Beyond Methods Reproducibility in Machine Learning

Leif Hancox-Li
Capital One
New York, NY 10011
leif.hancox-li@capitalone.com

Abstract

Reproducibility has become more of a concern in the machine learning community, as indicated by the implementation of reproducibility checklists at some major conferences. I argue that the criteria used in these checklists are insufficient for many ML research projects. I categorize different types of claims that ML experiments may support, and argue that reproducibility standards should differ for these different types of claims.

1 Introduction

As new research in machine learning (ML) continues to grow at a prodigious rate, ML conferences have started to pay more attention to the problem of reproducibility [Raff, 2019, Dodge et al., 2019, NeurIPS, 2020]. Reproducibility, as a signal of the credibility and scope of ML results, is epistemically and socially important for several reasons.

Firstly, much ML research receives public funds and is therefore subject to public accountability–it would be harder to justify public funding of a discipline that failed to produce epistemically reliable results. Secondly, ML research papers are often reported on in the media and subject to public interpretation, and these interpretations have effects on society. ML research that fails to meet particular epistemic standards can thus have negative effects related to misinformation and other harmful applications. Thirdly, the results of ML research are often used by an array of public and private actors, which could lead to negative consequences if the actors misunderstand the scope and reliability of the methods they use. Many recent ethical controversies around ML have been intertwined with the scientific and epistemic integrity of certain ML algorithms or results [Coalition for Critical Technology, 2020, Wang and Kosinski, 2018, Wu and Zhang, 2016].

In this paper, I draw distinctions between different types of claims that ML research based on computational experiments might be making, rendering explicit what is often left implicit. Then, I argue that those different types of claims demand different reproducibility standards. This is closely related to arguments from the philosophy of science that reproducibility should not be pursued as an overarching epistemic value regardless of context [Leonelli, 2018].

Roughly speaking, reproducibility in an epistemically meaningful sense requires that the results still hold when we tweak conditions in the experiment that ought to be irrelevant to the claim the paper is making. Depending on how robust or general a claim the paper is making, different sets of conditions ought to be tweaked. In some cases, we may not require much more than being able to get the same results when we rerun the same code on the same dataset. In other cases where stronger claims are at stake, the result should be reproducible over a much wider range of changed conditions. Unfortunately, the "reproducibility checklists" that are starting to be used at some ML conferences don't provide different definitions for different kinds of experimental claims.[1]

[1] They do offer a different definition for purely theoretical claims [NeurIPS, 2020].

NeurIPS 2020 Workshop: ML Retrospectives, Surveys & Meta-Analyses (ML-RSA).

One area in which having a stricter standard of reproducibility is essential is when ML is used to support scientific hypotheses about causes in the social or natural world. As ML is used increasingly in the social and natural sciences, it has been used to support ethically fraught hypotheses that come uncomfortably close to claims reminiscent of phrenology, physiognomy, and other discredited areas of study [Coalition for Critical Technology, 2020, Wang and Kosinski, 2018, Wu and Zhang, 2016]. In these cases, the failure to recognize the need for different reproducibility standards becomes especially dangerous, as the appearance of these types of research in "top" ML publication venues offers a popular respectability to scientific claims that have not been subject to the same level of vetting that, for example, a biological lab experiment would be subject to.

Ethical considerations also affect the type of reproducibility we should demand. Some philosophers of science have argued that pragmatic concerns about the risks of scientific projects should affect the evidentiary standards we apply to those projects [Rudner, 1953, Douglas, 2000, Miller, 2014]. Similarly, we might demand that ML results with higher risks for society satisfy more stringent standards for reproducibility.

