Sachbericht zum Verwendungsnachweis

Förderkennzeichen: 13N15342

Einsatz von KI zur Früherkennung von Straftaten (KISTRA)

Teilvorhabensbezeichnung: „Organisationale, rechtliche und gesellschaftliche Rahmenbedingungen“ im Projektverbund KISTRA – Einsatz von KI zur Früherkennung von Straftaten; **Akronym:** KISTRA

Technische Universität Berlin, Zentrum Technik und Gesellschaft (TUB)

Berichtsdatum: 26. Juni 2024
Berichtszeitraum: 01.07.2020 – 31.12.2023
Förderlaufzeit: 01.07.2020 – 31.12.2023

Förderer:
Bundesministerium für Bildung und Forschung

Projektträger:
VDI Technologiezentrum

Zuwendungsempfänger:
Technische Universität Berlin, Zentrum Technik und Gesellschaft
Kaiserin-Augusta-Allee 104
10553 Berlin

Projektleitung:
Dr. Robert Pelzer

Projektmitarbeitende: Michael Hahne, Jonatan Schewe (WM), Carolin Lambotte, Alina Kang (SHK)

Inhalt

I. Kurzbericht 3
I.1 Aufgabenstellung 3
I.2 Voraussetzungen, unter denen das Vorhaben durchgeführt wurde 4
I.3 Planung und Ablauf des Vorhabens 4
I.4 Wissenschaftlicher und technischer Stand, an den angeknüpft wurde 5
I.5 Zusammenarbeit mit anderen Stellen 6
II. Eingehende Darstellung 7
II.1 Verwendung der Zuwendung und des erzielten Ergebnisses im Einzelnen, mit Gegenüberstellung der vorgegebenen Ziele 7
Teil-AP 1.1: Bedarfsanalyse 7
Teil-AP 1.2: Praxisanalyse 8
Teil-AP 1.3.1: Behördliche Akzeptanz 10
Teil-AP 1.3.2: Gesellschaftliche Akzeptanz 20
Teil-AP 1.3.3: Risikokommunikation 25
II.2 Wichtigste Positionen des zahlenmäßigen Nachweises 34
II.3 Notwendigkeit und Angemessenheit der geleisteten Arbeit 35
II.4 Voraussichtlicher Nutzen, insbesondere Verwertbarkeit des Ergebnisses im Sinne des fortgeschriebenen Verwertungsplans 35
II.5 Während der Durchführung des Vorhabens dem ZE bekannt gewordener Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen 35
II.6 Erfolgte oder geplante Veröffentlichungen des Ergebnisses 36
Literatur 37

I. Kurzbericht

I.1 Aufgabenstellung

Das Teilprojekt der TUB „Organisationale, rechtliche und gesellschaftliche Rahmenbedingungen“ im Projektverbund „KISTRA: Einsatz von KI zur Früherkennung von Straftaten“ bezog sich auf die Bekanntmachung vom 15.10.2019 „Künstliche Intelligenz in der zivilen Sicherheitsforschung“ im Rahmen des Programms „Forschung für die zivile Sicherheit 2018 bis 2023“ der Bundesregierung, Bundesanzeiger vom 15.10.2019. Gesamtziel von KISTRA war die Erforschung der Möglichkeiten und Rahmenbedingungen für den ethisch und rechtlich vertretbaren Einsatz von Systemen Künstlicher Intelligenz (KI) durch polizeiliche Endanwender:innen zur Erkennung, Vorbeugung und Verfolgung von Straftaten. Dem lagen zwei konkrete Anwendungsszenarien zugrunde: Im Rahmen des novellierten NetzDG wurden Telemediendienstanbieter (TMDA) verpflichtet, strafrechtlich relevante Inhalte dem Bundeskriminalamt zu melden. Das BKA hat hierzu die Zentrale Meldestelle für Straftaten im Internet (ZMI) errichtet. Um die erwarteten massenhaften Meldungen effizient verarbeiten und eine zeitnahe Strafverfolgung zu ermöglichen, sollten in KISTRA eine Reihe von Machine Learning-Klassifizierer zur Bewertung der strafrechtlichen Relevanz der Meldungen entwickelt werden. Nachdem die Meldepflicht der Plattformbetreiber nach erfolgreicher Klage der Plattformbetreiber ausgesetzt wurde, wurde das NetzDG nun durch den europäischen Digital Service Act ersetzt. Das für KISTRA zentrale und antragskonstitutive ZMI-Anwendungsszenario hat sich im Projektverlauf somit als nicht mehr relevant gezeigt. Der Schwerpunkt wurde daher auf das Anwendungsszenario „Staatsschutz“ gelegt. Hierbei ging es darum, die Internetauswertung im Phänomenbereich PMK-rechts durch die zu entwickelnden KI-Modelle in Auswertungsprozessen um Zusammenhang mit Hasskriminalität zu unterstützen.

Ziele des Teilvorhabens waren die Untersuchung der Praxis polizeilicher Auswertetätigkeiten im Bereich Hasskriminalität, die Untersuchung von möglichen Akzeptanzwiderständen bei den Anwender:innen, aber auch bei relevanten Stakeholdern hinsichtlich des Einsatzes von KI bei polizeilichen Auswertetätigkeiten im Bereich Hasskriminalität sowie die Erarbeitung von gesellschaftlich-ethischen Rahmenbedingungen und Anforderungen für den Einsatz von bei entsprechenden Auswertetätigkeiten.

Eine wesentliche Voraussetzung zur Untersuchung von Akzeptanzwiderständen gegen KI-Systeme in der Praxis wie auch zur Formulierung von gesellschaftlich-ethischen Anforderungen an die Gestaltung derartiger Systeme ein genaues Verständnis der Arbeitsweisen und Handlungslogiken der polizeilichen Praktiker:innen und späteren Anwender:innen des in KISTRA zu entwickelnden KI-Systems, aber auch des Akteur:innennetzwerkes (Stakeholderanalyse). Im Mittelpunkt der ersten Phase des Vorhabens standen daher eine Praxis- und Stakeholderanalyse mittels Interviews und Workshops. Dabei ging es darum, relevante Praxiselemente wie Software, Werkzeuge, Daten, Informationen und Interaktionen zu identifizieren und deren so-

zio-technisches Zusammenwirken zu rekonstruieren. Die Praxis- und Stakeholderanalyse bildete die Grundlage für die Anforderungsanalyse an das KI-system, die im Mittelpunkt der zweiten Projektphase stand. Dabei wurden allgemeine gesellschaftlich-ethische Dimensionen von KI-Systemen für das im Fokus von KISTRA stehende Einsatzgebiet des polizeilichen Staatsschutzes konkretisiert und szenarienbezogen ausgearbeitet.

I.2 Voraussetzungen, unter denen das Vorhaben durchgeführt wurde

Das Teilprojekt „Organisationale, rechtliche und gesellschaftliche Rahmenbedingungen“ wurde unter Leitung von Dr. Robert Pelzer am Zentrum Technik und Gesellschaft der Technischen Universität Berlin (TUB) durchgeführt. Er verfügt über langjährige Erfahrung bei der Organisierung und Koordinierung von Forschungsvorhaben. Der Forschungsbereich Sicherheit – Risiko – Kriminologie des ZTG verfügt über umfangreiche Erfahrungen in verschiedenen Themenfeldern ziviler Sicherheit, insb. in der trans- und interdisziplinären Erforschung neuer technischer und nicht-technischer Sicherheitslösungen unter Einbeziehung von Praxispartner:innen, Entwickler:innen, Rechtsexpert:innen sowie zivilgesellschaftlichen Akteure:innen. Für das Vorhaben konnte die TUB auf Erfahrungen in der Übersetzung von sozialwissenschaftlichen Modellen in technische Verfahren und vielfältige Methodenkompetenz zurückgreifen.

Des Weiteren wurde an Vorarbeiten aus dem SIFO-Projekt INTEGER („Visuelle Entscheidungsunterstützung von Analysten bei der Auswertung von Daten aus Sozialen Netzwerken“) angeknüpft werden. Im Rahmen von INTEGER wurden Nutzer:innenszenarien für technisch unterstützte Internetauswertungen im polizeilichen Staatsschutz entwickelt. Die TUB hat Rahmenbedingungen für einen gesellschaftlich-ethisch akzeptablen Einsatz von Tools zur Unterstützung von Internetauswertungen skizziert.

I.3 Planung und Ablauf des Vorhabens

Das KISTRA-Projekt hatte eine Laufzeit von 3 Jahren (Juli 2020 bis Juni 2023) und ist zuwendungsneutral bis zum Dezember 2023 verlängert worden. Das Gesamtvorhaben gliederte sich in fünf Arbeitspakte, wobei das TUB-Teilvorhaben im Wesentlichen an dem ersten Arbeitspaket „Rechtliche, organisationale und ethische Rahmenbedingungen“ beteiligt war. Die fünf Arbeitspakete und die verschiedenen Teil-Arbeitspakete des AP 1 waren zwischen den Verbundpartner:innen eng verflochten, so dass eine stete Abstimmung zwischen ihnen erfolgen musste. Diese hier aufgeführte Kurzbeschreibung nimmt nur auf solche Teil-Arbeitspakete des AP 1 Bezug, an denen die TUB beteiligt war. Entsprechend wird nur auf die Arbeiten der TUB in den entsprechenden Arbeitspaketen Bezug genommen.

Teil-AP 1.1 (Bedarfsanalyse): Durchgeführt wurden eine initiale Bedarfsanalyse und Spezifikation des Anforderungsrahmens, damit das zu entwickelnde KI-System technisch, ethisch und rechtlich sowie für die Anwender bedarfsgerecht erforscht und prototypisch demonstriert werden kann.

Teil-AP 1.2 (Praxisanalyse): In diesem Teil-AP wurde die bestehende Praxis der Erkennung, Vorbeugung und Verfolgung von Hasskriminalität im BKA analysiert. Geplante Hospitationen konnten pandemiebedingt nicht stattfinden, weshalb auf Interviews und Workshops mit Praktiker:innen aus der Zentralen Meldestelle für Straftaten im Internet (ZMI) sowie dem polizeilichen Staatsschutz zurückgegriffen wurde. Die Praxisanalyse beinhaltete v.a. eine Analyse der verschiedenen Prozessschritte und der darin einbezogenen Praktiken, Wissensbestände, Ressourcen sowie organisationalen und interorganisationalen Kommunikationsbeziehungen. Die Ergebnisse der Praxisanalyse flossen in das übergreifende Dokument zur Anforderungsdokumentation ein.

Teil-AP 1.3.1 (Behördenakzeptanz): In diesem Teil-AP wurden die Sichtweisen von kooperierenden Behörden, insb. Staatsanwaltschaften und LKÄ, auf den Einsatz von KI-Systemen zur Verfolgung von Hasskriminalität erhoben. Dabei ging es darum Akzeptanzwiderstände zu identifizieren, um diese frühzeitig bei der Gestaltung der Technik und des organisationalen Umfelds zu berücksichtigen. Die Ergebnisse flossen ebenfalls in das Anforderungsdokument ein. Des Weiteren wurde ein Bewertungsschema für gesellschaftlich verantwortungsvolle und effiziente KI-Systeme in Sicherheitsbehörden entwickelt

Teil-AP 1.3.2 (Gesellschaftliche Akzeptanz): Untersucht wurde die Akzeptabilität des Einsatzes von KI-Systemen in der Polizei durch professionelle Stakeholder und die Bevölkerung. Durchgeführt wurden dazu Interviews mit zivilgesellschaftlichen Akteur:innen. Des Weiteren wurde bereits vorliegendes empirisches Material aus Gruppendiskussionen mit Vertreter:innen der allgemeinen Bevölkerung aus dem Projekt INTEGER in Hinblick auf Anforderungen an den Einsatz von KI-Systemen im polizeilichen Staatsschutz ausgewertet. Die Ergebnisse wurden zudem in das Bewertungsschema integriert.

Teil-AP 1.3.3 (Risikokommunikation): Im Rahmen dieses Teil-AP wurde eine Risikokommunikationsleitfaden zur Adressierung der Risikowahrnehmungen der professionellen Stakeholder und der Ängste und Erwartungen in der Bevölkerung in Bezug auf den Einsatz von KI-Systemen in Sicherheitsbehörden zur Bekämpfung von Hasskriminalität erarbeitet. Auf Grundlage der in den vorherigen Teil-AP erhobenen Risikoszenarien wurden entsprechend diesem Leitfaden abgeleitet Kommunikationsstrategien für die KI anwendenden Sicherheitsbehörden erstellt.

I.4 Wissenschaftlicher und technischer Stand, an den angeknüpft wurde

In Deutschland spielt der Einsatz von KI-Systemen in Sicherheitsbehörden derzeit kaum eine Rolle. Das Programm „Polizei 2020“ erwähnt den Einsatz von künstlicher Intelligenz nicht (BMI 2016). Vielmehr sollen zunächst Strukturen zur Harmonisierung von Datenbeständen, zur Verbesserung der Datenqualität, zum Zugang zu relevanten Daten und zur Gewährleistung des Datenschutzes aufgebaut werden. Da das Programm sich nun bereits seit einigen Jahren in der Umsetzung befindet, ist dies der ideale Zeitpunkt, entsprechende KI-Systeme zu entwickeln.

Der Einsatz von KI in der Polizeiarbeit wurde bisher im Zusammenhang mit „predictive policing“ diskutiert, also Verfahren zur Prognose zukünftigen Kriminalitätsgeschehens. Das Projekt KISTRA zielt jedoch nicht auf die Prädiktion kriminellen Verhaltens, sondern auf die Identifikation und Klassifikation strafrechtlich relevanter Inhalte. Unabhängig vom Einsatz zur Prädiktion oder zur verbesserten Erkennung von Straftaten stellen sich ähnliche Forschungsfragen: Wie kann verhindert werden, dass KI-Systeme Vorurteile reproduzieren und bestimmte Personengruppen stigmatisieren (Alexander 2012, Angwin et al. 2016)? Wie kann gewährleistet werden, dass diese Systeme einem kontinuierlichen Monitoring unterzogen werden, um Diskriminierung und Manipulation von Trainingsdaten zu vermeiden? (Joh 2017, Biggio et al. 2012, 2014, Huang et al. 2017). Zudem sollte bei der Anschaffung von KI-Systemen nicht nur die Sicht der Polizei berücksichtigt werden, sondern auch öffentliches Vertrauen, Privatsphäre und Persönlichkeitsrechte (Joh 2017).

KISTRA-Org baut zur Beantwortung dieser Fragen auf klassischen Akzeptanzmodellen auf und untersucht die Akzeptanz von KI im Arbeitsumfeld, geprägt durch einfache Nutzbarkeit und Nützlichkeit (Venkatesh & Davis 2000). Studien zur Akzeptanz von KI-Systemen betonen zudem Kriterien wie Fairness, Nachvollziehbarkeit und Transparenz. Vertrauen und Transparenz sind wesentliche Akzeptanztreiber für KI-Systeme (Joh 2016). Die Diversität der Akzeptanzfaktoren zeigt die Notwendigkeit einer Konsolidierung der vorhandenen Modelle und die Untersuchung ihrer Gültigkeit in Sicherheitsbehörden. Schließlich ist für die Akzeptanz eine angemessene Risikokommunikation unverzichtbar. Vertrauen, Transparenz, Fairness und Nachvollziehbarkeit sind dabei zentrale Kategorien, um frühzeitig Konflikte zu identifizieren und dialogisch zu adressieren (Renn & Levine 1990, Dombrowsky 1991, Wiedemann 2000, Renn 2003). Drews (2016) stellt jedoch fest, dass Risikokommunikation in den meisten Behörden nicht ausreichend etabliert ist.

