Predictive Analytics in der Vergabe und Kontrolle von staatlichen Transferleistungen

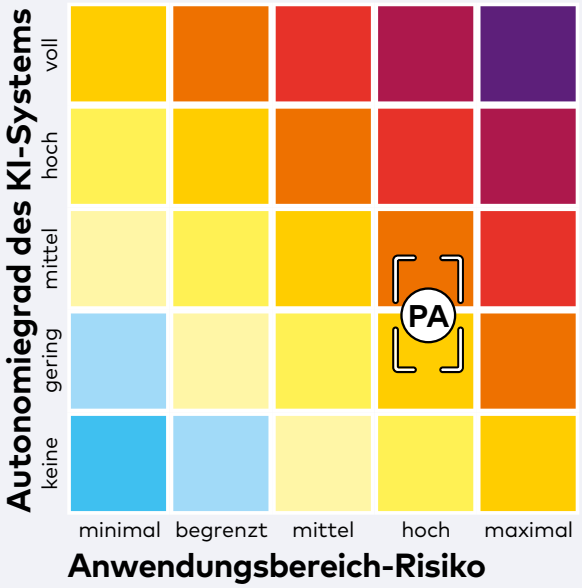
Executive Summary

Eine faire und transparente Vergabe von staatlichen Transferleistungen, wie etwa Familienbeihilfe, Sozialhilfe, Wohnbeihilfe uvm. stellt Behörden und staatliche Einrichtungen vor große Herausforderungen. Einerseits sollte den Betroffenen so rasch und treffsicher wie möglich geholfen werden, andererseits sind aber Fehler, Ineffizienz und Diskriminierung bei der Bearbeitung von Anträgen zu vermeiden. In der Praxis greift die Verwaltung oftmals auf Predictive Analytics-Systeme zurück, die zur Lösung dieser Probleme beitragen können. Da KI-gestützte Entscheidungen über die Gewährung von staatlichen Transferleistungen unmittelbaren Einfluss auf das Leben der Bürger:innen haben, sind sie mit der gebotenen Sorgfalt und Vorsicht zu treffen.

Im ALAIT Risikoradar wird daher das Gesamtrisiko bei der Nutzung von Predictive Analytics-Systeme für die Bewertung der Anspruchsberechtigung staatlicher Transferleistungen als hoch bis mittel eingestuft (dunkel- und hellorange, s. Grafik). Diese Gesamtbewertung erschließt sich aus einem hohen Risiko aus dem Anwendungskontext (in Übereinstimmung mit dem *EU AI-Act*) gemeinsam mit einem mittleren bis geringeren *Autonomiegrad* der KI-Systeme. In Bezug auf den Autonomiegrad sind zur Anwendung kommende Predictive Analytics-Systeme meist nicht voll autonom, da in der Regel Sachbearbeiter:innen den Prozess überwachen und jederzeit eingreifen bzw. Entscheidungen an sich ziehen können (Stufe 3, Human-on-the-Loop).

Die bisherigen praktischen Erfahrungen mit Predictive Analytics-Systemen in verschiedenen europäischen Ländern zeigen, dass damit mehr Output erzeugt und mehr Fälle bearbeitet werden können.[1,8,18] Die automatisierte Bearbeitung von Anträgen führt zu rascheren Entscheidungen und entlastet Sachbearbeiter:innen. Allerdings stehen diesen Vorteilen auch erhebliche Herausforderungen gegenüber: Predictive Analytics-Systeme können – anders als Menschen – Einzelentscheidungen nicht kontextualisieren und auf konkrete Lebenssituationen der Bürger:innen Rücksicht nehmen. Dies kann systematisch zu unfairen Entscheidungen führen. Bei mangelnder Qualität der Trainingsdaten kommt es bei KI-Systemen zu systematischen *Verzerrungen (KI-Bias)* und damit zu Diskriminierung und falschen Entscheidungen. Ein weiteres Problem stellt die mangelnde Transparenz und Nachvollziehbarkeit bei vielen KI-Systemen dar. Dieses sogenannte *Black-Box-Problem* ist insofern problematisch, als gerade Behörden dazu verpflichtet sind, ihren Bürger:innen begründete Entscheidungen vorzulegen und die gesetzlichen Anforderungen zum Schutz der Grundrechte, insbesondere des Rechts auf Privatsphäre und des Rechts auf Nichtdiskriminierung, einzuhalten.

ALAIT Risikoradar für Predictive Analytics in der Vergabe und Kontrolle von staatlichen Transferleistungen



PA KI-Anwendungen im Bereich Predictive Analytics

Das ALAIT Risikoradar ist ein wissenschaftlich entwickeltes Risikoanalysetool für Künstliche Intelligenz (KI), das KI-Anwendungen kontextbezogen und unter Berücksichtigung ihres technischen Autonomiegrades einstuft und so die Risikosphäre für Anwender:innen auf einen Blick sichtbar macht. Dabei gilt: Je höher das Einsatz-Risiko aus dem Anwendungskontext und je größer der Autonomiegrad des KI-Systems in Bezug auf Entscheidungen, desto riskanter ist der Einsatz einzustufen. Eine erweiterte Klammer weist auf eine Bandbreite in der Risikoeinstufung hin. Ein geringer Autonomiegrad eines KI-Systems bedeutet nicht, dass man sich zurücklehnen kann. Es erfordert eine starke Rolle der Menschen, die es anwenden. (Details zum Stufenmodell s. S. 8f).

**ALAIT ist ein vom Bundesministerium für Innovation, Mobilität und Infrastruktur (BMIMI) initiiertes Forschungsprojekt zur Entwicklung einer neuartigen Form von Technologiebewertung für KI-Technologien mit wissenschaftlichen Methoden.*

Behörden und Systementwickler:innen können das Risiko allerdings reduzieren, wenn sie entsprechende Maßnahmen zur Risikominderung umsetzen. Grundsätzlich dürfen nur Systeme, die der Europäische Menschenrechtskonvention entsprechen, zum Einsatz kommen. Sie müssen mit möglichst verzerrungs- und diskriminierungsfreien Datensätzen trainiert werden und dem Stand der Technik entsprechen. Außerdem müssen menschliche Kontrolle und Aufsicht durch Verwaltungsmitarbeiter:innen sicherstellen, dass Anfragen und Beschwerden von Bürger:innen adäquat bearbeitet werden. Es ist dafür zu sorgen, dass Sachbearbeiter:innen in Behörden und staatlichen Stellen kontinuierlich geschult werden – sowohl in Bezug auf die Funktionsweise und Grenzen der Predictive Analytics-Systeme als auch in Bezug auf soziale und rechtliche Anforderungen.

