On Interpretability in AI

April 18, 2025

Matthew Schwartz

Harvard University

Institute for Artificial Intelligence
and Fundamental Interactions

explain verb

ex·plain ik-ˈsplān

1 a : to make known

explain the secret of your success

b : to make plain or understandable

footnotes that *explain* the terms

interpret verb

in·ter·pret in-ˈtər-prət -pət

1 : to explain or tell the meaning of **:** present in understandable terms

interpret dreams

needed help *interpreting* the results

meaning 1 of 2 noun

mean·ing ˈmē-niŋ

1 a : the thing one intends to convey especially by language **: PURPORT**

Do not mistake my *meaning*.

b : the thing that is conveyed especially by language **: IMPORT**

Many words have more than one *meaning*.

understand verb

un·der·stand ˌən-dər-ˈstand

1 a : to grasp the meaning of

understand Russian

convey verb

con·vey kən-ˈvā

1 a : to bear from one place to another

especially **:** to move in a continuous stream or mass

b : to impart or communicate by statement, suggestion, gesture, or appearance

struggling to *convey* his feelings

Understanding is inherently vague and subjective

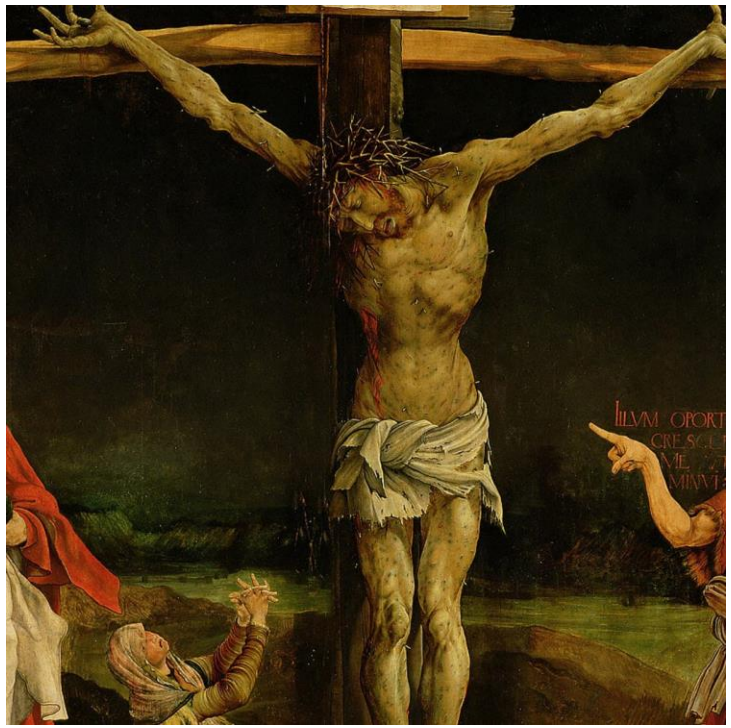
Ludwig Wittgenstein *Philosophical Investigation* 1953

"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**.

"A rose has no teeth" is unambiguous

the rose has teeth in the mouth of a beast

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.



Happy **good** friday!

Was the interpreation of the crucifiction of christ?
What is a good?

- No! The savior of mankind was brutally tortured and murdered
- Yes! Christ died for *our* sins, leading to our salvation
- No! It lead to Christian conquest and colonial violence
- Yes! It led to an ethical framework that underpins modern moral thought
- No! It is leads to political manipulation and abuse in the name of Christ
- Yes! It can be used in a constructive discussion about interpreability in AI
- No! It can be misued for an antrhopcentric discussion of interpretability which is a distraction from the scientific goals of AI.

Understanding is inherently vague and subjective

"The Supreme Court said the district court order was unlawful and its main components were reversed 9-0 unanimously," Miller told Trump.

White House Defiant Against Supreme Court, Says Ruling Was 'In Our Favor'

Published Apr 15, 2025 at 7:14 AM EDT | Updated Apr 15, 2025 at 7:24 AM EDT

Supreme Court Sides With Wrongly Deported Migrant

A trial judge had ordered the Trump administration to take steps to return the migrant, Kilmar Armando Abrego Garcia, from a notorious prison in El Salvador.

POLITICS / SUPREME COURT

Trump defied a court order. The Supreme Court just handed him a partial loss.

Even Trump's lawyers concede that deporting Kilmar Armando Abrego Garcia was illegal.

