Ethics Implications of Irreproducibility

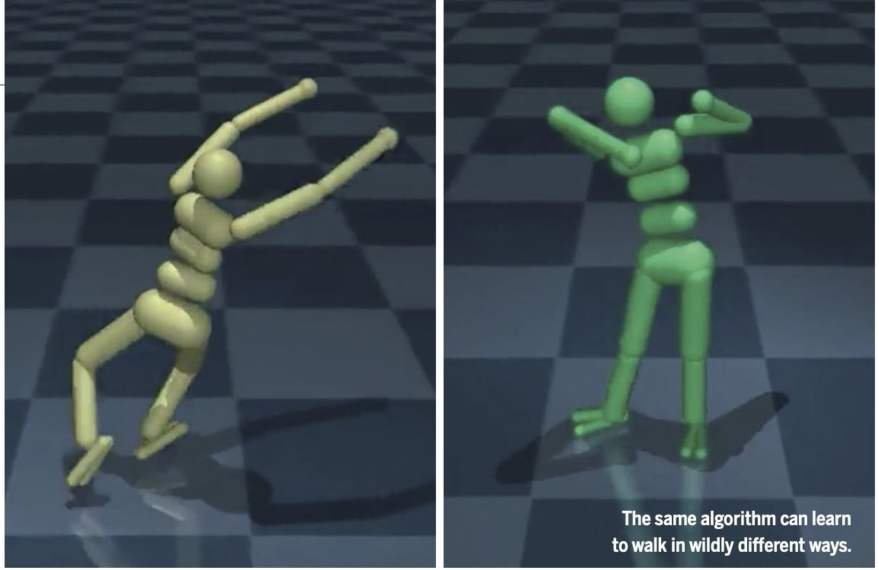
AGU - AI/ML Ethics Workshop Series - 2022

Kevin Coakley - SDSC/NTNU

Christine Kirkpatrick - SDSC

Overview

- Reproducibility and sources of irreproducibility
- Case study: reproducibility exploration using the Open Science Grid
- Implications of variance
- Takeaways


The same algorithm can learn to walk in wildly different ways.

COMPUTER SCIENCE

Artificial intelligence faces reproducibility crisis

Unpublished code and sensitivity to training conditions make many claims hard to verify

Hutson, Matthew. “Artificial Intelligence Faces Reproducibility Crisis.” *Science*, vol. 359, no. 6377, Feb. 2018, pp. 725–26. *DOI.org*

What is Reproducibility?

- Confusion between:
  - Repeatability (10x)
  - Replicability
  - Reproducibility

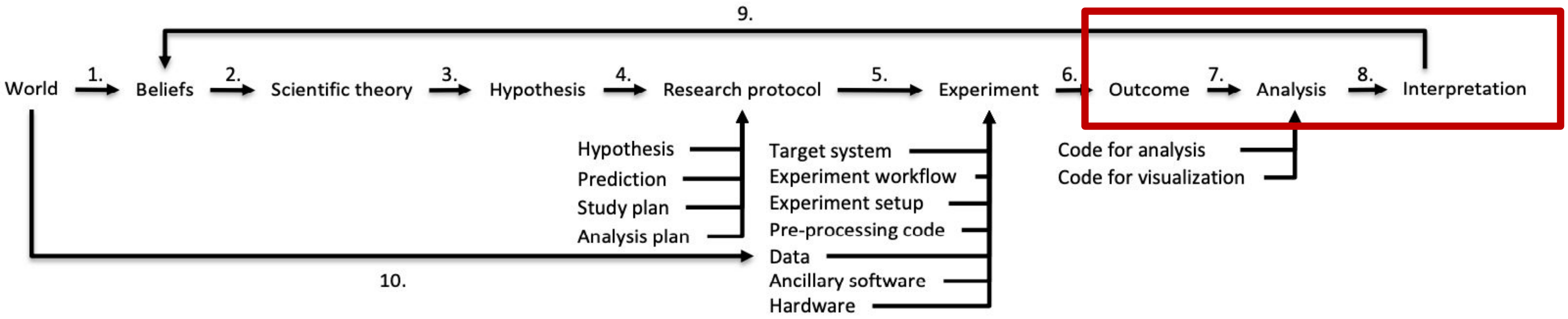
← repeat on same infra yourself / reproduce independently →

- Definitions can differ between scientific disciplines.

