Some reflections on AI science

Felipe Morales Carbonell
Universidad de Chile

XXIII Jornadas Rolando Chuaqui
PUC, 2023

Outline

1) Clearing up some terms about the possibility of AI science
2) The issue of defining the minimal class of agents able to do science
3) An argument against the possibility of AI science
4) A way out? The knowledge machine
5) Catching crumbs from the table? A Suitsian response

Some clearing up

- AI will be treated here as a countable term. *An* AI is an artificial agent that possesses intelligence to some degree.
- Here I am not interested in the question of whether AIs are possible. I will simply assume they are, and I am not interested in debating the point (except in Q&A).

AGI and the lesser AIs

- AGI: Artificial General Intelligence. An AI that can perform well in a human-comparable range of tasks.
  - There is an issue of course about what range of tasks is relevant here, but we can bypass it here.
- So, by definition, for any humanly-implementable task, an AGI can perform it.
- Now, there are classes of AIs that have a more limited scope.
  - Presumably, most current AI systems belong to this class.

AI science

- One such task is science.
  - In reality, ‘science’ encompasses a wide range of tasks that are often interlocked, but which also can be independent (disciplinary and sub-disciplinary divides, heterogeneity of goals, etc.)
  - Here, I want to focus on the loop of finding a problem, devising hypotheses, constructing tests, and checking those hypotheses
- Is it possible for an AGI to do science? Yes, by definition.
- **Is it possible for weaker AIs? *How* weak can they be?**

A simple argument against the possibility of AI science

- Inspired on Haugeland (2017):

  1) Doing science requires know-how/understanding, *but*

  2) There cannot be artificial know-how, *hence*

  3) There cannot be AI science

I will now make some comments on why we may want to accept those premises, and then sketch some ways to reject them.

Scientific know-how

- Doing science is a complex set of procedures directed towards a number of goals (both epistemic and pragmatic).
- To achieve those goals, those who engage in science need to control their behaviour in specific ways, either dictated by the nature of those goals or by procedural conventions shared by the community involved in scientific practice.
- To engage in science effectively is not just a matter of grasping certain information about particular subject matters. It also requires knowing how to perform a number of actions.
- Scientific know-how is also a collective achievement, as it applies to *working together* towards scientific goals.

Gathering evidence, however, is not the only way of learning or finding out about the world that is essential to empirical science. ***Scientists must also, for instance, learn or find out how to make or perform the observations, measurements, and experiments that yield that factual evidence.*** *Yet if that is so, then that learning-how cannot itself be just more evidence gathering, on pain of regress.*

(Haugeland 2017, 296)

[...] normal science is research in which scientists know their way around. Professional training and research experience give scientists a reliable sense of what they are dealing with, what can affect its relevant behavior, how it makes itself known, and ***what they can do with it****. These abilities are held together by their practical grasp of one or more paradigms, concrete scientific achievements that point toward an open-ended domain of possible research. Paradigms should not be understood as beliefs (even tacit beliefs) agreed upon by community members, but instead as exemplary ways of conceptualizing and intervening in particular situations.* ***Accepting a paradigm is more like acquiring and using a set of skills than it is like understanding and believing a statement.***

(Rouse 2003, 107)

cf. Kuhn on tacit knowledge in the Postscript to *Structure*.

Know-how requires more than possible procedural success

- A lot depends on how we develop this idea of know-how.
- In a particular sense of ability, an AI *can* be reliably successful in performing a task.
  - *How* is the question. One is reminded of the Frame problem, which involves recognition of context-sensitive task-relevance.
- But this, some argue, is not sufficient. An AI needs to be *responsible* for its actions, or it needs to *care*.

The trouble with artificial intelligence is that computers don’t give a damn.

(Haugeland 1998, 47)

Softly pushing back

OK, but *does it matter*?

