ERKLÄRBARE KI:

Die **Vertrauenslücke** schließen

Funktionalität und Komplexität geschäftsorientierter KI-Anwendungen haben exponentiell zugenommen. DevOps-Tools, Security-Response-Systeme, Suchtechnologien und vieles mehr haben von den Fortschritten der KI-Technologie profitiert. Vor allem Automatisierungs- und Analysefunktionen haben die betriebliche Effizienz und Leistung gesteigert, indem sie komplexe oder informationsdichte Situationen verfolgen und darauf reagieren … | VON ALEXANDER ZACHOW

Die ständig wachsende Komplexität der KI-Modelle bringt jedoch ein eklatantes Problem mit sich: Transparenz. Viele hochmoderne KI-Modelle sind in ihrer Entscheidungsfindung so komplex geworden, daß selbst Fachleute nicht mehr nachvollziehen können, wie oder warum ein Modell seine Entscheidungen trifft. Dies wird oft als Black-Box-Problem bezeichnet. Erklärbare KI zielt darauf ab, dieses Problem anzugehen, indem KI-basierte Ergebnisse transparenter und damit verständlicher werden.

Im Folgenden werden die Bedeutung erklärbarer KI-Entscheidungen, die Herausforderungen bei ihrer Umsetzung und die Schlüsselkomponenten, auf die man bei einer KI-gestützten Lösung achten sollte, näher beleuchtet.

Was ist erklärbare KI?

Erklärbare KI ist ein Aspekt der künstlichen Intelligenz, der darauf abzielt, KI für Menschen interpretierbar zu machen, was bedeutet, daß sich die Gründe für eine Entscheidung explizit beschreiben lassen und den Teams verständlich vermittelt werden können. In einem idealen Szenario kann ein robustes KI-Modell komplexe Aufgaben ausführen, während die Benutzer den Entscheidungsprozeß beobachten und etwaige Fehler oder Bedenken überprüfen.

Die Bedeutung der Nachvollziehbarkeit von KI-Ergebnissen nimmt zu – unabhängig von der Anwendung und dem Sektor, in dem ein Unternehmen tätig ist. So müssen beispielsweise Anwendungen im Finanz- und Gesundheitswesen möglicherweise gesetzliche Anforderungen an die Transparenz von KI-Tools erfüllen. Bei der Entwicklung autonomer Fahrzeuge stehen Sicherheitsaspekte im Vordergrund, und die Verständlichkeit von Modellen ist entscheidend für die Verbesserung und Aufrechterhaltung der Funktionalität solcher Technologien. Daher ist erklärbare KI oft mehr als nur eine Frage der Bequemlichkeit – sie ist ein entscheidender Bestandteil von Geschäftsabläufen und Branchenstandards.

Je mehr KI-gestützte Technologien entwickelt und eingeführt werden, desto mehr staatliche und branchenspezifische Vorschriften werden erlassen. In der EU beispielsweise schreibt der EU AI Act Transparenz für KI-Algorithmen vor, auch wenn der Geltungsbereich derzeit noch begrenzt ist. Da es sich bei KI um ein so leistungsfähiges Instrument handelt, wird erwartet, daß es immer gefragter und ausgefeilter wird, was zu weiteren Vorschriften und Erklärungsanforderungen führen wird.

Es gibt auch Bedenken hinsichtlich der Voreingenommenheit und Zuverlässigkeit von KI-Modellen. In den Medien waren in letzter Zeit vor allem Halluzinationen durch generative KI, bei denen Large Language Models Fehlinformationen fabrizieren, ein Thema. KI-Modelle haben in der Vergangenheit immer wieder Verzerrungen aufgrund von Rasse, Geschlecht und anderen Kriterien produziert. Erklärbare KI-Tools und -Praktiken sind wichtig, um solche Verzerrungen zu verstehen und auszuräumen und damit die Genauigkeit der Ergebnisse und die betriebliche Effizienz zu verbessern. >>

Letztlich geht es bei erklärbarer KI um die Vereinfachung und Verbesserung der Fähigkeiten eines Unternehmens. Mehr Transparenz bedeutet ein besseres Verständnis der verwendeten Technologie, eine bessere Fehlerbehebung und mehr Möglichkeiten zur Feinabstimmung der Tools eines Unternehmens.

Heutige Herausforderungen erklärbarer KI

Erklärbare KI kann verschiedene Bedeutungen haben, sodaß bereits die Definition des Begriffs per se eine Herausforderung darstellt. Für die einen ist es eine Design-Methodik – ein Grundpfeiler in der KI-Modellentwicklung. Erklärbare KI ist auch die Bezeichnung für eine Reihe von Funktionen oder Fähigkeiten, die von einer KI-basierten Lösung erwartet werden, wie etwa Entscheidungsbäume und Dashboard-Komponenten. Der Begriff kann auch auf eine Art und Weise hinweisen, wie ein KI-Tool verwendet wird, das die Grundsätze der KI-Transparenz aufrechterhält. All dies sind legitime Beispiele für erklärbare KI, doch ihre wichtigste Aufgabe besteht darin, die Interpretierbarkeit von KI in einer Reihe von Anwendungen zu fördern.

