

# KIEZ 4.0

## (Künstliche Intelligenz Europäisch Zertifizieren unter Industrie 4.0 Aspekten)



### Teilvorhaben Fraunhofer

Gefördert durch:



aufgrund eines Beschlusses  
des Deutschen Bundestages

### **ABSCHLUSSBERICHT** **FKZ: 20D1914I**

Laufzeit des Vorhabens: 01.09.20-31.12.23

Sujan Gannamaneni (IAIS)

Viraj Gala (FOKUS)  
Jürgen Großmann (FOKUS)  
Martin Schneider (FOKUS)

Joanna Polewczyk (FKIE)  
Arne Schwarze (FKIE)  
Felix Govaers (FKIE)

Wachtberg, Mai 2024



# Inhaltsverzeichnis

|  |           |
|--|-----------|
| <b>I. Kurzdarstellung .....</b>  | <b>3</b>  |
| <b>I.1 Aufgabenstellung .....</b>  | <b>3</b>  |
| <b>I.2 Planung und Ablauf des Vorhabens .....</b>  | <b>3</b>  |
| I.2.1 Aufgaben in Hauptarbeitspaket 1 .....  | 4         |
| I.2.2 Aufgaben in Hauptarbeitspaket 2 .....  | 5         |
| I.2.3 Aufgaben in Hauptarbeitspaket 3 .....  | 6         |
| I.2.4 Aufgaben in Hauptarbeitspaket 4 .....  | 7         |
| <b>I.3 Wissenschaftlicher und technischer Stand .....</b>                                | <b>9</b>  |
| <b>I.4 Zusammenarbeit mit anderen Stellen .....</b>                                      | <b>12</b> |
| <b>II. Eingehende Darstellung .....</b>  | <b>13</b> |
| <b>II.1 Projektergebnisse und Verwertung .....</b>                                       | <b>13</b> |
| II.1.1 Ergebnisse in Hauptarbeitspaket 1.....  | 13        |
| II.1.1.1. Synthese des bestehenden Know-hows für die Zertifizierung (AP 1.1) .....       | 13        |
| II.1.1.1.1. Vorgehen .....   | 13        |
| II.1.1.1.2. Ergebnisse .....   | 13        |
| II.1.1.2. Zertifizierungskonzepte für KI (AP 1.2).....                                   | 13        |
| II.1.1.2.1. Vorgehen .....   | 13        |
| II.1.1.2.2. Ergebnisse .....   | 6         |
| II.1.2 Ergebnisse in Hauptarbeitspaket 2.....  | 14        |
| II.1.2.1. Potentiale für Luftfahrzeug, Flugsicherung und Flughafen (AP 2.1) .....        | 14        |
| II.1.2.1.1. Vorgehen .....   | 14        |
| II.1.2.1.2. Ergebnisse .....   | 14        |
| II.1.2.2. Prototypen mit KI für Luftfahrzeug, Flugsicherung und Flughafen (AP 2.2) ..... | 14        |
| II.1.2.2.1. Vorgehen .....   | 14        |
| II.1.2.2.2. Ergebnisse .....   | 14        |
| II.1.3 Ergebnisse in Hauptarbeitspaket 3.....  | 15        |
| II.1.3.1. Infrastruktur (AP 3.1).....  | 7         |
| II.1.3.1.1. Vorgehen .....   | 7         |
| II.1.3.1.2. Ergebnisse .....   | 7         |
| II.1.3.2. Komponenten mit KI (AP 3.2).....   | 15        |
| II.1.3.2.1. Vorgehen .....   | 15        |
| II.1.3.2.2. Ergebnisse .....   | 15        |

|   |           |
|---|-----------|
| II.1.3.3. Komplexe Anwendungen (AP 3.3) .....   | 15        |
| II.1.3.3.1. Vorgehen .....  | 15        |
| II.1.3.3.2. Ergebnisse .....  | 15        |
| II.1.3.4. Systems of Systems Anwendungen (AP 3.4) .....   | 16        |
| II.1.3.4.1. Vorgehen .....  | 16        |
| II.1.3.4.2. Ergebnisse .....  | 16        |
| II.1.4 Ergebnisse in Hauptarbeitspaket 4 .....  | 17        |
| II.1.4.1. Testverfahren (AP 4.1) .....  | 17        |
| II.1.4.1.1. Vorgehen .....  | 17        |
| II.1.4.1.2. Ergebnisse .....  | 17        |
| II.1.4.2. Verifikation und Validierung (AP 4.2) .....   | 17        |
| II.1.4.2.1. Vorgehen .....  | 17        |
| II.1.4.2.2. Ergebnisse .....  | 17        |
| II.1.4.3. Methodik der Zertifizierung (AP 4.3) .....  | 18        |
| II.1.4.3.1. Vorgehen .....  | 18        |
| II.1.4.3.2. Ergebnisse .....  | 18        |
| II.1.4.4. Abgleich mit EASA (AP 4.4) .....  | 18        |
| II.1.4.4.1. Vorgehen .....  | 18        |
| II.1.4.4.2. Ergebnisse .....  | 18        |
| II.1.4.5. Neue Methodik für die Zertifizierung von KI (AP 4.5) .....  | 19        |
| II.1.4.5.1. Vorgehen .....  | 19        |
| II.1.4.5.2. Ergebnisse .....  | 19        |
| <b>II.2 Wichtigste Positionen des zahlenmäßigen Nachweises.....</b>   | <b>9</b>  |
| <b>II.3 Notwendigkeit und Angemessenheit der geleisteten Arbeit .....</b>   | <b>20</b> |
| <b>II.4 Voraussichtlicher Nutzen und Verwertbarkeit der Ergebnisse im Sinne des fortgeschriebenen Verwertungsplanes .....</b> | <b>20</b> |
| <b>II.5 Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen.....</b>   | <b>20</b> |
| <b>II.6 Erfolgte oder geplante Veröffentlichungen des Ergebnisses.....</b>  | <b>21</b> |

# I. Kurzdarstellung

Das Projekt KIEZ 4.0 wurde ab 1. September 2020 vom Bundesministerium für Wirtschaft und Energie (BMWi), das am 8. Dezember 2021 in Bundesministerium für Wirtschaft und Klimaschutz (BMWK) umbenannt wurde, im Rahmen des ersten Programmaufrufs des sechsten nationalen zivilen Luftfahrtforschungsprogramms (LuFo VI-1) gefördert. Die Bearbeitung erfolgte in einem Konsortium mit den folgenden Partnern:

- Airbus
- Deutsche Flugsicherung (DFS)
- fortiss
- Universität Stuttgart
- ILS
- Prof. Elmar Giemulla
- Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V., München

Unter dem Förderkennzeichen 20D1914I hat das Fraunhofer-Konsortium bestehend aus den drei Fraunhofer Instituten für

- Offene Kommunikationssysteme FOKUS,
- Intelligente Analyse- und Informationssysteme IAIS
- Kommunikation, Informationsverarbeitung und Ergonomie FKIE.

an dem Förderprojekt teilgenommen. Dieser Abschlussbericht beschreibt die Arbeiten und Ergebnisse dieser Institute.

## I.1 Aufgabenstellung

Die Arbeiten der Fraunhofer Institute in den verschiedenen Arbeitspaketen des KIEZ 4.0-Konsortiums konzentrierten sich auf die zentralen Herausforderungen und Anforderungen für die Zertifizierung von KI/ML-Systemen in der Luft- und Raumfahrt. Dazu gehört die Entwicklung eines Demonstrators für die Flugbahnvorhersage und die Schätzung der Ankunftszeit von Flugzeugen, die auf einem Flughafen landen sollen, mit dem Ziel, Probleme im Zusammenhang mit dem Informationsfluss an den Flughafen bei Flugplanänderungen zu lösen. Als Teil des Zertifizierungsprozesses des Demonstrators, der sich am EASA-Konzeptpapier orientiert, wurden Ansätze zur Bewertung der Robustheit durch gegnerische Angriffe und Unsicherheitsabschätzungen, Datenqualitätstests, Erklärbarkeitstechniken und andere Bewertungsmethoden entwickelt und umgesetzt. Diese Ansätze adressieren spezifische Ziele, die im EASA-Konzeptpapier umrissen wurden, und tragen zum Aufbau einer umfassenden Zertifizierungsmethodik für KI-basierte Systeme in der Luft- und Raumfahrt bei.

## I.2 Planung und Ablauf des Vorhabens

Abbildung 1 zeigt eine schematische Darstellung der Arbeitspakete des gesamten Vorhabens. Dieses gliedert sich in die vier Hauptarbeitspakete (HAP)

- Konzepte
- Anwendungen von KI
- KI Demos / Use Cases
- Analyse

In HAP 1 wurden die bestehenden Konzepte und Ansätze zur Zertifizierung von KI zusammengetragen, eingeordnet und bewertet. HAP 2 hat im Anschluss daran existierende und potentielle Anwendungen von KI im Luftfahrtbereich betrachtet und exemplarische Use Cases hinsichtlich einer Zertifizierbarkeit

untersucht. Diese wurden im HAP 3 experimentell als Software realisiert. Die Analyse der Umsetzbarkeit von Methoden der Zertifizierung sowie deren Bewertung auf Basis der implementierten Demonstratoren war der Fokus von HAP 4. Die Fraunhofer Institute haben zu allen HAPs beigetragen und waren verantwortlich für HAP 4.

