Lernende Systeme
DIE PLATTFORM FÜR KÜNSTLICHE INTELLIGENZ

KI-Systeme schützen, Missbrauch verhindern

Whitepaper

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

Kurzfassung

Künstliche Intelligenz wird bereits in vielen gesellschaftlichen Bereichen eingesetzt, sei es im Gesundheitsbereich, in der Arbeitswelt, im Straßenverkehr oder im öffentlichen Raum. 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 frühzeitig mitgedacht werden. Auf diese Weise können passende Maßnahmen strategisch ergriffen werden, um Missbrauch vorzubeugen oder um mögliche Risiken bereits frühzeitig zu minimieren. Damit kann letztlich auch das Vertrauen in die Zuverlässigkeit und Sicherheit von KI-Systemen sichergestellt und gestärkt werden.

Expertinnen und Experten unter Federführung der beiden Arbeitsgruppen Lebensfeindliche Umgebungen sowie IT-Sicherheit, Privacy, Recht und Ethik der Plattform Lernende Systeme nehmen im Whitepaper das Thema Missbrauch von KI-Systemen in den Fokus und stellen geeignete Maßnahmen vor, wie Missbrauch wirksam vorgebeugt werden kann, vorrangig unter technologischen Aspekten. Die Autorinnen und Autoren empfehlen vor dem Hintergrund des jeweiligen Anwendungskontextes beim Einsatz der KI-Anwendung, 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 führen. Aus dieser Gesamtsicht lassen sich erforderliche Schutzmaßnahmen ableiten, die eingebettet in einer Gesamtstrategie Missbrauch verhindern können. Damit liefern sie einen grundsätzlichen Beitrag, sich offen mit dem Thema Missbrauch beim Einsatz von KI-Technologien auseinanderzusetzen. Die theoretischen Überlegungen werden anhand

von realistischen Anwendungsszenarien, bei denen jeweils einem „Worst Case“ ein „Best Case“ gegenübergestellt wird, konkretisiert und veranschaulicht.

Missbrauch – Theoretische Überlegungen

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 – im Speziellen von KI-Technologien – ist (Kapitel 2.1). Missbrauch, und damit letztlich auch die Ausführung des Missbrauchs, lässt sich als „Zweckentfremdung mit negativen Folgen“ spezifizieren, bei der fundamentale Werte wie körperliche und psychische Unversehrtheit, (demokratische) Freiheiten und Rechte, Privatheit oder auch materielle und immaterielle Werte und die Umwelt verletzt werden.

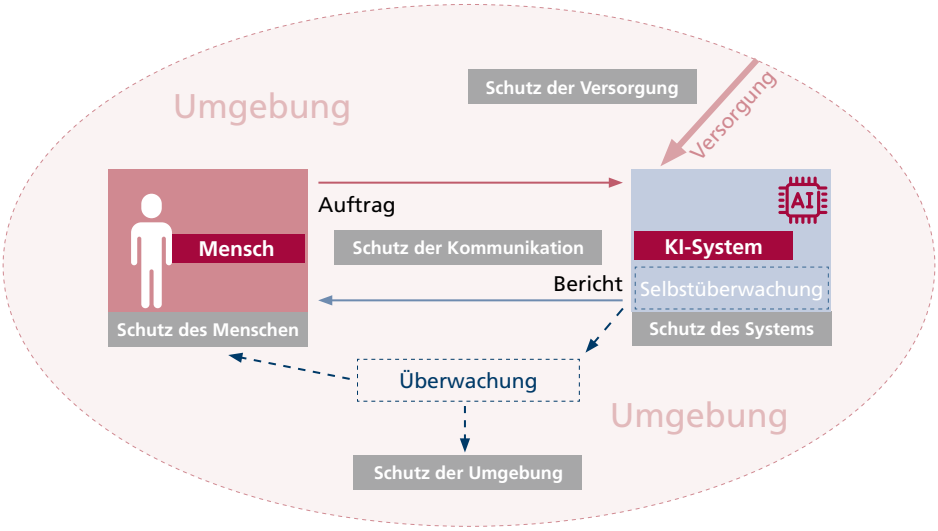
Beim Thema Schutz vor Missbrauch sind zuvorderst folgende Fragen zu beantworten: Was ist zu schützen? Welche Angriffsszenarien sind denkbar? Welche Schutzmaßnahmen sind möglich, angemessen und zulässig? Basierend auf den sich ergebenden Antworten und den daran anschließenden Analysen lassen sich konkrete technische wie organisatorische Schutzmaßnahmen ableiten. Dabei empfiehlt es sich, die verschiedensten Tätertypen, Motive sowie deren Angriffsziele genauer zu betrachten. Denn Auslöser des Missbrauchs ist menschliches Handeln, dessen Motor die verschiedensten Motive unterschiedlicher Akteure, wie Kriminelle oder Terroristen, sein können (Kapitel 2.2). Dieser Perspektivenwechsel, sich fiktiv in die Rolle der Tätertypen hineinzuversetzen, und zu antizipieren, wie diese versuchen könnten, den Missbrauch zu ihrem eigenen Vorteil oder sogar als Waffe zu nutzen, ermöglicht es, konkrete Schutzziele hinter den möglichen Angriffsszenarien zu erschließen (Kapitel 2.3).

