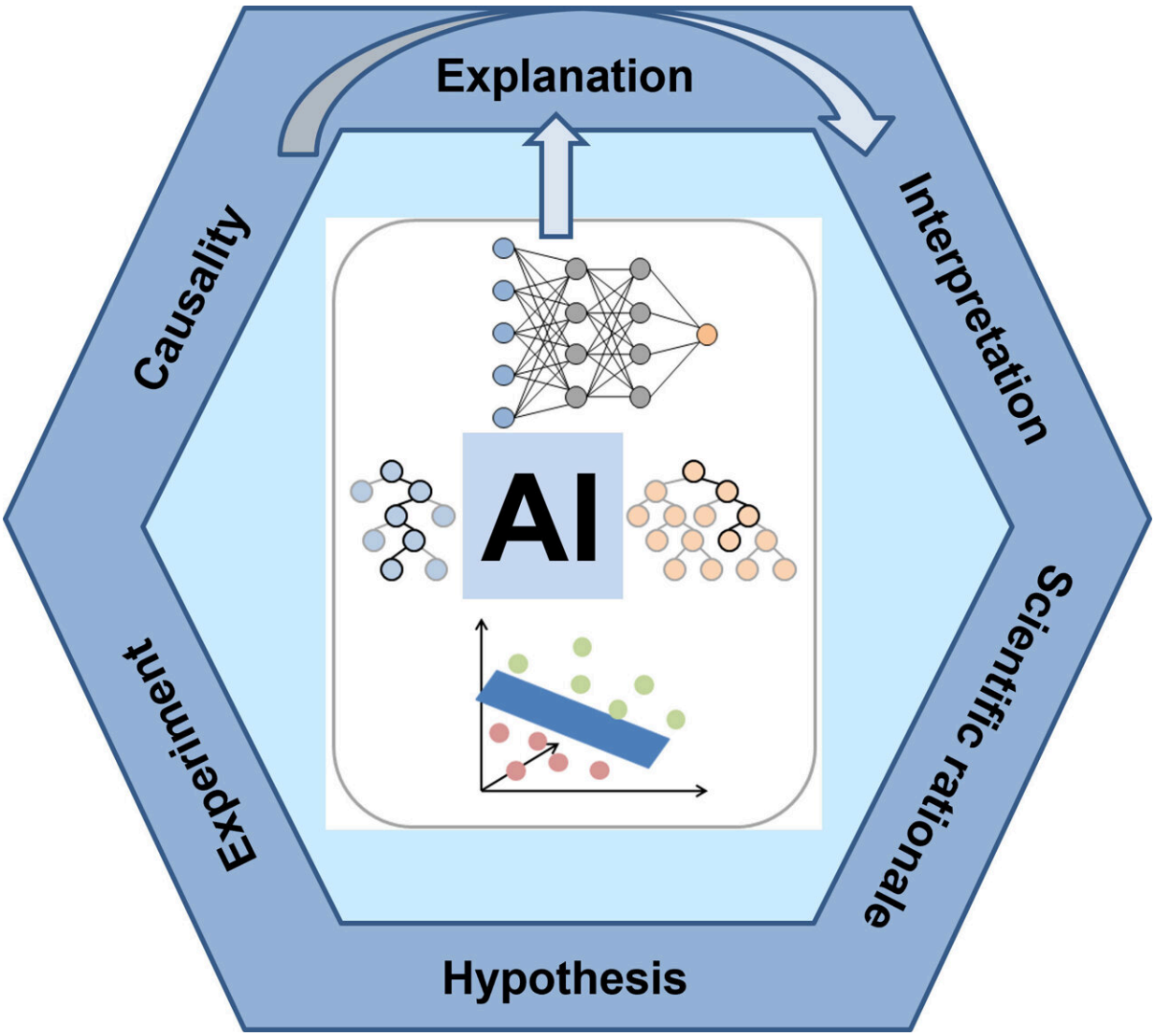
Expert warns of misinterpretations in AI-generated research hypotheses

April 4 2025, by Johannes Seiler



AI models in the natural sciences: From explaining predictions to capturing causal relationships. Credit: Jürgen Bajorath/University of Bonn

Researchers from chemistry, biology, and medicine are increasingly turning to AI models to develop new hypotheses. However, it is often unclear on which basis the algorithms come to their conclusions and to what extent they can be generalized.

