Vertrauenswürdigkeit von KI: Klare Anforderungen an die KI-Anbieter

Ulla Coester, Norbert Pohlmann

Grundsätzlich ist es erstrebenswert vertrauen zu können. Denn ohne Vertrauen wären Menschen nicht handlungsfähig. Insbesondere nicht im Kontext der Digitalisierung, da es aufgrund der zunehmenden Komplexität der Technologie ständig schwieriger wird diese zu bewerten und darauf basierend Entscheidungen zu treffen. Jedoch muss dieses Vertrauen gerechtfertigt sein und dafür bedarf es bestimmter Voraussetzungen, die es Anwendern ermöglicht, innovative Technologien wie KI zu nutzen. Die Frage, was diesbezüglich erfüllt sein muss und wann Unternehmen letztendlich dazu bereit sind, KI-Lösungen zu implementieren, steht im Mittelpunkt des Forschungsprojektes „TrustKI – Vertrauenswürdigkeits-Plattform für KI-Lösungen und Datenräume". Erste Antworten darauf liefert die kürzlich abgeschlossene Anwender-Studie „TrustKI".

Die Veränderungen, die mit der Nutzung von KI einhergehen, werden absehbar sowohl Unternehmen als auch die Gesellschaft insgesamt sowie jeden einzelnen Menschen zunehmend stärker tangieren. Diese Entwicklung lässt sich generell nicht mehr aufhalten. Denn die Verbreitung von ChatGPT – OpenAI gelang trotz offen kommunizierter Limitationen die bislang schnellste Marktdurchdringung einer Innovation: innerhalb von fünf Tagen meldeten sich eine Million Nutzer bei der Plattform an – oder aktuell des KI-Copilot, der aller Voraussicht nach auch schnellstens in den Unternehmen Einzug halten wird, zeigen bereits heute eindrucksvoll, dass Anwender sich mit KI-Lösungen auseinandersetzen müssen. Denn aus deren Anwendung können – teilweise auch unbeabsichtigte – Effekte resultieren.

Doch aufgrund der Komplexität der Technologie ist es für den Einzelnen nicht trivial, sich damit im notwendigen Umfang zu beschäftigen. Von daher gilt es zu klären, was für Anwender in Unternehmen tatsächlich hinsichtlich einer fundierten Einschätzung im Rahmen ihres Entscheidungsprozesses essenziell ist.

Es braucht reelles Vertrauen in KI

Zur Klärung der Fragestellung, welche Kriterien vor dem Hintergrund eines Entscheidungsprozesses aus Sicht der Anwender relevant sind, wurde im Rahmen des Forschungsprojekts „TrustKI – Vertrauenswürdigkeits-Plattform für KI-Lösungen und Datenräume" Ende vergangenen Jahres die Anwender-Studie „TrustKI" durchgeführt. Befragt wurden dabei insgesamt 263 Führungskräfte, was aus ihrer Sicht diesbezüglich notwendig ist. Als eines der zentralen Ergebnisse hat sich gezeigt, dass die befragten Anwender dem Aufbau eines reellen Vertrauensverhältnisses basierend auf Vertrauenswürdigkeit eine hohe Bedeutung beimessen.

TrustKI: Die Studie auf den Punkt gebracht

Im Zuge der Anwender-Studie haben die befragten Personen klare Anforderungen formuliert, was aus ihrer Sicht für einen nachhaltigen Vertrauensaufbau notwendig ist. Die wesentlichen Erkenntnisse lassen sich folgendermaßen zusammenfassen: Um einen Nachweis ihrer Vertrauenswürdigkeit zu erbringen, müssen KI-Anbieter unter anderem verbindlich relevante Kriterien hinsichtlich des Aspekts „Integrität" – der vorrangig beinhaltet, dass alle Kriterien der Vertrauenswürdigkeit und hier insbesondere die ethischen Dimensionen sowie die gebotene Sorgfaltspflicht Berücksichtigung finden – erfüllen.

