Lernende Systeme

DIE PLATTFORM FÜR KÜNSTLICHE INTELLIGENZ

KI-Systeme schützen, Missbrauch verhindern

Maßnahmen und Szenarien in fünf Anwendungsgebieten

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

acatech
DEUTSCHE AKADEMIE DER TECHNIKWISSENSCHAFTEN

WHITEPAPER

Beyerer, J. & Müller-Quade, J. et al.
AG Lebensfeindliche Umgebungen
AG IT-Sicherheit, Privacy, Ethik und Recht

Inhalt

Zusammenfassung

Künstliche Intelligenz wird bereits in einer Vielzahl von gesellschaftlichen Bereichen eingesetzt, sei es im Gesundheitsbereich, in der Arbeitswelt, im Straßenverkehr oder in öffentlichen Räumen. Trotz der vielfältigen Chancen, die die KI-Technologie mit sich bringt, wie etwa eine verbesserte Gesundheitsversorgung oder eine attraktive, individuelle Arbeitsplatzgestaltung, sollte das Potenzial für den Missbrauch von KI-Systemen nicht aus den Augen verloren und realistisch eingeschätzt werden. Auf diese Weise können passende Maßnahmen frühzeitig und strategisch ergriffen werden, um Missbrauch entweder von Beginn an zu erschweren oder im konkreten und akuten Fall zu verhindern. Dies, um letztendlich damit das Vertrauen in die Zuverlässigkeit und Sicherheit von KI-Systemen sicherzustellen und zu stärken.

Expertinnen und Experten der Arbeitsgruppen Lebensfeindliche Umgebungen sowie IT-Sicherheit, Privacy, Recht und Ethik der Plattform Lernende Systeme nehmen sich ganz gezielt des Themas Missbrauch im Whitepaper an und stellen geeignete Maßnahmen vor, wie dem Missbrauch von KI-Systemen wirksam vorgebeugt werden kann. Dies vorrangig unter technologischen Aspekten. Damit liefern sie einen grundsätzlichen Ansatz, sich offen mit dem Thema Missbrauch beim Einsatz von KI-Technologien auseinanderzusetzen. Hierzu empfehlen sie, den Blick unverstellt auf die Schwachstellen zu richten, die diese Technologie mit sich bringen kann, und sich dabei zugleich mögliche Motive sowie Täterperspektiven vor Augen zu halten; dies vor dem Hintergrund des jeweiligen Anwendungskontextes beim Einsatz der KI-Anwendung. Aus dieser Gesamtsicht lassen sich erforderliche Schutzmaßnahmen ableiten, die eingebettet in einer Gesamtstrategie Missbrauch verhindern können.

Zur Konkretisierung und Veranschaulichung werden die theoretischen Überlegungen zudem in realistische Anwendungsszenarien eingebettet. Anhand von sieben Szenarien aus dem Bereich Gesundheit, Freizeit, Mobilität oder Arbeitswelt zeigen die Autorinnen und Autoren exemplarisch auf, wo Einfallstore für Missbrauch vorliegen könnten und wie der Worst Case durch geeignete und effektive Schutz- und Abwehrmechanismen in einen Best Case bereits von Beginn an umgeleitet und damit letztlich verhindert werden kann.

Um ein grundlegendes Verständnis für Missbrauch im Zusammenhang mit KI-Technologie zu schaffen, gehen die Autorinnen und Autoren zunächst der Frage nach, was Missbrauch von Technologien – und im Speziellen von KI-Technologien – überhaupt ist (Kapitel 2.1). Missbrauch, und damit auch die Ausführung des Missbrauchs, lässt sich als „Zweckentfremdung mit negativen Folgen" spezifizieren, bei dem fundamentale Werte, wie körperliche und psychische Unversehrtheit, (demokratische) Freiheiten und Rechte, Privatheit oder auch materielle und immaterielle Werte und die Umwelt verletzt werden. Auslöser des Missbrauchs ist menschliches Handeln unterschiedlicher Akteure, wie Kriminelle oder Terroristen, deren Motor verschiedenste Motive sein können.

Grundsätzlich sind zuvorderst folgende Fragen zu beantworten: Was ist zu schützen? Welche Angriffsszenarien sind denkbar? Welche Schutzmaßnahmen sind möglich, angemessen und zulässig? Basierend auf diesen Antworten und den daran anschließenden Analysen lassen sich konkrete Schutzmaßnahmen ableiten (Kapitel 2). Dabei empfiehlt es sich auch, die verschiedensten Motive, Tätertypen sowie deren Angriffsziele genauer zu analysieren (Kapitel 2.2). Dieser Perspektivenwechsel, sich in die Rolle der Tätertypen hineinzuversetzen und zu antizipieren, wie diese versuchen könnten, den Missbrauch zu ihrem eigenen (finanziellen) Vorteil oder sogar als Waffe zu nutzen, ermöglicht es, konkrete Schutzziele hinter den Angriffszielen oder -szenarien zu erschließen (Kapitel 2.3). Daraus lassen sich im Weiteren verschiedenste Schutz- und Abwehrmaßnahmen

– technologische wie organisatorische – für einen zuverlässigen Schutz von KI-Systemen ableiten (Kapitel 2.4), die sowohl auf Systemen mit Künstlicher Intelligenz wie Anomalieerkennung, Video- und Spracherkennung oder Identitätserkennung als auch auf Systemen ohne Künstliche Intelligenz beruhen können.

Die theoretischen Überlegungen zum Missbrauch und zu möglichen Schutzmaßnahmen werden in Kapitel 3 anhand konkreter Anwendungsfälle wirklichkeitsnah dargestellt. Somit wird ein Einblick vermittelt, wie und wo Missbrauch schon heute oder in kommenden Jahren Realität werden könnte und welche konkreten und geeigneten Maßnahmen jeweils abzuleiten wären.

Mit dieser Fokussierung auf das Thema Missbrauch von KI-Systemen leistet das Paper einen wichtigen Beitrag, sich gegen Missbrauchsversuche zu wappnen. Auch wenn eine Gesamtstrategie für geeignete Schutz- und Abwehrmaßnahmen seitens des Papers nicht geleistet werden kann, eröffnen die vorgestellten Grundsatzüberlegungen sowie konkreten Maßnahmen Perspektiven für reale Handlungsmaßnahmen. Damit wird zugleich ein Orientierungsrahmen vorgegeben, Missbrauch von KI-Systemen strategisch wie methodisch anzugehen mit dem Ziel, Missbrauch effektiv vorzubeugen und entgegenzuwirken.

Auf der Website der Plattform Lernende Systeme (https://www.plattform-lernende-systeme.de/missbrauch-von-ki-systemen.html) finden sich Anwendungsszenarien, die den Missbrauch von KI-Systemen sichtbar machen und dazu einladen, sich interaktiv mit diesem Thema auseinanderzusetzen. Dabei zeigen die Szenarien in der Gegenüberstellung „worst case“ versus „best case“ auf, wie durch geeignete Maßnahmen missbräuchlichen Angriffen vorgebeugt werden kann.

1. Einleitung

KI-Systeme haben das Potenzial, auf vielfältige Weise zu innovativen Entwicklungen in unterschiedlichen Branchen und Anwendungsbereichen beizutragen und auch unseren Alltag zu erleichtern. Beispiele dafür sind autonome Fahrzeuge, die in Deutschland ab 2022 am Straßenverkehr teilnehmen können (DPA, 2021), autonome Drohnen(-schwärme), die bei der Zustellung von Paketen oder bei der Wiederaufforstung von Flächen unterstützen können (Kort, 2020; Niemann, 2020), oder auch medizinische KI-Assistenzsysteme, die beispielsweise bei der Auswertung von Computertomographie-Scans unterstützen können (Plattform Lernende Systeme, 2019a).

Die KI-Systeme selbst können jedoch sowohl Ziel eines Missbrauchsversuchs werden als auch als Mittel zur Realisierung von Missbrauchsversuchen eingesetzt werden (siehe Infobox, S. 6, und vgl. exemplarisch Brundage et al., 2018; Cladwell et al., 2020; UNICRI & UNOCT, 2021). So wurden bereits Flugdrohnen zur selbstständigen Lieferung von Drogen eingesetzt (BBC, 2015) oder das eingangs erwähnte autonome Fahrzeug könnte beispielsweise als „fahrende" Waffe missbraucht werden (siehe Kapitel 3.1).

Die möglichen Folgen des Missbrauchs von KI-Systemen haben eine besondere Tragweite, da letztlich Entscheidungen von Algorithmen oder sogar des Menschen manipuliert werden können. Hinzu kommt, dass KI-Systeme häufig adaptionsfähig und stark vernetzt sowie teilweise in andere Systeme eingebettet sind (vgl. System of Systems (SoS)). Im Gegensatz zu herkömmlichen technischen Systemen vereinfachen und automatisieren KI-Systeme Prozesse nicht nur, sondern unterstützen darüber hinaus oft Entscheidungsprozesse des Menschen (z. B. im Umgang mit großen Datenmengen) oder befähigen sogar Systeme, selbst Entscheidungen zu treffen (z. B. in der Robotik).

Diese spezifischen Charakteristika von KI-Systemen in Kombination mit ihrer zunehmenden Anwendung und Verbreitung in vielen Teilbereichen der Gesellschaft machen den Schutz vor Missbrauch von KI-Systemen zu einem besonders relevanten Thema – sowohl heute schon als auch zukünftig. Deshalb zeigt das Whitepaper unterschiedliche Angriffsszenarien und Missbrauchspotenziale auf, um daraus vielfältige mögliche Schutzmaßnahmen abzuleiten. Diese verdeutlichen, dass durch deren Einsatz Missbrauch wirksam vorgebeugt und verhindert werden kann. Damit richtet es sich an die fachlich interessierte Öffentlichkeit, Bürgerinnen und Bürger, an die Hersteller und Anwendenden von KI-Systemen sowie politische Entscheidungsträgerinnen und Entscheidungsträger.

Das Whitepaper setzt zwei Schwerpunkte: Erstens wird vor allem auf Schutzmaßnahmen fokussiert, die bei einem Missbrauch ergriffen werden können. Denn Szenarien zum Schutz vor Missbrauch von KI-Systemen können sehr vielfältig sein. Folglich erhebt die Auswahl der Anwendungsszenarien in diesem Whitepaper keinen Anspruch auf Vollständigkeit oder Repräsentation aller denkbaren Fälle. Vielmehr dienen die vorgestellten Szenarien der Illustration von Maßnahmen, um den Missbrauch in unterschiedlichen Anwendungsdomänen wie Gesundheit, Freizeit, Mobilität oder im Unternehmens- und Arbeitskontext zu verhindern. Zweitens werden vorrangig technologische Aspekte zur Verhinderung von Missbrauch betrachtet. Gleichzeitig weisen die Autorinnen und Autoren darauf hin, dass der Schutz vor Missbrauch von KI-Systemen stets in eine ganzheitliche Strategie, die neben technischen Maßnahmen auch organisatorische umfasst, eingebettet sein muss. In diesem Papier werden einzelne, allgemeine Komponenten vorgestellt, die Teil einer solchen Gesamtstrategie sein können. Die Erarbeitung und Aufstellung einer ganzheitlichen Strategie selbst wird nicht vorgenommen, da eine solche immer stark einzelfallabhängig (insbesondere bei organisatorischen Maßnahmen) und somit schwer generalisierbar ist, was über die Intention des Papiers hinausgeht.

2. Schutz vor Missbrauch von KI-Systemen: Grundlagen und Perspektiven

Bevor passende Schutzmaßnahmen vor Missbrauch von KI-Systemen identifiziert werden können, ist zunächst zu klären und zu definieren, was Missbrauch von KI-Systemen ist und was diesen kennzeichnet. Im Anschluss daran werden mögliche Motive unterschiedlicher Akteure dargelegt. Diese vorangestellten Überlegungen entfalten Perspektiven, die es erleichtern, im nächsten Schritt Schutzziele und mögliche Angriffsszenarien unter folgenden Fragestellungen zu betrachten: Was soll geschützt werden? Welche Angriffe auf das System sind denkbar? Welche Schutzmaßnahmen sind möglich, angemessen und zulässig? Aus den sich daran anschließenden Analysen werden in einem zweiten Schritt verschiedene technische und organisatorische Schutzmaßnahmen abgeleitet und vorgestellt.

2.1 Definition von Missbrauch

Der Missbrauch von KI-Systemen hängt eng mit weiteren Konzepten wie der Zweckentfremdung, dem Angriff oder der IT-Sicherheit zusammen. Im Folgenden wird der Missbrauch von KI-Systemen eingeordnet und abgegrenzt. IT-Sicherheitsmechanismen können die Zweckentfremdung oder den Missbrauch eines KI-Systems stark erschweren oder ganz verhindern, während mithilfe der Zweckentfremdung oder des Missbrauchs eines KI-Systems auch IT-Sicherheitsmechanismen umgangen werden können.[1]

„Missbrauch“ und „Zweckentfremdung“ sind keine Synonyme (siehe Infobox). Jeder Missbrauch ist eine Zweckentfremdung, aber nicht jede Zweckentfremdung ist ein Missbrauch. Kurzum: Missbrauch kann als „Zweckentfremdung mit negativen Folgen“ beschrieben werden. Angriffe können dazu dienen, den Missbrauch von KI-Systemen zu ermöglichen, und können auch mithilfe des Missbrauchs von KI-Systemen durchgeführt werden.

KURZINFO

Missbrauch von KI-Systemen

Missbrauch von KI-Systemen beschreibt zum einen die Nutzung eines KI-Systems entgegen dessen Zweck (siehe Zweckentfremdung) und zum anderen die Verletzung fundamentaler Werte, wie körperliche und psychische Unversehrtheit, (demokratische) Freiheiten und Rechte, Privatheit oder materielle und immaterielle Werte sowie die Umwelt.

Zweckentfremdung

Eine Zweckentfremdung von KI-Systemen beschreibt die Nutzung des Systems entgegen dessen Zweck – mit positiven oder negativen Folgen.

1 Nicht zuletzt kann KI sowohl in IT-Sicherheit als auch bei KI-Systemen selbst dazu beitragen, Missbrauch zu verhindern. Diese mehrdimensionale Beziehung ist auch bei KI-Systemen und IT-Sicherheit zu finden. Siehe weiterführend dazu: Müller-Quade et al., 2019.