I.5 Zusammenarbeit mit anderen Stellen

Es erfolgte während der gesamten Projektlaufzeit eine intensive Zusammenarbeit mit den Verbundpartner:innen. Zu diesen gehörten das Bundeskriminalamt (BKA), die Johannes Gutenberg-Universität Mainz (JGU), die Ludwig-Maximilians-Universität München (LMU), Munich Innovation Labs GmbH (MIL), die RWTH Universität Aachen (RWTH), die Technische Universität Darmstadt (TUD), die Fernuniversität Hagen und die Zentrale Stelle für Informationstechnik im Sicherheitsbereich (ZITiS), die den Verbund koordiniert hat. Der Austausch erfolgte in mehreren gemeinsamen Verbundtreffen, Workshops sowie spezifischen Arbeitstreffen im Rahmen des von der TUB geleiteten AP 1 (Rechtliche, organisationale und ethische Rahmenbedingungen) mit den Partner:innen JGU und RWTH. Im Rahmen der Verbundtreffen und der Praktiker:innen-Workshops erfolgte zudem ein regelmäßiger Austausch mit den assoziierten Partner:innen (Landeskriminalamt Berlin, Max-Planck-Institut zur Erforschung von Gemeinschaftsgütern in Bonn, Polizeipräsidium München, Generalstaatsanwaltschaft München, Landeskriminalamt Nordrhein-Westfalen, Landeskriminalamt Hamburg, Universität Lübeck).

II. Eingehende Darstellung

II.1 Verwendung der Zuwendung und des erzielten Ergebnisses im Einzelnen, mit Gegenüberstellung der vorgegebenen Ziele

Das Gesamtvorhaben wurde in vier Arbeitspaketen realisiert. Das TUB-Teilvorhaben war für das Arbeitspakt 1 verantwortlich, und dort wesentlich in den Teil-AP 1.1, 1.2, 1.3 tätig.

Die TUB konzentrierte sich in seinem Teilprojekt auf die Durchführung einer Praxisanalyse zur Erkennung von Hasskriminalität beim LKA Berlin und BKA ST 52 und erstellte auf dieser Grundlage eine Anforderungsdokumentation für den Einsatz von KI aus praktischer, ethischer und gesellschaftlicher Sicht. Schließlich wurde ein „Bewertungsschema für gesellschaftlich verantwortungsvolle und effiziente KI-Systeme in Sicherheitsbehörden“ entwickelt und ein Leitfaden zur Risikokommunikation erstellt.

Der Meilenstein für das TUB-Teilprojekt lag in Monat 18. Der teilvorhabenspezifische Meilenstein definiert sich wie folgt:

- Das Anforderungsdokument für das technische System aus Sicht der Anwender:innen und der Behördenvertreter liegt vor.
- Die Abschlussdokumentation zur Arbeitspraxis der Erkennung, Vorbeugung und Verfolgung von Hasskriminalität liegt als Bericht vor.

Die Meilensteinziele wurden erreicht.

Das Abbruchkriterium aus Sicht des Teilvorhabens, bei dessen Nichterreichen das Gesamtprojekt hätte beendet werden müssen, trat nicht ein.

Teil-AP 1.1: Bedarfsanalyse

Ziel von Teil-AP 1.1 war die Durchführung einer initialen Bedarfsanalyse und Spezifikation des Anforderungsrahmens, damit das gesamte System sowie Teilsysteme aus anderen AP technisch, ethisch und rechtlich sowie für die Anwender bedarfsgerecht erforscht und prototypisch demonstriert werden können. Das Teil-AP 1.1 stellte die Grundlage für die fortdauernde Detaillierung und Erweiterung des Anforderungsdokuments im Rahmen der nachfolgenden Teil-APe in AP 1 dar.

Arbeiten der TUB

Die TUB hat sich an den Workshops mit dem BKA beteiligt. Dabei wurde insbesondere ein Workshop mit Vertreter:innen verschiedener Abteilungen und des Leitungspersonals im BKA sowie ein Workshop mit Justiz-Praktiker:innen im Bereich der Strafverfolgung von Hasskriminalität zur Diskussion von Erwartungen und Einstellungen in Bezug auf KI-Systeme im Allgemeinen und dem zu entwickelnden System im Speziellen durchgeführt. die Workshopergebnisse wurden als Abschluss des Teil-AP 1.1 (Bedarfsanalyse) sowie als Beitrag zum Gesamtan-

forderungsdokument dokumentiert. Aus den Ergebnissen von Teil-AP 1.1 wurden (inter-)organisationale Übergabe- und Passagepunkte für die Rekonstruktion des Gesamtprozesses sowie Forschungsbedarfe für Teil-AP 1.2 (Praxisanalyse) und Teil-AP 1.3 (Akzeptanzanalyse) abgeleitet.

Teil-AP 1.2: Praxisanalyse

Ziel des Teil-AP war die Analyse der bestehenden Praxis der Erkennung, Vorbeugung und Verfolgung von Hasskriminalität, um in den bisherigen Abläufen relevante Praxiselemente sowie kritische Übergabe- und Passagepunkte zu identifizieren, die auch in der um das technische System ergänzten und veränderten Arbeitspraxis zu berücksichtigen sind. Zu den Praxiselementen gehören z.B. Technologien, Werkzeuge, Daten, zu berücksichtigende Regeln, soziale Interaktionen sowie die Prozesse in die sie eingebettet sind. Übergabepunkte beschreiben z.B. den Erhalt von Rohdaten oder die Weitergabe von Zwischen- oder Endergebnissen an andere Abteilungen, Behörden oder Dritte. Passagepunkte umfassen z.B. die Einholung von Genehmigungen für bestimmte Arbeitsschritte, Zwischenevaluationen oder Arbeitsprozesse, die eine Abstimmungen mit Kollegen voraussetzen.

Arbeiten der TUB

- Auf Basis der in Teil-AP 1.1 identifizierten Übergabe- und Passagepunkte wurden eine soziotechnische Kartierung bzw. Visualisierung des ZMI-Gesamtprozesses erstellt und geeignete Interviewpartner:innen innerhalb und außerhalb des BKA identifiziert.
- Durchführung und Auswertung eines Gruppeninterviews mit BKA-Endanwender:innen, eines Gruppeninterviews mit justiziellen Stakeholdern (gemeinsam mit der Ruhr-Uni Bochum) sowie eines Gruppeninterviews und eines Workshop mit Praktiker:innen des LKA Berlin. Einige Interviews konnten auf Tonband aufgezeichnet werden und wurden anschließend transkribiert, während bei anderen Interviews sowie bei den Workshops lediglich Notizen angefertigt wurden. Ziel der Auswertung war es, einen inhaltlichen Überblick über Prozess- und Arbeitsabläufe bei Internetauswertungen sowie diesbezügliche Anforderungen an KI-Systeme zu erlangen.
- Durchführung einer Literaturrecherche und Auswertung von Stellungnahmen der Telemediendienstanbieter aus öffentlich zugänglichen Quellen zum NetzDG.
- Durchführung einer Literaturrecherche zum Stand der Forschung zur Internetauswertepraxis und Analysetätigkeit bei unterschiedlichen Polizeibehörden. Auf dieser Grundlage wurde ein Konzept für die teilnehmende Beobachtung beim LKA Berlin sowie bei ST 52 im BKA erstellt. Dabei wurde insbesondere deutlich, dass ein umfassenderes soziotechnisches Prozessverständnis erforderlich ist, um die ethischen und gesellschaftlichen Herausforderungen des Einsatzes von KI bei der Polizei praxisadäquat evaluieren zu können. Entsprechend sollten bei der Praxisanalyse folgende Prozessschritte detaillierter in den Blick genommen werden:
    1. Auftragserteilung und Aushandlung
    2. Aushandlungspraktiken bei der Festlegung relevanter Kategorien, Faktoren und Quellenmerkmale

3. Identifikation und Auswahl geeigneter Datenquellen
4. Inferenz und Deduktionspraktiken bei der Datenauswahl
5. Technische Extraktion von Inputdaten
6. Kollektive „Sensemaking“-Prozesse in der Modellentwicklung (soweit im Rahmen der Praxisanalyse beobachtbar)
7. Konfigurationspraktiken bei der Analysestellung (hier insbesondere die Zusammenführung unterschiedlicher Analysetools und menschlichen Erfahrungswissens)
8. Technische Auswertungsverfahren
9. Darreichungsformen der Ergebnisse
10. Kollektive „Sensemaking“-Prozesse bei der Interpretation der visuell aufbereiten Ergebnisse, inkl. der Filterung, Sortierung, Kalibrierung und Rekonfiguration der Analysestellung
11. Kontroll- und Reparaturpraktiken zur abschließenden Freigabe oder Widerrufung der Ergebnisse
12. Berichtslegung der Ergebnisse (hier insbesondere die Zusammenführung menschlicher Interpretation und technisch generierter Daten und Zusammenhänge)
13. Interpretation und kollektive Aushandlung der in Berichtsform dargereichten Ergebnisse durch Entscheider:innen und/oder externe Prozessbeteiligte sowie daraus resultierender Entscheidungen
14. Kommunikation und Rezeption der Entscheidungen durch die Maßnahmenausführenden

- Beim BKA fand eine halbtägige Praxisbeobachtung am Standort Meckenheim statt. Die Mitarbeitenden der TUB hatten die Möglichkeit die operative und strategische Internetauswertung bei ST 52 zu begleiten und Nachfragen zur Auswertungspraxis, auftretenden analytischen und technischen Herausforderungen sowie Unterstützungspotentialen für den Einsatz von KI zu besprechen
- Zudem wurde ein Workshop mit Auswerter:innen von ST 52 im BKA durchgeführt, um ausgewählte Aspekte der Arbeitsweisen der strategischen und operativen Auswertearbeit vertiefend zu erörtern und um Nutzungsszenarien für den Einsatz der KISTRA-KI zu entwickeln. Die Nutzungsszenarien wurden ausformuliert und dem BKA zur Überarbeitung und Kommentierung vorgelegt.
- Im Rahmen eines zweiten Workshops wurden die Nutzungsszenarien inhaltlich angepasst. Im Rahmen des Workshops wurde zudem vereinbart, das Szenario „Aktionstag zur Verfolgung von Hasskriminalität“ beim BKA vertiefend weiterzuverfolgen.
- Darüber hinaus wurden die Rahmenbedingungen für die Durchführung weiterer Workshops zur detaillierten Anpassung des Nutzungsszenarios an die ethischen Anforderungen und die Möglichkeit einer Bewertung erörtert.
- In diesem Zusammenhang fanden gemeinsame prozessanalytische Modellierungen des ausgewählten Szenarios „Aktionstag zur Verfolgung von Hasskriminalität“ sowie ein Workshop zur Validierung mit einem Gruppenleiter in ST 52 statt. Das ausformulierte

Szenario stellte die Grundlage für die Entwicklung des ethischen Bewertungsschemas in Teil-AP 1.3 dar.

Abänderungen gegenüber der Planung

Im Projektverlauf hat sich gezeigt, dass die technische Entwicklung der KI unabhängig von der Anwendungsumgebung ZMI realisiert werden muss, da die Softwareumgebung zur Bearbeitung der Meldungen von Telemediendienstanbietern aufgrund des bestehenden Zeitdrucks durch das Inkrafttreten des Gesetzes am 01. Februar 2022 von einer externen Firma realisiert wurde.

Für die Entwicklung eines eigenständigen Softwaredemonstrators, wie es in KISTRA angestrebt wurde, haben die technischen Partner:innen vor diesem Hintergrund die Anwendungsumgebung auf die Arbeitsumgebung der Abteilung Staatsschutz im BKA ausgeweitet, da dort ebenfalls Auswertungen zu Hasskriminalität erfolgen. Hierfür waren jedoch neue Nutzungsszenarien ebenso erforderlich wie die Ausweitung der Praxisanalyse. Um diesen Mehraufwand der Ausweitung der Praxisanalyse in Teil-AP 1.2 zu kompensieren, wurde eine Verschiebung von Personalressourcen von Teil-AP 1.3.2 (Gesellschaftliche Akzeptanz) nach Teil-AP 1.2 im Umfang von 4 Personenmonaten beantragt und vom Projektträger genehmigt.

Im Rahmen der erweiterten Praxisanalyse sollten neben Interviews und Workshops zur Rekonstruktion der Arbeitsweise und Herausforderungen auch Praxisbeobachtungen nicht nur beim BKA, sondern zusätzlich auch beim LKA Berlin durchgeführt werden. Diese Praxisbeobachtungen konnten jedoch anders als ursprünglich geplant aufgrund verschärfter Corona-Auflagen bei der Polizei Berlin trotz intensiver Vorbereitung nicht durchgeführt werden. Zwar wäre eine Beobachtung gegen Ende des Projekts unter Umständen noch möglich gewesen, wovon jedoch abgesehen wurde, da die Auswertungsergebnisse erst zu spät vorgelegen hätten und nicht mehr in das Bewertungsschema hätten einfließen können. Dies hatte zur Folge, dass eine Praxisbeobachtung im geplanten Umfang beim LKA Berlin nicht erfolgte. Teilweise konnte dies durch Gruppeninterviews kompensiert werden (siehe Darstellung der Aktivitäten oben). Gleichwohl ist auf dieser Basis nur eine rudimentäre Praxisanalyse möglich gewesen. Unabhängig davon bestand der erhöhte Arbeitsaufwand größtenteils dennoch, da umfassende Vorarbeiten bereits durchgeführt wurden und die Praxis von ST 52 des BKA in Teilen durchgeführt wurde.

Die Verzögerungen beim Feldzugang haben sich auf den Zeitplan insgesamt ausgewirkt, da die Ausarbeitung detaillierter Nutzungsszenarien als Ergebnis des Teil-AP 1.2 erst später als geplant erfolgen konnte. Diese Verzögerungen haben sich auf die Folgearbeitspakete, in deren Mittelpunkt die Erarbeitung gesellschaftlich-ethischer Anforderungen steht, ausgewirkt.

Teil-AP 1.3.1: Behördliche Akzeptanz

Ziel war es abzuschätzen, welche Auswirkungen die Einführung der neuen Technologie auf vor- und nachgelagerte Abteilungen sowie andere Sicherheitsbehörden hat und welche Inkompatibilitäten und Akzeptanzwiderstände zu erwarten sind, um diese frühzeitig bei der Ge-

staltung der Technik und des organisationalen Umfelds zu berücksichtigen. Hierzu sollten Interviews mit ausgewählten Behördenvertreter:innen z.B. der Leitungsebenen des BKA, Staatsanwält:innen, Auswertenden in LKÄ, usw. hinsichtlich ihrer Aufgaben und Beiträge zur Arbeit der mit Hasskriminalität im BKA befassten Auswerter:inen und Ermittler:innen geführt werden. Auf dieser Grundlage sollte ein Bewertungsschema für gesellschaftlich verantwortungsvolle und effiziente KI-Systeme in Sicherheitsbehörden entwickelt werden.