- Definitions can change over time as the literature evolves.

The Scientific Method

Perhaps a better way to think of reproducibility to examine the scientific method:



The scientific method as a ten step process: 1) observe the world to form beliefs about it; 2) explain causes and effects by forming a scientific theory; 3) formulate a genuine test of the theory; 4) design an experiment to test the theory; 5) implement the experiment; 6) conduct the experiment; 7) analyse the outcome; 8) interpret the analysis; 9) update beliefs according to the result; and 10) observe the world systematically.

Gundersen, Odd Erik. "The Fundamental Principles of Reproducibility." *ArXiv:2011.10098 [Cs]*, Nov. 2020. *arXiv.org*,

Sources of Irreproducibility

Implementation Factors

- Initialization seeds
- Ancillary software
- Ancillary software version
- Bugs in software
- Non-deterministic ordering of floating-point operations
- Parallel execution
- Compiler settings
- Processing unit

Gundersen, Odd Erik, Kevin Coakley, and Christine Kirkpatrick. "Sources of Irreproducibility in Machine Learning: A Review." *arXiv preprint arXiv:2204.07610* (2022)

What Can Be Done?

Not all sources of irreproducibility can be controlled for.

However, many of the sources of irreproducibility can be examined by performing multiple runs in heterogeneous computing environments consisting of different hardware and software environments.

The biggest challenge for researchers can be getting access to multiple heterogeneous computing environments.

Sources of Irreproducibility

Implementation Factors

- Initialization seeds
- Ancillary software
- Ancillary software version
- Bugs in software
- Non-deterministic ordering of floating-point operations
- Parallel execution
- Compiler settings
- Processing unit

Gundersen, Odd Erik, Kevin Coakley, and Christine Kirkpatrick. "Sources of Irreproducibility in Machine Learning: A Review." *arXiv preprint arXiv:2204.07610* (2022)

Open Science Grid

- Distributed High-Throughput Computing
  - For problems that can be run as numerous and self-contained jobs
- Supports Machine Learning and AI executed with Multiple Independent Training Tasks, Different Parameters, and/or Data Subsets
- More than 20 Institutions Participating
- Uses HTCondor Batch Scheduler
  - Run Non-Interactively
  - Uses Bash Scripts to Setup Environment and Copy Output
- Free to qualifying researchers: https://www.osgconnect.net

Keras Examples Used

Computer Vision

- Simple MNIST convnet: https://github.com/keras-team/keras-io/blob/master/examples/vision/mnist_convnet.py

Natural Language Processing

- Bidirectional LSTM on IMDB: https://github.com/keras-team/keras-io/blob/master/examples/nlp/bidirectional_lstm_imdb.py

Structured Data

- Imbalanced classification: credit card fraud detection: https://github.com/keras-team/keras-io/blob/master/examples/structured_data/imbalanced_classification.py

Examples were modified to run deterministically

https://www.tensorflow.org/api_docs/python/tf/config/experimental/enable_op_determinism

Varying Hardware Test

Using Open Science Grid Servers

The job was configured with:

- Various CPUs (4 threads)
- Various GPUs (1 GPU)
- Singularity containers based on Docker containers:
  - tensorflow/tensorflow:2.9.1-gpu (via TensorFlow)
  - nvcr.io/nvidia/tensorflow:22.06-tf2-py3 (via Nvidia)

The placement of what hardware the jobs ran on was not controlled.

Varying Hardware Test

30 tests ran on 2 different sites and 8 different servers.

Hardware included Intel Xeon and AMD EPYC CPU and NVidia A40 & A100 GPUs.