Einleitung

Unter staatlichen Transferleistungen versteht man finanzielle Unterstützungen wie Familienbeihilfe, Sozialhilfe und Studienbeihilfen, die ohne eine direkte Gegenleistung an anspruchsberechtigte Bürger:innen geleistet werden. Diese staatlichen Transferleistungen sollen Bürger:innen absichern und Chancengerechtigkeit fördern.[22,23] Gleichzeitig müssen sie effizient, rechtmäßig und fair vergeben werden – ein Anspruch, der angesichts knapper öffentlicher Mittel und der öffentlichen Debatte um Verteilungsgerechtigkeit an Bedeutung zunimmt. Komplexe Anspruchsprüfungen werden daher zunehmend durch Künstliche Intelligenz (KI-)Systeme wie Predictive Analytics und Automatisierte Entscheidungssysteme (AES) unterstützt. Bei Predictive Analytics werden Daten aus der Vergangenheit herangezogen und mithilfe statischer Verfahren analysiert, um Muster zu erkennen und daraus zukünftige Ereignisse vorhersagen zu können.[1] Predictive Analytics-Systeme liefern somit Entscheidungsgrundlagen, treffen jedoch keine Entscheidungen selbst. Automatisierte Entscheidungssysteme (AES) gehen einen Schritt weiter: Sie nutzen die von Predictive Analytics oder anderen KI-Systemen erzeugten Informationen, um konkrete Entscheidungen auf Basis festgelegter Kriterien oder Zielvorgaben eigenständig umzusetzen – etwa bei der automatischen Bewilligung oder Ablehnung eines Antrags.[6]

Der Hauptvorteil dieser Technologien besteht darin, dass sie große Datenmengen in kurzer Zeit analysieren, um Muster zu erkennen oder Vorhersagen über künftige Ereignisse treffen zu können.[2] Mitarbeiter:innen können damit beispielsweise Anomalien wie Betrug und Geldwäsche schneller und genauer erkennen. In Österreich hat das Finanzministerium ein „Predictive Analytics Competence Center" (PACC) eingerichtet, welches die Behörden bei der Kontrolle von Steuer- von und Beihilfenzahlungen an Betrieben unterstützt.[3] Andere Länder setzen Predictive Analytics und AES bereits bei der Verteilung von staatlichen Transferleistungen oder zur Aufdeckung von Sozialbetrug ein.[4,5] Etwa setzt die schwedische Studienbeihilfebehörde (CSN) ein System ein, das in den meisten Fällen völlig automatisiert über die Zahlung von Finanzhilfen an Studierende entscheidet. Das System kombiniert dabei Informationen der Studienbeihilfenbehörde mit in Schweden öffentlich zugänglichen Informationen wie Steuerdaten und personenbezogenen Daten der Antragsteller:innen. Da nur im Falle eines Einspruchs Mitarbeiter:innen der Behörde eingreifen müssen, bietet das System im Idealfall eine hohe Servicequalität und Schnelligkeit und wird auch von Antragsteller:innen positiv bewertet.[5] Allerdings wird durch diese Vorgangsweise die Last der Überprüfung auf Richtigkeit von Entscheidungen auf Bürger:innen übertragen.

Trotz der Vorteile dieser Systeme können mit deren Einführung auch große Herausforderungen verbunden sein, die in jedem Fall eine sorgfältige Abwägung von Chancen und Risiken und entsprechende Maßnahmen erfordern.[5,9] Der Einsatz von KI-Systemen ist nicht immer eine geeignete Lösung, insbesondere, wenn Grundrechte verletzt oder vulnerable Bevölkerungsgruppen benachteiligt werden. Studien zeigen, dass viele dieser Systeme dazu neigen, sozial benachteiligte Bevölkerungsgruppen überproportional negativ zu treffen, da sie in den Datensätzen über- oder unterrepräsentiert sind und somit systematisch diskriminiert werden.[34] Ein prägnantes Beispiel ist das in den Niederlanden eingesetzte System „SyRI" zur Aufdeckung von Sozialbetrug. Es verstieß gegen das Recht auf Privatsphäre gemäß der Europäischen Menschenrechtskonvention und wurde nach massiver Kritik und einem Gerichtsurteil ausgesetzt. Zudem lieferte es fehlerhafte Ergebnisse, führte zu mangelnder Transparenz und Fairness und hatte gravierende Folgen für Tausende betroffene Familien.[10,11,12]

Technologiebeschreibung

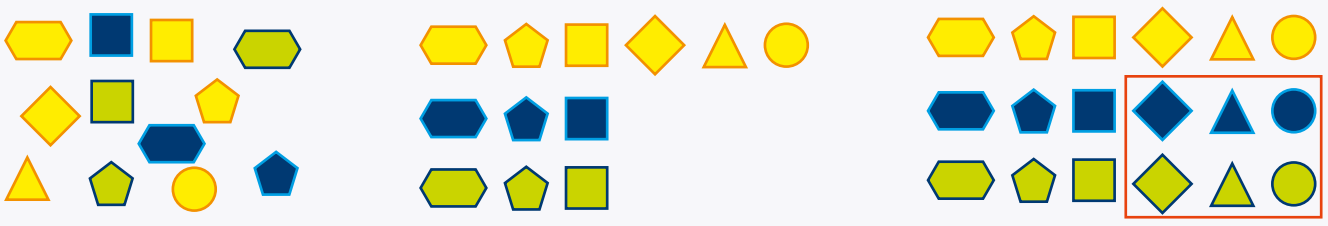
Für Predictive Analytics wird häufig maschinelles Lernen (ML) verwendet, eine Methode, die die Analyse von Trainingsdaten zur Vorhersage künftiger Ereignisse und zum Risikomanagement verwendet.[13,14] Im Kern basiert maschinelles Lernen auf Mustererkennung und Klassifizierung bei gekennzeichneten Daten (*überwachtes Lernen*) oder – im Bereich der Predictive Analytics weniger verbreitet – ungekennzeichneten Daten (*unüberwachtes Lernen*).[8,20] Zu den Werkzeugen des maschinellen Lernens zählen:[21]

1. Künstliche neuronale Netze: das sind Rechenmodelle, die aus vielen miteinander verbundenen Knoten (Neuronen) bestehen. Diese Knoten wenden gewichtete mathematische Funktionen auf Eingaben an, transformieren sie mit Aktivierungsfunktionen und geben die Ergebnisse an nachfolgende Schichten weiter.