Die aufgeführten Unterschiede zwischen Missbrauch und Zweckentfremdung lassen sich am Beispiel von KI-Systemen für die Auswertung von Bildern verdeutlichen. Solche Systeme werden bereits für die Auswertung von Videomaterial entwickelt, ebenso wie für die Auswertung von medizinischem Bildmaterial (z. B. Röntgenbilder) oder Satellitenbilder. Die Ergebnisse aus solchen Forschungs- und Entwicklungsprozessen können in einem Gebiet häufig auch in anderen Bereichen eingesetzt werden. Wird beispielsweise ein System entwickelt, das Abweichungen von „Normal-Bewegungen" in Videos erkennen kann, so könnte dieses im Sinne einer positiven Zweckentfremdung beispielsweise im Hochleistungssport eingesetzt werden. Ebenso kann eine KI-Technologie zur Auswertung von Videomaterial aber auch missbräuchlich eingesetzt werden, etwa zum Zweck unrechtmäßiger Überwachung von Minderheiten.

Positive Zweckentfremdung aufgezeigt am Beispiel Hochleistungssport:
Nutzung für einen anderen Zweck, beispielsweise zur Analyse der Bewegung von Hochleistungssportlerinnen und -sportlern. Das Trainingsteam einer Mannschaft könnte beispielsweise leichte Verletzungen einer Spielerin oder eines Spielers während eines Hand-, Basket- und Fußballspiels durch eine ausgewertete Video-Aufzeichnung angezeigt bekommen und diese Person aus dem Spiel nehmen, um sie für den nächsten Einsatz zu schonen.

Missbrauch aufgezeigt am Beispiel Überwachung von Minderheiten:
Missbrauch des KI-Systems zum Auffinden bzw. zur Überwachung von Minderheiten oder Dissidenten in autoritären und semi-autoritären Staaten in der Öffentlichkeit. So nutzt der chinesische Staat beispielsweise KI, um die muslimische Minderheit zu überwachen (vgl. Mozur, 2019; Feldstein, 2019).

2.2 Angreifende und ihre Motive

KI-Systeme können von ganz unterschiedlichen Tätertypen aus verschiedensten Gründen missbraucht werden (Hesse & Müller-Quade et al., 2021, S. 31). Tabelle 1 gibt einen Überblick über Täterinnen und Täter, ihre jeweilige Motivation für einen Missbrauch von KI-Systemen sowie über mögliche Ziele, die im Missbrauchsfall betroffen sein könnten.

Tabelle 1: **Missbrauch von KI-Systemen: Täter-Typ, Motive und Ziele**

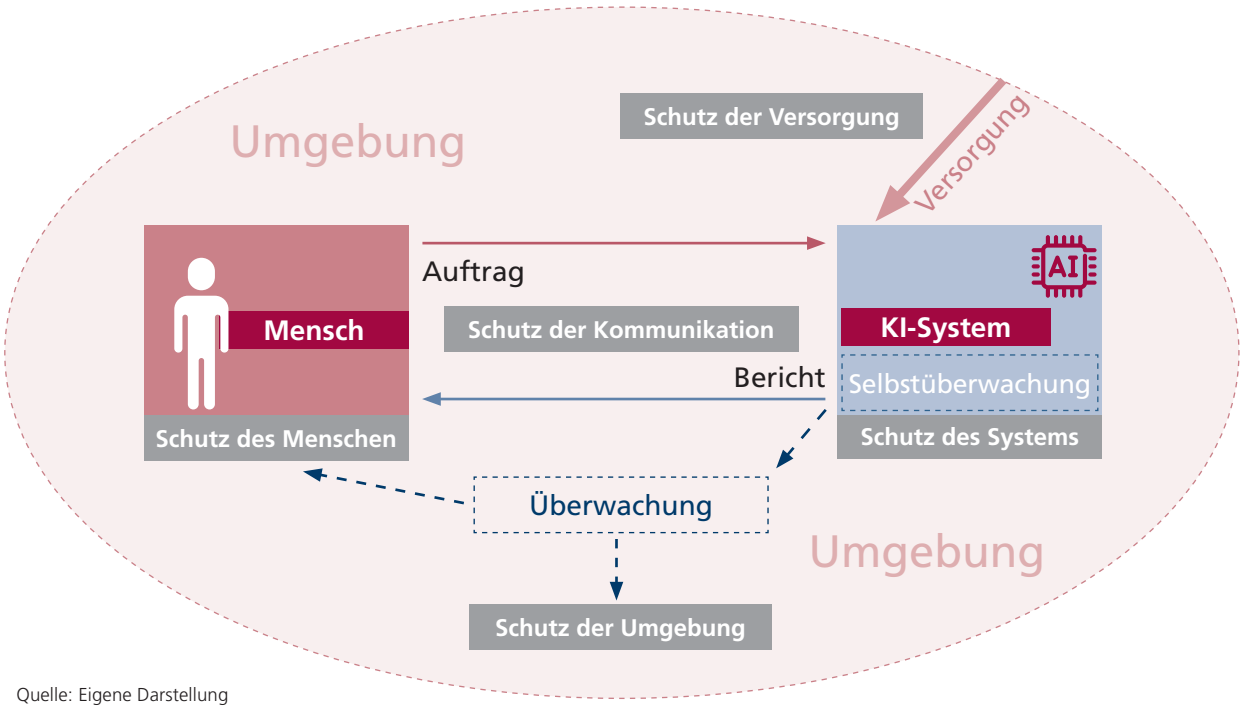
Täter-Typ	Motive	Ziele	Beispiele
(Ausländische) Staatliche Stellen	• Spionage • Sabotage • Politische Einflussnahme • Repression in autoritären Staaten	• Unternehmen • Öffentliche Einrichtungen • Infrastruktur • Politiker/-innen • Journalist/-innen • Dissident/-innen • Bürger/-innen • Minderheiten	• Überwachung von Minderheiten • Einflussnahme auf Wahlen mit Social-Media-Bots
Terroristen	• Anhängerschaft motivieren • Bewegung bewerben • Verunsicherung in der Bevölkerung verbreiten • Opponent/-innen beseitigen • Reaktionen provozieren	• Bürger/-innen • Menschenmengen • Bedeutende Veranstaltungen, Objekte und Einrichtungen	• Manipulation von autonomen Fahrzeugen • Manipulation von Drohnen
Wettbewerber	• Wettbewerbsvorteile erzielen oder Wettbewerbsposition unrechtmäßig verbessern (z. B. Industriespionage) • Schwächung von Sicherheitssystemen	• Andere Unternehmen	• Industriespionage (z. B. via Phishing oder durch Drohnen)
Kriminelle	• Monetäre Bereicherung (auch im Dienste Dritter)	• Unternehmen • Öffentliche Einrichtungen • Infrastruktur • Privatpersonen	• Erpressung von Unternehmen durch technische Täuschung (z. B. Spear-Phishing)
KI-nutzende Organisation oder Einrichtung	• Kontrolle zur Optimierung von Organisationszielen • Verhaltensbeeinflussung (z. B. zur Absatzförderung)	• Angestellte • Mitglieder • Bevölkerung	• Durchgängige individuelle Leistungsüberwachung von Angestellten in Unternehmen

Quelle: Zusammenstellung in Anlehnung an Hesse & Müller-Quade et al. (2021, S. 31); Thornton (1964) und BKA (2021, S. 29).

2.3 Schutzziele und Angriffsszenarien

Maßnahmen, um KI-Systeme vor potenziellem Missbrauch zu schützen, sind vielfältig und system- sowie situationsabhängig. Sie hängen sowohl von den Schutzzielen als auch von den Angriffsszenarien ab. Beide Aspekte sind wiederum von weiteren Faktoren abhängig: So steht ein KI-System nicht für sich allein, sondern ist stets in seinen Anwendungskontext eingebunden (siehe Abbildung 1). Hierzu zählen sowohl der Mensch, der den allgemeinen Auftrag an das System gibt und eine Rückmeldung über die Ausführung des Auftrags erhält (Bericht), als auch die Umgebung des KI-Systems selbst sowie die der Umgebung angehörigen vorliegenden Kommunikations-, Versorgungs- und Überwachungskanäle. All diese Komponenten stellen mögliche Angriffsziele dar und bedürfen deshalb eines besonderen Schutzes.

Abbildung 1: **Allgemeine Maßnahmen zum Schutz von KI-Systemen gegen Missbrauch**



Quelle: Eigene Darstellung

Schutzziele

Für die Einrichtung und Ergreifung von wirksamen Schutzmaßnahmen ist es essenziell, in einem ersten Schritt die zu schützenden Aspekte, Komponenten oder Bereiche zu identifizieren. Mögliche Schutzziele sind das KI-System (inklusive der enthaltenen Daten[2]), die Mission (also der Auftrag) des KI-Systems sowie die Umgebung des KI-Systems. Diese drei Haupt-Schutzziele umfassen jeweils weitere Unter-Schutzziele (siehe Abbildung 2).

Die vorgestellten Schutzziele können sowohl einzeln als auch in Kombination miteinander betrachtet werden. Grundsätzlich gilt: Es sollte am besten das geschützt werden, wodurch bei einem Missbrauchsfall die stärksten Schäden oder gefährlichsten Folgen entstehen können. Hierbei sind auch Angriffswahrscheinlichkeiten und die jeweils vorliegende Situation mit in Betracht zu ziehen.[3]

2 Es sind personenbezogene, personenbeziehbare und anonymisierte Daten zu unterscheiden: Unter personenbezogenen Daten werden „alle Informationen, die sich auf eine identifizierte oder identifizierbare natürliche Person (betroffene Person) beziehen" (vgl. BMJV, 2018), verstanden. Personenbeziehbare Daten können einer natürlichen Person zugeordnet werden. Beispiel: Bewohner im Einfamilienhaus in der Hauptstraße 1 in Werder (Havel). Sind Daten anonymisiert, kann eine betroffene Person nicht oder nur mit unverhältnismäßigem Aufwand identifiziert werden (vgl. Hesse & Müller-Quade et al., 2021, S. 21).

3 Beispiele für Täter-Typen und Motive in der Mobilität siehe Hesse & Müller-Quade et al. (2021, S. 31).

Abbildung 2: **Beispiele für Schutzziele**

Schutzziele		
KI-System	Mission	Umgebung
• Hardware (ganzes System oder Komponenten) • Datenbanken (inkl. gelernter Daten, Trainingsdaten) • Software (inkl. Algorithmen)	• Gesamte Mission • Teile der Mission • Zielwerte • Zwischenergebnisse	• Gesamte Umgebung • Bestimmte Situation oder Bereiche, z. B. andere KI-Systeme oder körperliche und psychische Unversehrtheit der Menschen • Rückkopplungen

Angriffsszenarien

Die unterschiedlichen Schutzziele von KI-Systemen können in verschiedenen Dimensionen angegriffen und/oder manipuliert werden und so zu einem realen Ziel eines Angriffs werden.[4] Ziel des Angriffs auf die Schutzziele ist es, das Gesamtsystem oder Teilsysteme zu manipulieren und dadurch Schaden zu verursachen. Abbildung 3 gibt einen Überblick über unterschiedliche Angriffsszenarien.

Abbildung 3: **Beispiele für Angriffsszenarien**

Angriffsszenarien		
KI-System	Mission	Umgebung
Änderung der Systemfähigkeiten durch Manipulation **Hardware** • Beschädigung oder Diebstahl des Systems oder seiner Komponenten • Störung und Manipulation von Komponenten • Austausch des Systems **Software und Daten** • Diebstahl • Manipulation • Austausch	• Verzögerungen, Unterbrechungen, Manipulation des Auftrages • Störung der Verbreitung oder Ausführung eines Auftrages • Unterbrechung der Versorgung oder Kommunikation • Störung der Auswertung bzw. Nachbereitung eines Auftrages	• Manipulation der Einsatzumgebung • Einwirken auf Menschen

2.4 Schutzmaßnahmen

Angriffsziele und -techniken können sehr unterschiedlich sein und es können mehrere Angriffe parallel stattfinden. Daher wird ein breites Spektrum an Schutz- und Abwehrmaßnahmen benötigt, um KI-Systeme zuverlässig zu schützen. Diese Maßnahmen sollten am System selbst, an dessen Umgebung sowie an den Menschen, die das System beauftragen und kontrollieren, ansetzen. Neben angemessenen, allgemeinen Schutzmaßnahmen (siehe Infobox) sollten in Abhängigkeit des Anwendungskontextes weitergehende Maßnahmen zum Schutz von KI-Systemen ergriffen werden.

4 Informationen zu Angriffszielen sind Anhang 1 zu entnehmen.

KURZINFO

Allgemeine Maßnahmen

Dazu gehören vor allem allgemeine Schutz- und Gegenmaßnahmen, die für eine Verhinderung oder Abwehr diverser Angriffe bzw. anderer Missbrauchsszenarien bei modernen IT-Systemen üblich sind: Kapselung, Firewalls, Antivirenprogramme, Sandboxes, ID- und Accountmanagement, Updates, Verschlüsselung, Schutz der Kommunikation etc. Zudem ist es wichtig, alle wesentlichen Aufträge, Situationen und Handlungen ausreichend zu protokollieren (sog. Logging), um einerseits nachträgliche Analysen und Verbesserungen durchführen zu können, und andererseits um Wiederholungen bekannter Angriffs- und Missbrauchsszenarien effizient zu unterbinden. Auch sind allgemeingültige Regelungen und Anweisungen im Umgang mit KI-Systemen notwendig, um Missbrauch zu verhindern. Beispiele sind das Datenschutz- und das Arbeitsrecht (siehe hierzu: Organisatorische Maßnahmen, S. 14).

Wie bei konventionellen IT-Systemen ist der Mensch auch bei KI-Systemen stets involviert, sei es als kontrollierende Instanz, als Interaktionspartner oder als Entwickler. Daher sollten Schutzmaßnahmen sowohl auf technischer als auch auf organisatorischer Ebene umgesetzt werden. Im Idealfall werden solche Maßnahmen ineinandergreifend eingesetzt, um ein möglichst hohes Schutzniveau zu erreichen. Während einige der nachfolgend dargestellten Maßnahmen lediglich nur einmal bei der Entwicklung eines KI-Systems durchgeführt werden müssen, wie die Implementation einer KI-gestützten Erkennung von Anomalien, bedürfen andere einer regelmäßigen Wiederholung sowie einer stetigen Optimierung, wie die Absicherung des Lernprozesses des KI-Systems und dessen Datenbasis. Bei der Festlegung auf Maßnahmen zum Schutz vor Missbrauch sollte in Abhängigkeit des jeweils vorliegenden Anwendungsfalls[5] immer auch das Verhältnis von Kosten und Nutzen der Maßnahmen betrachtet werden.

