

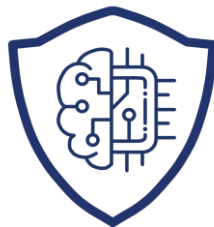
Sachbericht

Verbundvorhaben

SENSIBLE-KI

Sichere und vertrauenswürdige mobile KI

SENSIBLE-KI



Gefördert durch das Bundesministerium für Wirtschaft und Klimaschutz (BMWK)

Konsortialführer: Fraunhofer-Institut für Angewandte und Integrierte Sicherheit (AISEC)	Förderkennzeichen: 01MT21005A
Weitere Partner: Bundesdruckerei GmbH (bdr) Hochschule Darmstadt (h_da) neXenio GmbH	01MT21005B 01MT21005C 01MT21005D
Assoziierte Partner: Freie Universität Berlin (FUB) SpleenLab	
Laufzeit des Vorhabens: von: 01.03.2021 bis: 29.02.2024	

Inhaltsverzeichnis

1. Aufgabenstellung.....	3
2. Voraussetzungen, unter denen das Vorhaben durchgeführt wurde	3
3. Planung und Ablauf des Vorhabens.....	4
4. Wissenschaftlicher und technischer Stand, an den angeknüpft wurde	6
5. Zusammenarbeit mit anderen Stellen.....	6
6. Verwendung der Zuwendung und der erzielten Ergebnisse	8
7. Wichtigste Positionen des zahlenmäßigen Nachweises.....	10
8. Notwendigkeit und Angemessenheit der geleisteten Arbeit	11
9. Voraussichtlicher Nutzen und Verwertbarkeit.....	12
10. Fortschritt bei anderen Stellen.....	13
11. Erfolgte und geplante Veröffentlichungen.....	13
Literaturverzeichnis.....	15

I. Teil

1. Aufgabenstellung

Die Aufgabe des Verbundvorhabens SENSIBLE-KI bestand in der Erforschung von Lösungen, um Systeme der Künstlichen Intelligenz (KI) in mobilen Anwendungen und eingebetteten Systemen sicher und vertrauenswürdig zu gestalten. Unter anderem sollten dabei solche Sicherheitsaspekte berücksichtigt werden wie: Schutz vor Manipulation des KI-Systems (Integrität und Authentizität), Schutz vor ungewollter Extraktion oder Identifikation von Informationen aus dem KI-System (Vertraulichkeit und Privatheit) sowie sicheres Übertragen der KI auf ein mobiles Endgerät.

Methoden der Künstlichen Intelligenz kommen bereits in einer Vielzahl von Anwendungen zum Einsatz, beispielsweise in intelligenten Assistenzprogrammen oder in biometrischen Verifikationsverfahren. Bisher stehen noch keine einheitlichen Herangehensweisen zur Absicherung der KI-Systeme in mobilen und eingebetteten Systemen zur Verfügung, was zu Sicherheitslücken führen kann. Im Projekt SENSIBLE-KI wurden Methoden untersucht, um den Schutz von KI-Systemen und der von ihnen verarbeiteten Daten sicherzustellen und gleichzeitig eine sichere Übertragung der KI-Modelle auf mobile Endgeräte zu ermöglichen. Dafür wurden Lösungen entwickelt, um Angriffe auf KI-Systeme im mobilen und eingebetteten Kontext zu erschweren oder komplett zu verhindern. Konkret kamen hardwarebasierte Trusted-Computing-Verfahren sowie softwarebasierte Mechanismen zum Einsatz, die einerseits die Authentizität und Integrität, andererseits aber auch die Vertraulichkeit, Privatheit und Nachprüfbarkeit der Machine-Learning-(ML)-Modelle sicherstellen.

Die im Rahmen von SENSIBLE-KI entwickelten Methoden und Ansätze wurden von den Industriepartnern mit Unterstützung der beteiligten Forschungsinstitute in Form von zwei use-case-getriebenen Demonstratoren umgesetzt: Zum einen ein Forschungsprototyp zur Echtzeiterkennung von Deepfake-Angriffen in Videokonferenzen, welcher auf „Self-ID“, einer innovativen Technologie, basiert und visuelle Selbsterkennung als biometrischen Mechanismus zur Identitätsvalidierung nutzbar macht. Zum anderen der Demonstrator „SeamlessMe“, der die Authentifizierung anhand menschlicher Gangprofile ermöglicht. Die aus der Prototypentwicklung resultierenden Ergebnisse und Code-Beispiele wurden in einem Best-Practice-Dokument der Öffentlichkeit zur Verfügung gestellt¹.

2. Voraussetzungen, unter denen das Vorhaben durchgeführt wurde

Um innovative Lösungen zur Absicherung von mobilen und eingebetteten KI-Systemen zu entwickeln und prototypisch umzusetzen, schlossen sich im Projekt SENSIBLE-KI Partner aus Forschung, angewandter Forschung und Wirtschaft zusammen. Dank dieser breiten Aufstellung konnten die angestrebten Ziele sowohl wissenschaftlich fundiert als auch praxisnah verfolgt werden.

Die Projektleitung übernahm das **Fraunhofer-Institut für Angewandte und Integrierte Sicherheit AISEC**, welches über umfassende Expertise in den Bereichen Privatsphäre von künstlicher Intelligenz, eingebettete Systeme und hardwarebasierte Kryptografie verfügt. Zu den weiteren

¹ Siehe www.sensible-ki.de

Partnern des Projektes zählten außerdem die **Bundesdruckerei GmbH**, die **neXenio GmbH** sowie die **Hochschule Darmstadt**. Die beiden innovativen Industrieunternehmen – Bundesdruckerei und neXenio – waren primär für die Entwicklung der beiden Demonstratoren zuständig und unterstützten das Projekt mit ihrem großen Erfahrungsschatz im Bereich digitaler Identitäten und biometrischer Zugangssysteme. Sie nahmen die Rolle der Wirtschaftsvertreter ein und konnten damit einen wertvollen Input aus der Praxis geben. Die Hochschule Darmstadt – als zweite Forschungseinrichtung – fungierte im Vorhaben als Inputgeber und Experte während der Evaluierungsphase der Schutzmechanismen und konnte wissenschaftliche Erkenntnisse in den Bereichen Biometrie und Maschinelles Lernen in das Projekt einbringen.