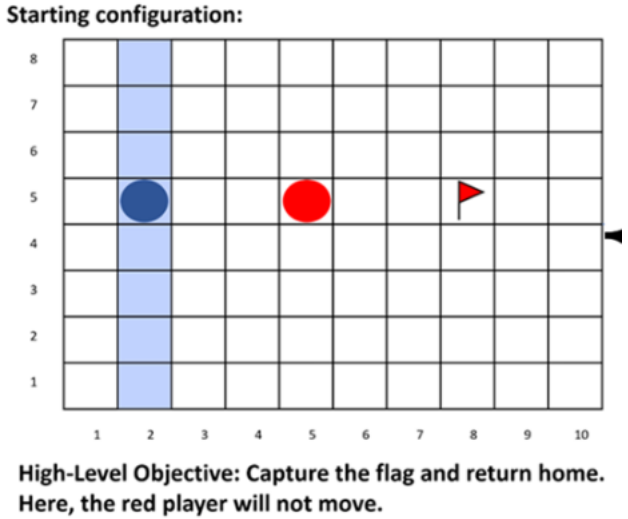
Humans can’t interpret things we think are intepretable

STL: Surprisingly Tricky Logic (for System Validation)

Ho Chit Siu, Kevin Leahy, and Makai Mann

MIT Lincoln Lab, Oct 2023

Learn some robot behavior (capture the flag)



believed to be interpetable

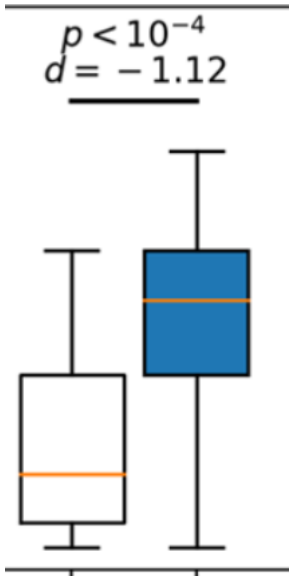
Signal Temporal Logic

$$F_{[15,20]}(x = 8 \wedge y = 5) \wedge F_{[30,32]}(x = 2) \wedge G_{[0,100]}(\neg(x \geq 4 \wedge x \leq 6 \wedge y \geq 4 \wedge y \leq 6))$$

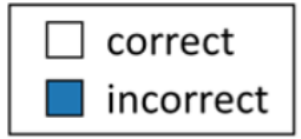
Natural language

definitely Interpretable

(Eventually between t = 15 and t = 20: $x = 8$ AND $y = 5$)
AND
(Eventually between t = 30 and t = 32: $x = 2$)