Arbeiten der TUB

- Durchführung und inhaltsanalytische Auswertung von Interviews mit der Staatsanwalt einer Spezialdienststelle sowie Internetauswertenden in einem weiteren LKA sowie in einem Polizeipräsidium zu Fragen der Akzeptanz des Einsatzes von KI zur Auswertung von Hasskriminalität im Rahmen des ZMI-Prozesses.
- Es wurde ein Workshop mit den AP 1-Partner:innen sowie einem Vertreter von MIL durchgeführt, um die technischen Rahmenbedingungen für die ethische und rechtliche Bewertung der Nutzungsszenarien abzuschätzen. Dabei ist deutlich geworden, dass die weitere Erarbeitung von Anforderungsprofilen für ethische und gesellschaftliche Präventions- und Kontrollkomponenten an eine höhere Fehlerrate der KI angepasst werden muss.
- Es wurde ein „Bewertungsschema für gesellschaftlich verantwortungsvolle und effiziente KI-Systeme in Sicherheitsbehörden“ entwickelt, um allgemeine Regeln für die Anschaffung, Implementierung und den Einsatz von KI-Systemen abzuleiten. Dazu wurden relevante ethische Dimensionen und Schutzgüter abgeleitet und beschrieben. Dies geschah einerseits auf Grundlage der Ergebnisse der Praxisanalyse, als auch mit Hilfe von umfassenden Literatur- und Online-Recherchen zur Identifikation von bereits durchgeführten Anwendungsfällen von KI in Sicherheitsbehörden auf nationaler und internationaler Ebene sowie deren Auswertung. Schließlich wurden die Ergebnisse der Interviewstudie mit Stakeholdern aus dem Jahr 2020 hinsichtlich weiterer Dimensionen der behördlichen Akzeptanz von KI-Systemen ausgewertet.
- Der Schwerpunkt der Arbeiten lag in der Spezifizierung allgemeiner ethischer Anforderungen für den Einsatz von KI-Systemen (in Polizeien bzw. Hochrisikobereichen im Allgemeinen) für polizeiliche Internetauswertungen im Staatsschutz, im Speziellen für da Szenario „Aktionstag zur Verfolgung von Hasskriminalität“. Erarbeitet wurde ein umfangreiches Anforderungsdokument, welches den gesamten Auswerteprozess, in den KI eingebunden werden soll, abbildet. Dieses Anforderungsdokument bildet die Grundlage für weitere Verwertungsschritte nach Projektende wie die Erstellung eines vereinfachten Leitfadens für die Praxis.
- Präsentation des entwickelten Bewertungsschemas auf zahlreihen Konferenzen und Fachtagen (siehe II.6); Austausch mit anderen Stellen (KI-Campus, ZiTIS), die sich mit der Bewertung von KI in Polizeibehörden befassen. Durchführung eines Validierungsworkshops mit Wissenschaftler:innen ohne Projekteinbindung.
- Das Bewertungsschema gliedert sich in folgende sieben Dimensionen und damit verbundenen Schutzgüter:

1. Effektivität o Unterstützung der Gefahrenabwehr & Strafverfolgung o Präzision o Zuverlässigkeit o Reproduzierbarkeit o Widerstandsfähigkeit 2. Autonomie o Selbstbestimmung o Menschliche Aufsicht o Abschaltmöglichkeit o Unterstützung und Entlastung 3. Freiheitsrechte o Privatsphäre o Datenschutz o Informationelle Selbstbestimmung o Meinungsfreiheit o Bewegungsfreiheit o Kontaktfreiheit	4. Fairness o Nichtdiskriminierung o Beteiligung marginalisierter Gruppen & passiver stakeholder 5. Transparenz o Rückverfolgbarkeit o Erklärbarkeit o Information über KI-Einsatz 6. Rechenschaftspflicht o Verantwortlichkeit o Gewährleistung der Minimierung negativer Folgen o Rechtsmittel 7. Gesellschaftliches Wohlergehen o Arbeitsbedingungen o Sicherheitsempfinden o Institutionenvertrauen

- Die ausformulierten Kriterien des Bewertungsschemas wurden mit den im EU AI Act-Entwurf beschriebenen Anforderungen an verantwortungsvolle KI abgeglichen. Zudem wurden Verweise auf den EU AI-Act im Anforderungsdokument eingearbeitet.
- Die konstruktive Bewertung innovativer Technologien setzt voraus, dass prozessspezifische Risikoszenarien entwickelt und den erwarteten Chancen gegenübergestellt werden. Um diesen Prozessschritt in der Praxis effizient durchführen zu können, wurden die genannten empirischen Daten hinsichtlich möglicher Risikoszenarien ausgewertet und entsprechend der ethischen Dimensionen und Schutzgüter systematisiert. Dazu wurde sowohl eine Literaturrecherche als auch ein Expertenworkshop mit vier Teilnehmern durchgeführt. In diesem Rahmen wurde das Anwendungsszenario vorgestellt und mögliche Risiken entlang eines Prozessmodells erörtert. Differenziert wurden die Risikoszenarien dahingehend, ob sie ursächlich im Rahmen der Datensammlung (D), der Analyse durch die KI (A) oder in Folge von menschlichen Fehlern im Umgang und der Kontrolle der KI (K) sind. Es wurden Risikoszenarien zu folgenden Aspekten herausgearbeitet:
- Effektivität
    - (A) Funktionsausfälle, Mangelnde Zuverlässigkeit, Hoher Wartungsbedarf → Hohe Betriebskosten bei geringer Effektivität; Legitimationsprobleme

- (A, D) Geringe precision der Ergebnisse (Hohe falsch positiv Rate im Verhältnis zu den positiv Klassifizierten) → Mehrarbeit durch zusätzliche Nachkontrollen
- (A, D) Geringe Spezifität der Ergebnisse (Hohe falsch positiv Rate im Verhältnis zu den korrekt negativ Klassifizierten) → Vergleichsweise hoher Anteil unschuldig Verdächtigter; ungerechtfertigte freiheitseinschränkende Maßnahmen; sinkendes Vertrauen in Polizei, Rechtstaat und Demokratie; chilling effect
- (A, D, K) Geringer Recall der Ergebnisse (Hohe falsch negativ Rate im Vergleich zu den korrekt positiv Klassifizierten) → Geringe Effektivität; Täter werden übersehen; Sicherheitsniveau beeinträchtigt; sinkendes Vertrauen in Polizei, Rechtstaat und Demokratie
- (A) Mangelhafte Reproduzierbarkeit der Ergebnisse → Misstrauen in die Technik; Vermeidungsverhalten; Mehrarbeit
- (A, K) Mangelhafte Prozessintegration → Vermeidungsverhalten; Mehrarbeit
- (A, K) Arbeitsentlastung bleibt aus → Mehrarbeit; Vermeidungsverhalten; Stress; Zunahme von Fehlentscheidungen; sinkende Arbeitsqualität
- (A, K) Bedienungsfehler durch mangelhafte Usability oder unzureichende Schulung → Misstrauen in die Technik; Vermeidungsverhalten; Mehrarbeit
- (K) Zweckentfremdung – function creep → Anwendung der KI für nicht legitime Zwecke; Ergebnisse minderer Qualität; Zunahme von Fehlentscheidungen; sinkendes Vertrauen in Polizei, Rechtstaat und Demokratie

- Transparenz
  - (A, K) Die KI Funktionsweise und/ oder das Zustandekommen der KI Ergebnisse können durch die Anwender und/oder andere Experten nicht nachvollzogen werden. → Zunahme von Fehlentscheidungen; Misstrauen in die Technik; Vermeidungsverhalten; Mehrarbeit; automation bias; blindes Vertrauen in KI
  - (A, K) Fehlerhafte Klassifizierungen oder Bewertungen werden aufgrund mangelhafter oder fehlender Erklärungen nicht erkannt. → Zunahme von Fehlentscheidungen; Mehrarbeit
  - (K) Gegenüber den von Datenerfassung, Auswertung oder Maßnahmen Betroffenen wir die Verwendung und die Art des KI Einsatzes nicht kommuniziert. →Verteidigung und Widerspruchsmöglichkeiten werden erschwert; Chilling effect; sinkendes Vertrauen in Polizei, Rechtstaat und Demokratie.
  - (K) Intransparenz der Nutzung und Auswertungspraxis von KI gegenüber der Öffentlichkeit. → sinkendes Vertrauen in Polizei, Rechtstaat und Demokratie
  - (K) Ungewollte Veröffentlichung der Auswertungspraxis→ Anpassung der Straftäter an die KI; sinkendes Sicherheitsniveau; sinkendes Vertrauen in Polizei, Rechtstaat und Demokratie
- Autonomie
  - (K) Automation bias → übermäßiges und falsches Vertrauen in KI; mangelhafte Kontrolle; Zunahme von Fehlentscheidungen; sinkendes Sicherheitsniveau; sinkendes Vertrauen in Polizei, Rechtstaat und Demokratie

- (K) Verstoß gegen Dienstvorschriften, um mit der Performance der KI bzw. den Erwartungen an den KI Einsatz gerecht zu werden; Scheinprüfungen → Stress; sinkende Arbeitszufriedenheit; mangelhafte Kontrolle; Zunahme von Fehlentscheidungen; sinkendes Sicherheitsniveau; sinkendes Vertrauen in Polizei, Rechtstaat und Demokratie
- (K) Mangelhafte Schulung der Anwender → mangelhafte Kontrolle; Zunahme von Fehlentscheidungen; sinkendes Sicherheitsniveau; sinkendes Vertrauen in Polizei, Rechtstaat und Demokratie

- Verhältnismäßigkeit
  - (A) Überschätzung des Nutzens bei der Bewertung der Geeignetheit → Entwicklung und Investionen in ineffektive Technolgien; Mehrarbeit; sinkende Arbeitszufriedenheit; Vermeidungsverhalten
  - (A; K) Fehlende /mangelhafte Technikfolgenabschätzung im Entwicklungsprozess → Unterschätzung der Folgen bei der Bewertung von Erforderlichkeit der KI-Anwendung; Unverhältnismäßige Abwägung zwischen Folgen und Nutzen bei der Bewertung der Angemessenheit der KI-Anwendung übermäßige freiheitseinschränkende Maßnahmen; übermäßige Zahl falscher Verdächtigungen; Mehrarbeit; sinkendes Vertrauen in Polizei, Rechtstaat und Demokratie
- Rechenschaftspflicht
  - (K) Verantwortung wird auf KI abgewälzt → Bei Fehlentscheidungen findet sich kein Verantwortlicher; Fehler werden nicht korrigiert; kein organisationales Lernen; sinkendes Vertrauen in Polizei, Rechtstaat und Demokratie
  - (K) Fehlende /mangelhaftes Monitoring der Güte, des KI Einsatzes oder durch KI beeinflussten Maßnahmen → Zunahme von Fehlentscheidung; sinkende Modellgüte z.B. durch model drift; sinkendes Sicherheitsniveau; Mehrarbeit; sinkende Arbeitszufriedenheit; sinkendes Vertrauen in Polizei, Rechtstaat und Demokratie
  - (K) Fehlende /mangelhafte Feedbackkanäle und Feedbackbearbeitung →eingeschränktes organisationales Lernen; Zunahme von Fehlentscheidungen
  - (K) Fehlende /mangelhafte Möglichkeiten oder Erschwerung des Rechtswegs → Eingeschränkte Möglichkeiten zum Widerspruch von Betroffenen; Zunahme von Fehlentscheidungen; sinkendes Vertrauen in Polizei, Rechtstaat und Demokratie
- Fairness
  - (A, K) Diskriminierung und/oder Stigmatisierung von Personengruppen (Verdächtige / Opfergruppen) → Mehrarbeit; Übernahme der KI Logik; automation bias; Zunahme von Fehlentscheidungen; einseitige Zunahme freiheitseinschränkende Maßnahmen zu Ungunsten bestimmter Bevölkerungsgruppen; Stigmatisierung von Personenggruppen durch öffentliche Wahrnehmung; selektiv sinkendes Vertrauen in Polizei, Rechtstaat und Demokratie
  - (A, K) Die Anwendenden übernehmen die in die KI Logik eingeschriebenen Vorurteile → automation bias; einseitige Zunahme freiheitseinschränkende Maßnah-

men zu Ungunsten bestimmter Bevölkerungsgruppen; Stigmatisierung von Personenggruppen durch öffentliche Wahrnehmung; selektiv sinkendes Vertrauen in Polizei, Rechtstaat und Demokratie
    - (K) Die Anwendenden korrigieren nicht-diskriminierende Vorauswahlen der KI auf Basis ihres vorurteilsbehafteten „Bauchgefühls“ → Vertrauensverlust in die KI; Vermeidungsverhalten; Fehlentscheidungen; Diskriminierung und Stigmatisierung von Bevölkerungsgruppen; selektiv sinkendes Vertrauen in Polizei, Rechtstaat und Demokratie
    - (A) Selektive Nichtberücksichtigung bestimmter Tatqualitäten (z.B.: Morde mit Messern werden nicht erkannt) → geringer Recall; Geringe Effektivität der KI; Täter werden übersehen; Sicherheitsniveau beeinträchtigt; Diskriminierung und Stigmatisierung von Bevölkerungsgruppen; selektiv sinkendes Vertrauen in Polizei, Rechtstaat und Demokratie
- Freiheitsrechte
    - (A) ungerechtfertigte Ausweitung ins strafrechtliche Vorfeld → Kriminalisierung von bis dato lediglich normabweichendem Verhalten; chilling effect; sinkendes Vertrauen in Polizei, Rechtstaat und Demokratie
    - (A) Entscheidungen auf Basis nicht validierter Bewertungskriterien /Features → mangelnde Güte der Ergebnisse; Fehlentscheidungen; ungerechtfertigte freiheitseinschränkende Maßnahmen; Diskriminierung und Stigmatisierung; sinkendes Vertrauen in Polizei, Rechtstaat und Demokratie; sinkendes Sicherheitsniveau
    - (D) Unverhältnismäßige Datenerhebung; Datenauswertung; Datenspeicherung → ungerechtfertigte freiheitseinschränkende Maßnahmen; Massenüberwachung; sinkendes Vertrauen in Polizei, Rechtstaat und Demokratie; sinkendes Sicherheitsniveau; chilling effect
    - (A, K) Verkettung / (Massen-) profiling → Kriminalisierung von bis dato lediglich normabweichenden Verhaltens; Ausweitung und Nutzung personenbeziehbarer Daten ohne entsprechende rechtliche Grundlage; Massenüberwachung; ungerechtfertigte freiheitseinschränkende Maßnahmen; Diskriminierung und Stigmatisierung von Bevölkerungsgruppen; sinkendes Vertrauen in Polizei, Rechtstaat und Demokratie; sinkendes Sicherheitsniveau; chilling effect
    - (A) Gefahreneinschätzung aufgrund nicht validierter Kriterien wird erstellt → Kriminalisierung von bis dato lediglich normabweichendem Verhalten; ungerechtfertigte freiheitseinschränkende Maßnahmen; Diskriminierung und Stigmatisierung von Bevölkerungsgruppen; sinkendes Vertrauen in Polizei, Rechtstaat und Demokratie; sinkendes Sicherheitsniveau; chilling effect

- Das Bewertungsschema wurde entsprechend der Prinzipien des „constructive technology assessment“ sowie des „responsible research and innovation“ operationalisiert und dokumentiert. Dazu wurde eine Matrix erstellt, in der Anforderungen auf der technischen

wie organisationalen Ebene formuliert wurden. Dabei wurden jeweils die ethischen Dimensionen und die zugeordneten Risikoszenarien mit den KI Prozessdimensionen (1.) Datensammlung und Training, (2.) Modellentwicklung und Betrieb sowie (3.) menschliche Kontrolle und Entscheidung ins Verhältnis gesetzt. Zentrale Erkenntnis hinsichtlich der Anforderungen ist, dass ein ganzheitliches Verfahren zur kontinuierlichen Bewertung des Einsatzes von KI nicht nur für die Erkennung von Hasskriminalität, sondern für den Einsatz von KI im Allgemeinen etabliert und belastbare und nachprüfbare Strukturen der Rechenschaftspflicht aufgebaut werden müssen. Hierfür haben wir ein Modell entwickelt, dass zentrale Aufgaben an drei verschiedenen Stellen – einer staatlichen Prüfstelle, einer polizeiinternen Prüfstelle und bei der internen fachlichen Aufsicht – entsprechend verfügbarer behördlicher Kompetenzen bündelt. So soll die zusätzliche Arbeitsbelastung der einzelnen Dienststellen minimiert und der Aufwand zum Aufbau neuer Strukturen gesenkt werden.