2. Entscheidungsbäume: das sind baumartige Strukturen, die auf Basis von Merkmalen der Daten schrittweise Entscheidungen treffen. Jeder Knoten stellt eine Bedingung dar, die zu unterschiedlichen Ästen und schließlich zu einer Klassifikation oder Vorhersage führt.

3. Support-Vektor-Maschinen (SVM): das sind überwachte Lernverfahren, die Datenpunkte durch das Finden einer optimalen Trennlinie (Hyperplane) in Klassen einordnen. Ein Beispiel ist die Unterscheidung zwischen Spam- und Nicht-Spam-E-Mails anhand von Merkmalen wie bestimmten Wörtern, Satzlängen oder Absenderinformationen.

Das folgende Diagramm zeigt das allgemeine Funktionsprinzip von Predictive Analytics mit maschinellem Lernen:



Daten sammeln **Muster erkennen** **Vorhersage treffen**

Zunächst werden große Mengen heterogener Daten aus verschiedenen Quellen gesammelt und aufbereitet. Anschließend werden mithilfe von ML-Algorithmen Muster, Korrelationen und wiederkehrende Strukturen innerhalb dieser Daten identifiziert. Auf Basis dieser Muster können zukünftige Ereignisse vorhergesagt und Handlungsempfehlungen abgeleitet werden.

Chancen

Die Nutzung von Predictive Analytics in der Vergabe und Kontrolle staatlicher Transferleistungen kann für die jeweiligen Stakeholder eine Reihe von Vorteilen bringen:

Nutzen für Bürger:innen:

- **Rasche Entscheidungen:** Im Gegensatz zu manueller Bearbeitung bieten maschinelle Entscheidungen den Vorteil hoher Geschwindigkeit. Bürger:innen können rasch Auskunft über Anspruchsberechtigungen erhalten.[1]
- **Konsistenz der Entscheidungsfindung:** Ein Predictive Analytics-System, das auf sorgfältigen und realistischen Modellannahmen basiert und mit hochwertigen Daten trainiert und richtig eingestellt wurde, bietet im Idealfall eine transparente, regelkonforme, faire und konsistente Entscheidungsgrundlage.[1]

Nutzen für Beschäftigte im öffentlichen Dienst:

- **Entscheidungsunterstützung und Arbeitsentlastung:** Mithilfe von Predictive Analytics-Systemen erzielte das österreichische Finanzministerium (PACC) eine Arbeitsentlastung bei seinen Mitarbeiter:innen von bis zu 40 %.[20] Insbesondere durch die Automatisierung zeitintensiver und sich wiederholender Aufgaben bieten automatisierte Entscheidungssysteme die Möglichkeit, administrative Routineaufgaben für Sachbearbeiter:innen zu minimieren und so mehr Zeit für aufwändigere Aufgaben zu schaffen.[1]

Chancen für staatliche Organisationen:

- **Umfang und Geschwindigkeit:** Maschinelles Lernen kann viele Verwaltungsaufgaben automatisieren und dadurch den Personalbedarf deutlich senken. Angesichts knapper Budgets und des bereits heute

wie künftig fehlenden Fachpersonals bietet der Einsatz von Predictive Analytics eine Möglichkeit, Kosten zu reduzieren und Verwaltungsprozesse um bis zu 31% beschleunigen.[1,13,14]

- **Verbesserte Missbrauchskontrolle:** Der entscheidende Fortschritt von Predictive Analytics-Systemen besteht darin, dass sie wesentlich größere und komplexere Datenmengen in Echtzeit analysieren können. Dadurch werden Risikomuster sichtbar, die sich mit herkömmlichen Kontrollmechanismen entweder gar nicht oder nur mit erheblichem Aufwand erkennen lassen. Das von der Europäischen Kommission entwickelte Risikobewertungstool ARACHNE unterstützt beispielsweise den Europäischen Sozialfonds (ESF) und den Europäischen Fonds für Regionale Entwicklung (EFRE) bei der Aufdeckung von Betrug und Misswirtschaft bei der Verwendung öffentlicher Fördergelder.[2] Predictive Analytics-Tools werden beispielsweise auch von der italienischen Regierung zur Aufdeckung von Korruption und Geldwäsche im öffentlichen Beschaffungswesen eingesetzt und helfen dadurch erfolgreich Mafia-Aktivitäten zu verhindern.[2,13] Darüber hinaus entwickelt die Europäische Union derzeit KI-Systeme zur Überwachung der Einhaltung des EU-Rechts. Beispielsweise hat das Claudette-Projekt ein Predictive Analytics-System entwickelt, das unrechtmäßige oder unfaire Klauseln in Verträgen und Datenschutzrichtlinien identifiziert.[35]

- **Evidenzbasierte Politikgestaltung:** Predictive Analytics-Systeme erlauben, aus historischen Daten auf wahrscheinliche zukünftige Entwicklungen zu schließen. Diese Informationen können in der Politikgestaltung auf vielfältige Weise genutzt werden (sogenannte „evidenzbasierte" Politikgestaltung):[8,10]
  1. Erkennung von Bedarf und Analyse von Wirkungen monetärer Programme.
  2. Erstellung von Prognosen und Szenarien zur Gestaltung von maßgeschneiderten Präventionsprogrammen.
  3. Permanente Neuintegration von Daten ermöglicht stetige Evaluierung und Anpassung der Programme (sog. policy learning).
  4. Entwicklung von Frühwarnsystemen zur Einleitung staatlicher Maßnahmen; beispielsweise experimentieren einige Gemeinden in Dänemark und Spanien mit Vorhersagemodellen über die Hilfsbedürftigkeit älterer Bürger:innen um entsprechende Maßnahmen zeitgerecht einleiten zu können.[10]

Herausforderungen und Risiken

Trotz der genannten Vorteile ist der Einsatz von Predictive Analytics zur Kontrolle und Vergabe von staatlichen Transferleistungen auch mit Kritik und Herausforderungen verknüpft.