Zudem besteht der Anspruch auf eine „holistischen Transparenz" – das heißt, die Vertrauenswürdigkeit resultiert nicht mehr einzig aus der Transparenz der KI-Lösung, sondern basiert ebenfalls auf einer transparenten Darstellung der Handlungsweise der KI-Anbieter.

Die wichtigsten Studien-Ergebnisse im Detail

Eine ethisch orientierte Handlungsweise der KI-Anbieter ist nachdrücklich erwünscht

Dem Aspekt „Integrität" wird seitens der Befragten eine hohe Relevanz beigemessen. Die Ergebnisse zeigen: Für Anwender ist es prinzipiell erforderlich, dass KI-Anbieter die gesellschaftlichen Werte und Normen anerkennen – entsprechend wollen 78 % wissen, wie der KI-Anbieter die Anforderungen der Gesellschaft bezüglich Ethik konkret umsetzt. In diesem Kontext sind sie besonders daran interessiert zu erfahren, ob „ein Ethik-Gremium im Unternehmen etabliert wurde und welche Verantwortung sowie Befugnisse dieses reell hat".

Ebenfalls als notwendig wird erachtet, mehr darüber zu erfahren, auf welchem Wege der KI-Anbieter sicherstellt, dass seine Mitarbeitenden die vorgegebenen ethischen Werte umsetzen können. Bemerkenswert ist in diesem Kontext, dass vornehmlich Befragte mit umfangreichen KI-Kenntnissen einen insgesamt hohen Informationsbedarf dahingehend haben, mit welchen Maßnahmen den ethischen Anforderungen der Gesellschaft nachgekommen wird.

Diese allgemeine Position lässt sich noch weitergehend präzisieren: So erwarten 68 % der Befragten eine wohlwollende Haltung von Seiten der KI-Anbieter. Diese offenbart sich nach Meinung der Befragten unter anderem darin, dass KI-Anbieter ihre Verantwortung gegenüber der Gesellschaft anerkennen und dementsprechend beim Inverkehrbringen von KI-Lösungen gemäß dem Nicht-Schädigungsprinzip verfahren. Hinsichtlich dieses Aspekts ist es für die Mehrzahl der Befragten unter anderem substanziell, dass sie darüber informiert werden, auf welche Funktionalitäten der KI-Anbieter zum Wohle des Kunden verzichtet.

Ebenso besteht ein hoher Informationsbedarf bezüglich der Folgenabschätzung. Hier wollen die befragten Führungskräfte nicht nur mehr darüber erfahren, ob der KI-Anbieter eine Folgenabschätzung vornimmt, sondern auch wie diese umgesetzt wird. Dass Anwender auf die Einhaltung der Sorgfaltspflicht großen Wert legen, zeigt sich auch daran, dass über 50 % eine zielgruppengerechte Aufklärung über potenzielle Folgen fordern.

Anwender erwarten eine holistische Transparenz

Grundsätzlich wird im Kontext von KI der Begriff der Vertrauenswürdigkeit in hohem Maße mit der Forderung nach Transparenz assoziiert. Wobei diese momentan vorrangig auf die KI-Lösung referenziert und die Anforderungen nach Erklärbarkeit, Interpretierbarkeit und Nachvollziehbarkeit umfasst – sodass die von der Anwendung getroffenen Entscheidungen auch klar zu deuten und darzulegen sind – oder die Reproduzierbarkeit der Ergebnisse sowie allgemein auf deren Funktionalität.

Bei dem Prozess zum Aufbau von Vertrauenswürdigkeit liegt die Herausforderung prinzipiell darin, die Perspektive der Vertrauensgeber – das heißt, der Anwender – zu erfassen und verstehen. Damit es Anwenderunternehmen möglich ist KI-Lösungen zu vertrauen gilt es somit, den allgemein anerkannten und bereits etablierten Maßstab zu erweitern – das heißt, die Vertrauenswürdigkeit sollte (oder kann) nicht einzig aus der Transparenz der KI-Lösung resultieren, sondern muss seitens der Anwender auf den KI-Anbieter ausgeweitet werden können.

