Künstliche Intelligenz für sichere Web-Infrastrukturen mit digitalem Identitätsmanagement

Akronym: KIWI

Teilvorhaben
Hochschule Karlsruhe Technik und Wirtschaft

Beherrschbarkeit und Einsetzbarkeit komplexer föderierter KI-Infrastrukturen

Förderkennzeichen: 16KIS1142K

Sachbericht Teil I
Kurzbericht

Ansprechpartner:

Prof. Dr.-Ing. Zoltán Nochta
Hochschule Karlsruhe - Technik und Wirtschaft
Moltkestr. 30, 76133 Karlsruhe
Tel.: +49-721-925-1578
E-Mail: zoltan.nochta@h-ka.de

1 Aufgabenstellung des Teilvorhabens

Das Verbundforschungsvorhaben „Künstliche Intelligenz für sichere Web-Infrastrukturen mit digitalem Identitätsmanagement“ (KIWI) beschäftigte sich mit der Einsetzbarkeit eines KI-gestützten Sicherheitsmanagements in komplexen Web-Infrastrukturen. Im Fokus standen KI-Verfahren und -Systeme zur Erkennung möglicher Angriffe gegen Web-Identitätsdienste. Die KI-Verfahren arbeiten auf Basis domänenspezifisch und auch domänen-übergreifend gesammelter Daten. Sie erfüllen dabei Anforderungen nach Autonomie, Geheimhaltung, Datenschutz und Selbstbestimmung der Nutzer und Betreiber der Systeme.

2 Projektablauf

Das Projekt startete am 01.06.2021. Die HKA übernahm die Koordination des Projektkonsortiums, und Prof. Nochta agierte als Gesamtprojektleiter. Um aufgrund der CoViD19-Pandemie entstandene Verzögerungen auszugleichen, wurde eine kostenneutrale Projektverlängerung von sechs Monaten beantragt und vom Projektträger genehmigt. Dementsprechend endete das Projekt am 30.11.2023, statt wie ursprünglich geplant am 31.05.2023.
Das Team der HKA hat zu allen sechs Arbeitspaketen des Gesamtprojekts Beiträge geleistet und mit den Konsortialpartnern, insbesondere adesso und KIT, intensiv zusammengearbeitet. Die Projektziele und -aufgaben wurden in dezidierten Teilarbeitspaketen der HKA verfolgt respektive erledigt. Die dabei vor allem durch die wissenschaftlichen Mitarbeiter erzielten Ergebnisse bilden die Grundlage ihrer geplanter Promotionsvorhaben. Ihre Arbeit wurde zeitweise durch studentische Hilfswissenschaftler/Innen unterstützt.

3 Wesentliche Ergebnisse

Die HKA beschäftigte sich mit drei Themenbereichen (s. auch Abb. 1) und erzielte dabei folgende wesentliche Ergebnisse:

- **Föderierte KI-Detektoren für sichere mobile Systeme:** Ziel war die Entwicklung neuer Authentifizierungsmethoden, mit besonderem Fokus auf mobile Geräte. Dadurch sollen Szenarien, wie beispielsweise der Diebstahl eines Gerätes und der Missbrauch von auf einem Gerät gespeicherten digitalen Identitäten verhindert werden.
  Es wurden KI-Detektoren und ein neues Authentifizierungsprotokoll entworfen, welche anhand des Nutzerverhaltens, Standorts, Netzwerkverbindung des mobilen Geräts bewerten, ob es sich um einen legitimen Nutzer handelt. Die Privatsphäre der Nutzer wird mithilfe von homomorpher Verschlüsselung geschützt. Die entstandenen neuronalen Netze unterstützen sowohl die horizontale- als auch vertikale Föderation von Modellen, sodass mehrere Anbieter mobiler Anwendungen gemeinsam KI-Detektoren trainieren können.

- **Föderierte Trainingsmethoden für KI-Modelle:** KI-Modelle sollen unter Mitwirkung mehrerer, voneinander unabhängiger, autonom agierender Teilnehmer (zum Beispiel Unternehmen) trainiert werden, möglichst ohne die Übermittlung eigener sensitiver Trainingsdaten oder auch Modellparameter an eine zentrale Stelle oder an andere im Training ebenfalls involvierte Teilnehmer.
  Es wurde hierfür ein Framework basierend auf Secure Multi-Party Computation konzipiert und implementiert, welches das föderierte Training neuronaler Netze ermöglicht. Zudem wurde die ebenfalls kollaborative Vorverarbeitung der Trainingsdaten untersucht und durch geeignete Maßnahmen, wie bspw. Private Set Intersection, abgesichert. Experimente mit diversen aus der Fachliteratur bekannten Angriffsmethoden gegen das System rundeten die Arbeit ab, um Restrisiken zu ermitteln.

- **Föderierte Data-Governance zur Qualitätssicherung von KI-Daten:** Ziel der Arbeit waren Methoden zur Steigerung der Zuverlässigkeit KI-basierter Systeme, insb. Detektoren zur Angriffserkennung, durch Steuerung und Kontrolle von Qualitätsfaktoren der für ihr Verhalten maßgeblichen Datenbasis. Insbesondere waren Ansätze für föderierte Umgebungen von Interesse, in denen die Basisdaten selbst nur partiell bekannt sind und der

Austausch der sie betreffenden Informationen einer strikten Regulierung unterliegt. Das erste Ergebnis dieser Forschung war die Definition eines Data-Governance-Lebenszyklus mit drei Hauptphasen: In der Verhandlungsphase werden das Ziel des FL-Projekts und die Trainingskonfiguration definiert. In der zweiten Phase wird das erstellte Modell evaluiert und auch die Beiträge der Teilnehmer bewertet. In der letzten Phase werden gesammelte Metadaten den Teilnehmenden zur Verfügung gestellt, ohne Datenschutzrechte zu verletzen. Für die erste und die letzte Phase wurden Komponenten entwickelt, die es den Teilnehmern ermöglichen, sich an der Verhandlung und Bewertung von Trainingsszenarien zu beteiligen. In der zweiten Phase wurden verschiedene Ansätze zur Bewertungen der Beiträge untersucht, die im erweiterten Bericht dargestellt werden.

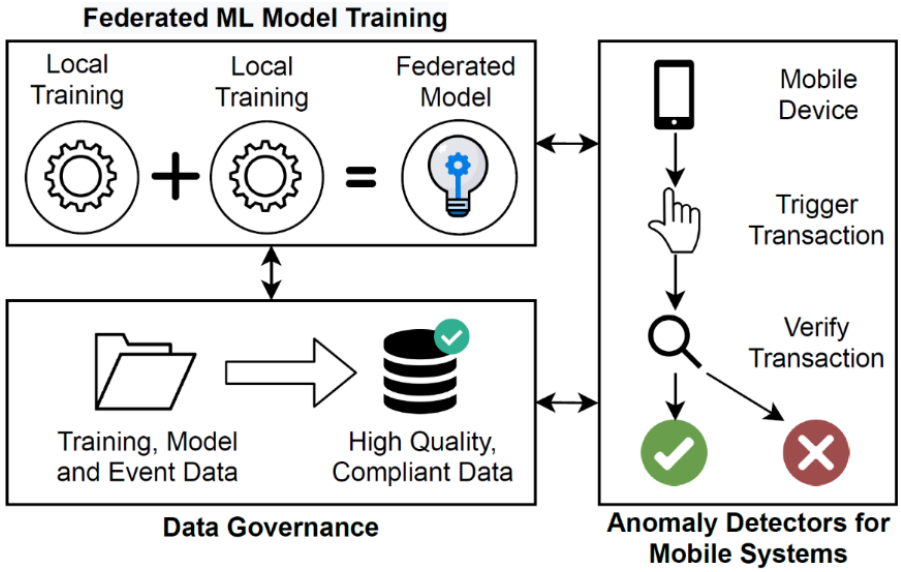

Abb. 1 Forschungsbereiche der HKA im Projekt KIWI

Insgesamt wurden zu den Ergebnissen 11 wissenschaftliche Fachartikel im Rahmen von Workshops und Konferenzen (darunter IEEE LCN 2021, IEEE ISCC 2021, IEEE DSN 2021, ESOCC 2022, IEEE LCN 2023) veröffentlicht. Außerdem wurde ein Patent angemeldet (EP4083838A1).

Im Rahmen der Konferenz ESOCC 2022 hat das KIWI-Konsortium unter Federführung der HKA den wissenschaftlichen Workshop "*International Workshop on AI for Web Application Infrastructure and Cloud Platform Security*" (kurz: AWACS) erfolgreich durchgeführt.

Verschiedene sicherheitskritische Vorfälle im Alltagsleben einer imaginierten Person samt KI-basierten Gegenmaßnahmen zur Verhinderung der Angriffe wurden im Erklärvideo „*The KIWI project explained*" (s. https://kiwi-project.org/) aufgezeigt.
Es illustriert die im Projekt adressierten Probleme und entwickelten Ansätze in einer auch für technisch weniger versiertes Zielpublikum verständlichen Form.

Im Verbundvorhaben aura.ai (s. https://www.h-ka.de/iaf/aura-ai), welches zum 01.01.2024 startete, greift die HKA ihre in KIWI erzielten theoretischen und technischen Ergebnisse auf, um maschinelles Lernen für die sichere kontinuierliche Benutzerauthentifizierung im Kontext des öffentlichen, grenzüberschreitenden Personenverkehrs einzusetzen.

