Anlage 2: „Vertrauenswürdige Künstliche Intelligenz für polizeiliche Anwendungen“ (VIKING)

Förder-kennzeichen	**Zuwendungs- / Zuweisungsempfänger**	**Erledigte Aufgaben**	**Offene Aufgaben**
13N16236	Landeskriminalamt Nordrhein-Westfalen	Es wurde einen Datensatz mit polizeilicher Relevanz zur Verfügung gestellt, ein Objektdetektor damit trainiert und die Entwicklung der Methoden zur erklärbaren künstlichen Intelligenz (XAI) vorangetrieben.	Die entwickelten Methoden sollen in den Demonstrator integriert werden und anschließend mit dem Demonstrator weiter optimiert und erprobt werden.
13N16237	Polizeipräsidium München	Die konkreten Bedarfe und Herausforderungen sowie Standardisierungsbedarfe einer Sicherheitsbehörde in Bezug auf vertrauenswürdige KI wurden ermittelt. Weiterhin wurden Test- und Trainingsdatensätze erstellt und die bisherigen Entwicklungen und Ergebnisse aus Anwendersicht getestet und bewertet.	In der zweiten Hälfte des Vorhabens wird nun der Schwerpunkt auf der weiteren Unterstützung des Entwicklungsprozesses des Objektdetektionsdemonstrators sowie der Testung und Evaluation der Ergebnisse liegen.
13N16238	IDEMIA Identity & Security Germany AG	Mögliche Methoden bzw. Methodenklassen zum Debiasing und der Erklärbarkeit tiefer neuronaler Netze wurden ausgewählt. Kategorien und benötigte Daten für die Implementierung und Testung der Methoden wurden spezifiziert. Anwendungsspezifische Klassifikatoren zur automatisierten Bestimmung von Charakteristika wurden implementiert. Benötigte Trainingsdaten, angereichert mit Informationen über die jeweiligen Kategorien wurden zusammengestellt. Eine Grobspezifikation des Demonstrators zum Debiasing wurde erstellt. Ein Methodensatz für das Debiasing im Trainingsprozess, der eine substanzielle Verbesserung der Fehlerraten der Gesichtserkennung ermöglicht, wurde implementiert. Es wurden Untersuchungen zur Übertragbarkeit der Methoden durchgeführt.	Eine Demonstratorversion zum Debiasing im Trainingsvorgang, die einen direkten und leicht interpretierbaren Vorher-/Nachher-Vergleich ermöglicht, soll implementiert werden. Zwei Demonstratoren zur Testung und Evaluierung der entwickelten Methoden mit den Anwendern sollen erstellt werden. Eine Kurzanleitung für die Demonstratorversionen wird erstellt. Eine Einweisung / Schulung für die Anwender an den Demonstratoren soll durchgeführt werden. Das BSI soll im Hinblick auf mögliche perspektivische Zertifizierungen eingebunden werden.
13N16239	Fraunhofer-Institut für Digitale Medientechnologie (IDMT) -	Es wurden Testdatensets zur Analyse von Sprechererkennungssystemen hinsichtlich Performance, Robustheit gegenüber Störeinflüssen und Fairness erstellt. Zur Auswertung und Darstellung der Ergebnisse auf den oben genannten Testdatensets wurde ein Demonstrator	Ein Framework zur Analyse der Modellparameter einer Netzwerkarchitektur mit integrierter Darstellungsform soll erstellt werden. Die Übertragbarkeit der erforschten Verfahren auf andere Anwendungsfälle soll