A publication by the University of Bonn now warns of misunderstandings in handling artificial intelligence. At the same time, it highlights the conditions under which researchers can most likely have confidence in the models. The study has now been published in the journal *Cell Reports Physical Science*.

Adaptive machine learning algorithms are incredibly powerful. Nevertheless, they have a disadvantage: How machine learning models arrive at their predictions is often not apparent from the outside.

Suppose you feed artificial intelligence with photos of several thousand cars. If you now present it with a new image, it can usually identify reliably whether the picture also shows a car or not. But why is that? Has it really learned that a car has four wheels, a windshield, and an exhaust? Or is its decision based on criteria that are actually irrelevant—such as the antenna on the roof? If this were the case, it could also classify a radio as a car.

AI models are black boxes

"AI models are black boxes," highlights Prof. Dr. Jürgen Bajorath. "As a result, one should not blindly trust their results and draw conclusions from them." The computational chemistry expert heads the AI in Life Sciences department at the Lamarr Institute for Machine Learning and Artificial Intelligence. He is also in charge of the Life Science Informatics program at the Bonn-Aachen International Center for

Information Technology (b-it) at the University of Bonn.

In the current publication, he investigated the question of when one can most likely rely on the algorithms. And vice versa: When not.

The concept of "explainability" plays an important role in this context. Metaphorically speaking, this refers to efforts within AI research to drill a peephole into the black box. The algorithm should reveal the criteria that it uses as a basis—the four wheels or the antenna. "Opening the black box currently is a central topic in AI research," says Bajorath. " Some AI models are exclusively developed to make the results of others more comprehensible."

Explainability, however, is only one aspect—the question of which conclusions might be drawn from the decision-making criteria chosen by a model is equally important. If the algorithm indicates that it has based its decision on the antenna, a human being knows immediately that this feature is poorly suited for identifying cars.

Despite this, adaptive models are generally used to identify correlations in large data sets that humans might not even notice. We are then like aliens who do not know what makes a car: An alien would be unable to say whether or not an antenna is a good criterion.

Chemical language models suggest new compounds

"There is another question that we always have to ask ourselves when using AI procedures in science," stresses Bajorath, who is also a member of the Transdisciplinary Research Area (TRA) "Modeling": "How interpretable are the results?"

Chemical language models currently are a hot topic in chemistry and pharmaceutical research. It is possible, for instance, to feed them with

many molecules that have a certain biological activity. Based on these input data, the model then learns and ideally suggests a new molecule that also has this activity but a new structure. This is also referred to as generative modeling. However, the model can usually not explain why it comes to this solution. It is often necessary to subsequently apply explainable AI methods.

Nonetheless, Bajorath warns against over-interpreting these explanations, that is, anticipating that features the AI considers important indeed cause the desired activity. "Current AI models understand essentially nothing about chemistry," he says. "They are purely statistical and correlative in nature and pay attention to any distinguishing features, regardless of whether these features might be chemically or biologically relevant or not."

In spite of this, they may even be right in their assessment—so perhaps the suggested molecule has the desired capabilities. The reasons for this, however, can be completely different from what we would expect based on chemical knowledge or intuition. For evaluating potential causality between features driving predictions and outcomes of corresponding natural processes, experiments are typically required: The researchers must synthesize and test the molecule, as well as other molecules with the structural motif that the AI considers important.

Plausibility checks are important

Such tests are time-consuming and expensive. Bajorath thus warns against over-interpreting the AI results in the search for scientifically plausible causal relationships. In his view, a plausibility check based on a sound scientific rationale is of critical importance: Can the feature suggested by explainable AI actually be responsible for the desired chemical or biological property? Is it worth pursuing the AI's suggestion? Or is it a likely artifact, a randomly identified correlation

such as the car antenna, which is not relevant at all for the actual function?

The scientist emphasizes that the use of adaptive algorithms fundamentally has the potential to substantially advance research in many areas of science. Nevertheless, one must be aware of the strengths of these approaches—and particularly of their weaknesses.

More information: Jürgen Bajorath, From Scientific Theory to Duality of Predictive Artificial Intelligence Models, *Cell Reports Physical Science* (2025). DOI: 10.1016/j.xcrp.2025.102516. www.cell.com/cell-reports-phys … 2666-3864(25)00115-8

Provided by University of Bonn