USC University of Southern California

Automating Assessment of Machine Learning Research: Revisiting Arrow's Impossibility Theorem

Victoria Stodden
Department of Industrial and Systems Engineering
University of Southern California

Departmental Seminar
Alfred-Weber-Institut Für Wirtschaftswissenschaften
Heidelberg University, Germany
November 30, 2023

USC

Agenda

1. The Model Checking Value Proposition for Machine Learning
2. Leveraging Steps in the Community
3. Reproscreener: A New Tool for ML Model Checking (Preliminary Empirical Results)
4. Reconciling Results: Revisiting Arrow's Impossibility Theorem

Interlude: My Educational Background

Undergraduate degree in economics from University of Ottawa

Master's degree in economics from University of British Columbia

Ph.D. in statistics from Stanford University

Master's in Legal Studies from Stanford Law School

1. Model Checking Value Proposition: Potential Gains

Scalability to large-scale complex models

- e.g. integration of multiple large data sources, deployment on large scale / high throughput computing systems.

Verification of model performance

- Consistency of results / predictions across time, systems, data / Reproducibility

Transparency / interpretability / reproducibility / re-executability

Efficiency in resource use / Discovery speedup

- Computing systems: compute time, appropriate benchmarking;
- Engineers: code re-use, reduction in effort duplication

Model Checking Value Proposition: Potential Drawbacks

Increased overhead

- Additional computational step(s) in model building, deployment

Culture change

- Increased emphasis on reproducibility, verification, correctness in ML models

Training and standards development

Hewing to the wrong goals

2. Community Efforts

ML model publication standards

- Gunderson (AAAI 2018)
- Pineau (JMLR 2020)

Formal Verification for ML models

- Abate (MEMOCODE 2017)
- Urban and Miné (arxiv 2021)

Research Publication Standards

- Willis and Stodden (HDSR 2020)
- ML Commons (Github)
- National Academies Reproducibility Report (NASEM 2019)

Many many more…

ML Model Checking: A Novel Approach

- Previous work in AI involves applying formal techniques using SMT (satisfiability modulo theory) solvers, constraint solving, or abstract numerical interpretation.

- We exploit specialized features of ML pipelines and propose a *reproducibility* approach (NASEM 2019):
  - Exposure of methods
  - Well-defined guarantees in correctness of results

3. Automating ML Model Checking: Reproscreener

- Automate ML model checking *at the point of publication*, to provide guarantees on correctness, scalability, and transparency.

- Reproscreener software tool verifies criteria and provides feedback[1].

Available at reproscreener.org and https://github.com/Machine-Learning-Pipelines/reproscreener/

Criteria used by Reproscreener

1. Automatically test for ML model criteria (Gunderson 2018) at publication.
2. Automatically test for code/repo criteria (Krafczyk et al 2020).

Implemented in a labelled testbed of arXiv publications

- 50 most recent arXiv preprint submissions in stat.ML and/or CS.GL from October 25 2022.

Reproscreener Performance on Testbed Preprints (Gunderson 2018)

Metric	Proportion Correct (n=50)
Code available	0.82
Hypothesis stated	0.60
Experimental setup	0.54
Dataset available	0.48
Problem stated	0.36
Predicted result	0.30
Research method	0.28
Objective/Goal	0.28
Research question	0.16

USC

Reproscreener Performance Code/Repos (based on Krafczyk et al 2020)

Metric	Proportion Correct (n=22)
Readme has dependencies info	0.45
Readme has setup instructions	0.45
Readme has requirements info	0.41
Readme has install instructions	0.41
Wrapper scripts	0.36
Dependency tracking files	0.32

ChatGPT4 Performance on Abstracts

Metric	Proportion Correct (n=50)
Code available	1.00
Research question	0.96
Hypothesis stated	0.88
Dataset available	0.88
Objective/Goal	0.88
Problem stated	0.82
Predicted result	0.52
Research method	0.46
Experimental setup	0.46

USC

ChatGPT mistake examples

“Problem stated”
GPT’s found phrase: Upcoming large astronomical surveys are expected to capture an unprecedented number of strong gravitational lensing systems.
Manually found phrase: The absence of large quantities of representative data from current astronomical surveys motivates the development
GPT’s conclusion: The problem is stated in the abstract. FALSE
Notes: The Problem here is the fact that large amounts of data is *missing* from surveys and not that the surveys are expected to capture a large number of systems.

