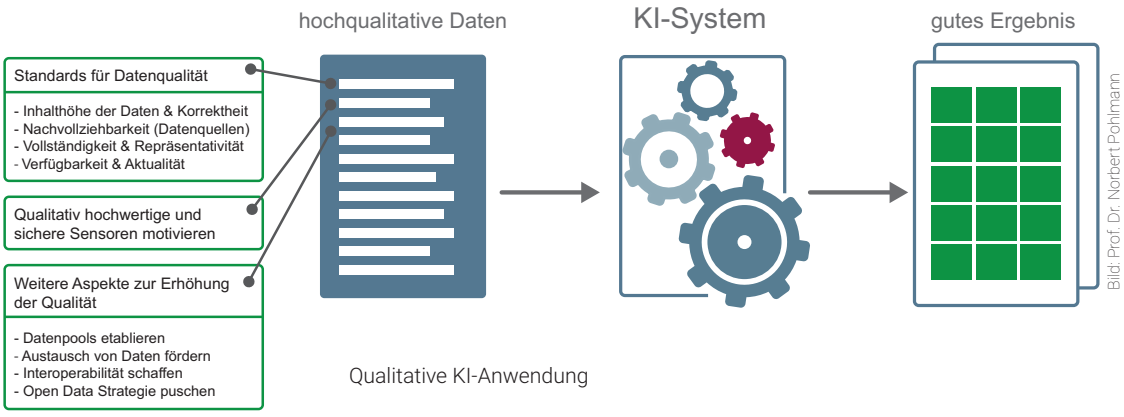
Mechanismus für gute Ergebnisse

Wie können wir der KI vertrauen?

Unternehmen setzen zunehmend auf KI oder planen, dies künftig zu tun. Doch die große Euphorie bleibt in der Industrie aus guten Gründen noch aus. Zum einen fehlt die kritische Masse an Einsatzszenarien, weswegen Unsicherheit besteht, welche Handlungsfelder nachhaltige Erfolge versprechen. Zum anderen ist die Frage der Zuverlässigkeit zu klären, also wie valide KI-generierte Ergebnisse wirklich sind. Im Folgenden geht es um die Mechanismen, die gute Ergebnisse sicherstellen helfen.



Bevor KI eine breite Akzeptanz in Unternehmen und Gesellschaft erfährt, müssen einige Herausforderungen gelöst werden. Doch letztendlich wird die Vertrauenswürdigkeit der KI-Technologie als Schlüssel für deren Erfolg gesehen. Aber wie kann diese aufgebaut werden? Ausgehend von der Definition, dass Vertrauen als die subjektive Überzeugung von der Richtigkeit einer Aussage und von Handlungen zu verstehen ist, kann ein KI-System generell als vertrauenswürdig eingestuft werden, wenn es sich für den vorgesehenen Zweck immer wie erwartet verhält. Daraus lässt sich folgern, dass Vertrauenswürdigkeit nachweisbar ist. In Bezug auf KI sind somit grundlegend folgende Faktoren relevant, die im Weiteren erläutert werden:

- Die Eingangsdaten der KI müssen eine hohe Qualität für den Anwendungsfall aufweisen.
- Die IT-Anwendung und das KI-System sind von KI- und Anwendungsexperten konzipiert sowie manipulationssicher und vertrauenswürdig umgesetzt.
- Ergebnisse nachzuvollziehen wird ermöglicht.
- Bei der Entwicklung und Anwendung werden jeweils ethische Grundsätze eingehalten.

Qualität der Eingangsdaten

Grundsätzlich basiert die Entwicklung und im Weiteren der Einsatz von KI-basierten Anwendungen auf Daten – etwa für das Trainieren des KI-Algorithmus sowie auch für dessen Nutzung. Unter dieser Prämisse ist eine differenzierte Analyse der Daten – bezüglich ihres Werts respektive ihrer Aussagekraft im Sinne der Aufgabenstellung – beider Kategorien ein essentieller erster Schritt zur Sicherstellung der Vertrauenswürdigkeit von KI-basierten Anwendungen. Denn aufgrund ihrer hohen Relevanz entscheidet deren Auswahl und Qualität maßgeblich über das Ergebnis. Aus diesem Grund sollte es obligatorisch sein, entsprechend Positionen im Unternehmen zu konstituieren, die für das Modell der Datengewinnung und -nutzung zuständig sowie für die Kontrolle der ordnungsgemäßen Umsetzung verantwortlich sind. Gemäß vorgegebener Kriterien lässt sich der Standard der Datenqualität für KI-Systeme sowohl etablieren als auch validieren. Im Einzelnen sind dabei unter anderem Vollständigkeit, Repräsentativität, Nachvollziehbarkeit, Aktualität und Korrektheit zu berücksichtigen.

