Why are the Social Sciences Hard? Two New Hypotheses.

Aydin Mohseni[1] Daniel Herrmann[2] Gabe Orona[3]

[1]Carnegie Mellon University
[2]University of Groningen
[3]University of Tübingen

The social sciences, it is claimed, are hard.

"Because they deal with systems that are highly complex, adaptive and not rigorously rule-bound, the social sciences are among the most difficult of disciplines, both methodologically and intellectually...."

—Via Dan Drezner, 2012, A Different Agenda, *Nature*, p. 271

"...the continuum of scientific problems should go from "hard" to "easy," and psychology is clearly a science that deals with hard problems. Perhaps the hardest."

—Sanjay Srivastava, 2009, lecture transcript from talk at *The Society for Personality and Social Psychology* conference

"Hard-to-do science is what the social scientists do and, in particular, it is what we educational researchers do. In my estimation, we have the hardest-to-do science of them all!"

—Berliner, 2002, "Educational Research: The Hardest Science of All," p. 18

"...In 10 minutes of superficial thought I easily came up with 20 features that make human psychology hard to scientize, I invite you to pick your own favorites."

—Meehl, 1978, "Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology," Journal of Consulting and Clinical Psychology, pp. 806.

In the literature, there are broadly three types of explanations for why the social sciences are hard.

The difficulty is due to:

1. Special subject
2. Poor methods
3. Perverse incentives

Existing explanations for SSH



Special subject:

- **Structural complexity of human behavior**
- Difficulty in representing or quantifying behavior
- Short-run adaptiveness of human behavior
- Medium-run change in human behavior
- Context sensitivity of human behavior
- Randomness/noise of human behavior

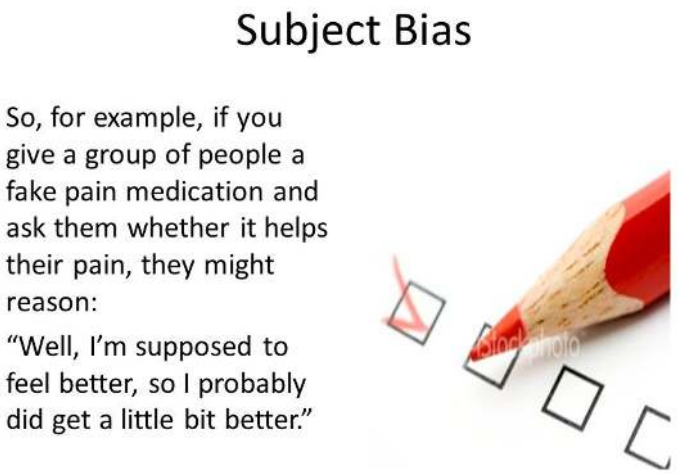
Existing explanations for SSH



Special subject:

- Structural complexity of human behavior
- **Difficulty in representing or quantifying behavior**
- Short-run adaptiveness of human behavior
- Medium-run change in human behavior
- Context sensitivity of human behavior
- Randomness/noise of human behavior

Existing explanations for SSH



Special subject:

- Structural complexity of human behavior
- Difficulty in representing or quantifying behavior
- **Short-run adaptiveness of human behavior.**
- Medium-run change in human behavior
- Context sensitivity of human behavior
- Randomness/noise of human behavior

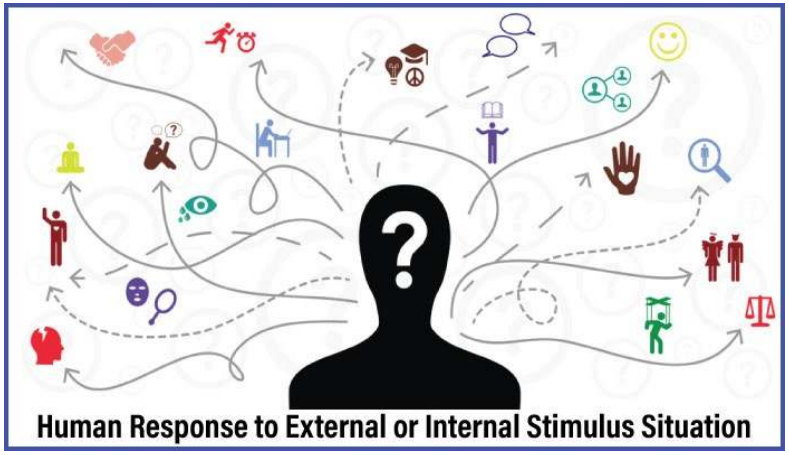
Existing explanations for SSH



Special subject:

- Structural complexity of human behavior
- Difficulty in representing or quantifying behavior
- Short-run adaptiveness of human behavior
- **Medium-run change in human behavior**
- Context sensitivity of human behavior
- Randomness/noise of human behavior

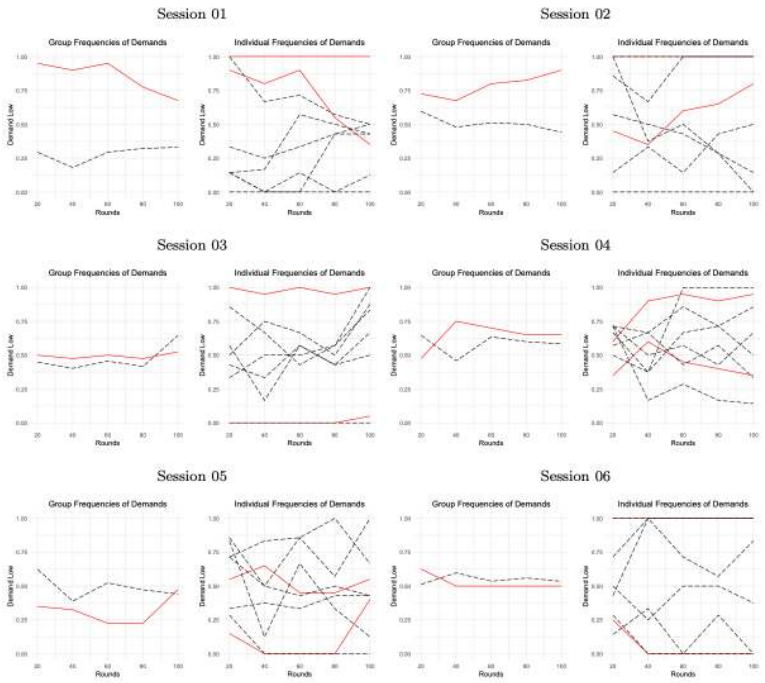
Existing explanations for SSH



Special subject:

- Structural complexity of human behavior
- Difficulty in representing or quantifying behavior
- Short-run adaptiveness of human behavior
- Medium-run change in human behavior
- **Context sensitivity of human behavior**
- Randomness/noise of human behavior

Existing explanations for SSH



Special subject:

- Structural complexity of human behavior
- Difficulty in representing or quantifying behavior
- Short-run adaptiveness of human behavior
- Medium-run change in human behavior
- Context sensitivity of human behavior
- **Randomness/noise of human behavior**

The **special subject** hypothesis

The hardness of the social sciences is due to the greater complexity of its subject—individual and collective human behavior—relative to other sciences.

The **special subject** hypothesis

Hardness, as it were, is in the human.

So, is this true?

Possibly.

But we don't think this is all that's at play.

We focus on *more general* features of sciences.

Aims of this talk

Present two novel hypotheses.

Aims of this talk

1. The **rigid demands** hypothesis

Aims of this talk

1. The **rigid demands** hypothesis

2. The **fruit in the hand** hypothesis

First, what does it mean to be ‘hard’.

SSH1. The social science are hard *relative* to other sciences.

SSH2. In the social sciences, we tend to exhibit *worse performance in prediction tasks of interest* relative to the performance in prediction tasks of other sciences.

SSH3. In the social sciences, we tend to exhibit *worse performance in our most successful prediction tasks of interest* relative to the performance in prediction tasks of other sciences.

Is this modified claim a good surrogate?

Clearly, it misses many important things—science is not only about predictive performance—we care about causation, practical intervention, unification, simplicity, explanation, and so on.

But let's start here.

Our alternative hypotheses

1. The **special subject** hypotheses
2. The **rigid demands** hypothesis
3. The **fruit in the hand** hypothesis

2 | Rigid demands hypothesis (RD)

A vignette



Hard scientist: finds an $R^2<0.8$



Social scientist: finds an $R^2>0.4$

Typically, a high R-square of above 60% is required for studies in the 'pure science' field because the behavior of molecules and/or particles can be reasonably predicted to some degree of accuracy in science research; while an R-square as low as 10% is generally accepted for studies in the field of arts, humanities and social sciences because human behavior cannot be accurately predicted.

—Ozili et al, 2022, "The Acceptable R-Square in Empirical Modelling for Social Science Research"

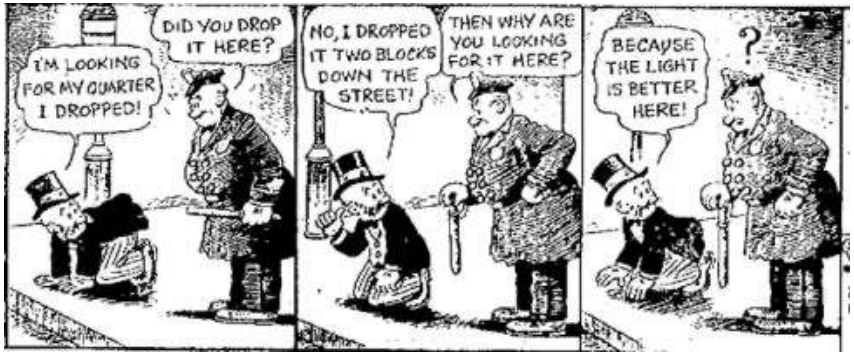
Idea

It may be that the social scientist has her questions more rigidly fixed and so she is in measure to this less able to use one of the great techniques in science:

Changing her research question to a more tractable one.

Idea

It may be that the social scientist has her questions more rigidly fixed and so she is in measure to this less able to use one of the great techniques in science: leaving her research question to pursue a more tractable one.

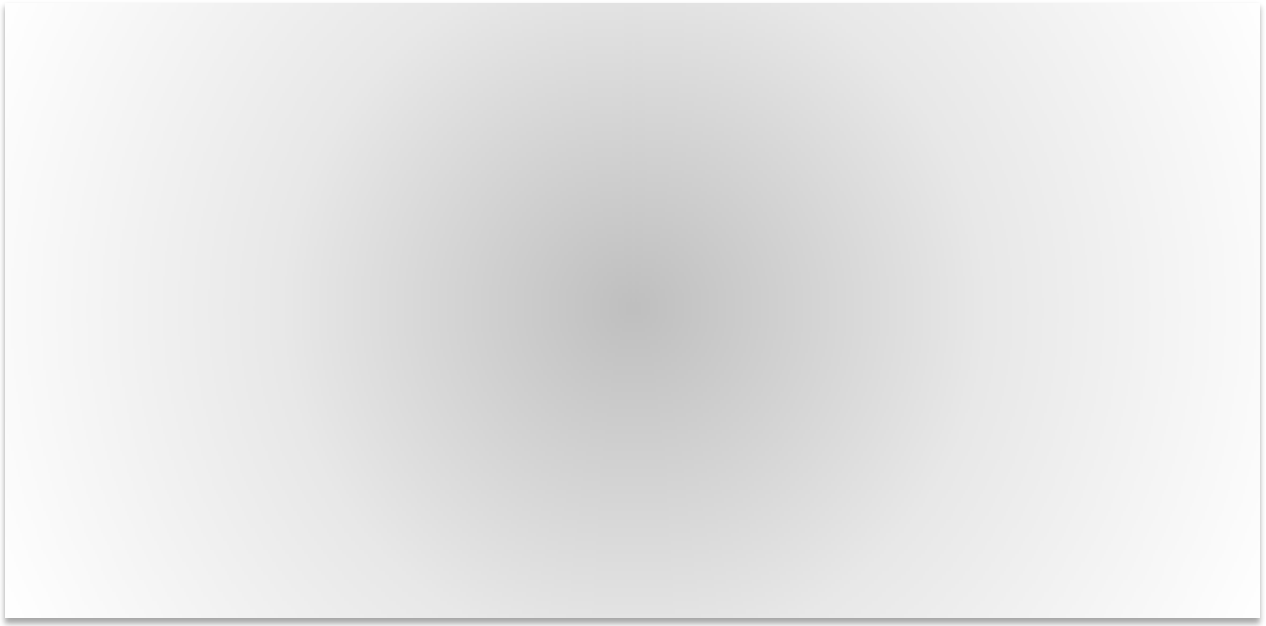




Why should being less able to change one's research direction tend to lead to worse predictive performance?