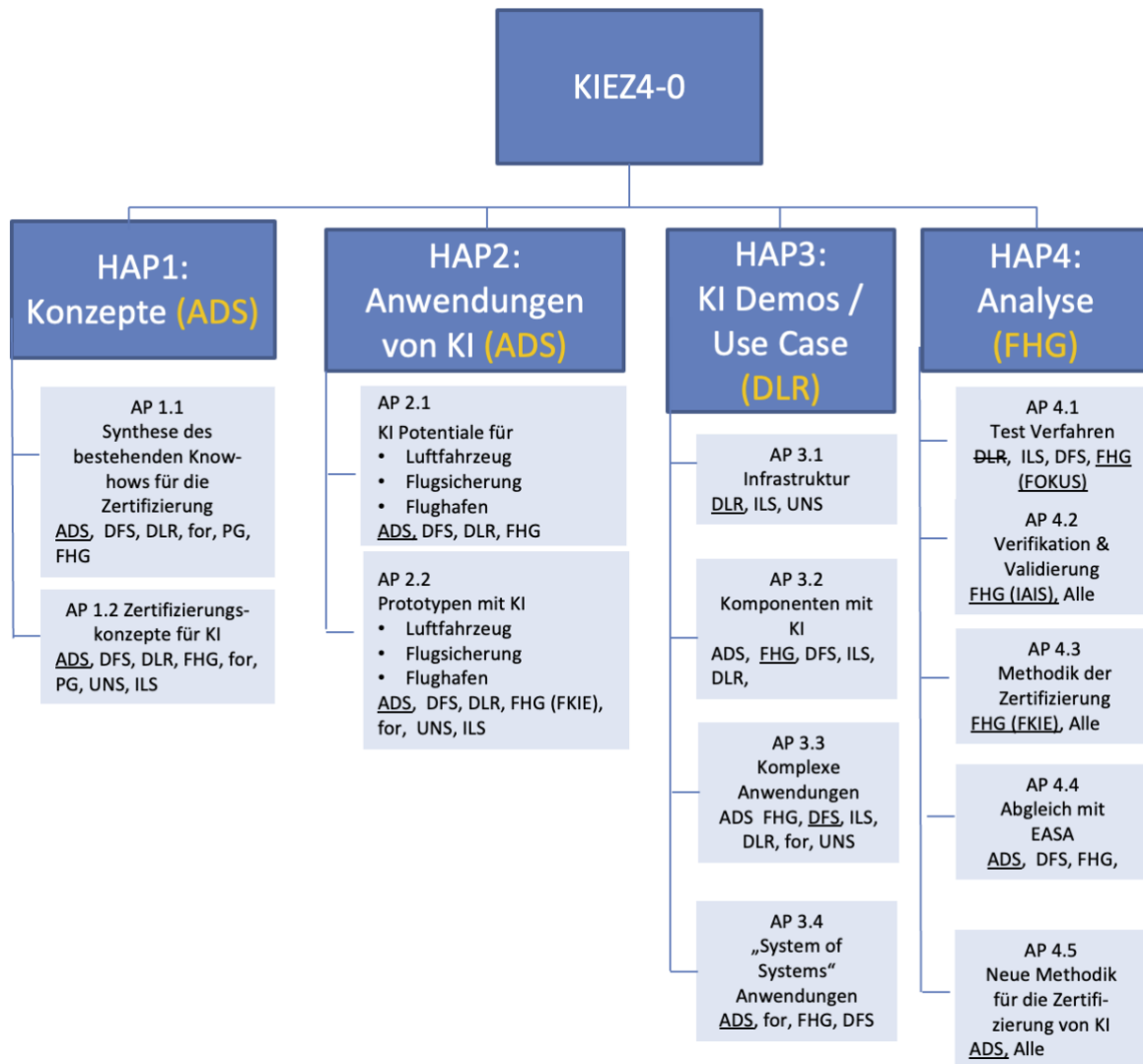


Abbildung 1: Arbeitspaketstruktur des Verbundvorhabens

### 1.2.1 Aufgaben in Hauptarbeitspaket 1

Tabelle 1: Teilarbeitspakete in HAP 1

|        |   |
|--------|---|
| HAP 1  | Konzepte  |
| AP 1.1 | Synthese des bestehenden Know-hows für die Zertifizierung |
| AP 1.2 | Zertifizierungskonzepte für KI                            |

Tabelle 1 zeigt eine Übersicht der Teilarbeitspakete in HAP 1 »Konzepte«, die von Fraunhofer bearbeitet wurden.

In HAP1 wurde vorhandenes Know-how zusammengeführt und neue Ansätze von Konzepten für die Zertifizierung unter Berücksichtigung des Stands der Technik entwickelt. Fraunhofer hat entlang der verschiedenen KI-Algorithmen neue, spezifische Risiken und Fehlerszenarien identifiziert, die durch

den Einsatz von KI in der Luftfahrt entstehen können. Auf Basis der identifizierten Risiken und Fehlerszenarien wurden KI-spezifische Qualitätsmerkmale bzw. KI-spezifische Ausprägungen bestehender Qualitätsmerkmale definiert und hinsichtlich ihrer Bedeutung für die Zertifizierung in der Luftfahrt eingeordnet. Risiken, Fehlerszenarien und Qualitätsmerkmale dienen dann in weiteren Arbeiten als systematische Grundlage für die Ableitung von Test- und Zertifizierungsstrategien. Der Schwerpunkt der Arbeiten in HAP1 lag weiterhin auf der Entwicklung neuer Ansätze, die von der bestehenden, auf Abdeckungskriterien basierenden Zertifizierungsmethodik im Hinblick auf die Qualitätssicherung abweichen und die Eigenschaften verschiedener KI-Methoden, -Techniken und -Algorithmen, deren jeweilige technische Spezifika, z.B. hinsichtlich Determinismus und Art des Lernens (überwacht, verstärkt, unüberwacht) und der Notwendigkeit der Pflege des zugrundeliegenden Modells, sowie Artefakte und Prozesse, die derzeit nicht im Fokus der Zertifizierung stehen, berücksichtigen. Besonderes Augenmerk wurde dabei auf die Lernprozesse und die beim Lernen verwendeten Datensätze gelegt, deren Qualität einen direkten Einfluss auf die spätere Systemfunktionalität hat.

Ziele AP 1.1:

- Anforderungskatalog hinsichtlich der Eignung für eine KI-Zertifizierung zusammentragen.
- Aktuelle Tendenzen wie z.B. der SORA-Prozess (Safety Oriented Risk Assessment) berücksichtigen.
- Prozess-relevante Vorschriften analysieren (wie z.B. EU-äquivalent APR 4761/4754A)

Ziele AP 1.2:

- Auf Basis von AP1.1 werden Konzepte mit den Partnern erstellt, die darauf ausgerichtet sind, KI zu zertifizieren. Verschiedene Entwicklungskonzepte (wie z.B. Reactive Model Based Programming auf der Basis temporaler Netzwerke oder On-Line-Monitoring nicht deterministischer Algorithmen) werden hinsichtlich der Luftfahrt-Entwicklungsstandards DO-178C (für Software) und DO254 (für Hardware) untersucht.

Es werden entlang der unterschiedlichen KI-Algorithmen neue, spezifische Risiken und Fehlerszenarien identifizieren, die durch den Einsatz von KI im Rahmen der Luftfahrt entstehen können.

## I.2.2 Aufgaben in Hauptarbeitspaket 2

Tabelle 2: Teilarbeitspakete in HAP 2

|        |   |
|--------|---|
| HAP 2  | Anwendungen von KI  |
| AP 2.1 | KI Potentiale für Luftfahrzeug, Flugsicherung und Flughafen     |
| AP 2.2 | Prototypen mit KI für Luftfahrzeug, Flugsicherung und Flughafen |

Tabelle 2 zeigt eine Übersicht der Teilarbeitspakete in HAP 2 »Anwendungen«, die von Fraunhofer bearbeitet wurden.

In HAP2 wurden Potentiale von KI-Methoden für die Luftfahrt, die aufgrund von rechtlichen oder sicherheitstechnischen Vorbehalten im Hinblick auf die Zulassung und Zertifizierung derzeit noch nicht in der Praxis eingesetzt werden können, ausgelotet und unter Reflexion der Ergebnisse aus HAP1 auf ihre Umsetzbarkeit hin diskutiert. Dabei wurden insbesondere Anwendungen für Flugsteuerung und Automatisierung von Luftfahrzeugen, Assistenzsysteme für die Flugsicherung und Luftraumüberwachung sowie Optimierungsunterstützung für Flughäfen untersucht. Fraunhofer hat dabei vor dem Hintergrund seiner Erfahrungen im Bereich der KI für Sensordatenfusion, Prozessoptimierung und Entscheidungsunterstützung mit der Bewertung von möglichen Anwendungen beitragen.

Ziele AP 2.1:

- In diesem Arbeitspaket werden für die Luftfahrzeuge und Air Traffic Management Systeme sowie Flughäfen die Potenziale der KI-Anwendungen identifiziert und auf Eignung geprüft.
- FHG wird Potentiale für den Bereich der Flughäfen zu Verbesserungen durch KI-Anwendungen identifizieren und auf Umsetzbarkeit prüfen.