Technische Maßnahmen

Maßnahmen zum Schutz von KI-Systemen gegen Missbrauchsversuche können sowohl auf Systemen mit Künstlicher Intelligenz als auch auf Systemen ohne Künstliche Intelligenz beruhen (vgl. Müller-Quade et al., 2019). Im Folgenden wird eine Reihe von Maßnahmen vorgestellt, die bei der Entwicklung, Herstellung und Anwendung ergriffen werden können, um KI-Systeme wirksam vor Missbrauchsversuchen zu schützen. Zum einen basieren diese Maßnahmen schwerpunktmäßig auf KI-Systemen wie Anomalieerkennung, Video- und Spracherkennung und Identitätserkennung. Zum anderen werden Maßnahmen behandelt, die für den Schutz von KI-Systemen (im Vergleich zu traditionellen IT-Systemen) besonders relevant sind, wie Absicherung der Datenbasis und des Lernprozesses sowie Einschränkung der Funktionalitäten und Fähigkeiten.

5 Die Art und der Umfang an Schutzmaßnahmen hängt auch vom Autonomiegrad des KI-Systems und dessen potenziellem Schadem im Missbrauchsfall ab. Insbesondere bei niedrigen Autonomiegraden kann der Mensch als steuernde oder kontrollierende Instanz viele der beschriebenen Angriffsmöglichkeiten unterbinden, da dieser kritische Situationen oder Vorgänge erkennen kann. Bei hohen Autonomiegraden kann der Mensch als Supervisor und Beobachter bei kritischen Situationen und Vorgängen nur dann rechtzeitig eingreifen, wenn das Systemverhalten genau beobachtet wird.

KI-gestützte Erkennung von Anomalien implementieren

Die Erkennung und die Blockierung von Angriffen und Missbrauchsversuchen kann generell durch die KI-gestützte Detektion von Anomalien erfolgen (Müller-Quade et al., 2019, S. 8), wie Abweichungen vom Umgebungsmodell beim autonomen Fahren oder Abweichungen von einem geplanten Verlauf eines Einsatzes. Handlungen ähnlicher KI-Systeme können beispielsweise in einer Cloudumgebung gesammelt werden. Auf diese Weise könnte abgeglichen werden, ob diese Handlungen bereits in ähnlichen Situationen vorkamen. Verdächtige bzw. ungewöhnliche Anfragen und Handlungen könnten so ermittelt und anschließend geprüft werden (z. B. durch Simulation in Cloudumgebungen, die erwartete Ergebnisse, wie System- oder Umweltveränderungen, analysieren). Die gesammelten Daten können wiederum in Lernprozesse überführt werden, um die Erkennung von Anomalien zu optimieren.

Regelsätze zur Erkennung und Verhinderung von Regelverstößen implementieren

Verbotene, schädliche oder gefährliche Handlungen eines KI- Systems können unter anderem dann erkannt werden, wenn Regelsätze in KI-Systeme implementiert werden. Durch den Abgleich des aktuellen Status des KI-Systems bzw. der vorliegenden Situation mit der implementierten Regel erkennt das System den Verstoß und kann Gegenmaßnahmen einleiten (z. B. Abschalten des Systems). So könnte zum Beispiel die Regel durchgesetzt werden, dass alle Handlungen untersagt sind, bei denen eine Person berührt oder verletzt wird. Grundsätzlich dürfen Regelsätze eines KI-Systems nicht durch sein Lernen direkt geändert werden. Genauso wenig dürfen unbefugte oder unbeabsichtigte Änderungen der Regelsätze eintreten. Vielmehr müssen diese technisch abgesichert werden. Sofern Änderungen notwendig sind, sollten diese zumindest durch ein „Vier-Augen-Prinzip“ von Menschen eingeleitet und freigegeben werden.

Systeme implementieren, die Identitäten eindeutig erkennen

Die Zusammenarbeit von Mensch und KI-System kann durch die Stimme, Gesten oder weitere Biosignale erfolgen. KI-Systeme müssen autorisierte Personen sicher erkennen (vgl. Müller-Quade et al., 2019, S. 8). Dies sollte auch dann gewährleistet sein, wenn sich das äußere Erscheinungsbild der Person verändert, etwa durch Frisur, Schmutz, Bekleidung etc. Besonderen Schutz vor unautorisierten Befehlen verspricht eine Kombination aus KI-basierten Video- und Spracherkennungssystemen und weiteren Authentifizierungsmaßnahmen, etwa eine Multi-Faktor-Authentifizierung (siehe Infobox). Für solche Authentifizierungsverfahren können biometrische Merkmale herangezogen werden (z. B. Gesicht, Iris, Stimme) oder Verfahren wie die gegenseitige Verhaltensanalyse (z. B. Prüfung des Tippverhaltens) oder Frage-Antwort-Verfahren, in denen zuvor hinterlegte Antworten abgefragt werden (z. B. Lieblingsessen, Name des Haustiers etc.). Die Authentifizierung kann kontinuierlich erfolgen, um auf diese Weise auch den Wechsel von Personen zu erkennen, die mit relevanten Komponenten eines KI-Systems in Berührung kommen bzw. an der Entwicklung solcher Komponenten mitarbeiten.

KURZINFO

Multi-Faktor-Authentifizierung (MFA)

MFA ist ein Verfahren, mit dem die Zugangsberechtigung zu einem IT-System über die Abfrage mehrerer unterschiedlicher Merkmale geprüft wird. Solche Faktoren können Informationen sein, die nur die berechtigte Person kennt oder besitzt, z. B. biometrische Merkmale wie Fingerabdrücke oder der Ort, an dem sich die Person gerade befindet, etc.

Vorkehrungen zur Absicherung der Datenbasis implementieren

Es ist unerlässlich, dass die Qualität, die Authentizität und die Integrität der Trainingsdaten gewährleistet werden. So sollten Besitzer von Datenquellen (z. B. Unternehmen, Behörden, Krankenhäuser etc.) die Auswahl der Trainingsdaten klar dokumentieren. Darüber hinaus sollte die Auswahl der Daten für den Zweck des Systems angemessen sein und unabhängig geprüft werden können (z. B. durch eine Kreuzvalidierung). Zu den Trainingsdaten sollten nur autorisierte Personen neue Daten hinzufügen können (Müller-Quade et al., 2020, S. 20). Neue und vom KI-System gesammelte Trainingsdaten können im Falle einer Unterbrechung der Energiezufuhr oder Kommunikation und bei falscher Authentifizierung gelöscht oder als unsicher gekennzeichnet werden. Schließlich können durch verteiltes, maschinelles Lernen (siehe Infobox), Datensicherheit und -schutz sowie Privatsphäre gewahrt bleiben. Dafür müssen solche Lernverfahren entsprechend abgesichert sein.

KURZINFO

Verteiltes Lernen (auch föderiertes Lernen, Federated Learning)

Das verteilte Lernen ist ein Verfahren, das Modelle dezentral, d. h. nahe an den Datenquellen, trainiert (z. B. auf lokalen Geräten), sodass Rohdaten nicht ausgetauscht werden müssen. Lediglich die lokal trainierten Modelle werden an eine Cloud gesendet und dort kombiniert (Wirtz et al., 2019; Müller-Quade et al., 2019, S. 19).

Vorkehrungen zur Absicherung des Lernprozesses implementieren

Um den Lernprozess von KI-Systemen zu schützen, kann es unter anderem sinnvoll sein, deren Lernfähigkeit einzuschränken. Dies können beispielsweise ort- und situationsbezogene Einschränkungen sein oder auch eine starke Verlangsamung des Lernens zwischen Systemwartungsschleifen, da so Einflüsse von schädlichen Lernbeispielen auf das KI-System minimiert werden können. Schließlich kann das selbstständige Lernen eines KI-Systems eingeschränkt werden, sodass neu Erlerntes erst durch eine menschliche Instanz bestätigt werden muss. Eine solche Überprüfung des Gelernten kann auch eingeleitet werden, wenn durch das KI-System oder den Menschen ungewöhnlicher Dateninput erkannt wird. In diesem Fall könnten die Lernergebnisse zunächst an einem Modell geprüft werden, bevor sie an das KI-System übergeben werden.[6]

Funktionalitäten und Fähigkeiten einschränken

Die Funktionalitäten und Fähigkeiten von KI-Systemen können für bestimmte Orte, Zeitfenster, Situationen und Umgebungen deaktiviert bzw. es kann der Grad der Autonomie eines solchen Systems reduziert werden (vgl. „Kompetenzanalyse", Beyerer et al., 2021, S. 24ff.). Beispiele hierfür sind das Geofencing und -targeting, die sich an geographischen Koordinaten orientieren, um beispielsweise Drohnenflüge in spezifische Gebiete zu unterbinden oder Funktionen von Baumaschinen auf einer Baustelle zu beschränken. Einschränkungen von Fähigkeiten können auch gekoppelt an eine Umgebungsanalyse auf Basis der Sensoren eines mobilen Systems erfolgen. Das heißt, dass bestimmte Fähigkeiten nur dann freigegeben sind, wenn Sensoren spezifische Merkmale eines Ortes oder einer Umgebung erkennen. Solche Umgebungs- und Ortsanalysen stellen eine gute Alternative oder Ergänzung dar, da Maßnahmen wie das Geofencing und -targeting unter Umständen technisch umgangen werden können etwa durch „Jamming" oder „Spoofing" (siehe Infobox). Die Grundlage hierfür kann der Abgleich von Informationen aus unterschiedlichen Sensoren eines mobilen KI-Systems bilden.

6 Dies kann z. B. durch ein Ende-zu-Ende Lernen in einer Cloud umgesetzt werden.

KURZINFO

Jamming und Spoofing

Jamming bezeichnet eine Störung eines Kommunikationsvorgangs (z. B. die Datenübertragung im Mobilfunk). Spoofing bezeichnet Verschleierungs- oder Täuschungsmethoden in Computernetzwerken.

Organisatorische Maßnahmen

Damit technische Schutzmaßnahmen effizient wirken können, sollten sie durch organisatorische Maßnahmen ergänzt werden. In jeder Lebenszyklusphase eines KI-Systems, von Beginn bis zu dessen Ende (Design, Herstellung, Inbetriebnahme, Einsatz, Wartung, Aktualisierung, Entsorgung), können Angriffsflanken entstehen, die das Missbrauchspotenzial erhöhen. Dies wäre der Fall, wenn zum Beispiel in der Designphase Fehler oder „Hintertüren" in die Soft- und Hardware des Systems eingebracht werden oder Fehlverhalten von Personen zu Schwachstellen im KI-System führen. Diesem Risiko kann auf verschiedene Weise mit organisatorischen Maßnahmen entgegengewirkt werden.

Regeln und Prozesse definieren und einhalten

Werden für Prozesse und Vorgänge in den verschiedenen Phasen Regeln definiert und durchgesetzt, kann dies dem Entstehen von möglichen Angriffspunkten entgegenwirken. So sollten etwa Regeln für den Zugang zur Soft- und Hardware des KI-Systems definiert werden oder auch für den Zugang zur Einsatzumgebung und zu Prozessen für die Beobachtung der Einsatzumgebung sowie für die Umsetzung der Prinzipien „Security- und Safety-by-Design" (siehe Infobox). So stellen beispielsweise Betriebsvereinbarungen eine wichtige Möglichkeit und damit ein wirksames Instrument dar, um den Einsatz von und den Umgang mit KI-Systemen schon mit der Einführung von KI-Systemen im Einvernehmen zwischen Unternehmen und ihren Beschäftigten zu definieren (siehe Kapitel 3.5). Auch der Faktor Mensch als Einfallstor für Missbrauch kann reduziert werden, wenn das Bewusstsein für Missbrauchspotenziale in der jeweiligen Arbeitsumgebung geschärft wird und entsprechende Verhaltensregeln definiert werden, die es einzuhalten gilt.

KURZINFO

Security- und Safety-by-Design

Prinzipien, die darauf abzielen, dass schon in der Entwicklungsphase eines (KI-)Systems sowohl bei der Software als auch bei der Hardware auf Sicherheitsanforderungen geachtet wird, um dadurch künftige Angriffsflanken zu vermeiden. Während das Security-by-Design dabei auf die Kriminalitätsprävention zielt, steht beim Safety-by-Design die Unfallprävention im Mittelpunkt.

KI-Systeme und Prozesse prüfen und zertifizieren

Sowohl das KI-System selbst als auch organisatorische Prozesse sollten geprüft und/oder zertifiziert werden (siehe auch Produkt- und Prozesszertifizierung, Heesen, Müller-Quade & Wrobel, 2020). Gegenstand der Prüfung und Zertifizierung können sein: technische Eigenschaften des KI-Systems, Trainingsdaten,

aufgezeichnete Handlungen, Reaktionen und Verhaltensweisen des Systems zur Einsatzzeit, aber auch in Prüfszenarien, oder erlernte und veränderte Fähigkeiten sowie die Lernfähigkeit selbst. Prozesse wie die Herstellungs- und Wartungsprozesse sollten ebenfalls Teil von Prüfungs- und Zertifizierungsmaßnahmen sein.

Organisatorische Maßnahmen sind oft komplex sowie aufwendig in der Umsetzung und daher kostspielig. So erfordern Prüfungen und Zertifizierungen meist einen zusätzlichen Aufwand und spezifische technische Voraussetzungen wie beispielsweise spezielle Systeme zur Umgebungsanalyse. Sie sind jedoch ein gutes Instrumentarium, um Missbrauch vorzubeugen und gerade für KI-Systeme unumgänglich, die im Missbrauchsfall einen potenziell großen Schaden verursachen können.[7]

Sowohl technologische als auch organisatorische Maßnahmen zum Schutz vor Missbrauch können künftig im Rahmen eines Advanced Systems Engineering (ASE) systematisch integriert werden. Intelligente cyberphysische Systeme werden zukünftig immer stärker zum Einsatz kommen. Dabei handelt es sich um Systeme, deren mechanische Komponenten, Software und moderne Informationstechnik über Netzwerke (z. B. das Internet) miteinander verbunden sind. Es wird davon ausgegangen, dass diese sich künftig durch einen hohen Grad an dynamischer Vernetzung, Autonomie und interaktiver, soziotechnischer Integration auszeichnen. Daher werden zum einen Herangehensweisen des Engineerings notwendig, die durch eine ganzheitliche und interdisziplinäre Perspektive der zunehmenden Komplexität Rechnung tragen und zum anderen kontinuierlich neue technologische und arbeitsorganisatorische Entwicklungen einbeziehen. Diese Sichtweise ist im ASE-Leitbild verankert (vgl. Dumitrescu et al., 2021).

7 Für eine Gesamtübersicht über die vorgestellten Schutz-, Angriffsziele und -szenarien sowie mögliche Gegenmaßnahmen siehe Anhang 1 und Anhang 2.