Vollständigkeit der Daten

Die Grundvoraussetzung für Vollständigkeit ist, dass ein Datensatz alle notwendigen Attribute und Inhalte enthält. Kann die Vollständigkeit der darin inkludierten Daten nicht garantiert werden, entsteht daraus potentiell das Problem von irreführenden Tendenzen, was letztendlich zu falschen oder diskriminierenden Ergebnissen führt. Dieses Phänomen tritt unter anderem bei Predictive Policing-Systemen auf: Wenn beispielsweise die Datenerhebung zu Kriminalitätsdelikten von vorneherein massiv in definierten Stadtvierteln stattfindet und dies im Kontext mit bestimmten Merkmalen wie Herkunft und Alter geschieht, ergibt sich daraus im Laufe der Zeit, dass dort bestimmte Bevölkerungsgruppen stärker überwacht und durch die häufiger durchgeführten Kontrollen letztendlich per se kriminalisiert werden. Der (vermeintliche) Tatbestand kann jedoch unter Umständen lediglich darauf basieren, dass entsprechende Vergleichswerte unter Berücksichtigung der gleichen Merkmalen aus anderen Stadtvierteln nicht im adäquaten Maße erhoben wurden. Vollständigkeit bedeutet somit keinesfalls, wahllos möglichst viele Daten zu erfassen – entscheidend ist die Auswahl.

Repräsentativität der Daten

Die Repräsentativität zeichnet sich dadurch aus, dass die Daten eine tatsächliche Grundgesamtheit und somit entsprechend die Realität abbilden, die stellvertretend im Sinne der Aufgabenstellung ist. Sind die Daten nicht repräsentativ, hat dies zur Folge, dass daraus ein Bias resultiert. Dieses Phänomen tritt beispielsweise im Recruiting von Führungskräften auf, wenn hier größtenteils Daten aus der Vergangenheit berücksichtigt werden und in dieser Zeit überwiegend Männer in Führungspositionen waren. Mit der Konsequenz, dass die KI-basierte Anwendung daraus folgern müsste, dass Männer für diese Positionen qualifizierter seien. Ergebnisse wie diese zeigen, dass durch KI-Systeme nicht zwangsläufig Objektivität erreichbar ist.

Nachvollziehbarkeit der Daten

Für die Überprüfung der Datenqualität ist es essentiell, dass nachvollzogen werden kann, aus welchen Quellen die verwendeten Daten stammen. Sind die Quellen nicht transparent, das heißt nicht nachvollziehbar, ist es nicht möglich eine notwendige Validierung der Daten vorzunehmen, was sich letztendlich auf deren Qualität negativ auswirken kann. Für eine bestmögliche Bewertung und Messung sowohl der Datenqualität als auch der Qualität der Quellen sowie der Ableitung gezielter Verbesserungsmaßnahmen, müssen im Vorfeld entsprechend Vorgaben definiert werden. Hierfür gilt es, die für den Prozess relevanten Kriterien zu bestimmen, etwa Konsistenz oder Einheitlichkeit. Anhand der gewählten Kriterien ist es dann möglich, die erhobenen Daten bezüglich ihrer konsistenten Qualität zu überprüfen. Hierbei sind noch zwei relevante Aspekte zu bedenken: Zum einen kommen oft Daten aus unterschiedlichen Quellen mit verschiedenen Formaten, die vor dem Einsatz auf ihre Utilität verifiziert werden müssen. Zum anderen ist die Nachvollziehbarkeit – gerade im Produktionsumfeld – auch durch die Förderung von qualitativ hochwertigen und sicheren Sensoren abhängig.

Aktualität der Daten

Die grundsätzliche Idee beim maschinellen Lernen oder KI ist die Extraktion von Wissen aus Daten. Aus diesem Grund muss sichergestellt werden, dass die generierten respektive verwendeten Daten auch die passenden Informationen und Erfahrungen enthalten, um mit den KI-Algorithmen das Problem richtig und vertrauenswürdig zu lösen. Nicht zuletzt aufgrund der Tatsache, dass Menschen sich nicht linear verhalten, können veraltete Daten zu falschen Ergebnissen führen. Aus diesem Grund sollten – in Abhängigkeit von der Anwendung – möglichst die aktuellsten Daten verwendet werden.

Korrektheit der Daten

Die Daten müssen mit der Realität übereinstimmen und damit für die Anwendung korrekt sein. Die Auswahl der Daten bedingt, dass diese Anforderungen mit einer detaillierten Analyse ermittelt wurden – als Methode kann hier das Mapping gegen Daten, deren Korrektheit bestätigt ist, eingesetzt werden oder definierte, abgestimmte Plausibilitätsregeln. So lässt sich sicherstellen, dass zwischen den – zur Entwicklung

KI nicht immer die beste Wahl

Anwendungsfall: Für die Berechnung der aktuellen Gefahrenlage im Online-Banking wurde für ein Alarmsystem mithilfe von Algorithmen des maschinellen Lernens entwickelt. Die Effektivität des Alarmsystems wurden anhand echter Betrugsfälle einer Bankengruppe in Deutschland evaluiert.

