USC University of Southern California

Verifying Correctness in AI-enabled Scientific Research: A New Frontier

Victoria Stodden
University of Southern California

HITS Colloquium
Heidelberg, Germany
March 24, 2025

USC

Agenda

1. **Where We Are**: The Open Research movement

2. **Everybody's Doing It, Why Can't We?** Leveraging (opaque and non-reproducible) AI in scientific research

3. **Disrupting** scientific norms: **Who wins? Who Loses?**

4. **Convergence**: The emergent *Digital Scholarly Record*

USC

0. My Background

Educational Experience

Ph.D. 2006. Statistics, Stanford University

Committee: Michael Saunders (Chair), David Donoho (Advisor), Jerry Friedman, Trevor Hastie, and Rob Tibshirani

Thesis title: "Model Selection When The Number of Variables Exceeds the Number of Observations"

M.L.S. 2007. Stanford Law School

M.S. 2000. Statistics, Stanford University

M.S. 1996. Economics, University of British Columbia

B.Soc.Sci. 1994. Economics (magna cum laude), University of Ottawa

1. Open Research in Academia

USC

A Transformation Driven by Shared Digital Research Objects

- Over the last several decades, we've seen a **global social experiment of data sharing** facilitated by immense mobile phone infrastructure.
- In academia, a meteoric rise in shared research data and code.
- Infrastructure development: innumerable institutional repositories, SaaS, platforms… e.g. Github, NIH GEO, NSF PAR, Kaggle, Colab…

Example:[1]

Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB

Table 1: **Pre-training data.** Data mixtures used for pre-training, for each subset we list the sampling proportion, number of epochs performed on the subset when training on 1.4T tokens, and disk size. The pre-training runs on 1T tokens have the same sampling proportion.

Observations:

- LLMs due, in part, to open data.
- Workflows now first class digital objects.
  - ⇒ Re-executable discovery pipelines for model estimation and prediction.

USC

1. Touvron et al., 2023, "LLaMA: Open and Efficient Foundation Language Models," https://arxiv.org/pdf/2302.13971

The Change Drivers

50 years of grass roots research community advocacy for digital research object sharing:[2]

- **Reports** (e.g. NASEM, Scientific Societies, …)
- **Policy Changes** (e.g. journal and conference reproducibility requirements)
- **Badging and Community Standards** (OSF, NISO, ACM, IEEE, ...)
- …

Researchers leverage any and all transformative tools that accelerate their work.

2. See Dewald WG, JG Thursby, and RG Anderson, “Replication in empirical economics: The Journal of Money, Credit and Banking project,” American Economic Review 76(4), **1986**; and
Claerbout J, Karrenbach M, “Electronic documents give reproducible research a new meaning.” In: Proceedings of the 62nd Annual International Meeting of the Society of Exploration Geophysics, **1992**; and
Buckheit J, Donoho DL, Antoniadis A, “Wavelab and reproducible research,” Wavelets and Statistics, New York, Springer, **1995**.

Example: Reproducibility Standards Development

Community Efforts: AAAS 2016 Workshop on Code and Modeling Reproducibility recommended:

REPRODUCIBLE RESEARCH

ADDRESSING THE NEED FOR DATA AND CODE SHARING IN COMPUTATIONAL SCIENCE

By the Yale Law School Roundtable on Data and Code Sharing

Roundtable participants identified ways of making computational research details readily available, which is a crucial step in addressing the current credibility crisis.

Progress in computational science is often hamp ers' inability reproduce or verif sults. Attendees at Yale Law School RoundtableNov212 a set of steps that s agencies, and journ improve the situati those steps here, alo for best practices available options term goals for the de tools and standards.

knowledge has long been scientific provide a long-term solution. We

Set the Default to "Open"

Reproducible Science in the Computer Age. Conventional wisdom sees computing as the "third leg" of science, complementing theory and experiment. That metaphor is outdated. Computing now pervades all of science. Massive computation is often required to reduce and analyze data; simulations are employed in fields as diverse as climate modeling and astrophysics. Unfortunately, scientific computing culture has not kept pace. Experimental researchers are taught early to keep notebooks or computer logs of every work detail: design, procedures, equipment, raw results, processing techniques, statistical methods of analysis, etc. In contrast, few computational experiments are performed with such care. Typically, there is no record of workflow, computer hardware and software configuration, or parameter settings. Often source code is lost. While crippling reproducibility of results, these practices ultimately impede the researcher's own productivity.