Künstliche Intelligenz für sichere Web-Infrastrukturen mit digitalem Identitätsmanagement

Akronym: KIWI

Teilvorhaben
Hochschule Karlsruhe Technik und Wirtschaft

Beherrschbarkeit und Einsetzbarkeit komplexer föderierter KI-Infrastrukturen

Förderkennzeichen: 16KIS1142K

Sachbericht TEIL II
Eingehende Darstellung

Berichtszeitraum 01.06.2020 – 30.11.2023

Ansprechpartner:

Prof. Dr.-Ing. Zoltán Nochta
Hochschule Karlsruhe - Technik und Wirtschaft
Moltkestr. 30, 76133 Karlsruhe
Tel: (0721) 925-1578
E-Mail: zoltan.nochta@h-ka.de

Inhalt

Abbildungen

1. Kurzüberblick und Einordnung des Teilvorhabens

Das Verbundforschungsvorhaben „Künstliche Intelligenz für sichere Web-Infrastrukturen mit digitalem Identitätsmanagement“ (KIWI) leistete Beiträge zur Entwicklung und praktischen Erprobung eines KI-gestützten, kognitiven Sicherheitsmanagements in komplexen Web-Infrastrukturen. Im Fokus standen die Entwicklung und Erprobung von KI-Verfahren und Systemen zur Erkennung von möglichen Angriffen gegen Web-Identitätsdienste, wobei eine Generalisierung der Ergebnisse für andere Web-Dienstleistungen angestrebt wurde. Die KI-Verfahren sollen auf Basis domänenspezifisch (d.h. horizontal) und auch domänen-übergreifend (d.h. vertikal) gesammelter Daten bzw. Beobachtungen arbeiten und dabei den Anforderungen nach Autonomie, Geheimhaltung, Datenschutz und Selbstbestimmung der Nutzer und Betreiber der Systeme bestmöglich gerecht werden.

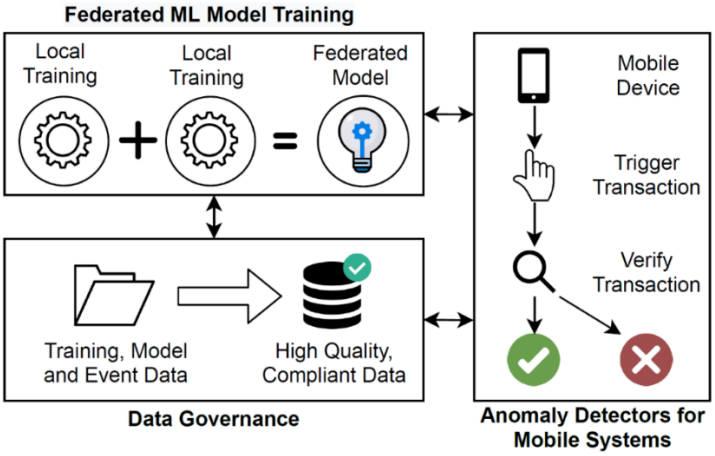

Abbildung 1: Forschungsbereiche der HKA im KIWI-Projekt

Die Forschungsarbeiten der HKA lassen sich in drei Bereiche einordnen (s. Abbildung 1):

- **Föderierte KI-Detektoren für sichere mobile Systeme:** Im Projekt wurden neue Authentifizierungsmethoden für mobile Geräte entwickelt. Diese basieren darauf, dass Verhaltensdaten der Nutzer analysiert werden, wie beispielsweise Standortdaten, Informationen über verwendete Netzwerke oder Interaktionen mit den Geräten. Die Analyse verwendet hierbei KI-Verfahren, um festzustellen, ob es sich um einen legitimen Nutzer handelt. Diese KI-Verfahren wurden so konzipiert, dass diese sowohl von horizontaler- als auch von vertikaler Föderation profitieren können. Hierdurch können Szenarien, wie beispielsweise der Diebstahl eines Gerätes und der Missbrauch von digitalen Identitäten effektiv verhindert werden. Außerdem wurde ein besonderes Augenmerk auf die Berücksichtigung der Privatsphäre der Nutzer und die Sicherstellung von adäquaten Skalierbarkeitseigenschaften gelegt.

- **Föderierte Trainingsmethoden für KI-Modelle:** KI-Modelle sollen unter Mitwirkung mehrerer, voneinander unabhängiger, autonom agierender Teilnehmer (zum Beispiel Unternehmen) trainiert werden, möglichst ohne die Übermittlung eigener sensitiver Trainingsdaten oder auch Modellparameter an eine zentrale Stelle oder an andere im Training ebenfalls involvierte Teilnehmer.

- **Föderierte Data-Governance zur Qualitätssicherung von KI-Daten:** Ziel der Arbeit sind Methoden zur Steigerung der Zuverlässigkeit KI-basierter Systeme, insb. Detektoren zur Angriffserkennung, durch Steuerung und Kontrolle von Qualitätsfaktoren der für ihr Verhalten maßgeblichen Datenbasis. Insbesondere sind Ansätze für föderierte Umgebungen von Interesse, in denen die Basisdaten selbst nur partiell bekannt sind und der Austausch der sie betreffenden Informationen einer strikten Regulierung unterliegt.

2. Im Teilvorhaben erzielte wesentliche Ergebnisse

Nachfolgend werden die in den einzelnen Arbeitspaketen erzielten Ergebnisse des Teilvorhabens erläutert.

2.1 Arbeitspaket 1: Gesamtarchitektur und Demonstrator

Im Rahmen von AP1 wurden zusammen mit den Projektpartnern unter Koordination der HKA die KIWI-Anwendungsfälle ausgearbeitet.

- **Anwendungsfall #1:** Dieser Use-Case konzentrierte sich auf die Erkennung von Missbrauch digitaler Identitäten in Geschäftsprozessen. „Missbrauch“ ist dabei definiert als die illegitime Annahme einer Identität, bzw. die illegitime Verwendung einer Identität in Geschäftsprozessen (z.B. nach AGB). Dabei kommen föderierte Verfahren sowohl für das Training von als auch für die Inferenz mit KI-Detektoren zum Einsatz. „Föderierung“ kann hier zunächst „fachlich“ als die Aufteilung/Verteilung der (Bewertung von) Domänenentitäten auf verschiedene Organisationen definiert werden. Es wird davon ausgegangen, dass ein Benutzer eine Identität beansprucht und für eine Session, die aus einer Menge von Anfragen mit den dazugehörigen Meta-Daten besteht, authentifiziert wird. Sowohl für die Identität an sich als auch für jeden Request soll dann entschieden werden, ob diese legitim oder missbräuchlich sind. Entsprechend kann der Zugriff auf angeforderte Prozesse und Ressourcen entsprechend eingeschränkt werden. Dabei wird der Ablauf einer Session im authentifizierten Kontext betrachtet und explizit nicht die Absicherung von Zugangsdaten sowie Angriffe auf den Authentifizierungsprozess (vgl. Anwendungsfall #5).

- **Anwendungsfall #2:** Der Use Case „Netzangriffe“ erforscht den Einsatz föderierter, maschineller Lernverfahren zur Erkennung von Angriffen gegen die Netzinfrastruktur von Web-basierten Identitätsmanagementsystemen. Im Rahmen des Use Cases werden (Distributed) Denial-of-Service-Angriffe betrachtet, die darauf abzielen die Verfügbarkeit Netz-gebundener Dienste einzuschränken. Ebenso werden typische Angriffe zur (illegitimen) Erhebung von Informationen über die Netzinfrastruktur und angebundene Dienste, wie etwa Port-Scans, in Betracht gezogen. Insbesondere werden Ansätze zum föderierten Training von KI-Detektoren zur Abwehr von Netzangriffen auf Basis heterogener Daten verschiedener Organisationseinheiten untersucht.

- **Anwendungsfall #3:** Der Use Case “Fraud Detection” befasst sich mit der Erkennung von betrügerischen Aktivitäten. Der Fokus liegt insbesondere auf Kreditkartenbetrug. Der Betrug besteht darin, dass die Kreditkartendaten des Karteninhabers gestohlen werden und mit den Daten dann unrechtmäßige Transaktionen getätigt werden. Jeder Zahlungsdienstleister muss sich mit der Entdeckung von Betrug beschäftigen. In diesem Anwendungsfall wird ein föderierter Ansatz angestrebt, bei dem verschiedenen Zahlungsdienstleister gemeinsam ein Fraud-Detection-Modell trainieren, ohne dass Trainingsdaten mit einem anderen Unternehmen aus-getauscht werden müssen. Dieses gemeinsam trainierte Modell sollte besser funktionieren als ein Modell, das nur mit den Daten eines einzelnen Unternehmens trainiert wird.

- **Anwendungsfall #4:** Der Use Case “Einschleusen/Manipulieren von Komponenten in Industrienetzen” beschäftigt sich mit dem Erkennen von Angriffen in Industrienetzen. Dabei sollen durch Analyse der Netzwerkkommunikation und Prozessdaten der OT-Komponenten, Manipulationen von Engineering Workstations sowie Industriespionage durch neu eingebrachte Komponenten erkannt werden. Obwohl Föderierung zwischen verschiedenen Industrieanlagen denkbar wäre, liegt der Fokus in diesem Use Case auf der Föderierung von Daten aus den Domänen Netzwerkdaten und Prozessdaten.