		entwickelt. Die Embeddings eines Sprechererkennungssystems wurden analysiert und verschiedene Erklärbarkeitsmethoden exploriert.	untersucht werden. Die Ergebnisse sollen für die Überführung in Normen und Standards aufbereitet werden. Die Ergebnisse und Demonstratoren werden zusammen mit den Anwendern getestet und evaluiert.
13N16240	DIN Deutsches Institut für Normung e. V.	Eine Übersicht zur Normungs- und Standardisierungslandschaft im Kontext zu VIKING wurde erstellt. Die Bedarfe wurden identifiziert, priorisiert und in einer Übersicht zusammengestellt.	Standardisierungspotentiale sollen identifiziert, entsprechende Standardisierungs-dokumente erstellt und bei ausreichend hohem Konsens im Gremium veröffentlicht sowie in relevanten Kreisen verteilt werden.
13N16241	Hochschule für Wirtschaft und Recht Berlin - Forschungsinstitut für öffentliche und private Sicherheit (FÖPS Berlin)	Es wurde ein Anforderungskatalog zu rechtlichen Aspekten der polizeilichen KI-Nutzung erstellt. Dieser beinhaltet erste rechtswissenschaftliche Analysen, insbesondere hinsichtlich der KI-Regulierungsvorschläge auf europäischer Ebene (KI-VO-E) sowie einen Abgleich mit grundrechtlichen Anforderungen und internationalen Empfehlungen, wie den UNESCO-Empfehlungen zur KI-Ethik oder den OECD KI-Leitlinien.	Der Anforderungskatalog soll kontinuierlich aktualisiert werden. Weiterhin ist kontinuierliche und enge Begleitung des Gesetzgebungsprozesses notwendig, um die künftigen, an KI-basierte Systeme zu stellenden, Anforderungen zu identifizieren und bei der Technikentwicklung zu berücksichtigen. Die Gerichtsverwertbarkeit der mittels KI analysierten Daten soll untersucht werden.
13N16242	Universität Konstanz - Fachbereich Informatik und Informationswissen-schaften	Es wurden Methoden zur Erklärung von künstlicher Intelligenz identifiziert und die Wichtigkeit der Erklärungen für die jeweiligen Benutzergruppen erarbeitet. Es wurden erste Konzepte für die Generierung von lokalen Erläuterungen erstellt und auch schon eine erste technische Teillösung erarbeitet. Es wurde ein Konzept für die Analyse von Netzwerkstrukturen vorgeschlagen und in einer Teillösung umgesetzt. Das XAI-Modul wurde erstellt und gemeinsam mit dem Objektdetektor evaluiert. Konzepte für die Übertragbarkeit und Evaluation der Methoden wurden erstellt.	Die Methoden und Teillösungen sollen in den Gesamtdemonstrator integriert und den Anwendern für Tests und Evaluierung zur Verfügung gestellt werden. Die dafür benötigten Schnittstellen werden erarbeitet. Die Übertagbarkeit der Ergebnisse auf andere Methoden wird untersucht und der Gesamtdemonstrator zusammen mit den Anwendern evaluiert.

13N16243	Eberhard Karls Universität Tübingen - Internationales Zentrum für Ethik in den Wissenschaften (IZEW)	Ein ethischer Anforderungskatalog in Absprache mit den Projektpartnern wurde als Basis für die Spezifikation der Demonstratoren erstellt. Eine Grobfassung einer Matrix zur ethischen Bewertung wurde erstellt. Die ethischen Standpunkte wurden in die Entwicklungsprozesse der technischen Demonstratoren integriert. Die ethischen Standpunkte wurden in den rechtlichen Anforderungskatalog und die Sondierungen zur Übertragbarkeit und Standardisierung der Methoden integriert.	Die Matrix zur ethischen Bewertung soll konkretisiert werden. Weiterhin ist kontinuierliche und enge Begleitung des Projektes und der Entwicklung für eine abschließende ethische Bewertung nötig.
13N16244	Universität der Bundeswehr München - Forschungsinstitut CODE	Es wurde ein Pflichtenheft und Methodenspezifikationen für die KI gestützte Textauswertung sowie die Spezifikation der benötigten Demonstratoren erarbeitet. Weiterhin wurden Benchmark-Datensätze erstellt und annotiert und die KI-Modelle trainiert und evaluiert. Erste Debiasing-Ansätze für KI-Modelle wurden identifiziert und evaluiert sowie erste Methoden zur Erklärbarkeit von Sprachmodellen entwickelt.	Die erforschten KI-Modelle, Ansätze zum Debiasing und Methoden zur Erklärbarkeit werden optimiert und in den Gesamtdemonstrator integriert. Die Ergebnisse der Generierung werden mit den Anwendungspartnern auf Akzeptanz und Nützlichkeit qualitativ evaluiert. Die Übertragbarkeit der Ergebnisse auf andere Anwendungen wird untersucht. Handlungsanweisungen und Empfehlungen für die Anwender werden erstellt.
13N16245	Zentrale Stelle für Informationstechnik im Sicherheitsbereich (ZITiS) - Abt. Big Data Analyse	Benchmark Datensätze für die Textauswertung wurden erstellt. Ein KI-Modell zur Klassifikation von Texten nach Anforderungen einer Sicherheitsbehörde wurde erarbeitet. Eine erste Untersuchung zur Auswirkung nicht-ausbalancierter Trainingsdaten auf die Genauigkeit von KI-Modellen zur Textklassifikation wurde durchgeführt und in einem Bericht dokumentiert.	Untersuchung der Robustheit der Modelle mittels Modellinversionsattacken werden durchgeführt. Eine Ableitung von Gegenmaßnahmen zur Reduktion der Angreifbarkeit wird erstellt. Eine Evaluierung des erstellten Demonstrators gemeinsam mit den Anwendern wird durchgeführt. Die Übertragbarkeit der Ergebnisse auf andere Anwendungen wird untersucht. Handlungsanweisungen und Empfehlungen für Entwickler und Anwender werden erstellt.