The State of Experimental and Computational Mathematics. Experimental mathematics[1]—application of high-performance computing technology to research questions in pure and applied mathematics, including automatic theorem proving—raises numerous issues of computational reproducibility.[2] It often pushes the bounds in very high precision computation (hundreds or

"It says it's sick of doing things like inventories and payrolls, and it wants to make some breakthroughs in astrophysics."

ScienceCartoonsPlus.com.

physicists, legal scholars, journal editors, and funding agency officials representing academia, government labs, industry research, and all points in between. While different types and degrees of reproducible research were discussed, an overwhelming majority argued that the community must move to "open research", research

INSIGHTS | POLICY FORUM

REPRODUCIBILITY

Enhancing reproducibility for computational methods

Data, code, and workflows should be available and cited

By **Victoria Stodden,[1] Marcia McNutt,[2] David H. Bailey,[3] Ewa Deelman,[4] Yolanda Gil,[5] Brooks Hanson,[6] Michael A. Heroux,[4] John P.A. Ioannidis,[7] Michela Taufer[8]**

Over the past two decades, computational methods have radically changed the ability of researchers from all areas of scholarship to process and analyze data and to simulate complex systems. But with these advances come challenges that are contributing to broader concerns over irreproducibility in the scholarly literature, among them the lack of transparency in disclosure of computational methods. Current reporting methods are often uneven, incomplete, and still evolving. We present a novel set of Reproducibility Enhancement Principles (REP) targeting disclosure challenges involving computation. These recommendations, which build upon more general

to understanding how computational results were derived and to reconciling any differences that might arise between independent replications (*4*). We thus focus on the ability to rerun the same computational steps on the same data the original authors used as a minimum dissemination standard (*5*, *6*), which includes workflow information that explains what raw data and intermediate results are input to which computations (*7*). Access to the data and code that underlie discoveries can also enable downstream scientific contributions, such as meta-analyses, reuse, and other efforts that include results from multiple studies.

RECOMMENDATIONS

Share data, software, workflows, and details of the computational environment that generate published findings in open trusted repositories. The minimal components that enable

Sufficient metadata should be provided for someone in the field to use the shared digital scholarly objects without resorting to contacting the original authors (i.e., http://bit.ly/2fVwJPH). Software metadata should include, at a minimum, the title, authors, version, language, license, Uniform Resource Identifier/DOI, software description (including purpose, inputs, outputs, dependencies), and execution requirements.

To enable credit for shared digital scholarly

1. Share data, software, workflows, and details of the computational environment that generate published findings in open trusted repositories.

2. Persistent links should appear in the published article and include a permanent identifier for data, code, and digital artifacts upon which the results depend.