1. **„Vermessung" von Bürger:innen:** Eine Europäische Studie unter Beteiligung der Österreichischen Akademie der Wissenschaften zeigt, dass es bei der Anwendung von Predictive Analytics zu einer „Vermessung" und damit Reduzierung von Bürger:innen auf Daten und Vergleiche mit Durchschnittswerten kommt. Wichtige Informationen aus dem Kontext konkreter Fälle werden möglicherweise außer Acht gelassen. Diese Diskrepanz zwischen den Algorithmen und der konkreten sozialen Realität von Menschen kann zu unfairen Ergebnissen führen.[29,32]

2. **Zunahme von Ungleichheit und *Verzerrungen (KI-Bias)*:** Voreingenommene Predictive Analytics-Anwendungen im Sozialleistungssystem richten sich häufig gegen genau die Menschen, die sie eigentlich schützen sollen. Mehrere Projekte in ganz Europa sind aufgrund von Problemen mit voreingenommenen automatisierten Algorithmen in erhebliche Kritik geraten.[4,7,10,12] Das oben genannte Beispiel des niederländischen SyRI-Systems ist der prominenteste Fall, der die drastischen Auswirkungen solcher Systeme auf einzelne Personen sowie die Gesellschaft durch Verzerrungen zeigt.

3. **Datenschutz:** Regierungen könnten durch die Erhebung von Daten aus Einkommens- und Beschäftigungsunterlagen, Reiseverläufen, Aktivitäten in sozialen Medien etc. die Grundlage für eine umfassende individuelle Überwachung ihrer Bürger:innen schaffen.[4,13] Selbst wenn die Datenerhebung rechtlich gedeckt wäre, erfolgt diese in der Regel ohne Wissen oder expliziter Zustimmung der betroffenen Personen.[4] Erschwerend kommt hinzu, dass anonymisierte Daten aus einzelnen Datenbanken durch die Kombination mit anderen Daten wieder de-anonymisiert werden können.[8]

4. **Datenqualität und Fehleinschätzungen:** Die gesammelten Daten müssen auf eine bestimmte Weise organisiert sein, um *genaue* Vorhersagen zu gewährleisten. Allerdings sind Datenbestände von Behörden oft unübersichtlich, veraltet oder über verschiedene Abteilungen verteilt. Das Europäische Parlament hat festgestellt, dass inkonsistente

Daten (z.B. in den Subventionsaufzeichnungen der EU-Mitgliedstaaten) eine einheitliche Anwendung von Predictive Analytics-Tools massiv erschweren.[2] Eine fehlerhafte Vorverarbeitung der Daten kann systematische Falschentscheidungen, z.B. Nichtgewährung von Unterstützung für Berechtigte (falsch negativ) oder Gewährung von Unterstützung für Nichtberechtigte (falsch positiv), mit erheblichen Konsequenzen für Antragsteller:innen zur Folge haben. So hat sich etwa gezeigt, dass die Vorhersagen des KI-Systems der Kopenhagener Stadtverwaltung für Pflegebedarf nach Krankenhausaufenthalten nur in vier von fünf Fällen zutreffend sind.[17] Die schwerwiegenden Folgen, die sich aus einer Falscheinschätzung durch ein KI-System für betroffene Bürger:innen ergeben, dürfen keinesfalls durch potenzielle Effizienzgewinne von Behörden gerechtfertigt werden.[10]

5. **Transparenz:** Die Transparenz von Predictive Analytics-Systemen soll sicherstellen, dass Mitarbeiter:innen Predictive Analytics-System-Entscheidungen nachvollziehen können, um Fehler und Diskrepanzen aufzudecken und um letztendlich ihre Kontrollfunktion ausüben und die Verantwortung für Entscheidungen tragen zu können. Aus technischer Sicht besteht allerdings nach wie vor das Problem, dass die meisten modernen KI-Systeme aufgrund ihrer Komplexität – selbst für Expert:innen – nicht transparent und nachvollziehbar sind, ein Phänomen das auch als „*Black-Box*" bezeichnet wird.[8,16] Darüber hinaus erfordert die Nutzung von KI-Tools eine entsprechende Expertise und Schulung, die von den Behörden und Unternehmen, die solche Tools einsetzen wollen, sichergestellt werden müssen (Siehe Artikel 4, *EU AI-Act*).[25]

6. **Übermäßiges Vertrauen in Technologien:** Mehrere Studien haben gezeigt, dass Menschen dazu neigen, KI-Systemen mehr zu vertrauen als Menschen. Ein Phänomen, das als „Automatisierungsbias" bezeichnet wird.[5] Dies gilt sowohl für Verwaltungsmitarbeiter:innen, die mit KI-Tools arbeiten, als auch für Bürger:innen, die staatlichen Hilfeleistungen und Services in Anspruch nehmen. Es besteht die Gefahr, dass Menschen Predictive Analytics-System-Vorschläge oder -Entscheidungen nicht mehr ausreichend hinterfragen, selbst dann, wenn sie widersprechende Informationen und Wissen über einen konkreten Fall haben.[8] Daher führt die Unterstützung durch KI-Systeme nicht automatisch zu besseren Entscheidungen, wie auch die aktuelle Übersichtsstudie des Massachusetts Institute of Technology (MIT) aus dem Jahr 2024 zeigt.[19]

Empfehlung zum Praxiseinsatz

Für Behörden (System-Anwender:innen):

1. **Prämisse menschliche Aufsicht:** Menschliche Aufsicht und Letztentscheidung an kritischen Punkten des Entscheidungsprozesses nutzt die Vorteile von Predictive Analytics-Systemen zur Vergabe/Kontrolle von staatlichen Transferleistungen und hilft, die bestehenden Risiken einzudämmen. Dieser Human-in-the-Loop-Ansatz betont die unterstützende Funktion von Predictive Analytics im Verwaltungswesen im Gegensatz zum Austausch von Menschen durch Maschinen.[1] Darüber hinaus sollte die Zusammenarbeit zwischen Mensch und Maschine im Fokus stehen. Künftig wird ein neues Kollaborations-Design zwischen mehreren Stakeholdern gefragt sein, das soziotechnische Synergien verbessert und nicht untergräbt.[29] Erklärbarkeit und *Transparenz* sollen ebenfalls eine Vorbedingung für die Aufsicht sein. Menschliche Aufsicht ist das Fundament eines ethischen KI-Ansatzes, um kontextualisierte Entscheidungen zu treffen, die die Rechte der Bürger:innen schützen und algorithmische Fehler verhindern.[16]