Consequentialism about science

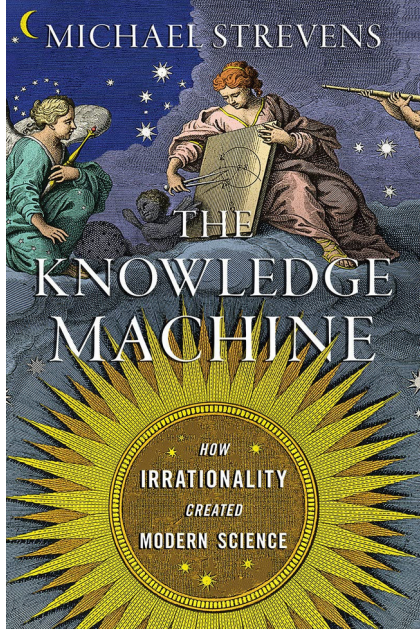
- We care about science, but do we care about doing science?
- *Consequentialism*: we care about the results of doing science, so any process that leads to those results is something that we could care about in the same way as we care for science.*
- Is there only one way to achieve those results?

* Maybe we will return to this point eventually.

The knowledge machine

- A suggestion from Strevens (2021): obeying the Iron Rule of Explanation is sufficient to achieve those goals.

1) Strive to settle all arguments by empirical testing.
2) To conduct an empirical test to decide between a pair of hypotheses, perform an experiment or measurement, one of whose possible outcomes can be explained by one hypothesis (and accompanying cohort) but not the other.



Grasp of the procedural structure of the Iron Rule

- The Iron Rule is a higher-order rule about how to proceed in the context of a community of inquirers.
- It moves the issue of the competences requires for engaging with science to the collective level.

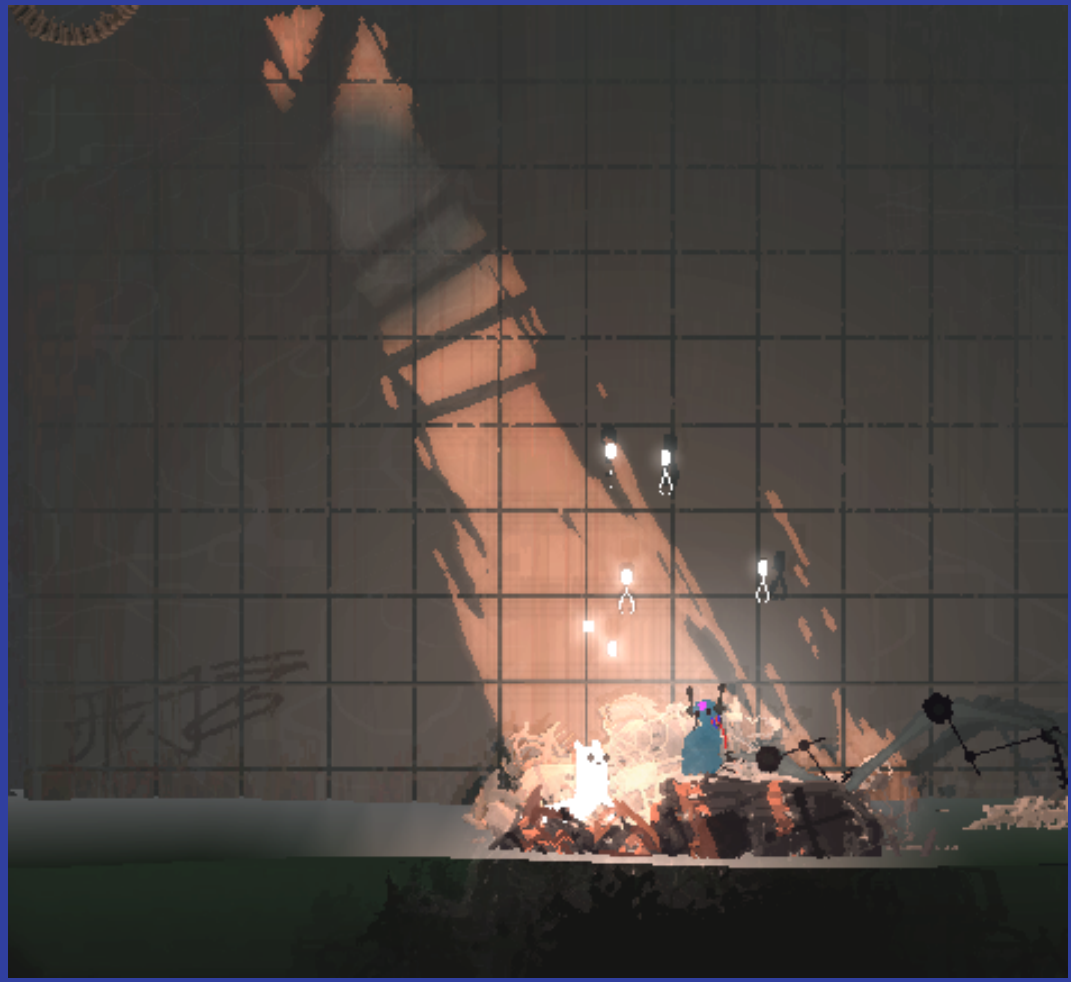
Mindless search as science enough

- From the consequentialist perspective I already sketched, science becomes a search problem: its practitioners search for solutions to problems.
- This can be automated.

