technische universität
dortmund

Medieninformation

Referat Hochschulkommunikation

Baroper Str. 285
D-44227 Dortmund
www.presse.tu-dortmund.de

2023-060 vom 30.08.2023

Studie der TU Dortmund stellt Menschen in den Mittelpunkt

Neue Erkenntnisse zur Erklärbarkeit von Künstlicher Intelligenz

Eine Untersuchung von Prof. Christian Janiesch von der Fakultät für Informatik der TU Dortmund stellt eine gängige Annahme der KI-Forschung in Frage: „Je leistungsfähiger die Methodik, umso schwerer ist sie erklärbar." Anhand von Tests mit medizinischer Bilddiagnostik konnte er gemeinsam mit Kollegen aus Würzburg und Magdeburg zeigen, dass Mediziner*innen einzelne KI-Analysen teils besser, teils schlechter verständlich fanden, als auf der Basis mathematischer und programmatischer Überlegungen bisher angenommen wurde. Die Ergebnisse sind im „International Journal of Information Management" veröffentlicht.

Hinter dem Oberbegriff Künstliche Intelligenz stecken verschiedene Verfahren, die heutzutage de facto alle auf Maschinellem Lernen fußen, d.h. die Programme haben eigenständig gelernt, wie sie entscheiden sollen. Die meisten bisherigen wissenschaftlichen Studien, die sich damit beschäftigen, wie erklärbar diese Systeme sind, verwenden dazu Annahmen, die auf mathematischen Prinzipien basieren. Dabei kamen sie zu dem Schluss, dass die Erklärbarkeit (explainability) sinkt, wenn die Leistung (performance) steigt. Sprich: Je komplexer die Anwendung, desto schwieriger ist sie zu erklären. Tiefe künstliche neuronale Netze sind demnach zwar überaus leistungsstark, aber für den Menschen nicht nachvollziehbar, während Entscheidungsbäume üblicherweise leistungsschwächer, aber gut erklärbar sind.

Prof. Janiesch und sein Team gingen die Frage nach der Erklärbarkeit aus einer anderen Perspektive an: Anstatt auf theoretische Modelle zu setzen, analysierten sie die Einschätzung von Fachleuten, die tagtäglich mit KI arbeiten. „Technologische Lösungen können nur dann dauerhaft erfolgreich sein, wenn sie von Menschen angewendet werden, die das Problem auch verstehen", erläutert der Wirtschaftsinformatiker. Dieser sozio-technische Ansatz ermöglichte es, die Erklärbarkeit aus einer realen Anwendungs- und Praxisperspektive zu betrachten. Dafür arbeiteten die Forscher mit Mediziner*innen zusammen, die einschätzten, wie nachvollziehbar verschiedene KI-Verfahren Symptome von Herzkrankheiten oder Hirnscans verarbeiteten.

Während die bisherigen Studien eine lineare oder kurvenförmige Beziehung zwischen Leistung und Erklärbarkeit vermuteten, zeigten die Untersuchungen von Prof. Janiesch und seinem Team ein gruppenartiges Muster in der Erklärbarkeit von KI-Systemen. Die Annahmen zur Leistungsfähigkeit der Systeme wurde grundsätzlich bestätigt, bei der Erklärbarkeit gab es jedoch abweichende Ergebnisse. Einige Modelle, die bisher als besser erklärbar galten, konnten die Fachleute eher schlecht nachvollziehen, oder umgekehrt.

Kontakt:
Nele Nafé
Telefon: (0231) 755-5524
Fax: (0231) 755-4664
nele.nafe@tu-dortmund.de

Während die tiefen künstlichen neuronalen Netze gleich bewertet wurden, gab es insbesondere bei Entscheidungsbäumen Unterschiede. Die Anwender*innen stuften sie als deutlich besser erklärbar ein, als die bisherigen Überlegungen dies suggerierten. Prof. Janiesch fügt an: „Unsere Forschung zeigt, dass die Erklärbarkeit von KI nicht nur auf mathematischen Analysen basieren sollte, sondern auf der Perspektive derjenigen, die mit dieser Technologie in der Praxis arbeiten. Denn wenn KI eingesetzt wird, müssen die Anwender*innen zunächst Vertrauen aufbauen und das geht nur, wenn nachvollziehbar gearbeitet wird."

Die Ergebnisse verdeutlichen die Notwendigkeit einer stärkeren Einbindung von Fachanwender*innen in den Prozess der KI-Entwicklung und -Einführung, um sicherzustellen, dass die Technologie nicht nur leistungsstark, sondern auch erklärbar und praktisch nutzbar ist. Dies ist besonders relevant, wenn es um kritische Entscheidungen auf Basis von KI-Vorhersagen geht, wie etwa in der Medizin.

Zur Publikation

Lukas-Valentin Herm, Kai Heinrich, Jonas Wanner, Christian Janiesch: Stop ordering machine learning algorithms by their explainability! A user-centered investigation of performance and explainability. *International Journal of Information Management*, Volume 69, 2023
(https://doi.org/10.1016/j.ijinfomgt.2022.102538)

Bildhinweis:

Prof. Christian Janiesch hat die Professur für Enterprise Computing an der Fakultät für Informatik der TU Dortmund inne. Foto: Anastasia Aulbach

Abbildung: Links die bisherige Annahme über den Zusammenhang zwischen Leistung und Erklärbarkeit, rechts die Ergebnisse der neuen Studie. Fünf gängige Klassen von KI-Algorithmen werden nach ihrer gemessenen Leistung (y-Achse) und wahrgenommenen Erklärbarkeit (x-Achse) geordnet. Foto: TU Dortmund

Ansprechperson für Rückfragen:

Prof. Christian Janiesch
Lehrstuhl für Enterprise Computing
Fakultät für Informatik
Telefon: (0231) 755-6634
E-Mail: christian.janiesch@tu-dortmund.de