Computational Brain & Behavior (2019) 2:242–246
https://doi.org/10.1007/s42113-019-00055-w

ORIGINAL PAPER

# Towards Replication in Computational Cognitive Modeling: a Machine Learning Perspective

Chris Emmery[1] · Ákos Kádár[1] · Travis J. Wiltshire[1] · Andrew T. Hendrickson[1]

Published online: 14 August 2019

## Abstract

The suggestions proposed by Lee et al. to improve cognitive modeling practices have significant parallels to the current best practices for improving reproducibility in the field of Machine Learning. In the current commentary on "robust modeling in cognitive science", we highlight the practices that overlap and discuss how similar proposals have produced novel ongoing challenges, including cultural change towards open science, the scalability and interpretability of required practices, and the downstream effects of having robust practices that are fully transparent. Through this, we hope to inform future practices in computational modeling work with a broader scope.

**Keywords** Reproducibility · Machine learning · Cognitive science

## Introduction

Over the last decade, Machine Learning (ML) has seen rapid performance gains, and exponentially increasing research interest, as a result of the success of (deep) neural networks, and openness of research. However, these advances have simultaneously brought forth increased scientific and societal concerns. In parallel to the continuous efforts on improving the interpretability of these purportedly "black box" model representations (Lipton and Steinhardt 2018; Rahimi and Recht 2017; Sculley et al. 2015, 2018b; Woods 2018), recent attention has shifted towards the lack of rigor in, and robustness of, the models achieving these results. More importantly, concerns have been raised about how the broader Artificial Intelligence community is affected by the speed of developments in these methods (Hutson 2018).

The issues with reproducibility in ML are perhaps surprising, given that one major driving force for the growth and interest has been the proliferation of powerful open-source libraries and pre-trained models that lower the bar for entry to the field. Furthermore, many of the properties of ML research are conducive to open, reproducible research: large standardized publicly available datasets, generally less noisy data collection, community-driven shared tasks and benchmarks, the ability to share complete research code, and a tendency towards open-access publication (Munafò et al. 2017). ML may arguably do so more than other fields applying computational modeling.

Despite these inherent advantages, (applied) ML research is still considered predominantly irreproducible (Gundersen and Kjensmo 2018). Efforts to reproduce the results (not just reuse published models) of seminal papers have been met with a multiplicity of issues, including unclear experimental descriptions (Pineau et al. 2018; Tian et al. 2019), failures to generalize to out-of-domain datasets (e.g., Recht et al. 2019; Zhang et al. 2019), and evaluation results that do not replicate (Melis et al. 2017). Significant blame has been allocated to the ecosystem of research in ML for exacerbating replication issues (Gebru et al. 2018; Henderson and Brunskill 2018; Hutson 2018; Lipton and Steinhardt 2018; Mitchell et al. 2019; Sculley et al. 2018a, b—many of which we deem relevant for any computational field.

✉ Chris Emmery
c.d.emmery@tilburguniversity.edu

Ákos Kádár
a.kadar@tilburguniversity.edu

Travis J. Wiltshire
t.j.wiltshire@tilburguniversity.edu

Andrew T. Hendrickson
a.hendrickson@tilburguniversity.edu

[1] Department of Cognitive Science and Artificial Intelligence, Tilburg University, Tilburg, Netherlands

We would accordingly like to emphasize that while part of our commentary is specifically about code reproducibility, in general, we uphold the Method Reproducibility framing of Gundersen and Kjensmo (2018), which assumes that any model conceptualization, given different data and an alternative implementation, should still produce the same results and support the same inferences and conclusions. This commentary highlights one primary and three secondary challenges for such reproducibility in machine learning and how they directly apply to research practices in computational cognitive modeling.