2. **Schaffung und Stärkung des KI-Wissens:** Die Rolle des Staates und seiner Verwaltung im Umgang von und der Nutzung mit KI ist mit einer starken Vorbildwirkung verbunden. Das bedeutet, dass mit der (steigenden) Nutzung von KI im Verwaltungsapparat auch das Wissen und die Kompetenzen der Verwaltungsmitarbeiter:innen gestärkt werden müssen. Artikel 4 des *EU AI-Act* besagt: „Anbieter und Entwickler von KI-Systemen müssen Maßnahmen ergreifen, um im Rahmen ihrer Möglichkeiten ein ausreichendes Maß an KI-Kompetenz ihrer Mitarbeiter und anderer Personen sicherzustellen, die in ihrem Auftrag mit dem Betrieb und der Nutzung von KI-Systemen befasst sind, wobei sie deren technische Kenntnisse, Erfahrungen, Ausbildung und Schulung sowie den Kontext, in dem die KI-Systeme eingesetzt werden sollen, berücksichtigen und die Personen oder Personengruppen berücksichtigen, auf die die KI-Systeme angewendet werden sollen".[25]

3. **Präzise, hochwertige Daten:** Die Zuverlässigkeit von Predictive Analytics-Systemen hängt zu einem wesentlichen Teil von der Qualität der Datensätze ab, mit denen sie trainiert bzw. im laufenden Betrieb gespeist werden. Die Datensätze sollen darüber hinaus diversifiziert werden, um die Genauigkeit zu verbessern. Für die Nutzung von Daten im Bereich Predictive Analytics sind weiters folgende Grundsätze einzuhalten:[11,13,24]

a. **Datenminimierung:** Es wird empfohlen, nur die für die Erfüllung der Aufgabe erforderliche Mindestmenge an personenbezogenen Daten zu verwenden.

a. **Zweckbindung:** Daten, die zur Erbringung von staatlichen Leistungen erhoben werden, sollten nicht willkürlich für andere Zwecke verwendet werden.

4. **Ausrichtung an Richtlinien und Implementierung von Risikobewertungstools vor der Einführung:** Alle Predictive Analytics-Systeme müssen geprüft werden, ob sie den Richtlinien der Datenschutz-Grundverordnung (DSGVO) und des ***EU AI-Act*** entsprechen. Die europäische KI-Regulierung ergänzt die Vorschriften der DSGVO und zwingt Behörden vor der Implementierung von Predictive Analytics Tools, die Risiken des Systems für Grundrechtsverletzungen abzuschätzen und entsprechende Maßnahmen zu ergreifen.[1,8,16,29] Darüber hinaus überwachen die Behördenstellen wie Europäisches Büro für Künstliche Intelligenz („AI Office") und Europäisches Gremium für Künstliche Intelligenz („AI Board") die Umsetzung und Anwendung des ***EU AI-Act*** auf Unionsebene.[33]

5. **Niederschwellige Beschwerdewege:** Es muss transparent gemacht werden, dass eine Person das Recht und die Möglichkeit hat, gegen die Entscheidung eines automatisierten Systems Beschwerde einzulegen (Art. 13 und 52 von ***EU AI-Act***).

Für Systementwickler:innen:

1. **Transparenz:** Neben der Schaffung von Wissen gilt Transparenz als der vertrauensbildende Faktor im Einsatz von KI. Es ist unbedingt sicherzustellen, dass der Einsatz von KI in der Verwaltung mit direkten Auswirkungen auf die Betroffenen, also auf die Bürger:innen, maximalen Transparenzkriterien folgt. ***Erklärbare KI (XAI),*** die Entscheidungen und Vorhersagen von KI-Systemen für Menschen verständlich und nachvollziehbar machen soll, kann hier nützlich sein.[36] Ein solches System kann für Menschen verständliche Begründungen liefern, indem es angibt, welche Eingabefaktoren den Output des Systems am stärksten beeinflusst haben (z. B. Einkommensniveau oder Wohnkostenbelastung). Auf diese Weise können Sachbearbeiter:innen automatisierte Vorschläge überprüfen, validieren und überschreiben, sodass die endgültige Entscheidung unter menschlicher Kontrolle bleibt.[36]

2. **Vermeidung spezifischer persönlicher Merkmale:** Die genaue Abwägung, welche Merkmale einer Person in algorithmischen Systemen verarbeitet werden und zur Entscheidungsfindung beitragen, ist keine Kür, sondern Pflicht.[29] So ist unbedingt darauf zu achten, dass die Nutzung von messbaren und nicht messbaren Merkmalen (proxy Variablen), wie z.B. Nationalität oder ethnische Zugehörigkeit, nicht unreflektiert in KI-Tools zur Vergabe und Kontrolle von staatlichen Transferleistungen eingewoben werden. KI-Tools generell, jedoch insbesondere in diesem Einsatzgebiet, unterstehen einer besonderen Pflicht, bestehende gesellschaftliche Vorurteile und Stereotype nicht zu reproduzieren bzw. diese gar zu verstärken.[11]

Für Bürger:innen:

1. **Schaffung und Stärkung des KI-Wissens:** Es ist wichtig, sich über KI-Technologien zu informieren, die richtigen Kompetenzen aufzubauen und seine Rechte zu kennen.[8] In Österreich besteht mit der Rundfunk- und Telekom Regulierungs-GmbH (RTR) eine Informations- und Anlaufstelle für die breite Öffentlichkeit. Die RTR-KI-Servicestelle unterstützt Bürger:innen bei Fragen zur Transparenz, zu Rechten im Umgang mit KI-Systemen oder zur Überprüfung von KI-Entscheidungen.[33]

2. **Recht auf Transparenz:** Bürger:innen haben ein Recht darauf, zu erfahren, wie und wann ihre personenbezogenen Daten genutzt und in welche Systeme sie eingespielt werden (Art. 15 DSGVO).[13,31] Anlaufstelle bei Verdacht auf die Verletzung von Datenschutzrechten ist die Österreichische Datenschutzbehörde (DSB).[27,28]

3. **Möglichkeit für die Überprüfung einer KI-Entscheidung:** Wenn eine Entscheidung mit Hilfe von automatisierten Systemen getroffen wurde und rechtliche oder ähnlich erhebliche Wirkungen hat, z. B. Bewilligung, Kürzung oder Rückforderung, haben Bürger:innen ein Recht darauf, nicht einer ausschließlich auf einer automatisierten Verarbeitung – einschließlich Profiling – beruhenden Entscheidung unterworfen zu werden, sondern auch eine menschliche Überprüfung zu erhalten (Art. 22 DSGVO).[26] Im Falle einer Beschwerde gegenüber einer Behörde kann die Beschwerde bei der Volksanwaltschaft eingereicht werden.[30]

Wichtige Begriffe

KI-Genauigkeit (AI-Accuracy): Bezieht sich auf die Fähigkeit eines KI-Systems, korrekte Vorhersagen oder Entscheidungen zu treffen. Sie ist ein wichtiger Maßstab für ihre Leistung und entscheidend für die Bestimmung ihrer Wirksamkeit und Zuverlässigkeit.