Ziele AP 2.2:

- Es wird eine Anwendung mit KI für den Bereich der Flughäfen in Bezug auf ihre Zertifizierbarkeit untersucht sowie ein eigener Demonstrator eingebracht.

### 1.2.3 Aufgaben in Hauptarbeitspaket 3

Tabelle 3: Teilarbeitspakete in HAP 3

|        |                                |
|--------|--------------------------------|
| HAP 3  | KI Demos / Use Case            |
| AP 3.2 | Komponenten mit KI             |
| AP 3.3 | Komplexe Anwendungen           |
| AP 3.4 | Systems of Systems Anwendungen |

Tabelle 3 zeigt eine Übersicht der Teilarbeitspakete in HAP 3 »KI Demos / Use Case«, die von Fraunhofer bearbeitet wurden.

Grundlage für Use Cases für die Infrastruktur des Luftverkehrs waren Arbeiten zum Aircraft Stand Assignment, also der Zuweisung von Abstellpositionen an Flughäfen. Als Komponente von KI wurde ein Fraunhofer-Demonstrator entwickelt, der auf Basis angelernter Information Trajektorien von zivilen Flugzeugen vorhersagen konnte. Dabei wurden auch Einflüsse von Wetterlagen und Ereignissen im Luftverkehr auf die Vorhersage untersucht. Der Demonstrator eignete sich als komplexe Anwendung im Zusammenspiel mit dem Arrival Manager (AMAN) von DLR und DFS und als Systems-of-Systems im Verbund mit der Routenplanung von Airbus. Anhand des Demonstrators wurden in HAP 4 Verfahren und Konzepte exemplarisch durchgeführt und evaluiert.

Ziele AP 3.2:

- In AP 3.2 wird der Demonstrator „Ressourcenzuweisung an Flughäfen“ in Zusammenarbeit mit dem Konsortium technisch aufbereitet und in Bezug auf die Optimierungspotentiale zur Nachvollziehbarkeit und Zertifizierbarkeit aus AP 2.2 erweitern. So ist unter anderem vorgesehen, dieselbe Datenbasis für unterschiedliche KI-Methoden zu nutzen und gegeneinander zu vergleichen.

Ziele AP 3.3:

- Bei den KI-Funktionen zugrunde liegenden maschinellen Lernverfahren wird unterschieden zwischen Trainings-, Test- und Inferenzphase (Anwendungsphase). In der Trainingsphase wird anhand geeigneter Trainingsdaten ein gewähltes Modell iterativ auf die anwendungsrelevanten Zusammenhänge ausgeprägt und anschließend getestet; während der Modellanwendung bleibt dieses jedoch üblicherweise statisch, so dass getestete Modelleigenschaften unverändert gültig bleiben. Unter einer komplexen KI-Anwendung wird in AP3.3. dagegen verstanden, dass die ML-Komponente auch im operativen Betrieb –

anhand der dort anfallenden Daten und Muster – mit dem Ziel der konstanten Selbstverbesserung weiter lernt. Entsprechend muss sich für diese KI-Anwendungsklasse eine Absicherung auf Aspekte des Monitoring und Alerting im Operativbetrieb erstrecken. Ziel dieses AP ist eine konzeptionelle Betrachtung der notwendigen Testszenarios und -umgebungen.

Ziele AP 3.4:

- Konzeption einer Systems-of-Systems-Anwendung, bei der eine Integration der Demonstration in ein Gesamtsystem vorgesehen ist. Dabei wird der FHG-Demonstrator aus AP 3.2 zur Ressourcenzuweisung an Flughäfen logisch mit dem AMAN-Demonstrator vom DLR zusammengeführt für eine automatisierte und intelligente Planung an Flughäfen.

#### I.2.4 Aufgaben in Hauptarbeitspaket 4

Tabelle 4: Teilarbeitspakete in HAP 4

|        |   |
|--------|---|
| HAP 4  | Analyse                                     |
| AP 4.1 | Testverfahren                               |
| AP 4.2 | Verifikation und Validierung                |
| AP 4.3 | Methodik der Zertifizierung                 |
| AP 4.4 | Abgleich mit EASA                           |
| AP 4.5 | Neue Methodik für die Zertifizierung von KI |

Tabelle 4 zeigt eine Übersicht der Teilarbeitspakete in HAP 4 »Analyse«.

Bei der Entwicklung von KI-basierten Algorithmen stehen in der gegenwärtigen Forschung meist Kriterien der Leistungsfähigkeit im Vordergrund, z.B. in Bezug auf Accuracy, Precision und Recall. Im Hinblick auf die kritischen Sicherheitsanforderungen darf in der Luftfahrt jedoch die Optimierung auf solche Kriterien nicht das einzige Entwicklungsziel sein. Einerseits beantworten Performanzkriterien naturgemäß nicht die Frage, wie mit dem (oft niedrigen) Anteil von Fehlerkennungen umgegangen werden soll. Andererseits müssen zur Gewährleistung der funktionalen und Gebrauchssicherheit auch andere Kriterien in den Fokus der KI -Algorithmen-Entwicklung rücken, die weniger intuitiv und v.a. kaum etabliert sind.

In HAP4 lagen die Schwerpunkte der Arbeiten entsprechend in der Untersuchung der bisherigen Testterminologie, Qualitätsattributen, Bewertungskriterien, und -Techniken und deren Adaption auf KI-Systeme im Hinblick auf ihre Verwendbarkeit in einem Zertifizierungsrahmenwerk. Testterminologie, Qualitätsattribute und Bewertungskriterien wurden mit bestehenden Standards wie ISO 29119, ISO 25010 sowie in aktuellen Standardisierungsaktivitäten von Fraunhofer wie der DIN SPEC 92001 des DIN-Ausschusses KI abgestimmt.

Auf Basis dieses Rahmenwerks und, ausgehend von den in HAP1 abgeleiteten Risiken, potenziellen Fehlerszenarien und Qualitätsattributen wurden von Fraunhofer im HAP4 eine Reihe von Testtechniken und -verfahren entwickeln, die geeignet sind, für KI-Algorithmen spezifischer Fehler zu erkennen und die geeignet sind, die Qualität von KI-Systemen bzw. Systemen mit KI-Funktionalität im Rahmen der Zertifizierung beleg- und bewertbar zu machen. Im Fokus der Arbeit standen neben der Prüfung von funktionalen Systemeigenschaften insbesondere die Prüfung nichtfunktionaler Eigenschaften wie Zuverlässigkeit, Robustheit und Cyber-Security sowie die Anwendung von Testverfahren und Techniken (bspw. Adversarial Machine Learning), die spezifische Eigenschaften von KI-Systemen wie beispielsweise Nichtdeterminismus und Instabilität adressieren. Weiterhin wurden

Verfahren und Techniken entwickelt, um die Prüfung von KI-spezifischen Artefakten wie Trainings- und Testdaten zu berücksichtigen. Des Weiteren wurde untersucht, in wie weit Methoden der lokalen oder globalen Erklärbarkeit von KI-Algorithmen als Bausteine eines generischen Zertifizierungsprozesses eingesetzt werden können. Dies wurde insbesondere auf der Basis von Ansätzen der „erklärbaren KI“ (XAI) durch Näherung mittels einfacher, nachvollziehbarer Methoden, durch Visualisierung von Datenflüssen und Strukturen sowie durch Sensitivitätsanalysen im Hinblick auf Klassifikationsresultate betrachtet.

Aufbauend auf diesen Ergebnissen haben die Fraunhofer Institute zu einer neuen Zertifizierungsmethodik für KI-Anwendungen im Luftfahrtbereich beitragen

#### Ziele AP 4.1:

- Hauptziel dieses Arbeitspakets ist die Weiterentwicklung von Testverfahren aus AP 4.1 im Hinblick auf eine Eignung für Zertifizierungsrahmenwerk.
- Untersuchung der bisherigen Testterminologie, Qualitätsattributen, Bewertungskriterien, und Techniken und deren Adaption auf KI-Systeme im Hinblick auf ihre Verwendbarkeit in einem Zertifizierungsrahmenwerk unter Berücksichtigung bestehender und in Entwicklung befindlicher Standards;
- Entwicklung einer Terminologie und Kriterien zur Bewertung von KI-Systemen in engem Austausch mit der DIN SPEC 92001 und unter Berücksichtigung der ISO 25010 und ISO 29119;
- Identifikation, Adaption und Weiterentwicklung von Testverfahren für Qualitätsbewertung von KI-basierten Systemen, unter Berücksichtigung der identifizierten relevanten Qualitätsattribute, sowie der in HAP1 identifizierten Gefahrenpotenziale und Risiken

#### Ziele AP 4.2:

- Hauptziel dieses Arbeitspakets ist die Anwendung der Testverfahren aus AP 4.1 auf die jeweiligen Use Cases im Rahmen eines V&V-Konzepts entlang der Use Cases aus AP 2.2.  
Arbeitsschritte:
  - Erstellung eines V&V-Konzepts für die jeweiligen Use Cases (AP 2.2)
  - Generierung von synthetischen Trainings- und Validierungsdaten unter Einbezug von bewussten Verunreinigungen für bestimmte kritische Situationen in den jeweiligen Use Cases (AP 2.2)
  - Definition eines hinreichenden Satzes an Verkehrsszenarien entsprechend der Testverfahren aus AP 4.1 für die jeweiligen Use Cases (AP 2.2)
  - Test funktionaler und nicht-funktionaler Anforderungen der KI-Komponenten nach den Qualitätsattributen und Kriterien aus AP 4.1 für die jeweiligen Use Cases (AP 2.2)

#### Ziele AP 4.3:

- Hauptziel dieses Arbeitspaket ist es, eine Methodik der Zertifizierung mit Hilfe von Indikatoren, Testverfahren und KPI zur Entwicklung eines Wirksamkeitsnachweises der Absicherungsmethoden von KI-basierten Systemen zu schaffen:
  - Neuartige Indikatoren, Testverfahren und Testmethoden und Entwicklung von standardisierbaren Tests und Testanforderungen;
  - Definition von KPI zur Bestimmung der Wirksamkeit der KI-spezifischen Analyse-Methoden und –Maßnahmen;
  - Entwicklung einer Argumentation für den Einsatz von Testmethoden zum Wirksamkeitsnachweis der Kombination von einzelnen Absicherungsmethoden, -maßnahmen und Nachweisstrategien zu einem ganzheitliche Argument.

