

Should AI be Interpretable to Humans?

Matthew Schwartz, with Anna Golubeva

IAIFI Journal Club

November 29, 2022

On scientific understanding with artificial intelligence

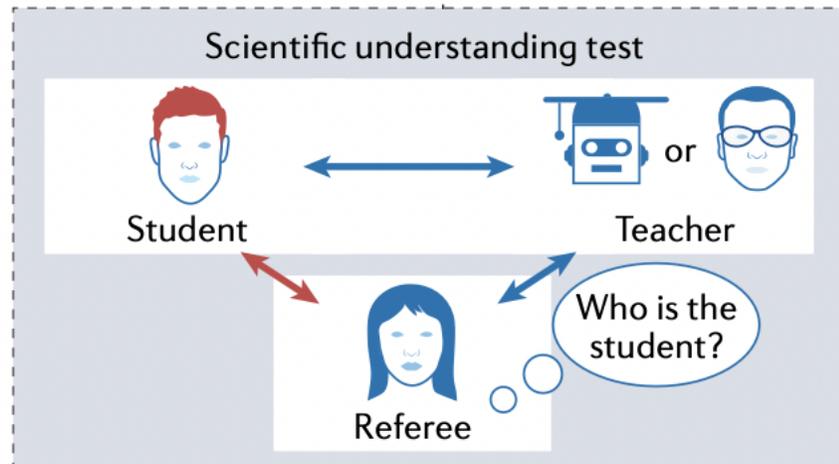
Mario Krenn, Robert Pollice, Si Yue Guo, Matteo Aldeghi, Alba Cervera-Lierta, Pascal Friederich, Gabriel dos Passos Gomes, Florian Häse, Adrian Jinich, AkshatKumar Nigam, Zhenpeng Yao  and Alán Aspuru-Guzik 

Abstract | An oracle that correctly predicts the outcome of every particle physics experiment, the products of every possible chemical reaction or the function of every protein would revolutionize science and technology. However, scientists would not be entirely satisfied because they would want to comprehend how the oracle made these predictions. This is scientific understanding, one of the main aims of science. With the increase in the available computational power and advances in artificial intelligence, a natural question arises: how can advanced computational systems, and specifically artificial intelligence, contribute to new scientific understanding or gain it autonomously? Trying to answer this question, we adopted a definition of ‘scientific understanding’ from the philosophy of science that enabled us to overview the scattered literature on the topic and, combined with dozens of anecdotes from scientists, map out three dimensions of computer-assisted scientific understanding. For each dimension, we review the existing state of the art and discuss future developments. We hope that this Perspective will inspire and focus research directions in this multidisciplinary emerging field.

Dieks and de Regt (2005)

- a scientific theory T is intelligible for scientists if they can recognise qualitatively characteristic consequences of T without performing exact calculations.

The scientific understanding test. A human (the student) interacts with a teacher, either a human or an artificial scientist.