This one is an old conversation between Five Pebbles and a friend of his. I'll read it to you.

"1591.290 - PRIVATE
Five Pebbles, Seven Red Suns

FP: Can I tell you something? Lately...
FP: I'm tired of trying and trying. And angry that they left us here. The anger makes me even less inclined to solve their puzzle for them. Why do we do this?
SRS: Yes, I'll spell this out - not because you're stupid or naive... Also, not saying that you're not ~
FP: Please, I'm coming to you for guidance.
SRS: Sorry, very sorry. I kid. Fact is, of course we are all aware of the evident futility of this Big Task. It's not said out loud but if you were better at reading between the lines there's nowhere you wouldn't see it. We're all frustrated.
FP: So why do we continue? We assemble work groups, we ponder, we iterate and try. Some of us die. It's not fair.
SRS: Because there's not any options. What else CAN we do? You're stuck in your can, and at any moment you have no more than two alternatives: Do nothing, or work like you're supposed to.
SRS: An analogy. You have a maze, and you have a handful of bugs. You put the bugs in the maze, and you leave. Given infinite time, one of the bugs WILL find a way out, if they just erratically try and try. This is why they called us Iterators.
FP: But we do die of old age.
SRS: Even more incentive! You know that nothing ever truly dies though, around and around it goes. Granted, our tools and resources get worse over time - but that is theoretically unproblematic, because in time even a miniscule chance will strike a positive. All the same to them, they're not around anymore!
FP: I struggle to accept being a bug."

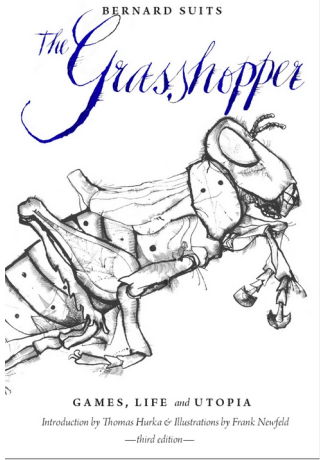


But what about questions?

- Science is also a search for problems.
- How to make an AI *look for* questions to address?
  - An AI can probably *generate* questions (again, *relevance* is the tricky part).
- Haugeland's point returns: how to create artificial curiosity?
  - Motivational states are necessary.

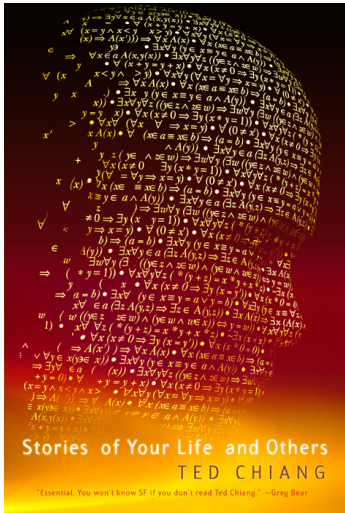
Against consequentialism

- We *do not just* care for the results of science.
- We enjoy solving puzzles, even if they are the wrong kind of puzzles (back to Kuhn).
- So I suspect there will never be a replacement of human science by AI science, just like there has not been a replacement of philosophy by science.
- We still play chess and go, even though AIs can outperform humans.
- Suits: the point of a game is that it makes a goal unnecessarily difficult.



Catching crumbs from the table?

- Humans may not contribute to the ‘hard goals’ of science (the pursuit of true general models of the world) anymore though.
- Ted Chiang’s ‘The Evolution of Human Science’ suggests that human science might turn into meta-human hermeneutics.
- But meta-human hermeneutics is (plausibly) a pseudo-science. Can that be avoided?



Thanks!

ef.em.carbonell@gmail.com
fmoralesc.github.io
okf@scholar.social