KI-Autonomie: Die Fähigkeit eines KI-Systems, eine Reihe von Zielen unter einer Reihe von Unsicherheiten in ihrer Umgebung selbstständig und ohne externe Eingriffe zu erreichen.

Systematische Verzerrungen (Bias): Bias ist eine systematisch unterschiedliche Behandlung bestimmter Objekte, Personen oder Gruppen im Vergleich zu anderen. Behandlung ist jede Art von Handlung, einschließlich Wahrnehmung, Beobachtung, Darstellung, Vorhersage oder Entscheidung.

Gesetz der Europäischen Union über künstliche Intelligenz (EU AI-Act): Eine europäische Verordnung über künstliche Intelligenz (KI) – die erste umfassende Verordnung über KI von einer großen Regulierungsbehörde. Sie konzentriert sich insbesondere auf KI-Systeme mit hohem Risiko.

Datenschutz in der KI: Die Gesamtheit der Praktiken und Bedenken im Zusammenhang mit der ethischen Erfassung, Speicherung und Nutzung personenbezogener Daten durch Systeme der künstlichen Intelligenz.

Überwachtes Lernen (supervised learning): Ein Teilgebiet des maschinellen Lernens, bei der die Trainingsdaten vorab gekennzeichnet werden und das System das Muster zwischen dem Inhalt und der Kennzeichnung erlernt. Die Aufgabe eines solchen KI-Systems besteht darin, eine Beziehung zu finden, die jede Eingabe des Trainingssatzes (die Daten) einer Ausgabe (der Kennzeichnung) zuordnet.

Transparenz: Bedeutet, dass die Funktionsweise, Entscheidungsprozesse und Einsatzbereiche eines KI-Systems nachvollziehbar, erklärbar und offen zugänglich sind – für Entwickler:innen, Nutzer:innen und andere Stakeholder.

Unüberwachtes Lernen (unsupervised learning): Ist eine Methode des maschinellen Lernens, bei der ein Modell selbstständig Strukturen und Muster in Daten findet, ohne dass richtige Antworten (Labels) vorgegeben werden.

Black-Box-Systeme: Bezeichnet alle Systeme, deren interne Entscheidungsprozesse für Menschen nicht transparent und nachvollziehbar sind. Das heißt, dass nur Eingaben und Ausgaben beobachtet werden können, ohne zu verstehen, wie die Verarbeitung dazwischen genau abläuft.

Erklärbare KI (XAI): Bezieht sich auf Methoden und Techniken, die es ermöglichen, die Entscheidungen und Ergebnisse von KI-Systemen besser zu verstehen und nachzuvollziehen.

Erklärung Stufenmodell des ALAIT Risikoradars

Im ALAIT KI-Risikoradar wird die Beziehung zwischen Anwendungsrisiko und Autonomie eines KI-Systems dargestellt. Die Risikostufen stützen sich auf das EU KI-Gesetz (*EU AI-Act*), insbesondere auf Artikel 6 und Annex III, die sich mit risikoreichen Anwendungsbereichen von KI befassen. Geringere System-Autonomie und Anwendungsrisiken werden durch kältere Farben (blau) und höhere System-Autonomie und Anwendungsrisiken durch wärmere Farben (rot) dargestellt. Der Farbwechsel vermittelt das erhöhte Risiko solcher Entscheidungen. Mithilfe dieser Farbskala lässt sich das Gesamtrisiko erkennen: Violett und dunkelrot – sehr hoch, rot und dunkelorange - hoch, hellorange und Gelbtöne - mittel, Blautöne - geringes Gesamtrisiko. Im Idealfall sollten hohe Anwendungsrisiken und System-Autonomie vermieden oder nur nach sehr sorgfältiger Abwägung eingesetzt werden.

Autonomiegrad des KI-Systems

Stufe 1: Keine Autonomie

KI ist ein passives Werkzeug; Menschen treffen alle Entscheidungen und leiten Maßnahmen ein.

Beispiel: Diagnosesysteme, die medizinische Rohdaten anzeigen oder die Daten analysieren (ohne Empfehlungen!)

Empfohlene Anwendungsfälle: Szenarien mit hohen Risiken oder bei denen ethische Entscheidungen von entscheidender Bedeutung sind (z. B. medizinische Diagnostik, Justizsystem).

Stufe 2: Geringer Autonomiegrad (Human-in-the-Loop)

Die KI gibt Empfehlungen oder Optionen, aber der Benutzer:innen bleibt für die Auswahl und Genehmigung von Maßnahmen verantwortlich.

Beispiel: KI schlägt optimale Routen für die Logistik vor oder Empfehlungssysteme im E-Commerce.

Empfohlene Anwendungsfälle: Aufgaben mittlerer Komplexität mit mäßigen Risiken (z. B. Optimierung der Lieferkette).

Stufe 3: Mittlerer Autonomiegrad (Human-on-the-Loop)

Die KI führt bestimmte Aufgaben autonom aus, wobei Menschen in Ausnahmefällen eingreifen.

Beispiel: KI-gestützte Fertigungsprozesse, bei denen das System Maschinen steuert, aber Nutzende bei Anomalien eingreifen.

Empfohlene Anwendungsfälle: Szenarien, in denen eine kontinuierliche menschliche Beteiligung nicht erforderlich ist, kritische Risiken jedoch eine menschliche Überwachung erfordern (z.B. industrielle Automatisierung, Überwachung von Finanztransaktionen).

Stufe 4: Hoher Autonomiegrad (Human in Control)

Das KI-System arbeitet weitgehend autonom, erlaubt es den Benutzern jedoch, es selbst zu übersteuern, um unerwünschte Ergebnisse zu vermeiden.

Beispiel: Autonome Fahrzeuge

Empfohlene Anwendungsfälle: Umgebungen mit geringem bis mittlerem Risiko (z.B. Logistik, einfaches Verkehrsmanagement).

