Getting philosophical

CONTENTS

! words - bodies

! words - perspective

AIs have perspective#

March 24, 2025

Do you feel uncomfortable saying "the AI misunderstood what I wanted"? Does that mean we think they are conscious?

Silva correctly draws attention to this and has some suggestions.

I think first we have to look at uses, not words. It isn't so useful to argue about generally whether some AI is intelligent or conscious or whatever. It's interesting to look at cases where we need some of these somewhat controversial words.

Often, when talking to each other about prompt engineering, we have to say things like:

> I realised it hadn't understood what I meant by 'repeat the first part'. It was still thinking of the previous instruction, which is fair enough, I guess. So I had to rewrite that part of the prompt. In this kind of case, we have to use words like 'understand' or 'think' (or, more often, 'misunderstand'). We can't get round that. You can put scare-quotes round the word, or like Silva put '-simila' on the end of it, but I'm not sure what you achieve.

The same goes for 'reasoning'. We often need to say to one another, when improving a prompt, 'ah, its reasoning was flawed here' or 'this newer model seems to do a lot more sophisticated reasoning before coming to a conclusion -- look at this part.'

I don't think I've ever, in contrast, had to use the word 'conscious' in any practical context when talking about AIs.

You could say, 'we shouldn't use the word understand about an AI because that implies they are conscious, and how can an AI be conscious'. But I'd say that these words mean no more and no less than how we use them in this context. I have no use (at the moment -- this might change) for saying an AI is conscious, or for that matter, not conscious.

Wittgenstein taught us to look at words in use. Beyond that is just words going on holiday: idle philosophising.

!700 Retraceable, reconstructable, operationalisable

[600! bricolage]

-
- the best criterion for rigour: replayability, retrospective reconstructability: even if it wouldn't have been obvious a priori how to answer this question, *afterwards* you can see how these conclusions were arrived at and can retrace the steps yourself.
- No method, not even RCTs, gives you a free pass to not do evaluative thinking. Because they're always embedded in the judgement of the evaluator that this is the right thing to do and I've checked all loopholes. In the same way our reasoning when doing bricolage is *evaluative* reasoning. The real-life application even of what seems like a monolithic tool like an RCT involves at different levels dozens of evaluative judgements about what to accept and what not to accept, whether this questionnaire or econometric measure is up to scratch, whether to trust this researcher, etc., whether this quantity of missing data is acceptable, etc.
- This appeal to evaluative reasoning is in some sense a transcendent one: you can never completely transfer responsibility to an algorithm. Because you always at least bear responsibility for using this particular algorithm in this particular way.
- At the same time algorithms can be incredibly helpful in breaking down our reasoning into transparent steps.
-
- Nachvollziehbarkeit also means it fits in your head
- Bricolage, rigour and *Nachvollziehbarkeit* in a partially evaluative summary of the Strategic Plans of 191 Red Cross / Red Crescent societies globally.

Steve Powell, Mon, May 20, 2024

- But what I find weaker in this paper is the just the generic weakness of the constructivist response to classic Cambellian validity:
- The bomb that's just dropped is suddenly we do have ways thanks to generative AI of making qualitative judgements (about qualitative constructions, about people's meanings, people's beliefs and understandings) which are in some sense reproducible and intersubjective and reliable. The methods aren't essentially any different from before, but they can be done quickly, reproducibly and at scale and in a way which vastly reduces noise due to the individual researcher. (Our chosen AI is of course also in some sense biassed, but it's the same bias each time, so we can compare results across timepoints and groups in a way which was never possible before: we can claim our methods are *reliable*).
- As a footnote here are two members of the extended family of the concept of validity:

!What is a system, really

In the context of M&E / MERL it mostly seems to mean: there is a causal network of factors which affect one another, and also new factors and connections may appear over time. That's fine, and causal mapping is well-placed to help, but that is not necessarily a system is it?

[Assessing systems change]

There is no hidden vocabulary

Our approach is minimalist -- we code only bare causation

It follows from this that we are sceptical about the idea of more sophisticated mid-range theory with blockers, enablers etc.

QCA is disappointing because it is frequentist about causation

TODO

QCA is disappointing because it thinks the world is a grid

TODO

[1c A minimalist approach to coding does not code absences]

Context and the transitivity trap

> From (Powell et al., 2024)

Transitivity is perhaps the single most important challenge for causal mapping. Consider the following example. If source P [pig farmer] states ‘I received cash grant compensation for pig diseases [G], so I had more cash [C]’, and source W [wheat farmer] states ‘I had more cash [C], so I bought more seeds [S]’, can we then deduce that pig diseases lead to more cash which leads to more seed (G à C àS), and therefore G à S (there is evidence for an indirect effect of G on S, i.e. that cash grants for pig diseases lead to people buying more seeds)?

The answer is of course that we cannot because the first part only makes sense for pig farmers, and the second part only makes sense for wheat farmers.

In general, from G --> C (in context P) and C --> S (in context W), we can only conclude that G --> S in the intersection of the contexts P and W. Correctly making inferences about indirect effects is the key benefit but also the key challenge for any approach which uses causal diagrams or maps, including quantitative approaches (Bollen, 1987).

References

Powell, Copestake, & Remnant (2024). *Causal Mapping for Evaluators*. https://doi.org/10.1177/13563890231196601.

Context, mechanisms and triggers 2

If termites cause a tree to fall in a forest where no-one can hear it, was it a causal event?

Maybe not.