“Dataset available”
GPT’s found phrase: Our investigation on 59 different USB flash drives---belonging to 17 brands, including the top brands purchased on Amazon in mid-2019---reveals a minimum classification accuracy of 98.2% in the identification of both brand and model, accompanied by a negligible time and computational overhead.
GPT’s conclusion: Dataset is available. FALSE

Extending “Gunderson 2018” Criteria

ReproScreener Goals:

- Automatically check specific guidances to improve correctness of ML models to predict error bounds, capture and identifies difference in model output at scale (due to architecture, non-determinism, etc.)
- Enable comparison of model code through:
  - Checking for modularity, file structure, dependencies.
  - Checking for steps/scripts to create figures & visualizations.
  - Tracking model benchmarks and provenance.

- Real world case studies to demonstrate ReproScreener’s functionality

Does a study conform to accepted norms? How do studies build knowledge?

"But then two new problems arose. The ***first*** problem was the ***appraisal of conjectural knowledge***… The ***second*** problem was the ***growth of conjectural knowledge***.

In this situation *two schools of thought emerged*. One school - *neoclassical empiricism* - started with the first problem and never arrived at the second. The other school- *critical empiricism* - started by solving the second problem and went on to show that this solution solves the most important aspects of the first too."

Lakatos, "Changes in the Problem of Inductive Logic," in: Imre Lakatos ed., The Problem of Inductive Logic. Amsterdam: North-Holland 1968, pp. 315-417. p. 322.

4. An Apparent(?) Paradox of Automation

Problem: ML pipelines can conform to accepted standards (appraisal), yet yield conflicting results (growth).

Example: Consider the LLM queries example:

- A high degree of accuracy against curated “truth” standards
- No notion of “temporal coherence” across queries over time
- Need consistency across queries.

Recall: *Arrow’s Impossibility Theorem*: Individual preferences can be consistent (e.g. transitive), however an aggregated overall preference may not exist that is also transitive. Consider the following preferences for three states A, B, and C:

Individual 1: ABC
Individual 2: BCA
Individual 3: CAB

In the aggregate, using simple majority voting:
A>B and B>C, and C>A => a contradiction.

An Apparent(?) Paradox of Automation

Paradox: Any particular ML pipeline can conform to good (best) scientific practices, yet the body of research may not provide reliable findings.

Simple example:

30 2-class classifiers have been independently estimated and applied to the famous acute lymphoblastic leukemia dataset (Golub ‘99, Stodden ‘18).

The classifiers perform inconsistently on test and new data.

Models can yield contradictory results in the aggregate, and be consistent at the model level. **Can model checking reconcile?**

An Apparent(?) Paradox of Automation

Traditional Responses:

1. (Positivistic) No paradox, there is one truth and we have yet to discover it. Models should (eventually) agree on it
2. (Empirical) *Noise and measurement error exist in models/data* and their findings and we can only have agreement in findings up to accurately estimated confidence intervals.

New sources of error:

3. (Computational) Not only do noise and measurement error exist in models/data but also in *computational implementations* and re-executions of ML pipelines.

Change perspective to resolve Impossibility Paradox: estimate errors from computational sources as well as mathematical models and data.

Conclusion: The Model Checking Value Proposition Revisited

Reproscreener:

1. Can assist in automatically checking individual manuscripts and associated code for the satisfaction of relevant criteria,
2. New paradigm of automated integration of research findings.

Goal: Boundedness guarantees regarding correctness of reproduced results compared to original ML pipeline, resulting *coherent model comparisons, a reliable body of knowledge*.