Stufe 5: Vollständige Autonomie mit minimaler Aufsicht

Das KI-System arbeitet unabhängig und erfordert nur minimale oder gar keine menschliche Intervention. Die Beteiligung des Menschen beschränkt sich auf die langfristige Aufsicht (Audits).

Beispiele: Autonome landwirtschaftliche Maschinen, KI für die Stromnetzverteilung, U-Bahnen, Flughafenbahnen

Empfohlene Anwendungsfälle: Umgebungen mit geringen Sicherheits- oder ethischen Risiken und hoher Zuverlässigkeit des KI-Systems (z. B. sich wiederholende Aufgaben in kontrollierten Umgebungen).

Anwendungsbereich-Risiko

Stufe 1: Minimales Risiko

Das KI-System hat keine Auswirkungen auf den Benutzer:innen oder die Entscheidungsfindung.

Beispiele: Filter, NPCs, Empfehlungsalgorithmen ohne schwerwiegende Folgen (DeepL, andere Übersetzungssysteme)

Kriterien: Keine direkte Auswirkung auf die Hochrisikobereichen des ***EU AI-Acts***.

Stufe 2: Begrenztes Risiko

KI-Systeme, die mit Benutzer:innen interagieren, aber keine Entscheidungen mit hohen Risiken treffen. Das Risiko steigt, wenn es an Transparenz über die Beteiligung von KI mangelt.

Beispiele: Chatbots und KI-generierte Inhalte ohne Offenlegung, einfache Automatisierungsaufgaben.

Kriterien: Bereiche, die nicht in der Liste der „hohen Risiken" des EU AI-Acts enthalten sind.

Stufe 3: Mittleres Risiko

KI-Systeme haben keine besonderen Auswirkungen auf einzelne Personen, aber sie entfalten Wirkung auf kollektiver oder gesellschaftlicher Ebene.

Beispiele: Generative KI wie ChatGPT und andere Systeme, die indirekt die Umgebung beeinflussen können, in der sie eingesetzt werden.

Kriterien: KI-Systeme, die für die öffentliche Nutzung verfügbar sind und das Potenzial haben, bestehende Gepflogenheiten zu beeinflussen und langfristig zu verändern.

Stufe 4: Hohes Risiko

Jeder Algorithmus, der in den laut EU AI-Act „Hochrisikobereichen" angewendet wird oder direkte Auswirkungen auf einzelne Personen hat.

Beispiele: Medizin, Biometrie, kritische Infrastruktur, Bildung und Berufsausbildung, Beschäftigung, Zugang zu Dienstleistungen im öffentlichen Sektor, Strafverfolgung, Migration.

Kriterien: Zugehörigkeit zum „Hochrisikobereich" des EU AI-Acts, nur wenn die Regeln für Transparenz und Datenqualität eingehalten werden.

Stufe 5: Extremes Risiko

Jeder Algorithmus, der in den laut EU AI-Act „Hochrisikobereichen" angewendet wird.

Beispiele: Medizin, Biometrie, kritische Infrastruktur, Bildung und Berufsausbildung, Beschäftigung, Zugang zu Dienstleistungen im öffentlichen Sektor, Strafverfolgung, Migration.

Kriterien: Zugehörigkeit zum „Hochrisikobereich", wenn die Regeln für Transparenz und Datenqualität NICHT eingehalten werden.

Quellen

1 OECD. (2024). Using AI to manage minimum income benefits and unemployment assistance: Opportunities, risks and possible policy directions (OECD Artificial Intelligence Papers No. 21; OECD Artificial Intelligence Papers, Vol. 21). https://www.oecd.org/en/publications/using-ai-to-manage-minimum-income-benefits-and-unemployment-assistance_718c93a1-en.html

2 Schwarscz, A. (2021). Use of big data and AI in fighting corruption and misuse of public funds—Good practice, ways forward and how to integrate new technology into contemporary control framework.

3 Bundesministerium Finanzen. (n.d.). Predictive Analytics Competence Center. Retrieved 15 July 2025, from https://www.bmf.gv.at/themen/betrugsbekaempfung/einheiten-betrugsbekaempfung/Predictive-Analytics-Competence-Center.html

4 Amnesty International. (2024, November 12). Denmark: AI-powered welfare system fuels mass surveillance and risks discriminating against marginalized groups – report. Amnesty International. https://www.amnesty.org/en/latest/news/2024/11/denmark-ai-powered-welfare-system-fuels-mass-surveillance-and-risks-discriminating-against-marginalized-groups-report/

5 Wihlborg, E., Larsson, H., & Hedström, K. (2016). 'The Computer Says No!' – A Case Study on Automated Decision-Making in Public Authorities. 2016 49th Hawaii International Conference on System Sciences (HICSS), 2903–2912. https://doi.org/10.1109/HICSS.2016.364

6 Turdibayeva, K. (2025, February 7). What is Automated Decision-Making? ProcessMaker. https://www.processmaker.com/blog/what-is-automated-decision-making/

7 Allhutter, D., Cech, F., Fischer, F., Grill, G., & Mager, A. (2020). Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective. Frontiers in Big Data, 3. https://doi.org/10.3389/fdata.2020.00005

8 Thapa, B. (2019). Predictive Analytics and AI in Governance: Data-driven government in a free society. The European Liberal Forum. https://liberalforum.eu/publication/predictive-analytics-and-ai-in-governance-data-driven-government-in-a-free-society

9 Hoffman, B. (2024, March 10). Automation Bias: What It Is And How To Overcome It. Forbes. https://www.forbes.com/sites/brycehoffman/2024/03/10/automation-bias-what-it-is-and-how-to-overcome-it/

10 AlgorithmWatch. (2019a). How Dutch activists got an invasive fraud detection algorithm banned. AlgorithmWatch. https://algorithmwatch.org/en/syri-netherlands-algorithm/

11 Dutch scandal serves as a warning for Europe over risks of using algorithms. (2022, March 29). POLITICO. https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/

12 Henley, J., & Booth, R. (2020, February 5). Welfare surveillance system violates human rights, Dutch court rules. The Guardian. https://www.theguardian.com/technology/2020/feb/05/welfare-surveillance-system-violates-human-rights-dutch-court-rules

13 Bhupatiraju, S., Chen, D., Jankin, S., Kim, G., Kupi, M., & Maqueda, M. R. (2023). Government Analytics Using Machine Learning.