Im Einzelnen sollte die **staatliche Prüfstelle** zunächst die Aufgabe haben, die grundsätzliche **Legitimität** neuer **Klassifizierer** zu überprüfen und die Einhaltung der Konformitätsprüfungen bei den Polizeibehörden zu überwachen. Aus unserer Sicht sind hier folgende Aufgaben erforderlich:

- **Prüfung der Geeignetheit:** Geprüft wird, ob der Klassifizierer geeignet ist, Text- oder Bilddaten (technische Gegenstände) hinsichtlich der vorgegebenen oder versprochenen Inhaltsklasse mit ausreichender **Güte,** also mindestens besser als der Zufall oder besser als im Status Quo (Präzision; Spezifität, Recall) zu klassifizieren. Dies erfordert einen Test des Klassifizierers anhand von **phänomenbereichsspezfischen Validierungsdaten** nach wissenschaftlichen Kriterien durch die staatliche Prüfstelle. Hier besteht jedoch eine grundsätzliche aber nicht für die Prüfstelle spezifische **Herausforderung**: Für jede Validierung bestehender oder neu angebotener oder selbst entwickelter KI-Anwendungen müssen Daten beschafft, getestet, gepflegt und erneuert werden. Wie diese Daten beschafft werden können, ist gerade im Kontext von Sicherheitsfragen eine bisher unbeantwortete, aber für den Erfolg von KI-Systemen generell zentrale Frage.
- **Fairnessprüfung**: Es ist eine Fairnessprüfung durchzuführen, bei der nachgewiesen werden muss, dass der Klassifizierer keine **Verzerrungen** hinsichtlich gesellschaftlicher möglicher Verdächtigen- und Opfergruppen erzeugt. Auch hierfür stellt sich wieder das Problem, dass gut annotierte Validierungsdaten vorliegen müssen, um die Fairness beurteilen zu können.
- **Festlegung** der **Phänomenbereiche** und ggf. **Anwendungsbereiche**, für die der Klassifizierer eingesetzt werden darf. Dabei sind wissenschaftliche Kriterien anzugeben, die nachweisen, dass der Gegenstand des Klassifizierers für den Phänomenbereich relevant und geeignet ist, z.B. die Aufklärung von Straftaten zu fördern oder Gefahren zu prognostizieren.
- **Ausweitung und Anpassung von Phänomenbereichen:** In der Praxis ist zu erwarten, dass der Wunsch nach Ausweitung der Phänomenbereiche, für die Klassifizierer angewendet werden dürfen, formuliert wird. Dies wird auch als **function creep** bezeichnet. Dies soll

nicht grundsätzlich ausgeschlossen werden. Es ist aber erforderlich, dass stets eine **erneute Prüfung von Effektivität und Fairness** durchgeführt wird, auch wenn dies bedeutet, dass ein weiteres Mal Validierungsdaten erforderlich sind.

- **Wartung der KI**: Für jede KI braucht es im laufenden Betrieb kontinuierlich aktuelle Validierungsdaten, um **„model drift"** zu identifizieren und die Geeignetheit auch langfristig sicherzustellen. Das heißt anhand aktueller Validierungsdaten ist in regelmäßigen Abständen die Güte und Fairness der Klassifizierer zu prüfen und wenn erforderlich auch ein Nachtraining oder die Außerbetriebsetzung zu veranlassen. Ein Nachtraining benötigt wiederum eigene Trainingsdaten. Wurde die KI von einem Drittanbieter erworben, sollte regelmäßiges Nachtraining im Rahmen eines Wartungsvertrag berücksichtigt werden, wobei Güte- und Transparenzkriterien hinsichtlich der Trainingsdaten auszuhandeln sind. Die staatliche Prüfstelle hat also nicht nur die Aufgabe, die Klassifizierer regelmäßig zu prüfen, sondern auch die Wartungskonditionen festzulegen: In welchem Intervall sind KI Systeme zu testen? Wer ist Verantwortlich und wer übernimmt die Kosten?
- **Kontrolle der internen Prüfung**: Neben der Prüfung der Güte und Fairness der Klassifizierer an sich, sollte der Staatlichen Prüfstelle auch die Aufgabe obliegen, die **auswertungsprojektspezifischen Prüfungen innerhalb der Auswerteeinheiten zu überwachen**. Dazu sollten regelmäßige **Audits** vor Ort durchgeführt werden.

Hat die staatliche Prüfstelle die Aufgabe, die prinzipielle Güte, Verzerrungsfreiheit und Geeignetheit eines KI-Systems zu prüfen, kann sie jedoch keine Aussagen über die Verhältnismäßigkeit des Einsatzes der KI für spezifische Auswertungsprojekte vornehmen. Dies erfordert vielmehr hinreichendes Kontextwissen über die verschiedenen Einsatzszenarien bei den anwendenden Polizeien. Hier muss entsprechend eine Verhältnismäßigkeitsprüfung für den Einsatz der KI in einem spezifischen Auswertungsprojekt vorgenommen werden. Die Verhältnismäßigkeitsprüfung umfasst die Prüfung der Legitimität des Auswertungszwecks sowie der Geeignetheit der Erforderlichkeit und der Angemessenheit der Prozessschritte. Diese Aufgaben benötigen unterschiedliche Kompetenzen. So erfordert die Prüfung der Legitimität, der Geeignetheit und Erforderlichkeit in erster Linie fachliche Expertise. Entsprechend sollten diese Aufgaben durch die jeweiligen fachlichen Vorgesetzte und Verantwortlichen für das Auswerteprojekt erfolgen. Demgegenüber erfordert die Prüfung der Angemessenheit aber eine fundierte rechtliche und ethische Expertise. Entsprechend sehen wir hier die Notwendigkeit einer eigenständigen internen Prüfstelle, die aber beispielsweise an den bereits vorhandenen Justiziariaten angesiedelt werden kann, die im Zuge der zunehmenden Einführung von KI Systemen ohnehin weiterqualifiziert werden müssen. Im Einzelnen sehen wir folgende Aufgaben:

Bei der **fachlichen Aufsicht**:

- **Zweckprüfung**: Sind die Annahmen hinsichtlich der angestrebten operativen und strategischen Ziele sowie der betrachteten Straftaten oder Gefahren plausibel und liegt ein hinreichender Anlass vor, der das Auswertungsprojekt rechtfertigt.
- **Legitimität der Datenquellen**: KI Systeme sind auf große Datenmengen angewiesen. Dennoch muss gewährleistet werden, dass die Auswertung einer Datenquelle legitim ist,

insofern mit dieser eine Auswertung personenbezogener Daten einhergeht. Für die zur Auswertung vorgesehenen Datenquellen ist daher jeweils zu prüfen, inwiefern hinreichend Anhaltspunkte oder eine valide Prognose bzgl. der Relevanz der Daten vorliegt. Nur so lässt sich die Aufnahme derselben in die Datenbasis rechtfertigen.

- **Geeignetheit der Datenquelle**: Eine weitere Aufgabe stellt die Prüfung der Geeignetheit der Datenquellen dar. Dies bedeutet in erster Linie, ob die erforderlichen Daten technisch extrahiert und die Daten im Sinne der Ziele weiterverarbeitet werden können (z.B. die Nutzeridentifikation bei Strafverfolgung, Scraping bei Trend- oder Sentimentanalysen).
- **Erforderlichkeit der Datenquelle**: Nachzuweisen sind die Erforderlichkeit des Datenumfangs, der jeweils einzelnen personenbezogenen Daten, der Speicherdauer, der Datenzugriffsrechte und der Verfügbarkeit von Weiterverarbeitungskapazitäten. Kurz gesagt geht es hier um die Einhaltung des Datenschutzes.
- **Legitimität der Klassifizierer**: (1.) Nachweis bzw. Dokumentation, dass die Phänomenbereiche des Auswerteprojekts den für den Klassifizierer legitimierten Phänomenbereichen entsprechen. (2.) Spezifische Prüfung der Phänomengerechtigkeit in Bezug auf die inhaltlichen Gegenstände des Auswerteprojekts: D.h. es müssen hinreichend Anhaltspunkte im Rahmen des Anlasses vorliegen, dass die auszuwertenden Gegenstände grundsätzlich im Gültigkeitsbereich des Klassifizierers liegen.
- **Erforderlichkeit der Klassifizierer**: Es ist nachzuweisen, inwiefern jeder einzusetzende Klassifizierer einen Mehrwert für die operativen Ziele bringt. Dies kann z.B. bedeuten zu plausibilisieren, dass KI-spezifische Fähigkeiten wie z.B. die Verarbeitung von Massendaten, die Mustererkennung oder die Detektion von bisher unbekannten Zusammenhängen einen Mehrwert gegenüber dem Status Quo für das spezifische Auswerteprojekt darstellt. Es muss deutlich werden, dass eine Datenverarbeitung tatsächlich auf KI angewiesen ist und weder manuell noch durch regelbasierte Systeme bessere Ergebnisse erzielt werden können.

Bei der internen Prüfstelle:

- **Angemessenheitsprüfung:** Bei der Angemessenheitsprüfung geht es um die Abwägung zwischen den möglichen negativen Folgen und dem möglichen Nutzen. Dazu ist zunächst erforderlich, mögliche Eingriffe in Freiheitsrechte im Rahmen des Auswerteprojekts zu identifizieren. Entsprechend dieser Eingriffe ist die Intensität möglicher negativer Folgen zu bewerten. Dabei soll gelten: Je größer die Zahl potentiell Betroffener, je seltener das Vorkommens einer Straftat oder eines Gefahrentatbestandes in der Datenquelle, je größer die Wahrscheinlichkeit als Person identifiziert zu werden weil z.B. keine Anonymität möglich ist, je unschärfer die Verdachtsschwelle und damit die Wahrscheinlichkeit in Verdacht zu geraten und je größer die Intensität möglicher realweltlicher Folgeeingriffe, desto größer ist die Intensität der negativen Folgen des Auswerteprojekts insgesamt. Es folgt analog die Bewertung des erwarteten Nutzens der mit dem Einsatz von KI verbundenen Eingriffe. Dabei soll gelten: je schwerer das Delikt oder die drohende Gefahr, je

größer die Anzahl der Opfer, je höher der Grad der erwarteten Erreichung der operativen Ziele, desto größer der ethische Nutzen des Auswerteprojekts. Schließlich sind diese Werte gegenüberzustellen und gegeneinander abzuwägen. An dieser Stelle wird deutlich, dass es eines Katalogs bedarf, wie die verschiedenen Einflussfaktoren in standardisierte Bewertungen überführt werden können. Wann ist also z.B. die Anzahl der Opfer groß und in welchem Verhältnis soll dies zu einer großen Anzahl von potentiell Betroffenen von Freiheitseingriffen stehen.

- **Planung und Prüfung der Kontrollpraxis:** Auch nachdem diese Fragen geklärt sind, ist nicht immer zu erwarten, dass eine eindeutige Bewertung möglich ist; oder es kommt zu Pattsituationen. Daher sollte es der internen Kontrollstelle obliegen, in unentscheidbaren Situationen oder Pattsituationen eine spezifische Kontrollpraxis zu planen. Dies kann z.B. bedeuten, eine erhöhte Kontroll- oder Evaluationsdichte anzuordnen oder vor dem Einsatz einer KI eine Erprobung an begrenzten Datensätzen vorzunehmen. Aber auch im Regelbetrieb sollte der internen Prüfstelle die Aufgabe zufallen, die menschliche Kontrollpraxis der KI-Ergebnisse zu überprüfen. Insbesondere ist erforderlich regelmäßig zu prüfen, dass sich bei den Auswertenden kein automation bias einschleicht, es also zu Scheinprüfungen kommt oder die Auswertenden von den Prüfaufgaben aus Gründen der Komplexität oder Kapazität überfordert sind.
- **Organisationales Lernen:** Um diesen Problemen entgegenzusteuern bedarf es schließlich eines kontinuierlichen Kompetenzaufbaus und kontinuierlicher Reflektion der eigenen KI Nutzung. Das heißt, dass die Anwender hinsichtlich der Risiken des KI Einsatzes und der Herausforderungen bei der Bewertung von Auswerteprojekten kontinuierlich sensibilisiert werden sollten. So kann auch verhindert werden, dass es zu einer ungewollten Übernahme von KI-Logiken und in der Folge zu Scheinprüfungen oder eine Zunahme des automation bias kommt. Daher sollte die interne Prüfstelle regelmäßig Reflektionsgespräche zu unklaren Auswertungssituationen aber auch zu abgeschlossenen Auswerteprojekten durchführen.

Abänderungen gegenüber der Planung

Ein Bewertungsschema, dass sich ausschließlich auf die behördliche Perspektive bezieht, erschien aus ethischer Perspektive nicht sinnvoll. Insbesondere die Dimensionen Freiheitsrechte, Fairness, gesellschaftliches Wohlergehen und in Teilen auch Transparenz machen eine über die organisationalen Anforderungen der Polizei hinausgehende Betrachtung erforderlich. Daher wurde ein integrierter Bewertungsansatz, der sowohl die behördliche als auch die in Teil-AP 1.3.2 verhandelte gesellschaftliche Akzeptanz umfasst, angestrebt.