2 Existing definitions of reproducibility

Debates over reproducibility are complicated by inconsistencies in terminology. Here's a small sample of the definitions available in the literature:

1. Goodman et al. [2016] differentiates between methods reproducibility, results reproducibility, and inferential reproducibility. The first means getting the same results when implementing the same computational procedures on the same data, the second means generating statistically similar results when re-implementing a method, and the third means drawing the same conclusion or finding from a different experimental setup.
2. Computational researchers have distinguished between "replicability" and "reproducibility". Initially, Peng [2011] intended the former to mean getting the same results with the orginal code and the latter to mean re-implementing the algorithm in one's own code to get similar results. However, some researchers like Chris Drummond have swapped the meanings of the two terms [Drummond, 2011].
3. Another distinction is between "direct" and "conceptual" replications, where the former aims to use experimental protocol and materials that are similar to those in the original experiment, and the latter uses different experimental protocols and materials [Guttinger, 2020]. Conceptual reproducibility is akin to Goodman et al's inferential reproducibility.

Goodman et al. [2016]'s "methods reproducibility" is typically what is emphasized in ML. However, in cases where ML is being used to support broader scientific conclusions, inferential reproducibility and conceptual replications may be more appropriate.

3 Reproducibility standards in ML conferences

Reproducibility standards in ML publication venues tend to emphasize what has been defined above as "methods reproducibility": getting the same results when implementing the same computational procedures on the same data, where "computational procedures" is taken to include the original code used. The emphasis is on making code and data accessible, and stating the metrics and parameters used to derive the reported results from the code.

The NeurIPS reproducibility checklist was first introduced during NeurIPS 2018 and has been updated for NeurIPS 2019 and 2020 [NeurIPS, 2020]. A report on the NeurIPS 2019 Reproducibility Program explicitly states that it focuses on reproducibility defined as producing the same results given the same data and same code [Pineau et al., 2020]. The 2020 NeurIPS 202 reproducibility checklist comports with this focus, as it asks authors to provide details about datasets, code, hyper-parameters, number of training and evaluation runs, definitions of metrics used, and the computing infrastructure used [NeurIPS, 2020].

The AAAI and EMNLP reproducibility checklists have a similar focus on methods reproducibility [AAAI, 2020, EMNLP]. Indeed, I was not able to find ML reproducibility checklists that went beyond methods reproducibility, which indicates that this is probably the dominant conception of reproducibility in the field.

I'll argue that methods reproducibility is an insufficient standard for many of the claims being made by ML experiments. But first, let's take a look at the variety of claims that ML experiments may make.

4 What are ML Experiments Trying to Show?

Many ML papers conform to the following form: they propose a new algorithm for some kind of task, they conduct "experiments" on one or more datasets using the new algorithm, and they demonstrate that the results are interesting in some way—either superior to that of other algorithms, or good enough to support some interesting hypothesis.

These experiments and their results, however, can be used to support very different kinds of claims. The specific type of claim is often not explicitly elucidated; instead, one has to infer this from the implications claimed by the researchers, and from other parts of the text outside of the actual results.

Here's an incomplete list of types of claims that papers based on ML experiments may be making:

1. **Method superiority**: The novel method being proposed is superior to other methods for certain types of problems. In the discussion, the authors may offer conjectures as to *why* the method is superior, but the computational experiments are not taken to support that claim, and the main claim is about the novel's method efficacy, rather than the reasons for the efficacy. Examples: He et al. [2013], Cormack et al. [2009]
2. **Theoretical virtues**: This method is superior (or inferior) to other methods, and we can attribute its superiority (or inferiority) to particular theoretical aspects of the method that are explicitly identified. This goes beyond the previous type of claim by showing why the method is superior (or inferior), rather than just showing *that* it is superior (or inferior). Examples: LeCun et al. [1990], Oakden-Rayner et al. [2020], Hassibi et al. [1993]
3. **Proof of concept**: This method *could* be useful for this type of problem. This differs from the previous two because it does not attempt to claim that the method is superior to other methods, nor does it put forward a theoretical reason for the method's superiority. Examples: Saxe et al. [2017], Tumasjan et al. [2010], Jean et al. [2016]
4. **Scientific hypotheses about natural or social phenomena**: The predictive efficacy of a particular ML method suggests hypotheses in a natural or social scientific domain. This could be used to improve existing causal theories, suggest new theories, or provide reality checks on existing theories. Examples: Burley et al. [2020], Tran et al. [2019], Thompson et al. [2020], Wang and Kosinski [2018], Liu and Xu [2016]
5. **Quantifying predictability**: The accuracy obtained by the ML application quantifies the potential predictability of the phenomenon.[2] This can serve as a benchmark for later attempts to predict the phenomenon, or induce scientists to study new areas that were thought to be unamenable to mathematical modeling. Studies that do this also effectively act as proofs-of-concepts. Examples: Tumasjan et al. [2010], Jean et al. [2016], Hosseinzadeh et al. [2013]