Let's approach this from a different angle.

The world can be thought of as composed of prediction tasks of various sorts.



The world can be thought of as composed of prediction tasks of various sorts.



The world can be thought of as composed of prediction tasks of various sorts.



Random fact about some
corner of a distant galaxy.

The world can be thought of as composed of prediction tasks of various sorts.



Random fact about some corner of a distant galaxy.

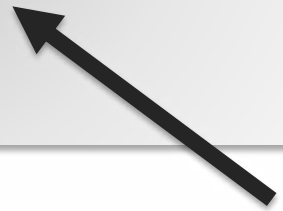
Relationship between the number of grains of sands on a beach on the Caspian sea and time since the big bang.

The world can be thought of as composed of prediction tasks of various sorts.

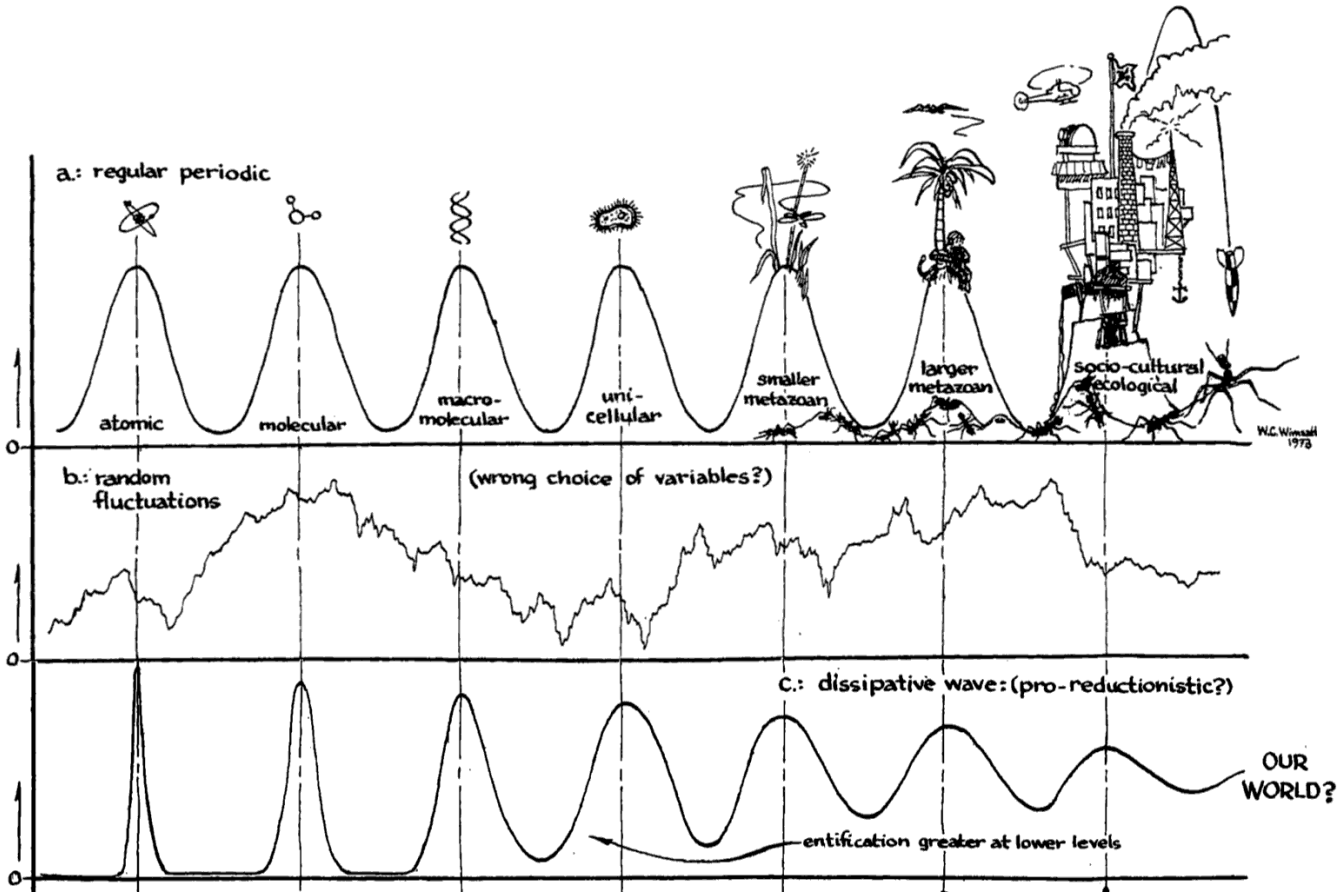
Most of these prediction tasks are going to be quite **hard** for us.



Random fact about some corner of a distant galaxy.



Relationship between the number of grains of sands on a beach on the Caspian sea and time since the big bang.