#### Ziele AP 4.4:

- Hauptziel ist es, die in AP 4.3 erarbeitete Methodik der Zertifizierung mit der EASA abzugleichen. Die Fraunhofer Institute diskutieren im Rahmen von AP4.4 mit der EASA begleiten und mit seiner Expertise unterstützen.

Ziele AP 4.5:

- Hauptziel ist es, die in AP 4.3 erarbeitete Methodik der Zertifizierung mit der EASA abzugleichen. Fraunhofer wird im Rahmen von AP4.5 neuartige Methoden der Erklärbarkeit für KI-Algorithmen durch Sensitivitätsanalysen, Visualisierungen sowie Analysen der Trainingsdaten in Bezug auf Robustheit erarbeiten.

### I.3 Wissenschaftlicher und technischer Stand

Die Forschung zu spezifischen Methoden zur Verifizierung und Validierung von ML steht trotz nicht unerheblicher Anstrengungen unterschiedlicher Fach- und Industriezweige derzeit am Anfang. Dennoch ist das Testen bereits Teil des Trainings, wobei die meisten Tests durchgeführt werden, um präzisere Modelle im Hinblick auf die ursprünglichen Trainingszielen zu erreichen. Beim überwachten Lernen beispielsweise werden Test- und Validierungsdatensätze verwendet, um eine unvoreingenommene Bewertung des ML-Modells zu ermöglichen. Validierungsdatensätze werden typischerweise während des Trainings zur Feinabstimmung der Modellparameter verwendet, während Testdatensätze am endgültigen Modell verwendet werden, um Generalisierungsfehler zu messen. Da die einzelnen Testsätze jedoch nur eine einzige Bewertung des Modells liefern und nur begrenzt in der Lage sind, die Unsicherheit in den Ergebnissen zu charakterisieren, werden für die Modellauswahl fortgeschrittenere statistische Testansätze wie Cross-Validation verwendet.

Klassische Verifizierungs- und Validierungsansätze, die in der Informatik schon lange Gegenstand der Forschung sind, wie Testen, Modellüberprüfung und Theorembeweis, sind bei der Verifizierung unvollständiger Spezifikationen im Allgemeinen begrenzt. In den meisten Fällen ist es entweder unpraktisch oder unmöglich, ML mit vorhandenen Werkzeugen und Techniken formal zu verifizieren<sup>[1]</sup>. Gegenwärtig wird an mehreren Stellen an der Anpassung von Werkzeugen zur Modellüberprüfung gearbeitet, wie z.B. Solver für die Satisfiability-Modulo-Theories (SMT) und Solver für die Mixed Integer Programming (MIP) zur Verifikation neuronaler Netze<sup>[2],[3]</sup>. In vereinfachter Form wird versucht, qualitative Aussagen über den Zusammenhang bestimmter Bereiche des Eingabedatenraums mit bestimmten Bereichen des Ausgangsraums eines neuronalen Netzes zu treffen. Gosh et. al.<sup>[4]</sup> kombinieren ML und Modellverifikation so, dass, wenn die gewünschten logischen Eigenschaften durch ein trainiertes Modell nicht erfüllt werden, das Modell ("Modellreparatur") oder die Daten, aus denen das Modell gelernt wird, systematisch modifiziert werden ("Datenreparatur"). Fulton et al.<sup>[5]</sup> schlagen vor, formale Verifikation mit verifizierter Laufzeitüberwachung zu kombinieren, so dass ein sicheres Lernen gewährleistet werden kann. Der Ansatz greift immer dann in den Lernprozess ein, wenn Sicherheitseigenschaften verletzt werden, und lenkt den Lernprozess so, dass das Ergebnis mit dem Verifikationsmodell konform ist. DeepXplore<sup>[6]</sup>, DLFuzz<sup>[7]</sup> und TensorFuzz<sup>[8]</sup> stellen verschiedene Metriken für die Quantifizierung der neuronalen Abdeckung zur Verfügung und vereinfachen die Testautomatisierung.

Darüber hinaus zeigen neuere Forschungen zum Adversarial Machine Learning<sup>[9]</sup>, dass nicht-funktionale Qualitätseigenschaften wie Security und Robustheit mehr Aufmerksamkeit benötigen. Pei et. al.<sup>[10]</sup> schlagen einen White-Box-Ansatz für tiefe neuronale Netze vor, der auf der Idee der Neuronenabdeckung und des differenziellen Testens basiert, um fehlerhafte Ausnahmeverhalten zu identifizieren, bei denen Klassifikationen in verschiedenen ML-Modellen zueinander umkehren. Carlini et al.<sup>[11]</sup> geben konkrete Empfehlungen und beschreiben Werkzeuge, um die Robustheit von ML-Systemen zu evaluieren. Forschungsbedarf besteht auch hinsichtlich der Metriken und Abbruchbedingungen. Dazu existieren zwar schon verschiedene Ansätze, bspw. zur neuronalen Überdeckung<sup>[12]</sup>, diese führen jedoch nicht per se zu einer verbesserten Robustheit<sup>[13]</sup>.

DeepTest<sup>[14]</sup> ermöglicht das systematische Testen von NN unter realistisch wechselnden Umgebungsbedingungen, insbesondere für den Einsatz im Automobilbereich. IBM stellt eine Reihe von Open-Source-Tools und Metriken zur Verfügung, die es ermöglichen, Datensätze und maschinelle Lernmodelle auf unerwünschte Verzerrungen zu prüfen<sup>[15]</sup> und Vertrauens- und Transparenzprüfungen für KI in der IBM Cloud durchzuführen<sup>[16]</sup>.

Das Testen hat Einschränkungen hinsichtlich der Dynamik von ML, der schieren Größe der Problemdomäne und des zugrunde liegenden Orakelproblems<sup>[17]</sup>. Der Dynamik von ML kann durch kontinuierliche und Online-Testansätze Rechnung getragen werden. Das Orakelproblem und die Größe der Problemdomäne werden derzeit durch metamorphe Tests angegangen. Murphy et. al.<sup>[18]</sup> und Xie<sup>[19]</sup> schlagen vor, metamorphes anzuwenden. Beim metamorphen Testen wird das zu testende System (eine DNN) gegen eine Beziehung (z. B. eine Ungleichheit) zwischen verschiedenen Paaren von DNN-Prädiktionen geprüft. Eine solche Beziehung wird als metamorphe Beziehung (MR) bezeichnet und beschreibt, wie die Ausgabe je nach den am Eingang vorgenommenen Änderungen variieren würde. MRs stellen eine leistungsstarke Technik zur Erstellung domänenbezogener Testfälle ohne menschliche Expertenunterstützung dar und könnten eine praktikable Option bei der Validierung von Modellen für Deep Learning darstellen.

### **Normungs- und Standardisierungsarbeiten zu KI/ML**

Der Bedarf, die Absicherung von KI-Funktionalität zu normieren und zu standardisieren, ist domänenübergreifend. So gibt es bereits Standards, welche die verschiedenen Aspekte von Qualität der KI adressieren, viele Standards befinden sich zurzeit in der Entwicklung:

- „ISO/IEC AWI 23894: Information Technology – Artificial Intelligence – Risk Management“ definiert Leitlinien, um das Risiko zu managen, mit dem die Organisationen bei der Entwicklung und beim Einsatz der KI konfrontiert sind.
- „ISO/IEC NP TR 24027 Information technology – Artificial Intelligence (AI) – Bias in AI systems and AI aided decision making“ adressiert Vorurteile/Bias in Relation zu KI-Systemen. Es werden Methoden und Techniken vorgestellt, um die Anfälligkeit für Vorurteile im gesamten KI-Lebenszyklus zu reduzieren.
- „ISO/IEC PDTR 24028 Information technology – Artificial Intelligence (AI – Overview of trustworthiness in Artificial Intelligence)“ identifiziert Bereiche zur Standardisierung, die die Vertrauenswürdigkeit von KI-Systemen adressieren, und beschreibt Vorgehen, um die Bedenken bzgl. der Vertrauenswürdigkeit zu reduzieren.
- „ISO/IEC NP TR 24029-1 Artificial Intelligence (AI) – Assessment of the robustness of neural networks – Part 1“ beschreibt den Hintergrund über existierende Methoden, welche die Robustheit von neuronalen Netzen sicherstellen.
- „ISO/IEC AWI TR 24368 Information technology – Artificial intelligence – Overview of ethical and societal concerns“ gibt einen Überblick über ethische und gesellschaftliche Bedenken beim Einsatz von KI-Systemen.
- „ISO/IEC TR 29119-11 Software and Systems Engineering — Software Testing — Part 11: Guidelines on the testing of AI-based systems“ beschreibt aus ganzheitlicher Sicht das Testen von KI-Systemen, u.a. Qualitätscharakteristika von KI-Systemen in Ergänzung zur ISO 25010, Workflows, Qualitätsprobleme und Herausforderungen beim Test von KI-Systemen sowie spezifische Testtechniken und Metriken.
- Das DIN hat den Standard DIN SPEC 92001-1: Artificial Intelligence - Life Cycle Processes and Quality Requirements - Part 1: Quality Metamodel veröffentlicht. Dieser definiert ein Qualitäts-Metamodell („Quality Metamodel“), das alle wichtigen Qualitätsaspekte von KI enthält und miteinander in Verbindung bringt.
- Der in Entwicklung befindliche Standard DIN SPEC 92001-2: Artificial Intelligence - Life Cycle Processes and Quality Requirements - Part 2: Robustness definiert in Ergänzung zu [12207]

Anforderungen Softwarequalitätsanforderungen bzgl. Robustheit und ordnet diese den verschiedenen Phasen des Lebenszyklus einer Anwendung zu.

UL 4600 ist ein Standard, der einen Sicherheitsnachweisansatz verfolgt, um Produktsicherheit bei autonomen Systemen im Allgemeinen und für autonom fahrende Fahrzeuge im Konkreten sicherzustellen. Dieser wurde von den Underwriters Laboratories, einer US-amerikanischen, unabhängigen Organisation für Produktuntersuchung und Zertifizierung bezüglich Sicherheit (Safety), veröffentlicht. UL 4600 verzichtet auf eine bestimmte Implementierung oder auf einen bestimmten technologischen Ansatz. Stattdessen wird ein sogenannter Sicherheitsnachweis (Safety Case) erstellt, welcher aus einem Ziel (goal), einer Erörterung (argument) und Nachweisen (evidence) besteht.

- [1] P. Wesel, A. E. Goodloe: Challenges in the Verification of Reinforcement Learning Algorithms, NASA/TM-2017-219628, 2017.
- [2] Huang, X., Kwiatkowska, M., Wang, S., & Wu, M. (2017, July). Safety verification of deep neural networks. In International Conference on Computer Aided Verification (pp. 3-29). Springer, Cham.
- [3] Katz, G., Barrett, C., Dill, D. L., Julian, K., & Kochenderfer, M. J. (2017, July). Reluplex: An efficient SMT solver for verifying deep neural networks. In International Conference on Computer Aided Verification (pp. 97-117). Springer, Cham.
- [4] Ghosh, S., Lincoln, P., Tiwari, A., Zhu, X., & Edu, W. (2016, June). Trusted machine learning for probabilistic models. In ICML Workshop on Reliable Machine Learning in the Wild.
- [5] Fulton, N., & Platzer, A. (2018, April). Safe reinforcement learning via formal methods: Toward safe control through proof and learning. In Thirty-Second AAAI Conference on Artificial Intelligence.
- [6] Pei, K., Cao, Y., Yang, J., & Jana, S. (2017, October). Deepxplore: Automated whitebox testing of deep learning systems. In proceedings of the 26th Symposium on Operating Systems Principles (pp. 1-18).
- [7] Guo, J., Jiang, Y., Zhao, Y., Chen, Q., & Sun, J. (2018, October). Difuzz: Differential fuzzing testing of deep learning systems. In Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (pp. 739-743).
- [8] Odena, A., & Goodfellow, I. (2018). Tensorfuzz: Debugging neural networks with coverage-guided fuzzing. arXiv preprint arXiv:1807.10875.
- [9] Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 427-436).
- [10] Pei, K., Cao, Y., Yang, J., & Jana, S. (2017, October). Deepxplore: Automated whitebox testing of deep learning systems. In proceedings of the 26th Symposium on Operating Systems Principles (pp. 1-18).
- [11] Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., ... & Kurakin, A. (2019). On evaluating adversarial robustness. arXiv preprint arXiv:1902.06705.
- [12] Ma, L., Juefei-Xu, F., Zhang, F., Sun, J., Xue, M., Li, B., ... & Zhao, J. (2018, September). Deepgauge: Multi-granularity testing criteria for deep learning systems. In Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering (pp. 120-131).
- [13] Dong, Y., Zhang, P., Wang, J., Liu, S., Sun, J., Hao, J., ... & Ting, D. (2019). There is Limited Correlation between Coverage and Robustness for Deep Neural Networks. arXiv preprint arXiv:1911.05904.
- [14] Tian, Y., Pei, K., Jana, S., & Ray, B. (2018, May). Deeptest: Automated testing of deep-neural-network-driven autonomous cars. In Proceedings of the 40th international conference on software engineering (pp. 303-314).
- [15] <https://developer.ibm.com/open/projects/ai-fairness-360/>
- [16] <https://www.ibm.com/cloud/watson-openscale>
- [17] Weyuker, E. J. (1982). On testing non-testable programs. The Computer Journal, 25(4), 465-470.
- [18] C. Murphy, G. E. Kaiser: Improving the Dependability of Machine Learning Applications. Columbia University, 2008. <https://doi.org/10.7916/D8P2761H>
- [19] Xie, X., Ho, J. W., Murphy, C., Kaiser, G., Xu, B., & Chen, T. Y. (2011). Testing and validating machine learning classifiers by metamorphic testing. Journal of Systems and Software, 84(4), 544-558.

#### **I.4 Zusammenarbeit mit anderen Stellen**

Von Beginn des Projekts an war eine Abstimmung mit der EASA (AP 4.4) vorgesehen, weswegen deren Vorarbeiten und Publikationen stets beachtet und verfolgt wurden. Insbesondere die erneuerte Version ihrer Roadmap (2.0)<sup>[1]</sup> gab neue Impulse, bestätigte jedoch auch das bis dahin konzipierte Vorgehen. Im Rahmen der Abstimmung in AP 4.4 wurden die Ansätze und Konzeptionen gegenseitig vorgestellt und diskutiert.

<sup>[1]</sup> EASA (2023). Artificial Intelligence Roadmap 2.0 (<https://www.easa.europa.eu/en/newsroom-and-events/news/easa-artificial-intelligence-roadmap-20-published>). European Union Aviation Safety Agency

## II. Eingehende Darstellung

Hier werden die Projektarbeiten und ihre Einbettung in zukünftige Arbeiten beschrieben.

### II.1 Projektergebnisse und Verwertung

Im Folgenden wird, aufgeteilt in die Arbeitspakete (AP), umfassend auf die im Projekt geleisteten Arbeiten und die erzielten Ergebnisse eingegangen. Die Arbeiten des FKIE wurden in den Hauptarbeitspaketen 3 und 4 (HAP 3 und HAP 4) des Projekts durchgeführt.

#### II.1.1 Ergebnisse in Hauptarbeitspaket 1

##### II.1.1.1. *Synthese des bestehenden Know-hows für die Zertifizierung (AP 1.1)*

###### II.1.1.1.1. *Vorgehen*

Für die Entwicklung von neuen Zertifizierungsmethoden für die Anwendung von KI bildet die Grundlage das bestehende Wissen zu bisher verwendeten und in der Entwicklung befindlichen Konzepten und Verfahren für die Zertifizierung von Softwaresystemen. Ziel des AP 1.1 war es daher, bestehende und neue Zertifizierungskonzepte und -verfahren innerhalb und außerhalb der Luftfahrtindustrie zu identifizieren, die damit eine wesentliche Grundlage für die Arbeiten in HAP 4 bilden.

###### II.1.1.1.2. *Ergebnisse*

Im Rahmen der Synthese des bestehenden Know-Hows wurden grundlegende Herausforderungen bei der Bewertung und Zertifizierung von KI-basierten Systemen mit Sicherheitsbezug identifiziert. Diese umfassten die Themenbereiche Zertifizierung, Sicherheitsargumentation, KI-spezifische Unterschiede zu klassischen Systemen und resultierende Herausforderungen bei deren Zertifizierung. Insbesondere wurde hier drei grundlegende, neue Ansätze für die Zertifizierung von KI-basierten Systemen vorgestellt, die sich auf wesentliche Eigenschaften von KI-Modellen oder auf den Vergleich mit bereits zertifizierten Systemen oder statistischen Systemen mit ähnlichen Eigenschaften wie sie KI-Systeme innehaben, bspw. Kalman-Filter, stützten. Die beiden neuartigen Ansätze umfassen dabei die von der NASA entwickelten Overarching Properties sowie sog. High-Level-Properties. Beide Ansätze wurden als vielversprechende Ansätze für das HAP identifiziert. Zudem wurden Entwicklungs- und Sicherheitsbasierte Prozesse sowie Prozesse zur Reifegradbewertung identifiziert. Zudem wurden spezifische Methoden zur Unsicherheitsmessung (Uncertainty Measurement) und Erklärbarkeit von KAI (Explainable AI, XAI) zusammengefasst, die wesentliche Herausforderungen bei der Zertifizierung von KI adressieren.