5 Matching Reproducibility Types to Claims

Depending on which type of claim is being made in a paper, the standards for reproducibility should be different. Here, I match each type of claim to the types of reproducibility we should demand from research making that claim. In all cases, methods reproducibility is the bare minimum level of reproducibiltiy we should expect, but in most cases we need stronger types of reproducibility to substantiate the claims being made.

[2] This possible function of ML has been pointed out by Shmueli [2010].

5.1 Method Superiority

If the claim being made in the paper is about the superiority of the method for a certain class of problems, then we should expect reproducibility of the method's success over a representative sample of that class of problems. We'll also want to rule out other reasons for the superiority that aren't inherent to the method. This means reproducing over at least the following conditions:

- Different conditions of computational power and hyperparameter optimization, within the stated range of computational power that the method is optimal at. This prevents spurious claims of method superiority when some experiments' superior results are actually due to more comprehensive hyperparameter optimization relative to previous experiments [Dodge et al., 2019].
- A representative sample of problems in the specified class of problems. This potentially includes using different data sets, different tasks, etc. For example, if a computer vision algorithm claims to be superior at object recognition in images, the superiority should ideally be demonstrated across diverse image datasets, not just ImageNet.

5.2 Theoretical Virtues

If the claim is not just that a method performs better than others but that the reason for its performance is a particular property possessed by the method, then we need to show not just that the method outperforms others, but that we can attribute that superior performance to one component of the method. This can be done by trying different variants of the method with the alleged theoretical virtue omitted or added, or comparing the performances of these variants to those of other methods lacking the theoretical virtue. If the theoretical virtue is genuine, we should see significant performance differences across multiple ways of "adding" the virtue to a model. For example, [Hassibi et al., 1993] support their claim about the virtues of their method of network pruning by conducting experiments comparing their method to other methods that lack that virtue. To reproduce that claim, one would have to go beyond re-implementing the original code and obtaining the same numerical results. Ideally, a reproduction would try implementing the virtue in a variety of different conditions and see if improvements in performance still result.

5.3 Proofs of Concept

This is the "weakest" kind of claim one can make with an ML experiment. It demonstrates the possibility that a method can achieve reasonable results. For example, a new classification algorithm may be shown to achieve reasonably high accuracy on a dataset. A performance of this kind, assuming it can be replicated by rerunning the same code, would constitute a preliminary reason for considering using the algorithm in other similar problem contexts.

The weakness of this kind of claim is that a reasonable performance on one or a few datasets does not say much about how the performance can generalize, and most algorithms are useful only when they can generalize. Acquiring more information on generalizability would require observing the algorithm's performance under a wider variety of conditions, including not just different datasets but also different computational resources. Information about which theoretical virtues of the algorithm are responsible for its good performance, as described in the previous section, would also help determine the algorithm's generalizability.

Precisely because this kind of claim is relatively weak, methods reproducibility can suffice for it. The less a piece of research asserts a claim to generalizability, the less we need to show that it works well under a variety of different conditions.

5.4 Natural or Social Scientific Hypotheses

While the field of ML has a strong predictive slant, ML is also sometimes used to provide evidence for or against scientific hypotheses. This may be a more common ML application in the social and natural sciences, which have stronger emphases on confirming hypotheses, in contrast to more engineering-oriented fields.