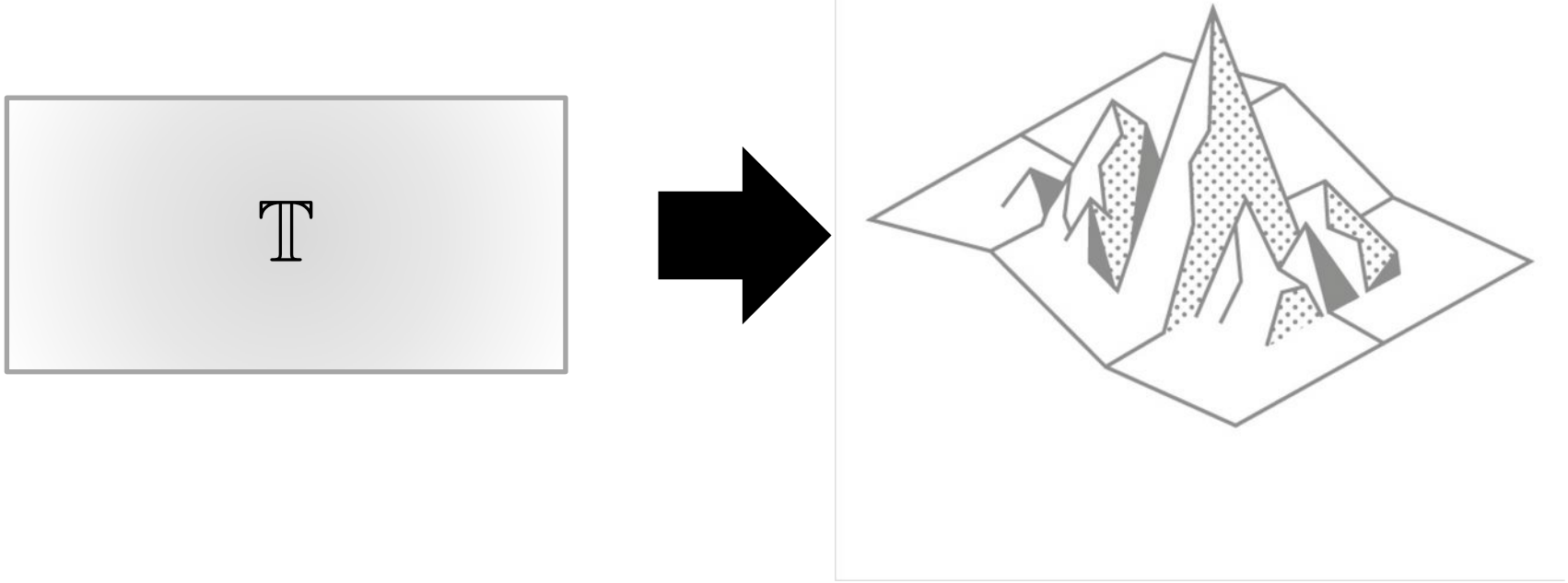
[l]evels of organization can be thought of as local maxima of regularity and predictability in the phase space of alternative modes of organization of matter. — William Wimsatt, 1976, p. 209

Let the hardness of a prediction task be given by a real number. (This might be an expected loss/error score.)

Hardness of a task $t \in \mathbb{T}$

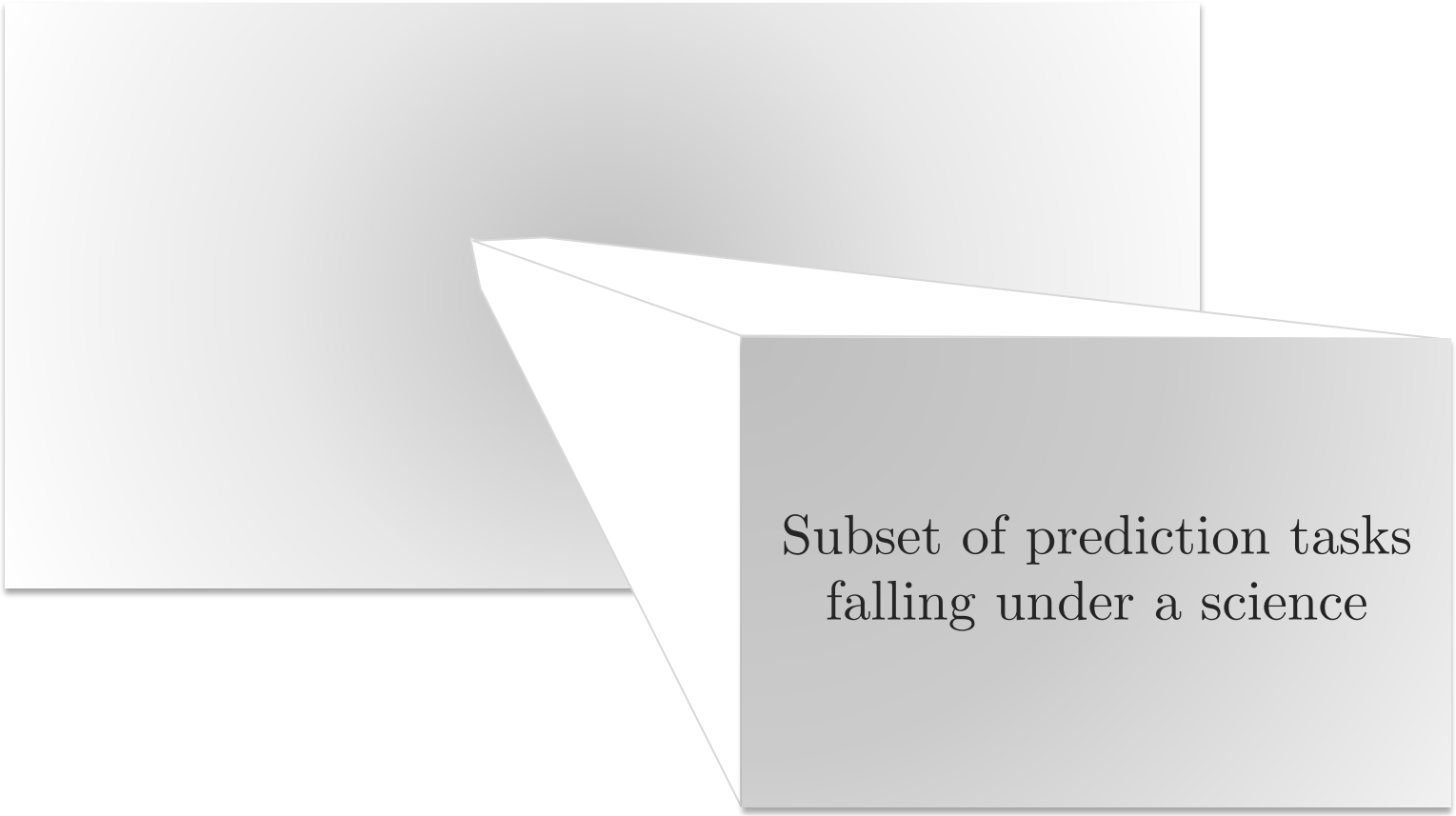
$$H := \mathbb{T} \rightarrow \mathbb{R}^+$$

This can be visualized.