##### II.1.1.2. *Zertifizierungskonzepte für KI (AP 1.2)*

###### II.1.1.2.1. *Vorgehen*

Für die Entwicklung eines robusten Zertifizierungsrahmens für ML/AI-Systeme in der Luft- und Raumfahrtindustrie ist das Verständnis von ML-induzierten Risikofaktoren sowie von ML-Fehlern und deren Ursachen entscheidend. Im Gegensatz zu klassischer Software, bei der das Verhalten explizit programmiert und vorhersehbar ist, lernen ML-Systeme aus Daten, um Aufgaben auszuführen, was einzigartige Risiken und Fehler mit sich bringt. Im Rahmen dieses Arbeitspakets haben wir uns auf Risiken konzentriert, die sich auf die Sicherheitsargumentation auswirken könnten, wenn klassische Zertifizierungsansätze in Betracht gezogen werden.

###### II.1.1.2.1. *Ergebnisse*

Als Teil des Arbeitspakets haben wir ein Dokument erstellt, das Risikofaktoren auf System-, Prozess-, Daten-, Integrations-, Design- und Implementierungsebene aufzeigt. Für jede Systemebene werden

Risiken wie fehlende Robustheitsgarantien, mangelnde Interpretierbarkeit und Transparenz berücksichtigt. Für jedes Risiko wird eine Beschreibung gegeben, und es werden mögliche Konsequenzen und Strategien zur Risikominderung erörtert. Anschließend werden die Ursprünge und Ausfälle von ML anhand einer facettierten Taxonomie erörtert, die auf zwei bestehenden Arbeiten aufbaut. Hier werden 4 Facetten in Bezug auf Ursprung, Zeitpunkt der Einführung, Zeitpunkt der Entdeckung und Auswirkung betrachtet.

## **II.1.2 Ergebnisse in Hauptarbeitspaket 2**

### **II.1.2.1. *Potentiale für Luftfahrzeug, Flugsicherung und Flughafen (AP 2.1)***

#### **II.1.2.1.1. *Vorgehen***

In diesem Arbeitspaket war es unser Ziel, das Potenzial von KI-Anwendungen für Flugzeuge, Flugverkehrsmanagementsysteme (engl. Air Traffic Management ATM) und Flughäfen zu untersuchen. Wir konzentrierten uns auf die Identifizierung von Bereichen und KI-Methoden, die machbar sind, aber derzeit aufgrund von Sicherheitsbedenken nicht genutzt werden. Ziel des war es, die Expertise der Fraunhofer-Institute im Bereich der KI einzubringen, insbesondere bei der Prozessoptimierung und Entscheidungsunterstützung an Flughäfen.

#### **II.1.2.1.2. *Ergebnisse***

Es wurde eine umfassende Liste von Projekten und möglichen Ansätzen zur Verfügung gestellt, die KI zur Bewältigung verschiedener Herausforderungen in der Luftfahrt einsetzen. Dazu wurde jeweils ihre Machbarkeit bewertet. Darüber hinaus wurden die kritischsten Aspekte (Verkehrsprognosen und -modellierung, Ressourcenmanagement und -optimierung, Leistung von Flughäfen und Fluggesellschaften, Arbeitsbelastung und Automatisierung, Passagiererlebnis, Infrastrukturüberwachung und -sicherheit sowie Datenmanagement) identifiziert und die damit verbundenen Probleme, Herausforderungen und Anforderungen detailliert untersucht. Darüber hinaus wurden für jeden Bereich vielversprechende KI-Methoden hervorgehoben, die aufgrund rechtlicher oder sicherheitstechnischer Bedenken im Zusammenhang mit der Lizenzierung und Zertifizierung noch nicht in der Praxis eingesetzt werden. Der Schwerpunkt lag auf KI-Techniken aus dem Bereich des maschinellen Lernens.

### **II.1.2.2. *Prototypen mit KI für Luftfahrzeug, Flugsicherung und Flughafen (AP 2.2)***

#### **II.1.2.2.1. *Vorgehen***

Dieses Arbeitspaket konzentriert sich darauf, wie die Methoden, die für die Zertifizierung von KI/ML-Komponenten identifiziert wurden, auf ein ML-basiertes Flugbahnvorhersagesystem angewendet werden können. In diesem Arbeitspaket war es das Ziel, einen Plan für unseren Demonstrator (das Flugbahnvorhersagesystem) zusammen mit der Art der verwendeten Daten vorzustellen und zu untersuchen. Zudem sollten die im EASA-Bericht beschriebenen Zertifizierungskriterien eingeordnet werden.

#### **II.1.2.2.2. *Ergebnisse***

Es wurden mehrere Anwendungen beleuchtet, exemplarisch jedoch die Vorhersage der Flugbahn von Flugzeugen mit LSTM-Netzen und Social-LSTM im Detail durchdekliniert. Es wurde das Konzept des Social-LSTM betrachtet, welches an der Stanford University für die Vorhersage von Fußgängerbewegungen entwickelt wurde. Es wurde insbesondere eine mögliche Anwendung in der Luftfahrt erörtert, bei der benachbarte Flugzeuge bei der Vorhersage der Flugbahn eines bestimmten Flugzeugs berücksichtigt werden. Dieser Ansatz bietet besonderes Potential bei der

Prädiktionsgenauigkeit, da wichtige Informationen aus seiner Umgebung ableitet werden. Das Ergebnis ist eine Reduzierung der Dimensionalität der Daten. Das Konzept sieht ein Training mittels Daten aus dem OpenSky Network vor. Es wurde die Architektur des Ansatzes untersucht und im Hinblick auf Hindernisse, die sich bei der Zertifizierung ergeben können insbesondere mit Bezug auf die EASA-Konzepte und die brauchbaren Anleitungen für ML-Anwendungen der Stufe 1, beleuchtet.

### **II.1.3 Ergebnisse in Hauptarbeitspaket 3**

#### **II.1.3.1. Komponenten mit KI (AP 3.2)**

##### **II.1.3.1.1. Vorgehen**

Das Ziel dieses Arbeitspakets war es, den Ansatz weiterzuentwickeln und ihn für die Zusammenarbeit mit dem Konsortium und die Erweiterung im Hinblick auf Optimierung und Zertifizierbarkeit vorzubereiten. Dieser Abschnitt konzentrierte sich auch darauf, sicherzustellen, dass die Methode für die nächsten Arbeitspakete bereit ist und dass die erforderlichen Daten gesammelt, hinsichtlich ihrer Qualität sichergestellt und bereinigt werden.

##### **II.1.3.1.2. Ergebnisse**

In diesem Abschnitt haben wir eine beträchtliche Menge an Daten aus der OpenSky ADS-B-Datenbank gesammelt, die sich über zwei Monate erstrecken (Juni und November 2019). Alle Daten decken dasselbe geografische Gebiet ab (Nordrhein-Westfalen, mit Breitengraden von 50 bis 52 Grad Nord und Längengraden von 5 bis 9 Grad Ost). Wir haben uns dafür entschieden, Daten aus zwei verschiedenen Wettersaisonen (Sommer und Winter) vor der Pandemie zu verwenden. Die Daten wurden reduziert, vorverarbeitet, getrimmt und für das weitere Training gesichert, so dass sie von unseren Partnern leicht verwendet werden können. Wir trainierten unseren Ansatz mit diesem größeren Datensatz und stellten alle erforderlichen Gewichte und Parameter zur Verfügung, um die Nutzung zu erleichtern. Außerdem haben wir drei verschiedene Schnittstellen/Zugangspunkte für die Anwendung geschaffen: eine grafische Benutzeroberfläche (GUI), eine Zugangs-API und Argparse. Diese drei Zugriffsmethoden ermöglichen es unserem Konsortium, die Methode auf verschiedene Weise zu nutzen, um die Benutzerfreundlichkeit zu gewährleisten. Die Access API bietet eine Reihe von Methoden, die über das Terminal aufgerufen werden können, während der Argparse-Code eine Liste von Parametern liefert, die leicht zu lesen und für zukünftige Experimente leicht zu ändern sind.

#### **II.1.3.2. Komplexe Anwendungen (AP 3.3)**

##### **II.1.3.2.1. Vorgehen**

Das Ziel dieses Arbeitspakets besteht darin, die erforderlichen Testszenarien und -umgebungen zu konzipieren. Ziel war es, Herausforderungen zu identifizieren und eine umfassende Strategie zur Verbesserung der Vorhersagefähigkeiten unseres Modells zu entwickeln. Durch die Bewältigung dieser Herausforderungen und die Umsetzung proaktiver Strategien in den Bereichen Datenauswahl, Modellarchitektur, Netzwerkanpassungen und Erklärbarkeit wollen wir die Vorhersagegenauigkeit und Zuverlässigkeit unseres Modells in realen Anwendungen verbessern.

##### **II.1.3.2.2. Ergebnisse**

Als Ergebnis haben wir die erforderlichen Testszenarien und -umgebungen konzeptionell umrissen, die notwendig sind, um dynamisches Lernen zu ermöglichen und die kontinuierliche Verbesserung von KI-Systemen zu gewährleisten. Wir haben fünf Hauptbereiche und deren Ziele identifiziert:

Datenauswahl und Qualitätssicherung: Sicherstellung der Angemessenheit und Zuverlässigkeit der Dateneingaben.