3. To enable credit for shared digital scholarly objects, **citation** should be standard practice.

4. To facilitate reuse, adequately **document** digital scholarly artifacts.

5. Use **Open Licensing** when publishing digital scholarly objects.

6. Funding agencies should instigate new research programs and pilot studies.

7. Journals should conduct a **reproducibility check** as part of the publication process.

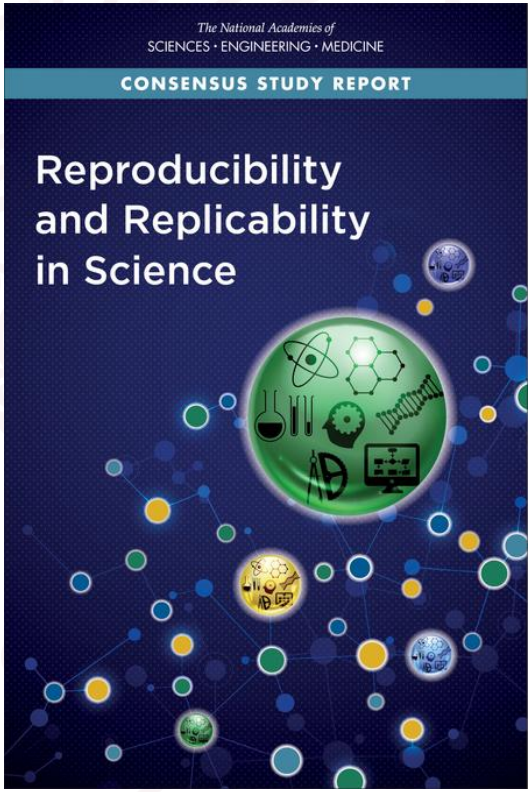
USC

Stodden, McNutt, Bailey, Deelman, Gil, Hanson, Heroux, Ioannidis, Taufer (2016). Enhancing Reproducibility for Computational Methods. Science.

Example: National Academies Consensus Report 2019

"Reproducibility and Replication in Science"

- 15 committee members (including myself)
- Chair: Harvey Fineberg,
  President of Gordon and Betty Moore Foundation
- Stakeholder input: over 50 individuals representing a range of disciplines
- Produced key definitions and several recommendations.



Report and white papers available at https://www.nap.edu/catalog/25303/reproducibility-and-replicability-in-science

Report Reproducibility Definitions (2019)

- ***Reproducibility*** is obtaining consistent results using the same input data, computational steps, methods, and code, and conditions of analysis. This definition is synonymous with "**computational reproducibility**."

- ***Replicability*** is obtaining **consistent results across studies** aimed at answering the same scientific question, each of which has obtained its own data. Two studies may be considered to have replicated if they obtain consistent results given the level of uncertainty inherent in the system under study.

AI: Reproducibility in Research

LLM training is only possible in industry due to the required scale of compute and data warehousing e.g:

- OpenAI's GPT4 model trained on 20,000 GPUs for 3 months in 2022;
- Google's Gemini Pro (Ultra) currently trained on 40,000 (80,000) TPU v4 chips.

$50 billion spent on GPUs by AI industry in 2023.[3]

Model training on this scale is not possible in academia.

Reproducibility in LLM model building is not possible, nor is transparency in output interpretation.

Yet LLMs are relied upon in scientific research at accelerating rates.[4]

3. B., Jin, "A Peter Thiel-Backed AI Startup, Cognition Labs, Seeks $2 Billion Valuation," WSJ, March 30, 2024. https://www.wsj.com/tech/ai/a-peter-thiel-backed-ai-startup-cognition-labs-seeks-2-billion-valuation-998fa39d

4. See e.g. Giglou et al., 2024, "LLMs4Synthesis: Leveraging Large Language Models for Scientific Synthesis," https://arxiv.org/abs/2409.18812

2. Leveraging AI and Digital Assets in Scientific Research and Discovery

USC

Science 2025: Unstoppable Forces

Highly complex computation, data, and integrated scientific workflows:

- Open, transparent, re-executable machine learning pipelines, shared on common infrastructure.[5]
- Open Data.

Deeply disruptive innovation in scientific discovery:

- Widely used radically different discovery methods:
  Common Task Framework: an "Olympics" of benchmarked competitions between machine learning models using common data and redeployment of complex scientific discovery workflows, often leveraged by the opaque results of LLMs (e.g. feature selection, data preprocessing).

→ **Research shared natively digitally or not at all**

→ **Publishing a pdf is an afterthought…**

5. For hundreds of examples see e.g. https://github.com/skrish13/ml-contests-conf

A Successful Machine Translation Research Program

In **1986** Charles Wayne, a DARPA program manager, reinvigorated the defunct machine translation program with:

1. A well-defined, *objective evaluation metric* applied by a neutral agent (NIST) on shared data sets, to protect against "glamour and deceit" from "mad inventors" and "untrustworthy engineers" (Pierce, 1969);
2. *Participants reveal their methods* to the neutral agent and to one another when the evaluation results are revealed, to ensure that "simple, clear, sure knowledge is gained" (Pierce, 1969).