The same 3 Karas examples were used and each example was ran 5 times to ensure repeatability. The accuracy from the imbalance example and the mnist example were repeatable on the same hardware. The following heat map shows the accuracy of these two examples:

Varying Hardware Test

imbalanced_classification and mnist_convnet set_random_seed GPU Runs

Server	imbalanced_classification	mnist_convnet
GP-ARGO-langston:NVIDIA A100-40GB:tensorflow_22.06	0.9880971312522888	0.9918000102043152
GP-ARGO-usd:NVIDIA A100-40GB:tensorflow_22.06	0.9880971312522888	0.9918000102043152
GP-ARGO-oru:NVIDIA A100-40GB:tensorflow_22.06	0.9880971312522888	0.9918000102043152
GP-ARGO-langston:NVIDIA A100-40GB:tensorflow_22.06	0.9880971312522888	0.9918000102043152
GP-ARGO-sdsu:NVIDIA A100-40GB:tensorflow_22.06	0.9880971312522888	0.9918000102043152
GP-ARGO-ku:NVIDIA A100-40GB:tensorflow_22.06	0.9880971312522888	0.9918000102043152
GP-ARGO-uark:NVIDIA A100-40GB:tensorflow_22.06	0.9880971312522888	0.9918000102043152
Rice-RAPID:NVIDIA A40:tensorflow_22.06	0.9949439167976379	0.9918000102043152
Rice-RAPID:NVIDIA A40:tensorflow_22.06	0.9949439167976379	0.9918000102043152
Rice-RAPID:NVIDIA A40:tensorflow_22.06	0.9949439167976379	0.9918000102043152
Rice-RAPID:NVIDIA A40:tensorflow_2.9.1-gpu	0.9906251430511475	0.9918000102043152
Rice-RAPID:NVIDIA A40:tensorflow_2.9.1-gpu	0.9906251430511475	0.9918000102043152
Rice-RAPID:NVIDIA A40:tensorflow_2.9.1-gpu	0.9906251430511475	0.9918000102043152
Rice-RAPID:NVIDIA A40:tensorflow_2.9.1-gpu	0.9906251430511475	0.9918000102043152
Rice-RAPID:NVIDIA A40:tensorflow_2.9.1-gpu	0.9906251430511475	0.9918000102043152
Rice-RAPID:NVIDIA A40:tensorflow_2.9.1-gpu	0.9906251430511475	0.9918000102043152
Rice-RAPID:NVIDIA A40:tensorflow_2.9.1-gpu	0.9906251430511475	0.9918000102043152
Rice-RAPID:NVIDIA A40:tensorflow_2.9.1-gpu	0.9906251430511475	0.9918000102043152
Rice-RAPID:NVIDIA A40:tensorflow_2.9.1-gpu	0.9906251430511475	0.9918000102043152
Rice-RAPID:NVIDIA A40:tensorflow_2.9.1-gpu	0.9906251430511475	0.9918000102043152
Rice-RAPID:NVIDIA A40:tensorflow_2.9.1-gpu	0.9906251430511475	0.9918000102043152
Rice-RAPID:NVIDIA A40:tensorflow_2.9.1-gpu	0.9906251430511475	0.9918000102043152
Rice-RAPID:NVIDIA A40:tensorflow_2.9.1-gpu	0.9906251430511475	0.9918000102043152
Rice-RAPID:NVIDIA A40:tensorflow_2.9.1-gpu	0.9906251430511475	0.9918000102043152
Rice-RAPID:NVIDIA A40:tensorflow_2.9.1-gpu	0.9906251430511475	0.9918000102043152
GP-ARGO-oru:NVIDIA A100-40GB:tensorflow_2.9.1-gpu	0.9922052025794983	0.9922000169754028
GP-ARGO-langston:NVIDIA A100-40GB:tensorflow_2.9.1-gpu	0.9922052025794983	0.9922000169754028
GP-ARGO-usd:NVIDIA A100-40GB:tensorflow_2.9.1-gpu	0.9922052025794983	0.9922000169754028
GP-ARGO-ku:NVIDIA A100-40GB:tensorflow_2.9.1-gpu	0.9922052025794983	0.9922000169754028
GP-ARGO-sdsu:NVIDIA A100-40GB:tensorflow_2.9.1-gpu	0.9922052025794983	0.9922000169754028

Varying Hardware Test

The bidirectional Keras test didn’t have repeatable results when trained on GPU hardware with the tensorflow/tensorflow:2.9.1-gpu container.

Determinism on GPU hardware can be difficult. The code and the framework can support determinism but if the ancillary software doesn’t support it then the results will include randomness.