Die Ausarbeitung des Bewertungsschemas im Sinne einer Operationalisierung der Anforderungen für das in Fokus stehende Anwendungsfeld polizeilicher Internetauswertungen im Zusammenhang mit Hasskriminalität hat sich als komplex und herausfordernd erwiesen und bildete daher den Schwerpunkt der Arbeiten im Zeitraum der kostenneutralen Laufzeitverlängerung. Nach Durchführung eines testweisen Validierungsworkshops mit nicht in das Projekt eingebundene, aber fachlich nahe Wissenschaftler:innen zeigte sich, dass eine Validierung des

vorliegenden Leitfadens ein hohes Maß an Vorwissen zur polizeilichen Auswertepraxis erfordert, welches im Rahmen der Validierungsworkshops mit Stakeholdern aus der Zivilgesellschaft entsprechend zu vermitteln gewesen wäre. Aufgrund der knappen zeitlichen Ressourcen, die für die Durchführung einer Validierung zur Verfügung standen, wurde daher von der Durchführung von Workshops abgesehen. Im Fokus stand vielmehr die Rückkoppelung des entwickelten allgemeinen Bewertungsmodells mit Verbundpartner:innen sowie Akteur:innen, die sich mit der Umsetzung der Anforderungen aus der Europäischen KI-Verordnung beschäftigen.

Teil-AP 1.3.2: Gesellschaftliche Akzeptanz

Ziel war die Untersuchung der Akzeptabilität des Einsatzes von KI-Systemen in Sicherheitsbehörden im Allgemeinen und im speziellen Anwendungsfall durch professionelle Stakeholder und die Bevölkerung.

Arbeiten der TUB

- Entsprechend der mit dem Projektträger abgestimmten Änderung des Arbeitsplans wurde zur Analyse der Akzeptabilität des Einsatzes von KI-Systemen in Sicherheitsbehörden im Allgemeinen und im speziellen Anwendungsfall durch professionelle Stakeholder und die Bevölkerung auf Fokusgruppendaten aus dem BMBF-Projekt INTEGER zurückgegriffen. Diese wurden hinsichtlich der in KISTRA relevanten Fragestellungen zur Identifikation von „Risikowahrnehmungen, Einstellungen, Ängste sowie vorhandene Wert- und Normkonflikte sowie Forderungen und offene Fragen gegenüber dem Einsatz von KI-Systemen in Sicherheitsbehörden im Allgemeinen und zur Bekämpfung der Hasskriminalität im Speziellen“ ausgewertet.
- Bemerkenswert ist dabei in erster Linie, dass derzeit noch kein tiefgreifendes Verständnis von modernen KI-Systemen und deren Möglichkeiten und Grenzen in der Bevölkerung existiert. Vielmehr tendieren die Befragten dazu, den Systemen entweder eine dem Menschen vergleichbare Intelligenz zuzugestehen oder sie ihnen gänzlich abzusprechen. Eine sachstandsorientierte Bewertung seitens der Bevölkerung ist jedoch nicht möglich. Bewertet wurden in erster Linie Fiktionen und Gerüchte vom Einsatz von „intelligenter Technik“ bei der Polizei. Zwar mag sich dies in den kommenden Jahren durch die zunehmende Verbreitung von Large Language Modells (LLM) ändern, jedoch repräsentieren gerade diese ebenfalls nicht die Systeme, die seitens der Polizei derzeit in der Entwicklung sind. Inwiefern LLMs in den kommenden Jahren auch bei der Polizei die Basis für KI-Anwendungen darstellen wird, ist derzeit nicht abzusehen. Vor diesem Hintergrund ist entsprechend auch nicht davon auszugehen, dass erneut durchgeführte Fokusgruppen einen differenzierteren Kenntnisstand in der Bevölkerung angetroffen hätten. Aus den gewonnenen Erkenntnissen lässt sich mit Blick auf die Akzeptanz polizeilicher Internetauswertungen folgendes Ergebnis ableiten: Die Zustimmung zu polizeilicher Auswertungsmaßnahmen in sozialen Medien ist umso größer je stärker a) die Anlässe polizeilichen Handelns von den Teilnehmer:innen klassischen polizeilichen Aufgabengebieten

wie Strafverfolgung und Gefahrenabwehr zugeschrieben werden können und b) die Maßnahmen auf „extremistische“ Gruppierungen oder Personen begrenzt sind.

- Darüber hinaus konnten folgende Erkenntnisse zu Risikoszenarien und Anforderungen bzgl. des Einsatzes von KI zur Bekämpfung der Hasskriminalität aus Sicht der Bevölkerung abgeleitet werden:

Risikoszenario: Wer Daten im Netz öffentlich macht, muss damit rechnen, dass diese von der Polizei ausgewertet werden, auch wenn Sie keine Hassrede enthalten.

Anforderungen an die Polizei:

- Transparenz der Polizei bzgl. der Daten und Inhalte die gesichtet, gespeichert, ständig beobachtet werden. Zur Balancierung von Sicherheitsinteressen sollte diese mit einem zeitlichen Versatz von 2 Jahren aber mindestens bis zum Ende der Ermittlungen erfolgen. Um die informationelle Selbstbestimmung zu gewährleisten, sollten lediglich die Datenarten, der Umfang, die im Fokus der Ermittlungen stehenden Delikte und die Zahl der von Sichtung und Speicherung Betroffenen genannt werden.

- Transparenz hinsichtlich der Plattformen auf denen Quellen ausgewertet wurden. So soll an die Social-Media-Erfahrungen möglichst vieler Personen angeschlossen werden. Um die informationelle Selbstbestimmung zu gewährleisten, werden lediglich die Datenquellen (z.B. einschlägige Telegramgruppen, Webforen, Kommentarfunktionen von Zeitungen oder Medienportalen) benannt. So soll durch Transparenz eine Sensibilisierung der Bevölkerung erreicht werden.

Risikoszenario: Die Maßnahmen stellen einen zu starken Eingriff in die Rechte von Personen dar, die mit ihren Daten bzw. Inhalten im Netz nicht ausreichend vorsichtig sind.

Anforderungen an die Polizei:

- Transparenz hinsichtlich der Verfahren, die sicherstellen, dass Personen (z.B. Jugendliche) nicht versehentlich in den Fokus der Polizei geraten. Hier sind insbesondere der Umgang mit Falsch-Positiven und deren Daten, das Prinzip der Rechenschaftspflicht und Maßnahmen zur Vermeidung der Zweckentfremdung darzulegen.

Risikoszenario: Es besteht Sorge vor Überwachung willkürlich ausgewählter Personen gegenüber denen kein Anlass oder Anfangsverdacht vorliegt

Anforderungen an die Polizei:

- Transparenz in Bezug auf die notwendigen Bedingungen, die für einen Anlass vorliegen müssen, damit eine Person Gegenstand polizeilicher Internetauswertungen werden kann.

- Rechenschaftspflicht hinsichtlich der Überwachung und Kontrolle entsprechender Entscheidungen über die Durchführung bestimmter Auswertungen. Insbesondere müssen

diese Informationen leicht zugänglich und für die Bevölkerung die Möglichkeit zur Einspruchserhebung bestehen.

Risikoszenario: Es besteht Sorge, dass die Schwere der Anlässe, die zu einer Internetüberwachung führen, im Laufe der Zeit abnimmt und im schlimmsten Fall sogar Ordnungswidrigkeiten oder Normverletzungen zum Anlass erklärt werden.

Anforderungen an die Polizei:

- Es bedarf einer klaren rechtlichen Beschränkung des Einsatzes von KI in der Internetauswertung. Neben schweren Straftaten sollten dies vor allem internetspezifische Delikte sein. Um der Bevölkerung das Verständnis für die Schwellen zu erleichtern sollten Beispiele für entsprechende Anlässe genannt werden. Dabei sollte an Referenzereignisse angeschlossen werden, die den Teilnehmer:innen z.B. aus den Medien bekannt sind.

Risikoszenario: Es besteht die Sorge, dass bestimmte Bevölkerungsgruppen, insbesondere Minderheiten, häufiger aufgrund von gruppenspezifischen Merkmalen, die mit bestimmten Straftaten korrelieren von Internetauswertungsmaßnahmen betroffen sind.

Risikoszenario: Es besteht die Sorge, dass bestimmte Bevölkerungsgruppen seltener von Maßnahmen betroffen sind, weil die Polizei ihnen seltener Straftaten zutraut oder ihre Aussagen im Internet eher als Kritik denn als Hassrede interpretiert. Insbesondere im Bereich Rechtsextremismus wurden hier Bedenken geäußert.

Anforderungen an die Polizei:

- Es bedarf geeigneter Maßnahmen zur Gewährleistung der Vermeidung von Ungleichbehandlungen. Dies umfasst z.B. die Kontrolle der Trainings- und Validierungsdaten hinsichtlich Verzerrungen, die Kontrolle der vorgegebenen oder abgeleiteten Features und Featurekorrelationen hinsichtlich ihrer Deliktspezifität. Zudem sind die menschlichen Auswerter:innen darin zu unterstützen, aus den KI-Ergebnissen keine vorurteilsbehafteten Schlussfolgerungen abzuleiten.
- Zum Schutz Unbeteiligter sollten alle Ergebnisdatensätze hinsichtlich offengelegter personenbezogener Daten möglichst streng reglementiert sein. Entsprechend sollte eine De-Anonymisierung erst nach Prüfung der strafrechtlichen Relevanz erfolgen. Auch sollte die Einhaltung von Datenschutzvorschriften streng überwacht werden. Schließlich sollte der Einsatzbereich der Daten auf die festgelegten legitimen Zwecke beschränkt werden.

Risikoszenario: Es besteht die Sorge, dass ein Algorithmus mehr Fehler macht als ein:e Polizist:in mit Berufserfahrung und Augenmaß.

Anforderungen an die Polizei:

- Das Prinzip, dass ein Mensch stets die letzte Entscheidungsinstanz vor der Einleitung über die Datenauswertung hinausgehender freiheitseinschränkender Maßnahmen hat, muss stets eingehalten werden. Gegenüber der Bevölkerung muss dieses Prinzip transparent und proaktiv kommuniziert und die Rolle der Polizeipraktiker:innen als maßgebliche Entscheider:innen im Prozess erläutert werden.
- Die menschliche Autonomie der Auswertenden muss stets gewährleistet und unterstützt werden. So könnten Fälle z.B. zunächst ohne eine Erklärung der KI ausgegeben werden, um den Auswertenden einen unvoreingenommenen Blick auf den Fall zu ermöglichen. Erst nachdem diese eine Entscheidung getroffen haben, wird die Erklärung der KI angezeigt.

Risikoszenario: Es besteht die Sorge, dass es keine ausreichende Kontrolle der Polizei gibt. Wenn die Polizei sagt, dass eine Maßnahme sinnvoll ist, dann sehen die Befragten nicht, wer dies überprüft und im Zweifelsfall frühzeitig unterbindet.

Anforderungen an die Polizei:

- Es bedarf wie oben beschrieben eines wirkungsvollen Regimes der Rechenschaftspflicht. Neben auswerteprojektspezifischen Kontrollen gehört hierzu insbesondere eine frühzeitige Festlegung auf legitime Zwecke, die auf einer klaren Rechtsgrundlage beruhen.

Risikoszenario: Es besteht die Sorge, dass die Polizei angesichts mangelnder personeller Ressourcen nicht mit der Kontrolle der Auswertungsergebnisse hinterherkommt. Die Nutzung von modernen Technologien mag zwar in der Lage sein, mehr Straftaten oder Gefahren sichtbar zu machen, wenn gleichzeitig aber keine oder nur eine qualitativ minderwertige Auswertung stattfindet, kann der Einsatz neuer Technologien mehr schaden als helfen.

Anforderungen an die Polizei:

- Bei der Planung von Auswertungsprojekten muss die personale Kapazität stets berücksichtigt werden. Die in den Datenkorpus einbezogenen Daten müssen durch die Polizei mit hoher Qualität bearbeitbar sein. Es ist besser eine hochklassige Stichprobe zu bearbeiten, die eine generalpräventive Wirkung entfaltet, als eine pauschale Überwachung weiter Teile der sozialen Medien vorzunehmen, die nur zu einem Bruchteil ausgewertet werden können.

Risikoszenario: Es besteht die Sorge, dass moderne Auswertungswerkzeuge zu einem Überwachungsstaat führen.

Risikoszenario: Es besteht die Sorge, dass mit den Auswertewerkzeugen alle unter Generalverdacht gestellt werden.

Risikoszenario: Es besteht die Sorge, dass Personen sich nicht länger trauen, ihre politische Meinung im Internet kundzutun.

Anforderungen an die Polizei:

- Teil des Rechenschaftsregimes muss die Etablierung demokratischer Kontrollinstanzen sein, die alle Kontrollgremien regelmäßig hinsichtlich ihrer rechtlichen und ethischen Konformität überwacht.
- Es bedarf einer hohen Transparenz hinsichtlich der Legitimität der Zwecke. Es muss stets deutlich werden, auf welcher Rechtsgrundlage die einzelnen Auswerteprojekte beruhen.
- Die Ergebnisse der Verhältnismäßigkeitsprüfungen hinsichtlich der Geeignetheit, der Erforderlichkeit und der Angemessenheit der Maßnahmen sollen nach Möglichkeit offengelegt werden. Daraus muss insbesondere hervorgehen, welchen Nutzen die Polizei annimmt, welche möglichen Folgen berücksichtigt und wie diese zum Nutzen ins Verhältnis gesetzt wurden.

Risikoszenario: Es besteht die Sorge, dass der Schutz personenbezogener Daten durch moderne Auswertungsmöglichkeiten umgangen werden kann.

Anforderungen an die Polizei:

- Es ist transparent gegenüber der Öffentlichkeit darzustellen, unter welchen Bedingungen personenbezogene Daten ausgewertet werden dürfen, bzw. wie die Polizei mit anderen Daten umgeht, die einen Personenbezug indirekt ermöglichen. Moderne KI-Systeme sind besonders geeignet, aus pseudonymisierten Daten (z.B. unter Pseudonym geposteten Inhalten) einen Personenbezug abzuleiten. Hier bedarf es einer gesetzlichen Regulierung unter welchen Umständen z.B. bei schweren Straftaten dies legitim sein kann.

Risikoszenario: Es besteht die Sorge, dass die derzeitigen demokratischen Parteien eine hinreichende Kontrolle der Polizei zwar noch gewährleisten können, dass bei weniger demokratische Parteien in der Zukunft diese Werkzeuge auch gegen politische Gegner:innen verwenden.

Anforderungen an die Polizei:

- Der Datenschutz muss in hohem Maße gesichert und gewährleistet werden. Insbesondere die Speicherung von Daten und die Einhaltung von Löschfristen muss streng überwacht werden. Dies kann einen Missbrauch der Technik zwar nicht verhindern, aber zumindest eine rückwirkende Datenauswertung erschweren.