3. Schutz vor Missbrauch von KI-Systemen: Konkrete Anwendungsszenarien

Das vorangegangene Kapitel hat gezeigt, dass ein Missbrauch von KI-Systemen auf unterschiedlichen Ebenen möglich ist – gleichzeitig gibt es aber eine Reihe von Maßnahmen, um solchen Bestrebungen wirksam zu begegnen und damit den Missbrauch zu verhindern. Die folgenden einzelnen, exemplarischen Szenarien aus den Anwendungsfeldern Gesundheit, Mobilität aus komplexen Einsatzumgebungen, wie sie etwa bei Hafenbecken oder Großveranstaltungen gegeben sind, sowie aus dem Unternehmens- und Arbeitskontext sollen die Anwendung von Schutzmaßnahmen der vorangestellten theoretischen Überlegungen an konkreten Beispielen aufzeigen. Hierfür wird ausgeführt, wie sich zum einen der Missbrauch eines KI-Systems im konkreten Anwendungsszenario äußern kann und zum anderen, welche Schutzmaßnahmen greifen, um diesen im jeweiligen Fall verhindern zu können. Dabei wird jeweils ein möglicher Entwicklungsverlauf mit unterschiedlichen, ausreichenden Schutzmaßnahmen (Best Case) einem möglichen Verlauf ohne ausreichende Schutzmaßnahmen (Worst Case) gegenübergestellt.

3.1 Gesundheit: KI-gestütztes Matching für die Organtransplantation

Dank des Fortschritts in der medizinischen Diagnostik können Ärztinnen und Ärzte heute bei der Behandlung auf eine Vielzahl an Daten zurückgreifen. So helfen beispielsweise molekulargenetische Untersuchungen, ein noch detaillierteres Verständnis für die komplexe Wirkweise des menschlichen Immunsystems zu erhalten. Bei einer Organtransplantation können Spendende und Empfangende noch besser aufeinander abgeglichen werden, um das Risiko für Abstoßungsreaktionen zu reduzieren. Bei der Transplantation von Nieren wird ein Lymphozytotoxizitätstest (LCT) durchgeführt, um zu prüfen, ob spezielle Humane Leukozyten-Antigene (HLA) auf dem Spenderorgan zu einer möglichen Abstoßung beim Empfänger oder bei der Empfängerin führen können.

In Szenario 1 wird ab dem Jahr 2030 zusätzlich bei Spendenden und Empfangenden das molekulargenetische Profil von Zellen des Immunsystems im Labor sequenziert: Aus den gewonnenen Informationen zum Erbgut erstellt ein KI-Algorithmus eine Prognose über das Risiko für Zellen des Empfänger-Immunsystems, bindungsfähige Antikörper gegen die HLA-Merkmale der Spenderniere zu bilden. Die Manipulation der Datenbank des Laborsystems stellt ein mögliches Missbrauchsszenario dar (siehe Infobox zu Adversarial-Machine-Learning).

SZENARIO 1

Manipulation eines Laborinformationssystems verhindern

Ausgangssituation

Ein Patient leidet an einer chronischen Niereninsuffizienz. Die Transplantation einer Spenderniere eröffnet ihm die Chance, ein Leben ohne Dialyse zu führen. Von seiner behandelnden Ärztin lässt er sich auf die Warteliste für Spenderorgane setzen. Für die molekulargenetische Untersuchung seines Immunsystems wird sein Blut ins Labor geschickt. Das Transplantationszentrum nutzt zur Entscheidungsfindung für eine Organtransplantation auch die Ergebnisse des Labor-internen KI-Systems zur Bestimmung des Abstoßungsrisikos.

Der Missbrauch

Der Patient beauftragt einen Hacker, seine Daten zu seiner verbleibenden Nierenfunktion sowie die zusätzlichen molekulargenetischen Daten im Labor so zu manipulieren, dass der KI-Algorithmus für diese eine möglichst hohe Kompatibilität zu aktuellen Spenderorganen errechnet und so seine Position auf der Warteliste erheblich verbessert.

Vorsicht!

Ein KI-System, das Analyseentscheidungen ermittelt, wird missbraucht, um sich einen persönlichen Vorteil zu erschleichen.

Worst Case

Schutzmaßnahmen nicht vorhanden

Der Hacker schleust eine Schadsoftware in das Laborinformationssystem ein. Der KI-Algorithmus ermittelt auf Basis der manipulierten Daten ein extrem geringes Risiko für eine Abstoßungsreaktion.

- Die Wartelistenposition des Patienten verbessert sich unverhältnismäßig.
- Andere Betroffene, die dringender auf ein Spenderorgan angewiesen sind, werden auf der Warteliste zurückgesetzt.

Best Case

Schutzmaßnahmen vorhanden

Mögliche Gegenmaßnahmen sind Schutzsysteme, die Schadsoftware erkennen, Protokolle der Änderungen sowie Erklärungen des KI-Algorithmus.

- Die Verantwortlichen im Labor erkennen die Manipulation der IT-Systeme.
- Eine Analyse der betroffenen Systeme und Daten wird durchgeführt.

Technische Maßnahmen

- Intrusion Detection Systems
- Protokollieren von Änderungen/Zugriffen
- Ausgeben detaillierter Erklärung zur Ergänzung des KI-Algorithmus
- Föderiertes Lernverfahren

Organisatorische Maßnahmen

- Überwachen der Wartelisten durch unabhängige Stellen
- Definition klinischer Prozesse zum Umgang mit entscheidungsunterstützenden KI-Systemen
- Validierung und Festlegung klinischer Aktionen durch medizinisches Fachpersonal (nicht durch KI-Systeme allein)

Szenario-Details: Die vertiefende Darstellung der jeweiligen Folgen und ihre Gegenüberstellung Worst und Best Case zeigen im vorliegenden Fall, wie organisatorische und technische Maßnahmen situativ ineinandergreifen können.

Details zum Szenario

Folgen Worst Case

- Das Laborinformationssystem passt, in Abhängigkeit der verfügbaren Spenderorgane, die Laborwerte des Patienten immer weiter an und prüft mit dem KI-Algorithmus, bis dieser eine Kompatibilität von 95 Prozent oder höher errechnet.
- Auf Grundlage der manipulierten Labordaten ermittelt der KI-Algorithmus dem Patienten ein extrem geringes Risiko für eine Abstoßungsreaktion.
- Dessen Wartelistenposition wird unverhältnismäßig verbessert. Andere Patientinnen und Patienten, die möglicherweise dringender auf ein Spenderorgan angewiesen sind, werden auf der Warteliste schlechter gestellt.
- Im Transplantationszentrum wird die durch den KI-Algorithmus ermittelte Kompatibilität entscheidungsunterstützend eingesetzt.
- Die Entscheidungsverantwortlichen verlassen sich – entgegen gängiger Praxis einer unabhängigen Überprüfung des Sachverhalts durch den Menschen – einzig auf das Ergebnis des KI-Systems.
- Die Patientin bzw. der Patient wird zur Transplantation ausgewählt und das Transplantationszentrum leitet die Operation ein.
- Wenige Tage nach der Operation treten Komplikationen auf, sodass der Patient hohe Dosen an Immunsuppressiva einnehmen muss, um die körpereigene Immunantwort gegen die Spenderniere zu unterdrücken.
- Binnen weniger Monate hat der Patient immer wieder mit eigentlich harmlosen Infektionen zu kämpfen, die durch das heruntergefahrene Immunsystem jedoch schwer in den Griff zu bekommen sind.
- Immer wieder müssen die Immunsuppressiva abgesetzt werden, damit das Immunsystem voll eingreifen kann. Dabei entsteht jedoch immer wieder eine Immunreaktion gegen das Spenderorgan, die letztlich zum Verlust führt. Der Patient ist nun mehrfach wöchentlich auf die Dialyse und ein neuerliches Spenderorgan angewiesen.

Folgen Best Case

- Das Labor erkennt die Manipulation an seinen IT-Systemen und ordnet eine Analyse der betroffenen Systeme und Daten an.
- Die automatische Überwachung der Zugriffsprotokolle des KI-Algorithmus zeigt eine Häufung während der Nachtstunden an, obwohl zu dieser Zeit üblicherweise eher wenige Anfragen verarbeitet werden. Außerdem wird eine spezielle Quelle identifiziert, die in kurzen Abständen immer wieder auf den KI-Algorithmus zugreift, dabei molekulargenetische Daten sendet, die sich aber nur sehr wenig voneinander unterscheiden.
- Durch die Kombination des KI-Algorithmus mit einer ausführlichen Erklärung fällt auf, dass der ausgewählte Patient unverhältnismäßig kurz auf der Warteliste für ein Spenderorgan steht. Eine Rückfrage bei der behandelnden Nephrologin deckt auf, dass seine aktuelle Nierenfunktion als verhältnismäßig gut einzustufen ist; eine Transplantation ist daher nicht sofort erforderlich.
- Das Transplantationszentrum fordert zusätzlich zur KI-Prognose einen traditionellen LCT-Test im Labor an, der eine übermäßig hohe Interaktion der HLA und des Immunsystems des Empfängers offenbart. Es wird keine Transplantation in Betracht gezogen. Zusätzlich wird eine Überprüfung der molekulargenetischen Labordaten und der Prognose des KI-Algorithmus angeordnet.
- Das Organ wird einem passenden Empfänger/einer passenden Empfängerin zugeführt.

Zur Verhinderung der Manipulation der Spender- und Empfängerdatenbanken sind folgende Maßnahmen möglich:

- **Definition klinischer Prozesse für den Umgang mit entscheidungsunterstützenden KI-Systemen:** Insbesondere die Validierung und Festlegung der klinischen Aktionen durch medizinische Expertinnen und Experten und nicht durch KI-Systeme allein.
- **Implementation eines Intrusion Detection Systems:** Dieses erkennt die fremde, eingeschleuste Schadsoftware des Hackers im eigenen Netzwerk.

- **Protokollieren aller Änderungen:** Das Laborinformationssystem protokolliert innerhalb der Insert-Only-Datenbank transparent alle durchgeführten Änderungen. So können Daten weder unbemerkt gelöscht noch manipuliert werden, sondern alle Änderungen werden mit einem Zeitstempel dauerhaft dokumentiert. Eine Software prüft Änderungen regelmäßig auf Konsistenz, zum Beispiel ob sie während der typischen Arbeitszeiten der Mitarbeitenden oder zu einer konkreten Laboranforderung passen.
- **Protokollieren aller Zugriffe:** Alle Zugriffe auf den KI-Algorithmus werden protokolliert und regelmäßig ausgewertet, um etwaige Häufungen von bestimmten Quellen oder zu bestimmten Uhrzeiten zu überprüfen.
- **Überwachen der Wartelisten:** Wartelisten für Spenderorgane werden durch unabhängige Stellen durch eigene Algorithmen überwacht, um ungewöhnliche Änderungen in der Wartelistenposition und somit Manipulationen auszuschließen.
- **Ausgeben einer Erklärung:** Der KI-Algorithmus wird um eine detaillierte Erklärung ergänzt, die neben der Risikowahrscheinlichkeit nun die konkreten molekulargenetischen Merkmale des Empfängers und deren erwartete Häufigkeit innerhalb der Population aufzeigt. Dazu werden Transplantationsdaten aus internationalen Transplantationszentren durch föderierte Lernverfahren miteinander datenschutzkonform kombiniert (siehe Infobox zu Verteiltes Lernen, S. 13; Plattform Lernende Systeme, 2019a, S. 27). Es herrscht somit mehr Transparenz darüber, welche konkreten Faktoren der Algorithmus zur Bestimmung der Kompatibilität herangezogen hat und wie häufig diese Merkmalskombination bisher in der Population erkannt wurde.

Im Gesundheitsbereich allgemein können darüber hinaus weitere Maßnahmen zum Schutz vor Missbrauch hilfreich sein:

- **Erklärbare KI:** Ärztinnen und Ärzte, die ein KI-basiertes, medizinisches Unterstützungssystem nutzen, benötigen eine gute Aufbereitung und Darstellung, wie das KI-System zu einem spezifischen Ergebnis gekommen ist, um eine sinnvolle Therapieentscheidungen treffen und das Risiko einer Manipulation reduzieren zu können. Verfahren erklärbarer KI können dazu beitragen, Ergebnisse eines Systems für Anwendende nachvollziehbar und -prüfbar zu machen.
- **Datensicherheit durch Anonymisierung oder Pseudonymisierung:** Dabei werden identifizierende Merkmale, wie Namen oder Geburtsdaten, aus den Daten entfernt und durch randomisierte Parameter ersetzt. Wichtig dabei ist, dass die relevanten medizinischen Charakteristiken in den Daten erhalten bleiben, um sie für das Training von KI-Systemen sinnvoll einsetzen zu können (vgl. Müller-Quade et al., 2020).
- **Automatisierte Verfahren zur kontinuierlichen Qualitätssicherung:** Betrachtet man KI-Systeme im Gesundheitswesen als medizinisches Produkt, gelten für sie dieselben strikten Zulassungsbestimmungen, wie zum Beispiel für medizinische Geräte und Medikamente. Mit einer zunehmenden Anzahl von KI-Anwendungen wird die manuelle Prüfung solcher Verfahren zum Engpass des wissenschaftlichen Fortschritts. Stattdessen bedarf es automatisierter Verfahren zur kontinuierlichen Qualitätssicherung von medizinischen Entscheidungsunterstützungssystemen, wie sie schon heute in Laboren zum Einsatz kommen. Regelmäßige Kontrollen durch Verwendung von speziell dazu ausgewählten Pseudofällen sowie die Berücksichtigung von Feedback der Nutzenden helfen dabei, die Vorhersagequalität zu überprüfen und etwaige Fehler frühzeitig zu erkennen. Dazu zählt auch eine detaillierte Charakterisierung der für das Training verwendeten Daten.
- **Zulassung:** Nach der Medizinprodukteverordnung der Europäischen Union (EU) müssen Konformitätsprüfungen von Medizinprodukten durch eine Drittstelle durchgeführt werden. Dies gilt für Hardware ebenso wie für Software. Wie andere Software auch, können KI-Algorithmen

eine CE-Kennzeichnung erlangen. Eine erprobte und allgemein verbindliche Konformitätsbewertung von Medizinprodukten mit selbstlernenden KI-Funktionen gibt es allerdings noch nicht. Die Medical Device Coordination Group (MDCG) der EU arbeitet gegenwärtig an einem MDCG-Leitfaden (vgl. MDCG).

KURZINFO

Adversarial-Machine-Learning

Adversarial-Machine-Learning-Angriffe können beispielsweise als Fehlklassifikationseingaben oder als Datenvergiftung (engl. data poisoning) verstanden werden, das bedeutet die Manipulation des Trainingsdatensatzes eines KI-Systems. Häufiger kommen die Fehlklassifikationseingaben vor. Hierbei schleust der Angreifer schädliche Inhalte in den Filter eines Machine-Learning-Algorithmus ein. Ein solcher Angriff zielt darauf ab, dass das System einen bestimmten Datensatz falsch klassifiziert (vgl. whatis.techtarget.com, 2021).