The following is a heat map of all 5 runs of all 30 test runs:

Varying Hardware Test

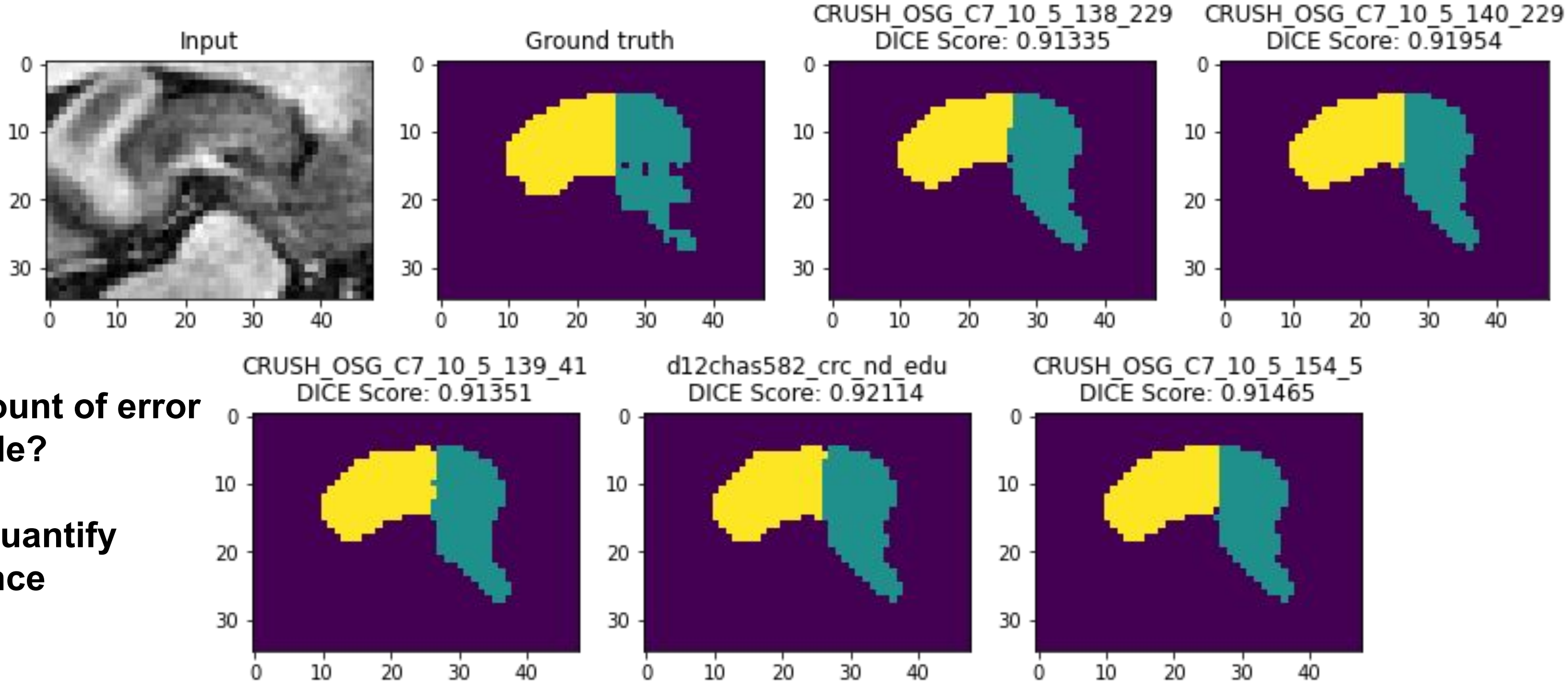
Bidirectional LSTM on IMDB set random seed on Different Hardware