Die genannten Risikowahrnehmungen machen in erster Linie deutlich, dass eine stärkere Transparenz und Aufklärung gegenüber der Bevölkerung und ein rigides Regime der Rechenschaftspflicht erforderlich sind. So sollte die Polizei beim Einsatz von KI-Systemen proaktiv und transparent über die Verwendung entsprechender Systeme kommunizieren und dabei auch die grundlegende Funktionsweise der Systeme offenlegen. Gleichwohl ist beim Einsatz von KI

in der Sicherheitsproduktion stets eine Balance zwischen Transparenz und notwendiger Geheimhaltung der genauen Funktionsweise und insbesondere der Grenzen der Systeme zu finden, um potentiellen Straftäter:innen die Vermeidung von Strafverfolgung zu erschweren. Damit diese Form der Geheimhaltung dennoch nicht zu einem Gefühl unkontrollierter Überwachung führt, sind im Sinne der oben dargestellten Organisation von Rechenschaftspflicht, regelmäßige Berichte zu veröffentlichen.

Abänderungen gegenüber der Planung

Um den Mehraufwand in Teil-AP 1.2 zu kompensieren, wurden mit Zustimmung des Projektträgers Personalressourcen im Umfang von 4 Personenmonaten von Teil-AP 1.3.2 nach 1.2 verschoben. Die Erkenntnisziele von Teil-AP 1.3.2 würden aus Sicht des Gesamtprojekts weiterhin erreicht. In Teil-AP 1.3.2 waren ursprünglich 8 Fokusgruppen mit Personen aus unterschiedlichen Bevölkerungsgruppen geplant, mit dem Ziel, die „Risikowahrnehmungen, Einstellungen, Ängste sowie vorhandene Wert- und Normkonflikte sowie Forderungen und offene Fragen gegenüber dem Einsatz von KI-Systemen in Sicherheitsbehörden im Allgemeinen und zur Bekämpfung der Hasskriminalität im Speziellen“ zu untersuchen und in einem Bericht zur Risikowahrnehmung der Bevölkerung zusammenzufassen.

Das Ziel die Risikowahrnehmung in Bezug auf den Einsatz von KI-Systemen in Sicherheitsbehörden im Allgemeinen zu untersuchen konnte erreicht werden, indem auf Fokusgruppendaten aus dem bereits abgeschlossenen BMBF-Projekt INTEGER zurückgegriffen wurde. Dort wurden bereits Fokusgruppen mit Personen aus unterschiedlichen Bevölkerungsgruppen zur Frage der Wahrnehmung und Risikobewertung polizeilicher Internetauswertung durchgeführt.

Teil-AP 1.3.3: Risikokommunikation

Ziel war die Erarbeitung einer Risikokommunikationsstrategie zur Adressierung der Risikowahrnehmungen der professionellen Stakeholder und der Ängste und Erwartungen in der Bevölkerung in Bezug auf den Einsatz von KI-Systemen in Sicherheitsbehörden zur Bekämpfung von Hasskriminalität sowie im Allgemeinen.

Es wurde eine Literaturrecherche zum Stand der Forschung in der Risikokommunikation durchgeführt, um allgemeine Kriterien guter Risikokommunikation zu entwickeln. Auf dieser Grundlage sowie auf Grundlage der Ergebnisse aus Teil-AP 1.3.2 wurde nachfolgender Leitfaden guter Risikokommunikation entwickelt:

I. Leitfaden zur Risikokommunikation

Die Bedeutung der Risikokommunikation hat in den letzten Jahren, insbesondere bei der Implementierung neuer Technologien, deutlich zugenommen. Der Einsatz von Künstlicher Intelligenz (KI) in der Sicherheitsproduktion ist ein besonders relevantes Thema, da es weitreichende Konsequenzen für die Gesellschaft hat. Die Planung und Umsetzung von Projekten zur Integration von KI in sicherheitsrelevante Bereiche stößt häufig auf Bedenken und Wider-

stände seitens der Bevölkerung. Um solche Projekte erfolgreich umzusetzen, ist eine frühzeitige und effektive Risikokommunikation unerlässlich. Risikokommunikation wird als der zielgerichtete Austausch von Informationen über die möglichen Auswirkungen von Ereignissen, Handlungen oder Technologien auf die menschliche Gesundheit und die gesellschaftliche Sicherheit verstanden (VDI, 2000).

Vertrauen in die Kommunikator:innen, also diejenigen Personen oder Gruppen, die risikorelevante Botschaften kommunizieren, ist ein wesentlicher Faktor für den Erfolg der Risikokommunikation (Carius und Renn, 2003). Ohne ein solches Vertrauensverhältnis ist ein dialogischer Austausch über differierende Risikowahrnehmungen und -bewertungen kaum möglich (VDI, 2000). In der Bevölkerung gibt es jedoch nur selten eine einheitliche Risikowahrnehmung. Vielmehr besteht das Problem, dass die Risiko- und Gefahrenpotentiale, die von den einzelnen Bürger:innen aber auch Expert:innen kaum noch beurteilt werden können. Die Komplexität der zugrundeliegenden Wirkungsgefüge führt zu kaum abschätzbaren Nebenfolgen und zu einer entsprechenden Unübersichtlichkeit bzw. Unsicherheit Entscheidungen betreffend. Mussten in den vergangenen Jahren Staat und Wissenschaft das „Monopol“ ihrer Deutungshoheit, Risiken gegenüber der Bevölkerung zu erklären und zu kommunizieren, einbüßen (Beck 1986), haben eine Vielzahl neuer Akteur:innen wie Interessen- und Aktivist:innengruppen die Bühne betreten, um die gesellschaftliche Risikowahrnehmung zu beeinflussen (Hempel/Lammerant 2014). In diesem Zusammenhang erscheint dann die Bevölkerung nicht als monolithischer, homogener Block, sondern muss selbst differenziert betrachtet werden. Es ist in der Literatur üblich, diese grob in drei Gruppen zu unterteilen: Technologiebefürworter:innen, Technologiegegner:innen und die schweigende Mehrheit. Dabei stellen die ersten beiden Gruppen gleichsam die Extreme zur zunächst indifferenten bzw. schweigenden Mehrheit dar.

Die Technologiebefürworter sind durch hohe **Risikotoleranz** gekennzeichnet. Die Mitglieder dieser Gruppe betrachten Innovationen in der Regel vom Nutzen her und folgen insofern häufig einer ökonomischen Perspektive. Dieser Gruppe diametral entgegengesetzt zeichnen sich die Mitglieder der Technologiegegner:innen sodann entsprechend durch **Risikoaversion** aus. Spielt auch hier Nutzen eine Rolle, so erfolgt die Ablehnung konkreter Vorhaben aber häufig auf Grundlage bestimmter tradierter Werturteile und sozialer Normen, die die Wahrnehmung des Risikos prägen. So kann beispielsweise „eine mehr oder minder radikale Ablehnung des industriellen Projekts der Moderne und seinen wahrgenommenen Auswüchsen: Kapitalismus, Ellenbogengesellschaft, Wachstumslogik und industrielle ‘Mega-dreads‘“ im Vordergrund stehen (Renn und Zwick 1997: 3f.). Die dritte Gruppe bildet schließlich die „schweigende Mehrheit“. Dieser ist „eine fundamentalistische Technik- oder Risikogegnerschaft“ ebenso fremd „wie eine besondere Vorliebe für Technik“. Entsprechend können aber Angehörige dieser Gruppe zu „Betroffenheitsaktivisten“ werden, beispielsweise wenn in ihrem sozialen Umfeld Vorhaben geplant werden, die auf die eine oder andere Weise in ihren Alltag eingreifen (ebd.).

Für die Konstituierung von Risikowahrnehmung spielen fakten-, institutionen-, und wertbezogene Faktoren eine Rolle, deren Ausprägung in verschiedenen Bevölkerungsgruppen aber stark differieren kann.

So spielt für die Konstitution der Risikowahrnehmung die **persönliche Betroffenheitswahrnehmung** eine zentrale Rolle. Der Verweis auf die Wahrnehmung ist dabei entscheidend, denn in der Praxis zeigt sich immer wieder, dass nicht die faktische, sondern in erster Linie die gefühlte Betroffenheit für die Risikowahrnehmung entscheidend ist. Gerade im Kontext des Einsatzes von KI zur Erkennung von Hasskriminalität ist dies für den überwiegenden Teil der Bevölkerung derzeit der Fall. Selbst wenn eine Auswertung durch die Polizei erfolgt, werden dies die wenigsten Betroffenen jemals erfahren. Umgekehrt gilt, dass Personen, die sich wissentlich im Umfeld von radikaler oder hasserfüllter Kommunikation bewegen, eher eine negative Betroffenheit annehmen und das Risiko negativer Folgen für sich oder die Gruppe in der sie sich bewegen also wesentlich höher einschätzen als Personengruppen, die sich nicht in entsprechenden Gruppen bewegen.

Die subjektive Gefahreneinschätzung spielt insofern eine zentrale Rolle, als sie weniger durch harte Fakten sondern durch **Faktengeschichten**, also medial vermittelten Bildern und Narrative, Meinungen von Verwandten, Freunden oder Kolleg:innen geprägt ist, die in vielen Fällen nicht auf wissenschaftlichen Erkenntnissen beruhen. Daher verwundert es auch nicht, dass gerade die Bewertung von KI derzeit noch mehr auf Science fiction als auf science beruht. Dies erklärt auch die geringe Spezifität der oben (siehe Teil-AP 1.3.2) beschriebenen Risikoszenarien der Bevölkerung. Die mehr oder weniger gerechtfertigten Annahmen und Faktengeschichten bleiben aber meist nicht auf das engste soziale Umfeld beschränkt, sondern werden von strategischen Akteur:innen genutzt, um in den (sozialen) Medien für oder gegen den Einsatz von KI zu mobilisieren.

Erweist sich Risikowahrnehmung also deutlich als kontextabhängig, spielen Fragen der **institutionellen Vermittlung und Kommunikation** von Risiken durch Behörden dennoch eine wichtige Rolle und können die Risikowahrnehmung entscheidend beeinflussen. Die Bilder und Botschaften, die institutionelle Akteure vermitteln, haben einen entscheidenden Einfluss auf die Wahrnehmung der Bevölkerung. Gewinnen Bürger:innen den Eindruck, dass sie nicht hinreichend informiert werden, dass ihnen Informationen fehlen oder gar bewusst vorenthalten werden, beginnen sie Risiken höher einzuschätzen. Behörden werden dann als unfähig wahrgenommen, sich um den Schutz und die Sicherheit der Bevölkerung adäquat zu kümmern. Gerade für peripher betroffene Bürger:innen erweisen sich Erfahrungen und Geschichten über Inkompetenz und Manipulationsversuche der zuständigen Behörden als wahrnehmungsprägend (Zimmer et al. 2013: 49). Es wird dann häufig nur noch graduell zwischen dem Fehlverhalten einzelner Behörden und allen anderen unterschieden (Renn und Zwick 1997). In der Folge kann dann nicht nur das Vertrauen in die Polizei, sondern darüber hinaus auch in den Rechtsstaat und die Demokratie insgesamt sinken.

Schließlich sind es **Werte- und Nutzenfragen**, die für die kontextspezifische Risikowahrnehmung entscheidend sind. Die Bürger:innen fragen sich beispielsweise „warum es notwendig und gerechtfertigt ist, dass sie einem Risiko ausgesetzt werden“, „welchen Mehrwert es für sie bringt, wenn sie sich dem wahrgenommenen Risiko aussetzen“ und „welche Prioritäten bei der Planung und Umsetzung von entsprechenden Maßnahmen angelegt wurden“. Wurden

dabei nur einseitig behördliche Interessen berücksichtigt, oder auch ethische Aspekte wie Freiheit, Gleichheit und Kontrolle? (Renn und Levine 1991). Wo ethische Aspekte nicht wahr und ernst genommen werden, steigt das Frustrationspotential. Es entsteht die Gefahr der „inneren Kündigung“ und der kommunikativen Totalverweigerung. Entsprechende Personengruppen sind für rationale Argumentationen dann so gut wie nicht mehr erreichbar. Jeder Versuch beispielsweise, wissenschaftlich zu argumentieren, wird mit dem Verweis abgelehnt, dass sich für alle Expert:innen Gegenexpert:innen finden lassen oder die Wissenschaftler:innen schlicht gekauft seien (Renn und Zwick 1997). Rechtliche Schranken verlieren ihre legitimierende Wirkung, weil davon ausgegangen wird, dass durch sie „Gefahren“ künstlich heruntergespielt werden, statt die Bevölkerung zu schützen.

Stellen Faktengeschichten, institutionelle Kommunikation und die Berücksichtigung von Wert- und Nutzenfragen zentrale Rahmenbedingungen gelingender Risikokommunikation dar, werfen sie gleichzeitig die Frage nach der konkreten Gestaltung gelingender Risikokommunikationsmaßnahmen auf. Gegenstand von Risikokommunikation ist die Information und der Dialog über Risikowahrnehmungen und -bewertungen sowie hiermit verbundene Konflikte. Die wichtigsten Werkzeuge sind hierfür die Schaffung von Vertrauen und Glaubwürdigkeit, die Informationsdarstellung sowie eine gelungene Zwei-Wege-Kommunikation zwischen Kommunikator:innen und der Bevölkerung.

Vertrauen kann auf verschiedenen Ebenen hergestellt werden, neben der kognitiven und der habituellen spielt auch die emotionale Ebene eine entscheidende Rolle (Endreß 2005). Vertrauen basiert dabei auf vier Grundprinzipien, derer sich alle Kommunikationsmaßnahmen und Kommunikator/innen verpflichtet fühlen sollten. Dies sind **Transparenz** (Kurzenhäuser et al. 2010; Wiedemann et al. 2000; Renn 2004, 2005; Carius und Renn 2003; Epp et al. 2008) im Umgang mit Informationen und Plänen, **Fairness** (Wiedemann et al. 2000; Renn 2004, 2005) gegenüber anderen Meinungen, **Verlässlichkeit** (Kurzenhäuser et al. 2010; Wiedemann et al. 2000; Renn und Levine 1991) in Bezug auf Argumentationen, Erwartungen und Verpflichtungen und **Offenheit** (Wiedemann et al. 2000; Renn 2005; Carius und Renn 2003) für die Anliegen der Bevölkerung. Zielen Maßnahmen auf die persönliche Ansprache der Bürger:innen ab, bedeutet dies, dass diese vier Prinzipien in besonderer Weise für jede Komponente einer Kommunikationsstrategie berücksichtigt und entsprechend verinnerlicht werden müssen. Folgende Komponenten lassen sich unterscheiden:

1. **Kommunikator:in:** In der Literatur werden eine Vielzahl von Empfehlungen ausgesprochen, wie Kommunikator:innen sich praktisch verhalten sollten (Renn 2005; Renn und Levine 1991):
    - Die Kommunikator:innen sollten dem Publikum die Möglichkeit geben, seine Erfahrungen und Überzeugungen darzustellen und einzubringen. Sie sollten zuhören und ernst nehmen. Sie sollten als Persönlichkeiten und nicht einfach nur als Sprecher:innen einer Institution auftreten.