There is *some* (potentially unknown) distribution of hardness of prediction tasks.

A particular scientific discipline—chemistry, biology, economics—can be thought of as focusing investigation into some subset of the set of possible prediction tasks.

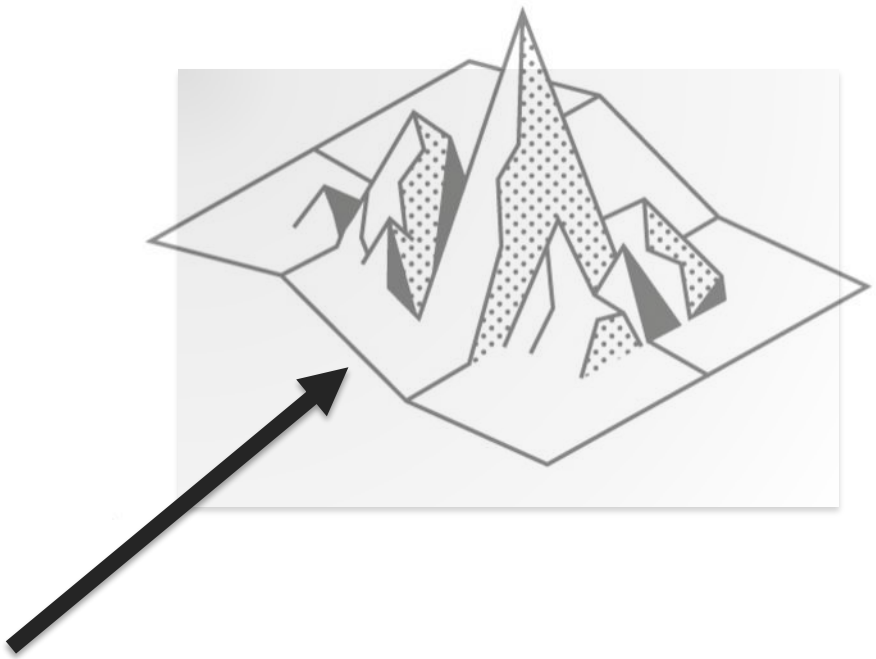


For some disciplines—for whatever reasons—the set of acceptable tasks may be larger or smaller.



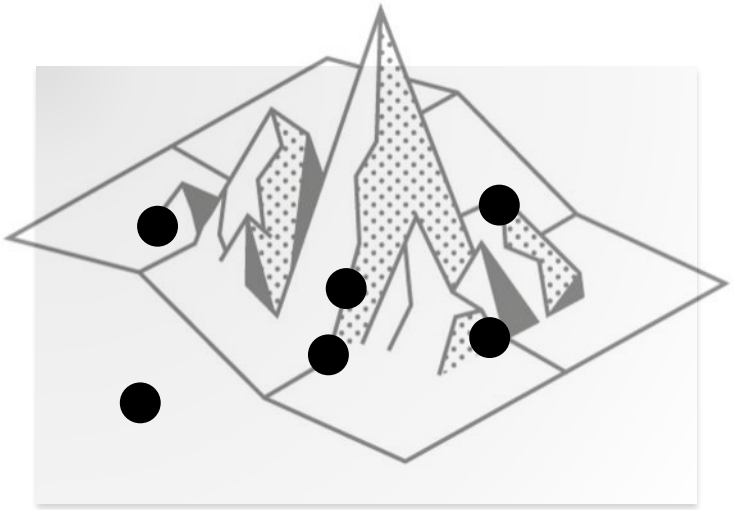
Indeed, some may be strict subsets of others—e.g., genetics' tasks will be a subset of biology tasks.

The success of a given science then will be in finding tractable, reliable regularities.



These can be thought of as valleys—low hardness points (or sequences of points) in the set of tasks.

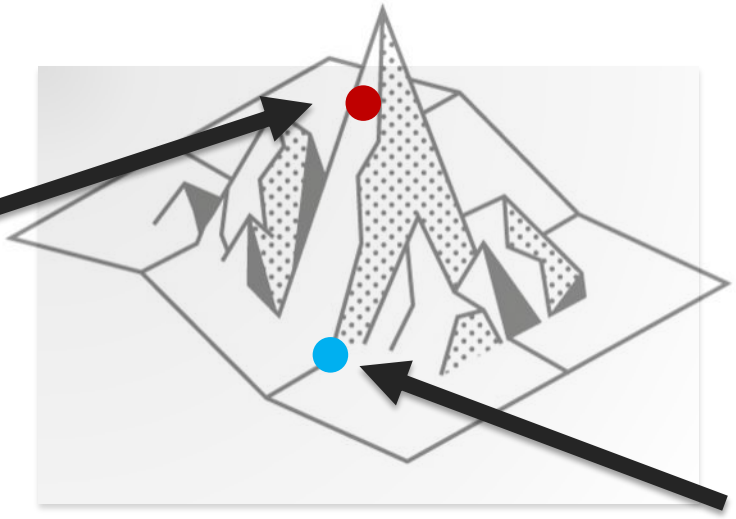
As a first pass, let's say scientists explore this set by selecting elements at random to investigate.



The scientist will make investigations with degrees of success inversely proportional to their difficulty.

A science may plausibly be evaluated by its *most successful performances*.