- **Anwendungsfall #5:** Der Use Case “Angriffe auf Basis von Auth Protokollen” agiert im Bereich der Authentisierung von Web-App-Benutzern. Im Speziellen werden die Protokolle OAuth 2.0 und OIDC untersucht, die von NetID eingebunden werden. Ziel dabei ist es, mögliche Angriffe zu erkennen, die aufgrund von Fehlkonfiguration oder fehlerhaften Anwendung der verwendeten Protocol Flows ermöglicht wurden. Hierfür sollen auf Basis der horizontalen Föderierung Daten von allen beteiligten Protokoll-Endpunkten zusammengetragen werden. Ein Monitor soll möglichst zur Laufzeit den aktuellen Flow beobachten und eine Aussage über seine Vertrauensstufe treffen. Diese Information soll auf Basis der vertikalen Föderierung an die nächste Partei weitergegeben werden, z.B. den Webseiten Provider. Im Falle einer niedrigen Vertrauensstufe könnte diese Information zum Beispiel im Anwendungsfall #3 genutzt werden, um eine Aussage über einen möglichen Kreditkartenbetrug zu treffen und diesen ggf. zu unterbinden.

Nach einer Konsolidierung wurden die Anwendungsfälle von einzelnen Partnern in ihren technischen Teilarbeitspaketen, je nach Relevanz, verfolgt. Die HKA betrachtete dabei insbesondere die Anwendungsfälle “Erkennung von Missbrauch digitaler Identitäten” (mit Fokus auf Benutzer mobiler Endgeräte) und “Fraud Detection” (föderiertes Modelltraining zur Erkennung von Kreditkartenbetrug).

Ebenfalls im Kontext von AP1 wurde der KIWI-Gesamtdemonstrator konzipiert. Diesem liegt eine „Storyline“ zugrunde, die verschiedenen sicherheitskritischen Vorfälle im Alltagsleben einer imaginierten Person samt KI-basierten Gegenmaßnahmen zur Verhinderung der Angriffe aufzeigt. Die Funktionsweise solcher Maßnahmen wurde in einem Erklärvideo auch einem technisch weniger versierten Zielpublikum verständlich illustriert. Mit der professionellen Realisierung wurde seitens der HKA die Firma Mynd beauftragt.

In den Teilarbeitspaketen der HKA wurden die nachfolgend dargestellten Ergebnisse erzielt:

- **TAP 141 Beiträge zur KI-Infrastruktur mit Mobilitätsaspekten in der Gesamtarchitektur:** Es wurden fundamentale Architekturentwürfe, die sich auf neuartige Föderationsaspekte konzentrieren, durch die Analyse der KIWI-Anwendungsfälle unter Beteiligung aller Projektpartner identifiziert und aggregiert (M1.2). Ein Teilergebnis ist die KIWI-Gesamtarchitektur (s. Zwischenbericht 2021). Zudem wurden praktische Anteile der in den Teilarbeitspaketen behandelten Fragestellungen konkretisiert. Insbesondere wurden spezialisierte Architekturen für die sichere Peer-to-Peer Interaktion (siehe Verwertungsplan, [3, 4, 6]) sowie Data Governance Mechanismen (siehe Verwertungsplan, [3, 5, 6]) im Kontext von föderiertem maschinellem Lernen sowie die Authentifizierung mobiler Nutzer anhand ihrer Verhaltensmuster (siehe Verwertungsplan, [1, 2, 3, 6, 8]) erarbeitet. Die Basis für die Authentifizierung mobiler Nutzer bildet ein neu entwickeltes Authentifizierungsprotokoll (siehe Verwertungsplan, [8]), was einen speziellen Fokus auf den Schutz der Privatsphäre der Nutzer legt. Hierbei werden die personenbezogenen Daten direkt auf dem mobilen Gerät des Nutzers homomorph verschlüsselt und nur in verschlüsselter Form während des Authentifizierungsprozesses verarbeitet. Dadurch kann die Privatsphäre der Nutzer beweisbar geschützt werden (siehe Verwertungsplan, [8]). In diesem Kontext wurde auch ein neuartiges Verfahren vorgestellt, welches es ermöglicht das Ergebnis der Analyse der verschlüsselten Daten sicher und manipulationssicher offenzulegen (siehe Verwertungsplan, [12]).

- **TAP 142 POC und Demonstrator für föderierte KI-Data-Governance Methoden:** Basierend auf der Arbeit in TAP 442 wurden Metadaten von Experimenten in Kombination mit dem Flower-Federated Learning Framework gesammelt. Diese Metadaten spiegeln die verschiedenen Aspekte wider, die für das föderierte Training vereinbart wurden, sowie Informationen über die operativen Aufgaben der einzelnen Teilnehmer. Was die Vereinbarungen betrifft, so konzentrierten wir uns auf diejenigen, die zur Definition von tabellarischen Daten zwischen den Teilnehmern verwendet wurden. Es wurde eine Vielzahl

von Tabellendatensätzen verwendet, um das gesamte Spektrum der Hauptmerkmale von Tabellendaten zu erfassen. Ziel war es, alle möglichen Merkmale abzubilden, die Dateningenieuren oder Forschern im Sicherheitsbereich von webbasierten Systemen begegnen können. In Zukunft werden wir daran arbeiten, Datensätze zu finden, die besser auf den Sicherheitsbereich abgestimmt sind. Schließlich speichern wir auch alle von den Teilnehmern durchgeführten Operationen. Dies geschieht, um Informationen über solche Aktivitäten zu überprüfen, die die Teilnehmer im Verbund durchführen. Dieser Aspekt wurde in einem Artikel veröffentlicht [11].
Ebenfalls auf der Grundlage der Arbeit in TAP 442, aber in Form von Richtlinien, wurden verschiedene Experimente durchgeführt, um Metadaten zu sammeln und die Ergebnisse der föderierten Ausbildung zu überprüfen. Insbesondere werden Metadaten über die Leistung des Modells für jeden der Teilnehmer aufgezeichnet, sowohl auf einem zentralen Server als auch dezentral (jeder Teilnehmer testet das Modell auf der Grundlage seiner eigenen Daten). Darüber hinaus messen wir auch die Beiträge jedes Teilnehmers auf einem zentralisierten Server. Diese Aspekte sind auch in der Publikation [10] zu finden.
Anhand eines Demonstrators, der auf der Basis der Flower-Architektur realisiert wurde, können verschiedene Algorithmen getestet werden. Der Demonstrator enthält Beispiele für ausgewogene und unausgewogene Verteilungen der Daten zwischen den Teilnehmern und zeigt die Leistung der Modelle im Vergleich zur Verteilung der Daten. Außerdem wurden Arbeiten aus TAP242 und TAP443 in den Versuchsrahmen einbezogen. Dabei handelt es sich um Methoden, die die Beiträge der verschiedenen Teilnehmer zum gemeinsamen Problem des föderierten Lernens berechnen. Außerdem dienen solche Methoden dazu, zu überprüfen, wie sich die Unausgewogenheit der Daten auf solche Berechnungen auswirken könnte.
Was schließlich den Verhandlungsprozess betrifft, so ermöglicht die Entwicklung des Data-Governance-Cockpits nun, sowohl die Data-Governance-Modelle (Strategie und Qualitätsanforderungen, keine restriktiven Richtlinien) als auch Trainingskonfigurationen (zu verwendende Datensätze und ML-Modelle) zu definieren. Mehrere Operationen können sowohl direkt im Cockpit über die REST-API als auch über das Web-Dashboard (bevorzugte Methode für Kunden) durchgeführt werden, um die verschiedenen Governance-Elemente der Diskussion zu erzeugen. Diese Operationen entsprechen dem in der Veröffentlichung [11] vorgeschlagenen theoretischen Rahmen.

- **TAP 143 POC und Demonstrator Erkennung von Bedrohungen durch mobile Systeme:** Anhand verschiedener öffentlicher Datensätze wurde dargelegt, dass die entwickelten KI-Verfahren, welche das Nutzerverhalten analysieren, dafür geeignet sind, um unberechtigte Zugänge zu erkennen (siehe Verwertungsplan, [1, 2, 3, 6]). Weiterhin wurde durch ein umfassendes Experiment (siehe Verwertungsplan, [8]) demonstriert, dass das neu entworfene Authentifizierungsprotokoll (TAP 141) in der Praxis eingesetzt werden kann. Hierbei wurden im Detail die Genauigkeit der Authentifizierung, der anfallende Rechenaufwand, die notwendige Netzwerkkommunikation und der Schutz der Privatsphäre untersucht. Zusammen mit dem prototypisch umgesetzten Monitoring, welches in der Lage ist, Verhaltensdaten in mobilen Anwendungen zu sammeln, können existierende Anwendungen auf einfache Weise auf eine innovative Authentifizierung umgerüstet werden.

- **TAP 144 POC und Demonstrator für additives, horizontales Lernen von KI-Modellen:** Im Rahmen des Projekts wurde anhand einer Architektur ein Miniframework für das horizontale Lernen von KI-Modellen entwickelt. Dieses ermöglicht es mehreren Parteien / Organisationen zusammen ein föderiertes KI-Modell zu trainieren, um ein gemeinsames Problem zu lösen. Die lokalen sensitiven Daten der Teilnehmer werden dabei nicht zwischen den Teilnehmern ausgetauscht. Das Training adaptiert ein P2P-Protokoll, bei dem die Teilnehmer zu jedem Zeitpunkt in das Training einsteigen oder das Training verlassen können. Hierbei trainieren die Teilnehmer ein KI-Modell jeweils eine volle Trainings-

runde lokal auf ihren eigenen Daten und tauschen anschließend ihre errechneten Modelgewichte aus. Dabei wird das in Abbildung 2 dargestellte Broadcast Kommunikationsschema verwendet. Anschließend aggregiert jeder Teilnehmer die erhaltenen Gewichte mit seinen eigenen und erhält dadurch das föderierte Modell, welches als Basis für die neue lokale Trainingsrunde dient. Das iterative Training endet sobald eine gewisse Anzahl Trainingsrunden absolviert ist oder das Modell eine vorab festgelegte Qualität erreicht hat. Um die sensitiven Daten zusätzlich zu schützen wurden zudem Sicherheitsmethoden, wie Secure Multi-Party Computation (SMPC), Differential Privacy (DP) und Private Set Intersection in das Framework integriert. Diese wurden ausführlich auf ihre Schwachpunkte und Nutzen untersucht, um ein finales Fazit der besten Methoden zu erhalten.