Server	0	1	2	3	4
GP-ARGO-langston:NVIDIA A100-40GB:tensorflow_22.06	0.8658000230789185	0.8658000230789185	0.8658000230789185	0.8658000230789185	0.8658000230789185
GP-ARGO-usd:NVIDIA A100-40GB:tensorflow_22.06	0.8658000230789185	0.8658000230789185	0.8658000230789185	0.8658000230789185	0.8658000230789185
GP-ARGO-oru:NVIDIA A100-40GB:tensorflow_22.06	0.8658000230789185	0.8658000230789185	0.8658000230789185	0.8658000230789185	0.8658000230789185
GP-ARGO-langston:NVIDIA A100-40GB:tensorflow_22.06	0.8658000230789185	0.8658000230789185	0.8658000230789185	0.8658000230789185	0.8658000230789185
GP-ARGO-sdsu:NVIDIA A100-40GB:tensorflow_22.06	0.8658000230789185	0.8658000230789185	0.8658000230789185	0.8658000230789185	0.8658000230789185
GP-ARGO-ku:NVIDIA A100-40GB:tensorflow_22.06	0.8658000230789185	0.8658000230789185	0.8658000230789185	0.8658000230789185	0.8658000230789185
GP-ARGO-uark:NVIDIA A100-40GB:tensorflow_22.06	0.8658000230789185	0.8658000230789185	0.8658000230789185	0.8658000230789185	0.8658000230789185
Rice-RAPID:NVIDIA A40:tensorflow_22.06	0.8658400177955627	0.8658400177955627	0.8658400177955627	0.8658400177955627	0.8658400177955627
Rice-RAPID:NVIDIA A40:tensorflow_22.06	0.8658400177955627	0.8658400177955627	0.8658400177955627	0.8658400177955627	0.8658400177955627
Rice-RAPID:NVIDIA A40:tensorflow_22.06	0.8658400177955627	0.8658400177955627	0.8658400177955627	0.8658400177955627	0.8658400177955627
Rice-RAPID:NVIDIA A40:tensorflow_2.9.1-gpu	0.8675600290298462	0.8675600290298462	0.8675600290298462	0.8675600290298462	0.8675600290298462
Rice-RAPID:NVIDIA A40:tensorflow_2.9.1-gpu	0.8613200187683105	0.8626000285148621	0.8573200106620789	0.8610799908638000	0.8649200201034546
Rice-RAPID:NVIDIA A40:tensorflow_2.9.1-gpu	0.8628799915313721	0.8525599837303162	0.8678799867630005	0.8515599966049194	0.8679999709129333
Rice-RAPID:NVIDIA A40:tensorflow_2.9.1-gpu	0.8495200276374817	0.8572800159454346	0.8664399981498718	0.8625199794769287	0.8621199727058411
Rice-RAPID:NVIDIA A40:tensorflow_2.9.1-gpu	0.8602399826049805	0.8642799854278564	0.8643199801445007	0.8637199997901917	0.8629199862480164
Rice-RAPID:NVIDIA A40:tensorflow_2.9.1-gpu	0.8610799908638000	0.8507199883460999	0.8378000259399414	0.8429200053215027	0.8574000000953674
Rice-RAPID:NVIDIA A40:tensorflow_2.9.1-gpu	0.8686000108718872	0.8612400293350220	0.8637199997901917	0.8675600290298462	0.8678399920463562
Rice-RAPID:NVIDIA A40:tensorflow_2.9.1-gpu	0.8638799786567688	0.8600000143051147	0.8575999736785889	0.8620799779891968	0.8352400064468384
Rice-RAPID:NVIDIA A40:tensorflow_2.9.1-gpu	0.8675600290298462	0.8675600290298462	0.8675600290298462	0.8675600290298462	0.8675600290298462
Rice-RAPID:NVIDIA A40:tensorflow_2.9.1-gpu	0.8586000204086304	0.8561199903488159	0.8648800253868103	0.8633199930191040	0.8621199727058411
Rice-RAPID:NVIDIA A40:tensorflow_2.9.1-gpu	0.8675600290298462	0.8675600290298462	0.8675600290298462	0.8675600290298462	0.8675600290298462
Rice-RAPID:NVIDIA A40:tensorflow_2.9.1-gpu	0.8363199830055237	0.8628799915313721	0.8389199972152710	0.8572400212287903	0.8527600169181824
Rice-RAPID:NVIDIA A40:tensorflow_2.9.1-gpu	0.8579599857330322	0.8658800125122070	0.8632799983024597	0.8652399778366089	0.8438400030136108
Rice-RAPID:NVIDIA A40:tensorflow_2.9.1-gpu	0.8641600012779236	0.8557199835777283	0.8619599938392639	0.8493199944496155	0.8092399835586548
Rice-RAPID:NVIDIA A40:tensorflow_2.9.1-gpu	0.8685200214385986	0.8543199896812439	0.8555999994277954	0.8573200106620789	0.8597999811172485
GP-ARGO-oru:NVIDIA A100-40GB:tensorflow_2.9.1-gpu	0.8597999811172485	0.8661999702453613	0.8315600156784058	0.8641200065612793	0.8456400036811829
GP-ARGO-langston:NVIDIA A100-40GB:tensorflow_2.9.1-gpu	0.8521999716758728	0.8642399907112122	0.8579599857330322	0.8552399873733521	0.8626800179481506
GP-ARGO-usd:NVIDIA A100-40GB:tensorflow_2.9.1-gpu	0.8619999885559082	0.8166800141334534	0.8648800253868103	0.8613200187683105	0.8586800098419189
GP-ARGO-ku:NVIDIA A100-40GB:tensorflow_2.9.1-gpu	0.8680800199508667	0.8500400185585022	0.8637599945068359	0.8516399860382080	0.8610399961471558
GP-ARGO-sdsu:NVIDIA A100-40GB:tensorflow_2.9.1-gpu	0.8632799983024597	0.8689600229263306	0.8511599898338318	0.8640800118446350	0.8553599715232849