Struktur des Modells: Untersuchung der Architektur und des Designs unseres Social-LSTM-Modells, um potenzielle Schwachstellen oder Ineffizienzen zu identifizieren.

Sicherheitskomponenten: Bewertung der Robustheit und Integrität der in unseren Rahmen integrierten Sicherheitsmaßnahmen zum Schutz vor Schwachstellen und Verstößen.

Anpassungen und -Änderungen des Netzwerks: Untersuchung der Anpassungsfähigkeit und Reaktionsfähigkeit unserer Netzinfrastruktur, um Schwankungen der Datenlast und den betrieblichen Anforderungen gerecht zu werden.

Erklärbarkeit und Interpretierbarkeit [12]: Untersuchung der Klarheit und Verständlichkeit der erzeugten Ergebnisse von unserem Social-LSTM-Modell, um Transparenz und einfache Interpretation zu gewährleisten.

Wir haben diese Bereiche eingehend untersucht und eine umfassende Analyse der spezifischen Herausforderungen und Szenarien erstellt, die in jedem Bereich auftreten können, zugeschnitten auf den von uns gewählten Ansatz (Social-LSTM). Darüber hinaus untersuchten wir potenzielle Lösungen für die Bewältigung der oben genannten Herausforderungen.

### **II.1.3.3. *Systems of Systems Anwendungen (AP 3.4)***

#### **II.1.3.3.1. *Vorgehen***

Ziel dieses Abschnitts war es, die Integration verschiedener Demonstratoren in das Gesamtsystem der Systeme konzeptionell zu skizzieren und anhand von Zertifizierungskriterien, Herausforderungen und potenziellen Risiken zu bewerten. Außerdem soll eine Zertifizierungsmethodik formuliert werden, die auf den gesammelten Erfahrungen für komplexe Anwendungen und vernetzte KI-Funktionalitäten basiert.

#### **II.1.3.3.2. *Ergebnisse***

In diesem Arbeitspaket haben wir die potenziellen Verbindungen zwischen den Demonstratoren KI-AMAN, ADS TPN und Social-LSTM zur Vorhersage von Trajektorien untersucht. Wir identifizierten potenzielle Verbindungen, Austauschmöglichkeiten und Anwendungsfälle und definierten die notwendigen Schnittstellen und Verbindungspunkte zwischen den Demonstratoren. Wir entwickelten sowohl sequentielle als auch parallele Prozesse: Beim sequentiellen Ansatz empfängt ein System den Output eines anderen Systems und nutzt ihn als Grundlage für die weitere Verarbeitung oder Entscheidungsfindung, während beim parallelen Konzept beide Systeme gleichzeitig arbeiten und synergetisch zum Benutzererlebnis beitragen, indem sie gleichzeitig zusätzliche Erkenntnisse oder Informationen liefern. Auf der Grundlage dieser beiden möglichen Prozesse haben wir eine Liste von Anwendungsfällen, Akteuren und Betriebsabläufen zusammengestellt. Darüber hinaus haben wir die Vorteile von vernetzten KI-Systemen hervorgehoben und uns mit den Herausforderungen in Bezug auf Sicherheit und Erklärbarkeit befasst.

## **II.1.4 Ergebnisse in Hauptarbeitspaket 4**

### **II.1.4.1. Testverfahren (AP 4.1)**

#### **II.1.4.1.1. Vorgehen**

Das Ziel dieses Arbeitspaket bestand darin, bestehende Testverfahren auf ihre Eignung für die Zertifizierung von KI-basierten Anwendungen der Luftfahrt zu untersuchen und ggf. weiterzuentwickeln bzw. neue Methoden zu entwickeln. Da die Ergebnisse dieses Arbeitspakets die Grundlage für die weiteren Arbeitspakete von HAP 4 bildeten, war nicht nur deren Eignung für die Zertifizierung entscheidend, sondern auch deren Anwendbarkeit auf den FhG-Demonstrator.

#### **II.1.4.1.2. Ergebnisse**

Als Teil dieses Arbeitspakets wurden Testverfahren für High-Level-Properties (HLP) entwickelt, nämlich Robustheit, Datenqualität und Erklärbarkeit. Um den Anforderungen hinsichtlich der Eignung für die Zertifizierung gerecht zu werden, wurden einerseits Testverfahren gewählt, die die drei zuvor genannten Properties des FhG-Demonstrators adressieren, und andererseits diese weiterentwickelt, um die spezifischen Anforderungen an die Zertifizierung zu adressieren. Bezüglich der HLP-Eigenschaft „Robustheit“ wurden dazu bestehende Verfahren zur Bewertung der Robustheit von Modellen des Maschinellen Lernens weiterentwickelt, nämlich Adverserielle Angriffe und Robustheitsmetriken, und dahingehend weiterentwickelt, dass die Herausforderungen der Zertifizierung, nämlich ein ganzheitliches Bild der Modell-Robustheit zu gewinnen, aufgegriffen und umgesetzt werden. Die Ergebnisse mündeten in eine neue Methode zur Messung der Robustheit eines Modells. Weitergehend wurden bestehende Methoden zur Erkennung von Anomalien in Datensätzen identifiziert und für die Evaluation am FhG-Demonstrator ausgewählt. Hinsichtlich der Erklärbarkeit von KI wurde der FhG-Demonstrator um eine Schnittstelle erweitert, die eine lokale Erklärbarkeit über lineare Approximation zulässt. Dabei wurde der LIME-Ansatz abgewandelt und auf LSTM-Architekturen übertragen. Die Ergebnisse dieses APs fanden Einzug in AP4.2.

### **II.1.4.2. Verifikation und Validierung (AP 4.2)**

#### **II.1.4.2.1. Vorgehen**

In diesem Arbeitspaket wurden die im Rahmen von AP4.1 entwickelten Testverfahren angepasst, um die in AP2.2 entwickelten Demonstratoren zu bewerten. Aufgrund der Komplexität von KI-Systemen sind neuartige Testtechniken erforderlich, um sie zu bewerten. Darüber hinaus sollten V&V-Ansätze, die diese neuartigen Testtechniken beinhalten, auch für KI-Systeme etabliert werden. Zu diesem Zweck und mit einem Schwerpunkt auf Demonstratoren im Projekt.

#### **II.1.4.2.2. Ergebnisse**

Im Rahmen dieses Arbeitspakets haben wir die drei High-Level Properties (HLP), die in AP1.1 behandelt wurden, – Robustheit, Datenqualität und Erklärbarkeit – an einem Demonstrator der FhG untersucht. Für die Robustheit wurden zwei Testtechniken entwickelt, nämlich adversarielle Robustheitsmessung und Unsicherheitsabschätzung (Uncertainty Estimation). Der erste Ansatz wurde für den FhG-Demonstrator und den zugehörigen Datensatz implementiert und evaluiert. Der Ansatz der Unsicherheitsabschätzung (Uncertainty Estimation) wurde als Proof-of-Concept für LSTMs an einem Beispieldatensatz demonstriert. Für die Datenqualität wird ein Ansatz zur Erkennung von Ausreißern vorgeschlagen und am für den FhG-Demonstrator verwendeten Datensatz wird evaluiert. Für Explainability, haben wir verschiedene Dimensionen identifiziert (Transparenz, Interpretierbarkeit, Kausalität und Vertrauenswürdigkeit) und versucht, diese mit verschiedenen Testmethoden (quantitative Testfortschritts-Metriken, Regularisierungstechniken, Evaluierungstechniken und Nutzerstudien) sowie durch die Bereitstellung verschiedener Schnittstellen und Zugangspunkte zu berücksichtigen.

Der Arbeitspaketbericht enthält Beschreibungen der Testverfahren, Bewertungen des Demonstrators und des Datensatzes.

#### **II.1.4.3. Methodik der Zertifizierung (AP 4.3)**

##### **II.1.4.3.1. Vorgehen**

Die Arbeiten dieses APs umfassen die Identifizierung relevanter Normen und Standards und deren Anwendbarkeit auf Systeme, die auf maschinellem Lernen basieren, um Wirksamkeitsnachweise zur Absicherung von KI-basierten Systemen zu ermöglichen. Um die Arbeiten des Projekts mit den aktuellen regulatorischen Entwicklungen im Luftfahrtsektor auszurichten, wurde bereits in diesem Arbeitspaket ein intensiver Austausch mit der zugehörigen Aufsichtsbehörde EASA aufgenommen. Dieser mündete in zwei von FhG organisierten Workshops, die mit dem Projektpartnern bei der Deutschen Flugsicherung in Langen und bei Airbus in Manching in Präsenz durchgeführt wurden. Das Ziel bestand zum einen in der Identifikation der vom Konsortium angewandten Methoden für die Zertifizierung und ihrer Relation zu den regulatorischen Entwicklungen hinsichtlich einer Zertifizierung von KI-basierten Systemen.