Das Verbundvorhaben SENSIBLE-KI griff nicht nur auf das Expertenwissen aus der Cybersicherheitsforschung und der Industrie zurück, sondern setzte gleichzeitig auf eine Entwickler-Community aus verschiedenen Branchen und Domänen. So konnte sichergestellt werden, dass die entwickelten Ansätze und Methoden auch den Schutz liefern, der in der Praxis benötigt wird.

3. Planung und Ablauf des Vorhabens

Um das primäre Ziel vom Verbundvorhaben SENSIBLE-KI – die Absicherung von mobilen und eingebetteten KI-Systemen mithilfe von Trusted-Computing-Verfahren und softwarebasierten Sicherheitsmechanismen – zu erreichen, wurden die geplanten Projektarbeiten in drei Phasen unterteilt, denen jeweils drei Arbeitspakete zugeordnet waren (siehe **Abb. 1**)²:

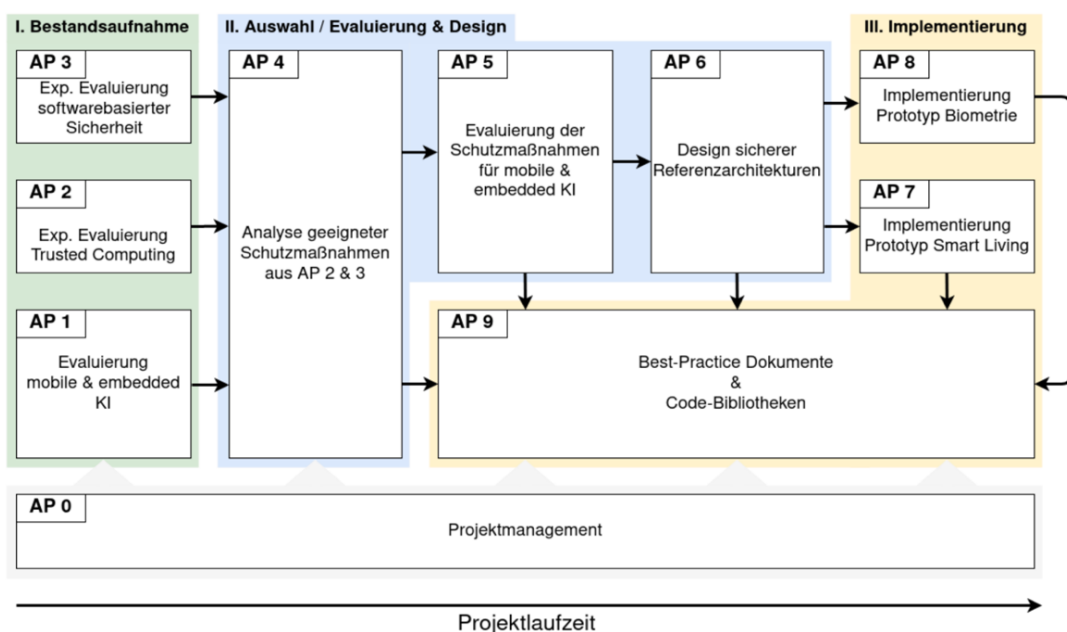


Abbildung 1: Übersicht der Projektphasen und Arbeitspakete im Verbundprojekt SENSIBLE-KI

I. Bestandsaufnahme – AP1, AP2 und AP3:

Eines der Hauptziele der initialen Projektphase war bestehende und am Markt präsente mobile sowie eingebettete KI-Systeme zu evaluieren und in funktionale Anwendungsklassen ein-

² Eine detaillierte Beschreibung der Arbeitspakete sowie der zu erwartenden Ergebnisse der einzelnen Arbeitspakete waren in der Gesamtvorhabenbeschreibung ausführlich dargestellt.

zuteilen, für die dann – basierend auf ihren vorher bestimmten Schutzziele – der individuelle Schutzbedarf definiert wurde (**AP1**). Auf diese Weise konnte sichergestellt werden, dass eine möglichst große Anzahl von praxisnahen KI-Anwendungsfällen in den darauffolgenden Projektphasen betrachtet und abgesichert wird. Analog zur Sichtung der KI-Systeme wurden während der Bestandsaufnahme state-of-the-art Verfahren des Trusted Computing (**AP2**) und softwarebasierte Sicherheitsmechanismen (**AP3**) untersucht, welche sich potenziell zur Absicherung von KI-Systemen eignen könnten. Hier wurden auch die Aspekte der Zugänglichkeit für Drittentwickler sowie die Marktabdeckung berücksichtigt.

II. Auswahl, Evaluierung & Design – AP4, AP5 und AP6:

Der erste Schritt der zweiten Projektphase bestand darin, den funktionalen Anwendungsklassen auf Grundlage ihres Schutzbedarfs passende Mechanismen des Trusted Computings und der softwarebasierten Sicherheitsmethoden zuzuordnen (**AP4**). In einer darauffolgenden Evaluation wurde dann geprüft, ob die zugeordneten Sicherheitsmechanismen dem Schutzbedarf der jeweiligen Anwendungsklassen genügen und eine technische Machbarkeit vorliegt (**AP5**). Auf Grundlage dieser Evaluation wurden sichere Referenzarchitekturen für ausgewählte Anwendungsklassen konzeptioniert (**AP6**).

III. Implementierung – AP7, AP8 und AP9:

In der finalen Projektphase wurden die Referenzarchitekturen zur Implementierung von konkreten Prototypen genutzt. Hierbei wurden zwei Prototypen im mobilen und eingebetteten Kontext aus dem Bereich Biometrie (**AP7** und **AP8**) entwickelt. Diese Implementierungen stellten sicher, dass die zuvor konzeptionierten Referenzarchitekturen sowohl die intendierten Schutzziele wirksam umsetzen als auch in der Praxis einfach zu realisieren sind. Die während der zweiten und dritten Projektphase gewonnenen Erfahrungen und Erkenntnisse wurden in einem öffentlich verfügbaren Best-Practice-Dokument gesammelt (**AP9**).