## Main Challenge: Inertia of Cultural Change

Publications of computational research have long been framed as a medium of promotion, where the main contribution is not provided on paper, but in the data and software (Ince et al. 2012; Buckheit and Donoho 1995; Claerbout and Karrenbach 1992). Over the years, several reports have highlighted the importance of transparent research practices to guarantee reproducibility including adequate documentation, open-source licensing, and persistent hosting on trusted repositories (e.g., Gundersen and Kjensmo 2018; Peng 2011; Stodden et al. 2016; Tatman et al. 2018). Part of those reports extend the Transparency and Openness Promotion (TOP)[1] guidelines to fit computational research, others introduced metrics for measuring transparency of publications (Gundersen and Kjensmo 2018). As a result, the ML community has developed an understanding of reproducibility as "(...) the ability of an independent research team to produce the same results using the same [computational] method based on the documentation made by the original research team" (Gundersen and Kjensmo, 2018, p. 1645), which implies fully executable code needs to be made available (Peng 2011), and data collection and annotation practices to be transparent.

On a surface level at least, it would seem that reproducibility has been part of the ML research agenda for quite a while. Journals have started special issues (Branco et al. 2017), conference workshops (AAAI, NeurIPS, ICLR, ICML), and special tracks (e.g., COLING) to promote these efforts. We have seen heavy use of open access pre-prints (arXiv), code repositories (GitHub, Papers With Code), and e-publishing (Distill).[2] Work that has been well-documented and provided open-source is more likely to have higher citation numbers and impact (e.g., Devlin et al. 2018; Isola et al. 2017; Mikolov et al. 2013; Russakovsky et al. 2015). Yet, survey data shows a historical lack of researchers adhering to this agenda in practice (Gundersen and Kjensmo 2018). While pseudo-code (54% shared) and training data (56%) seem more common, troubling numbers arise for test data (30%), and most notably code (7% avg).

We argue that the main challenge here is the disconnect between the pace at which best practices are introduced, the rate at which a lack of adherence to these practices are identified, and finally the time it takes to correct these issues and apply in practice. One major factor is that appropriate credit attribution (as argued in favor of in Stodden et al. 2013) is not easily entrenched in the ML research community. Several of the mentioned reports have argued proper attribution requires a structural culture change—both in the practices of researchers, as well as that of journals, conferences, and funding agencies. A system without such rewards heavily favors researchers who do a minimum amount of work. Reproducibility therefore remains absent as a strict criterion for publishing in ML. We illustrate some plausible causes for this in the following secondary challenges.

## Sub-challenge #1: Scalability of Producing Reproducible Code

Our first persistent issue relates to the effect extensive reproducibility requirements have on different researchers, and applies to most of the pre-registration and reporting practices proposed in Lee et al. (2019). Diverse levels of interest and expertise might cause great variation in the workload that full reproducibility requires. Running a single well-understood model in a popular programming or scripting language on readily available data requires much less documentation to be written than the introduction of an algorithm or a novel combination of methods and/or collected data. In the latter cases, significant requirements are imposed on the same pool of researchers that are more akin to that of a software engineer; well-documented (according to accepted standards) and reusable code is expected to be modular. Otherwise, the experiment cannot be repeated in isolation, on new data, or with a different method. Sharing this code as software is not a trivial task, and requires detailed familiarity with computational infrastructure, as it should be reproducible between different machines with varying hardware and operating systems (Tatman et al. 2018).

Furthermore, an emphasis on fine-grained logging and creation of research-related documentation also quickly expands the administrative complexity introduced in this process. When version numbers of higher-level software libraries, lower-level hardware drivers, and model numbers of that hardware all have effect on running particular models

[1] https://cos.io/top or (archived): https://web.archive.org/web/20190617070618/https://cos.io/top/

[2] See: https://arxiv.org, https://github.com, https://paperswithcode.com, and https://distill.pub respectively.

(as is often the case in ML), it becomes increasingly important for researchers to be able to rely on straightforward, standardized ways of reporting research. If reproducibility practices are not in a mature stage, and the infrastructure for doing so is not public, the overhead related to preparation of the required materials is likely to be inversely proportional to the size of the research groups and the degree of past experience with such procedures. Hence, it could well turn into counter-productive practices when open science is deemed important, favoring larger labs with established workflows and higher chances of credit for their efforts.