14 Boulieris, P., Pavlopoulos, J., Xenos, A., & Vassalos, V. (2024). Fraud detection with natural language processing. Machine Learning, 113(8), 5087–5108. https://doi.org/10.1007/s10994-023-06354-5

15 Joshi, R., Nair, V., Singh, M., & Nair, A. (2021). Leveraging Natural Language Processing and Predictive Analytics for Enhanced AI-Driven Lead Nurturing and Engagement.

16 Santiso, C. (n.d.). Public Governance in the Age of Artificial Intelligence. Retrieved 1 July 2025, from https://www.chandlerinstitute.org/governancematters/public-governance-in-the-age-of-artificial-intelligence

17 AlgorithmWatch. (n.d.). DENMARK. AlgorithmWatch. Retrieved 1 September 2025, from https://algorithmwatch.org/en/automating-society-2019/denmark/

18 AlgorithmWatch. (2019b). NETHERLANDS. AlgorithmWatch. https://algorithmwatch.org/en/automatingsociety-2019/netherlands/

19 Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. Nature Human Behaviour, 8(12), 2293–2303. https://doi.org/10.1038/s41562-024-02024-1

20 Kölem, Ö. Interview über Predictive Analytics in PACC. Interview by Ilya Faynleyb, 19. Aug. 2025. Personal interview.

21 Kumar, V., & L., M. (2018). Predictive Analytics: A Review of Trends and Techniques. International Journal of Computer Applications, 182(1), 31–37. https://doi.org/10.5120/ijca2018917434

22 Stadt Wien. (n.d.). Förderhandbuch. Retrieved 27 August 2025, from https://www.wien.gv.at/spezial/foerderhandbuch/allgemeines-zu-forderungen/begriff-der-forderung-was-ist-eine-forderung/

23 Bundeszentrale für politische Bildung. (n.d.). Transferzahlungen. bpb.de. Retrieved 27 August 2025, from https://www.bpb.de/kurz-knapp/lexika/lexikon-der-wirtschaft/20866/transferzahlungen/

24 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA Relevance), 119 OJ L (2016). http://data.europa.eu/eli/reg/2016/679/oj/eng

25 Article 4: AI literacy | EU Artificial Intelligence Act. (n.d.). Retrieved 5 September 2025, from https://artificialintelligenceact.eu/article/4/

26 Art. 22 GDPR – Automated individual decision-making, including profiling. (n.d.). General Data Protection Regulation (GDPR). Retrieved 5 September 2025, from https://gdpr-info.eu/art-22-gdpr/

27 Unternehmensberatung, A. (n.d.). § 77 DSGVO (Datenschutz-Grundverordnung)—JUSLINE Österreich. Retrieved 5 September 2025, from https://www.jusline.at/gesetz/dsgvo/paragraf/77

28 Datenschutzbehörde, Ö. (n.d.-b). Österreichische Datenschutzbehörde. Österreichische Datenschutzbehörde. Retrieved 5 September 2025, from https://dsb.gv.at/

29 Allhutter, D., Alushi, A., Cavalcanti De Alcântara, R., Männiste, M., Pentzold, C., & Sosnowski, S. (2024). Public value in the making of automated and datafied welfare futures. Internet Policy Review, 13(3). https://doi.org/10.14763/2024.3.1803

30 Volksanwaltschaft—Beschwerdewegweiser. (n.d.). Retrieved 7 September 2025, from https://volksanwaltschaft.gv.at/fuer-probleme-mit-behoerden/beschwerdewegweiser/

31 Art. 15 DSGVO – Auskunftsrecht der betroffenen Person. (n.d.). Datenschutz-Grundverordnung (DSGVO). Retrieved 7 September 2025, from https://dsgvo-gesetz.de/art-15-dsgvo/

32 Macht Digitalisierung den Sozialstaat gerechter? (2024, May 6). scilog - Das Magazin des Wissenschaftsfonds FWF. https://scilog.fwf.ac.at/magazin/macht-digitalisierung-den-sozialstaat-gerechter

33 KI-Servicestelle der RTR. (n.d.). RTR. Retrieved 16 September 2025, from https://www.rtr.at/rtr/service/ki-servicestelle/ki-servicestelle.de.html

34 Petheram, A. (2019, September 19). Predictive Analytics, Public Services and Poverty. Oxford Insights. https://oxfordinsights.com/insights/predictive-analytics-public-services-and-poverty/

35 Sartor, G. (2025). Artificial Intelligence for Monitoring the Application of EU Law

36 Mehdiyev, N., Houy, C., Gutermuth, O., Mayer, L., & Fettke, P. (2021). Explainable Artificial Intelligence (XAI) Supporting Public Administration Processes – On the Potential of XAI in Tax Audit Processes. In F. Ahlemann, R. Schütte, & S. Stieglitz (Eds), Innovation Through Information Systems (pp. 413–428). Springer International Publishing. https://doi.org/10.1007/978-3-030-86790-4_28

Projekt ALAIT

Das Austrian Lab for AI Trust (ALAIT) ist ein vom österreichischen Bundesministerium für Innovation, Mobilität und Infrastruktur (BMIMI) initiiertes Forschungs- und Entwicklungs-Projekt zur Schaffung von Vertrauen durch Wissen im Bereich Künstliche Intelligenz (KI). Das Projekt ALAIT zielt darauf ab, Interessierte und wichtige gesellschaftliche Gruppen zu befähigen, KI-Technologien verantwortungsvoll zu nutzen und ethische sowie qualitativ hochwertige Standards für den Einsatz von AI zu etablieren.

Das Projekt wird von **winnovation** geleitet (Gertraud Leimüller und Lena Müller-Kress) und im Konsortium mit **leiwand.ai** (Rania Wazir und Silvia Wasserbacher-Schwarzer), **TU Wien** (Sabine Köszegi und Ilya Faynleyb) und **Austria Presse Agentur – APA** (Verena Krawarik und Sophia Marecek) umgesetzt.

Die ALAIT-Dossiers sind auf der Projekthomepage abrufbar: https://science.apa.at/project/alait/

Die Inhalte des Dossiers entsprechen dem aktuellen Stand der Technik und wurden sorgfältig nach wissenschaftlichen Kriterien erstellt. Sie dienen jedoch nicht als rechtsverbindliche Auskunft oder Beratung.

Impressum

Medieninhaberin und Herausgeberin:
winnovation consulting gmbh
Linke Wienzeile 42/1, Top 5
1060 Vienna

Danksagung:

Wir danken folgenden Expert:innen für ihr hilfreiches Feedback zu Vorversionen dieses Dossiers:
Doris Allhuter