Im Verlauf des dreijährigen Projektes³ mussten sowohl der Arbeits- als auch der Zeitplan an einigen Stellen entsprechend angepasst werden. Die meisten Änderungen betrafen lediglich die Reihenfolge bzw. die Zeitdauer der durchzuführenden Arbeitsschritte. Die erforderlichen Anpassungen resultierten aus einigen inhaltlichen Abhängigkeiten zwischen den einzelnen Arbeitspaketen bzw. -schritten, welche erst im Laufe des Projektes erkannt werden konnten, und hatten keinen großen Einfluss auf den weiteren Verlauf des Vorhabens.

Das Verbundvorhaben SENSIBLE-KI wurde von allen Projektpartnern planmäßig am 29.02.2024 abgeschlossen. Sowohl der Arbeits- als auch der Kostenplan wurden eingehalten.

³ Das Projekt SENSIBLE-KI wurde zum großen Teil während der COVID-19-Pandemie durchgeführt. Aufgrund der dadurch bedingten Beschränkungen sowie immer wieder steigenden Infektionszahlen entschied sich das Konsortium, die Projekt- und Meilensteintreffen – so wie auch die regelmäßigen Arbeitstreffen – online stattfinden zu lassen, was allerdings keine negative Auswirkung auf die Qualität bzw. Effektivität dieser Termine hatte. Nur das allerletzte Konsortialtreffen am 29.02.2024 war ein Präsenztreffen.

4. Wissenschaftlicher und technischer Stand, an den angeknüpft wurde

Die während der Projektlaufzeit aktuellen Forschungsbestreben im Bereich der Sicherheit und Privatsphäre von KI-Systemen fokussierten sich weitestgehend auf Schutzmaßnahmen der Robustheit und Privatheit für überwachtes Deep Learning [1, 2, 3]. Die vorgeschlagenen Lösungsmechanismen bieten oft nur eine softwareseitige Absicherung und führen zu erheblichen Einbußen in der Nutzbarkeit [4, 5]. Der Einfluss von erhöhter Privatsphäre auf die Robustheit war unklar. Zur Evaluierung der Softwaremaßnahmen, insbesondere im Bereich der Privatsphäre und Robustheit, liegen bekannte Methoden vor [6, 7]. Die sichere Integration in ein Gesamtsystem, wie z. B. einem mobilen Endgerät, bleibt hier meist außen vor.

Bestehende Techniken und Mechanismen aus dem Bereich Trusted Computing (z. B. Trusted Execution Environment oder Secure Element) weisen bereits eine weite Verbreitung auf, sind aber noch nicht auf die Bedürfnisse und Anforderungen von KI-gestützten Lösungen ausgelegt [8]. Einzelne, bereits existierende Vorarbeiten für sichere und vertrauenswürdige mobile KI [9] betrachten nur einen kleinen Teil des Themenkomplexes.

Die beteiligten Projektpartner konnten für das geplante Vorhaben auf einen tiefen Erfahrungsschatz in der Erforschung, Konzeption und dem Betrieb von KI- sowie vertrauenswürdigen Systemen zurückgreifen. So lag ein Kompetenzschwerpunkt der Forschungspartner im Konsortium auf der Evaluierung der Sicherheit und Privatsphäre von eingebetteten und mobilen Systemen [10, 11] sowie auf der Erforschung von neuen Ansätzen und Konzepten für zukünftige Vertrauensanker im digitalen Raum [12, 13]. Einen weiteren Schwerpunkt, der von den Forschungspartnern in das Projekt eingebracht wurde, war das Thema sichere und nachvollziehbare KI-Algorithmen [14, 15, 16, 17]. Weiter haben die projektbeteiligten Unternehmen umfangreiche und langjährige Erfahrungen bei der Entwicklung sowie Einbindung von KI-gestützten Ansätzen in innovative Produkte auf mobilen und eingebetteten Systemen.

5. Zusammenarbeit mit anderen Stellen

Mit Partnern aus Forschung, angewandter Forschung und Wirtschaft war das SENSIBLE-KI-Projektkonsortium breit aufgestellt. Im Zuge der Projektarbeiten wurde zusätzlich die Firma **SpleenLab** als beratende KI-Experten aus dem Bereich eingebettete sichere KI unterbeauftragt. SpleenLab lieferte Input aus der praktischen Anwendung von KI-Systemen in eingebetteter Hardware und erweiterte mit ihrem Know-How die Ausarbeitungen in der ersten Projektphase.

Der zweite assoziierte Partner – **Freie Universität Berlin** – unterstützte das Konsortium vor allem in der letzten Projektphase. Der Partner lieferte insbesondere bei den Ansätzen zum Schutz zahlreiche Beiträge in Form von theoretischen Ansätzen und experimentellen Evaluierungen. Insbesondere die Methoden zum Erhöhen der Robustheit entstanden in intensiver Zusammenarbeit. Darüber hinaus wurden die Forschungsergebnisse auch in die Lehrveranstaltungen Computersicherheit und Embedded Security sowie in Seminare zu ausgewählten Themen der IT-Sicherheit an der Freien Universität Berlin integriert, wodurch eine zusätzliche Verbreitung der Projektergebnisse bei zukünftigen Entscheidern, Anwendern und Wissenschaftlern erreicht wird.

Im Rahmen der Öffentlichkeitsarbeit wurde mit dem **Fraunhofer Heinrich-Hertz-Institut** zusammengearbeitet. Das Institut stellte verschiedene Räumlichkeiten beziehungsweise Ausstellungsflächen für die Demonstratoren bereit. Im Ort für Technologie- und Informationstransfer konnte ein Exponat ausgestellt werden, welches den Kontakt zu den Besuchergruppen ermöglicht.