Ziel: Tagesaktuelle Warnungen sollten auf eine punktuell erhöhte Gefahrenlage für das Online-Banking hinweisen.

Ansatz des Alarmsystems: Sicherheitskennzahlen zum Betrug identifizieren und in das System integrieren. Auf dieser Basis sollte mittels KI die Gefahrenlage bestimmt werden.

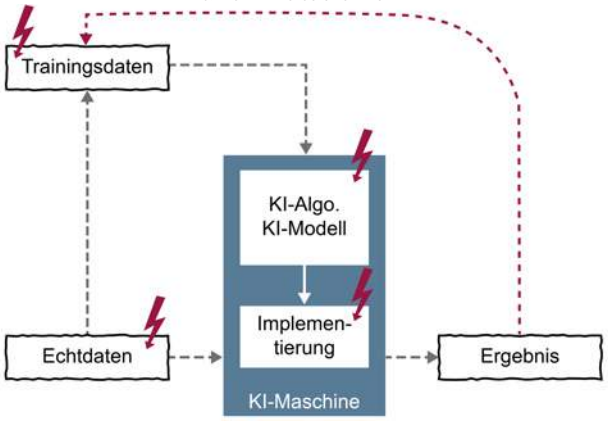
Auszug der Sicherheitskennzahlen zum Betrug:
- E-Mail (Phishing-Angriff) – Daten aus dem Projekt: Spam Archive
- Phishing-Webseiten – Daten aus dem Projekt: PhishTank
- Infektionen von Banking-Trojaner – Daten von großen Herstellern von Antivirus-Produkten
- Relevante Schwachstellen – Daten wurden aus der National Vulnerability Database NVD extrahiert.

Resultat: Mit dem Alarmsystem für Online-Banking konnte sehr gut bewiesen werden, dass sich die Bedrohungslage und somit das aktuelle Risiko eines Angriffs auf einen Bankkunden beim Online-Banking aufzeigen lässt. Würde dieses Ergebnis den Bankkunden mit dem Hinweis zur Verfügung gestellt, dass die Transaktion aufgrund der aktuellen Situation später besser durchzuführen wäre, könnten viele Schadensfälle vermieden werden.

Ergebnisverwertung: Trotz dieser positiven Ergebnisse wurde das Alarmsystem für Online-Banking nicht eingeführt. Da der momentan entstehende Schaden jährlich (nur) zirka 65 Millionen beträgt, wird dieses Risikopotential von den Verantwortlichen als weniger gravierend bewertet, als der unkalkulierbare Schaden, der entstünde, wenn die Bankkunden das Online-Banking aufgrund des Alarmsystems als nicht vertrauenswürdig einschätzen.

Entscheidung Mensch gegen KI: Auch wenn das Einsatzszenario der KI durchaus nützlich wäre, ist deren Einsatz in manchen Fällen aufgrund höher zu bewertender weicher Faktoren nicht empfehlenswert.

Bild: Prof. Dr. Norbert Pohlmann



Manipulation von KI-Systemen

oder im Weiteren in der Anwendung – genutzten Daten und der Realität keine Diskrepanz besteht.