Note: This program/approach resulted in the development of Siri..

Recap: The Common Task Framework [6]

Elements:

1. **A detailed evaluation plan**
   - Developed in consultation with researchers
   - Published as the *first step* in the project.
2. **Automatic evaluation software**
   - Written and maintained by NIST
   - Published at the start of the project.
3. **Shared data**
   - Training data is published at start of project.
   - Test data is withheld for periodic public evaluations.

6. Mark Liberman (2011) Lessons for Reproducible Science from DARPA's Progams in Human Language Technology, The Digitization of Science: Reproducibility and Interdisciplinary Knowledge Transfer, AAAS Annual Meeting 2011.

A Thought Experiment..

Say, in some future of shared executable workflows and data:

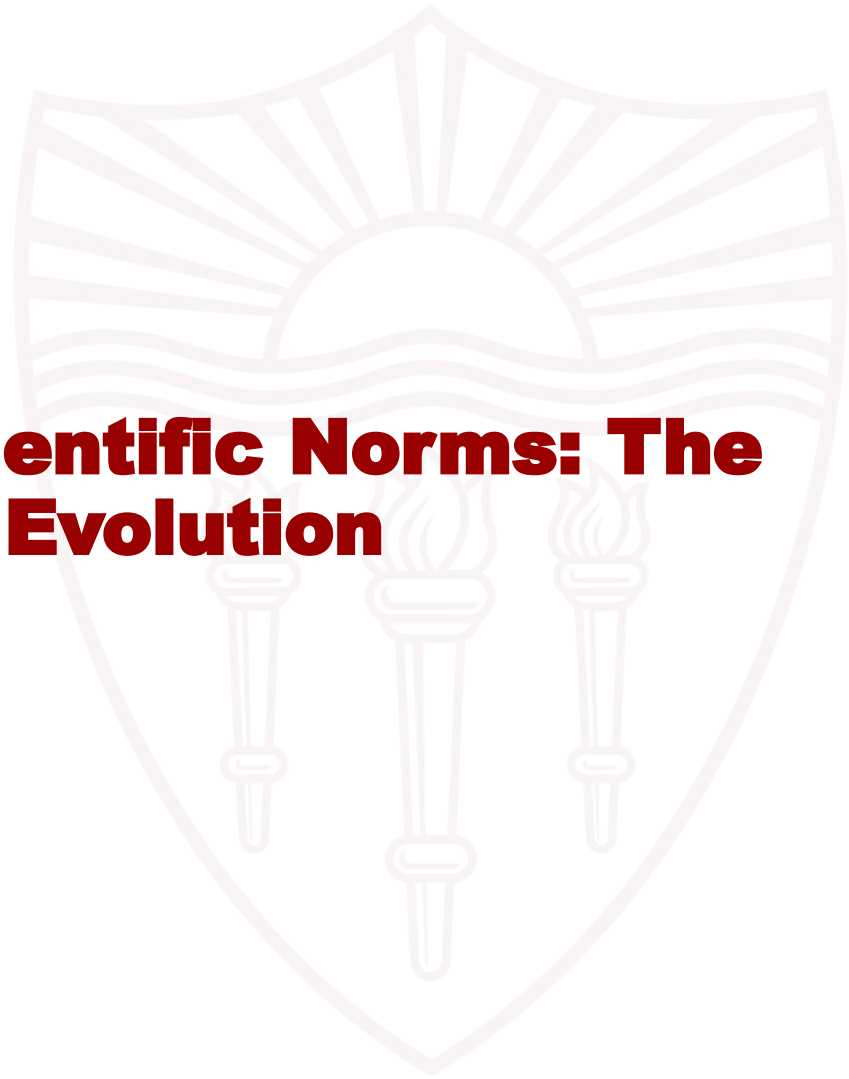
- LLMs are routinely leveraged for metadata completion for integrated and accessible datasets.
- Research has standard benchmarking of clear and concise solution definitions.
- Black box pipelines are routinely crawled to find “best” solutions.