Schutz- und Abwehrmaßnahmen

Da ein KI-System nicht für sich allein steht, sondern stets in einen spezifischen Anwendungskontext eingebettet ist (siehe Abbildung), sollten die Maßnahmen zur Missbrauchsabwehr grundsätzlich an den Hauptschutzzielen – dem KI-System selbst, dem Menschen sowie der Umgebung – ansetzen.

Aus der Analyse und Identifizierung, welche Motive und Angriffsziele auf ein KI-System – stets in Abhängigkeit zu seinem speziellen Anwendungskontext – möglich sein könnten, lassen sich verschiedenste Schutz- und Abwehrmaßnahmen technischer und organisatorischer Natur ableiten. Diese Maßnahmen können einerseits auf Systemen mit Künstlicher Intelligenz, wie Anomalieerkennung, Video- und Spracherkennung oder Identitätserkennung, beruhen oder andererseits auf Systemen ohne Künstliche Intelligenz (Kapitel 2.4). Die technischen Maßnahmen können dabei sowohl auf das eingesetzte KI-System selbst als auch auf die um das KI-System umgebende technische Infrastruktur, die selbst nicht KI-basiert ist, abzielen.

Allgemeine Maßnahmen zum Schutz von KI-Systemen gegen Missbrauch



Quelle: Eigene Darstellung

Damit die technischen Maßnahmen effektiv und zuverlässig greifen können, sollten diese durch organisatorische ergänzt werden, beispielsweise durch definierte einzuhaltende Regeln und Prozesse oder Prüf- und Zertifizierungsmaßnahmen.

Szenarien in realistischen Anwendungsgebieten

Die vorangestellten theoretischen Überlegungen zum Missbrauch und zu möglichen Schutzmaßnahmen gegen Missbrauch von KI-Systemen werden anhand von sieben Anwendungsszenarien aus dem Bereich Gesundheit, Freizeit, Mobilität und Arbeitswelt realistisch abgebildet (Kapitel 3). Unter anderem beschreibt das Whitepaper wie Kriminelle Ransomware einsetzen, um mit einem Phishing-Angriff auf ein Unternehmen Lösegeld zu erpressen. Im Worst Case-Szenario schaffen sie es, eine E-Mail in den Kommunikationsverkehr an alle Mitarbeitenden einzuschleusen, über die sie das gesamte Abrechnungssystem lahmlegen. Beim Szenario aus dem Bereich Mobilität führt eine terroristische Organisation einen Angriff auf das autonome Fahrzeug einer Geschäftsfrau über eine CAR-2X-Schnittstelle durch. So können sie direkt auf die Steuerungselektronik eingreifen und das Auto gezielt lenken. Eine mögliche Gefahr des Missbrauchs liegt auch bei einem Einsatz von Drohnen vor. Kriminelle manipulieren bei einem Fußballweltmeisterspiel die Drohnen über eine Schadsoftware so, dass diese als „Waffe“ gegen die Stadionbesucherinnen und Stadionbesucher missbraucht werden können.

Die aufgeführten Szenarien veranschaulichen realistisch, wie und wo Missbrauch schon heute oder in kommenden Jahren Realität werden könnte und welche konkreten und geeigneten Maßnahmen jeweils abzuleiten sind, um den „Worst Case“ durch geeignete und effektive Schutz- und Abwehrmechanismen in einen „Best Case“ umzuleiten bzw. von Beginn an dem Missbrauch vorzubeugen.