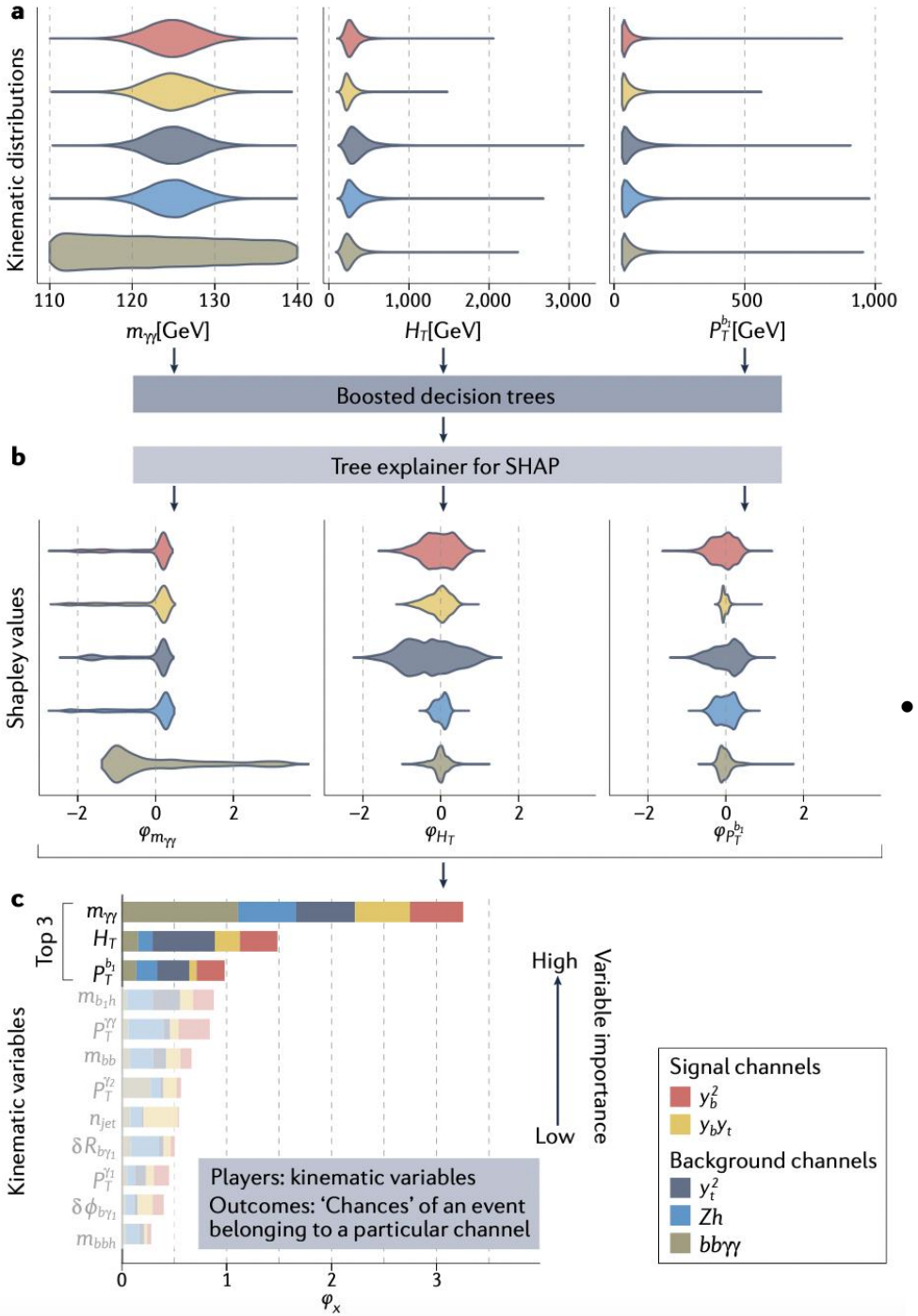


people failed miserably at interpreting either



Our data do not support the belief that formal specifications are inherently human-interpretable to a meaningful degree for system validation. We recommend ergonomic improvements to

Additionally, participants, particularly those familiar with formal methods, tended to be overconfident in their answers, and be similarly confident regardless of actual correctness.



Lessons on interpretable machine learning from particle physics

2022

Christophe Grojean [1,2], Ayan Paul [1,2] ✉, Zhuoni Qian [3] and Inga Strümke [4]

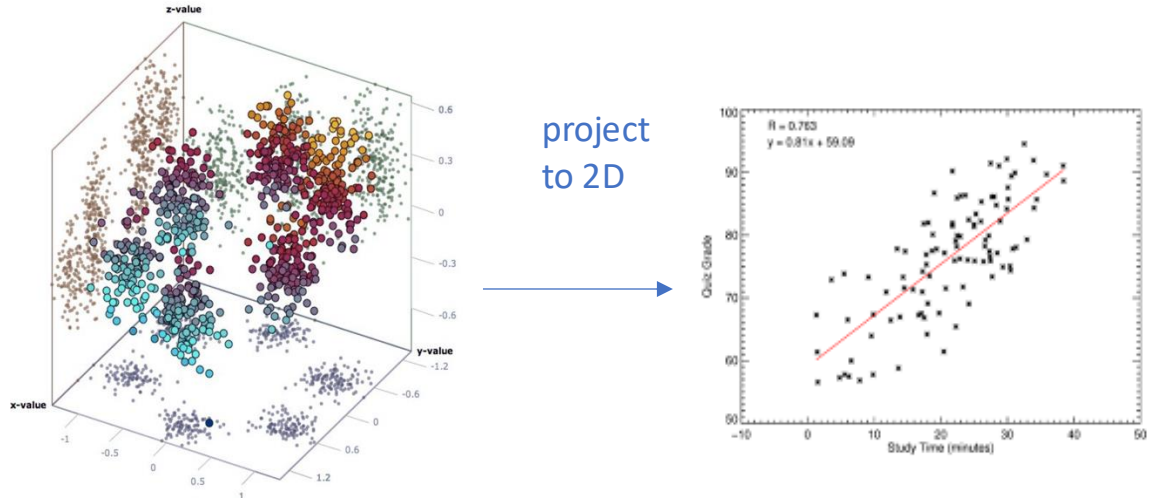
Machine learning methods have proved powerful in particle physics, but without interpretability there is no guarantee the outcome of a learning algorithm is correct or robust. Christophe Grojean, Ayan Paul, Zhuoni Qian and Inga Strümke give an overview of how to introduce interpretability to methods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

- Example: measure Higgs-bottom-quark coupling with BDT
  - Shapley values correlate with observables
  - Reduces to comfortable categories: $m_{\gamma\gamma}$
- 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

Interpretabillity is not a **scientific** problem
It is a **human** problem

Humans like to "**visualize**"



Why do we do this? Because we have eyes

- 2D is not special to a machine.
- Machines can "visualize" in d dimensions

Eyes have **nothing to do** with fundamental physics!