- ***What problems do not lend themselves to benchmarking?***
- ***How to choose research problems?***
- ***Who wins who loses?***
- ***How can we judge correctness if it is impossible to understand the chain of reasoning that gave the results?***[7]

USC

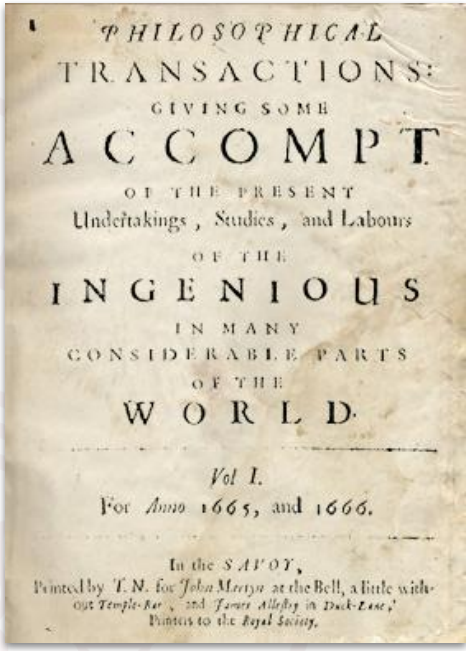
7. Stodden, V. (2024). On Emergent Limits to Knowledge—Or, How to Trust the Robot Researchers: A Pocket Guide. *Harvard Data Science Review*, *6*(1). https://doi.org/10.1162/99608f92.dcaa63bc

3. A Disruption in Scientific Norms: The History and Ongoing Evolution



USC

The Scientific Record: Touching the *Spring of the Air*


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.

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

- Members discussed Francis Bacon's "new science" of 1645,
- Society correspondence reviewed by the first Secretary, Henry Oldenburg, who became the founder, editor, and publisher of the first scientific journal in 1665; Philosophical Transactions.

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

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

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

USC

Standards for Scientific Publication Are a Function of Research Technology

1660's:

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

1900's

- Standards for journal publication: e.g. Introduction, Methods, Results, Discussion.

2025+: (Phase transition)

- Executable workflows with links to corresponding data and results.. Open accessibility.

USC

Change 1. Wither Authorship

Since Boyle, authorship has been key to 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..[8]

8. Stodden, V. (2024). On Emergent Limits to Knowledge—Or, How to Trust the Robot Researchers: A Pocket Guide. *Harvard Data Science Review*, *6*(1). https://doi.org/10.1162/99608f92.dcaa63bc

Change 2. Knowledge as Utility not Understanding

In an AI-enabled future, results are believed correct due to trust in the (opaque) discovery process.

Our task: Develop a Scientific Method for Challenges:
Is the quantitative goal of the challenge problem is sufficient to support scientific conclusions?