Auf der Website der Plattform Lernende Systeme 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.

Gestaltungsoptionen

Vor diesem Hintergrund lassen sich zu den bereits genannten Aspekten einige konkrete Gestaltungsoptionen (Kapitel 4) formulieren, die darauf abzielen, Missbrauch von KI-Systemen wirksam vorzubeugen. Letztlich bedarf es eines offenen gesellschaftlichen Diskurses, der Transparenz und Vertrauen in den Mittelpunkt stellt, ohne dass beim Thema Missbrauch Ängste geschürt werden.

- **Die Politik** sollte durch staatliche Maßnahmen Verantwortlichkeiten sowie Schadensregulierungen im Falle von Missbrauch klar definieren, auf die Entwicklung und Einhaltung von Standards und Zertifizierungen bei KI-Systemen mit einem hohen Missbrauchspotenzial einwirken und Forschungsvorhaben dahingehend fördern, frühzeitig die Chancen und Gefährdungspotenziale von KI-Systemen zu erkennen, um rechtzeitig Maßnahmen zur Vorsorge und zum Minimieren möglicher Risiken einzuleiten.

- **Forschungsinstitutionen** sollten sich verstärkt des Themas Missbrauch von KI-Systemen unter Berücksichtigung der Anforderungen an Sicherheit und Zuverlässigkeit annehmen, diesen Aspekt stärker bei der Qualifizierung von Fachpersonal berücksichtigen, publizierte Fälle untersuchen und geeignete Maßnahmen transparent nach außen kommunizieren.

- **Bildungsinstitute** sind aufgerufen, vermehrt Aufklärungsarbeit zum Thema Missbrauch von KI-Systemen zu leisten sowie das Bewusstsein für KI und den Umgang mit KI zu fördern. Einerseits, um das Thema gesellschaftsweit zu setzen und andererseits, um diesem bereits frühzeitig in schulischer und beruflicher Aus- und Weiterbildung einen festen Platz zukommen zu lassen.

- **Unternehmen** sind, auch in Zusammenarbeit mit Forschungseinrichtungen, aufgerufen, gemeinsam neue Lösungen frühzeitig zu entwickeln und umzusetzen, mögliche Ziele eines Missbrauchs zu bewerten, für einen sicheren und zuverlässigen Einsatz von KI-Systemen zu sorgen und dies bei der Weiterentwicklung des Personals zu berücksichtigen.

- **Die Zivilgesellschaft** ist aufgerufen, sich aktiv einzubringen, um die gesellschaftliche und nutzbringende Perspektive auf das Thema Missbrauch klar zu platzieren.

Mit der Fokussierung auf das Thema Missbrauch von KI-Systemen leistet das Paper einen wichtigen Beitrag, einen frühzeitigen und proaktiven Schutz nicht nur in den Blick zu nehmen, sondern ‚tatsächlich' durch geeignete Maßnahmen zu implementieren, um sich gegen Missbrauchsversuche beim Einsatz von KI-Systemen zu wappnen. Die vorgestellten Grundsatzüberlegungen sowie die konkreten Maßnahmen eröffnen Perspektiven für Handlungsmaßnahmen und bieten zudem einen Orientierungsrahmen, Missbrauch von KI-Systemen effektiv vorzubeugen und entgegenzuwirken.

Impressum
Herausgeber: Lernende Systeme – Die Plattform für Künstliche Intelligenz | Geschäftsstelle | c/o acatech | Karolinenplatz 4 | D-80333 München | kontakt@plattform-lernende-systeme.de | www.plattform-lernende-systeme.de | Folgen Sie uns auf Twitter: @LernendeSysteme | Stand: März 2022 | Bildnachweis: Peera_Sathawirawong/Adobe Stock/Titel

Diese Kurzfassung entstand auf Grundlage des Whitepapers *KI-Systeme schützen, Missbrauch verhindern – Maßnahmen und Szenarien in fünf Anwendungsgebieten*. München, 2022. Es wurde erstellt von Mitgliedern der Arbeitsgruppen Lebensfeindliche Umgebungen sowie IT-Sicherheit, Privacy, Recht und Ethik. https://doi.org/10.48669/pls_2022-2

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

acatech
DEUTSCHE AKADEMIE DER TECHNIKWISSENSCHAFTEN