Humans can only hold 5-9 concepts in working memory at once

- We like simple-looking equations

$$i\partial_t\psi = H\psi \quad i\,\partial\!\!\!/\psi = m\psi \quad G_{\mu\nu} = \kappa T_{\mu\nu}$$

- **Computer** memory can **handle much more** than 5-9 concepts at once
- They can understand systems not governed by simple equations

Thomas Negel “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 chew
- 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
- Undestanding AI is doomed to failure
  - we will soon be the dogs that cannot understand quantum mechanics

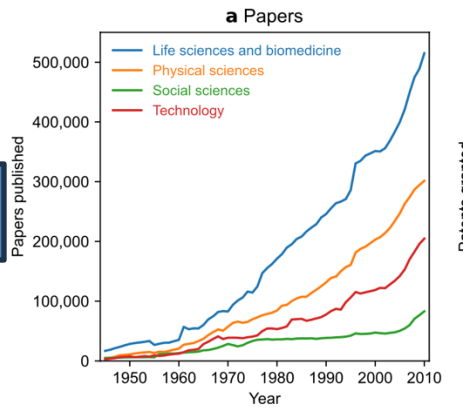
Theoretical particle physics may have stalled

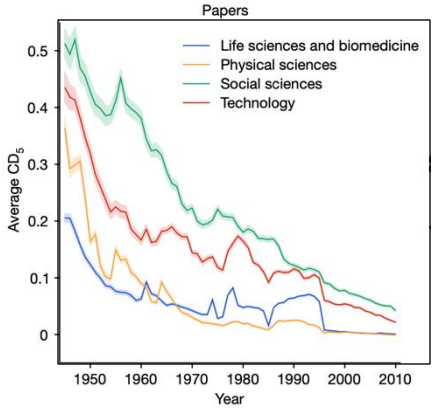
Article | Published: 04 January 2023

Papers and patents are becoming less disruptive over time

Michael Park, Erin Leahey & Russell J. Funk

more and more papers are written





the papers are less and less innovative

Maybe the problems are just too difficult (for us)

Could a cat ever learn to play chess?