When ML is used to support scientific hypotheses, methods reproducibility is too weak an epistemic standard. In these cases, it becomes key to make sure that the algorithm's success on one or a few datasets is actually due to relevant features of the world as represented in the dataset, rather than due to quirks of the datasets, preprocessing, or model that are irrelevant to the scientific hypothesis. Showing that the algorithm "works" is not enough in these cases, because we also have to ensure that the model is in fact reflecting the underlying pattern being hypothesized, rather than reflecting some other aspect of the dataset that is unrelated to the hypothesis. While we do want methods reproducibility for ML experiments that claim to support scientific hypotheses, inferential reproducibility or conceptual replications (see definitions in Section 2) may also be called for in these cases.

5.4.1 Example: Using ML to model odorant receptors

Tran et al. [2019] is one example of an ML paper providing evidence for a scientific hypothesis. They explicitly state their hypothesis to be about causal structures in nature: "Our main hypothesis is that the [odorant receptor] ensemble forms a set of 3D filters that interact with molecules in real space." Their finding supports the causal hypothesis: "we can associate the latent variables produced by our autoencoder with the responses of the ensemble of [odorant receptors]."

Given this type of hypothesis, inferential reproducibility is key—we should focus on obtaining the same *finding* about the responses of the ensemble of odorant receptors. This could be done by a physical experiment, or, in an ML context, could also be done by making changes in the original experiment that ought to be irrelevant to the finding. For example, one may vary the datasets used or try alternative ways of operationalizing human olfactory percepts.

5.4.2 Example: Predicting sexual orientation from facial images

This next example demonstrates how having appropriate reproducibility standards can help guard against the resurgence of scientifically dubious attempts to use ML to revive ethically dubious fields like phrenology and physiognomy. The example project is arguably unethical for reasons independent of reproducibility, but it also demonstrates how methods reproducibility would be insufficient to validate the results if the claim being made is a social scientific hypothesis.

Wang and Kosinski [2018] created a social media firestorm when they claimed that an algorithm was more accurate than humans at predicting sexual orientation from faces. The authors link this work to the long history of physiognomy and state that they are testing the hypothesis that "some of our intimate traits are prominently displayed on the face, even if others cannot perceive them." They link their investigation to prenatal hormone theory, suggesting that any physiognomic differences they detect could be due to prenatal hormones.

In a case like this, where the intent is to provide support for a scientific hypothesis, reproducibility should go beyond methods reproducibility to require obtaining the same results under more widely varying set of conditions, including:

- Using different datasets. The original study used publicly available images from a dating website and Facebook, but the hypothesis is making claims about faces in general, so one should make sure that the results hold across faces in general, not just faces as presented in these venues.
- Using different ways of operationalizing the construct being measured. Other ML researchers have made the point that a lot of ML research does not consider the importance of construct validity and construct repeatability, as defined in the social sciences [Jacobs and Wallach, 2019]. Construct validity and construct repeatability are concepts that describe whether a measurement of a theoretical construct (of which sexual orientation is an example) is measured reliably and correctly. A construct may be measured unreliably if it cannot be repeated in different circumstances, or if the way it's measured does not match what the theory demands of it. Wang and Kosinski [2018] do not consider if their method of estimating the "ground truth" of sexual orientation has construct validity or construct repeatability in these senses. To support the scientific hypothesis in a robust way, reproducing the results by re-running the code is not enough—the measurement of the "ground truth" should also be validated by attempting to measure it in different ways and validating that different methods of measurement agree (among other things).

In the absence of further experiments reproducing the paper's results under relevantly different sets of conditions, the results of the paper should be treated only as a proof-of-concept, not as support for a scientific hypothesis.[3] For the purposes of confirming a scientific hypothesis, we are not interested in algorithmic "successes" that are due to quirks of, for example, data collection or preprocessing, that cannot be attributed to an underlying "reality" in actual faces. However, the evidence presented in the paper leaves wide open the possibility that the model is latching on to quirks in the datasets rather than properties of real faces.

