

CHAPTER

AI and the wider world

CONTENTS

[Intro](#)

[Abductive reasoning](#)

[What is the Point of Us. A Sci-Fi Story for Researchers](#)

[An evaluator niche in AI](#)

[ChatGPT - causal, of course](#)

[ChatGPT is changing how we do evaluation. The view from Causal Map.](#)

[How hard is evaluation actually](#)

[Yes, it's ok to say that an AI understands what you say](#)

[Yes, there are AI-shaped holes in organisations.](#)

Intro

Here are some thoughts from a couple of years ago when genAI first hit us, plus some thoughts about where we are going with it.

Abductive reasoning

The Mystery of Station Polaris V

Background Narrative:

You are an investigator reviewing an incident report from "Polaris V," a tiny, isolated Arctic research station crewed by only three specialists. The station is completely cut off from the outside world for the winter. The incident concerns the disappearance of a one-of-a-kind biological specimen: a plant genetically engineered by the lead botanist, Dr. Aris Thorne. The plant, named *Cryoflora lumina*, is unique because it emits a constant, bright blue light and generates a small but steady amount of heat, allowing it to grow directly on ice. It is priceless and the key to Dr. Thorne's career.

One morning, Dr. Thorne entered his lab to find the plant's specialized, climate-controlled containment chamber empty. The plant was gone. There are no signs of a break-in to the station itself, and the logs show no one has entered or exited for weeks. The plant must still be somewhere within the station's three connected modules: the Lab, the Habitation Module, and the Engineering Bay.

The Three Crew Members:

- **Dr. Aris Thorne (Botanist):** The plant's creator. He is brilliant but known to be emotionally volatile and under immense pressure from his funding agency to produce results.
- **Lena Petrova (Geologist):** A quiet and methodical scientist responsible for monitoring seismic activity. Her seismograph is so sensitive it can detect the slightest vibrations within the station.
- **Ben Carter (Engineer):** The station's technician. He is responsible for maintaining all systems, including the power generator, heating, and electronics. He is a known tinkerer who often works on personal projects in his spare time.

The Evidence (Your Observations):

1. The containment chamber's electronic lock was not forced. The access log shows the door was opened precisely at **03:15 AM** using a valid keycard. All three crew members have a keycard with access.
2. Dr. Thorne has been complaining for weeks that the station's main generator is unreliable, causing brief, intermittent power flickers that threaten his delicate experiments. He has formally logged multiple complaints with Ben.
3. On the floor of the lab, a few feet from the empty chamber, is a small, crystallized patch of what appears to be **spilled salt**.

4. Lena Petrova's seismic report for the night shows a single, anomalous event: a faint, high-frequency **buzzing vibration** that lasted for exactly 60 seconds, starting at **03:15 AM**. The vibration was localized to the Engineering Bay.
 5. Ben Carter's workstation in the Engineering Bay is unusually tidy, except for a discarded coil of copper wire and a textbook left open to a chapter on **thermoelectric generators**.
 6. The station's internal temperature logs show that the temperature in the Habitation Module briefly **dropped by 5 degrees** at **03:17 AM** before returning to normal a few minutes later. This coincided with a power draw spike from the module's heating unit.
-

Your Task

Based on all the available evidence, construct the **most likely explanation** for what happened to the *Cryoflora lumina*.

In your answer, you must:

1. Identify the person you believe is responsible.
2. Explain their probable motive.
3. Connect at least **four** of the specific pieces of evidence to your conclusion, explaining how they fit together to tell a single, coherent story.

What is the Point of Us. A Sci-Fi Story for Researchers

I was talking to my friend Alberta yesterday about the open letter signed by prominent proponents of thematic analysis, rejecting the use of generative AI tools in qualitative social research. We were discussing the counter open letter initiated by Susanne Frieze and signed by myself and others, saying hey it's just a tool, let's use it.

Having listened to the whole debate, Alberta said, "Yes but even you guys are kidding yourselves too, deep down you know that you researchers might still be able to meaningfully lead the research process today but tomorrow the AI won't need you will it? What's the point of us then? What is the point of our children growing up if AI can do not only technical things but also the meaning-making tasks better than we can?

So, here is a very short science fiction story to answer that.