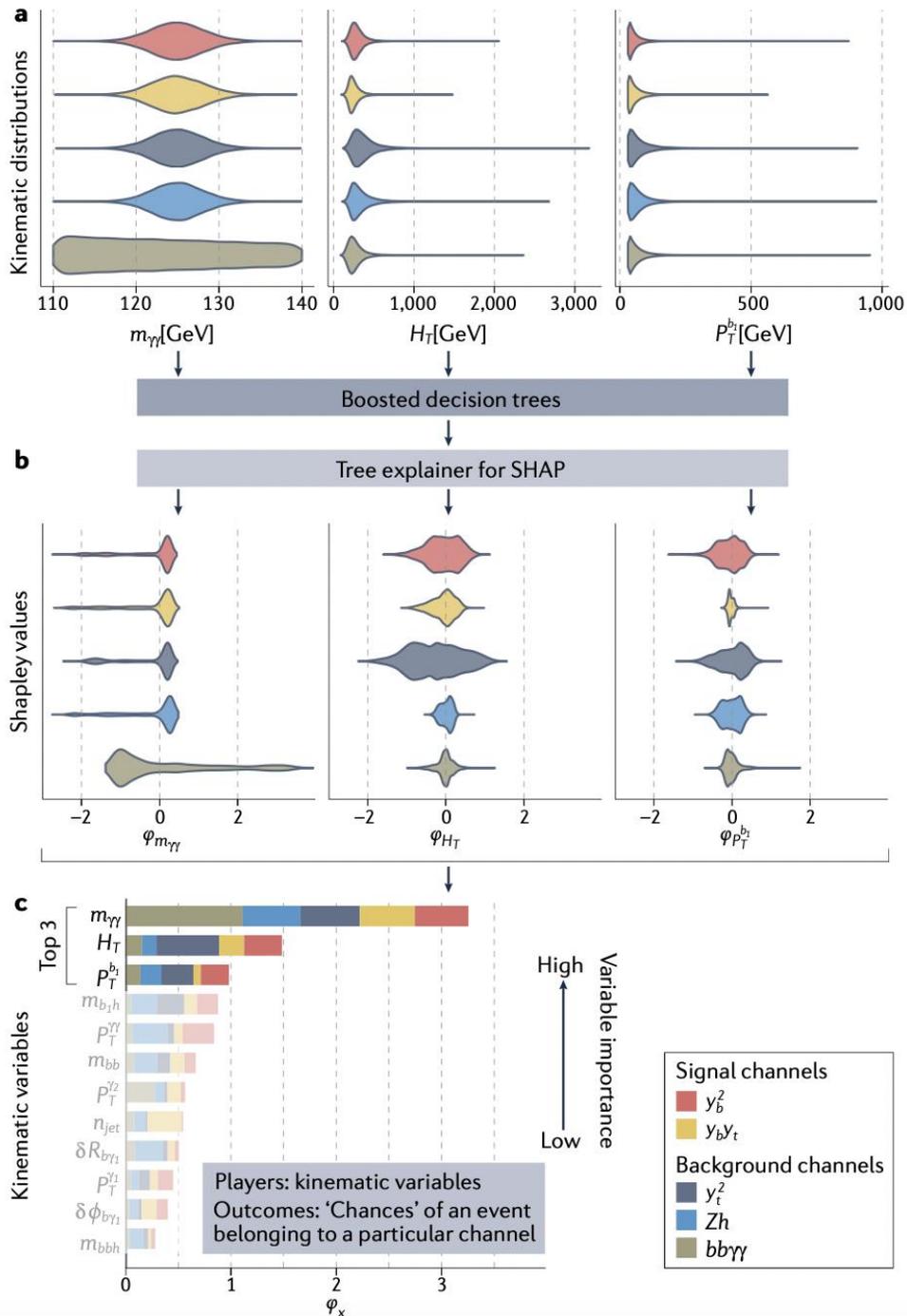


The teacher's goal is to explain a scientific theory and its qualitative, characteristic consequences to the student. Another human (the referee) tests both the student and the teacher independently. **If the referee cannot distinguish** between the qualities of their non-trivial explanations in various contexts, we argue that **the teacher has scientific understanding.**

Lessons on interpretable machine learning from particle physics

Christophe Grojean^{1,2}, Ayan Paul^{1,2}, Zhuoni Qian³ and Inga Strümke⁴

Machine learning methods have proved powerful in particle physics, but without interpretability there is no guarantee the outcome of a learning algorithm is correct or robust. Christophe Grojean, Ayan Paul, Zhuoni Qian and Inga Strümke give an overview of how to introduce interpretability to methods commonly used in particle physics.