The broader ethical ramifications of research like this could also be a reason for demanding more stringent standards of reproducibility. Some philosophers of science have argued that scientific projects that present higher risks ought to meet higher evidentiary standards [Rudner, 1953, Douglas, 2000, Miller, 2014]. Given the risks of this type of research being used to oppress sexual minorities, it may be appropriate to impose higher-than-normal reproducibilty standards on it.

5.5 Quantifying Predictability

As remarked above, many studies that attempt to quantify the predictability of a phenomenon also effectively act as proofs-of-concepts: if successful, they show that it's possible for a prediction algorithm to "work" for a certain task. If the "phenomenon" whose predictability is being quantified is sufficiently narrow, methods reproducibility may be sufficient as a standard of reproducibility.

To take as an example, the case of quantifying the predictability of poverty from satelite images [Jean et al., 2016] may be framed in different ways. If the aim is to quantify the predictability of poverty from that specific set of satelite images, then methods reproducibility, where one attempts to obtain the same results from the original dataset and code, is enough. But one could also have a broader aim of quantifying the predictability of poverty from aerial images in general, rather than from a specific set of images. In that case, a stronger form of reproducibility, using different datasets of aerial images, would be called for.

6 Conclusion

I've argued that ML computational experiments are used to make many different types of scientific claims, and our standards of reproducibility should vary based on the type of claim being made. This is in contrast with conferences suggesting only one type of reproducibility standard for ML experiments. Furthermore, the reproducibility standards enforced by ML conferences capture only a weak form of reproducibility, namely methods reproducibility. A lot of ML research makes claims that are broader than proof-of-concept-style claims about a method "working" on a particular dataset. When making broader claims, such as a method being superior on a class of problems, or a method's success providing evidence for a scientific hypotheses, the standards of reproducibility should be correspondingly stronger. Methods reproducibility alone is insufficient, in these cases, to assure us of the reliability of these broader claims.

The obvious solution might be to have different reproducibility checklists for different types of claims. But here we confront a deeper problem: ML papers are not always clear on the scope of the claim being made. As I discussed using an example in Section 5.5, the same paper can be interpreted to be making claims of different scopes, as ML publishing norms do not yet require authors to be clear on the scopes of their claims. Scope could also be something that should be specified in advance as part of a pre-registration procedure [Nosek et al., 2018]. Thus, a re-evaluation of publicaiton norms, in addition to making more sophisticated checklists, is worth considering.

References

AAAI. AAAI reproducibility checklist. https://aaai.org/Conferences/AAAI-21/reproducibility-checklist/, 2020. Accessed: 2020-09-13.

Timothy Burley, Lorissa Humble, Charles Sleeper, Abigail Sticha, Angela Chesler, Patrick Regan, Ernesto Verdeja, and Paul Brenner. NLP workflows for computational social science: Understanding triggers of state-led mass killings. In *Practice and Experience in Advanced Research Computing*, PEARC '20, page 152–159, New York, NY, USA, 2020. Association for Computing Machinery. URL https://doi.org/10.1145/3311790.3397343.

[3] Again, I'm not advocating for scientific resources to be spent actually doing this reproduction, as one may hold the view that physiognomy has been so thoroughly debunked that we don't need to try and verify it again.

Coalition for Critical Technology. Abolish the #techtoprison-pipeline. https://medium.com/@CoalitionForCriticalTechnology/abolish-the-techtoprisonpipeline-9b5b14366b16, 2020. Accessed: 2020-09-20.

Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759, 2009.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A Smith. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, 2019.

Heather Douglas. Inductive risk and values in science. *Philosophy of Science*, 67(4):559–579, 2000.

Chris Drummond. Replicability is not reproducibility: nor is it good science. http://cogprints.org/7691/, 2011.