## Sub-challenge #2: Interpretability of Reproducible Research

Even after solid, reproducible work has been created, there are significant challenges remaining. Lee et al. propose documentation in the style of registered reports (Chambers et al. 2015; Hardwicke and Ioannidis 2018) and activity logging in Jupyter (Kluyver et al. 2016) or R Markdown (Baumer et al. 2014). However, concerns have been raised in the ML community about the degree to which these highly flexible tools promote actual reuse of code.

Notebooks have been applauded by many, yet have also seen pushback from an engineering point of view (see, e.g., Grus 2018; Pimentel et al. 2019). We argue that they work well for free-form documentation, and demonstrating selected parts of research. However, we also partly agree with aforementioned critiques; hosting the majority of the written code in this format encourages bad practice in terms of structured software engineering, discourages modularity, and is disconnected from the way mature open-source projects have agreed on documenting their computational work (see, e.g., Pedregosa et al. 2011), code versioning is handled, and code is run as large compute jobs on servers.[3] Most importantly, notebooks do not offer a structured way to search or analyze all resulting information.

Documentation of methods, data properties, performance of (bookend) models in several metrics, research logs, and extensive error analyses—as proposed by Lee et al.—quickly ramp up the human effort required to analyze all results. Discovery of errors, similarities to other research, or unexplored avenues in these results can only be reasonably facilitated if search time is optimized. This implies the need to rigorously provide structure and documentation standards to present all this information, thus making it humanly interpretable. One promising direction would be to focus on community-accepted tools for the automatic generation of summaries of all this meta-data. Failing to provide interpretation of time-consuming documentation would dismiss the academic potential of the proposed practices, and the credit for investing in reproducibility.

## Sub-challenge #3: Downstream Effects of Reproducibility

Our final issue again assumes prior challenges have been dealt with; research is fully reproducible and interpretable, adhering to the practices as presented by Lee et al. (2019). This leaves dealing with the impact that such a significant culture change has on the research community. Hutson (2018), for example, mentions the delicate dynamics of communicating observed errors in code between junior and senior researchers. Particularly in computational research, a small bug could have significant effect on the impact of a publication. Research methods becoming more widely documented, accessible, and better understood have proven a double-edge blade for ML: facilitating a faster research pace, with openly accessible pre-prints introducing additional issues such as flag planting with preliminary (often incomplete) results (Feldman et al. 2018). On the upside, it has established a unique position regarding the ability to reassess many of these novel methods, and to compare them against established ones. We would like to end this commentary with a few of such examples from the last year, and finally the noticeable positive change in culture these have caused.

A relevant phenomenon in applied ML is the heavy focus on increasing performance on delimited tasks and benchmarks, combined with the introduction of increasingly complex models tested on these. Often, such work forgoes spending time on optimizing baselines or critically assessing the task that is being tackled, as these are not seen as part of the main contributions. When they are critically investigated in large-scale studies, however, surprising results emerge: in Language Modeling for example (the task of predicting the next word given a span of text), Merity et al. (2018) compared ten models of papers applied on the same data. They concluded that models performing well on the most popular benchmarks exhibit pathological behavior due to biases in these data. More importantly, the models introduced more than twenty years ago were still state-of-the-art with proper tuning and small tricks. In Reinforcement Learning, models were shown to be sensitive to minor changes in implementation details as small as changing random seeds (Henderson et al. 2018). In the same field, the reliance on complex evaluation protocols when

[3] For more practical notebooks issues, we recommend watching Grus's "I don't like notebooks." talk (see the references for the YouTube link). Slides are available from: https://conferences.oreilly.com/jupyter/jup-ny/public/schedule/detail/68282.

evaluating on popular benchmarks resulted in Mania et al. (2018) proposing a particularly straight-forward method based on finite-difference random search (Matyas 1965). They found its performance competitive with the most sophisticated current models. Lastly, recently, Locatello et al. (2018) performed a large-scale reproduction study and critical analysis of the field of disentangled representations learning. Their study points out that there is no agreed upon definition of disentanglement, and the current models do not help in reducing the number of training examples required for training domain-specific classifiers. Finally, the performance of current state-of-the-art models was shown to depend on hyperparameter settings and minute details, more than proposed architectural choices.