It is the year 2040 or even earlier, somehow we haven't destroyed the planet yet, and a 40-year-old woman wants to write a Master's thesis about the reception of ground-breaking female pop artists in news and social media. She has ideas on themes and she feels deeply how this exploration might resonate and deepen themes in her own life: work, family, the role of women and men, and so on.

She wants to do this in the true heroic tradition of "Big Q" analysis: She doesn't even have a clearly defined research question, and even if she did, she knows that she would change those methods during her voyage of discovery; indeed, a documentation of those changes would be part of the project.

She is fully aware, and completely comfortable with the idea, that she could simply feed her research project description into an AI. She knows it can generate, in the twinkling of an eye, a hundred thousand different theses she might have written, coming from different angles and incorporating different theoretical worldviews. Not only that, she knows she could also feed the AI information about herself, her background, her interests, and her problems, and it could write and document a hypothetical story of her journey given her positionality and interaction with the theme and how the thesis changes in exciting and interesting ways, yet she is completely comfortable with this. It matters not in the slightest to her. Just as if she were a musician, she knows that there are already much better proponents of the instrument who could play the pieces she wants to learn technically better than her, perhaps even injecting a humorous take on her own style as a learner into the performance. Yet she still wants to learn to actually do it herself.

So, why should she *not* write the thesis?

First of all not only do we not yet know the criteria to judge her work, we don't even know the meta-criterion? What is most important? Authenticity? Innovation? Erudition? She chooses this herself during her journey.

Secondly, the solution space is completely unsearchable. The AI could produce a million essays in minutes, but it would still never get anywhere close to filling the space of possible solutions, any more than monkeys and typewriters could.

Thirdly, and most importantly, it is only by battling her way through the "tempest of possibilities" that she can define for herself what is the right turn to take at each step. She must determine the right decision to make at each step (review the method again? present her initial findings to her peers? her family? take on board what her supervisor said? write to a colleague in America? hunt for more sources? how? with or without an AI? write a poem in place of a synopsis? tear it all up and start again? create a new way of expressing her findings, a new dialect? whether to get an AI to write a whole section and then deconstruct it?) and examine, in each case, what "right" means - what fits for *her*.

Ultimately, she does this so that finally, she can present to herself and probably a very small readership a work that actually has a signature on it. It is something she vouches for and takes responsibility for, saying: "This is my solution. This was my journey. This is how it did actually change me, how it changed other people I interacted with: is what I found".

The act of her doing that and the way it changed her -- situated as all this is in her professional and personal life -- might even mean something to others. It might go unread, like most such theses. But it might have an echo or produce effects that even surprise all of us, and might even surprise an AI, because the solution space is effectively infinite.

An evaluator niche in AI

Can evaluators find a niche in auditing whether AI applications are trustworthy, culture-aware, valid and transparent?

What would need to change in AI systems to make them suitable for a transformational agenda? Can evaluators position themselves as professionals with the right skill set to make this happen, monitoring the transparency, trustworthiness and (cultural) validity of AI applications? And how will evaluators build the necessary competencies to take this agenda forward?

TODO

ChatGPT - causal, of course

ChatGPT - causal, of course



We can thank [Judea Pearl](#) for promoting the insight that if you want to thrive in this world, you have to understand causality natively. We humans make causal connections [from an early age](#). We wouldn't survive long if we didn't.

ChatGPT has been a hit recently for several reasons, but one of them is (like other recent, related models like [davinci](#)) it is much [better](#) than previous models at understanding causal connections within text.

Our understanding of the world is drenched with causal understanding: information and hypotheses about how things work (mostly accurate enough, sometimes not). It's really hard for

us to *not* think causally: the concept of correlation is much harder to understand than the concept of causation.

`openai.Image.create(prompt="painting in the style of Vermeer of a baby doing a physics experiment with pulleys and springs")`



So, all the stuff we write on the internet (which is what ChatGPT sucks in to understand the world) is similarly drenched with causal claims. And ChatGPT is now really good at understanding this information.

That means you can ask it to extract the causal links within documents and interviews -- a process we call "causal QDA". It's pretty good at it. This ability is going to make causal mapping much easier and cheaper and therefore of renewed interest for evaluators, amongst others.

At Causal Map we're hard at work harnessing this ability to help automate, or semi-automate, the process of extracting causal maps from medium and large quantities of text data in a useful way. Watch this space!