EMNLP. EMNLP call for papers. https://2020.emnlp.org/call-for-papers. Accessed: 2020-09-13.

Steven N. Goodman, Daniele Fanelli, and John P. A. Ioannidis. What does research reproducibility mean? *Science Translational Medicine*, 8(341):341ps12–341ps12, 2016.

Stephan Guttinger. The limits of replicability. *European Journal for Philosophy of Science*, 10(2):10, January 2020.

Babak Hassibi, David G Stork, and Gregory J Wolff. Optimal brain surgeon and general network pruning. In *IEEE International Conference on Neural Networks*, pages 293–299. IEEE, 1993.

Kaiming He, Fang Wen, and Jian Sun. K-means hashing: An affinity-preserving quantization method for learning binary compact codes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2938–2945, 2013.

Arian Hosseinzadeh, Masoumeh Izadi, Aman Verma, Doina Precup, and David Buckeridge. Assessing the predictability of hospital readmission using machine learning. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, AAAI'13, page 1532–1538. AAAI Press, 2013.

Abigail Z Jacobs and Hanna Wallach. Measurement and fairness. *arXiv preprint arXiv:1912.05511*, 2019.

Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.

Yann LeCun, John S Denker, and Sara A Solla. Optimal brain damage. In *Advances in Neural Information Processing Systems*, pages 598–605, 1990.

Sabina Leonelli. Rethinking reproducibility as a criterion for research quality. In *Research in the history of economic thought and methodology, volume 36B: Including a symposium on Mary Morgan: curiosity, imagination, and surprise*. Emerald Publishing Limited, 2018.

Y. Liu and S. Xu. Detecting rumors through modeling information propagation networks in a social media environment. *IEEE Transactions on Computational Social Systems*, 3(2):46–62, 2016.

Boaz Miller. Science, values, and pragmatic encroachment on knowledge. *European Journal for Philosophy of Science*, 4(2):253–270, 2014.

NeurIPS. The machine learning reproducibility checklist (v2.0, apr.7 2020). https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf, 2020. Accessed: 2020-09-13.

NeurIPS. Designing the reproducibility program for NeurIPS 2020. https://medium.com/@NeurIPSConf/designing-the-reproducibility-program-for-neurips-2020-7fcccaa5c6ad, 2020. Accessed: 2020-09-13.

Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018.

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Re. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, CHIL '20, page 151–159, New York, NY, USA, 2020. Association for Computing Machinery.

Roger D Peng. Reproducible research in computational science. *Science*, 334(6060):1226–1227, 2011.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). `https://arxiv.org/abs/2003.12206`, 2020.

Edward Raff. A step toward quantifying independently reproducible machine learning research. In *Advances in Neural Information Processing Systems*, pages 5485–5495, 2019.

Richard Rudner. The scientist qua scientist makes value judgments. *Philosophy of Science*, 20(1): 1–6, 1953.

Glenn N Saxe, Sisi Ma, Jiwen Ren, and Constantin Aliferis. Machine learning methods to predict child posttraumatic stress: a proof of concept study. *BMC Psychiatry*, 17(1):223, 2017.

Galit Shmueli. To explain or to predict? *Statistical Science*, 25(3):289–310, 2010.

Bill Thompson, Seán G Roberts, and Gary Lupyan. Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour*, pages 1–10, 2020.

Ngoc Tran, Daniel Kepple, Sergey Shuvaev, and Alexei Koulakov. DeepNose: Using artificial neural networks to represent the space of odorants. volume 97 of *Proceedings of Machine Learning Research*, pages 6305–6314, Long Beach, California, USA, 09–15 Jun 2019.

Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Fourth International AAAI Conference on Weblogs and Social Media*, 2010.

Yilun Wang and Michal Kosinski. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*, 114(2):246, 2018.

Xiaolin Wu and Xi Zhang. Automated inference on criminality using face images. `https://arxiv.org/pdf/1611.04135v1.pdf`, 2016. Accessed: 2020-09-20.