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

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

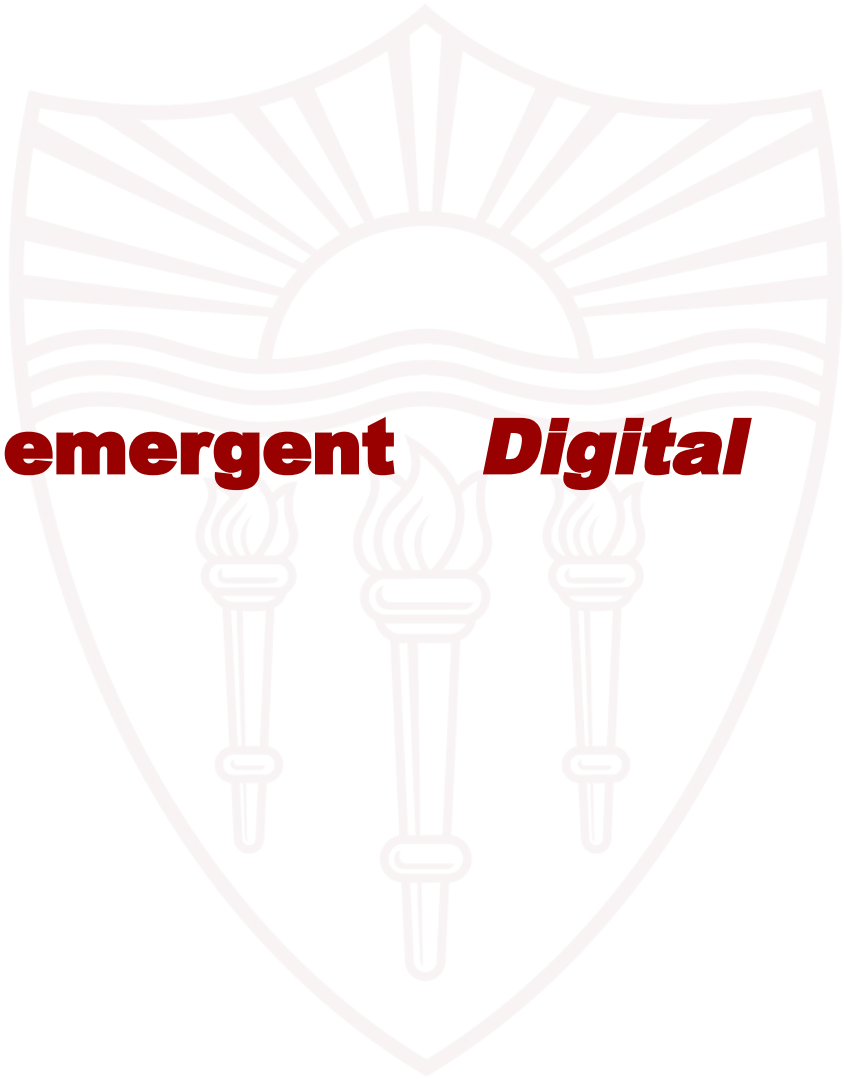
USC

Data Evaluation for Scientific Inference

Information about the data can increase trust in the result, for example:

- Measures of how well the data span the true underlying population of interest.
- Measures of sample representativeness can guide where we expect applications of the resulting model to be reasonable, in other words, trustworthy.
- Was the test set sampled from the input data or is it a new sample, likely with greater variability?

4. Convergence: The emergent *Digital Scholarly Record*



USC

The Digital Scholarly Record: Integration of Data, Code, Results

A verifiable and extensible database in a systematic and open way, facilitating:

1. Regeneration of computational results/models;
2. Comparisons and reconciliations of different hypotheses;
3. The reimplementation of methods on new data and updating of methods;
4. The generation and evolution of benchmarks and standardized testbeds for the assessments of models and inference methods;
5. Appropriate policies regarding data privacy, ethics, and meta-research on the scholarly record.

USC

Corollary: A De Facto Digital Scholarly Record (for the robots)

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 methodology:

1. verification of conclusions by direct experiment **on black boxes**,
2. 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.

A Computable Digital Scholarly Record II

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.

Integrates AI-enabled affordances:

- Document and trace contributions and authorship,
- Expose pipelines and (automated) testing.
- Automated discovery verification through replication.
- …

USC

The Emergent Digital Scholarly Record

Integrate computational knowledge for:

- a queriable and extensible open knowledge base,
- synthesizes computational and data-enabled discoveries,
- reusable discovery pipelines / regeneration of a computational results or models,
- reimplementation of methods on new data,
- comparison and reconciliation of different conclusions,
- the generation and evolution of benchmarks and standardized testbeds,
- development and application of appropriate policies regarding data privacy, ethics.