One might conclude these events are closely related to the replication crisis in the field of psychology (Pashler and Wagenmakers 2012), although we believe an important distinction here is the speed of which a field with a high publication pace and a rising open-source culture is able to adapt to such impactful work. Very recently major venues such as NeurIPS introduced a soft Code Submission Policy,[4] and an author-prepared Reproducibility Checklist[5] made available to reviewers. Moreover, a Reproducibility Chair was assigned to guide the analyses of the impact of these experimental measures—all clear indications of culture shifts in the right direction.

## Conclusion

In this commentary, we highlighted several issues that have stalled progress towards fully open science in Machine Learning. While fields dealing with computational research have not advanced enough to deal with all of these, a direct effect on the ML community and a positive outlook for future practices are clearly present. We are therefore confident that gradual introduction of similar best practices suggested by Lee et al. might bring about similar effects. We do however hope the remaining challenges presented in this commentary can serve as a recommendation to move at a cautionary pace implementing the proposed practices.

[4] https://medium.com/@NeurIPSConf/call-for-papers-689294418f43 or (archived) https://web.archive.org/web/20190530143750/ https://medium.com/@NeurIPSConf/call-for-papers-689294418f43

[5] https://cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf

## References

Baumer, B., Cetinkaya-Rundel, M., Bray, A., Loi, L., Horton, N.J. (2014). R markdown: integrating a reproducible analysis tool into introductory statistics. arXiv:14021894.

Branco, A., Cohen, K.B., Vossen, P., Ide, N., Calzolari, N. (2017). Replicability and reproducibility of research results for human language technology: introducing an lre special section.

Buckheit, J.B., & Donoho, D.L. (1995). Wavelab and reproducible research. In *Wavelets and statistics* (pp. 55–81): Springer.

Chambers, C.D., Dienes, Z., McIntosh, R.D., Rotshtein, P., Willmes, K. (2015). Registered reports: realigning incentives in scientific publishing. *Cortex*, *66*, A1–A2.

Claerbout, J.F., & Karrenbach, M. (1992). Electronic documents give reproducible research a new meaning. In *SEG technical program expanded abstracts 1992, society of exploration geophysicists* (pp. 601–604).

Devlin, J., Chang, M.W., Lee, K., Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding. arXiv:181004805.

Feldman, S., Lo, K., Ammar, W. (2018). Citation count analysis for papers with preprints. arXiv:180505238.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumeé, I.H., Crawford, K. (2018). Datasheets for datasets. arXiv:180309010.

Grus, J. (2018). I don't like notebooks. https://www.youtube.com/watch?v=7jiPeIFXb6U, accessed 07/19/19.

Gundersen, O.E., & Kjensmo, S. (2018). State of the art: reproducibility in artificial intelligence. In *Thirty-second AAAI conference on artificial intelligence*.

Hardwicke, T.E., & Ioannidis, J.P. (2018). Mapping the universe of registered reports. *Nature Human Behaviour*, *2*(11), 793.

Henderson, P., & Brunskill, E. (2018). Distilling information from a flood: a possibility for the use of meta-analysis and systematic review in machine learning research. arXiv:181201074.

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., Meger, D. (2018). Deep reinforcement learning that matters. In *Thirty-second AAAI conference on artificial intelligence*.

Hutson, M. (2018). Artificial intelligence faces reproducibility crisis.

Ince, D.C., Hatton, L., Graham-Cumming, J. (2012). The case for open computer programs. *Nature*, *482*(7386), 485.

Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1125–1134).

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B.E., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J.B., Grout, J., Corlay, S., et al. (2016). Jupyter notebooks-a publishing format for reproducible computational workflows. In *ELPUB* (pp. 87–90).