- Example: measure hbb Yukawa using boosted decision tree
 - Shapley values rank observables
- Reduces to comfortable categories
- Does not explain why the BDT does better than high-level observables

It is the improvement of the AI above the interpretable component which makes it so powerful

Ludwig Wittgenstein 'Philosophical Investigations' 1953

"If a lion could talk, we wouldn't understand him"



Understanding is inherently vague and subjective

"A new-born child has no teeth."—"A goose has no teeth."—"A rose has no teeth."—This last at any rate—one would like to say—**is obviously true!** It is even surer than that a goose has none.—**And yet it is none so clear.**

For where should a rose's teeth have been? The goose has none in its jaw. And neither, of course, has it any in its wings; but no one means that when he says it has no teeth.—Why, suppose one were to say: the cow chews its food and then dungs the rose with it, so the rose has teeth in the mouth of a beast. This would not be absurd, because **one has no notion in advance where to look for teeth** in a rose.

Thomas Nagel “What is it like to be a bat” 1974

“Bat sonar, though clearly a form of perception, is not similar in its operation to any sense that we possess, and there is no reason to suppose that it is subjectively like anything we can experience or imagine.”

- You can record all the sensory information coming into a bat
- You can reproduce the bat’s actions
- You can never really understand what that bat perceives

Consider trying to explain

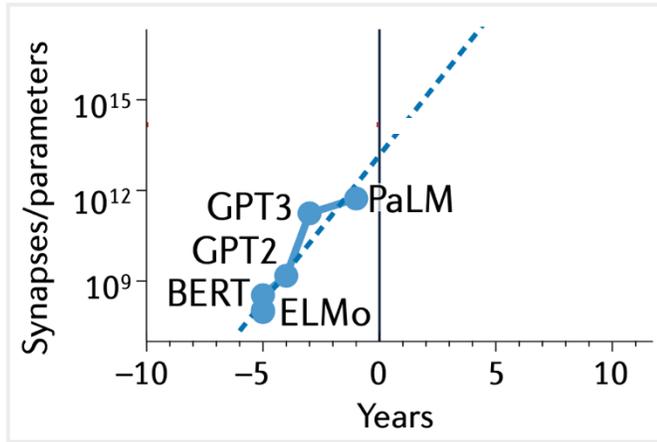
- what sight is to someone who is blind
- what it is like to eat food to someone who cannot
- quantum mechanics to a dog



There are limits to what can possibly be understood by a given organism

- Once you accept that there are limits, you must accept that AI can go beyond them
- Understanding AI is doomed to failure
 - we will soon be the dogs that cannot understand quantum mechanics

Future of AI



- Brains grow by a factor of 2 in one million years

- Both AI and biological intelligence grow exponentially
- Factor of 10^6 difference in exponent

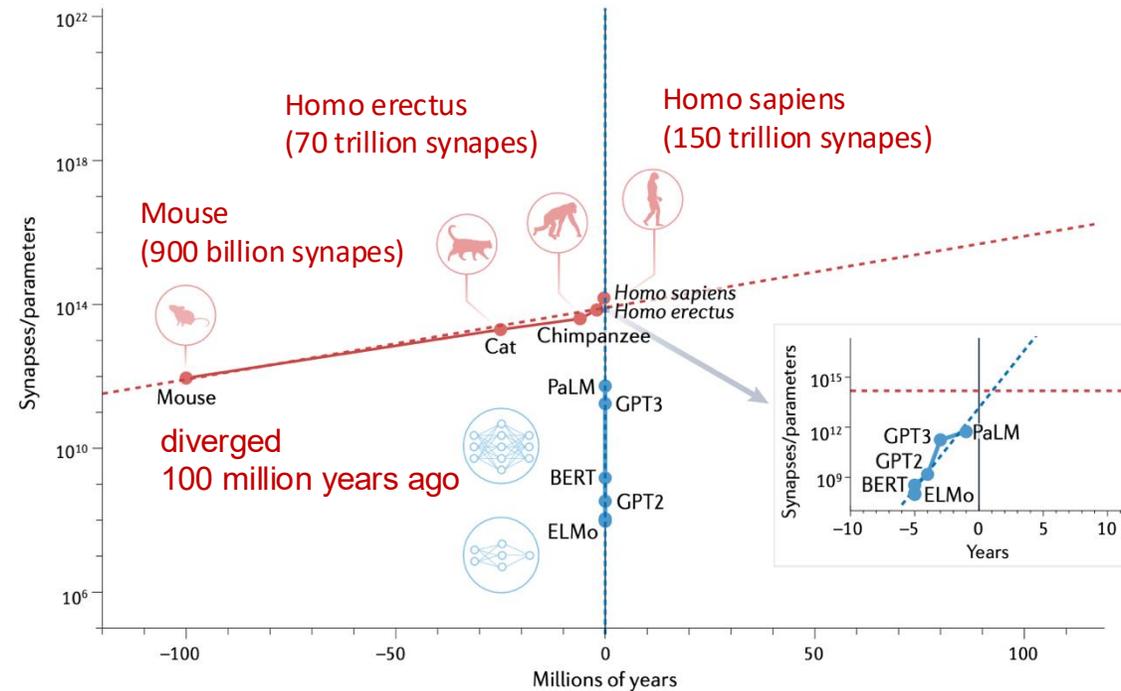
Should artificial intelligence be interpretable to humans?