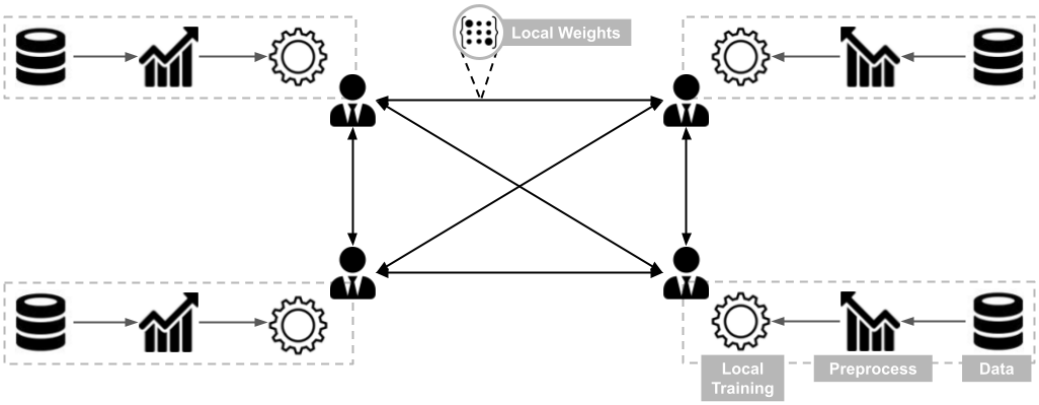

Abbildung 2: P2P-Föderiertes Lernen - Protokoll für vier Teilnehmern

2.2 Arbeitspaket 2 Einsetzbarkeit KI Verfahren

- **TAP 241 Identifikation von KI-Detektoren zur Erkennung von Bedrohungen durch mobile Systeme:** Anhand verwandter Arbeiten und explorativen Auswertungen von KI-Detektoren wurden Verfahren des maschinellen Lernens ausgewählt, die dafür geeignet sind, um Nutzer anhand ihres Verhaltens zu erkennen (siehe Verwertungsplan, [1] u. [2]). Dies wiederrum ermöglicht die Authentifizierung von Nutzern in mobilen Anwendungen und damit auch die Verwendung der Modelle für die Erkennung von Identitätsmissbrauch. Der Fokus lag hierbei auf Recurrent Neural Networks (RNNs) (siehe Verwertungsplan, [1]) und auf Decision Trees (siehe Verwertungsplan, [2]). Durch den Einsatz von homomorpher Verschlüsselung zum Schutz der Privatsphäre der Nutzer ergaben sich jedoch Einschränkungen, da nicht alle Modelltypen effizient auf homomorph verschlüsselten Daten arbeiten können. Aus diesem Grund wurden im Kontext des Authentifizierungsprotokolls auf Convolutional Neural Networks (CNNs) zurückgegriffen (siehe Verwertungsplan, [8]). Darüber hinaus wurden Methoden erforscht, wie die Inferenzen von neuronalen Netzen auf homomorph verschlüsselten Daten beschleunigt werden können. Beim Design der KI-Detektoren wurde darauf geachtet, dass diese für Szenarien mit horizontaler Föderation geeignet sind.
  Dies wurde dadurch ermöglicht, dass die Modelle nicht nutzerspezifisch trainiert werden (ein Modell pro Nutzer, das spezifisch dessen Verhalten erkennt), sondern für alle Nutzer das gleiche Modell trainiert und verwendet wird. Außerdem wird der Aspekt der vertikalen Föderation berücksichtigt, indem verschiedene Modelle zusammengeschalten werden können, welche verschiedene Aspekte des Nutzerverhaltens betrachten (bspw. Standorte und Netzwerkverbindungen des mobilen Geräts).

- **TAP 242: Qualitätsfaktoren föderierbarer KI-Methoden:** Die Arbeiten begannen mit Verzögerung am zweiten Oktober 2020 und wurden dann wie geplant durchgeführt. Die Bearbeitung endete im Juli 2021 mit einem geschätzten Aufwand von sechs Monaten. Die Forschungsarbeiten wurden mit einem Vorschlag für ein Modell zur Verwaltung der Datenqualität von ML-Modellen abgeschlossen (M2.1). In erster Instanz haben sich die

traditionellen Metriken der Datenqualität als unzureichend erwiesen, um sie unmittelbar für föderiertes Lernen anzuwenden. Daher haben wir uns auf die Erforschung anderer Möglichkeiten zur Messung der Datenqualität konzentriert. Als möglicher alternativer Ansatz mit großem Potential wurden Shapley-Werte identifiziert. Shapley-Werte sind ein Konzept aus der Spieltheorie, mit dem die Beiträge der verschiedenen Teilnehmer zum Ergebnis der gemeinsamen Bemühungen (Training) berechnet werden können. Durch Verwendung dieses Konzepts auf lokaler Modellebene kann die Eignung der Daten durch Mischung des Modellwerts mit zusätzlichen Informationen der Daten abgeleitet werden. Darüber hinaus hat dieser Ansatz die weitere Forschung zur Bewertung der Qualität von ML-Modellen motiviert. Aus diesem Grund wurden auch einige Informationen darüber, wie ML-Modelle zu bewerten sind, in das Qualitätsmodell aufgenommen. Die letzte gesammelte Information in Qualitätsmodellen betrifft die Notwendigkeit einer guten Rückverfolgbarkeit von Experimenten, wenn es um die Bewertung von ML-Modellen geht. Daher muss auch ein geeigneter Weg gefunden werden, um alle erzeugten Metadaten zu speichern und sie abzufragen, um die Folgen einer Änderung zu verstehen. Alle Ergebnisse wurden in einen internen Bericht zusammengefasst, der später in TAP 442 und TAP 443 verwendet werden wird. Dieser Bericht wurde bereits verfasst und ist als interne Ressource für die übrigen Teilnehmer verfügbar.

2.3 Arbeitspaket 3: Operationalisierung föderierter KI Detektoren

- **TAP 341 Horizontale Föderation von KI-Detektoren für mobile Systeme in der Netzwerk-Domäne:** Die Struktur der trainierten Modelle, welche zur Nutzerauthentifizierung verwendet werden (vgl. TAP 241), ermöglichen die horizontale Föderation. Außerdem hat das neu entwickelte Authentifizierungsprotokoll keinen Einfluss auf den Ablauf des Trainings, wodurch die Möglichkeit zur horizontalen Föderation unangetastet bleibt. Diese ist eng gekoppelt mit den Erkenntnissen aus den Teilarbeitspaketen TAP 343 und TAP 344. Mit Hilfe der entwickelten Protokolle können mehrere Entitäten föderiert KI-Detektoren trainieren. Dadurch können robustere Modelle zum Authentifizieren von Nutzern etabliert werden. Beispielsweise ist es hierdurch auf einfache Weise möglich, dass Anbieter verschiedener mobiler Anwendungen gemeinsam KI-Detektoren für die Nutzerauthentifizierung trainieren.

- **TAP 342 Vertikale Föderation von KI-Detektoren für mobile Systeme:** Die Möglichkeit zur vertikalen Föderation ist als zentrale Eigenschaft in das Konzept für die Nutzerauthentifizierung eingebettet und damit im entwickelten Authentifizierungsprotokoll berücksichtigt (siehe Verwertungsplan, [8]). Diese basiert darauf, dass verschiedene Aspekte des Nutzerverhaltens mit Hilfe unterschiedlicher KI-Detektoren analysiert werden. Hierbei können auch verschiedene Verfahren des maschinellen Lernens eingesetzt werden. Ein Beispiel hierfür ist die Analyse der Standorte und die Analyse von Touchscreen-Eingaben anhand von spezialisierten neuronalen Netzen, was in den Publikationen (siehe Verwertungsplan, [1], [2] u. [8]) demonstriert wurde. Die Zusammenführung der Ergebnisse der verschiedenen Modelle als Basis für eine finale Authentifizierungsentscheidung wurde experimentell umgesetzt (siehe Verwertungsplan, [8]). Hierbei wurden einfache arithmetische Verfahren verwendet (z. B. Bildung des Durchschnitts) und es wurde gezeigt, dass dadurch die Genauigkeit der Authentifizierung verbessert werden kann. Abschließend wurde daher ermöglicht, ein multimodales Authentifizierungssystem zu etablieren, das die Schwächen einzelner Authentifizierungsfaktoren (z.B. Bewegungsmuster des Nutzers) durch die Einbindung weiterer Faktoren abmildert.