3.2 Mobilität: Autonomes Fahren und Verkehrsflüsse

Das Mobilitätssystem ist ein Beispiel für ein System of Systems (SoS, siehe Infobox). Denn es existieren starke Interaktionen über verschiedene Systeme hinweg, die jeweils von unterschiedlichen Stakeholdern verantwortet werden, wie beispielsweise Fahrzeuge, die Verkehrsinfrastruktur, die Gesellschaft oder der Markt. Aus der Perspektive eines SoS werden die hohen Anforderungen an Funktionssicherheit von Hard- und Software beim autonomen Fahren sowie die Sicherheit vor Cyberangriffen nochmals deutlich. Autonomes Fahren muss vor allen denkbaren Angriffs- und Manipulationsversuchen geschützt werden, wie die beiden folgenden, fiktiven Szenarien zeigen. Dabei muss jedoch eine geeignete Balance zwischen unseren Freiheitswerten und erforderlicher Kontrolle gewahrt werden.

KURZINFO

System of Systems (SoS)

Ein SoS ist ein System, bestehend aus einer Menge interagierender Systeme, von denen jedes einzelne System in sich selbst als System betrachtet wird. Die wichtigsten SoS-Charakteristiken, die unterschiedlich ausgeprägt sein können, sind: Jedes System kann unabhängig agieren und einen eigenen Zweck besitzen. Die individuellen Systeme der Menge werden unabhängig organisiert, um ihre eigenen Zwecke zu erfüllen. Die Systemkombination liefert Ergebnisse, die von den einzelnen Systemen nicht erreicht werden können (IPEK, 2021).

Wichtiger Bestandteil vieler Szenarien für das Mobilitätssystem der Zukunft sind autonome fahrerlose Fahrzeuge im Straßenverkehr. Die Fahrzeuge können als Pkw, für den ÖPNV oder auch als Shuttles eingesetzt werden. Autonome Fahrzeuge sind in sehr komplexe Mobilitätssysteme eingebettet, das heißt zum Beispiel in Datenökosysteme, um das Verarbeiten von Sensor- und Verkehrsinformationen oder auch das Lernen der Fahrzeuge zu gewährleisten. Ein mögliches Szenario ist der Missbrauch des Fahrzeugs als Waffe[8], ohne dass dabei Menschen direkt in die Ausführung des Missbrauchs involviert sein müssen (siehe Szenario 2).

SZENARIO 2

Angriff durch ein autonom fahrendes Fahrzeug verhindern

Ausgangssituation

Zahlreiche autonome fahrerlose Fahrzeuge sind auf den Straßen unterwegs, um Personen und Waren an ihr Ziel zu bringen. Die autonome Fahrzeugführung wird auf verschiedenen Ebenen des Mobilitätssystems neben regelbasierten KI-Systemen auch durch datengetriebene KI-Systeme bestimmt. Beispiele dafür sind die Wahrnehmung der Außenwelt durch Machine-Vision-Algorithmen oder die Planung, Optimierung und Regelung des Fahrzeugverhaltens.

Der Missbrauch

Mitglieder einer terroristischen Organisation manipulieren ein autonomes, fahrerloses Fahrzeug einer Geschäftsfrau mit dem Ziel, anderen Verkehrsteilnehmerinnen und -teilnehmern, Fahrzeugen und der Infrastruktur gezielt Schaden zuzufügen.

Vorsicht!

Das autonome Auto, dessen Zweck es ist, Personen sicher von A nach B zu befördern, wird für terroristische Zwecke missbraucht.

Worst Case

Schutzmaßnahmen nicht vorhanden

Über die Car-2-X-Schnittstelle manipulieren Angreifende die Sensorik und Systeme der internen Datenübertragung oder greifen direkt in die Steuerungselektronik des Autos ein.

- Das autonom fahrende Fahrzeug fährt auf dem Heimweg gezielt in eine Menschenmenge.
- Menschen werden verletzt, einige teilweise sehr schwer.

Best Case

Schutzmaßnahmen vorhanden

Durch vorbeugende Maßnahmen in der Entwicklungsphase des autonomen Autos sowie akute Gegenmaßnahmen, die situationsbedingt greifen und ausgelöst werden können, kann der Missbrauch verhindert werden.

- Das autonome Fahrzeug verletzt niemanden.
- Die Geschäftsfrau wird sicher nach Hause gefahren.

Technische Maßnahmen

- Übergang in Safe-State durch Fahrende selbst: manuell ausgelöst
- Anomalieerkennung
- Implementierte Regelsätze

Organisatorische Maßnahmen

- Security- und Safety-by-Design in der Entwicklungsphase des autonomen Autos bei der Software und Hardware berücksichtigen

Werden entsprechende Anforderungen an Security und Safety bereits früh im Produktentstehungsprozess berücksichtigt (vgl. Security- und Safety-by-Design), kann auch bei hochvernetzten sowie hochkomplexen Systemen Missbrauch verhindert oder zumindest erschwert werden. *Safety* kann in solchen Systemen nur durch ausreichende Security erreicht werden. Da nicht alle Missbrauchsfälle antizipiert werden können, besteht zusätzlich die Möglichkeit, akute Gegenmaßnahmen in der jeweils vorliegenden Situation zu ergrei-

8 Die Nutzung von KI-Systemen durch terroristische Organisationen wurde bisher noch nicht beobachtet. Allerdings nutzen solche Organisationen Fahrzeuge als Waffen. Ein Video aus dem Jahr 2016 zeigt, wie ISIS offenbar mit einem rudimentären System experimentierte, um Autos fernzusteuern (UNICRI & UNOCT, 2021).

fen. Im vorgestellten Beispiel könnte die Fahrerin den Fahrmodus manuell in einen Safe-State wechseln, das heißt, sie kann den autonomen Modus abschalten und das Auto in eine sichere Lage steuern. Zudem kann durch ein System der Anomalieerkennung in der Fahrzeugflotte eine frühzeitige Warnung abgegeben werden und durch implementierte, nicht überschreibbare Regelsätze zum Schutz von Menschen können Folgen für Leib und Leben verhindert oder erschwert werden.

KI-Systeme autonomer Fahrzeuge erheben und verarbeiten große Mengen an anonymisierten, personenbeziehbaren[9] (pseudonymisierten), aber auch personenbezogenen Daten, die über Car-2-X-Schnittstellen mit anderen autonomen Fahrzeugen und übergeordneten Systemen geteilt werden. Ein mögliches Missbrauchsszenario liegt in der Manipulation der Verkehrsdaten (siehe Szenario 3 und Adversarial-Machine-Learning). Die Möglichkeit zur Manipulation der Stauanzeige eines Kartendienstes zeigte ein Künstler 2020 im Rahmen einer Aktion auf, in der er mit einem einfachen Handwagen und 99 Handys einen Stau vortäuschte (Hoppenstedt, 2020).

SZENARIO 3

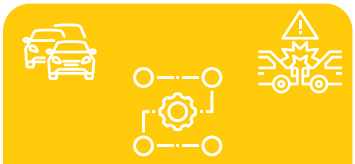
Missbrauch von Daten zur Umlenkung von Verkehrsflüssen verhindern

Ausgangssituation

Zahlreiche autonome, also fahrerlose Fahrzeuge sind auf den Straßen unterwegs. Die für einen sicheren und reibungslosen Verkehrsfluss notwendigen Daten teilen die Fahrzeuge untereinander und mit übergeordneten Systemen wie beispielsweise Verkehrsleitstellen.

Der Missbrauch

Ein Angreifer hat das Ziel, Chaos zu stiften, indem er Verkehrsdaten manipuliert, um autonome Fahrzeuge gezielt umzulenken.



Vorsicht!

Das KI-System, das als Verkehrsleitsystem eingesetzt ist, wird dazu missbraucht, Chaos zu stiften.

Worst Case

Schutzmaßnahmen nicht vorhanden

Der Angreifer dringt in das System der Verkehrsleitstelle ein und erlangt Zugriff auf die für die Streckenplanung notwendigen Verkehrsdaten auf einer übergeordneten Ebene des Mobilitätssystems.

- Zahlreiche autonome Fahrzeuge werden gezielt umgelenkt.
- Künstliche Staus und Menschenansammlungen entstehen, deren Auflösung lange Zeit in Anspruch nimmt.

Best Case

Schutzmaßnahmen vorhanden

KI-Systeme können genutzt werden, um andere KI-Systeme zu überwachen und so frühzeitig Angriffe zu erkennen.

- Keine künstlichen Staus und/oder Menschenansammlungen entstehen.
- Der Verkehr fließt ruhig weiter.

Technische Maßnahmen

- Advanced Systems Engineering (ASE)
- Anomaliedetektion

Organisatorische Maßnahmen

- Security-by-Design in Kombination mit ASE

Ansatz für das Leitbild des **Advanced Systems Engineering (ASE):** ganzheitliche und interdisziplinäre Perspektive, die der Komplexität Rechnung trägt und zugleich neue technologische und arbeitsorganisatorische Entwicklungen miteinbezieht

9 Pseudonymisierte bzw. personenbeziehbare Daten sind nach Art. 4 DSGVO nicht mehr einer spezifischen Person zuzuordnen, ohne zusätzliche Information hinzuzuziehen.

Auch in diesem Anwendungsfall stellt die Detektion von Anomalien in den Verkehrsdaten sowie das Vorgehen im Sinne des Security-by-Design eine wirksame Maßnahme für den Schutz vor Missbrauch dar. Besonders in komplexen System of Systems (SoS) ist es schon jetzt eine Herausforderung, kritische Systeme und deren Schnittstellen zu identifizieren, das heißt, Systeme mit potenziellen Angriffsflanken. Methoden des Advanced Systems Engineering (ASE) können vorausschauend die Identifikation solcher kritischen Systeme unterstützen (siehe hierfür ASE).

3.3 Drohneneinsatz in komplexen Umgebungen: Hafenbecken und Großveranstaltungen

KI-Systeme können in schwer zugänglichen und/oder sich dynamisch verändernden komplexen Umgebungen[10] Tätigkeiten übernehmen, die für den Menschen gefährlich, unzumutbar oder gesundheitsschädlich sind. Dies können beispielsweise Routineaufgaben wie die Überwachung von Infrastrukturen, die Erkundung von Gebieten oder die Unterstützung von Kommunikation sein. Damit sie ohne Menschen operieren und interagieren können, werden die eingesetzten KI-Systeme zunehmend mit speziellen Fähigkeiten ausgestattet (vgl. Beyerer et al., 2021). Dazu zählen Methoden für die Manipulation von Objekten, die Untersuchung bzw. Reparatur von Infrastrukturen und die weitreichende Interaktion. Hinzu kommt, dass die KI-Systeme sowohl als Einzelsysteme als auch in Interaktion miteinander, wie etwa Flugdrohnen in einem Schwarm, operieren können – und somit sowohl einzeln als auch in ihrem Miteinander Ziel von Missbrauchsversuchen werden können.

Häfen müssen jederzeit die wasserseitige Zugänglichkeit ermöglichen, wofür die Fahrrinnen und Hafenbecken regelmäßig von Ablagerungen befreit werden müssen. Dazu werden die Hafenanlagen mit Schiffen und ferngesteuerten Unterwasserdrohnen detailliert vermessen. Neben Peilschiffen werden heutzutage auch ferngesteuerte Drohnen eingesetzt. Basierend auf der Datenlage werden die Ablagerungen abgebaggert. Ein mögliches Missbrauchsszenario bestünde im Ausstatten der Drohnen mit Sprengkörpern (siehe Szenario 4).

10 Vgl. hierzu auch lebensfeindliche Umgebungen. Diese umfassen beispielsweise die Tiefsee, das Weltall, kontaminierte Umgebungen oder Krisengebiete (Plattform Lernende Systeme, 2019b). In lebensfeindlichen Umgebungen ist ein direkter Eingriff in bzw. Zugriff auf die KI-Systeme im Falle eines Missbrauchs oft nicht sofort möglich, sondern erst nach dem Ende des Einsatzes, welcher in einigen Fällen und unter gewissen Umständen bis zu mehreren Monaten dauern kann.

SZENARIO 4

Sabotage durch Unterwasserdrohnen verhindern

Ausgangssituation

Die Hafenbecken des größten europäischen Hafens in Rotterdam werden durch einen Schwarm KI-basierter Unterwasserdrohnen überwacht. Diese überprüfen Wassertiefen und die Bodengegebenheiten eigenständig und fordern bei Bedarf Saugdrohnen an, die Ablagerungen entfernen können. Dank Unterwassergaragen und einem Kommunikationsnetz sind sie ständig im Einsatz.

Der Missbrauch

Eine militante Gruppe fängt mehrere Unterwasserdrohnen ab und stattet sie mit zeitcodierten Sprengladungen aus. Dazu werden die Unterwasserdrohnen nur kurz aufgehalten und nehmen im Anschluss ihre eigentlichen Aufgaben wieder wahr.

Vorsicht!

Die Unterwasserdrohnen, die zur Pflege der Hafenbecken im Einsatz sind, werden für terroristische Zwecke missbraucht.

Worst Case

Schutzmaßnahmen nicht vorhanden

Die angebrachten Sprengladungen explodieren zwei Monate später an unterschiedlichen Orten am Hafenbecken.

- Hafenstrukturen und mehrere Schiffe werden beschädigt.
- Ein Schiff sinkt in einer der Hauptzufahrtsrinnen und blockiert den Hafenverkehr.

Best Case

Schutzmaßnahmen vorhanden

Änderungen an den Drohnen und der Umgebung werden erkannt; durch regelmäßiges Scannen der Hülle der autonomen Fahrzeuge an definierten Orten wird die angebrachten Sprengladung erkannt.

- Die Sprengsätze werden entschärft.
- Die Angreifer werden identifiziert und festgenommen.

Technische Maßnahmen

- Anomaliedetektion (z. B. Änderung der Route, des Gewichts etc.)
- Regelmäßiges Abscannen der Drohnenhülle

Organisatorische Maßnahmen

- Security-by-Design in Kombination mit ASE

Die im Beispiel dargestellte Anomaliedetektion erkennt, dass die Unterwasserdrohnen ihre Route verlassen haben. Dies geschieht konkret über eine modellbasierte Missionsüberwachung. Zudem ist in der Drohne ein System der (Selbst)-Überwachung implementiert, das Veränderungen am Gehäuse und des Gewichts erkennen kann, genauso wie die Annäherung von Personen und weiteren Drohnen.