- Kommunikator:innen sollten des Weiteren versuchen eine Sprache zu sprechen, die das Gegenüber versteht. Dazu sollte auf allgemein bekannte Symbole und ansprechende Formate zurückgegriffen werden und das Publikum sollte durch Offenheit und Ehrlichkeit überrascht werden. Daten, komplexe Sachverhalte oder Wahrscheinlichkeiten sollten an Alltagsbeispielen erläutert werden und es sollte vermieden werden, negativ über andere zu sprechen.
- Jede Position sollte entsprechend an den Erfahrungen und Problemen der Zuhörer:innen anschließen. Das bedeutet auch, dass man in der interpersonalen Kommunikation nicht nur auf Fakten, Rationalitäten oder Argumentationen, sondern auch auf Mitgefühl und Empathie setzen sollte.
- Darüber hinaus lässt sich Transparenz am besten dadurch vermitteln, dass man bei der Kommunikation von Fakten und Entscheidungen immer auch auf das Zustandekommen derselben hinweist und Zugänge zu vertiefenden Informationen bereithält.
- Um schließlich dem Prinzip der Fairness gerecht zu werden, sollte eigeninitiativ darauf hingewiesen werden, wenn es zu bestimmten Aspekten oder Einschätzungen auch abweichende Meinungen oder Interessenkonflikte gibt. Idealerweise lässt sich auch erklären, wie diese zustande kommen, ohne die Repräsentant:innen dieser Positionen persönlich zu diskreditieren.

2. Gestaltung der Kommunikationssituation: Kommunikationssituationen sind vielgestaltig und ändern sich je nach gesellschaftlichem Anlass und kommunikativem Kontext. Dabei spielen zahlreiche Dimensionen von der zeitlichen, räumlichen bis hin zur Objekt- und Mediendimension eine maßgebliche Rolle (Ziemann 2013). Im Hinblick auf die Risikokommunikation empfiehlt es sich grundsätzlich, dass sämtliche Maßnahmen einen interaktiven Charakter haben. Unabhängig davon, ob es sich um eine interpersonale Face-to-face-Kommunikation handelt oder massenmediale Kommunikationskanäle wie z.B. Kommunikation über Websites oder in sozialen Medien genutzt werden, entscheidend ist, dass durch die Gestaltung der Kommunikationssituation Vertrauen aufgebaut wird. Folgende Aspekte sind hierbei von Relevanz:

- Der Aufbau einer funktionierenden **Zwei-Wege-Kommunikation** (Ulbig et al. 2010; Wiedemann und Clauberg 2005; Scheer et al. 2010) zeichnet sich einerseits dadurch aus, dass den angesprochenen Rezipient/innen entweder direkt oder durch Angabe einer Kontaktmöglichkeit ein Weg aufgezeigt wird, wie sie ihre Anliegen gegenüber der kommunizierenden Institution vortragen können. Das Kommunikationsangebot wird aber erst dann zu einer echten Zwei-Wege-Kommunikation, wenn die Kommunikator:innen in der Lage sind, auf die Fragen und Einwände mit einer individuellen Antwort zu reagieren. Es muss also auf beiden Seiten Sprachfähigkeit hergestellt und garantiert werden. Diese stellt die Voraussetzung für eine Kommunikation auf Augenhöhe dar.
- Die **Gestaltung der interpersonalen Kommunikationssituationen** (Johnson et al. 1995; Rogers 1983) erfordert, die Anliegen der Bürger:innen zu hören als auch auf diese eine individuelle Antwort zu geben. Neben der persönlichen Kompetenz, der Empathie und

dem Charisma der Kommunikator:innen ergibt sich das größte Potential interpersonaler Kommunikation aus der Tatsache, dass Menschen grundsätzlich leichter zu überzeugen sind, wenn sie in einer Kommunikationssituation ihre:n Gesprächspartner:in als präsent, einfühlend, warmherzig und aufmerksam wahrnehmen. Interpersonale Gesprächssituationen erleichtern die Erklärung von besonders komplexen Sachverhalten und damit die Überzeugung einzelner Personen oder kleinen Gruppen, da auf Nachfragen direkt reagiert werden kann.

- Im Falle **massenmedialer** Kommunikation muss die Organisation entsprechende Strukturen schaffen, um in der Lage zu sein, eingehende Anfragen individuell bearbeiten zu können. Hierzu gehört beispielsweise eine Hotline oder ein Ticketsystem, die mit dem Versprechen verbunden sind, zeitnah zu jeder Anfrage eine Antwort zu liefern.

3. Darstellung von Information: Wissensvermittlung stellt einen notwendigen Eckpfeiler einer gelungenen Risikokommunikation dar. Informationen zu neuen Technologien bzw. Auswerteprojekten müssen entsprechend aufbereitet und dargestellt werden, um zielgruppengerecht kommuniziert werden zu können. Auf der inhaltlichen Ebene sollten deshalb Aussagen zu den drei zentralen Aspekte der Risikokommunikation enthalten sein (Renn und Levine 1991; Wiedemann et al. 2000):

- Erstens sollten **Wert- und Nützlichkeitsfragen** adressiert werden. Dies kann zum Beispiel ein Hinweis darauf sein, dass eine Kommunikationsmaßnahme durchgeführt wird, um der Bevölkerung eine Stimme zu geben, ihr zuzuhören und ihre Belange ernst zu nehmen. Darüber hinaus kann der allgemeine oder spezifische Nutzen adressiert werden. Hier könnte zum Beispiel darauf hingewiesen werden, welche Sicherheitsgewinne durch den Einsatz von KI zu erwarten sind und wie Werte wie Transparenz und Rechtskonformität umgesetzt werden.
- Zweitens sollte die Kommunikation die **institutionelle Ebene** adressieren und beschreiben, wie welche Strukturen zur Kontrolle und Überwachung des Einsatzes von KI bei der Polizei implementiert wurden, bzw. welche Formen demokratischer Kontrolle etabliert wurden, um die Bevölkerung vor übermäßigen Eingriffen in ihre Freiheitsrechte oder Ungleichbehandlung zu schützen.
- Drittens sollte die **Sachverhaltsebene** angesprochen werden. Informationen und Zusammenhänge bzgl. des KI Einsatzes sollten leicht verständlich sein, und es sollte erklärt werden, wie und durch wen die Informationen zusammengetragen und deren Richtigkeit überprüft wurde. Um die Bürger:innen darin zu unterstützen, Zusammenhänge besser zu verstehen, sollte auf Alltagsbeispiele und Vergleiche zurückgegriffen und vertiefende Informationen zur Verfügung gestellt werden können (Renn 2005; Jungermann 1991).

II. Entwicklung von Kommunikationsmaßnahmen

Auf Basis dieses Leitfadens wurden für die Risikowahrnehmungen der Bevölkerung passende Kommunikationsmaßnahmen entwickelt. Die oben dargestellten Risikoszenarien wurden zu diesem Zweck noch einmal zu sechs Themenfeldern zusammengefasst.

Datenschutz und Privatsphäre: Nutzer:innen befürchten, dass ihre Daten ohne ausreichenden Grund überwacht und ausgewertet werden, selbst wenn sie keine Hassrede enthalten.

Kommunikationsmaßnahmen:

- **Wert- und Nützlichkeitsfragen:** Es sollte hervorgehoben werden, dass der Schutz der Privatsphäre ein zentraler Wert ist und dass alle Maßnahmen im Einklang mit den geltenden Datenschutzgesetzen stehen. Die Bürger:innen sollten informiert werden, dass ihre Belange ernst genommen werden und ihre Daten nur unter strikten gesetzlichen Vorgaben ausgewertet werden.
- **Institutionelle Ebene:** Die institutionelle Ebene sollte transparent gemacht werden, indem beschrieben wird, welche Strukturen zur Kontrolle und Überwachung des KI-Einsatzes im Sinne des Rechenschaftspflichtmodells implementiert wurden. Unabhängige Kontrollinstanzen und regelmäßige Audits sollten betont werden, um zu zeigen, dass die Bevölkerung vor Missbrauch geschützt ist.
- **Sachverhaltsebene:** Auf der Sachverhaltsebene sollten leicht verständliche Erklärungen über die Funktionsweise der eingesetzten KI-Systeme gegeben werden. Es sollten Beispiele aus dem Alltag verwendet werden, um die Vorteile und die Sicherheitsgewinne durch den KI-Einsatz zu verdeutlichen.

Eingriff in Persönlichkeitsrechte: Es wird befürchtet, dass Maßnahmen der KI-gestützten Sicherheitsproduktion einen zu starken Eingriff in die Rechte von Personen darstellen.

Kommunikationsmaßnahmen:

- **Wert- und Nützlichkeitsfragen:** Es sollte betont werden, dass der Einsatz von KI dazu dient, die allgemeine Sicherheit zu erhöhen und dass alle Maßnahmen darauf abzielen, die Persönlichkeitsrechte der Bürger:innen zu respektieren und zu schützen.
- **Institutionelle Ebene:** Die institutionelle Ebene sollte aufzeigen, wie demokratische Kontrollmechanismen implementiert sind, um sicherzustellen, dass keine übermäßigen Eingriffe in die Freiheitsrechte der Bürger:innen erfolgen bzw. jeder über die Datenauswertung hinausgehende Eingriff ausschließlich nach menschlicher Bewertung und Entscheidung erfolgt. Beispiele für solche Mechanismen sind die Existenz unabhängiger Aufsichtsstellen (siehe in unserem Modell die staatliche Prüfstelle), innerbehördliche Reflektionsmaßnahmen und Maßnahmen zur Qualitätssicherung ((siehe in unserem Modell die inerne Prüfstelle) sowie gerade bei innovativen Maßnahmen unabhängige wissenschaftliche Evaluationen.

- **Sachverhaltsebene:** Auf der Sachverhaltsebene sollten die technologischen und rechtlichen Grundlagen des KI-Einsatzes erläutert werden. Dabei sollte erklärt werden, wie die Daten gesammelt und verarbeitet werden und welche Maßnahmen getroffen werden, um Missbrauch zu verhindern.

Willkürliche Überwachung: Es besteht Sorge vor einer Überwachung willkürlich ausgewählter Personen ohne Anlass oder Anfangsverdacht.

Kommunikationsmaßnahmen:

- **Wert- und Nützlichkeitsfragen:** Es sollte kommuniziert werden, dass der Einsatz von KI auf klar definierten und transparenten Kriterien basiert, die willkürliche Überwachung ausschließen. Die Bürger:innen sollen wissen, dass ihre Rechte gewahrt bleiben und die Maßnahmen auf einem nachvollziehbaren rechtlichen Rahmen beruhen. Zudem soll den Bürger:innen der Sinn von Internetauswertungen nachvollziehbar erklärt werden. Welche Personen oder Gruppen werden aufgrund welcher Ereignisse mit welchem Ziel und für wie lange im Internet beobachtet? Welche Abbruchkriterien gibt es für die Überwachung? und wie und in welchen Intervallen werden diese evaluiert?
- **Institutionelle Ebene:** Auf der institutionellen Ebene sollte dargestellt werden, welche internen und externen Kontrollen existieren, um sicherzustellen, dass die Auswertungsmaßnahmen gerecht und rechtmäßig sind. Dazu sollten die Prozeduren und Kriterien regelmäßiger Überprüfung beschrieben und erklärt werden, welche unabhängigen Gremien die Einhaltung dieser Überprüfung gewährleisten.
- **Sachverhaltsebene:** Es sollten verständliche Informationen über die Algorithmen und Entscheidungsprozesse bereitgestellt werden, die bei der Überwachung zum Einsatz kommen. Beispiele aus der Praxis können helfen, das Vertrauen der Bürger:innen zu gewinnen.

Diskriminierung bestimmter Gruppen: Nutzer:innen befürchten, dass bestimmte Bevölkerungsgruppen häufiger überwacht werden, während andere aufgrund von Vorurteilen seltener betroffen sind.

Kommunikationsmaßnahmen:

- **Wert- und Nützlichkeitsfragen:** Es sollte betont werden, dass alle Auswertungen auf der Grundlage von Gleichbehandlung und Gerechtigkeit durchgeführt werden. Die Bürger:innen sollen wissen, dass die KI-Systeme darauf ausgelegt sind, Diskriminierung zu vermeiden und alle Personen gleich zu behandeln.
- **Institutionelle Ebene:** Auf der institutionellen Ebene sollte beschrieben werden, welche technischen und organisationalen Maßnahmen implementiert wurden, um zum einen sicherzustellen, dass in den Trainings- und Validierungsdaten keine Verzerrun-

gen enthalten sind und auch in die Modellen keine verzerrenden Annahmen eingeflossen sind. Aber auch hinsichtlich der polizeilichen Praxis sollte verdeutlicht werden, welche institutionellen Maßnahmen wie Schulungen, Reflektionsgespräche, Supervision, etc. getroffen wurden, um diskriminierungsfreie Bewertungen und Maßnahmenentscheidungen zu gewährleisten.

- **Sachverhaltsebene:** Es sollten klare und verständliche Informationen über die Funktionsweise der KI und die getroffenen Maßnahmen zur Verhinderung von Diskriminierung bereitgestellt werden. Dabei sollten die Bürger:innen die Möglichkeit haben, sich über technische, regulatorische und prozedurale Verfahren zur Vermeidung von Verzerrungen in den Daten oder den KI-Ergebnissen sowie über Maßnahmen zur Gewährleistung von Diskriminierungsfreiheit polizeilicher Maßnahmen insgesamt zu informieren. Alltagsbeispiele und konkrete Fallstudien können dabei helfen, die Erklärungen zu verdeutlichen.

Fehleranfälligkeit der KI: Es wird angenommen, dass Algorithmen mehr Fehler machen als erfahrene Polizist:innen.

Kommunikationsmaßnahmen:

- **Wert- und Nützlichkeitsfragen:** Es sollte vermittelt werden, dass die Einführung von KI-Systemen nicht dazu dient, menschliche Polizist:innen zu ersetzen, sondern sie zu unterstützen und ihre Arbeit zu erleichtern. Die Vorteile der KI, wie die Fähigkeit große Datenmengen schnell zu analysieren, sollten hervorgehoben werden.
- **Institutionelle Ebene:** Auf der institutionellen Ebene sollte betont werden, dass es strenge Qualitätskontrollen und kontinuierliche Verbesserungsprozesse gibt, um die Genauigkeit sowie Einhaltung ethischer Prinzipien durch die KI-Systeme zu gewährleisten. Es sollte zudem aufgezeigt werden, wie menschliche Expertise und technologische Lösungen zusammenarbeiten, um die Vorteile der menschlichen und der technischen Beiträge für eine effektive Sicherheitsproduktion zu verdeutlichen.
- **Sachverhaltsebene:** Es sollte erklärt werden, warum es unvermeidlich ist, dass KI-Systeme Fehler machen und welchen Mehrwert dies mit sich bringt. Es sollten Beispiele aus der Praxis bereitgestellt werden, die zeigen, wie KI-Systeme in Kombination mit menschlicher Aufsicht erfolgreich eingesetzt werden. Die Bürger:innen sollten über die Verfahren zur Fehlererkennung und -korrektur informiert werden.

Fehlende Kontrolle: Es besteht die Sorge, dass es keine ausreichende Kontrolle der Polizeiarbeit mit KI gibt.