- **TAP 343 Anwendbarkeit des horizontalen, additiven Trainings:** Im Rahmen des Projekts wurden Architektur-, Laufzeit- als auch Sicherheitsanalysen durchgeführt. Hierbei wurden die Unterschiede zwischen zentralem und dem entwickelten dezentralen P2P-Framework für föderiertes Lernen verdeutlicht. Die Forschung bezüglich der Architektur ergab, dass für eine sinnvolle und sichere Zusammenarbeit der Teilnehmer, jeder Teil-

nehmer sowohl als Server als auch als Client agieren muss. Die Teilnehmer müssen entsprechend dazu in der Lage sein, die Funktionalität eines zentralen Servers selbst bereitzustellen und entsprechende Kommunikation als auch Aggregation zu ermöglichen. Entsprechend der Funktionalitätsreplikation auf jedem Teilnehmer, bedeutet dies eine höhere Laufzeit und Rechenlast pro Teilnehmer. Hingegen der ursprünglichen Erwartungen führte die Dezentralisierung der Föderation zu Sicherheitsproblemen. Statt einem potenziellen Angreifer in Form des Servers, kann nun jeder Teilnehmer durch die erbrachte Serverfunktionalität als Angreifer agieren. Dennoch ist durch die Dezentralisierung eine Abwehr gegenüber Isolationsangriffen gewährleistet. Um die tatsächliche Gefahr durch Angriffe zu evaluieren, wurde das Framework unter Verwendung von verschiedenen Datensätzen und Neuronalen Netzwerk Architekturen untersucht. Die Experimente ergaben, dass ein Großteil der zur Verfügung stehenden Angriffe tatsächlich keine große Gefahr für unser Framework darstellt. Rekonstruktionsangriffe sind in der Forschung meist in Kombination mit FedSGD-Ansätzen aufzufinden. Die Föderation ist bei FedSGD so konfiguriert, dass die Teilnehmer bereits ihre Modelgewichte im Anschluss an eine Trainingsrunde auf einer kleinen Schnittmenge der lokalen Daten austauschen. Im Gegenzug verwendet das hier Konzipierte Framework einen FedAVG-Ansatz, bei welchem alle lokalen Daten in ein Modelupdate einfließen. Somit sind zu viele verschiedene Informationen in einem Update vorhanden, um Trainingsdaten zu rekonstruieren. Trotz FedAVG gelang es in einigen Angriffen die verwendeten Bildtrainingsdaten zu rekonstruieren, während eine ähnliche Rekonstruktion bei tabellarischen Daten aktuell unmöglich erscheint. Stattdessen ergaben die Experimente, dass bereits bei der Vorverarbeitung tabellarischer Daten sensitive Daten preisgegeben werden. So müssen beispielsweise textbasierende Features in numerische Daten transformiert und codiert werden. Hierbei müssen die Teilnehmer Klartext-Daten austauschen, um gemeinsam ein Wörterbuch mit einem zugehörigen numerischen Mapping für jeden Eintrag zu erstellen. Entsprechende gegenmaßen wurden in TAP 344 untersucht und implementiert.

- **TAP 344 Protokolle fürs horizontale, additive Training von KI-Modellen:** Mit Hilfe der Sicherheitsanalyse aus TAP 343 konnte die Effektivität von Angriffen auf das entwickelte Framework gemessen werden. Die tatsächlich funktionierenden Angriffe sollten mit Hilfe von Verfahren wie Secure Multi-Party Computation (SMPC) oder Differential Privacy (DP) verhindert werden. Diesbezüglich wurden verschiedene Varianten implementiert und eine erneute Evaluierung der Angriffe unter Verwendung der Sicherheitsmechanismen wurde durchgeführt. In Rekonstruktionsattacken auf Bilder gelang trotz Verwendung von SMPC, wie zuvor, Klassenpräsentanten der Trainingsteilnehmer zu generieren. Somit war keine tatsächliche Rekonstruktion möglich, allerdings besteht dennoch eine Verletzung der Privatsphäre. Ein ähnliches Resultat konnte bei den Membership-Attacken festgestellt werden. Hierbei gelang es tatsächlich noch bessere Aussagen über das Vorhandensein einzelner Trainingsdateninstanzen zu bestimmen. Die Verwendung von DP konnte einen solchen Eingriff in die Privatsphäre unterbinden. Es fand sich in den Experimenten ein Angriff, der durch beide Verfahren nicht verhindert werden konnte. Dabei handelt es sich um die sogenannte One-Shot Attacke, welche eine exakte Rekonstruktion von Bildern ermöglicht, wie in Abbildung 3 dargestellt.

Abbildung 3: Ergebnis einer One-Shot Attacke (rekonstruierte Bilder unten)

Wie zuvor erwähnt, lag das Sicherheitsproblem bei tabellarischen Daten an der Vorverarbeitung und nicht an dem tatsächlichen Training. Zur Behebung des Problems wurde PSI verwendet, um identische, kategoriale Merkmalswerte aller Trainingsteilnehmer einheitlich zu kodieren. Darüber hinaus wurde eine Methode zur Sicherung des Z-Score-Normalisierungsschritts entwickelt, die eine konsistente und einheitliche Vorverarbeitung numerischer Merkmalswerte für alle Teilnehmer garantiert. Zum Schutz der Berechnung der erforderlichen Metriken, wie Standardabweichung und Mittelwerte der numerischen Merkmale, diente wiederum SMPC.

2.4 Arbeitspaket 4: Datenmanagement und -analyse

- **TAP 441 Föderierte KI-Data-Governance:** Abbildung 4 Das wichtigste Ergebnis dieses Arbeitspakets ist die Definition eines Data-Governance-Lebenszyklus, der in Abbildung 4 dargestellt ist. Der Lebenszyklus umfasst drei Phasen. Die erste Phase besteht aus der Aushandlung des Data-Governance-Modells, das die zu erreichenden Ziele enthält, und der Trainingskonfiguration. In der zweiten Phase folgt eine Bewertung des erstellten Modells und der Qualität der Beiträge der Teilnehmer. Da es wahrscheinlich ist, dass die ersten beiden Schritte einige Iterationen erfordern werden, um ein zufriedenstellendes Modell zu erhalten, werden die Metadaten verwaltet und den Teilnehmern zur Verfügung gestellt. So können sie beurteilen und entscheiden, was bei der nächsten Iteration zu ändern ist. Alle Inhalte, die sich auf das Data-Governance-Modell und den Lebenszyklus beziehen, wurden in der Publikation [11] veröffentlicht. Seitdem hat es zwei kleine Änderungen gegeben: Erstens ist während der Qualitätsphase kein Symbol mehr in der Abbildung zu sehen. Da die Metadaten in den ersten beiden Phasen gespeichert werden, könnte das Symbol zu einer gewissen Verwirrung führen. Zweitens gibt es jetzt insgesamt fünf Arten von Metadaten: Data Governance Model, Training Configuration, Results, Descriptive und Provenance. Die ersten beiden repräsentieren alle Instanzen aller Elemente, die während des Diskussionsprozesses diskutiert wurden. Die dritte Gruppe stellt die Ergebnisse des Schulungsprozesses nach den Vereinbarungen der ersten beiden Phasen dar. Der vierte Bereich enthält alle zusätzlichen Informationen, die nach Ansicht der Teilnehmer zum Verständnis der ausgetauschten Metadaten beitragen können. Schließlich werden in Provenance die Änderungen und Autoren der Interaktionen während des Lebenszyklus festgehalten. Diese Änderungen bilden den Abschluss der Data Governance für das Modell des föderierten maschinellen Lernens.

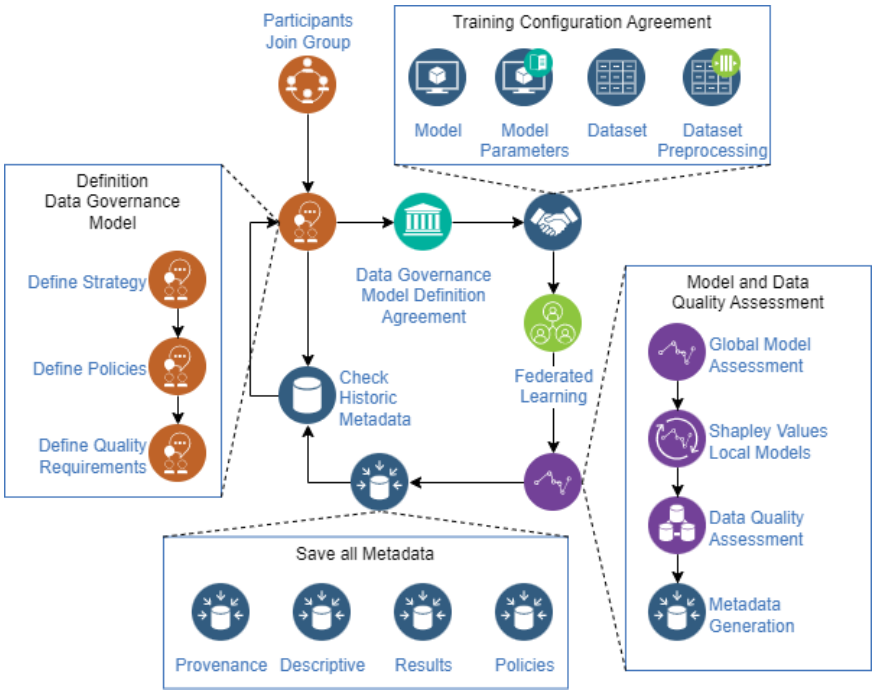

Abbildung 4: Federated Machine Learning Data Governance Lifecycle.