##### **II.1.4.3.2. Ergebnisse**

Die Grundlage dieses Arbeitspakets bildete das Concept Paper der EASA, das Richtlinien für die Zulassung von KI-basierten Systemen beschreibt. Im Rahmen dieses Arbeitspakets wurden daher die Projektarbeiten mit den Ansätzen der EASA verknüpft, um die Zertifizierbarkeit von KI-basierten Systemen durch Vorschläge für den Einsatz konkreter Methoden zu verbessern. Im Ergebnis wurden die vom EASA Concept Paper vorgeschlagenen Ziele (Objectives) für den Einsatz von KI in sicherheitskritischen Luftfahrtanwendungen um konkrete Methoden zur Einhaltung der Vorschriften (Means of Compliance) ergänzt, da diese bisher nur in wenigen Fällen Teile des EASA Concept Paper waren. Zudem wurden, soweit möglich, bereits konkrete Akzeptanzkriterien für die Bewertung der Means of Compliance dokumentiert. Hinweise für Aufsichtsbehörden zur Bewertung der angewandten Methoden vervollständigten vorgeschlagenen Means of Compliance. Abschließend wurden Anforderungen an eine Zertifizierungsmethodik dokumentiert.

#### **II.1.4.4. Abgleich mit EASA (AP 4.4)**

##### **II.1.4.4.1. Vorgehen**

Im Rahmen dieses Arbeitspaketes wurden verschiedene Workshops mit der EASA durchgeführt, um die Ergebnisse von AP4.1 - 4.3 zu diskutieren und zu bewerten. Mit dem Schwerpunkt auf dem EASA-Konzeptpapier für AI, den Zertifizierungsstrategien, den fehlenden Prüftechniken und V&V-Ansätzen wurde die Gültigkeit bestehender Evaluierungen diskutiert.

##### **II.1.4.4.2. Ergebnisse**

In zwei Workshops mit der EASA haben wir Feedback zu allen Demonstratoren des Projekts erhalten, indem wir ihnen eine strukturierte Vorlage zur Verfügung gestellt haben. Die Vorlage wurde von allen Partnern mit Prüftechniken ausgefüllt, aus denen hervorging, welche der im EASA-Konzeptpapier genannten Ziele erfüllt wurden. Darüber hinaus wurden auch Verbindungen zwischen den Testverfahren und den Demonstratoren hergestellt. Die EASA gab jedem Partner eine Rückmeldung zu der Vorlage.

#### **II.1.4.5. Neue Methodik für die Zertifizierung von KI (AP 4.5)**

##### **II.1.4.5.1. Vorgehen**

Dieses Arbeitspaket fasst die entwickelten Methoden für die Zertifizierung zusammen, diskutiert das von der EASA bereitgestellte Feedback und hebt die Herausforderungen hervor, die bei der Durchführung von Zertifizierungen auf Komponenten- und Systemebene für auf maschinellem Lernen basierende Systeme auftreten.

##### **II.1.4.5.2. Ergebnisse**

Das Ergebnis verweist auf die Herausforderungen bei der Übertragung der Systemzertifizierung für KI-basierte Systeme auf verschiedene Flughäfen aufgrund von Unterschieden in den Operational Design Domains (ODD). Die Ergebnisse heben auch hervor, dass einige Aspekte des Systems, wie z. B. Methoden zur Unsicherheitsabschätzung, möglicherweise keine vollständige Neubewertung erfordern, während andere Aspekte, wie z. B. die Modellschulung für eine genaue Vorhersage, bei einem erneuten Einsatz an anderen geografischen Standorten neu behandelt werden müssen.

## **II.2 Notwendigkeit und Angemessenheit der geleisteten Arbeit**

In der Luftfahrt hat die Sicherheit oberste Priorität. KI-Systeme, die in der Luftfahrt eingesetzt werden, müssen nachweislich sicher und zuverlässig sein, um Risiken für Passagiere und Besatzung zu minimieren. KI-Systeme können komplexe Aufgaben übernehmen, aber es ist entscheidend, sicherzustellen, dass sie unter allen Umständen korrekt funktionieren und keine gefährlichen Fehlentscheidungen treffen. Des Weiteren müssen KI-Systeme unter verschiedenen Bedingungen konsistent und zuverlässig arbeiten, was durch Zertifizierungsverfahren überprüft und garantiert werden kann. Die Systeme müssen robust gegen unerwartete Eingaben und Situationen sein und dennoch sicher funktionieren.

In der Luftfahrtindustrie gibt es strenge regulatorische Anforderungen und Standards. Forschung hilft dabei, geeignete Zertifizierungsprozesse zu entwickeln, die sicherstellen, dass KI-Systeme diese Anforderungen erfüllen. Einheitliche Zertifizierungsstandards sind notwendig, um weltweit einheitliche Sicherheits- und Leistungsanforderungen für KI-Systeme in der Luftfahrt zu gewährleisten. Für die breite Akzeptanz von KI in der Luftfahrt ist es wichtig, dass Passagiere und die Öffentlichkeit Vertrauen in die Sicherheit und Verlässlichkeit dieser Technologien haben. Zertifizierungen können dieses Vertrauen stärken. Auch innerhalb der Luftfahrtindustrie müssen Ingenieure, Piloten und andere Stakeholder Vertrauen in die KI-Systeme haben. Zertifizierung kann dieses Vertrauen fördern.

Zertifizierungsforschung kann Innovationen fördern, indem sie klare Richtlinien und Rahmenbedingungen für die Entwicklung neuer KI-Technologien bietet. Durch Forschung und Zertifizierung können Best Practices für die Entwicklung und Implementierung von KI-Systemen in der Luftfahrt etabliert werden. Die Luftfahrtumgebung ist komplex, und KI-Systeme müssen in der Lage sein, diese Komplexität zu bewältigen. Zertifizierungsprozesse helfen sicherzustellen, dass die Systeme entsprechend entwickelt und getestet werden. KI-Systeme müssen nahtlos mit vorhandenen Luftfahrtsystemen und -prozessen interagieren können, was durch standardisierte Zertifizierungsverfahren sichergestellt werden kann.

Die Forschung im Bereich der Zertifizierung von KI in der Luftfahrt ist notwendig, um die Sicherheit und Verlässlichkeit der Systeme zu gewährleisten, regulatorische Anforderungen zu erfüllen, Vertrauen bei der Öffentlichkeit und in der Industrie zu schaffen, technologische Innovationen zu fördern und die spezifischen Herausforderungen der Luftfahrt zu bewältigen.

## **II.3 Voraussichtlicher Nutzen und Verwertbarkeit der Ergebnisse im Sinne des fortgeschriebenen Verwertungsplanes**

Wie in der Verwertungsplanung des Fraunhofer Konsortiums vorgesehen, konnten arbeitswissenschaftliche Kompetenzen und domänenspezifisches Wissen gewonnen werden. Besonders herauszuheben sind hier methodische Kompetenzen und praktische Erfahrung im Bereich Überprüfbarkeit von KI, Verifikation und Validierung sowie Zertifizierung im Luftfahrtbereich. Außerdem konnten tiefe Einblicke in verwandte Arbeiten, vordergründig bei der EASA, nachgeordnet jedoch auch in Bereiche des straßengebundenen Verkehrs gewonnen werden. Diese domänenspezifischen Kenntnisse bieten eine sinnvolle Erweiterung der bisherigen Kompetenzen im Bereich der Luftfahrt und KI. Durch die enge Zusammenarbeit mit den Partnern, aber auch durch den Austausch mit Experten und die Möglichkeit, die Projektergebnisse auf unterschiedlichen Veranstaltungen einem Fachpublikum zu präsentieren, konnte die angestrebte Netzwerkbildung gestärkt werden.

## **II.4 Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen**

Zum Zeitpunkt der Drucklegung sind keine für die Forschungsergebnisse relevanten Fortschritte auf dem Gebiet des Vorhabens bei anderen Stellen bekannt.

## II.5 Erfolgte oder geplante Veröffentlichungen des Ergebnisses

F. Govaers and P. Baggenstoss, "On a Detection Method of Adversarial Samples for Deep Neural Networks," *2021 IEEE 24th International Conference on Information Fusion (FUSION)*, Sun City, South Africa, 2021, pp. 1-5, doi: 10.23919/FUSION49465.2021.9627060.

V. R. Gala and M. A. Schneider, "Evaluating the Effectiveness of Attacks and Defenses on Machine Learning Through Adversarial Samples," *2023 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*, Dublin, Ireland, 2023, pp. 90-97, doi: 10.1109/ICSTW58534.2023.00027

V. R. Gala, M. A. Schneider and M. Vogt, "Towards an Empirical Robustness Assessment Through Measuring Adversarial Subspaces", *AST International Conference on Automation of Software Test, ICSE 2024 - Conference Lisbon, Portugal, April 14-20, 2024 IEEE, 2024*, doi: 10.1145/3644032.3644464

V. R. Gala, M. A. Schneider and M. Vogt, "Towards an Evaluation Methodology of ML Systems from the Perspective of Robustness and Data Quality", *International Workshop on System Testing and Validation (STV), QRS Conference, Cambridge, United Kingdom, July 1-5, 2024*.

Joanna Polewczyk, Arne Schwarze and Felix Govaers, "Requirements specification for certification of artificial intelligence applications in aviation", *Poster, Fraunhofer FKIE, 2023*.

Joanna Polewczyk, Arne Schwarze and Felix Govaers, "Aircraft Trajectory Prediction and Time Estimation adapting Social-LSTM", *Preprint, 2024*.