Kommunikationsmaßnahmen:

- **Wert- und Nützlichkeitsfragen:** Es sollte klargestellt werden, dass der Einsatz von KI-Systemen strengen gesetzlichen und ethischen Vorgaben unterliegt. Die Bürger:innen

sollen wissen, dass es klare Regeln und Verantwortlichkeiten gibt, um den Missbrauch der Technologie zu verhindern.

- **Institutionelle Ebene:** Auf der institutionellen Ebene sollten die verschiedenen oben beschrieben im Rahmen der Rechenschaftspflicht verankerten Kontroll- und Überwachungsmechanismen detailliert beschrieben werden. Zudem sollte auf die Unabhängigkeit und demokratische Legitimation der entsprechenden Überwachungsorgane, und Kontrollprozesse eingegangen werden.
- **Sachverhaltsebene:** Es sollten verständliche und zugängliche Informationen über die Kontrollprozesse und -strukturen bereitgestellt werden. Beispiele und Vergleiche mit der Kontrollpraxis anderer kritischer Technologien wie z.B. Atomenergie, Massenvernichtungswaffen, etc. können helfen, die Wirksamkeit von Kontrollmaßnahmen zu veranschaulichen.

Abänderungen gegenüber der Planung

Aufgrund der Verzögerungen im Zeitplan des Gesamtprojektes, konnten zum Projektende das vorgesehenen Dialogformat und die Fokusgruppen zur Evaluation der Kommunikationsstrategie nicht mehr durchgeführt werden. So lag der Schwerpunkt der Arbeiten im Zeitraum der kostenneutralen Verlängerung auf der Weiterentwicklung der gesellschaftlich-ethischen Kriterien/Anforderungen, der Einarbeitung von Anforderungen aus der Europäischen KI-Verordnung, dem Austausch mit staatlichen Stellen, die sich mit der Umsetzung der KI-Verordnung befassen sowie Präsentationen des entwickelten Bewertungsschemas auf Tagungen und Konferenzen.

II.2 Wichtigste Positionen des zahlenmäßigen Nachweises

Position	Entstandene Ausgaben in €	Erläuterungen
0812 (wiss. Mitarbeitende)	204.801,15	Wiss. Mitarbeiter (ca. 75%)
0822 (stud. Mitarbeitende)	18.610,17	Stud. Mitarbeitende 40h/mtl.
0843 (Sonstige Kosten)	3.703,64	Catering und Reisekosten für Referent:innen für Konferenz „Künstliche Intelligenz in der Kriminalitätskontrolle: Gesellschaftliche und rechtliche Dimensionen“ am 5. Mai 2023 an der TUB
0846 (Dienstreisen)	4.762,15	Verbundtreffen in München, Konferenzen in München und im Harz
Summe	231.877,11	

II.3 Notwendigkeit und Angemessenheit der geleisteten Arbeit

Die Arbeitsschritte entsprachen im Wesentlichen den in dem Antrag zum Teilvorhaben skizzierten Arbeitsschritten und Arbeitspaketen. Sie waren sowohl in diesem Sinne des Arbeitsplans notwendig und angemessen als auch im Sinne der Sichtbarmachung des Vorhabens für Öffentlichkeit und Fachpublikum.

II.4 Voraussichtlicher Nutzen, insbesondere Verwertbarkeit des Ergebnisses im Sinne des fortgeschriebenen Verwertungsplans

Es bestehen hohe kurz- und mittelfristige wissenschaftliche Erfolgsaussichten, da die Projektergebnisse in einen Zeitraum fallen, in dem Politik und Praxis aufgefordert sind, einen regulatorischen Rahmen für den Einsatz von KI-Lösungen in Sicherheitsbehörden zu definieren, der die Anforderungen aus der europäischen KI-Verordnung erfüllt. Das entwickelte Bewertungsschema und der entwickelte Leitfaden für den Einsatz von KI bei Internetauswertungen im polizeilichen Staatsschutz sollen im ersten Jahr nach Projektabschluss in Form eines praxisorientierten Gestaltungs- und Bewertungsleitfadens der Praxis und Politik zur Verfügung gestellt werden. Zudem sollen Vorschlägen zur Gestaltung des regulatorischen Rahmens an Entscheidungsträger:innen kommuniziert werden. Des Weiteren dient das entwickelte Bewertungsschema der konstruktiven Technikgestaltung im Rahmen von zukünftigen Forschungs- und Entwicklungsprojekten von KI für Sicherheitsbehörden.

Die Ergebnisse des Projektes sollen zudem zur Entwicklung eines Verfahrens zur Bewertung und wissenschaftlichen Evaluation des KI-Einsatzes sowie zur Validierung von KI-Modellen im Phänomenbereich der Politisch Motivierten Kriminalität nach ethischen Gesichtspunkten wie Fairness genutzt werden. Hierfür ist ein Zeitrahmen von etwa 3 Jahren nach Projektende vorgesehen.

Die Ergebnisse sollen zudem in den ersten zwei Jahren nach Projektende in Beiträgen für polizeiwissenschaftliche und kriminologische Fachzeitschriften (z.B. European Journal of Policing Studies, Kriminalistik, siehe auch geplante Veröffentlichungen weiter unten) aufbereitet und auf weiteren einschlägigen anwendernahen Konferenzen (Europäischer Polizeikongress, Deutscher Präventionstag) präsentiert werden.

II.5 Während der Durchführung des Vorhabens dem ZE bekannt gewordener Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen

Seit Projektbeginn sind fortlaufend Veröffentlichungen zum Themenfelder der ethischen Bewertung von KI-Anwendungen erschienen, die für das Teilvorhaben in konzeptueller Hinsicht relevant waren und entsprechend ausgewertet wurden. Zum Thema des KI-Einsatzes in Polizeibehörden sind keine unmittelbar relevanten Veröffentlichungen erschienen. Mit dem BMBF-Projekt VIKING, Teilvorhaben HWR, in dessen Rahmen ebenfalls ein Bewertungsschema für den Einsatz von KI in Polizeibehörden entwickelt wird, fand ein Austausch im Rahmen der

von TUB mitveranstalteten Tagung „Künstliche Intelligenz in der Kriminalitätskontrolle: Gesellschaftliche und rechtliche Dimensionen“ am 5. Mai 2023 an der TUB statt. Dabei zeigten sich Übereinstimmungen in den allgemeinen Bewertungsdimensionen. Während bei VIKING der Schwerpunkt eher auf rechtlichen Fragestellungen lag, lag dieser bei TUB auf ethisch-gesellschaftlichen Anforderungen. Der Ansatz unterscheidet sich ebenfalls darin, dass TUB detaillierte Gestaltungsanforderungen für das spezifische Anwendungsszenario Internetauswertungen im polizeilichen Staatsschutz entwickelt hat. KISTRA jedoch das Anwendungsszenario einen Leitfaden für den Einsatz von KI im Staatsschutz. Des Weiteren fand im Rahmen einer Videokonferenz am 21. Juli 2023 ein Austausch mit dem Projekt PABOS (Planung einer Algorithmenbewertungsstelle für Behörden und Organisationen mit Sicherheitsaufgaben) der ZITIS, dem Projekt KIRP (KI-relevante Plattformen) im Rahmen von P20 und dem KI-Campus statt. Es bestehen demnach Überschneidungen in den Zielsetzungen (Bewertung von KI-Systemen), wobei TUB einen stärkeren Fokus auf die Bewertung des soziotechnischen Gesamtsystems, in das ein KI-Modell integriert wird, legt. Auch unterscheidet sich der Bewertungsansatz der TUB durch eine dezidiert polizeiwissenschaftlich-kriminologische Perspektive auf polizeiliche Internetauswertungen (siehe auch das AP zur Praxisanalyse) und die damit verbundene stärkere Akzentuierung einer sozialwissenschaftlich-kriminologischer Phänomenexpertise im Bewertungsprozess, etwa was Anforderungen an Validierungsdaten betrifft. Nicht zuletzt liegt den anderen Projekten ein breiter Fokus hinsichtlich polizeilicher Anwendungsszenarien zu Grunde.

II.6 Erfolgte oder geplante Veröffentlichungen des Ergebnisses

- Vortrag „Herausforderungen der Strafverfolgung von Rechtsextremismus und Hasskriminalität im Kontext des novellierten NetzDG“ (Jonatan Schewe & Robert Pelzer), Konferenz „Im toten Winkel – Wie Rechtsextreme alternative Online-Plattformen zur Radikalisierung nutzen“, Jahreskonferenz 2021 des Projektes »Radikalisierung in rechtsextremen Online-Subkulturen entgegentreten« des ISD vom 25. bis 26. November 2021 in Berlin
- Poster „Bewertungsschema für gesellschaftlich verantwortungsvolle und effiziente KI-Systeme in Sicherheitsbehörden“ (Robert Pelzer & Michael Hahne), SIFO-Innovationsforum vom 3. bis 4. Mai 2022 in Berlin
- Vortrag „KI-Unterstützung polizeilicher Internetauswertungen im Bereich PMK: Potentiale und Herausforderungen aus praktischer, ethischer und rechtlicher Sicht“ (Robert Pelzer & Elias Tiemann), Konferenz „TechZoom Hasskriminalität“ der Zentrale Stelle für Informationstechnik im Sicherheitsbereich am 24. Mai 2022
- Pelzer, Robert (2022): Verfolgung und Prävention von Hasskriminalität im Internet: Benötigt es „mehr Polizei“ in Sozialen Medien?. In: PRIF blog, online verfügbar unter: https://blog.prif.org/2022/10/14/verfolgung-und-praevention-von-hasskriminalitaet-im-internet-benoetigt-es-mehr-polizei-in-sozialen-medien/

- Vortrag „Bewertung von KI-Systemen bei der Polizei“ (Michael Hahne), Tagung „Künstliche Intelligenz in der Kriminalitätskontrolle: Gesellschaftliche und rechtliche Dimensionen“ am 5. Mai 2023 in Berlin
- Vortrag „Rahmenbedingungen für den ethisch und rechtlich vertretbaren Einsatz von Künstlicher Intelligenz – Wie können Sicherheitsbehörden der besonderen Verantwortung beim Einsatz von KI gerecht werden?“ (Robert Pelzer), Kloster-Klausur des Behörden Spiegel „Digitale Kriminalistik als komplexe Herausforderung für die Kriminalpolizei“ vom 21. bis 23. August 2023 im Kloster Drübeck/Harz
- Vortrag „Chancen und Risiken von Künstlicher Intelligenz in der Polizeiarbeit“ (Robert Pelzer), Führungskräftetagung der Direktion 11 der Bundespolizei, 14. Februar 2024 in Berlin
- Vortrag: „Herausforderungen des Einsatzes von KI-Tools zur polizeilichen Auswertung von neuen digitalen Kommunikationsräumen“ (Robert Pelzer & Stefan Taing), Side event „Ein technischer, rechtlicher und gesellschaftlicher Blick auf die Auswirkungen von KI auf die zukünftige Polizeiarbeit“ auf dem Europäischen Polizeikongress am 17. April 2024

Geplante Veröffentlichungen:

Pelzer, Robert & Taing, Stefan (2025): „Herausforderungen des Einsatzes von KI-Tools zur polizeilichen Auswertung von neuen digitalen Kommunikationsräumen“. In: Honekamp/Kemme (Hg.): „Ein technischer, rechtlicher und gesellschaftlicher Blick auf die Auswirkungen von KI auf die zukünftige Polizeiarbeit“, Springer VS.

„Künstliche Intelligenz in polizeilichen Auswerteprozessen – ein Leitfaden zur konstruktiven Bewertung ethischer Anforderungen“ (Zielgruppe: Praxis, Journal: Kriminalistik)

„Ethical requirements for artificial intelligence in policing hate crime: a socio-technical-criminological perspective” (Zielgruppe: Polizeiforschung, Journal: European Journal of Policing Studies)

Literatur

Beck, U. (1986). Risikogesellschaft: Auf dem Weg in eine andere Moderne. Suhrkamp Verlag.

Carius, A., & Renn, O. (2003). Umweltkonfliktmanagement: Ein Handbuch. Oekom Verlag.

Endreß, M. (2002). Vertrauen. Transcript Verlag.

Epp, A., Brauerhoch, F. O., Ewen, C., Sinemus, K., Hertel, R., & Böl, G. F. (2008). Formen und Folgen behördlicher Risikokommunikation. Bundesinstitut für Risikobewertung.

Hempel, L., & Lammerant, H. (2014). Impact Assessments as Negotiated Knowledge. In S. Gutwirth, R. Leenes, & P. De Hert (Eds.), Reforming European Data Protection Law (pp. 125-145). Springer Netherlands.

Johnson, B. B., & Covello, V. T. (Eds.). (1995). The social and cultural construction of risk: Essays on risk selection and perception (Vol. 3). Springer.

Jungermann, H. (1991). Inhalte und Konzepte der Risiko-Kommunikation. In Risiko-Kommunikation (pp. 335-354). Springer Berlin Heidelberg.

Kurzenhäuser, S., Epp, A., Schütz, H., & Bender, M. (2010). Risikokommunikation bei neuen Technologien: Ein kommunikationswissenschaftlicher Ansatz. Springer VS.

Renn, O. (2004). Risikokommunikation: Der Umgang mit Risiken in Politik, Verwaltung und Technik. Springer VS.

Renn, O., Carius, R., Kastenholz, H., & Schulze, M. (2005). ERiK - Entwicklung eines mehrstufigen Verfahrens der Risikokommunikation. Bundesinstitut für Risikobewertung (BfR-Wissenschaft, 2005/2).

Renn, O., & Levine, D. (1991). Credibility and trust in risk communication. In R. E. Kasperson & P. J. M. Stallen (Eds.), Communicating Risks to the Public. Technology, Risk, and Society, vol 4. Springer, Dordrecht.

Renn, O., & Zwick, M. M. (Eds.). (1997). Risiko- und Technikakzeptanz. Springer.

Rogers, E. M. (1983). Diffusion of Innovations. Free Press.

Scheer, D., Gold, S., Benighaus, C., Benighaus, L., Ortleb, J., & Renn, O. (2010). Kommunikation von Risiko und Gefährdungspotenzial aus Sicht verschiedener Stakeholder. Bundesinstitut für Risikobewertung.

Ulbig, E., Hertel, R. F., & Böl, G. F. (2010). Evaluierung der Kommunikation über die Unterschiede zwischen „risk“ und „hazard“ Abschlussbericht. Bundesinstitut für Risikobewertung.

Wiedemann, P. M., Carius, R., Henschel, C., Kastenholz, H., Nothdurft, W., Ruff, F., & Uth, H.-J. (2000). Risikokommunikation für Unternehmen. VDI Verlag.

Wiedemann, P. M., & Clauberg, M. (2005). Risikokommunikation. In R. Fehr, H. Neus, & U. Heudorf (Hrsg.), Gesundheit und Umwelt. Ökologische Prävention und Gesundheitsförderung (S. 316–328). Huber.

Ziemann, A. (2013). Zur Philosophie und Soziologie der Situation–eine Einführung. In Offene Ordnung? Philosophie und Soziologie der Situation (pp. 7-18).

Zimmer, R., Kloke, S., & Gaedtke, M. (2013). Der Streit um die Uckermarkleitung - Eine Diskursanalyse. UfU-Paper 3.