- **TAP 442 Föderierte Metadaten- und Regelverwaltung:** Es gliedert sich in zwei Teile: 1) Metadaten im Zusammenhang mit der Governance des föderierten Lernens zur Verifizierbarkeit und Unterstützung von Artefakten und 2) Definition von Richtlinien zur Festlegung von Zielen und zur Beschränkung des Zugangs zu Artefakten im Verbund.
  Die Metadaten im Zusammenhang mit der Governance werden in Form eines Metadatenmodells und eines Metadatenmanagementsystems konzipiert. Ersteres definiert die Metadaten, die den Teilnehmern zur Verfügung stehen, sowohl im Hinblick auf die Provenienz für die Nachvollziehbarkeit als auch auf die Verfolgung von Experimenten für kontinuierliche Verbesserungen. Es wurden verschiedene Experimente durchgeführt und veröffentlicht [10]. Die weitere Implementierung des Metadaten-Managementsystems wird in TAP 142 dargestellt.
  Was die Richtlinien betrifft, so wurden mehrere Ansätze für die Richtlinien verfolgt. Im Hinblick auf die Kontrollfunktionen wurde Keycloak integriert. Keycloak stellt sowohl Authentifizierungs- als auch Autorisierungsmechanismen für das Data Governance Cockpit zur Verfügung und blockiert jeden Versuch eines unautorisierten Zugriffs von externen Agenten der Föderation.
  In einem parallelen Ansatz analysierten wir, ob KI-Modelle für alle Teilnehmer einer Föderation gut funktionieren. Je nach der Verteilung der Daten auf die Teilnehmer eines Verbunds kann sich dies auf das resultierende KI-Modell auswirken. Dies geschieht insbesondere in statistisch nicht unabhängigen und identisch verteilten (non-IID) Szenarien. Der Einsatz von selbstüberwachten Lerntechniken mildert das Problem, löst es aber nicht vollständig. Daher wurde eine Kombination aus zentralisierten und dezentralisierten Metriken entwickelt, um sicherzustellen, dass die Teilnehmer das KI-Modell in ihrem eigenen Bereich ordnungsgemäß testen können. Darüber hinaus wurden zusätzliche Tests mit einer Client-Auswahl für den Bewertungsprozess durchgeführt. Die Ergebnisse zeigen, dass bei Cross-Silo-Szenarien die Client-Auswahl in der Regel dazu führt, dass alle Teilnehmer ausgewählt werden, selbst bei stark unausgewogenen Verteilungen. Dies beweist, wie ungeeignet solche Verfahren für siloübergreifende Szenarien sind.

- **TAP 443 KI-Daten-Management*:*** Auf der Grundlage des in TAP 441 definierten Datenverwaltungsmodells und der in TAP 442 entwickelten Infrastruktur konzentrierte sich die-

ses Arbeitspaket auf die Entwicklung einer Softwareplattform zur Durchsetzung der Datenverwaltung in verschiedenen Anwendungsfällen. Dabei werden auch Ergebnisse aus TAP 242 einbezogen, um die Qualität der Daten der Teilnehmer zu bewerten.
Das Governance-Cockpit wurde entwickelt, um die Abläufe zu unterstützen, die für die Verhandlung von Federated Learning-Schulungen erforderlich sind. Hier können die Teilnehmer Strategien erstellen, Vorschläge für Qualitätsanforderungen machen und darüber abstimmen, wie mit den Trainingseinstellungen verfahren werden soll. Die Vorschläge umfassen auch die Trainingskonfigurationen mit allen Parametern, die für die Definition der Daten und des beim Training zu verwendenden Modells für maschinelles Lernen erforderlich sind. Schließlich werden sowohl die Ergebnisse als auch die Beiträge gesammelt und den Teilnehmern zur Verfügung gestellt, indem das Cockpit mit dem Metadaten-Management-System verbunden wird.
Der Entwurf einer dezentralen, auf Blockchain-Technologien basierenden Governance-Variante wurde zugunsten einer zentralisierten Variante zurückgestellt. Der zentralisierte Entwurf wurde so konzipiert, dass er auch auf dezentrale Architekturen angewendet werden kann. Sie würde eine Replikation von Daten und Metadaten zwischen den Teilnehmern erfordern, was jedoch im Rahmen des Projekts als nicht möglich erachtet wurde.
Die Anwendung der Ergebnisse von TAP 242 in diesem Arbeitspaket bezieht sich auf eine Lücke in der Literatur zur Auswertung von Daten. Die Auswertung von Teilnehmerdaten an einer zentralen Stelle ist mit Hilfe der so genannten Shapley-Werte unter der Bedingung identisch und unabhängig verteilter Daten (IID) möglich. In einem IID-Szenario haben die Daten aller Teilnehmer ähnliche Werte und alle Klassen sind bekannt. Das bedeutet aber auch, dass die Teilnehmer nur wenig voneinander lernen können. Non-IID-Fälle hingegen spiegeln Szenarien wider, in denen einige Teilnehmer völlig andere Werte haben oder neue Klassen einführen. Die übrigen Teilnehmer sind daran interessiert, Muster aus diesen Daten zu lernen, die sie selbst nicht besitzen. Aber auch dieses Szenario führt zu Problemen in der Lernphase, und das KI-Modell kann am Ende schlechter abschneiden als in IID-Szenarien. Selbst wenn ein KI-Modell mit zusätzlichen Klassen gut abschneiden würde, könnte der entsprechende Teilnehmer mit speziellen Daten unterschätzt werden, wenn der Bewertungsprozess die speziellen Klassen nicht ausreichend berücksichtigt.
In dieser Hinsicht wurden bereits zahlreiche Untersuchungen durchgeführt. Es hat sich jedoch als ungeeignet für Szenarien erwiesen, in denen heterogene tabellarische Daten verwendet werden. Daher ist zum jetzigen Zeitpunkt die einzige geeignete Option die Verwendung eines zentralen Testdatensatzes als Goldstandard für die Datenqualität und Datenbewertung. Ein solcher zentraler Datensatz muss von den Teilnehmern vereinbart werden, entweder durch die Berücksichtigung eines öffentlichen Datensatzes oder durch die gemeinsame Nutzung synthetischer Daten.

- **TAP 444 Anforderungen ans Meta-Datenmanagement für das sichere Modelltraining:** Hierbei wurde das föderierte Training auf Schwachstellen unter Berücksichtigung verschiedener Datensätze, wie Bilddaten oder tabellarische Daten, untersucht. Diese Schwachstellen galt es zu beheben oder mit entsprechenden Sicherheitsmechanismen zu unterbinden. Insbesondere bei der Vorverarbeitung tabellarischer Daten müssen Informationen über einzelne Merkmale der Trainings- und Testdatensätze vorliegen. Bei tabellarischen Daten müssen die Teilnehmer diskrete Merkmale, wie Wörter und Bezeichnungen, austauschen. Dies ermöglicht die Bildung eines gemeinsamen Wörterbuchs, sodass solche diskreten Merkmale fürs Training in numerische Werte umgewandelt werden können. Der Klartextaustausch stellt jedoch ein Datenschutzproblem dar, denn bei diesen Merkmalen kann es sich um sensitive Daten handeln, welche die Teilnehmer nicht offenlegen wollen. Basierend auf den Metadaten muss unterschieden werden können, welche Merkmale eine spezifische abgesicherte Vorverarbeitung benötigen. Die in TAP 343 entwickelten Sicherheitsmechanismen verwenden bei kategorischen Werten eine Private Set Intersection (PSI) während numerischer Werte den Einsatz von Secure Multi-Party Computation (SMPC) erfordern, um eine sichere Stauchung der Werte mittels Z-Score Normalisierung zu garantieren.

- **TAP 445 Framework fürs Lebenszyklusmanagement föderiert trainierter KI-Modelle:** Für das Live Deployment eines föderiert trainierten KI-Modells wurde ein Konzept ausgearbeitet. Ein KI-Modell erreicht dabei das Ende seines Lebenszyklus, sobald die Entitäten der Föderierung ein neues Modell trainieren, welches eine gleiche oder höhere Qualität gegenüber dem aktuellen Modell bietet. Die Qualität selbst wird anhand des F1-Scores basierend auf Prognosen von Testdaten gemessen. Das alte Modell wird anschließend als Artefakt gespeichert. Dadurch bietet sich eine Fall-back Option an, falls das neue Modell unzureichende Resultate im Live Deployment erzielt. Um den stetigen Wandel der zu prognostizierenden realen Daten entgegenzuwirken, sollten die Trainings- und Testdaten entsprechend häufig angepasst bzw. aktualisiert werden.