Eine weitere Einschränkung der derzeitigen erklärbaren KI-Technologien besteht darin, daß ihre Wirksamkeit je nach Modell variiert. Einige Modelle, wie Deep Learning oder auf neuronalen Netzen basierende Modelle, sind dicht und komplex, was ihre Interpretation erschwert. Entscheidungsbäume und lineare Modelle hingegen lassen sich leichter verständlich und transparent machen, da der Entscheidungsprozeß über die Abbildung von Abhängigkeiten einfacher ist.

Alexander Zachow, Regional Vice President EMEA Central bei Dynatrace. Foto: Dynatrace

Erklärbare KI-Methoden befinden sich noch in einem frühen Stadium der Entwicklung. In fünf Jahren wird es neue Tools und Methoden für das Verständnis komplexer KI-Modelle geben, auch wenn diese Modelle weiterwachsen und sich weiterentwickeln. Schon jetzt ist es entscheidend, daß KI-Experten und Lösungsanbieter sich kontinuierlich um die Erklärbarkeit von KI-Anwendungen bemühen, um Unternehmen sichere, zuverlässige und leistungsstarke KI-Tools zur Verfügung zu stellen.

Die Schlüsselkomponenten erklärbarer KI

Erklärbare KI ist ein weites Feld. Daher ist es schwierig, eine definierte Liste von Merkmalen für alle erklärbaren KI-Lösungen zu erstellen. Einige Ansätze bevorzugen bestimmte Aspekte der Methodik gegenüber anderen oder gelten nur für bestimmte maschinelle Lernmodelle. Jeder umfassende Ansatz für erklärbare KI muß jedoch die folgenden Komponenten berücksichtigen:

- **Interpretierbarkeit:** Eine Basisfunktionalität für die Interpretation eines KI-Modells ist erforderlich. KI-Vorhersagen, -Entscheidungen und -Outputs müssen für einen Menschen verständlich und sollten zumindest durch den Entscheidungsprozeß des Modells nachvollziehbar sein. Die Tiefe der Interpretierbarkeit, die ein Unternehmen benötigt, hängt wahrscheinlich von dem Modell ab, das es nachvollziehbarer machen möchte, sowie von den jeweiligen Anwendungsfällen.

- **Kommunikationsmethoden:** Wie eine Lösung, die sich an erklärbarer KI orientiert, Informationen vermittelt, ist ebenfalls entscheidend. Starke Visualisierungstools sind notwendig, um die Vorteile jeder Methode für erklärbare KI gänzlich auszuschöpfen. Entscheidungsbäume und Dashboards sind zwei gängige Visualisierungsmethoden, die komplexe Daten in einem leicht lesbaren Format darstellen. Mit diesen Tools können Daten in umsetzbare Erkenntnisse umgewandelt werden. Auch hier hängt der Nutzen der verschiedenen Visualisierungstools vom jeweiligen KI-Modell ab.

- **Globale versus lokale Verständlichkeit:** Schließlich gibt es eine wichtige Unterscheidung zwischen globalen und lokalen Erklärungen. Globale Erklärungen sind Analysen und Informationen, die den Benutzern einen Einblick in die Funktionsweise des Modells als Ganzes geben. Zum Beispiel, wenn dargestellt wird, welche Datenbereiche während einer Reihe von Jobs verwendet werden, wo automatisierte Systeme agieren, was sie tun und mehr. Lokale Erklärungen sind Einblicke in die einzelnen Entscheidungen eines KI-Modells. Diese sind wichtig, wenn ein Unternehmen eine seltsame oder falsche Ausgabe nachvollziehen muß oder aus Gründen der Branchenregulierung transparente Informationen zur Hand haben möchte.

Mehr Transparenz durch Einsatz kausaler und vorhersagender KI

Erklärbare KI ist ein sich schnell verändernder Bereich in der Entwicklung von KI-Technologien, obwohl es aufgrund der zunehmenden Erklärbarkeit von Modellen bereits jetzt interessante neue Möglichkeiten für den Einsatz von KI gibt. So ermöglicht etwa die Kombination von generativer KI mit anderen, besser erklärbaren Formen wie der kausalen und vorhersagenden KI mehr Transparenz. Im Gegensatz zu generativer KI folgen sie keinem probabilistischen Ansatz, sondern sind auf Präzision ausgelegt und verwenden graphenbasierte und statistische Modelle, die bereichsspezifische, kontextbezogene Daten nutzen. Hierdurch eignen sie sich besser für spezielle Anwendungsfälle und sind resistenter gegen Halluzinationen und Verzerrungen. Durch ihre transparente Konzeption sind deren generierte Erkenntnisse für Nutzer nachvollziehbar, statt in einem Blackbox-Verfahren verborgen zu bleiben. <<

Noch Fragen?
https://www.dynatrace.com/de/