II. Teil

6. Verwendung der Zuwendung und der erzielten Ergebnisse

Die Zuwendung wurde von allen Verbundpartnern hauptsächlich für den Personaleinsatz verwendet, um die im Arbeitsplan beschriebenen Teilaspekte zu erforschen und die Projektziele zu erreichen. Die dem Projekt SENSIBLE-KI zugrundeliegenden Ziele waren im Antrag wie folgt beschrieben worden:

Methoden der Künstlichen Intelligenz (KI) – wie Machine Learning (ML) (z. B. Deep Learning (DL)) – kommen bereits in einer Vielzahl stetig wachsender heterogener mobiler und eingebetteter Plattformen zum Einsatz. Dort werden sie z. B. in intelligenten Assistenzprogrammen aber auch für biometrische Verifikationsverfahren oder zur Anomalieerkennung [18] verwendet. Aufgrund ihrer Heterogenität ist es schwierig, solche Plattformen gegen Cyberangriffe zu schützen und die Manipulation der ML/DL-Modelle zu verhindern. Dies kann, z. B. im Bereich des autonomen Fahrens, lebensgefährliche Folgen haben [19], aber auch in weniger kritischen Umgebungen, wie z. B. auf Smartphones, kann ein gezielter Angriff auf KI-Systeme weitreichende Folgen haben. So kann z. B. durch gezielte Manipulation des Inputs die biometrische Gesichtserkennung von Authentifizierungsverfahren umgangen [20] oder aber aus einem lokal gespeicherten KI-Modell zur Kreditkartenerkennung die zum Training verwendeten Kreditkartennummern extrahiert werden [21]. Da es sich bei bisherigen Schutzmechanismen und Frameworks ausschließlich um proprietäre Lösungen handelt, bleibt den meisten Unternehmen – insbesondere KMUs – der Zugang verwehrt. Insbesondere gibt es keine einheitlichen Herangehensweisen zur Absicherung der KI-Systeme im mobilen und eingebetteten Kontext, was zu Sicherheitslücken führt.

Das Ziel des Projektes SENSIBLE-KI war es, Angriffe auf die KI-Systeme im mobilen und eingebetteten Kontext zu erschweren bzw. komplett zu verhindern. Hierfür sollten bewährte Mechanismen des Trusted Computings (TC) eingesetzt und mithilfe wissenschaftlicher Methoden die Hypothese, dass derartige Technologien effektiven und zuverlässigen Schutz in den beschriebenen Angriffsszenarien bieten, untersucht werden. Insbesondere sollten dafür die klassischen Sicherheitsziele sichergestellt werden. Der Ansatz, der im Rahmen des Vorhabens SENSIBLE-KI entwickelt werden sollte, beruht auf dem Einsatz hardwarebasierter kryptographischer Verfahren, die einerseits die Authentizität und Integrität, andererseits aber auch die Vertraulichkeit und Privatheit der ML-Modelle sicherstellen. Zusätzlich zu diesen hardwareseitigen TC-Methoden sollten auch Softwaremethoden für Privatsphärenschutz und Robustheit entwickelt werden. Um die Effektivität und Effizienz des entwickelten Ansatzes zu evaluieren, war die Entwicklung von zwei unterschiedlichen Prototypen aus dem Bereich Biometrie geplant, die einerseits praktische Beispiele darstellen, wie die im Projekt erarbeiteten Referenzarchitekturen und Best Practices implementiert werden können und andererseits dazu dienen, die Einhaltung der Sicherheitsziele praktisch evaluieren zu können. Die zu erstellenden Referenzarchitekturen und Dokumente sollten als Basis für zukünftige Standardisierungsvorhaben im Bereich sicherer und privater mobiler eingebetteter KI dienen.

Zusammenfassend kann man die wissenschaftlichen und technischen Herausforderungen des Projektes SENSIBLE-KI wie folgt darstellen:

Umfassende Identifizierung von funktionalen Klassen, in die sich bestehende KI-Systeme einteilen lassen und praxisnahe Evaluierung des Schutzbedarfs der einzelnen Klassen

- Initialer Multi-Level-Entwurf zur Einteilung von KI-Systemen in funktionale Klassen. Für spätere Arbeitspakete wurde dieses Modell aufgrund zu hoher Komplexität überarbeitet und auf charakteristische Merkmale, die einen bestimmten Schutzbedarf hervorrufen, reduziert.
- Anhand dieser Merkmale konnte der individuelle Schutzbedarf bestimmt werden.

Analyse der Eignung von bestehenden software- und hardwareseitigen Schutzmechanismen für KI-Systeme im eingebetteten und mobilen Kontext

- Analyse des Stands der Technik und anschließende prototypische Implementierung von hard- und softwarebasierten Schutzmaßnahmen für Android-Mobilgeräte und eingebettete Systeme. Anschließend Evaluation des erfolgversprechendsten Ansatzes TrustyTEE hinsichtlich Umsetzbarkeit, Herausforderungen, Geschwindigkeit und Erweiterbarkeit des Ansatzes auf der Android-Plattform. Weiterhin wurde ausgewählte Tamper Resistant Hardware und OP-TEE auf eingebetteten Systemen bzgl. ihrer Auswirkung auf Performance und Stromverbrauch evaluiert. Zusammenfassend sind grundlegende Schutzmaßnahmen, wie z. B. Modellattestierung, ohne weiteres umsetzbar. Tiefergehende Schutzmaßnahmen, wie z. B. die Attestierung von Sensordaten, sind jedoch aufgrund der vom Hersteller gewählten Systemarchitektur und des beschränkten Funktionsumfangs nicht einfach umsetzbar und erfordern Expertenwissen.
- Evaluation von etablierten Softwaremaßnahmen für Android-Mobilgeräte hinsichtlich Implementierbarkeit, Wirksamkeit, Einfluss auf die Modellgenauigkeit und Performance. Hierbei wurde ermittelt, dass der aktuelle Stand der Technik hinsichtlich der Schutzmaßnahmen grundsätzlich auf die Android-Plattform übertragen werden kann, jedoch der Einfluss auf die Performance des Gesamtsystems signifikant ist. Weiterhin soll angemerkt werden, dass im Allgemeinen eine Abwägung zwischen Modellgenauigkeit und Schutzniveau getroffen werden muss. Eine direkte praktische Umsetzbarkeit in operativen Systemen ist nach aktuell gewonnen Erkenntnissen nicht gegeben.