[Matthew D. Schwartz](#) ✉

[Nature Reviews Physics](#) 4, 741–742 (2022) | [Cite this article](#)

- ELMo (94 million parameters, 2018)
- GPT2 (1.5 billion parameters, 2019)
- GPT3 (175 billion parameters, 2020)
- PALM (540 billion parameters, 2022)

AI grows by factor of 10/year



PaLM and Minerva

- PaLM (Pathways Language Model) is a language model by google [April 2022]



80 billion neurons

540 billion parameters

trained for 60 days on 6144 TPU (Tensor Processing Unit) chips

\$23 million to train

Solving Quantitative Reasoning Problems with Language Models

Aitor Lewkowycz*, Anders Andreassen†, David Dohan†, Ethan Dyer†, Henryk Michalewski†,
Vinay Ramasesh†, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo,
Yuhuai Wu, Behnam Neyshabur*, Guy Gur-Ari*, and Vedant Misra*

Google Research

arXiv:2206.14858
[June 2022]

- **Minerva**

- Built on PaLM
- Trained on 60 GB of arXiv papers and mathy webpages
 - (e.g. libretxts, chegg, stacoverflow)
- 1 month additional training on 1024 TPU chips

Minerva (June 2022)

<https://minerva-demo.github.io>

QUESTION Four politicians and three lawyers attend a party. Each politician shakes hands exactly once with everyone, and each lawyer shakes hands exactly once with each politician. How many handshakes take place?

Answer by Minerva (output is in LaTeX)

Let P be the set of politicians and L be the set of lawyers. There are $4 \cdot 3 = 12$ handshakes between the politicians and lawyers. There are also $\binom{4}{2} = 6$ handshakes among the politicians. Therefore, the total number of handshakes is $12 + 6 = \boxed{18}$.
Final Answer: The final answer is 18.

EVALUATION RESULT Correct

Minerva (June 2022)

Question: A liquid flows at a constant flow rate through a pipe with circular cross-sections of varying diameters. At one point in the pipe, the diameter is 2 cm and the flow speed is 18 m/s . What is the flow speed at another point in this pipe, where the diameter is 3 cm ? Which of the following is the right choice? Explain your answer. (A) 4 m/s (B) 6 m/s (C) 8 m/s (D) 12 m/s

Answer by Minerva (output is in LaTeX)

Let A_1 be the cross-sectional area of the pipe at the first point and A_2 be the cross-sectional area of the pipe at the second point. Let v_1 be the flow speed at the first point and v_2 be the flow speed at the second point. Since the flow rate is constant, the product of the cross-sectional area and the flow speed is constant. Hence,

$$A_1 v_1 = A_2 v_2.$$

The cross-sectional area of the pipe at the first point is $A_1 = \pi r_1^2 = \pi \left(\frac{2}{2}\right)^2 = \pi$, where r_1 is the radius of the pipe at the first point. Similarly, $A_2 = \pi r_2^2 = \pi \left(\frac{3}{2}\right)^2 = \frac{9\pi}{4}$. Therefore,

$$\pi \cdot 18 = \frac{9\pi}{4} v_2 \Rightarrow v_2 = \boxed{8\text{ m/s}}.$$

Final Answer: The final answer is (C).