Entsprechend gibt es seitens der Anwender dedizierte Kriterien, die im Sinne einer holistischen Transparenz von den KI-Anbietern erfüllt und dargestellt werden müssen. Unter anderem legen die Befragten Wert darauf, dass sie mehr Informationen zum verantwortungsvollen Umgang mit der Technologie, vor allem explizit im Kontext der Privatheit erhalten, auch weil dies wesentlich im Hinblick auf mögliche Manipulation ist.

Auf die Frage, ob der KI-Anbieter erläutern sollte, welche Prozesse er etabliert hat, um zu gewährleisten, dass sich kein Mitarbeitender über die festgelegten (ethischen) Werte und Regeln hinwegsetzen kann, gab die Mehrzahl der Befragten an, dass die Beantwortung dieser Frage für sie relevant ist.

Anwender wollen über die Kompetenz von KI-Anbietern informiert werden

Zutrauen und Zuverlässigkeit sind die beiden Vertrauenswürdigkeits-Aspekte, in denen sich die Kompetenz der KI-Anbieter manifestiert. Aus den Ergebnissen der Anwender-Studie lässt sich ableiten, dass die befragten Führungskräfte der Anwender-Studie an den Kriterien, die in diesem Kontext eine Rolle spielen, sehr interessiert sind. So ist es für einen Großteil von ihnen relevant, mehr über die Qualifikation der Mitarbeitenden zu erfahren, unter anderem hinsichtlich deren Erfahrung im speziellen Einsatzbereich der KI-Lösung.

Bemerkenswert ist die hohe Relevanz des Produktionsstandorts Deutschland – diesen erachten 79,5 % als wichtig oder sehr wichtig. Dies legt nahe, dass die verlässlichen gesetzlichen Regelungen – u. a. bezüglich des Datenschutzes und der EU-Produktsicherheitsvorschriften – dem Anwender die Gewissheit geben, dass er in geregelter Form zu seinem Recht kommen kann, wenn der KI-Anbieter das in ihn gesetzte Vertrauen missbraucht. Zudem ist dies potenziell ein weiterer Anhaltspunkt für die hohe Bedeutung, die einem gemeinsamen Wertverständnis hinsichtlich Entwicklung und Einsatz von KI-Lösungen beigemessen wird. Dies lässt wiederum darauf schließen, dass der übergeordnete Begriff der Transparenz und die damit inhärent assoziierte Zuverlässigkeit im Rahmen der Vertrauenswürdigkeit der KI-Anbieter als signifikant bewertet wird.

Fazit und Ausblick

Wie die Anwender-Studie zeigt, gibt es mittlerweile ein gängiges Verständnis dahingehend, dass die Implikationen resultierend aus dem Einsatz von KI allgemein, sowohl die Anwenderunternehmen als auch potenziell deren Kunden unmittelbar, sowie mittelbar die Gesellschaft, betreffen. Aufgrund dessen ist der Wunsch nach einer holistischen Transparenz – die sich in allen sieben Vertrauenswürdigkeits-Aspekten manifestiert – nachvollziehbar. Die Anwender möchten sichergestellt wissen, dass die Werte des KI-Anbieters mit den ihren vereinbar sind.

Die Ergebnisse der Anwender-Studie TrustKI werden zeitnah im Rahmen von Workshops mit KI-Anbieter diskutiert, mit dem Ziel die Interessen von Anwendern und KI-Anbietern auszutarieren.

TrustKI-Forschungsbericht:
Alle Ergebnisse der TrustKI-Umfrage zum Download.
(www.vertrauenswuerdigkeit.com)

Ulla Coester
Projektleiterin TrustKI
if(is) - Institut für Internet-Sicherheit an der Westfälischen Hochschule
45879 Gelsenkirchen

Prof. Dr. Norbert Pohlmann
Informatikprofessor und Direktor
if(is) - Institut für Internet-Sicherheit an der Westfälischen Hochschule
45879 Gelsenkirchen