Für die Routen- und Verhaltensüberwachung der Unterwasserdrohnen werden historische und simulierte Verhaltensdaten genutzt. Diese werden mit dem tatsächlichen Verhalten der Unterwasserdrohnen verglichen. Dabei können Abweichungen und Unregelmäßigkeiten bestimmt und bewertet werden. Die Anomaliedetektion löst ein Warnsignal aus und informiert den Hafenbetreiber, der konkrete Schritte einleiten kann, um die Situation zu entschärfen.

Zusätzlich können die Unterwasserdrohnen regelmäßig auf Fremdkörper überprüft werden. Dazu wird mithilfe eines Sonars oder Lasers die Hülle einer Unterwasserdrohne gescannt und mit den Modelldaten verglichen. Fallen Unregelmäßigkeiten auf, muss die Drohne überprüft werden. Der Scanvorgang kann beispielsweise bei der Einfahrt in die Unterwassergarage oder an ausgewählten Hafenbereichen stattfinden.

Flugdrohnen werden häufig für die Überwachung bei Großveranstaltungen genutzt oder in Krisengebieten eingesetzt. Die integrierten Sensoren können wichtige Informationen erfassen und weiterleiten. Der Luftraum verfügt jedoch kaum über effiziente Abwehrsysteme und mechanische Schutzanlagen bzw. -vorrichtungen, die Personen, kritische Bereiche und Objekte vor möglichen Angriffen von Flugdrohnen schützen. Flugdrohnen könnten jedoch auch für gezielte Angriffe auf Personen missbräuchlich zum Einsatz kommen. (siehe Szenario 5).

SZENARIO 5

Angriff durch Flugdrohnen verhindern

Ausgangssituation

Finale der Fußballweltmeisterschaft in einer Großstadt: Das Stadion wird mithilfe eines Schwarms KI-basierter Flugdrohnen aus der Luft überwacht. Dazu kommunizieren die KI-Systeme untereinander und kooperieren im Falle einer eventuellen Verfolgung von bestimmten Zielpersonen.

Der Missbrauch

Hacker dringen in die Einsatzzentrale des Stadions ein und schleusen eine Schadsoftware anhand eines manipulierten Speicherchips in das System ein. Am Tag des Finales, kurz nach der Halbzeit, wird diese automatisch aktiviert.

Vorsicht!

Drohnen, die für den Zweck des Veranstaltungsschutzes eingesetzt werden, werden für terroristische Zwecke missbraucht.

Worst Case

Schutzmaßnahmen nicht vorhanden

Durch die Aktivierung der Schadsoftware wird das Update des Wertes für Normalnull routinemäßig in alle Systeme gesendet. Die Luftdrohnen senken sich um **minus** 200 Meter.

- Alle Luftdrohnen senken sich um 200 Meter und fliegen in vollem Flug in die Zuschauermenge.
- Es gibt zahlreiche Verletzte.

Best Case

Schutzmaßnahmen vorhanden

Ausreichender Schutz der Operationsbasis verhindert das Eindringen der Hacker.
Durch das Erkennen von Unregelmäßigkeiten – das gleichförmige Absenken der Flugdrohnen – wird ein Schutzszenario eingeleitet.

- Der Regelsatz zwingt die Drohnen zum Abbremsen.
- Es kommen keine Personen zu Schaden.

Technische Maßnahmen

- Implementierter Regelsatz, der vom restlichen System isoliert ist
- Anomaliedetektion

Organisatorische Maßnahmen

- Regelsatz ist zertifiziert

Die rettende Schutzmaßnahme ist ein implementierter Regelsatz, der den Drohnen verbietet, näher als fünf Meter an Menschen heranzufliegen und diese anzugreifen. In dem Moment, in dem die Schadsoftware aktiviert und der „Normalnull-Wert" manipuliert wird, melden Sensoren (Ultraschall- oder Radarsensoren), dass die Einhaltung des Minimalabstandes zu Menschen und auch zum Boden unterschritten wird. Dieser Regelsatz, der dann greift, ist vom Rest des Systems isoliert, zertifiziert und im Idealfall auch von einem anderen Hersteller entwickelt und eingebaut worden. Der Regelsatz lässt sich nicht überprogrammieren und schützt deshalb die Menschen vor einem Angriff durch die Flugdrohne.

3.4 Unternehmenskontext: E-Mail-Kommunikation

Prominente Vorfälle wie die Erpressung des Pipeline-Betreibers Colonial im Jahr 2021 zeigen, wie folgenreich eine Attacke mit einer sogenannten Ransomware für ein Unternehmen und weit darüber hinaus sein kann (Bertrand et al., 2021). So musste das Unternehmen fünf Millionen Dollar an die Erpresser zahlen und so konnten zeitweise mehrere US-Bundesstaaten nicht mehr mit ausreichend Treibstoff versorgt werden. Es ist davon auszugehen, dass kriminelle Gruppen künftig mithilfe von KI immer ausgefeiltere Methoden entwickeln werden, um Schadsoftware in Unternehmen einzuschleusen und sich auf diese Weise zu bereichern oder Informationen auszuleiten. Da Unternehmen für ihre interne wie externe Kommunikation auf E-Mails angewiesen sind, sind diese Kommunikationswege und die Personen, die daran teilnehmen, häufig das Angriffsziel solcher Attacken.

Während jedes handelsübliche Smartphone über automatisierte Wortvorschlagsfunktionen verfügt, können Computer, versehen mit deutlich höherer Rechenleistung, mit KI-Algorithmen komplette Texte vollautomatisiert erzeugen. Aus der Wortwahl und der statistischen Nutzung der Wörter, aber auch aus häufig benutzten Satzbauten kann KI einen Text selbstständig zusammenstellen. Liegen diesem Computer mit der KI-Funktion als Trainingsdaten die echten Textmuster von Personen der Vorstandsebene eines Unternehmens (CxOs, d. h. CEO, COO, CFO und andere) vor, wie etwa aus Interviews, Pressemeldungen oder Social Media Accounts, können daraus neue Textblöcke künstlich erstellt und für kriminelle Zwecke angewendet werden (siehe Szenario 6). IT-Sicherheitsforschende können sogar aufzeigen, dass neueste Sprachmodelle überzeugendere E-Mails für Spear-Phishing-Attacken bzw. Business E-Mail-Compromise-Attacken (siehe Infobox) generieren können als Menschen (Newman, 2021).

KURZINFO

Business-E-Mail-Compromise-Attacke (BEC)

Bei einer Business-E-Mail-Compromise-Attacke werden gefälschte, geschäftliche E-Mails verwendet, um beispielsweise Daten zu erbeuten oder Systeme funktionsuntüchtig zu machen. Cyberkriminelle versenden hierfür E-Mails, die scheinbar von Kolleginnen oder Kollegen, Vorgesetzten oder Geschäftspartnern stammen. In diesen E-Mails fordern sie die Zielperson zu einer bestimmten Tätigkeit auf (beispielsweise Öffnen eines Links, Überweisen von Geld, Mitteilen von Daten etc.). BEC-Attacken greifen zur Informationsbeschaffung oft auf Spear-Pishing, das heißt, das Sammeln von Informationen über das Internet und über soziale Netzwerke, zurück, weshalb BEC-Attacken oft auch Spear-Pishing-Attacken genannt werden. Weitere gängige Bezeichnungen sind „Man-in-the-eMail Attack“ oder „CxO-Fraud“.

SZENARIO 6

Einschleusen von Schadsoftware mit Lösegeldforderung verhindern

Ausgangssituation

Ein gut situiertes Unternehmen kommuniziert intern wie extern über E-Mail. Über ihre Website sind Interviews sowie Pressemeldungen von Personen aus der Vorstandsebene des Unternehmens online frei zugänglich.

Der Missbrauch

Eine kriminelle Gruppe schleust mithilfe einer E-Mail eine Schadstoffsoftware (Spear-Phishing) in das Abrechnungssystem der Buchhaltungsabteilung ein. Die E-Mail, die von einer vertrauenswürdigen Quelle zu stammen scheint – dem CEO –, wurde mithilfe von KI-basierten Sprachmodellen und Textproben des CEO als Trainingsdaten täuschend echt nachgestellt und fordert die Mitarbeitenden auf, umgehend zu antworten sowie den Anhang zu öffnen.

Vorsicht!

KI-basierte Sprachverarbeitungssysteme werden missbraucht, um Erpressungen zu optimieren.

Worst Case

Schutzmaßnahmen nicht vorhanden

Mitarbeitende im Unternehmen erkennen die Täuschung nicht und öffnen den Anhang (sog. Ransomware).

- Die Ransomware attackiert das Abrechnungssystem.
- Die Abrechnungen können wegen fehlender Eingabedaten nicht mehr automatisiert erstellt werden.
- Lösegeld wird zur Deaktivierung der Ransomware gefordert.
- Dem Unternehmen entsteht ein wirtschaftlicher/ideeller Schaden.

Best Case

Schutzmaßnahmen vorhanden

Das Unternehmen setzt einen modernen Spamfilter mit KI-Funktionalität ein, der Alarm schlägt und die Mitarbeitenden warnt, sodass die Täuschung erkannt und der Erpressungsversuch vereitelt wird.

- Es entsteht kein finanzieller Schaden.
- Es entsteht kein Reputationsschaden.

Technische Maßnahmen

- Spamfilter mit KI-Funktionalität

Organisatorische Maßnahmen

- Prozess im Umgang mit E-Mails definieren
- Beschäftigte regelmäßig durch Schulungen sensibilisieren

Unternehmen, insbesondere Großunternehmen, benutzen seit vielen Jahren Spamfilter für die E-Mail-Korrespondenz. Moderne Spamfilter sind mit KI-Funktionalität ausgestattet. Sie können gefälschte E-Mail-Adressen und Domänen sowie verdächtige Texte, Wörter, Missbrauchsmuster und Abweichungen von der Normalkorrespondenz, aber auch Auffälligkeiten in E-Mail-Kontexten, wie etwa den Meta-Daten, unterscheiden und erkennen. Als Trainingsdaten werden erkannte verdächtige, leicht modifizierte bzw. gefälschte E-Mails verwendet. Um die Wirksamkeit dieser Systeme zu verbessern, werden Muster der Normalkorrespondenz des jeweiligen Accounts aufgenommen und als Profil hinterlegt. Dadurch kann die KI-Funktion des Spamfilters leichte Abweichungen gegenüber dem „Normal-Zustand“ automatisiert und permanent auf gleich hohem Qualitätsniveau erkennen. Werden Abweichungen erkannt, wird ein Warnhinweis an die Nutzenden versandt.

Unternehmen sollten zudem Prozesse für praktische Verhaltensweisen im Umgang mit E-Mails definieren und Mitarbeitende dadurch für die Problematik von Spear-Phising und BEC-Attacken immer wieder sensibilisieren. Zwei praktische Verhaltensweisen können Teil solcher Prozesse sein. Wird eine gefälschte CxO-Mail durch den Spamfilter nicht erkannt, aber die Mitarbeitenden sind sich unsicher, was die Echtheit angeht, sollte eine kurze Rückfrage bzw. Bestätigungsnachfrage erfolgen. Dazu sollte die original CxO-Adresse genutzt und nicht auf die eingegangene E-Mail geantwortet werden. Es sollte weiterhin die Möglichkeit bestehen, dass die Mitarbeitenden eine Meldung an interne Zentralabteilungen versenden, wie etwa Continuity (BC), Security Operation Center (SOC) oder Computer Emergency Response Teams (CERT). Diese Abteilungen versenden Warnhinweise über neue, zum Teil massenhaft auftretende Phishing-, Spear-Phishing- oder BEC-Mails an alle Mitarbeitenden eines Unternehmens, sofern derartige Angriffe erwartet werden.

3.5 Arbeitskontext: Mensch-Maschine-Interaktion

KI-Systeme eröffnen Beschäftigten vielfältige Potenziale für ein sicheres, eigenverantwortliches und selbstbestimmtes Arbeiten. So können sie beispielsweise Beschäftigte von sich wiederholenden oder mühsamen Tätigkeiten entlasten oder dazu eingesetzt werden, Arbeitsprozesse zu beobachten, um Fehler zu vermeiden und so die Arbeitssicherheit zu erhöhen. Durch KI-Systeme erhobene und ausgewertete Daten können auf einer individuellen Ebene für die einzelnen Beschäftigten belastungsärmere und bessere Arbeitsprozesse erwirken, indem sie sich an die individuelle Arbeitsweise und speziellen Routinen der Beschäftigten anpassen: Dazu erfassen beispielsweise in der Produktion eingesetzte, selbstlernende Roboterwerkzeuge Schrittfolgen und Arbeitstechniken einer Facharbeiterin und führen richtige Arbeitsschritte zum optimalen Zeitpunkt aus. Durch die Erhebung physischer Daten – wie Stress-Level, Müdigkeit, Konzentration und Aufmerksamkeit – können KI-Systeme der Zukunft die Kollaboration zwischen Mensch und Roboter individualisieren und so die mentale Gesundheit der Beschäftigten verbessern. Zugleich gilt es zu verhindern, dass durch den Einsatz von Robotern eine Demotivation bei den Beschäftigten auftritt und so Leistungs- und Qualitätskosten entstehen (Stock-Homburg & Merkle, 2019, S.11). KI-Systeme können so auch vor Burn-out oder vor Bore-out schützen, wenn die erhobenen Daten konkrete Hinweise auf Selbstausbeutung und Überarbeitung oder Demotivation aufzeigen. Auf einer aggregierten Ebene können diese Maßnahmen einerseits zu einer Steigerung der Qualität und Effizienz für die Unternehmen beitragen, andererseits die Zufriedenheit und Motivation der Beschäftigten im Arbeitsprozess erhöhen.

Auch wenn die Nutzung von KI-Systemen für die beschriebenen Zwecke große Chancen für die Beschäftigten und letztlich für die Arbeitswelt bietet, können die erhobenen Daten im Missbrauchsfall zu durchgängiger, individueller Leistungskontrolle führen (siehe Szenario 7):

SZENARIO 7

Individuelle Leistungsüberwachung verhindern

Ausgangssituation

Ein Unternehmen setzt KI-basierte, selbstlernende Werkzeuge in der Produktion ein, die die einzelnen Arbeitsschritte, Routinen und mentale Verfasstheit der Beschäftigten erfassen, um sich ihnen individuell anzupassen. Um voneinander beschleunigt zu lernen, sind die KI-basierten Werkzeuge miteinander verknüpft und können sich so auf die Erfahrungswerte und Fähigkeiten der anderen Maschinen stützen.

Der Missbrauch

Das Unternehmen nutzt die KI-Systeme in der Produktion missbräuchlich, um aus den gesammelten Daten der Mitarbeitenden Nutzerprofile zur weiteren Auswertung, Überwachung und Leistungskontrolle zu erstellen.