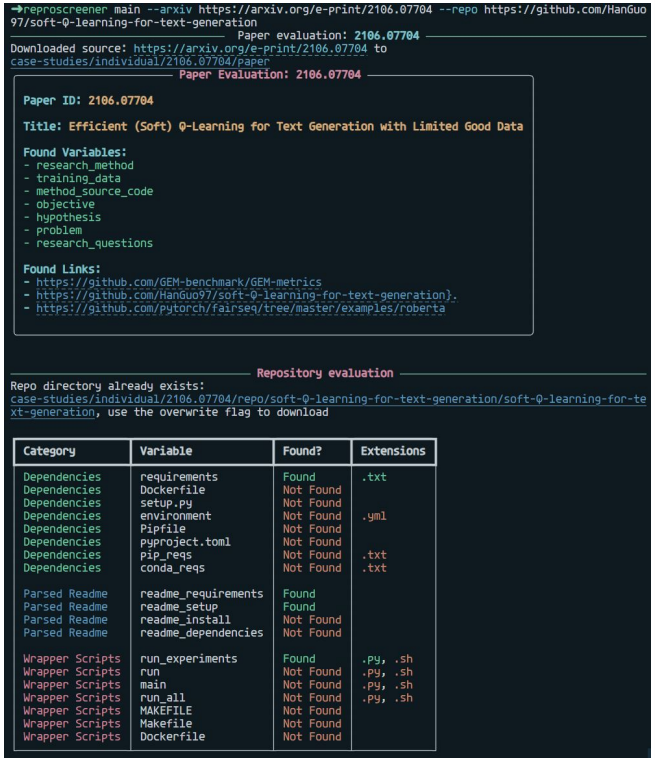
Thank you!

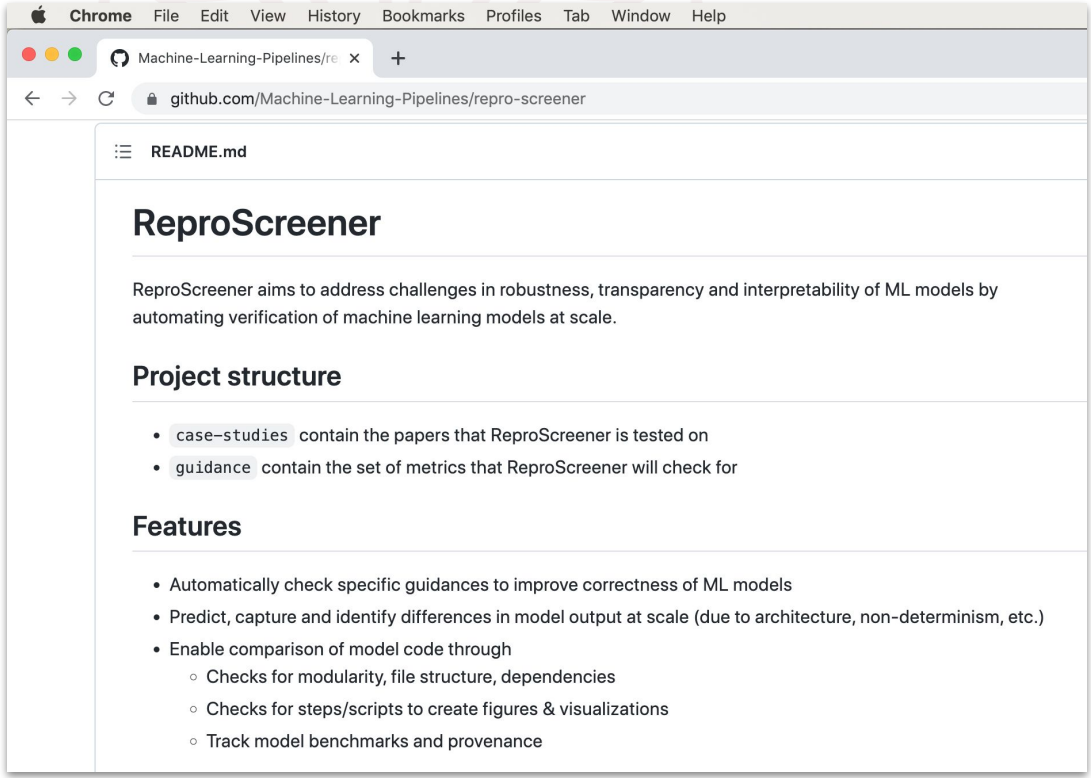
Joint work with **Adhithya Bhaskar**, Ph.D. student
Department of Industrial and Systems Engineering
University of Southern California

This material is based upon work supported by the REAL@USC-META Center and National Science Foundation Grant No 2138776

USC

Reproscreener Open Source Development (work in progress)





USC