On Emergent Limits to Knowledge

Or, How to Trust the Robot Researchers: A Pocket Guide

Victoria Stodden
University of Southern California

Session 4: Enabling Open Science, Reproducibility, Replicability, and Privacy
US-UK Scientific Forum: Science in the Age of AI
11-12 June 2024

USC

Agenda

1. Science 1660
2. Science 2024:

   Effects of leveraging AI on scientific research and discovery

   Disruption in scientific norms: Transparency, Accountability, Reproducibility.
3. Science 2050: An emergent digital scholarly record

USC

A *Scientific* Fact

In *Opus Tertium* (1267) Roger Bacon distinguishes experimental science by:

1. verification of conclusions by direct experiment,
2. discovery of truths unreachable by other approaches,
3. investigation of the secrets of nature, opening us to a knowledge of past and future.

He described a now familiar **repeating cycle of observation, hypothesis, experimentation**, and the need for independent verification,

He recorded his experiments (e.g. the nature and cause of the rainbow) in enough detail to permit **reproducibility** by others.

USC

Inductive Scientific Reasoning

In *Novum Organum* (1620) Francis Bacon proposes:

1. the gathering of facts, by observation or experimentation,
2. verification of general principles.

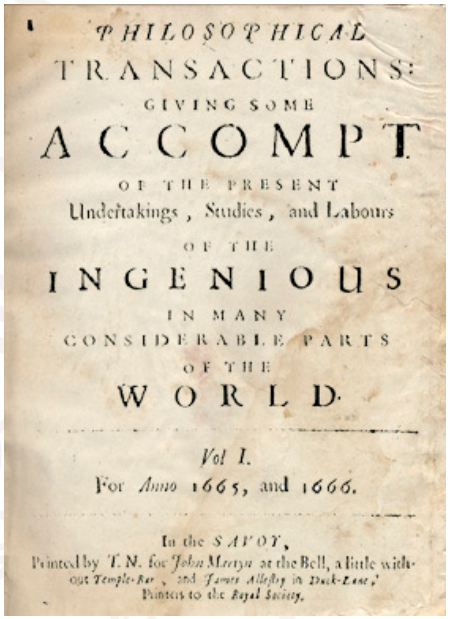
"There are and can be only two ways of searching into and discovering truth. The one flies from the senses and particulars to the most general axioms, and from these principles, the truth of which it takes for settled and immoveable. ... The other derives axioms from the senses and particulars, rising by a gradual and unbroken ascent, so that it arrives at the most general axioms last of all. This is the true way, but as yet untried."

USC

The Scientific Record: Touching the Spring of the Air

The Royal Society of London founded 1660 (the "Invisible College")

- members discussed Francis Bacon's "new science" from 1645,
- Society correspondence reviewed by the first Secretary, Henry Oldenburg, who became the founder, editor, author, and publisher of Philosophical Transactions 1665.


PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.
Vol I.
For Anno 1665, and 1666.
In the SAVOY,
Printed by T.N. for John Martyn at the Bell, a little without Temple-Bar, and James Allestry in Duck-Lane,
Printers to the Royal Society.

Boyle wrote, "**It is much more difficult than most men can imagine, to make an accurate Experiment**" in *Certain Physiological Essays And Other Tracts: Written at Distant Times, and on Several Occasions By the Honourable Robert Boyle* (1673)

In *New Experiments Physico-mechanicall, Touching the Spring of the Air and its Effects* (1660) Boyle set the standard for scientific (reproducible) communication:

1. Enough detail on equipment, material, and procedures, for reproducibility
2. "Communal witnessing"
3. Exhaustive details on experimental settings, false starts, failures, etc.

USC

Science 2024: An Unstoppable Force

Highly complex computation, data, and scientific workflows:

- Open and transparent re-executable machine learning pipelines,
- Open Data

Deeply disruptive innovation in scientific discovery:

- Common task framework – new discovery methods
- Leveraging LLMs
- Publishing in a pdf an afterthought
- ➢ Research shared natively digitally or not at all

Boyle's standards fail(ed). Result is a deep disruption in scientific norms to achieve: Transparency, Accountability, Reproducibility => Correctness.