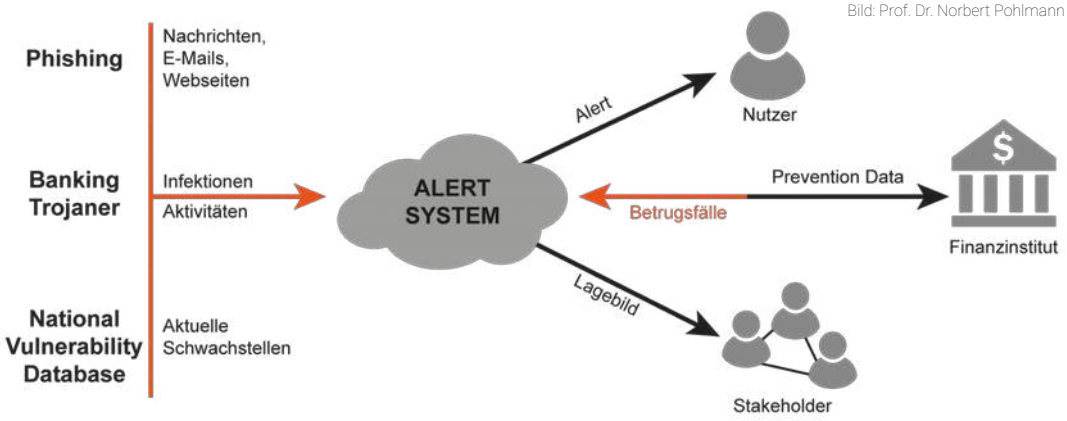
Vertrauenswürdige Umsetzung

Die Vertrauenswürdigkeit wird im Kern durch die Infrastruktur des KI-Systems manifestiert. Neben dem Einsatz qualitativ hochwertiger KI-Technologien aus sicherer Herkunft ist auch die Kooperation von Experten der jeweiligen Anwendungsdomänen zu fördern sowie mit den Spezialisten im Bereich KI-Entwicklung und IT-Sicherheit, die für das entsprechende Einsatzgebiet KI-Systeme vertrauenswürdig konzipieren und umsetzen können. So ist für den physischen Schutz der KI-Systeme der Stand der Technik an IT-Sicherheitsmaßnahmen bezüglich Integrität, Vertraulichkeit, Datenschutz sowie Verfügbarkeit zu definieren und umzusetzen. Dadurch lassen sich die Manipulationsmöglichkeiten sowie der Missbrauch von KI-Anwendungen und der genutzten Daten reduzieren.

Nachvollbare Ergebnisse

Künftig wird KI zunehmend in automatisierte IT-Systeme eingebunden sein. Da hier die Eingriffsmöglichkeiten extrem restriktiv sind, muss mit Tests, Simulationen und Validierungen sichergestellt werden, dass die KIs die intendierten Funktionen korrekt durchführen. Ebenso kommt der Definition der Verantwortung sowie der daraus resultierenden Haftung eine besondere Bedeutung zu, auch um Vertrauen aufzubauen. Zudem sollte der Mensch – wann immer dies möglich ist – als kontrollierender Faktor gemäß dem Konzept 'Keep the human in the loop' in den Kreislauf der KI eingebunden werden. Diese Forderung

Bild: Prof. Dr. Norbert Pohlmann



Alarmsystem für Online-Banking (siehe ganz links)

macht die grundlegende Übereinkunft notwendig, dass das Ergebnis der KI als Handlungsempfehlung zu verstehen ist und somit dem Mensch die Entscheidungsfreiheit obliegt, ob er dieser folgt oder auch nicht. Dieser Grundsatz dient im Weiteren dazu, die Selbstbestimmtheit der Nutzer zu erhalten, wodurch unter anderem die Vertrauenswürdigkeit erhöht wird.

Vertrauen einhandeln

In erster Linie liegt es an den Unternehmen, durch schlüssiges Handeln Vertrauen aufzubauen – bei ihren Mitarbeitern, bei ihren Zielgruppen und in der Gesellschaft insgesamt, um die Akzeptanz für die KI-Anwendung bei den Nutzern zu erzielen. Dies lässt sich über die Deklaration grundlegender Prinzipien zur Entwicklung sowie zum Einsatz von KI-basierten Anwendungen erreichen, in der wesentliche Handlungsprämissen wie der Vorrang des menschlichen Handels sowie die Verpflichtung zu gesellschaftlichem und ökologischen Wohlergeben von dem Unternehmen niedergelegt werden.

Fazit

Der Einsatz von KI macht bestimmte Analysen, die als Basis einer komplexen Urteilsfindung dienen, und viele andere Aktivitäten erst jetzt möglich oder führt dazu, dass diese verbessert werden können. Doch aufgrund der Tatsache, dass die Prozesse hin zur KI-Entscheidung in einer Blackbox ablaufen, kann es unmittelbar keine Gewissheit darüber geben, wie die Ergebnisse zustande kommen und ob sie valide sind. Eine Vielzahl von bislang veröffentlichten Beispielen bezüglich Verzerrungen oder im Hinblick auf das Manifestieren von Vorurteilen macht deutlich, dass Entwicklung und Einsatz KI-basierter Anwendungen einer hohen Methodenkompetenz bedarf – insbesondere in Bezug auf die Gestaltung von Modellen zur Erfassung und Nutzung von Daten. Insgesamt müssen diese Herausforderungen gelöst werden, denn die daraus resultierende Vertrauenswürdigkeit ist maßgeblich für die Akzeptanz der KI-Anwendung. Zusätzlich sollte hier durch eine ernst gemeinte Aufklärungsarbeit aller beteiligten Parteien bezüglich der Chancen und Risiken der KI-Technologie Transparenz und damit Verständnis geschaffen werden. Letztendlich ist es ebenso notwendig auch die Ethik zu berücksichtigen – Unterstützung bei der Umsetzung von KI-Systemen gemäß ethischer Kriterien wird mittlerweile auch Tool-basiert angeboten. ■

www.xethix-empowerment.de

Autoren

Ulla Coester ist CEO bei Xethix-empowerment, Digitale Ethik.
Professor Dr. Norbert Pohlmann
ist Leiter des Institut für Internetsicherheit if(is).