Hard,
higher error,
less success



Easy,
lower error,
more success

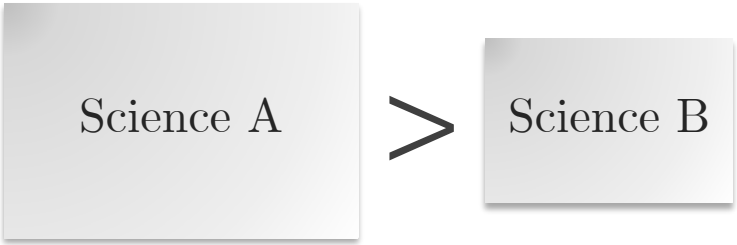
Hardness of a discipline $D_i \subset \mathbb{T}$

$$H(D_i) := \min\{h | \exists t \in D_i : h = H(t)\}$$

This can be state as "a science may be evaluated by its *least hard prediction task performed*."

What might inform our expectations for the value of the most successful prediction tasks within a domain?

Actually, mere size can.



$$\rightarrow$$

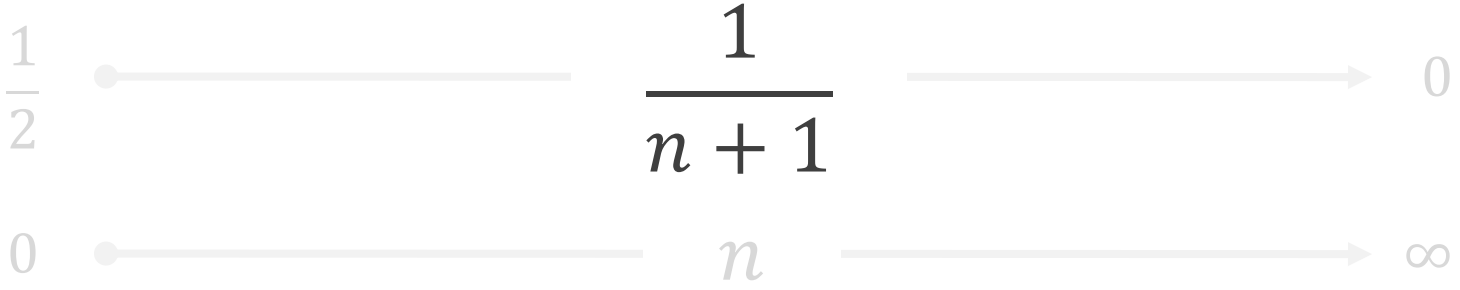
$$\mathbb{E}[\min\{H(t_i \in A)\}] < \mathbb{E}\left[\min\left\{H\left(t_j \in B\right)\right\}\right]$$

In other words, all else equal, we should expect that the most successful predictions will grow with the size of the set of possible prediction tasks.

Toy example. Imagine all tasks are binary prediction tasks and the hardness of a task is given by the Brier score attainable on that task, and let hardness be uniformly distributed.

It follows that the expected value of the most successful performance (minimum error) on a prediction task across a set of n prediction tasks is