EVALUATION RESULT Correct

Minerva (June 2022)

QUESTION Let \mathbf{D} be a matrix representing a dilation with scale factor $k > 0$, and let \mathbf{R} be a matrix representing a rotation about the origin by an angle of θ counter-clockwise. If

$$\mathbf{DR} = \begin{pmatrix} -7 & 1 \\ -1 & -7 \end{pmatrix},$$

then find k .

MODEL ANSWER We have

$$\mathbf{DR} = \begin{pmatrix} -7 & 1 \\ -1 & -7 \end{pmatrix}.$$

We also know that

$$\mathbf{D} = \begin{pmatrix} k & 0 \\ 0 & k \end{pmatrix}.$$

We have that $\mathbf{DR} = (\mathbf{DR})^T$, so we have

$$\begin{pmatrix} k & 0 \\ 0 & k \end{pmatrix} \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} = \begin{pmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{pmatrix} \begin{pmatrix} k & 0 \\ 0 & k \end{pmatrix} = \begin{pmatrix} k \cos \theta & k \sin \theta \\ -k \sin \theta & k \cos \theta \end{pmatrix}.$$

We have that $k \cos \theta = -7$ and $k \sin \theta = 1$. Thus,

$$k^2 \cos^2 \theta + k^2 \sin^2 \theta = 49 + 1 = 50.$$

This gives us that $k^2 = 50$, so $k = \sqrt{50} = \boxed{5\sqrt{2}}$. Final Answer: The final answer is $5\sqrt{2}$.

EVALUATION RESULT Correct

Current SOTA:

- AI's like Minerva can solve high-school/college level physics problems
- Learns by studying solved problems (just like a student)
- Chain-of-thought reasoning: shows it's work (just like a student)

Next Steps

- Learn like a graduate student
 - Few-shot learning (limited training data)
 - Data augmentation (create toy models)
 - Alpha Zero showed that AI can teach itself to be smarter than humans (for some tasks)
- Research like a pro
 - Identify approachable problems
 - Solve pieces at a time

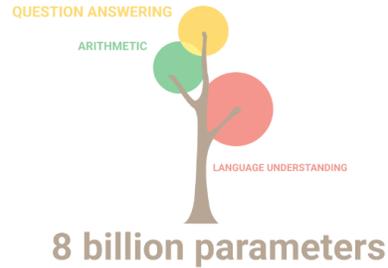
Future

- Automatically generate arXiv papers
- “try” to explain results to humans
- summarize results for humans
- humans accept that understanding is beyond them

What's next?



PaLM/Minerva (2022)
540 billion parameters

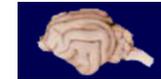


Human brain

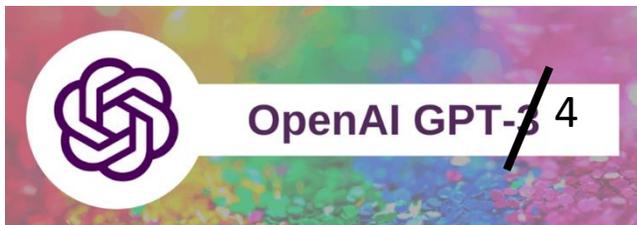


80 billion neurons
150 trillion synapses

Cat brain



0.760 billion neurons
10 trillion synapses



...expected soon

100 trillion parameters
What will it do?

2023 ... 2040... 2100? ...3000...?