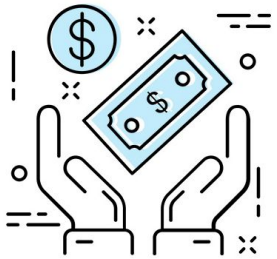
All perhaps in automated ways…

USC

No One Knows How and Nobody Cares

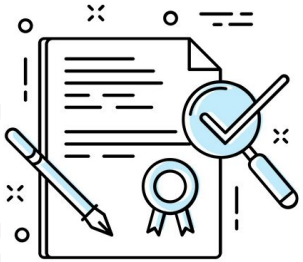
1. Models routinely contain billions of parameters, so a direct interpretation of the mechanism of response creation is far out of reach.
2. A focus on producing results that satisfy widely recognized benchmark performance goals, not a cognitively tractable explanation of the underlying phenomena.
3. The guardrails for knowledge production processes become incredibly important, as we can no longer solely rely on our usual mechanisms to assess research:
   - peer review,
   - disclosure of transparent methods for (human) verification,
   - the independent reproduction of findings.

Stakeholders

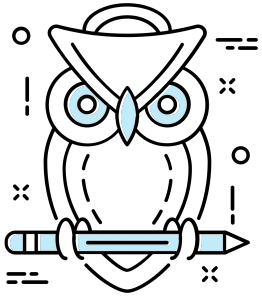


Funders

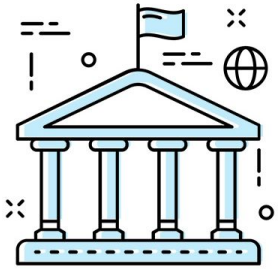
Researchers



Publishers

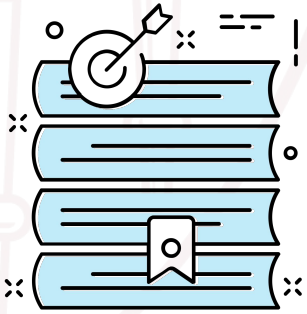


Scientific Societies



Regulatory Bodies

Institutions



Libraries

USC

Ongoing Research

January 21 Workshop at Triangel Studio KIT: student workshop where we investigate reproducible research pipelines at KIT.

Key questions:

1. What does the KIT environment enable well?
2. What changes or additions to the environment would accelerate research, if any?

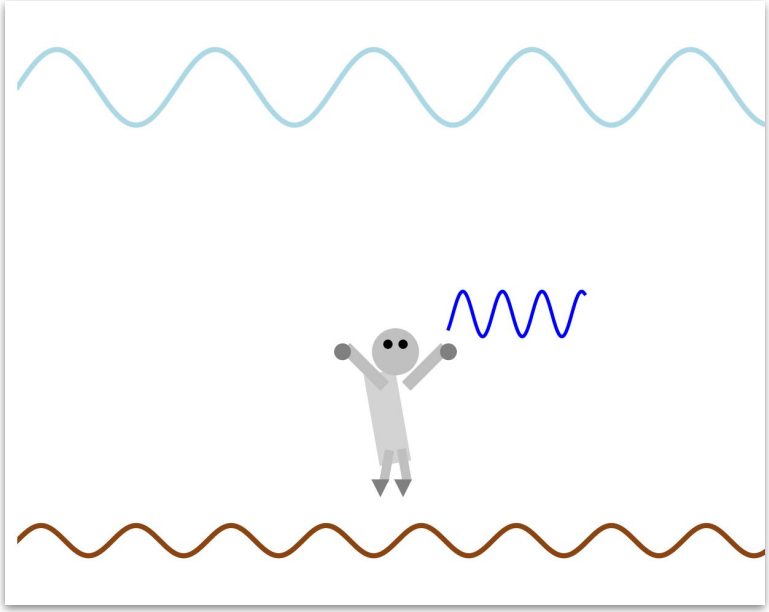
Background on the current state of reproducibility goals and how they can be institutionally dependent.

Conclusions

Challenge problems and large opaque models are coming to dominate scientific research, leveraging massive compute and data infrastructure.

Implications of leveraging AI in scientific discovery:

- *Authorship and accountability disrupted,*
- *Non-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." ChatGPT4

Our task:

Develop a Scientific Method for AI-enabled Research