The difference is the need for an explanation. That's the trigger. The trigger says why here and now if forest fires are happening all the time then we lose the sense of the trigger and the explanation could involve anything Oxygen or gravity or whatever. We don't have a question to answer so we don't know what factors to mention right the way out to the orbits of Pluto and beyond

Context, mechanisms and triggers

> What delineates a mechanism and its contents, the circle in the CMO diagram, is not so much facts about the world but the things which I happen to need to call upon to make a causal explanation

I found the concept of a trigger in realist evaluation totally baffling because RE is supposed to be somehow scientific, yet most forms of scientific explanation don't involve actual triggers.

But then I realised, triggers (and mechanisms) are best understood from an epistemic perspective.

"This mechanism gets triggered here" can be parsed as: there is a need for an explanation here: "I am invoking this mechanism to explain something that needs explaining".

We don't invoke oxygen to explain the forest fire, although both oxygen and the spark are (let's say) necessary.

What does a causal coding mean

The way we do causal mapping, a coded causal claim does not mean:

- "X occurred and then Y occurred"
- "Y frequently occurs after X"
- "X has the causal power to affect Y"

It means something like

- "X affected Y via its causal power to do so"

Counterfactuals are part of the meaning of causation but are not necessarily part of how we know about it

Just about everything is complex

TODO

> Putting a man on the moon was merely a complicated task, not a complex one?

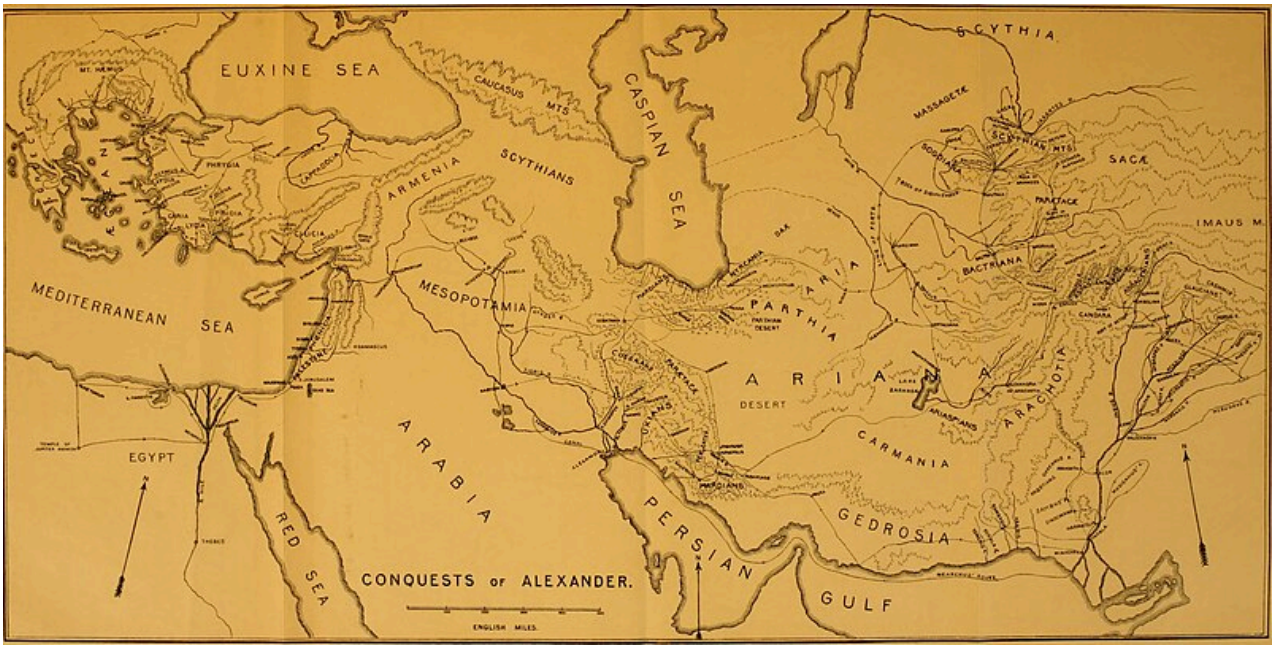
“Putting a man on the moon” is really often given as an example of a merely complicated but not complex task - Glouberman and Zimmerman (2002), cited in Rogers (2008).

But while that task was certainly very complicated, it was often complex too. There were plenty of conflicting sub-goals, arguments about how to solve a particular problem, interdependencies and conflicts between means, and between ends... You can *see* it as merely complicated, in order to make a point, and maybe that’s OK. But if you’re an actual space scientist you’d probably disagree.

There has always been complexity

Irene Ng speaks for many who write about "complex systems" when she says: "What has happened in the last 50 years is that we've been trying to use deterministic tools to achieve emergent outcomes, essentially because those are the only tools we have learnt (systems thinkers are still a minority unfortunately). We treat complex systems like complicated systems. We try to design, specify, impose, dictate when we should be designing, enabling, intervening, stablising."

Is there any historical truth in this at all? Did, say, a midwife 50 years ago only know how to impose and dictate rather than intervene and stablise? Was, say, managing the Mongol Empire, or Alexander's conquests, a merely complicated, not complex task?



An often-used example of a complex task is bringing up a child (and I'd agree, loosely). Well, did we have no children to bring up until our frightfully modern era?

Perhaps Irene Ng is writing about *our writing about* management, not how it is or was actually done. But there are ancient books like "The Art of War" about how to lead, and manage, and reach goals. Were they all merely guides to snapping together simple solutions? Of course not.