Technische Integration und Zusammenführung von Verfahren des Trusted Computing mit Methoden des softwareseitigen Schutzes

- Es wurden keine Wechselwirkungen zwischen den evaluierten soft- und hardwareseitigen Maßnahmen festgestellt, sodass diese problemlos zusammengeführt werden konnten.

Anwendungsnahe Sicherheitsevaluierung der erarbeiteten Methoden

- Die theoretisch erarbeiteten Methoden wurden auf konkreter Hardware evaluiert. Für die softwareseitigen Maßnahmen wurden u. A. die Auswirkungen der Maßnahme auf Genauigkeit, Robustheit und Privatsphäre evaluiert. Die Trusted-Computing-Maßnahmen wurden nicht zusätzlich auf ihre Sicherheit evaluiert, da hierzu eine Überprüfung der Korrektheit der Verschlüsselungs- und Signaturverfahren notwendig wäre. Da dies

den Umfang des Projektes übersteigen würde, wurde diesbezüglich den Herstellerangaben vertraut.

Implementierung und Evaluierung der zwei Demonstratoren aus dem Bereich Biometrie unter Einbindung der entwickelten Schutzmaßnahmen

- Der Demonstrator zur Erkennung von Deepfake-Angriffen in Videokonferenzen wurde im Rahmen des Projektes entwickelt. Dieser basiert auf der experimentellen Technologie "Self-ID", welche visuelle Selbsterkennung misst, um Identitätsvalidierung durchzuführen. Im Zentrum befindet sich ein Machine-Learning-Modell, welches Eye-Tracking-Daten klassifiziert, um die Selbsterkennung des Nutzers zu überprüfen. Dadurch kann das System potenzielle Deepfake-Angriffe erkennen. Dieses zentrale Modell, insbesondere dessen Training und Verwendung, wurde im Rahmen der zuvor erarbeiteten Sicherheitsmaßnahmen abgesichert. Dabei wurden die geeigneten Maßnahmen, z. B. Bereinigung von Eingabe- und Ausgabedaten, verprobt und es konnten wertvolle Erkenntnisse hinsichtlich deren praktischen Einsatzes gewonnen werden. Über das Projekt hinaus wird der Prototyp zur Datensammlung und anwendungsorientierten Forschung im Bereich der neuartigen Biometrie-Technologien weitere Verwendung finden.
- Der Demonstrator, welcher die Authentifizierung von Benutzern anhand des Gangprofils vornimmt, wurde innerhalb des Projektes entwickelt. Der im Demonstrator "SeamlessMe" eingesetzte Machine-Learning-Algorithmus ist eine Einzelklassen-Klassifizierung (One Class Classification), die durch eine Ausreißererkennungsmethodik (Novelty Detection) erweitert wird. Ein Machine-Learning-Modell dient dazu, die Gangdaten zu klassifizieren, um Nutzer zu erkennen. Verschiedene Sicherheitsmaßnahmen konnten implementiert und erprobt werden. Beispielsweise wurde die Anomalieerkennung implementiert und protokolliert. Für weitere Forschung konnte eine Sammlung von Daten bereitgestellt werden. Die Evaluierung konnte beispielsweise anhand verschiedener Metriken zur Bestimmung der Klassifizierungsgenauigkeit durchgeführt werden.

Erstellung von Referenzarchitekturen, Best-Practice-Dokumenten und Code-Libraries für eine breite Entwickler-Community in verschiedensten Anwendungsdomänen

- Auf Basis des entwickelten Modells zur Bestimmung des Schutzbedarfs wurden Referenzarchitekturen für generelle Anwendungsszenarien und die zwei Prototypen erstellt.
- Ein Best-Practice-Dokument inkl. Code-Libraries wurde auf Basis der im gesamten Projekt gesammelten Erkenntnisse erstellt.

7. Wichtigste Positionen des zahlenmäßigen Nachweises

Bei den Projektpartnern entstanden folgende Personalausgaben:

AP-Nr.	AISEC	bdr	h_da	neXenio
AP0	12 PM	-	-	-
AP1	9 PM	2 PM	4 PM	4 PM
AP2	9 PM	1 PM	3 PM	4 PM
AP3	9 PM	-	3 PM	-
AP4	9 PM	-	4 PM	-
AP5	9 PM	-	4 PM	-
AP6	9 PM	1 PM	4 PM	4 PM
AP7	-	5 PM ⁴	-	3 PM
AP8	-	-	2 PM	19 PM
AP9	6 PM	1 PM	-	2 PM
Insgesamt	72 PM	10 PM	24 PM	36 PM

Tabelle 1: Ressourcenplanung vs. Personalausgaben (in Personenmonaten)

Darüber hinaus ergaben sich im Laufe des Projektes SENSIBLE-KI weitere planmäßige Kosten:

- Entwicklungshardware zur planmäßigen Erfassung des präexistenten verfügbaren Funktionsumfangs und Evaluation der Schutzmaßnahmen (u. A. eingebettete Plattformen mit KI-Hardwarebeschleunigern und verschiedene Trusted-Computing-Lösungen, welche im Kontext von mobilen oder eingebetteten Plattformen genutzt werden können; Smartphones und PCs für Test- und Demonstrationszwecke),
- Server-Infrastruktur bzw. Cloud-Ressourcen,
- Reisekosten für Konferenzbesuche, um den Austausch mit Fachexperten zu ermöglichen und Projektergebnisse vorzustellen,
- Kommunikationsmaterialien zur anschaulichen Verbreitung von Projektergebnissen.