2.5 Arbeitspaket 5: Evaluation und Erprobung

- **TAP 541 Evaluation föderierter KI-Data-Governance Mechanismen***:* Die Bewertung des Data-Governance-Cockpits bestand aus mehreren Simulationen. Die Simulationen bestanden aus mehreren Parteien, die sich auf das Training eines föderierten Lernmodells einigten, indem sie als Referenz einen der zahlreichen Datensätze verwendeten, die für die Experimente in den oben erwähnten TAPs verwendet wurden. Diese Datensätze stellen realistische Datensätze dar, die in den Szenarien für die verschiedenen Anwendungsfälle des Projekts zu finden sind. Dann wird um diese Datensätze herum ein Verhandlungsprozess sowohl für die Strategie als auch für die Trainingskonfiguration simuliert.
  Verschiedene Bewertungsmetriken wurden unter verschiedenen Heterogenitätsgraden analysiert. Mehrere Szenarien mit ausgeglichener und unausgeglichener Datenverteilung wurden mit der Flower-Architektur ausgeführt. Der aktuellen Literatur zufolge sollte die Leistung von KI-Modellen bei hoher Heterogenität abnehmen. Dies gilt für bestimmte Fälle, in denen die Kennzeichnungen nicht gleichmäßig auf die Teilnehmer verteilt sind. Dies ist jedoch nicht immer der Fall, und es kann von Vorteil sein, Teilnehmer mit zusätzlichen Daten zu haben. Es ist erwähnenswert, dass Techniken des selbstüberwachten Lernens dazu beitragen können, diese Leistungsverschlechterung abzumildern und das Training lokal zu verbessern, wenn die Daten für bestimmte Klassen zu knapp sind. Allerdings ist das Verfahren nicht in der Lage, die Datenvielfalt der Teilnehmer zu messen, weder bei der Auswertung noch bei der Bewertung. Es ist auch nicht möglich zu sehen, ob ein Teilnehmer von neuen oder externen Daten profitiert, die er (erst) in der Zukunft selbst finden könnte. Darüber hinaus zeigen andere getestete Metriken, wie z. B. Abstandsverteilungen und Kosinusähnlichkeit, eine schlechte Leistung bei der Arbeit mit heterogenen Tabellendaten. Dies schränkt die Möglichkeiten der Teilnehmer ein, wenn sie das Modell und die Beiträge richtig bewerten wollen. Zurzeit gibt es keine bessere Lösung für die Bewertung als einen zentralen Datensatz, der als Referenzpunkt für die Leistung dient.

- **TAP 542 Evaluation Erkennung von Bedrohungen durch mobile Systeme:** Im Projekt wurde gezeigt, dass Recurrent Neural Networks (siehe Verwertungsplan, [1]) geeignet sind, um das Touchscreen-Verhalten der Nutzer zu analysieren. Der Vorteil der von uns entwickelten Modelle liegt hierbei in den Skalierbarkeitseigenschaften, da kein spezifisches Modell pro Nutzer etabliert werden muss (siehe Verwertungsplan, [1]). Des Weiteren wurde untersucht, wie Nutzer anhand der Netzwerkverbindungen ihrer mobilen Geräte mit Entscheidungsbäumen (XGBoost) identifiziert werden können (siehe Verwertungsplan, [2]). Dabei wurden Methoden erprobt, die den Schutz der Privatsphäre der Nutzer verbessern können. Es zeigte sich zum Beispiel, dass konkrete Standorte oder Netzwerkinformationen durch Hashing verschleiert werden können (siehe Verwertungsplan, [2]). Schlussendlich wurde in einem umfassenden Experiment unser Authentifizierungsprotokoll auf Basis von homomorpher Verschlüsselung evaluiert. Es wurde ein existierender Datensatz, der innerhalb einer mobilen Anwendung mit echten Nutzern erhoben

wurde, verwendet, um ein kontinuierliches Authentifizierungsszenario zu simulieren. Zur Authentifizierung wurden Convolutional Neural Networks (CNNs) verwendet und es wurde neben der Genauigkeit auch die Performanz der Authentifizierung ausgewertet (siehe Verwertungsplan, [8]). Es zeigte sich, dass sowohl die Genauigkeit als auch die Performanz ausreichend sind, damit unser Ansatz in der Praxis eingesetzt werden kann. Außerdem wurde der Schutz der Privatsphäre analysiert und belegt (siehe Verwertungsplan, [8]). In diesem Zusammenhang wurde auch die Sicherheit und die Effizienz des neuartigen Verfahrens zur Offenlegung der finalen Authentifizierungsentscheidung, welches Teil des entwickelten Protokolls ist, untersucht und nachgewiesen (siehe Verwertungsplan, [12]).

- **TAP 543 Evaluation der Einsetzbarkeit des additiven Lernens in KIWI-Beispielszenarien:** Es wurden Evaluationstests des P2P-Training-Prototyps auf einer GPU durchgeführt. Dabei wurden in verschiedenen Szenarien IID und non-IID Datensätze sowie eine wechselnde Anzahl Teilnehmer verwendet. Die Resultate zeigten, dass der aktuelle Ansatz sehr gut funktioniert, falls die Daten gemäß IID bei den Teilnehmern vorliegen. Allerdings bestehen diverse Probleme in Non-IID-Anwendungsfällen, in denen nicht alle Teilnehmer über alle Kategorien des Klassifikationsproblems verfügen. Weiterhin wurde die Kommunikation zwischen den Teilnehmern optimiert, indem Remote Procedure Calls (RPC) statt einem REST-API über HTTP/S verwendet werden. Ebenfalls wurde die Verwendung von Gradienten-Stauchung untersucht, diese hatte jedoch gravierende Einflüsse auf die Modellqualität. Somit konnte eine Beschleunigung des Trainings nicht ohne Verlust der Qualität gewährleistet werden.

2.6 Arbeitspaket 6: Projektmanagement

- **TAP 641 Allgemeines Projektmanagement:** Die Aktivitäten umfassten das übergeordnete Projektmanagement bzw. Koordination des Gesamtkonsortiums. Es wurden während der gesamten Projektlaufzeit mehrere Gesamtprojekttreffen bei verschiedenen Gastgebern in Karlsruhe (in Präsenz) sowie monatliche turnusgemäße Treffen (meistens virtuell) auf Konsortiums- und auch Arbeitspaketebene organisiert und durchgeführt.

- **TAP 642 Dissemination:** Im Bereich der Öffentlichkeitsarbeit wurde unter Koordination des HKA-Teams der wissenschaftliche Workshop mit dem Titel "*International Workshop on AI for Web Application Infrastructure and Cloud Platform Security*" (kurz: AWACS), als Teil der Konferenz ESOCC 2022 erfolgreich durchgeführt.
  Darüber hinaus konnten mehrere Publikationen aus dem Gesamtkonsortium bzw. auch seitens des HKA-Teams an Fachtagungen veröffentlicht werden (s. Liste als Teil der Verwertungstabelle).
  Verschiedene sicherheitskritische Vorfälle im Alltagsleben einer imaginierten Person samt KI-basierten Gegenmaßnahmen zur Verhinderung der Angriffe wurden in einem Erklärvideo aufgezeigt. Das Erklärvideo (s. https://www.youtube.com/watch?v=lo-Vqje06NYs) illustriert die Funktionsweise solcher Maßnahmen auch einem technisch weniger versierten Zielpublikum verständlich.

3. Vergleich des Vorhabenverlaufs mit der ursprünglichen Planung

3.1 Einhaltung des Zeitplans

Die ursprünglich geplante Projektdauer war 01.06.2020 – 31.05.2023. Bedingt durch die CoViD19-Pandemie und die damit zusammenhängenden Einschränkungen verlangsamten sich Prozesse im Verwaltungswesen der HKA erheblich, was zu Anlaufschwierigkeiten unseres Teilvorhabens geführt hat. Insbesondere konnten unter den erschwerten Bedingungen keine neuen für die Projektdurchführung fachlich geeigneten Mitarbeiter fristgerecht eingestellt, Arbeitsplätze eingerichtet, Arbeitsmittel beschafft werden. Die dadurch entstandenen Verzögerungen konnten bis zum regulären Projektende nicht mehr ausgeglichen werden, weshalb eine kostenneutrale Projektverlängerung von sechs Monaten beantragt wurde. Dementsprechend endete das Projekt am 30.11.2023.

3.2 Angaben zur Mittelverwendung

Die projektbezogenen Ausgaben wurden in folgende Positionen des Zuwendungsbescheids vom 14.05.2020 unterteilt, wobei die Positionen 0817 und 0835 erst während der Projektlaufzeit, durch Umwidmung ursprünglich anderweitig eingeplanter Mittel relevant wurden.

- **0812 Beschäftigte E12-E15:** Die Arbeiten an der HKA wurden von drei in Vollzeit eingestellten wissenschaftlichen Mitarbeitern durchgeführt. Die beantragten Personalkosten dienten ausschließlich der Finanzierung der wissenschaftlichen Mitarbeiter gemäß der TV-L Entgelttabelle 2019 im Tarifgebiet West. Die Arbeitgeberleistungen wurden dabei mit 30% kalkuliert.

- **0817 Beschäftigte E1-E11:** Um finale, seitens der HKA vorgesehenen Implementierungsarbeiten fristgerecht abzuschließen, wurde ein weiterer Mitarbeiter zum 1.10.2023 (TV-L Entgelttabelle, Tarifgebiet West) eingestellt und im Projekt beschäftigt.

- **0822 Beschäftigungsentgelte:** Diese Mittel wurden für die Finanzierung von wissenschaftlichen Hilfskräften verwendet. Angesetzt wurden dabei ursprünglich monatliche Kosten in Höhe von 585 € pro Hilfskraft. Die Hilfskräfte haben die wissenschaftlichen Mitarbeiter hauptsächlich bei Recherche-, und Implementierungsarbeiten, der Betreuung technischer Systeme, Durchführung von Software- und System-Tests, etc. unterstützt. Zusätzlich wurden zur Entlastung der im Projekt involvierten Hochschulprofessoren (insbesondere bedingt durch die Koordination des Verbundvorhabens) durch die Fakultätsleitung externe Lehrbeauftragte engagiert bzw. aus Projektmitteln dieser Position (gemäß aktuell gültigen Stundensätzen) vergütet.

