
Wärtsilä Voyage Autonomy Solutions	(Wärtsilä, 2022)	x	x	x		x	x	x	x
Yara Birkeland[1]	(Kongsberg, 2017; Skredderberget, 2018)	x	x	x	x	x	x		

[1] Vereint KI-basierte Produkte als Systemintegrator.

Tabelle 6: In Marktanalyse gesichtete Unternehmen oder Produkte und ihre Anwendungsfälle.

Unternehmen / Produkt	COLREGs-Evaluation	Kollisions-vermeidung	Hindernis- und Küsten-erkennung	Schiffs-erkennung	Routen-planung	Andock- und Ablege-Assistenz
ABB Ability Marine Pilot Vision	x		x	x		
Avikus AiBOAT	x	x	x	x	x	x
Avikus HiNAS				x	x	
Captain AI		x		x	x	
Kongsberg Situation Awareness			x	x		x
Kongsberg Maritime Autonomous Shipping	x	x	x	x	x	
Marine AI Guardian Autonomy		x	x	x	x	
Mayflower Autonomous Ship[1]	x	x	x	x	x	
Orca AI		x	x	x		
Oscar	x	x	x	x		
Roboat		x	x	x		
Saildrone	x	x	x	x	x	
Sea Machines SM300	x	x	x	x		
Seadronix AVISS			x	x		x
Seafar	x	x	x	x	x	
SVessel Samsung Heavy Industries				x		
Wärtsilä Voyage Autonomy Solutions			x			x
Yara Birkeland[1]	x	x	x	x	x	x

[1] Vereint KI-basierte Produkte als Systemintegrator.