MedWiss.Online

Wie können Naturwissenschaften von KI profitieren?

Publikation der Uni Bonn warnt vor Missverständnissen im Umgang mit Vorhersage-Algorithmen

Forschende aus Chemie, Biologie oder Medizin setzen zunehmend auf KI-Modelle, um neue Hypothesen zu entwickeln. Doch häufig ist unklar, auf welcher Basis die Algorithmen zu ihren Ergebnissen kommen und wie sehr diese verallgemeinerbar sind. Eine aktuelle Publikation der Universität Bonn warnt nun vor Missverständnissen im Umgang mit Künstlicher Intelligenz. Zugleich zeigt sie auf, unter welchen Bedingungen Forschende den Modellen am ehesten Vertrauen schenken können. Die Studie ist nun in der Zeitschrift „Cell Reports Physical Science“ erschienen.

Lernfähige Computeralgorithmen sind erstaunlich leistungsfähig. Doch sie haben einen Nachteil: Von außen betrachtet ist oft nicht ersichtlich, auf welcher Basis die Modelle ihre Schlüsse ziehen.

Mal angenommen, man füttert eine Künstliche Intelligenz mit Fotos von mehreren Tausend Autos. Wenn man ihr nun ein neues Bild vorlegt, kann sie in der Regel treffsicher erkennen, ob darauf ebenfalls ein Pkw zu sehen ist. Doch woran liegt das? Hat sie wirklich gelernt, dass ein Auto vier Räder, eine Windschutzscheibe und einen Auspuff hat? Oder orientiert sie sich an Kriterien, die eigentlich völlig irrelevant sind – etwa der Antenne auf dem Dach? Falls dem so wäre, könnte es sein, dass sie auch ein Radio als Auto klassifiziert.

KI-Modelle sind eine Black Box

„KI-Modelle sind eine Black Box“, betont Prof. Dr. Jürgen Bajorath. „Daher sollte man ihren Schlussfolgerungen nicht blind trauen.“ Der Chemieinformatiker leitet am Lamarr-Institut für maschinelles Lernen und künstliche Intelligenz den Bereich KI in den Lebenswissenschaften. Zudem verantwortet er das Life Science Informatics Programm am Bonn-Aachen International Center for Information Technology (b-it) der Universität Bonn. In der aktuellen Publikation ist er der Frage nachgegangen, wann man sich auf die Algorithmen am ehesten verlassen kann. Und umgekehrt: wann nicht.

Eine wichtige Rolle spielt dabei das Konzept der „Erklärbarkeit“. Darunter versteht man bildlich gesprochen die Bemühungen der KI-Forschung, ein Guckloch in die Black Box zu bohren. Der Algorithmus soll preisgeben, an welchen Kriterien er sich orientiert – an den vier Rädern oder an der Antenne. „Dieser Ansatz ist momentan ein zentrales Thema in der KI-Forschung“, sagt Bajorath. „Es gibt sogar KIs, die nur dafür entwickelt wurden, die Ergebnisse anderer KIs nachvollziehbarer zu machen.“

Die Erklärbarkeit ist aber nur ein Punkt – ebenso wichtig ist die Frage, welche Schlüsse sich aus dem von der KI gewählten Entscheidungskriterium ableiten lassen. Wenn der Algorithmus angibt, sich an der Antenne orientiert zu haben, weiß man als Mensch sofort, dass dieses Merkmal sich zur Identifikation von Autos schlecht eignet. Lernfähige Modelle werden aber meist genutzt, um in großen Datensätze Zusammenhänge zu erkennen, die uns Menschen gar nicht auffallen würden. Uns geht es dann wie einem Außerirdischen, der nicht weiß, was ein Auto ausmacht: Der könnte gar nicht sagen, ob die Antenne ein gutes Kriterium ist oder nicht.

Chemische Sprachmodelle schlagen neue Verbindungen vor

„Das ist eine zweite Frage, die wir uns beim Einsatz von KI-Verfahren in der Wissenschaft immer stellen müssen“, betont Bajorath, der auch Mitglied im Transdisziplinären Forschungsbereich (TRA) „Modelling“ ist: „Wie interpretierbar sind die Ergebnisse überhaupt?“ In der Chemie und Wirkstoffforschung sorgen momentan chemische Sprachmodelle für Furore. Man kann sie zum Beispiel mit vielen Moleküle füttern, die eine bestimmte biologische Aktivität haben. Im Idealfall schlägt das Modell auf dieser Basis dann ein neues Molekül vor, das diese Aktivität ebenfalls besitzt. Man spricht daher auch von einem „generativen“ oder „prädiktiven“ Algorithmus. Das Modell kann in der Regel allerdings nicht erklären, warum es zu dieser Lösung kommt. Dazu müssen oft nachträglich Methoden erklärbarer KI angewendet werden.

Bajorath warnt aber davor, diese Erklärungen – also die Merkmale, die die KI für wichtig hält – als kausal für die gewünschte Aktivität zu interpretieren. „KI-Modelle verstehen nichts von Chemie“, sagt er. „Oft achten sie auf Dinge, die chemisch oder biologisch irrelevant sind.“ Dennoch können sie mit ihrer Einschätzung sogar richtig liegen – vielleicht hat das vorgeschlagene Molekül also die gewünschten Fähigkeiten. Die Gründe dafür können aber ganz andere sein, als wir aufgrund chemischer Kenntnisse oder Intuition erwarten würden. Um das zu prüfen, sind Experimente notwendig: Die Forschenden müssen das Molekül synthetisieren und testen, ebenso wie andere Moleküle mit dem Strukturmotiv, das die KI für wichtig erachtet.

Plausibilitätsprüfungen sind wichtig

Derartige Tests sind zeitaufwändig und teuer. Bajorath warnt daher vor Überinterpretationen der KI-Ergebnisse auf der Suche nach wissenschaftlich plausiblen kausalen Zusammenhängen. An erster Stelle müsse eine Plausibilitätsprüfung stehen: Kann das von erklärbarer KI vorgeschlagene Merkmal tatsächlich für die gewünschte chemische oder biologische Eigenschaft verantwortlich sein? Lohnt es sich, den Vorschlag der KI weiterzuverfolgen? Oder handelt es sich um ein Artefakt, eine zufällig gefundene Korrelation wie die Autoantenne, die für die eigentliche Funktion gar nicht relevant ist?

Grundsätzlich habe der Einsatz lernfähiger Algorithmen das Potenzial, die Forschung in vielen Bereichen der Naturwissenschaften deutlich voranzubringen, betont der Wissenschaftler. Dazu müsse man aber die Stärken dieser Ansätze kennen – und besonders auch ihre Schwächen.

Publikation: Jürgen Bajorath: From Scientific Theory to Duality of Predictive Artificial Intelligence Models; Cell Reports Physical Science; DOI: 10.1016/j.xcrp.2025.102516, Internet: https://www.sciencedirect.com/science/article/pii/S2666386425001158