8. Notwendigkeit und Angemessenheit der geleisteten Arbeit

Die im Projektantrag ausgeführten Arbeiten beschrieben bereits die Notwendigkeit zur Erreichung der Projektziele des Gesamtvorhabens SENSIBLE-KI, die nur auf Basis der Aufarbeitung des Standes der Wissenschaft, der Evaluation der am Markt präsenten mobilen und eingebetteten KI-Systeme sowie der Prüfungen der state-of-the-art Verfahren des Trusted Computing und

⁴ Weitere 10 PM wurden von der Bundesdruckerei an den Projektpartner neXenio unterbeauftragt. Dies geschah aufgrund eines unerwarteten, kurzfristigen Wegfalls eines Entwickler-Teams in der Bundesdruckerei.

softwarebasierte Sicherheitsmechanismen erreicht werden konnten. Das Ziel, mit dem Projektende zwei Demonstratoren aus dem Bereich Biometrie zu präsentieren, bei deren Umsetzung die im Projekt entwickelten Schutzmaßnahmen eingebunden wurden, um ihre Produkteigenschaften zu verbessern, war nur möglich, indem alle im Arbeitsplan beschriebenen Arbeitspakete abgearbeitet und deren Resultate kontinuierlich sowohl für die Entwicklung der Demonstratoren als auch für die Erstellung des Best-Practice-Dokuments genutzt wurden. Dies betrifft nach der ausführlichen Bestandsaufnahme und der experimentellen Evaluierung hard- und softwareseitiger Sicherheitsmaßnahmen für KI-Systeme die Analyse und Auswahl geeigneter Schutzmaßnahmen sowie anschließend das Design der sicheren Referenzarchitektur.

Mit Bezug auf die Angemessenheit der geleisteten Arbeiten kann festgehalten werden, dass die Arbeiten so durchgeführt wurden, dass die im Projektantrag geplanten Ziele der jeweiligen Arbeitsphasen erreicht werden konnten. Dabei wurde innerhalb einzelner Arbeitsschritte aufgrund von Erkenntnissen aus vorherigen Arbeitsphasen vom ursprünglichen Plan abgewichen. Wie in den Zwischenberichten angegeben, wurden unangemessene Arbeiten durch sinnvolle Arbeitsschritte ersetzt, sodass die Angemessenheit gegeben war.

9. Voraussichtlicher Nutzen und Verwertbarkeit

Im Verlauf des Projektes SENSIBLE-KI haben sich keine Änderungen gegenüber den im ursprünglichen Antrag festgehaltenen Punkten bzgl. voraussichtlicher Nutzen und Verwertbarkeit ergeben.

Um die wissenschaftliche Anschlussfähigkeit abzusichern, werden die Ergebnisse der beiden Teilvorhaben in die Lehrveranstaltungen an der Hochschule Darmstadt und in die Lehre des eng mit dem Fraunhofer AISEC verbundenen Lehrstuhls an der Freien Universität Berlin (AG Informationssicherheit) integriert, wodurch eine zusätzliche Verbreitung der Projektergebnisse bei zukünftigen Entscheidern, Anwendern und Wissenschaftlern erreicht wird. Hierzu wurden bereits Softwareprojekte an der FU Berlin angeboten, die auch die Erkenntnisse aus dem Projekt wiederverwenden und vermitteln. Zudem wurde einer der Demonstratoren im Rahmen einer Masterarbeit ausführlich behandelt und die zugrundeliegenden Methoden wissenschaftlich weiterentwickelt, sodass die Projektinhalte unmittelbar zur Verbesserung der Lehre beitragen haben und werden. Weiterhin ist es geplant, das während des Verbundvorhabens SENSIBLE-KI gewonnene Know-how neben dem internen Kompetenzaufbau auch weiterhin direkt für die Förderung des wissenschaftlichen Nachwuchses einzusetzen. Die Projektpartner AISEC und h_da sehen in Bezug auf Forschung als Verwertungsaussicht die Nutzung der innerhalb ihrer Teilprojekte gewonnenen wissenschaftlichen Erkenntnisse als Basis für weitere darauf aufbauende wissenschaftliche Arbeiten (wie z. B. wissenschaftliche Veröffentlichungen und Dissertationen) sowie drittmittelgeförderte nationale und internationale Anschlussprojekte.

Die wissenschaftlichen und technischen Erfolgsaussichten nach Projektende werden als durchaus gut beurteilt. Dies begründet sich zum einen durch die kontinuierlich steigende Verbreitung von KI-Anwendungen vor allem auch in sicherheitskritischen und Privatsphäre-relevanten Bereichen und zum anderen durch eine gesteigerte Sensibilisierung der Gesellschaft gegenüber dem Datenschutz und der Sicherheit von IT-Systemen. SENSIBLE-KI liefert hier als eines der ersten Förderprojekte konkrete Vorschläge zur Umsetzung von sicherer und vertrauenswürdiger KI, auf welchen weitere Beiträge bspw. wissenschaftliche Artikel aufsetzen können.

Die durchgeführten Evaluationen zeigten, dass die Technologien grundsätzlich wirksam sind, jedoch zu gegebenen Zeitpunkt noch Entwicklungsbedarf im Bereich der Hardwareplattform und Forschungsbedarf zur besseren Vereinbarkeit von Sicherheit und Genauigkeit gibt. Aktuelle Arbeiten zeigen aber eine schnelle Entwicklung des Forschungsbereichs, welche aller Wahrscheinlichkeit nach auch technisch umgesetzt werden können.

Mit der Beteiligung am Verbundvorhaben SENSIBLE-KI zielten die Projektpartner darauf ab, ihre führende Stellung in der angewandten Forschung auf dem Gebiet sicherer und vertrauenswürdiger KI zu behaupten und weiter zu stärken. Damit wird gewährleistet, dass auch nach Projektende erfolgreich Industrieprojekte eingeworben werden können und auf diese Weise ein wichtiger Beitrag zu Erhalt und Ausbau des High-Tech-Standorts Deutschland geleistet wird. Die Entwicklungen im Bereich sicherer und vertrauenswürdiger mobiler KI-Architekturen sind somit mit dem Projektende nicht abgeschlossen, sondern werden sowohl in Forschungsprojekten weitergeführt als auch in Projekten mit Unternehmen und öffentlichen Einrichtungen gebraucht.