Lee, M., Criss, A., Devezer, B., Donkin, C., Etz, A., Leite, F., Matzke, D., Rouder, J., Trueblood, J., White, C., Vandekerckhove, J. (2019). Robust modeling in cognitive science. PsyArXiv https://psyarxiv.com/dmfhk/.

Lipton, Z.C., & Steinhardt, J. (2018). Troubling trends in machine learning scholarship. arXiv:180703341.

Locatello, F., Bauer, S., Lucic, M., Gelly, S., Schölkopf, B., Bachem, O. (2018). Challenging common assumptions in the unsupervised learning of disentangled representations. arXiv:181112359.

Mania, H., Guy, A., Recht, B. (2018). Simple random search provides a competitive approach to reinforcement learning. arXiv:180307055.

Matyas, J. (1965). Random optimization. *Automation and Remote control*, *26*(2), 246–253.

Melis, G., Dyer, C., Blunsom, P. (2017). On the state of the art of evaluation in neural language models. arXiv:170705589.

Merity, S., Keskar, N.S., Socher, R. (2018). An analysis of neural language modeling at multiple scales. arXiv:180308240.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D., Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220-229): ACM.

Munafò, M.R., Nosek, B.A., Bishop, D.V., Button, K.S., Chambers, C.D., Du Sert, N.P., Simonsohn, U., Wagenmakers, E.J., Ware, J.J., Ioannidis, J.P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 0021.

Pashler, H., & Wagenmakers, E.J. (2012). Editors' introduction to the special section on replicability in psychological science: a crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in python. *Journal of machine learning research*, *12*(Oct), 2825–2830.

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

Pimentel, J.F., Murta, L., Braganholo, V., Freire, J. (2019). A large-scale study about quality and reproducibility of jupyter notebooks. In *Proceedings of the 16th international conference on mining software repositories* (pp. 507–517): IEEE Press.

Pineau, J., Fried, G., Ke, R., Larochelle, H. (2018). Iclr 2018 reproducibility challenge. In *ICML workshop on reproducibility in machine learning*.

Rahimi, A., & Recht, B. (2017). Reflections on random kitchen sinks.

Recht, B., Roelofs, R., Schmidt, L., Shankar, V. (2019). Do imagenet classifiers generalize to imagenet? arXiv:190210811.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.F., Dennison, D. (2015). Hidden technical debt in machine learning systems. In *Advances in neural information processing systems* (pp. 2503–2511).

Sculley, D., Snoek, J., Wiltschko, A. (2018a). Avoiding a tragedy of the commons in the peer review process. arXiv:190106246.

Sculley, D., Snoek, J., Wiltschko, A., Rahimi, A. (2018b). Winner's curse? on pace, progress, and empirical rigor.https://openreview.net/forum?id=rJWF0Fywf.

Stodden, V., Borwein, J., Bailey, D.H. (2013). Setting the default to reproducible. *Computational Science Research SIAM News*, *46*(5), 4–6.

Stodden, V., McNutt, M., Bailey, D.H., Deelman, E., Gil, Y., Hanson, B., Heroux, M.A., Ioannidis, J.P., Taufer, M. (2016). Enhancing reproducibility for computational methods. *Science*, *354*(6317), 1240–1241.

Tatman, R., VanderPlas, J., Dane, S. (2018). A practical taxonomy of reproducibility for machine learning research. https://openreview.net/forum?id=B1eYYK5QgX.

Tian, Y., Ma, J., Gong, Q., Sengupta, S., Chen, Z., Pinkerton, J., Zitnick, C.L. (2019). Elf opengo: an analysis and open reimplementation of alphazero. arXiv:190204522.

Woods, B. (2018). Expanding search in the space of empirical ml. arXiv:181201495.

Zhang, C., Bengio, S., Hardt, M., Singer, Y. (2019). Identity crisis: memorization and generalization under extreme overparameterization. arXiv:190204698.