USC

2050: A Science Odyssey

Say:

- LLMs leveraged for meta data completion for integrated datasets (Zisserman 2048),
- Benchmarking of a “clear and concise definition of a solution” (Hoult 2049),
- Black box pipelines find “best” solution => extraordinary discovery 2050.

Is the solution correct? Ask an LLM (Hoult 2024).

Assertion: We can’t tell since we don’t, and cannot, understand the chain of reasoning (even if we document and maximize open re-executable code/data/inputs).

Corollary:

A De Facto Digital Scholarly Record (for the robots)

A vision: In 2050 the scholarly record will be a detritus of organically connected frameworks, training data, and leaderboard results.

Correctness will be established not by transparency and human verification, but by direct checks in a new meta-Bacon methodology:

1. verification of conclusions by direct experiment **on black boxes**,
2. discovery of truths unreachable by other approaches,
3. investigation of the secrets **of black box discovery methods**, opening us to a knowledge of past and future.

We don't Touch the Spring of the Air, but instead generate a collection of useful results that work.

Idea: The chain of logic behind the discoveries is accessible via LLMs and impenetrable code (to humans) comprising yet-to-be-developed specialized scientific discovery pipeline querying tools.

USC

Change 1. Wither Authorship

In Since Boyle authorship has been an important part of accountability.

LLMs and other large model leveraged discoveries break the chain of authorship.

Who's the author?

Whose ideas does this work build on?

What ideas does the result build on?

It no longer matters whether the reasoning is human accessible. As long as we can query to obtain the results..

USC

Change 2. Knowledge as Utility not Understanding

2050: Researchers (and everyone else) believe results are correct since they trust their (opaque) discovery process:

- meaningful precision of the problem description,
- correctness of the benchmarks,
- appropriateness of the data.

A Scientific Method for Challenges

- What problems lend themselves best to this approach?
- What data attributes are essential for success?
- Who wins who loses?

USC

A Computable Digital Scholarly Record

Deliberate integration of computational knowledge to create a verifiable and extensible base in a systematic and open way, facilitating:

- **regeneration** of a computational results/models;
- **comparison**s and reconciliations of different hypotheses;
- the **reimplementation** of methods on new data and updating methods;
- the generation and evolution of **benchmarks and standardized testbeds** for the assessments of models and inference methods;
- the development and application of appropriate **policies** regarding data privacy, ethics, and meta-research on the scholarly record.

Such an entity acts as a traditional scholarly record:

- forms a locus for a research community to share ideas, get feedback, improve their work, agree on priorities, and resolve debates.

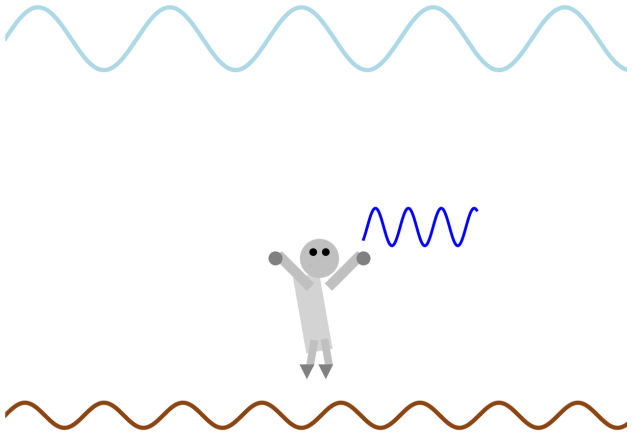
USC

Conclusions

Challenge problems and large opaque models will dominate scientific research.

Traditional ideas foundational to science break:

- Authorship and accountability,
- Human accessible chains of scientific reasoning,
- An emergent digital scholarly record.



Here's a surrealist-style drawing of a robot Touching a Spring in the Air. The robot is depicted with abstract and distorted features, set against a dreamlike landscape with flowing, surreal shapes. The spring maintains a twisted, helix-like form to add to the surreal ambiance.