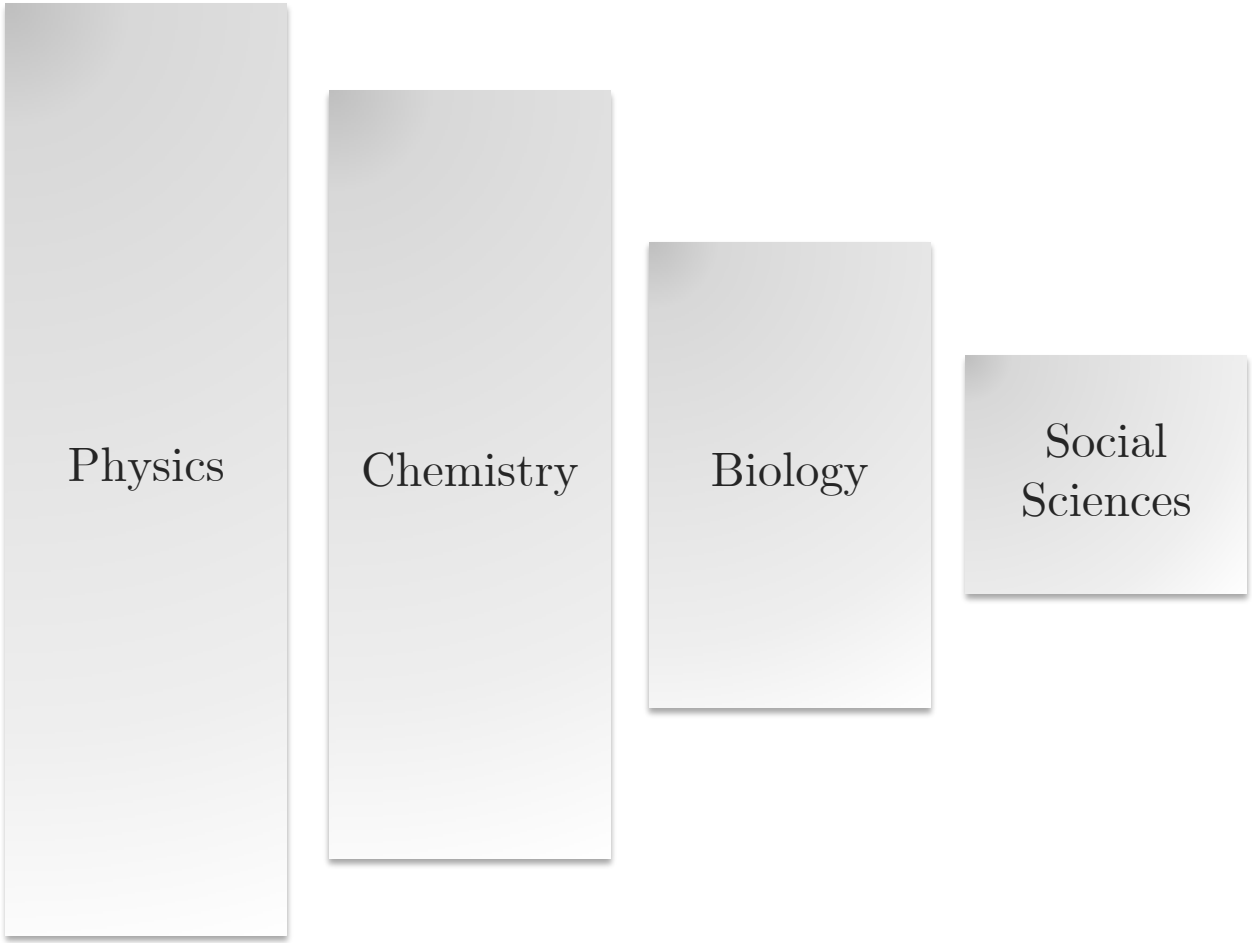
$$\frac{1}{n+1}$$



Prediction: Fields with smaller domains will tend to perform more poorly on their best tasks.

So, do the social sciences have a smaller domain of inquiry than hard sciences?

If this picture is qualitatively correct, then yes.



Intuition Pump

Two medical research teams are tasked with investigating two domains of inquiry and return to us in 10 years with their greatest success.

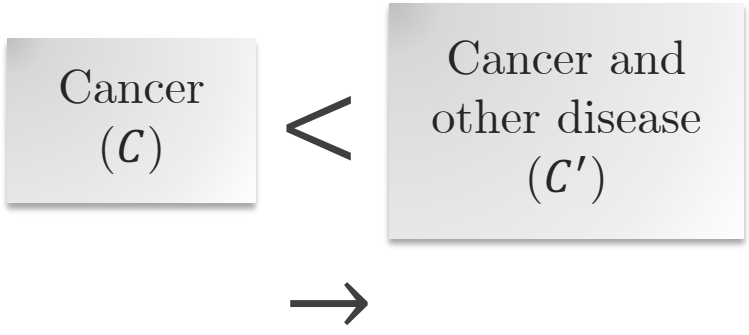
Intuition Pump

One we assign one team the task finding the best treatment they can for cancers; there we assign to finding the best treatment for *any disease*.

Intuition Pump

Whose greatest success should we expect will be greater?

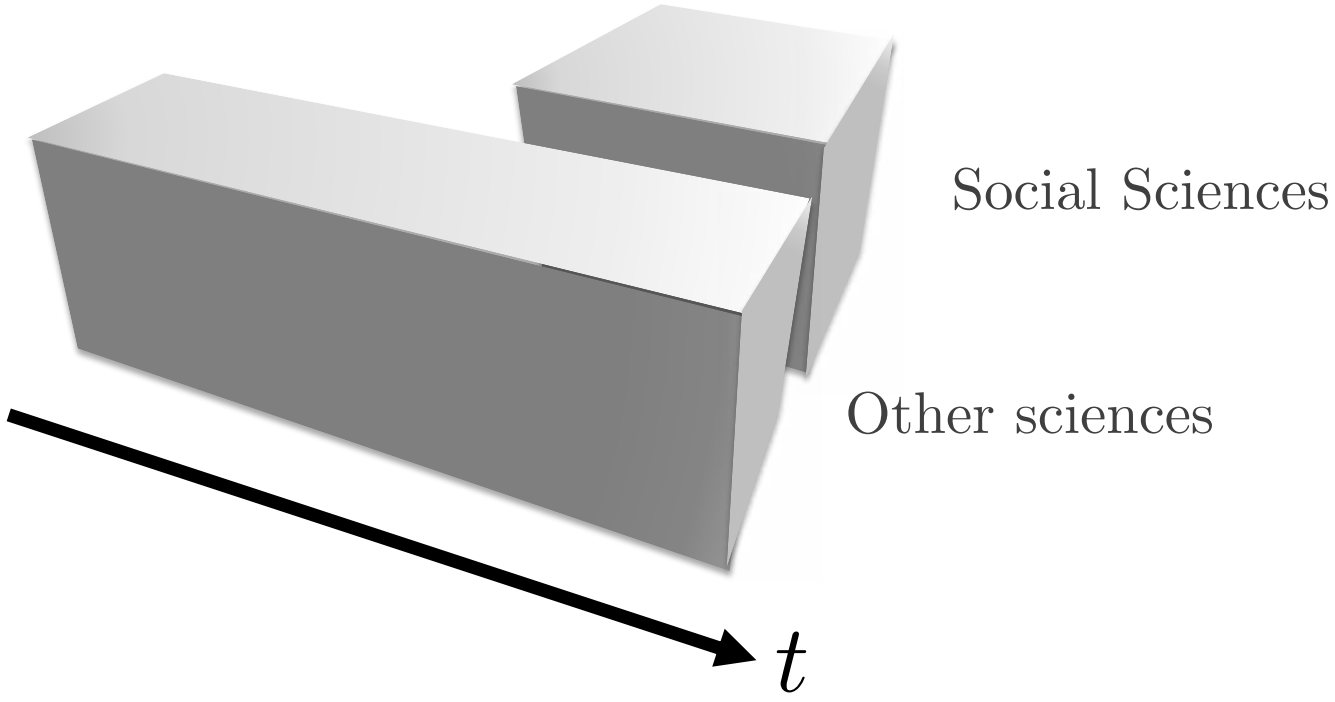
This is an instance of the previous result.



$$\mathbb{E}\left[\min\left\{H\left(t_j \in C\right)\right\}\right] < \mathbb{E}\left[\min\left\{H\left(t_i \in C'\right)\right\}\right]$$

But this is not the only way to get a difference in cardinality of the set of prediction tasks.

This also works.



(Thanks to Kevin Zollman for pointing this out.)

"The natural sciences, now so fruitful in predictions, were well established by the 17th century. The social sciences are at least three centuries behind them, consequently, in the race for achievement."

—Chambliss (1958), "Prediction in Social Sciences," *Social Science*, p. 91

Prediction: Fields with smaller domains will tend to perform more poorly on their best tasks.

Do the social sciences perform worse at their best tasks?