Zur wirtschaftlichen Anschlussfähigkeit ist festzuhalten, dass seitens des Partners bdr vorgesehen ist, eine mögliche Verwertung des entwickelten Demonstrators zu testen. Dies geschieht mittels verschiedener Business Units (BUs), welche jeweils den Fokus auf einen bestimmten Markt, wie beispielsweise die öffentliche Verwaltung oder die Privatwirtschaft, legen. neXenio war in der Lage, im Verbundvorhaben SENSIBLE-KI Wissen auf den Gebieten der sicheren KI-Systeme sowie der Nutzbarkeit in echten Anwendungen aufzubauen. Über den Demonstrator konnten die gewonnenen Kenntnisse in den Wirkbetrieb übernommen werden. Ebenso konnte neXenio die Ergebnisse einsetzen, um eine Grundlage für zukünftige Softwarelösungen zu etablieren. neXenio erwartet eine Stärkung des Produktes im Markt und strebt den Aufbau des KI-Teams sowie die Fortführung der engen Kommunikation mit der aufgebauten KI-Community an. Somit tragen die beiden Industrieprojektpartner ebenfalls weiterhin dazu bei, den Wirtschaftsstandort Deutschland zu stärken und einen deutschen Wissensvorsprung im Bereich Security und KI zu ermöglichen.

10. Fortschritt bei anderen Stellen

Zur Projektlaufzeit sind für die Durchführung des Verbundvorhabens SENSIBLE-KI keine relevanten F&E-Ergebnisse von dritter Seite bekannt geworden.

11. Erfolgte und geplante Veröffentlichungen

Für die Außendarstellung des Verbundprojektes SENSIBLE-KI wurde insbesondere die Projektwebseite genutzt, welche kontinuierlich um die Ergebnisse aus den jeweiligen Arbeitspaketen aktualisiert wurde.

Darüber hinaus erfolgte zur Projektlaufzeit folgende Veröffentlichung:

- wissenschaftliche Konferenz "7th IEEE/IAPR International Joint Conference on Biometrics (IJCB) 2023" (Ljubljana, Slovenien): Vortrag und Artikel "*Unconventional Biometrics: Exploring the Feasibility of a Cognitive Trait based on Visual Self-Recognition*".

Folgende Veröffentlichungen sind noch geplant:

- beim Fraunhofer-Institut für Angewandte und Integrierte Sicherheit AISEC:
 - *„Verbesserung der Softwareschutzmaßnahmen für Privatsphäre durch alternative Trainingsansätze“*,
- bei der Bundesdruckerei:
 - wissenschaftliche Konferenz “22nd International Conference on Applied Cryptography and Network Security (ACNS) 2024” (Abu Dhabi, Vereinigte Arabische Emirate): Posterpräsentation und Artikel *“Applying Self-Recognition Biometrics to Live Deepfake Detection in Video Conferences”*,
 - Industriekonferenz “Optical & Digital Document Security (ODDS) 2024” (Lissabon, Portugal): Vortrag und Artikel *“Leveraging In-Brain Identity Validation Mechanisms for Detection of Live Video Deepfake Attacks”*.

Literaturverzeichnis

- [1] Zhou, J., et al. "Transparent machine learning—revealing internal states of machine learning." Proceedings of IUI2013 Workshop on Interactive Machine Learning. 2013.
- [2] Ganin, Y., et al. "Domain-adversarial training of neural networks." In Domain Adaptation in Computer Vision Applications, pp. 189-209. Springer, Cham, 2017.
- [3] Papernot, N., et al. "Scalable private learning with pate." arXiv preprint arXiv:1802.08908 (2018).
- [4] Abadi, M., et al. "Deep learning with differential privacy." In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 308-318. ACM, 2016.
- [5] Papernot, N., et al. "Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning." arXiv preprint arXiv:1803.04765 (2018).
- [6] Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., & Tramer, F. (2022, May). Membership inference attacks from first principles. In 2022 IEEE Symposium on Security and Privacy (SP) (pp. 1897-1914). IEEE, Fredrikson, M., Jha, S., & Ristenpart, T. (2015, October)
- [7] Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC conference on computer and communications security (pp. 1322-1333)
- [8] Ohrimenko, O., et al. "Oblivious multi-party machine learning on trusted processors." In 25th USENIX Security Symposium (USENIX Security 16), pp. 619-636. 2016.
- [9] Bayerl, Sebastian P., et al. "Offline model guard: Secure and private ML on mobile devices." DATE 2020 (2020).
- [10] Auer, Lukas, Christian Skubich, and Matthias Hiller. "A Security Architecture for RISC-V based IoT Devices." 2019 Design, Automation & Test in Europe Conference & Exhibition (DATE). IEEE, 2019.
- [11] Jacob, Nisha, et al. "How to break secure boot on FPGA SoCs through malicious hardware." International Conference on Cryptographic Hardware and Embedded Systems. Springer, Cham, 2017.
- [12] Morbitzer, Mathias. "Scanclave: Verifying Application Runtime Integrity in Untrusted Environments." 2019 IEEE 28th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE). IEEE, 2019.
- [13] Hristozov, Stefan, et al. "Practical runtime attestation for tiny iot devices." Proceedings of the 2018 Workshop on Decentralized IoT Security and Standards, San Diego, CA, USA. Vol. 18. 2018.
- [14] Böttinger, Konstantin, Patrice Godefroid, and Rishabh Singh. "Deep reinforcement fuzzing." 2018 IEEE Security and Privacy Workshops (SPW). IEEE, 2018.
- [15] Kolosnjaji, Bojan, et al. "Empowering convolutional networks for malware classification and analysis." 2017 International Joint Conference on Neural Networks (IJCNN). IEEE, 2017.