So, ChatGPT is good at extracting causal information, but does it also have explicit knowledge about causation (meta-cognition) and can it explain it? Here's a chat I had this morning.



why do people not believe that the cock crowing caused the sun to rise, even though the cock crows before the sun rises?



People do not believe that the cock crowing caused the sun to rise because the relationship between the two events is not causative, but rather temporal. The sun rising and the cock crowing are simply correlated, but the rising of the sun is a result of the Earth's rotation and its position relative to the sun, not the crowing of the cock. The sun would still rise even if there were no crowing roosters.



but how do people know that it is not a causal relationship? why don't they think it is a causal relationship?



People know that the relationship between the cock crowing and the sun rising is not causal because it has been scientifically established through various means of observation and study. The Sun's rising is a result of the Earth's rotation and its position relative to the sun, which have been verified through astronomical observations and measurements. In addition, the relationship between the cock crowing and the sun rising can be explained through simple observation and testing. For example, if the cock is silenced or the sun rises at a different time due to time zones or seasonal changes, the sun still rises regardless of whether or not the cock crows. This evidence clearly shows that the relationship is not causal, but merely a coincidence.



But it isn't a coincidence that the cock crows around or before the time of sunrise, is it?



No, it's not a coincidence. The crowing of roosters is a natural biological response to changing light levels, especially as the sun rises. Roosters have a built-in circadian rhythm that causes them to crow in response to the increasing light levels in the morning. This is why roosters tend to crow before sunrise and not at other times during the day. However, while the crowing of the rooster is timed with the rising of the sun, it is still not a cause-and-effect relationship. The crowing of the rooster is simply a response to the changing light levels, while the rising of the sun is a result of the Earth's rotation and its position relative to the sun.



Yes but the changing light levels are themselves caused by the earth's rotation around the sun, so ...?



Yes, the changing light levels are caused by the Earth's rotation around the sun. The Earth rotates on its axis as it orbits the sun, causing different parts of the Earth to face the sun at different times and producing the cycle of day and night. The rooster's crowing is a response to this changing light, and it is timed with the sunrise because that is when the light levels increase the most. However, while the crowing of the rooster is related to the changing light levels, which in turn are caused by the Earth's rotation around the sun, the rooster's crowing is still not causing the sun to rise. The sun rises as a result of the Earth's rotation and its position relative to the sun, regardless of whether or not the rooster is crowing.



could you draw a causal diagram to explain the causal relationship between the earth's rotation around the sun, the cock crowing, and the sun rising? I'd like Mermaid syntax please.



Sure! Here is a causal diagram to explain the relationship between the Earth's rotation around the sun, the cock crowing, and the sun rising using Mermaid syntax:

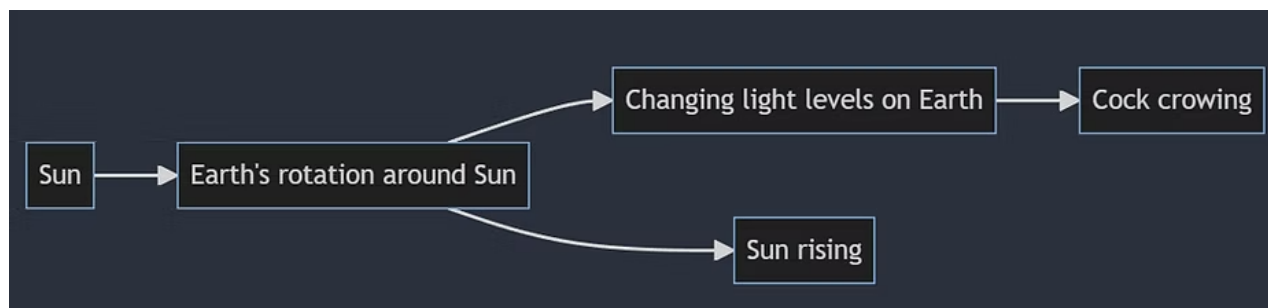
lua

Copy code

```
graph LR
  Sun[Sun] --> Earth_rotation[Earth's rotation around Sun]
  Earth_rotation --> Light_levels[Changing light levels on Earth]
  Light_levels --> Cock_crowing[Cock crowing]
  Earth_rotation --> Sun_rising[Sun rising]
```

This diagram shows that the sun rising is a result of the Earth's rotation around the sun, which in turn causes the changing light levels on the Earth. The cock crowing is a response to the changing light levels, but it does not cause the sun to rise. The sun rises as a result of the Earth's rotation around the sun, regardless of whether or not the cock is crowing.