Run Number

Tolerance for Erroneous Conclusions: Low

Running nnUNet with Medical Imaging



What amount of error is tolerable?

Need to quantify the variance

Tolerance for Erroneous Conclusions: High

Image classification

Find the four-leafed clover →

0187683105	0.8626000285148621	0.8573200106620789	0.8610799908638000	0.864920020
9915313721	0.8525599837303162	0.8678799867630005	0.8515599966049194	0.867999970
0276374817	0.8572800159454346	0.8664399981498718	0.8625199794769287	0.86211997
9826049805	0.8642799854278564	0.8643199801445007	0.8637199997901917	0.86291998
9908638000	0.8507199883460999	0.8378000259399414	0.8429200053215027	0.857400000
0108718872	0.8612400293350220	0.8637199997901917	0.8675600290298462	0.86783999
9786567688	0.8600000143051147	0.8575999736785889	0.8620799779891968	0.83524000
0290298462	0.8675600290298462	0.8675600290298462	0.8675600290298462	0.86756002
0204086304	0.8561199903488159	0.8648800253868103	0.8633199930191040	0.86211997
0290298462	0.8675600290298462	0.8675600290298462	0.8675600290298462	0.86756002
9830055237	0.8628799915313721	0.8389199972152710	0.8572400212287903	0.85276001
9857330322	0.8658800125122070	0.8632799983024597	0.8652399778366089	0.84384000
0012779236	0.8557199835777283	0.8619599938392639	0.8493199944496155	0.80923998
0214385986	0.8543199896812439	0.8555999994277954	0.8573200106620789	0.85979998
9811172485	0.8661999702453613	0.8315600156784058	0.8641200065612793	0.84564000
9716758728	0.8642399907112122	0.8579599857330322	0.8552399873733521	0.86268001
9885559082	0.8166800141334534	0.8648800253868103	0.8613200187683105	0.85868000
0199508667	0.8500400185585022	0.8637599945068359	0.8516399860382080	0.86103999
9983024597	0.8689600229263306	0.8511599898338318	0.8640800118446350	0.85535997

Experimentation Conclusions

- There are sources of irreproducibility that cannot be controlled for.
- Running the experiment in multiple heterogeneous environments during the analysis stage will help validate the conclusions and also help validate the conclusions will be reproducible.
- Having resources like the Open Science Grid and access to Cloud resources can be used to improve reproducibility.

Takeaways

- AI/ML under constant development: new tools, new implementations of algorithms, new versions up and down the stack (operating system, framework, processors)
    - Consider usage carefully for application with impact
- Commitment to repeatability needed with AI/ML
    - Kevin repeats each experiment 10x
- For reproducibility, must document all applicable ‘implementation factors’
    - Publishers should consider guidelines for AI/ML driven work
    - Nanopublications to cite for replicable documentation
- Awareness building needed
    - Especially when data with embedded bias is used in ML

Contact

Kevin Coakley - kcoakley@sdsc.edu

Christine Kirkpatrick - christine@sdsc.edu

Sources of Irreproducibility in Machine Learning: A Review

https://arxiv.org/abs/2204.07610