- **0835 Vergabe von Aufträgen:** Mittel aus dieser Position wurden für die Beauftragung eines externen Dienstleisters verwendet, um das Erklärvideo über das Gesamtprojekt zu konzipieren und zu produzieren. Der Dienstleister wurde nach Vergleich von insgesamt drei Angeboten in Rücksprache mit dem Projektträger ausgewählt.

- **0846 Dienstreisen:** Bei der ursprünglichen Reiseplanung und Kostenkalkulation wurden Dienstreisen mit Projektbezug im Inland (Projekttreffen) und im Ausland (Fachtagungen, Konferenzen) vorgesehen. Insbesondere bedingt durch die Covid19-Pandemie fanden einige Projekttreffen sowie Präsentationen von Beiträgen an Fachtagungen im Rahmen von Videokonferenzen statt, wodurch weniger Dienstreisen als geplant getätigt wurden.

4. Notwendigkeit und Angemessenheit geleisteter Projektarbeiten

Der Ansatz von KIWI stellt ein innovatives Konzept dar, insbesondere vor dem Hintergrund einer zunehmenden Fokussierung von Unternehmen auf KI-gestützte Anwendungen in datenintensiven Geschäftsprozessen. Eine notwendige Voraussetzung ist die Absicherung relevanter Datenbestände unter Sicherstellung des Datenschutzes und der Datensouveränität der Unternehmen. Die Umsetzung des föderierten Ansatzes ist allerdings kosten- und zeitintensiv, da entsprechende Vorgehensweisen, KI-Modelle und Architekturen entworfen aber auch konkrete Systeme (seitens der Unternehmenspartner) entwickelt werden müssen, die später zu marktreifen Produkten oder Services weiterentwickelt werden könnten. Hiermit waren und sind, vor allem seitens der industriellen Partner, hohe wirtschaftliche und wissenschaftlich-technische Risiken verbunden.
Infolge dessen konnte eine Entwicklung ohne Förderung nicht innerhalb der finanziellen Möglichkeiten der Projektpartner vorangetrieben werden.
Die Möglichkeit einer EU-Förderung wurde initial geprüft, war aber seinerzeit für das Vorhaben nicht in Sicht, weshalb das Konsortium eine Förderzuwendung durch das BMBF beantragt hat.

5. Verwertbarkeit des Ergebnisses

Die Verwertungstabelle fasst den Stand zum Projektende am 30.11.2023 bzgl. der einzelnen, im Projektantrag angegebenen Verwertungsziele der HKA zusammen.

Verwertungsziel	lfd. Nr.	Beschreibung
Erfindungsmeldungen, Patente	1	METHOD AND SYSTEM TO COLLABORATIVELY TRAIN DATA ANALYTICS MODEL PARAMETERS Application, Publication/Patent Number: EP4083838A1, Publication Date: 2022-11-02, Application Number: EP21171572.7, Filing Date: 2021-04-30
Wissenschaftliche Publikationen im Rahmen von Konferenzen und Workshops	2	Insgesamt wurden 12 wissenschaftliche Beiträge im Rahmen von Workshops und Konferenzen veröffentlicht, s. Liste im Kap. 6.
Wissenschaftliche Publikationen in Journals	3	Während der Projektlaufzeit wurden keine Journalartikel publiziert.
Anreicherung der Lehrinhalte	4	Durchführung insg. 17 studentischer Projekt- und 6 Thesis-Arbeiten im Projektkontext.
Festigung des Forschungsthemas KI an der Hochschule	5	Etablierung der Forschergruppe Datenzentrierte Softwaresysteme (DSS) am IAF (https://www.h-ka.de/iaf/dss). Erfolgreiche Akquise weiterer Drittmittel und Initiierung von Nachfolgeprojekte durch die in KIWI beteiligten HKA-Professoren, wie bspw. aura.ai (https://www.h-ka.de/en/iaf/aura-ai) oder ILKA (https://www.h-ka.de/iaf/ilka).
OSS Initiativen	6	Prototypische Implementierungen des HKA-Teams wurden teilweise in GitHub der interessierten Entwickler-Community zur Verfügung gestellt.
Ausgründungen	7	Während der Projektlaufzeit gab es keine Ausgründungen.
Meet-ups, Workshops, Diskussionsforen	8	Das Team der HKA agierte federführend bei der Durchführung des wissenschaftlichen AWACS Workshop (https://kiwi-project.org/AWACS) im Rahmen der wissenschaftlichen Konferenz ESOCC 2022.

6. Veröffentlichungen des Ergebnisses nach Nr. 5 der NABF

Im Rahmen des Teilvorhabens erfolgten folgende Veröffentlichungen der (Zwischen-) Ergebnisse:

[1] David Monschein, Oliver P. Waldhorst: “SPCAuth: Scalable and Privacy-Preserving Continuous Authentication for Web Applications.”, In Proc. 46th IEEE Conf. on Local Computer Networks (LCN), Virt. Conf., 2021.
[2] David Monschein, Oliver P. Waldhorst: “Privacy-Preserving and Scalable Authentication based on Network Connection Traces.” In Proc. GI/ITG Conf. on Networked Systems (NetSys), Virt. Conf., September 2021.
[3] David Monschein, José Antonio Peregrina Pérez, Tim Piotrowski, Zoltán Nochta, Oliver P. Waldhorst, Christian Zirpins: “Towards a Peer-to-Peer Federated Machine Learning Environment for Continuous Authentication.” In Proc. 1st IEEE Int. Workshop on Distributed and Intelligent Systems (DistInSys), co-located with 26th IEEE Symp. on Computers and Communications (ISCC), Athens, Greece, 2021.
[4] Tobias Wink, Zoltán Nochta: “An Approach for Peer-to-Peer Federated Learning”, In Proc. 51st Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W), 2021, Taipeh, Taiwan, June 2021.
[5] José Antonio Peregrina, Guadalupe Ortiz, Christian Zirpins: “Data Governance for Federated Machine Learning in Secure Web-based Systems”, In Actas de Las Jornadas de Investigación Predoctoral en Ingeniería Informática, 2021, pages 36-39
[6] David Monschein, José Antonio Peregrina Pérez, Tim Piotrowski, Zoltán Nochta, Oliver P. Waldhorst and Christian Zirpins: “KIWI: Artificial Intelligence in Secure Web Infrastructures”, Forschung aktuell 2021, Hochschule Karlsruhe, p. 40-43.
[7] Piotrowski T and Nochta Z. 2023. Towards a Secure Peer-to-Peer Federated Learning Framework. In Advances in Service-Oriented and Cloud Computing : International Workshops of ESOCC 2022 ; Revised Selected Papers (Communications in Computer and Information Science). Springer International Publishing, 19–31 https://doi.org/10.1007/978-3-031-23298-5_2
[8] D. Monschein and O. P. Waldhorst, mPSAuth: Privacy-Preserving and Scalable Authentication for Mobile Web Applications, in https://arxiv.org/abs/2210.04777, 2022.
[9] Zirpins C, Ortiz G, Nochta Z, Waldhorst O, Soldani J, Villari M, Tamburri D: Advances in Service-Oriented and Cloud Computing: International Workshops of ESOCC 2022; Revised Selected Papers. International Workshop on AI for Web Application Infrastructure and Cloud Platform Security (AWACS 2022) (Wittenberg, Germany, 22.-24.03.2022), Cham: Springer 2023 (Communications in Computer and Information Science 1617), X, 117 S.- ISBN 978-3-031-23297-8 (Elektronische Veröffentlichung: http://dx.doi.org/10.1007/978-3-031-23298-5)
[10] Peregrina Pérez JA, Ortiz G, Zirpins C: Towards a Metadata Management System for provenance, reproducibility and accountability in Federated Machine Learning. In: Zirpins C, Ortiz G, Nochta Z, Waldhorst O, Soldani J, Villari M, Tamburri D (Hrsg.): Advances in Service-Oriented and Cloud Computing : International Workshops of ESOCC 2022 ; Revised Selected Papers. International Workshop on AI for Web Application Infrastructure and Cloud Platform Security (AWACS 2022) (Wittenberg, 22.-24.03.2022), Cham: Springer 2023 (Communications in Computer and Information Science 1617), S. 5-18.- ISBN 978-3-031-23297-8 (Elektronische Veröffentlichung: http://dx.doi.org/10.1007/978-3-031-23298-5_1)
[11] Peregrina Pérez JA, Ortiz G, Zirpins C: Towards Data Governance for Federated Machine Learning. In: Zirpins C, Ortiz G, Nochta Z, Waldhorst O, Soldani J, Villari M, Tamburri D (Hrsg.): Advances in Service-Oriented and Cloud Computing : International Workshops of ESOCC 2022 ; Revised Selected Papers. International Workshop on AI for Web Application Infrastructure and Cloud Platform Security (AWACS 2022) (Wittenberg, 22.-24.03.2022), Cham: Springer 2023 (Communications in Computer and Information Science 1617), S. 59-71.- ISBN 978-3-031-23297-8 (Elektronische Veröffentlichung: http://dx.doi.org/10.1007/978-3-031-23298-5_5)
[12] Baumstark P, Monschein D, Waldhorst O: Secure Plaintext Acquisition of Homomorphically Encrypted Results for Remote Processing, 2023 IEEE 48th Conference on Local Computer Networks (LCN), Los Alamitos, CA, USA: IEEE Computer Society, Oct. 2023, pp. 1–4. doi: 10.1109/LCN58197.2023.10223372. (Elektronische Veröffentlichung: https://doi.ieeecomputersociety.org/10.1109/LCN58197.2023.10223372)