If the strength of relationships (R^2) meriting publication is a reasonable surrogate measure for performance, then yes.





Mean R^2 of publications in in physical science journal.

Mean $R^2 \approx .04$ of publications in social psychology journals. $(n = 25k)$

Richard et al, 2003, "One Hundred Years of Social Psychology Quantitatively Described," *Review of General Psychology*, p. 331–363

"If you needed to do a statistical test, you just did a bad experiment," and "If you needed statistics, you are studying something so trifling it doesn't matter"

—Reinagel 2022 "The limited role of null hypothesis testing in biology"

Proposition. Let D be a discipline and assume that the tasks in D are IID and the distributions non-degenerate. Then, as $|D|$ decreases, holding the distribution of the difficulty of tasks in D fixed, $\mathbb{E}(H(D))$ increases.

Our vignette revisited





This makes sense under the **rigid demands hypothesis**.
In virtue of the smaller scope of her domain of inquiry the social scientist has to make do with weaker results.

The RD hypothesis implies that the size of the prediction tasks composing the domain of inquiry of a discipline should inform our expectation of the maximal performance attained by that discipline.

The social sciences—both in focusing on human behavior and in being a relatively younger discipline—already focus on a constrained subset of all possible tasks.

3 | Fruit in the hand hypothesis

Fruit in the Hand Hypothesis

Due to our evolutionary endowment and distinctive enculturation, many of the easy prediction tasks in the social domain have already been solved; we have already picked all or most of the low hanging fruit.

But we do not count these successes are part of the successes of the social sciences. This leaves the harder problems, high up in the tree, for the social scientists to tackle.

Further, our evolutionary and cultural inheritance has not given us the same gifts when it comes to the other sciences.

Intuition pump

Imagine: You are in a room full of philosophers quietly listening to a talk...

We do not include such prediction tasks as part of the human sciences, even though they concern the behavior of humans in response to stimuli.

Plucking the Fruit

One way of making FTH precise is to imagine starting with a discipline, and then removing some of the easiest prediction tasks from it.

As the easiest tasks are removed, the expected difficulty of the discipline goes up.

Proposition 2. Let D be a discipline, and assume that the tasks in D are IID and the distributions non-degenerate. Let $l \in \arg min_{t \in D} h(t)$, and let t be a task in D chosen uniformly at random. Then $E(H(D - \{l\})) \geq E(H(D - \{t\}))$ and $E(H(D - \{l\})) \geq E(H(D))$.

Domain of social science prediction tasks

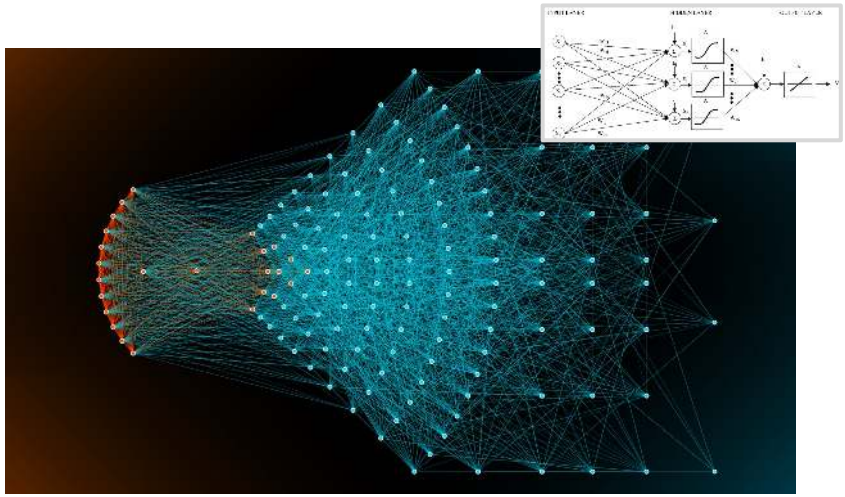
Human behavior prediction tasks



Human behavior prediction tasks not already tackled by our evolved cognition

$$E = \begin{cases} \sum_{n=1}^{\infty} \frac{M^{\frac{n}{3}}}{n!} \lim_{\theta \to 0} \left(\frac{\mathrm{d}^{n-1}}{\mathrm{d}\theta^{n-1}} \left[\left(\frac{\theta}{\sqrt[3]{\theta - \sin(\theta)}} \right)^n \right] \right), & \epsilon = 1 \\ \sum_{n=1}^{\infty} \frac{M^n}{n!} \lim_{\theta \to 0} \left(\frac{\mathrm{d}^{n-1}}{\mathrm{d}\theta^{n-1}} \left[\left(\frac{\theta}{\theta - \epsilon \cdot \sin(\theta)} \right)^n \right] \right), & \epsilon \neq 1 \end{cases}$$

Functional complexity of successful solution to prediction task: predict the orbital behavior of some medium-sized dry good.



Functional complexity of successful solution to prediction task: predict the continuation of some input text.

GPT−3

Ex. 175 B parameters

Intuition pump

Imagine an alien that it did not have our cognitive tool kit for predicting human behavior, or anything much like it from its own experience.

The alien may consider many prediction tasks we perform routinely and reliably and consider trivial as at least as difficult to replicate as our performance in prediction tasks in the hard sciences.

$$H\Big(\text{Mid-sized dry good } x \text{ will proceed on this orbit } y.\Big) \leq H\Big(\text{Input text } u \text{ can be sensibly continued by the sequence of tokens } v.\Big)$$



Discussion

1. The **special subject** hypothesis
2. The **poor methods** hypothesis
3. The **bad incentives** hypothesis

Discussion

1. The **special subject** hypothesis
2. The **poor methods** hypothesis
3. The **bad incentives** hypothesis
4. The **rigid demands** hypothesis
5. The **fruit in the hand** hypothesis

Discussion

Our hypotheses do not negate other, specific causes of difficulty; they are logically compatible.

Discussion

What are we talking about when we talk about hardness? Many metrics (like average success) are genuinely quite difficult to understand in this context.

Discussion

What evidence might be used to test these hypotheses?

Thank you.