Vorsicht!

Ein KI-System zur Unterstützung in der Produktionsline wird als Überwachungsinstrument für Beschäftigte missbraucht.

Worst Case

Schutzmaßnahmen nicht vorhanden

Das Unternehmen kann ungehindert KI-Systeme zur durchgängigen, individuellen Überwachung einsetzen, da keinerlei betriebliche Einschränkungen vorherrschen.

- Die Mitarbeitenden werden permanent überwacht.
- Leistungsergebnisse werden vorausberechnet.
- Lohn/Leistungsprämien werden mit Daten verknüpft.
- Für die Beschäftigten entsteht ein hoher Leistungsdruck mit psychischen Auswirkungen („gläserne Beschäftigte“).
- Wegen Fehlschlüssen kommt es zu Diskriminierungen.

Best Case

Schutzmaßnahmen vorhanden

Es gibt ausreichend technische Maßnahmen und betriebliche Instrumente, um eine individuelle Leistungskontrolle zu unterbinden.

Das Unternehmen setzt KI-Systeme ein:

- Zur Effizienzsteigerung.
- Zur attraktiven individuellen Arbeitsplatzgestaltung.
- Für sichere Arbeitsumgebung, die die Beschäftigten in der produktiven Ausführung ihrer Aufgaben unterstützt.

Technische Maßnahmen

- Anonymisierung der Daten
- Verteiltes Lernen

Organisatorische Maßnahmen

- Betriebsvereinbarungen zum Umgang mit individuellen Leistungskontrollen
- Wahrung der Datensparsamkeit
- Klar geregelte Zugriffsrechte auf Daten
- Change Management unter Einbeziehung der Mitarbeitenden
- Transparenzregelungen

Um zu verhindern, dass KI-Systeme dazu eingesetzt werden, Beschäftigte zu überwachen, ihre Leistung zu bewerten oder künftige Leistungsprognosen zu erstellen, müssen Unternehmen in enger Kooperation mit den Betriebsräten und Arbeitnehmervertretungen entsprechende Maßnahmen ergreifen: Dies betrifft zum einen technische Maßnahmen, wie eine Anonymisierung der erhobenen Daten direkt dort, wo die Daten anfallen, und, damit verbunden, klar geregelte Zugriffsrechte, um Rückschlüsse durch anonymisierte Daten auf Beschäftigte durch Kontextfaktoren zu verhindern. Weiterhin betrifft dies die ausschließliche Verwendung von Daten mit einer Löschfunktion nach Auswertung oder die Erhebung von nur zwingend notwendigen Daten im Sinne des Prinzips der Datensparsamkeit. Alternativ können Unternehmen auf „Federated Learning“ setzen (siehe S. 13), einer technischen Lösung, mit der auf Anonymisierung und Datensparsamkeit verzichtet werden kann: Damit überschreiten sensible Daten (z. B. personenbezogene Daten) keine Systemgrenzen und können auf diese Weise für die KI genutzt werden, während gleichzeitig die Datenhoheit bei den betroffenen Personen verbleibt.

Auf einer anderen Ebene der betrieblichen Regelungen kann die Verwendung der KI-Systeme und der entsprechenden Daten zusammen mit den Beschäftigten in Betriebsvereinbarungen festgehalten werden. Ebenso kann definiert werden, welche Systeme Beschäftigte im Arbeitskontext nutzen dürfen, und ein Bewusstsein für die Problematik der KI-basierten Datenverarbeitung kann bei den Beschäftigten geschaffen werden. Der Einsatz von KI-Systemen in den Unternehmen kann von der Unternehmensführung und den Beschäftigten nur gemeinsam bewältigt werden. Insgesamt geht es dabei um die Gestaltung eines neuen Verhältnisses zwischen Mensch und Technik, in dem Mensch und KI-System produktiv zusammenwirken und die jeweiligen Stärken betont werden. Ein Change Management, das die Mitarbeitenden sukzessiv miteinbezieht, ist ein entscheidender Faktor für die erfolgreiche Einführung von KI-Systemen und kann einem Missbrauch von KI in Unternehmen vorbeugen. Eine solche System-Einführung kann sich durchaus komplex darstellen, denn sie erstreckt sich von der Ziel- und Folgeneinschätzung über die Planung und Gestaltung bis hin zur Vorbereitung und Implementierung und schließlich der Evaluierung und Anpassung des Einsatzes von KI-Systemen (für detaillierte Ausführungen zum Change Management siehe Stowasser & Suchy et al., 2020; vgl. auch Schnalzer et al., 2021, S. 8).

Neben der schon angesprochenen Thematisierung in Betriebsvereinbarungen erscheint es wichtig, dass die betroffenen Beschäftigten Informationen darüber erhalten, welche Daten über sie in welchem Zusammenhang existieren und für welche Zwecke erhoben werden. Dieses grundlegende Prinzip der sogenannten inversen Transparenz ist möglicherweise einer der wesentlichen Erfolgsfaktoren für eine menschengerechte Einführung von KI-Systemen. Schließlich sollte für Transparenz über die getroffenen Maßnahmen zur Verhinderung von Missbrauch gesorgt werden.

4. Gestaltungsoptionen

Um künftig einen wirkungsvollen Schutz vor Missbrauch von KI-Systemen zu erzielen, der Vertrauen in die KI-Technologie gewährleistet und zudem deren Nutzen voll ausschöpft, schlagen die Autorinnen und Autoren des Whitepapers folgende Gestaltungsoptionen vor.

Die politischen Entscheidungsträgerinnen und Entscheidungsträger könnten …

- sich für die Einführung gesetzlicher Regelungen auf europäischer Ebene engagieren, welche zum einen die Verantwortlichkeiten von Entwickelnden, Herstellenden, Betreibenden, Reparierenden, Prüfenden und Nutzenden im Falle eines Fehlers oder Missbrauchs regeln und zum anderen Vorgaben für die Schadensregelung klar definieren.

- sich dafür einsetzen, dass bei ausgewählten KI-Systemen eine regelmäßige Überprüfung auf mögliche Missbrauchspotenziale durch Drittstellen – auch nach einer Zulassung – verpflichtend ist (vgl. Heesen, Müller-Quade & Wrobel, 2020). Solche Drittstellen könnten auch entsprechende Nachweise ausstellen. Hierfür ist insbesondere der Auf- bzw. Ausbau entsprechender Stellen, die solche Kontrollen durchführen können, notwendig. Hierunter sollten vor allem solche KI-Systeme fallen, deren Auswirkungen im Missbrauchsfall potenziell schwerwiegend sein können und zugleich sehr viele Nutzende betreffen würden. Mögliche Beispiele sind KI-Systeme in kritischen Infrastrukturen wie beispielsweise in Energiesystemen oder autonomen Beförderungssystemen (z. B. Bahn, Auto, Fluggeräte) sowie Anwendungen in der Medizin.

- gezielt Forschungsvorhaben und zugehörige Entwicklungseinrichtungen fördern, die sich mit der Identifikation und dem Vereiteln von Missbrauchsversuchen von KI-Systemen sowie allgemein mit vertrauenswürdiger KI beschäftigen.

- vertrauensbildende Maßnahmen für den Einsatz und das Verständnis von KI-Systemen fördern. Ziel ist es, die Sensibilisierung und das Rechtsbewusstsein aller Betroffenen und insbesondere der Bevölkerung zu erhöhen.

- in ausgewählten Anwendungsbereichen von hohem öffentlichen Interesse interdisziplinär besetzte Fachbeiräte einberufen, die regelmäßig die Weiterentwicklungen von KI-Systemen hinsichtlich des möglichen Missbrauchs bewerten.

Die Forschungsinstitutionen könnten …

- kommerziell erhältliche KI-Systeme für den Konsumenten auf Missbrauchspotenziale untersuchen. Solche Untersuchungen könnten im Auftrag des Herstellers oder durch Verbraucherverbände erfolgen.

- publizierte Missbrauchsfälle hinsichtlich der Art und Weise des Missbrauchs untersuchen und geeignete Gegenmaßnahmen entwickeln. Dies kann auch für Herstellende durchgeführt werden.

- die Forschung zu Fragen des Missbrauchs und zum Schutz vor Missbrauch von KI-Systemen sowie allgemein zu vertrauenswürdiger KI intensivieren und Simulations-, Test- und Zertifizierungsumgebungen entwickeln.

- zielgerichtete Aus- und Weiterbildung des Fachpersonals anbieten.

Die Bildungsinstitutionen könnten …

- sich dafür einsetzen, dass in schulischen und betrieblichen Bildungseinrichtungen entlang der Bildungskette über den Schutz vor Missbrauch sowie mögliche Gegenmaßnahmen aufgeklärt wird, sowie die generelle Auseinandersetzung mit KI-Systemen gezielt fördern.

- ein Bewusstsein für KI sowie einen reflexiven, kritischen Umgang mit KI-Systemen von früher Kindheit an gezielt schulen und fördern.

- Wissen und Verständnis von KI sowie KI-Systemen fördern. So könnten etwa in Industrie und Handelskammern sowie Volkshochschulen gezielt Angebote durchgeführt werden, um große Teile der Gesellschaft zu erreichen.

Die Unternehmen könnten …

- bei der (Weiter-)Entwicklung von KI-Systemen mögliche Ziele von Missbrauch bewerten und geeignete Gegenmaßnahmen ableiten, die als Prävention implementiert oder im Ernstfall umgesetzt werden können.

- für die (Weiter-)Entwicklung von KI-Systemen die Auftragsforschung zum Schutz vor Missbrauch ausbauen.

- in der Anwendung zusammen mit dem Herstellenden des KI-Systems in festgelegten Zeitabständen Testreihen durchführen, um mögliche Anomalien in der Anwendung zu erkennen. So könnten beispielsweise KI-Systeme sowohl bei Anwendenden als auch bei Herstellenden redundant genutzt werden. Aufbauend auf gemeinsam festgelegten Testdaten kann so die Bewertung der Ergebnisse beider Systeme verglichen werden.

- durch Pilot-Anwendungen praktische Erfahrungen im Umgang mit und in der Anwendung von KI-Systemen sammeln. So können sich die Unternehmen besser vor einer möglichen missbräuchlichen Anwendung der KI-Systeme durch Externe schützen und zusammen mit den Beschäftigten betriebsspezifische Regelungen für den KI-Einsatz aufstellen.

- bei der Anwendung von KI-Systemen erkannte Vorfälle unter anderem mit Verdacht auf Wirtschaftsspionage und -sabotage an die Strafverfolgungsbehörden melden. So können Angreifende identifiziert und weitere Anwendende über Vorkommnisse informiert werden (im Sinne des IT-Sicherheitsgesetz 2.0 via BSI).

- für die Anwendung von KI-Systemen die Infrastruktur und Umgebung für den sicheren Einsatz von KI rechtzeitig schaffen.

- sowohl bei der (Weiter-)Entwicklung als auch bei der Anwendung dafür Sorge tragen, dass ihr Fachpersonal zielgerichtet aus- und weitergebildet wird.

Die Zivilgesellschaft könnte …

- gesellschaftliche Herausforderungen identifizieren und mit verhältnismäßigen Mitteln Ideen für den Einsatz von KI entwerfen, die einen tatsächlichen gesellschaftlichen Nutzen bringen.

- entsprechende, gemeinsam erarbeitete Vorschläge im engen Austausch mit der Politik einbringen und die Politik dabei unterstützen, die Entwicklung und den Einsatz von KI so zu regulieren, dass gesamtgesellschaftliche Interessen gewahrt werden.

- ein breites Ökosystem von Stakeholdern aus Zivilgesellschaft und Wissenschaft weiter ausbauen, welches die KI-Politikgestaltung in Europa mit ihren Fähigkeiten unterstützt und vorantreibt.

Es bedarf eines gesellschaftlichen Diskurses über …

- das Verhältnis von technischen, organisatorischen, rechtlichen und gesellschaftlichen Maßnahmen zum Schutz vor Missbrauch. Denn es gilt beispielsweise zu berücksichtigen, dass nicht alle möglichen Schutzmaßnahmen und -techniken rechtlich, ethisch oder politisch zulässig bzw. wünschenswert sind. Werden beispielsweise Einschränkungen von Fähigkeiten eines mobilen KI-Systems implementiert, könnten diese zeitweise außer Kraft gesetzt werden, wenn Gefahr für Leib und Leben von Personen besteht und dadurch die Gefahr abgewendet werden kann.

- künftige Veränderungen durch und mit KI-Systemen, der die Transparenz und das Bewusstsein hinsichtlich der KI-Technologie fördert und alle relevanten Stakeholder miteinbezieht. Dieser Diskurs sollte pragmatisch und sachlich geführt werden, sodass über Funktionsweise und Missbrauchsmöglichkeiten von KI-Systemen sowie über wirksame Gegenmaßnahmen und die Innovationskraft von KI-Systemen aufgeklärt wird, ohne dabei Ängste zu schüren.

Literatur

BBC (2015): Drug delivery drone crashes in Mexico. https://www.bbc.com/news/technology-30932395 (abgerufen am 10.08.2021).

Bertrand, N. et al. (2021): Colonial Pipeline did pay ransom to hackers, sources now say. CNN. https://edition.cnn.com/2021/05/12/politics/colonial-pipeline-ransomware-payment/index.html (abgerufen am 28.06.2021).

Beyerer, J. et al. (Hrsg.) (2021): Kompetent im Einsatz – Variable Autonomie Lernender Systeme in lebensfeindlichen Umgebungen. Whitepaper aus der Plattform Lernende Systeme, München 2021. https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG7_Whitepaper_Autonomiegrad.pdf (abgerufen am 02.08.2021).

BKA (2021): Lagebild Cybercrime 2020. https://www.bka.de/SharedDocs/Downloads/DE/Publikationen/JahresberichteUndLagebilder/Cybercrime/cybercrimeBundeslagebild2020.html;jsessionid=B4BEAF11B5A24B949116906EB8ADAAE7.live2291?nn=28110 (abgerufen am 10.08.2021).

BMJV (2018): Bundesdatenschutzgesetz (BDSG), § 46. https://dejure.org/gesetze/BDSG/46.html (abgerufen am 20.04.2021).

Brundage, M. A. et al. (2018): The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. https://arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf (abgerufen am 28.05.2021).

Cladwell, M. et al. (2020): AI-enabled future crime. Crime Science, 14. https://crimesciencejournal.biomedcentral.com/articles/10.1186/s40163-020-00123-8 (abgerufen am 28.05.2021).

DPA (2021): Deutschland soll Vorreiter beim autonomen Fahren werden. https://www.zeit.de/mobilitaet/2021-05/bundestag-autonomes-fahren-gesetz-verkehrspolitik-strassenverkehr (abgerufen am 25.05.2021).