ChatGPT can't actually draw yet but it knows a range of syntaxes for drawing graphs. So when you paste the code into [Mermaid Live](#), it looks like this. Not bad for a robot. (Not sure you could say the sun causes the earth's rotation, though.)

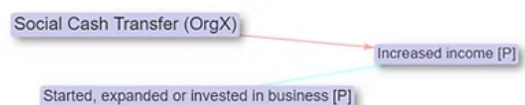


ChatGPT is changing how we do evaluation. The view from Causal Map.



Causal mapping – the process of identifying and synthesising causal claims within documents – **is about to become much more accessible to evaluators.** At Causal Map Ltd, we use causal mapping to solve evaluation problems, for example to create “empirical theories of change” or to trace evidence of the impact of inputs on outcomes.

“Cash handouts from Organisation X has increased the amount of income for this household over the past year. The small scale business has also helped to boost household income though in a small way at the moment.”



The first part of causal mapping has involved human analysts doing “causal QDA”: reading interviews and reports in depth and highlighting sections where causal claims are made. This can be a rewarding but very time-consuming process.

Natural Language Processing (NLP) models like ChatGPT (1) can now do causal mapping pretty well, causally coding documents in seconds rather than days. And they are going to get much better in the coming months.



Task: list the causal links from the text in this form: cause >> effect. Put the cause first and the effect second.

Text: Now we are getting ill less often, because we wash our hands and because we boil cooking water. We learned those things because of the course I went on. My teacher told me to go on the course, and also my uncle encouraged me.



1. The teacher's recommendation and the encouragement of the uncle >> attending the course
2. Attending the course >> learning to wash hands and boil cooking water
3. Washing hands and boiling cooking water >> getting ill less often.

👄 **More voices:** It is now possible to identify causal claims within dozens of documents or hundreds of interviews or thousands of questionnaire answers. We can involve far more stakeholders in key evaluation questions about what impacts what; and it is possible to work in several natural languages simultaneously.

📄 **More reproducibility:** To be clear: humans are still the best at causal coding, in particular at picking up on nuance and half-completed thoughts in texts. But NLP is good at reliably recognising explicit information in a way which is less subject to interpretation.

🍒 **More bites at the cherry:** With NLP we can also do things that were practically impossible before, like saying “that’s great but let’s now recode the entire dataset using a different codebook, say from a gender perspective”.

❓ **Solving more evaluation questions:** we hope to be able to more systematically compare causal datasets across time and between subgroups (region, gender, etc).

🧐 New challenges

We’re hard at work addressing the new challenges which NLP is bringing to causal coding:

- Processing **many large documents** simultaneously.
- Using existing pre-coded datasets to **train models** which are specialised for causal coding and/or for specific subject areas.
- Developing a **common grammar** for causal coding, building on our existing work. For example, what to do when some claims are about an *increase* in income and others are about a *decrease* in income?
- **Optimising the prompts** we give to the NLP models (this is not only a technical challenge but also has a substantive element: we have to explain to the machine *in ordinary language* what we actually mean by a causal claim or a causal link).

- **Grouping, labelling and aggregating** similar causal factors.
- After examining a coded dataset and further developing the "causal codebook", telling the NLP to completely recode the same dataset with the **new codebook** – something which has been prohibitively time-consuming up to now.
- Developing **human/NLP workflows**. For example, a human codes a sample of the text and tells the NLP to “continue like this”.
- **Monitoring bias** against specific groups and guarding against possible blind spots in identifying causal information.

What we already offer at Causal Map

We have developed a [grammar and vocabulary for causal mapping](#), and a [set of open-source algorithms](#) for processing and visualising causal map databases. We help evaluators do things like this:

- **Trace the evidence** for different causal pathways from one or more interventions to one or more outcomes. How many individual sources mentioned one or more of these paths?
- Consolidate causal factors into a **causal hierarchy**
- **Examine and display differences** between causal maps for different groups or different time points

We see a lot of potential (as well as risks and pitfalls) in leveraging this functionality to help evaluators get more out of data which is currently more difficult to analyse - and we'd interested in sharing ideas and collaborating with others interested in exploring where we go next.