- [16] Müller, Nicolas M., et al. "Distributed Anomaly Detection of Single Mote Attacks in RPL Networks." (2019).
- [17] Schneider, Peter, and Konstantin Böttinger. "High-performance unsupervised anomaly detection for cyber-physical system networks." Proceedings of the 2018 Workshop on Cyber-Physical Systems Security and PrivaCy. 2018.
- [18] Wu, D., et al. "Droidmat: Android malware detection through manifest and api calls tracing." 2012 Seventh Asia Joint Conference on Information Security. IEEE, 2012.
- [19] Eykholt, K., et al. "Robust physical-world attacks on deep learning models." arXiv preprint arXiv:1707.08945 (2017).
- [20] Sharif, M., et al. "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition." Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2016.
- [21] Carlini, N., et al. "The secret sharer: Evaluating and testing unintended memorization in neural networks." In 28th USENIX Security Symposium (USENIX Security 19), pp. 267-284. 2019.

Berichtsblatt

1. ISBN oder ISSN	2. Berichtsart (Schlussbericht oder Veröffentlichung) Verbundschlussbericht
3. Titel Sachbericht SENSIBLE-KI – Sichere und vertrauenswürdige mobile KI	
4. Autor(en) [Name(n), Vorname(n)] Prof. Dr. Margraf, Marian; Dr. Schnjakin, Maxim; Prof. Dr. Busch, Christoph; Berger, Philipp	5. Abschlussdatum des Vorhabens 29/02/2024
	6. Veröffentlichungsdatum
	7. Form der Publikation
8. Durchführende Institution(en) (Name, Adresse) Fraunhofer-Institut für Angewandte und Integrierte Sicherheit (AISEC), Lichtenbergstraße 11, 85748 Garching b. München Bundesdruckerei GmbH, Kommandantenstraße 18, 10969 Berlin Hochschule Darmstadt (Fachbereich Informatik, da/sec Gruppe), Haardtring 100, 64295 Darmstadt neXenio GmbH, Charlottenstraße 59, 10117 Berlin	9. Ber. Nr. Durchführende Institution
	10. Förderkennzeichen 01MT21005A 01MT21005B 01MT21005C 01MT21005D
	11. Seitenzahl 16
12. Fördernde Institution (Name, Adresse) Bundesministerium für Wirtschaft und Klimaschutz (BMWK) 53107 Bonn	13. Literaturangaben 21
	14. Tabellen 1
	15. Abbildungen 1
16. Zusätzliche Angaben	
17. Vorgelegt bei (Titel, Ort, Datum)	
18. Kurzfassung Im Projekt SENSIBLE-KI wurden Methoden am Beispiel von Android-Systemen untersucht, um den Schutz von KI-Systemen und der von ihnen verarbeiteten Daten sicherzustellen und gleichzeitig eine sichere Übertragung der KI-Modelle auf mobile Endgeräte zu ermöglichen. Dafür wurden Lösungen entwickelt, um Angriffe auf KI-Systeme im mobilen und eingebetteten Kontext zu erschweren oder komplett zu verhindern. Konkret kamen hardwarebasierte Trusted-Computing-Verfahren sowie softwarebasierte Mechanismen zum Einsatz, welche unabhängig von der Plattform auf die KI-Anwendung angewendet werden können. Diese Maßnahmen verbessern einerseits die Robustheit, andererseits aber auch die Vertraulichkeit und Privatheit der Machine-Learning-Modelle und deren Daten. Die innerhalb der Projektlaufzeit gewonnenen Erkenntnisse wurden anhand zweier Use-Case-getriebenen Demonstratoren praktisch veranschaulicht.	

19. Schlagwörter	
Sicherheit, Robustheit, Privatheit, Künstliche Intelligenz, KI, mobile Endgeräte, Sicherheitsmaßnahmen, Biometrie, Erkennung von Gangmustern bei Zutrittskontrollen, Echtzeiterkennung von Deepfake-Angriffen in Videokonferenzen, Verhaltensbiometrie, Ganganalyse, Maschinelles Lernen, eingebettete Systeme, Trusted Computing	
20. Verlag	21. Preis

Document Control Sheet

1. ISBN or ISSN	2. type of document (e.g. report, publication) Final report
3. title Sachbericht SENSIBLE-KI – Sichere und vertrauenswürdige mobile KI	
4. author(s) (family name, first name(s)) Prof. Dr. Margraf, Marian; Dr. Schnjakin, Maxim; Prof. Dr. Busch, Christoph; Berger, Philipp	5. end of project 29/02/2024
	6. publication date
	7. form of publication
8. performing organization(s) (name, address) Fraunhofer-Institut für Angewandte und Integrierte Sicherheit (AISEC) , Lichtenbergstraße 11, 85748 Garching b. München Bundesdruckerei GmbH , Kommandantenstraße 18, 10969 Berlin Hochschule Darmstadt (Fachbereich Informatik, da/sec Gruppe), Haardtring 100, 64295 Darmstadt neXenio GmbH , Charlottenstraße 59, 10117 Berlin	9. originator's report no. -
	10. reference no. 01MT21005A 01MT21005B 01MT21005C 01MT21005D
	11. no. of pages 16
12. sponsoring agency (name, address) Bundesministerium für Wirtschaft und Klimaschutz (BMWK) 53107 Bonn	13. no. of references 21
	14. no. of tables 1
	15. no. of figures 1
16. supplementary notes	
17. presented at (title, place, date)	
18. abstract In the SENSIBLE-KI project, methods were investigated using the example of Android systems to ensure the protection of AI systems and the data processed by them, while at the same time enabling the secure transfer of AI models to mobile devices. To this end, solutions were developed to make attacks on AI systems in mobile and embedded contexts more difficult or to prevent them completely. Specifically, hardware-based trusted computing procedures and software-based mechanisms were used, which can be applied to the AI application regardless of the platform. These measures improve both the robustness and the confidentiality and privacy of the machine learning models and their data. The insights gained during the project were illustrated in practice using two use case-driven demonstrators.	

19. keywords	
safety, robustness, privacy, artificial intelligence, AI, mobile devices, security measures, biometrics, recognition of gait patterns for access control, real-time detection of deepfake attacks in video conferences, behavioural biometrics, gait analysis, machine learning, embedded systems, Trusted Computing	
20. publisher	21. price