Dumitrescu, R. et al. (Hrsg.) (2021): Engineering in Deutschland – Status quo in Wirtschaft und Wissenschaft. Ein Beitrag zum Advanced Systems Engineering, Paderborn.

Feldstein, S. (2019): The global expansion of AI surveillance. https://carnegieendowment.org/files/WP-Feldstein-AISurveillance_final1.pdf (abgerufen am 13.08.2021).

Heesen, J., Müller-Quade, J. & Wrobel, S. et al. (Hrsg.) (2020): Zertifizierung von KI-Systemen – Kompass für die Entwicklung und Anwendung vertrauenswürdiger KI-Systeme. Whitepaper aus der Plattform Lernende Systeme, München. https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG1_3_Whitepaper_Zertifizierung_KI_Systemen.pdf (abgerufen am 21.06.2021).

Hesse, T. & Müller-Quade, J. et al. (Hrsg.) (2021): Mit KI sicher reisen. Datenmanagement und Datensicherheit bei KI-basierten Reiseassistenten. Whitepaper aus der Plattform Lernende Systeme, München. https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG3_5_WP_Sicheres_Reisen.pdf (abgerufen am 06.07.2021).

Hoppenstedt, M. (2020): Was der Bollerwagen-Hack über Google Maps verrät. https://www.sueddeutsche.de/digital/google-maps-hacks-stauanzeige-1.4784081 (abgerufen am 13.08.2021)

IPEK (2021): System of Systems. https://www.ipek.kit.edu/glossar/index.php?title=System_of_Systems (abgerufen am 01.12.2021).

Kort, K. (2020): Amazon darf Pakete mit Drohnen ausliefern. https://www.handelsblatt.com/unternehmen/handel-konsumgueter/onlinehaendler-amazon-darf-pakete-mit-drohnen-ausliefern/26145978.html?ticket=ST-2913157-Ylk9md322435rdKfGBFr-ap4 (abgerufen am 06.07.2021).

Mozur, P. (2019): One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority. https://www.nytimes.com/2019/04/14/technology/china-surveillance-artificial-intelligence-racial-profiling.html (abgerufen am 10.08.2021).

Müller-Quade, J. et al. (Hrsg.) (2019): Künstliche Intelligenz und IT-Sicherheit. Bestandsaufnahme und Lösungsansätze. Whitepaper aus der Plattform Lernende Systeme, München. https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/20190403_Whitepaper_AG3_final.pdf (abgerufen am 28.06.2021).

Müller-Quade, J. et al. (Hrsg.) (2020): Sichere KI-Systeme für die Medizin. Whitepaper aus der Plattform Lernende Systeme, München. https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG3_6_Whitepaper_07042020.pdf (abgerufen am 06.07.2021).

Newman, L. H. (2021): AI Wrote Better Phishing Emails Than Humans in a Recent Test. https://www.wired.com/story/ai-phishing-emails/ (abgerufen am 10.08.2021).

Niemann, A.-L. (2020): Wald liegt in der Luft. https://www.faz.net/aktuell/technik-motor/technik/so-koennen-baeume-mit-drohnen-gepflanzt-werden-16783098.html (abgerufen am 06.07.2021).

Plattform Lernende Systeme (Hrsg.) (2019a): Lernende Systeme im Gesundheitswesen – Bericht der Arbeitsgruppe Gesundheit, Medizintechnik, Pflege, München.

Plattform Lernende Systeme (Hrsg.) (2019b): Retten, schützen, erkunden. Lernende Systeme in lebensfeindlichen Umgebungen. Potenziale, Herausforderungen und Gestaltungsoptionen. Bericht der Arbeitsgruppe Lebensfeindliche Umgebungen, München. https://www.plattform-lernende-systeme.de/files/Downloads/Publikationen/AG-7_Bericht_web_final.pdf (abgerufen am 06.07.2021).

Schnalzer, K. et al. (2021): TransWork – Transformation der Arbeit durch Digitalisierung. In: Bauer, W. et al. (Hrsg.), Arbeit in der digitalisierten Welt. Praxisbeispiele und Gestaltungslösungen aus dem BMBF-Förderschwerpunkt. Springer-Nature, Berlin, S. 1 – 15.

Stock-Homburg, R. & Merkle, M. (2019): Roboter und KI in der Arbeitswelt – Szenarien, Chancen und Herausforderungen. In: BMWi (Hrsg.), KI und Robotik im Dienste der Menschen. Eine Herausgeberschrift der AG 5 – Arbeit, Aus- und Weiterbildung der Plattform Industrie 4.0. https://www.bmwi.de/Redaktion/DE/Publikationen/Industrie/industrie-4-0-ki-und-robotik.pdf?__blob=publicationFile&v=4 (abgerufen am 10.08.2021).

Thornton, T. P. (1964): Terror as a Weapon of Political Agitation. In: Harry Eckstein (Hrsg.), Internal War: Problems and Approaches, London: Free Press of Glencoe, S. 87.

UNICRI & UNOCT (2021): Algorithms and Terrorism: The Malicious Use of Artificial Intelligence for Terrorist Purposes. https://www.un.org/counterterrorism/sites/www.un.org.counterterrorism/files/malicious-use-of-ai-uncct-unicri-report-hd.pdf (abgerufen am 10.08.2021).

Wirtz et al. (2019): Maschinelles Lernen »On the Edge«. https://www.iais.fraunhofer.de/de/publikationen/studien/whitepaper-machine-learning-on-the-edge.html (abgerufen am 24.11.2021).

Über dieses Whitepaper

Vorliegendes Whitepaper wurde auf der Basis der Expertise und Diskussion von Mitgliedern der Plattform Lernende Systeme erstellt. Federführend waren Prof. Dr. Jürgen Beyerer für die Arbeitsgruppe Lebensfeindliche Umgebungen und Prof. Dr. Jörn Müller-Quade für die Arbeitsgruppe IT-Sicherheit, Privacy, Recht und Ethik. Darüber hinaus waren Mitglieder weiterer Arbeitsgruppen der Plattform Lernende Systeme beteiligt: Arbeit/Qualifikation, Mensch-Maschine-Interaktion – Mobilität und intelligente Verkehrssysteme – Gesundheit, Medizintechnik, Pflege.

Autorinnen und Autoren der Plattform Lernende Systeme
Prof. Dr. Jürgen Beyerer, Fraunhofer-Institut für Optronik, Systemtechnik und Bildauswertung (Projektleitung)
Prof. Dr. Jörn Müller-Quade, Karlsruher Institut für Technologie (Projektleitung)
Prof. Dr. Albert Albers, Karlsruher Institut für Technologie – IPEK
Dr. Detlef Houdeau, Infineon Technologies AG
Dr. Igor Tchouchenkov, Fraunhofer-Institut für Optronik, Systemtechnik und Bildauswertung
Dr. Jeronimo Dzaack, ATLAS ELEKTRONIK GmbH
Dr. Matthieu-P. Schapranow, Hasso-Plattner-Institut für Digital Engineering gGmbH
Nadine Reißner, KUKA Deutschland GmbH
Dr. Rahild Neuburger, Ludwig-Maximilians-Universität München
Prof. Dr. Sascha Stowasser, Institut für angewandte Arbeitswissenschaft (ifaa)

Gastautoren
Simon Rapp, Karlsruher Institut für Technologie – IPEK
Sebastian Terstegen, Institut für angewandte Arbeitswissenschaft (ifaa)

Redaktion
Jan Biehler, Geschäftsstelle der Plattform Lernende Systeme
Stephanie Dachsberger, Geschäftsstelle der Plattform Lernende Systeme
Dr. Maximilian Hösl, Geschäftsstelle der Plattform Lernende Systeme
Alexander Mihatsch, Geschäftsstelle der Plattform Lernende Systeme
Christine Wirth, Geschäftsstelle der Plattform Lernende Systeme

Anhang 1: Übersicht Angriffsziele

KI-Systeme können in verschiedenen Dimensionen angegriffen und manipuliert werden. Eine Grundlage für die Auswahl und Gewichtung passender Schutz- und Abwehrmaßnahmen bilden 1) mögliche konkrete Angriffsziele und 2) mögliche Angriffe und Durchführungsstörungen einzelner Missionen, Aufträge und Handlungen.

1 Mögliche konkrete Angriffsziele für KI-Systeme:

Änderung der Systemfähigkeiten durch Manipulation von

- Lerndaten
- Wissen/Daten
- Eigensoftware (z. B. Fähigkeiten oder Lernalgorithmen)
- Abwehrkomponenten
  - Missbrauchsdetektoren
  - Kontrollkomponenten
  - Schutzmechanismen
- Kommunikation

Zugang zu geschützten Daten verschaffen

- Personendaten
- Technische Daten (interne Daten, Daten über Umgebung, Schlüssel etc.)

Beschädigung oder Diebstahl des Systems oder von Systemkomponenten
Störung und Manipulation von Komponenten

- Sensoren (z. B. Akustik, Vision)
- Effektoren (z. B. Greifen, Transport)
- Interne Module und Regelsätze (z. B. Navigationsregeln)

2 Mögliche Angriffe und Durchführungsstörungen einzelner Missionen, Aufträge und Handlungen beispielsweise durch:

Einschleusen (und ggf. auch Ausführungsvorbereitung) von verbotenen oder schädlichen/gefährlichen Aufträgen

- von nicht befugten Personen/Stellen
- von befugten Personen/Stellen

Störung der Vorbereitung oder Ausführung eines Auftrags (bzw. einzelner Handlungen)

- Störung der Vorbereitung
- Aufschiebung oder Aufhebung der Ausführung
- Manipulation

Störung der Auswertung bzw. Nachbereitung eines Auftrags

- Manipulation der Ergebnisse (z. B. durch Fremdeingriffe oder -einflüsse)
- Manipulation der Logeinträge (Files bzw. Datenbanken)
- Manipulation der Interpretation der Ergebnisse (z. B. zur Ausführungsänderung weiterer Aufträge)

Anhang 2: Übersichtstabelle Schutz vor Missbrauch

Tabelle 2: **Beispiele für Schutzziele, Angriffsszenarien und Gegenmaßnahmen bei der Missbrauchsabwehr von KI-Systemen**

Schutzziele	Unterziele	Angriffsszenarien	Gegenmaßnahmen
KI-System selbst	Gesamtsystem	Diebstahl, Beschädigung, Manipulation	• Schutz von Wartungssystemen[11] • Zugangsbeschränkungen • Schutz des Systems
	Systemteile	Diebstahl, Beschädigung, Manipulation, Austausch	
	Daten/Wissen	Diebstahl, Manipulation, Austausch	• Verschlüsselung • Verteilung der Daten • Schutz von Wartungssystemen
	Eigensoftware	Manipulation, Diebstahl, Austausch	• Absicherung (etwa durch Kapselung und Verschlüsselung) • Verteilung der Software • Schutz von Wartungssystemen
Auftrag/Mission	Gesamtauftrag	• Manipulation der Einsatzumgebung • Manipulation/Beschädigung des KI-Systems (Hardware/Software/Daten) • Verzögerungen, Unterbrechungen, Manipulationen des Auftrags • Unterbrechung der Versorgung/Kommunikation	• Detektion und Vermeidung verbotener bzw. schädlicher/gefährlicher Handlungen des Systems • Detektion von Abweichungen vom geplanten Verlauf
	Teile des Auftrags		
Umgebung	Gesamte Umgebung	Manipulation der Einsatzumgebung	• Detektion von Abweichungen vom Umgebungsmodell • Detektion und Vermeidung verbotener bzw. schädlicher/gefährlicher Handlungen des Systems • Geofencing/Geotargeting zum Schutz bestimmter Bereiche
	Menschen	Einwirkung auf Menschen	
	Andere KI-Systeme	Initiierung/Verhinderung bestimmter Aufträge/Handlungen des KI-Systems	
	Bestimmte Situationen bzw. Gebiete/ Bereiche	Manipulation der Einsatzumgebung	

Anmerkung: In der Tabelle wird davon ausgegangen, dass angemessene allgemeine Schutzmaßnahmen (wie etwa Zugangsbeschränkungen; siehe hierzu auch Seite 7 des Whitepapers) durchgeführt werden. Es werden vorwiegend Technologien und Maßnahmen aufgelistet, die auf spezifische Besonderheiten von KI-Systemen zielen.

11 Führen Wartungssysteme Arbeiten an mehreren KI-Systemen durch, können durch ihre Manipulation alle Systeme gefährdet werden. Besitzen Wartungssysteme Lernfähigkeiten, sind für ihren Schutz alle in der Tabelle aufgeführten Maßnahmen zielführend.

Impressum

Herausgeber
Lernende Systeme –
Die Plattform für Künstliche Intelligenz
Geschäftsstelle | c/o acatech
Karolinenplatz 4 | 80333 München
www.plattform-lernende-systeme.de

Gestaltung und Produktion
PRpetuum GmbH, München

Stand
März 2022

Bildnachweis
Peera_Sathawirawong/Adobe Stock/Titel

Empfohlene Zitierweise
Empfohlene Zitierweise: Beyerer, Jürgen & Müller-Quade, Jörn et al. (2022): KI-Systeme schützen, Missbrauch verhindern – Maßnahmen und Szenarien in fünf Anwendungsgebieten. Whitepaper aus der Plattform Lernende Systeme, München.
https://doi.org/10.48669/pls_2022-2

Bei Fragen oder Anmerkungen zu dieser Publikation kontaktieren Sie bitte Johannes Winter (Leiter der Geschäftsstelle):
kontakt@plattform-lernende-systeme.de

Über die Plattform Lernende Systeme

Die Plattform Lernende Systeme ist ein Netzwerk von Expertinnen und Experten zum Thema Künstliche Intelligenz (KI). Sie bündelt vorhandenes Fachwissen und fördert als unabhängiger Makler den interdisziplinären Austausch und gesellschaftlichen Dialog. Die knapp 200 Mitglieder aus Wissenschaft, Wirtschaft und Gesellschaft entwickeln in Arbeitsgruppen Positionen zu Chancen und Herausforderungen von KI und benennen Handlungsoptionen für ihre verantwortliche Gestaltung. Damit unterstützen sie den Weg Deutschlands zu einem führenden Anbieter von vertrauenswürdiger KI sowie den Einsatz der Schlüsseltechnologie in Wirtschaft und Gesellschaft. Die Plattform Lernende Systeme wurde 2017 vom Bundesministerium für Bildung und Forschung (BMBF) auf Anregung des Hightech-Forums und acatech – Deutsche Akademie der Technikwissenschaften gegründet und wird von einem Lenkungskreis gesteuert.