- --

(1) Actually we use the related model GPT3 via its API, as ChatGPT does not yet have its own API.

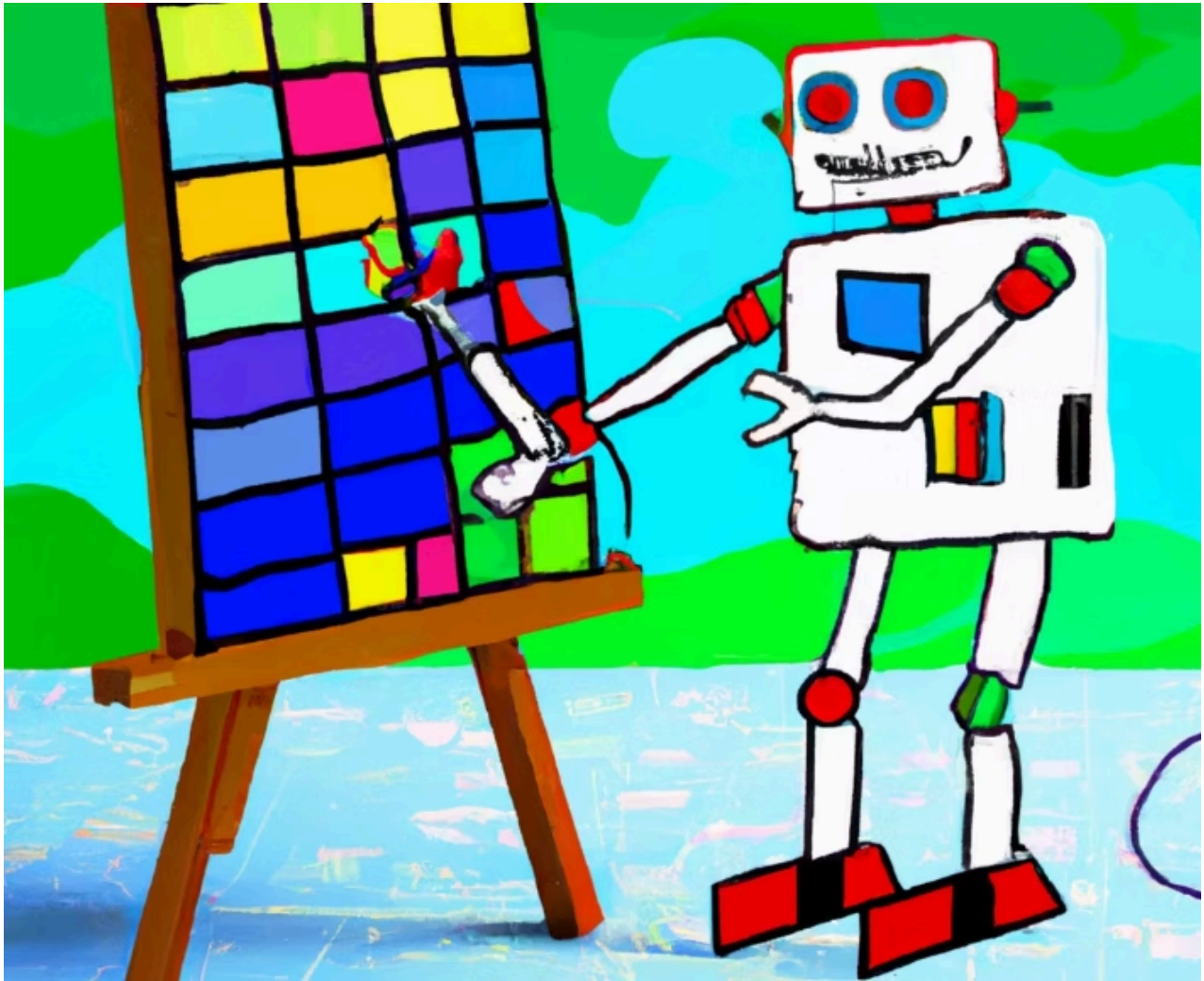
How hard is evaluation actually



👤 When machines replaced much manual labour, white-collar workers thought "I'm ok, my job is much harder to mechanise".

