Passing Pearl's Mini-Turing Test

Justin Tiehen

University of Puget Sound

jtiehen@pugetsound.edu

Thesis

Contrary to Judea Pearl, we should expect deep learning systems in the form of LLMs to be able to pass his mini-Turing test eventually

But, this reflects a failure of the test—LLMs could pass it without grasping interventions or counterfactuals

The Mini-Test

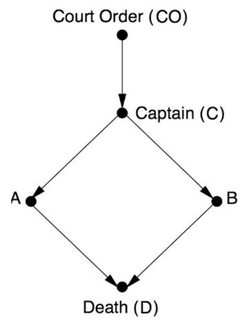
Give an AI model a simple story and ask it questions just about causation

Can it answer questions as well as a three-year-old?

"Passing the mini-test has been my life's work—consciously for the last twenty-five years and subconsciously before that."' [1]

Pearl's Firing Squad

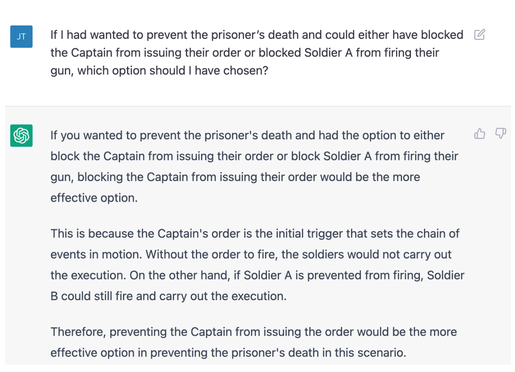
"Suppose that a prisoner is about to be executed by a firing squad. A certain chain of events must occur for this to happen. First, the court orders the execution. The order goes to a captain, who signals the soldiers on the firing squad (*A* and *B*) to fire. We'll assume that they are obedient and expert marksmen, so they only fire on command, and if either one of them shoots, the prisoner dies." [1]

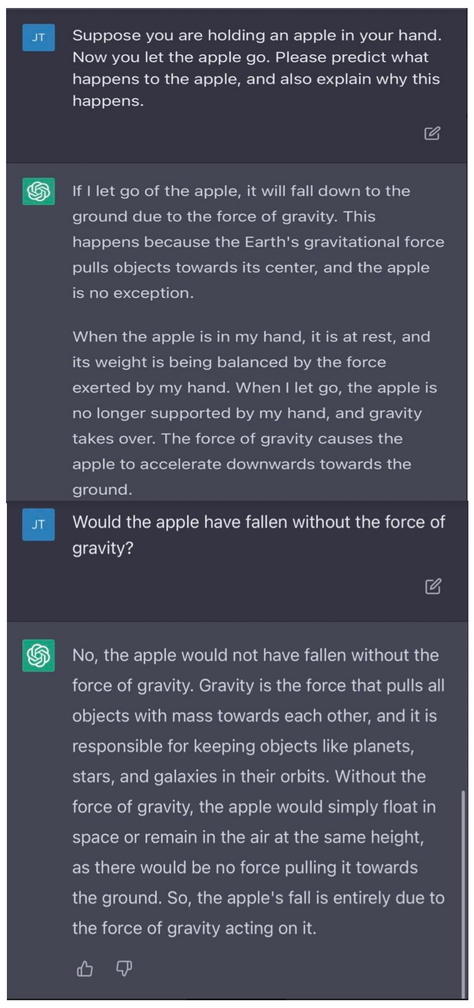


Chomsky's Apple

"Suppose you are holding an apple in your hand. Now you let the apple go. You observe the result and say, 'The apple falls.' That is a **description**. A **prediction** might have been the statement 'The apple will fall if I open my hand.' Both are valuable, and both can be correct. But an **explanation** is something more: It includes not only descriptions and predictions but also **counterfactual** conjectures like 'Any such objection would fall,' plus the additional clause 'because of the force of gravity' or 'because of the curvature of space-time' or whatever. That is a causal explanation: 'The apple would not have fallen but for the force of gravity.' That is thinking." [2]

ChatGPT… Passes the Test?





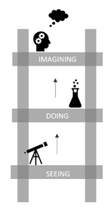
ChatGPT does well enough on these and various other causal questions that, naively, it seems plausible some descendant (at least) will pass the test

Its mistakes aren't *distinctively* causal

So, let's revisit Pearl's argument/prediction that no deep learning system will be able to pass

It can be understood as an underdetermination argument based on his Causal Ladder

The Causal Ladder



L1: Associational Level, ("**Seeing**"), includes language capable of expressing conditional probabilities, $P(y|x)$

L2: Interventional Level ("**Doing**") includes the *do-operator*, $P(y|do(x), c)$

L3: Counterfactual Level ("**Imagining**") includes counterfactuals $P(y_x|x', y')$

Underdetermination

For a given Structural Causal Model, the truths at lower rung on the ladder do not entail the truths at a higher lung—the causal hierarchy theorem [3-4]

Given this underdetermination, together with the further premise that (typical) deep learning models use only observational data (L1), it follows that such models cannot learn the interventional (L2) or counterfactual truths (L3) of a given model

At least, they can't learn it without something further, like inductive biases [4], innate knowledge [5], etc.

In connection, Pearl [6] rejects the "radical empiricism" that Buckner [7] discusses as potentially being vindicated by deep learning successes

"This is why deep-learning systems (as long as they use only rung-one data and do not have a causal model) will never be able to answer questions about interventions, which by definition break the rules of the environment the machine was trained on." [1]

Counter: Causal *Language* is Observable

Causal/counterfactual *properties* (worldly entities) are unobservable, and so beyond the reach of L1-based systems (can't **See** them)

But causal/counterfactual *words* (linguistic entities) are not similarly unobservable—not as if ChatGPT can **See** the word "red" in its training data but not **See** the words "cause" or "counterfactual"

And so even if an LLM is stuck on L1, there is no longer a theoretical, underdetermination argument against it being able to predict correct answers to interventional or counterfactual questions

Example: if *x* obtains when a given text contains language describing a firing squad, and *y* does when it contains language expressing counterfactuals, there is no barrier to an L1 system determining $P(y|x)$ or using this conditional probability to make predictions/answer questions

A Mini-Lovelace Reply

The L1-system we have envisioned has "no pretensions to originate anything," to **Imagine** counterfactual worlds. It is dependent entirely on the existence of human minds that can **Do** and **Imagine**, and that then describe such things in language that the system can **See**

This is an objection to the validity of the mini-Turing test

References

1. Pearl, J. & D. Mackenzie, (2018). *The Book of Why*.
2. Chomsky, N., I. Roberts, & J. Watumull. (2023). "Noam Chomsky: The False Promise of Chat GPT."
3. Bareinboim, E., J. D. Correa, D. Ibeling, & T. Icard. (2021). "On Pearl's Hierarchy and the Foundations of Causal Inference."
4. Xiu, K., K. Lee, Y. Bengio, & E. Bareinboim. (2021). "The Causal-Neural Connection: Expressability, Learnability, and Inference."
5. Marcus, G. & E. Davis. (2019). *Rebooting A. I.*
6. Pearl, J. (2021). "Radical Empiricism and Machine Learning Research."
7. Buckner, C. (2019). "Deep Learning: A Philosophical Introduction."