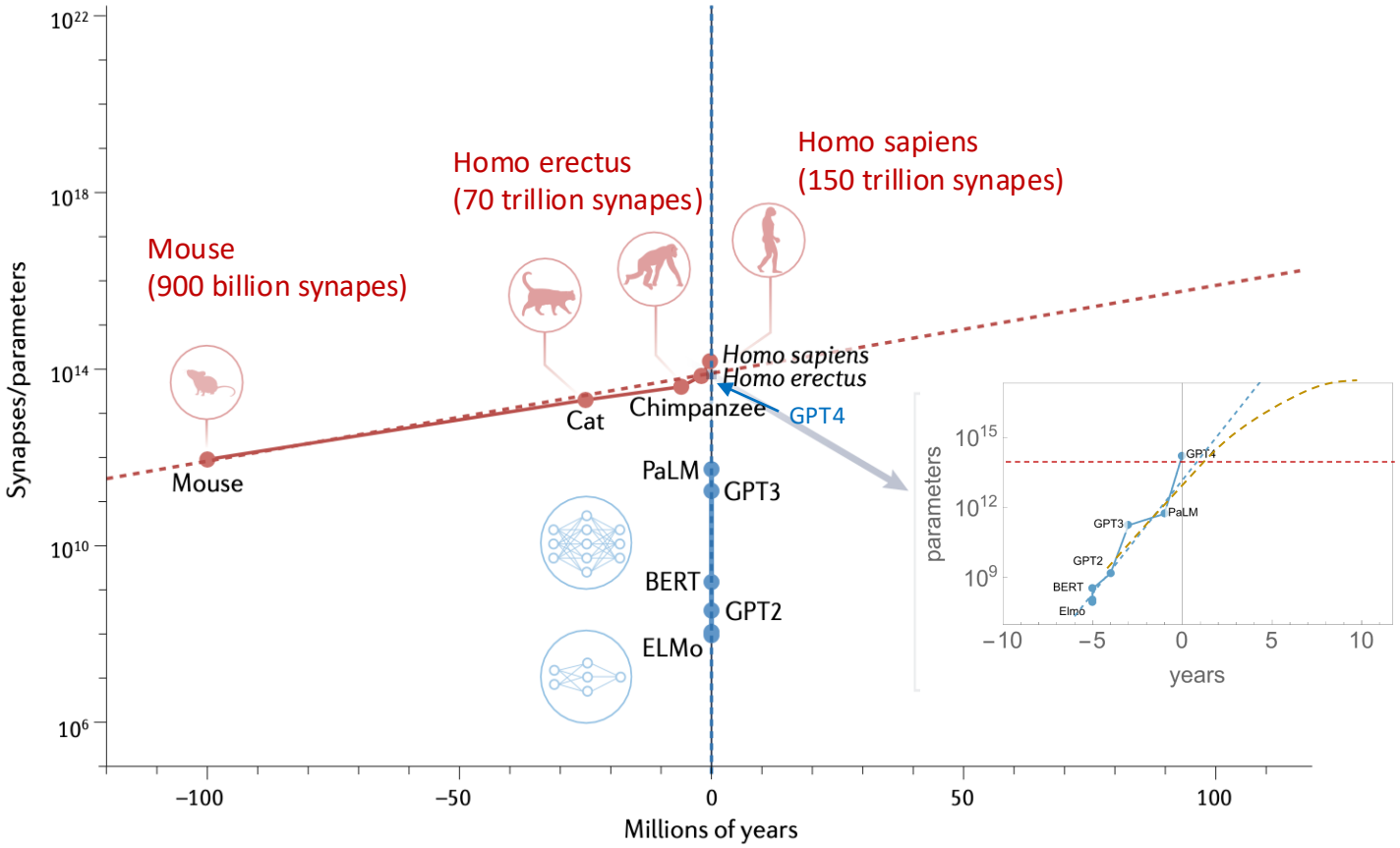
- Humans have limits too

Why should Homo sapiens be able to understand the theory of everything?

Machine vs. Biological intelligence

- Machine intelligence grows by a factor of 10 in 1 year
- Biological intelligence grows by a factor of 2 in 20 million years

MDS, "Should artificial intelligence be interpretable to humans?" Nature reviews physics (2022)



even with subexponential growth machines will soon have superhuman intelligence

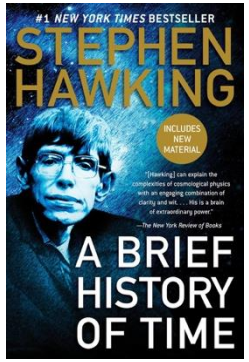
- Both AI and biological intelligence grow exponentially
- **Factor of 10^7 difference in *exponent***
- Intersection time, when machines and biology have comparable "intellegence" is **now**

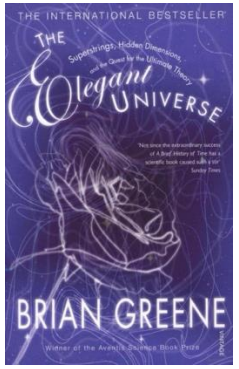
Superhuman intelligence

Suppose a machine understands the theory of everything but we don't

- e.g. can calculate electron mass from scratch
- e.g. can explain dark matter

Is this enough or do we need to understand it too?





- The authors of **Popular science book** understand the details; we just get the general idea

I don't understand the proof of Fermat's last theorem

- I'm glad that somebody does
- Does it matter that the person is human?

If a machine understands fundamental physics it can

1. Dumb it down so we can get the general idea
   - Provide sujective intepretation for humans

2. Find practical applications
   - True scientific goal

NATURE | VOL 405 | 1 JUNE 2000 | www.nature.com

Catching crumbs from the table

Ted Chiang

It has been 25 years since a report of original research was last submitted to our editors for publication, making this an appropriate time to revisit the question that was so widely debated then: what is the role of human scientists in an age when the frontiers of scientific inquiry have moved beyond the comprehensibility of humans?

No one denies the many benefits of metahuman science, but one of its costs to human researchers was the realization that they would probably never make an original contribution to science again. Some left the field altogether, but those who stayed shifted their attentions away from original research and toward hermeneutics: interpreting the scientific work of metahumans.

We need not be intimidated by the accomplishments of metahuman science. We should always remember that the technologies that made metahumans possible were originally invented by humans, and they were no smarter than we. ■

Conclusions

1. Interpretability is subjective
   - People are inconsistent about interpreting *everything*

2. Interpretabilty requires **over**simplifying
   - It is the non-interpretable part that makes ML so powerful

3. Interpretability is athropocentric
   - ML can do things we cannot
   - There are things humans cannot understand
   - ML can do things we cannot understand