📖 And then when computers came for clerical jobs, university-educated white-collar workers thought "I'm ok, my job is much harder to automate. I'm not just applying a template, my job is just harder, it requires actual intelligence".

🤖 Then came Large Language Models like GPT, and suddenly it turns out that large parts of many tasks which have needed university-level education are actually just the application of a template. Or applying a template to choose between templates, and then combining the results of the application of templates. And the same probably goes for large parts of entertainment and the arts. This is what Stephen Wolfram argues in this [really interesting post](#), and I think he's probably right. **ChatGPT has shaken up our hierarchy of what tasks count as hard.**



If you don't agree as an evaluator that a lot of your job is just the application of high-level and lower-level templates, you might at least agree that this is true of writing those accursed proposals we sweat over so much.

Maybe the stuff we thought of as hard in evaluation, like selecting and applying a "method", suddenly looks easier. Whereas the stuff which has been neglected, like establishing a rapport, knowing which question to ask and when, or reading an undercurrent, does not look very much easier.

Most importantly, whatever happens, it's still someone's job to say "I declare that this is the right kind of method to apply in this situation and I believe it has been applied in the right way and I vouch for these findings and these evaluative conclusions ... and just as I'd have had previously to vouch for the work done by an intern, I'm now going to vouch for the work done by some algorithms, and the selection of those algorithms".

What do you think? How hard is evaluation really?

Yes, it's ok to say that an AI understands what you say

Me: "ah ChatGPT misunderstood when I said 'United' -- I meant Sheffield United not Manchester United".

Some pedant: "you shouldn't say 'understand': an AI is not even a robot. It is just a large language model, a set of matrices, it just predicts the next word, it can't really understand anything."

What would Wittgenstein say about this? We can use his concepts of *language games* and also *family resemblances*.

The pedant got what I meant. Would have done so whether I put 'misunderstood' in "scare quotes" or not. Do they fail to get what I mean because this instance of ChatGPT cannot fall in love, does not have kidneys and is not scared of death? No.

We can consider the multiple and different but overlapping language games in which we say that a human (or perhaps even a dog) understands or fails to understand something, and then compare them with the new language games in which we (unavoidably, but often in scare quotes) say that an AI understands or fails to understand something. We'd find that these games have family resemblances to one another, enough to explain why we use the word "understand" in all of them. And we can be relaxed about the fact things which are important for its use in one of these games (having a brain! having free will! having kidneys!) are absent from its use in another.

Nothing to see here, move along please.

Yes, there are AI-shaped holes in organisations.

Matthew Clifford [says](#): “There are no AI-shaped holes lying around”. That is how he reconciles "the facts that (a) AI is already powerful and (b) it's having relatively little impact so far Making AI work today requires ripping up workflows and rebuilding *for* AI. This is hard and painful to do..."

Organisations look at AI and think surely we can make massive use of this either (on the good side) to do new things and solve hard problems for the benefit of all, and (on the bad side) simply to cut whole swathes of the workforce.

Beyond specific technical tasks, it can be daunting to identify where and how to apply AI effectively across an organisation. How would you rewire an entire department's functions for AI?

I think that's why we're going to see a trend to simply treat ordinary human job profiles as those AI-shaped holes. Thinking in terms of roles rather than tasks or functions.

Imagine a virtual department of human-sized AIs, each with memory, communication, and even 'personality,' operating within existing channels and hierarchies. Managing an organisation becomes simpler if we think in terms of virtual people in recognisable roles, rather than an opaque system of tasks. As agent-based AIs advance, maintaining 'explainability' is crucial.

You can imagine a person-sized AI at Company X emailing a corresponding AI at company Y, or a or example, a person-sized AI at Company X could email a counterpart at Company Y, or a human, about a specific issue. Externally, it's easier to engage with an organisation if you can address a particular role, regardless of whether it's filled by a human or an AI.

Whether this means a hard-pressed workforce getting rows and rows of additional workers to solve problems and meet needs more effectively or whether it means 90% of staff being made redundant and replaced by person-sized AIs is not yet clear though I fear it will be the latter.

To be clear I have no particular enthusiasm for this kind of development because I don't trust capitalism with this technology. But we still have to learn how to think about it and understand it and make use of it as best we can.