The Reproducibility Crisis Is Real

Odd Erik Gundersen

■ *The reproducibility crisis is real, and it is not only the field of psychology that has to deal with it. All the sciences are affected; the field of artificial intelligence is not an exception.*

The reproducibility crisis is real, and it is not only the field of psychology that has to deal with it. All the sciences are affected; the field of artificial intelligence (AI) is not an exception. To recover from this crisis, one has to accept that there is a problem. This is the first step. Say after me: "The reproducibility crisis is real, even for AI." You might not be convinced yet, so let me try to convince you.

In 2016, a poll was conducted on *Nature*'s web site and the results were reported in the journal *Nature* (Baker 2016). The poll was conducted as a brief online questionnaire in which 1,576 researchers participated. Fifty-two percent of the respondents answered that there is a significant reproducibility crisis going on, while thirty-eight percent thought there is a slight crisis. (This makes me wonder about the term *crisis*. Could there be such a thing as a "slight crisis"?) Only three percent believes that there is no crisis, and seven percent do not know. This means that ninety percent of those taking the poll believe that there is an ongoing reproducibility crisis.

Other questions were asked as well. Almost ninety percent of the scientists doing chemistry had failed to reproduce other researchers' experiments, and the numbers were just above sixty percent for those respondents belonging to the group of sciences other than those mentioned specifically in the *Nature* article. For all groups, it was found that between forty percent and sixty percent had failed to reproduce their own experiments! The respondents rated "selective reporting" as the factor that contributed the most to irreproducible research, while other important factors included "pressure to publish," "low statistical power," and "poor analysis."

Of course, there are problems related to online polls. *Nonresponse bias* is one such problem. Not all researchers visit *Nature*'s website, and for those that do, there is a response bias. People feeling strongly about something are more likely to take the poll. The article says nothing about how they were sure that only actual researchers responded to the poll, so there is a coverage bias as well. At least the sampling size is fair, so sampling bias should not be too problematic.

So, how about AI then? As part of the 2018 International Conference on Learning Representations, the Reproducibility Challenge was organized. The challenge was to reproduce papers submitted through their open review process — while it was ongoing. This allowed the participants of this challenge to easily communicate with the authors of the papers they tried to reproduce. In the end, ninety-eight different researchers participated in the challenge. They were asked more or less the same questions as were asked in the *Nature* poll.

Before the challenge started, twenty-two percent of the participants believed that there was a significant crisis, while forty-nine percent considered it to be slight. Seventeen percent were not sure, and eleven percent thought there was no crisis at all. Interestingly, the participants were asked whether their opinion had changed after participating in the Reproducibility Challenge. Fifty-one percent stated that their opinion had not changed, eleven percent were not sure, eight percent were less convinced, and thirty percent were more convinced that there was a reproducibility crisis.

The biases of this study are less problematic than those of the *Nature* website poll. Also, the study shows that most of the AI researchers partaking in the challenge believed there is a significant or slight reproducibility crisis going on, and even more so after trying to actually reproduce the results presented in papers. Joelle Pineau presented these results as part of her keynote talk for the 2018 International Conference on Learning Representations. (You can easily find this talk on YouTube. If you have forty-five minutes to spare, I suggest you give it a try. I found the keynote very interesting.)

It is clear that reproducibility is tightly connected to documentation of experiments. Any physics or chemistry student would know. They spend their first years at the university writing detailed laboratory reports. Needless to say, sharing is important as well. In what other way could fellow colleagues know about the research? To evaluate the results, they need to know what exactly was investigated and how the experiments were conducted. The more details the documentation contains, the easier it is for independent researchers to reproduce the results. In itself, good documentation builds trust in the results. Also, it lowers the barriers for others to actually run the experiment themselves, as more detailed documentation reduces the effort required to conduct the experiment.

Given that reproducibility requires good documentation, it is alarming how poorly top AI research is documented. Sigbjørn Kjensmo and I conducted a study where we reviewed 400 papers from two installments of the International Joint Conference on Artificial Intelligence and the Association for the Advancement of Artificial Intelligence Conference, which are considered to be among the most prestigious conferences in our field of AI (Gundersen and Kjensmo 2018).

Our survey shows that AI research is not well documented. Around seventy percent of the research that is published in AI experiments is empirical, but neither hypotheses nor predictions are explicitly stated. These elements are the basis of the scientific method. The same goes for explicitly stating research questions and which research methods were used. Few explicitly state the objective of the research, while the problem that was being solved was stated in less than half of the papers we reviewed.

Both the International Joint Conference on Artificial Intelligence and the Association for the Advancement of Artificial Intelligence Conference are general AI conferences where top research in very narrow domains is presented. When writing a paper for a general conference, one should state why the presented research is relevant and important, even if everyone in the subfield is fully aware of this. I have read papers presented at these top conferences where I did not understand why the authors conducted the research; they never even hinted at which problem they solved or why it was relevant to me.

Given that AI is a fairly young field of science, the research methodology and analytics methods are still being experimented with. John P. A. Ioannidis mentioned this in his famous paper "Why Most Research Findings Are False" (Ioannidis 2005). He presents several reasons for why most research findings are false for most research designs and for most fields. Let me mention a few.

There is no surprise that *small sample sizes* present a problem, but *small effect sizes* are a problem as well. *Effect size* is related to how much better one method is when compared with another. (This is worth remembering when a new method is only 0.5-to-1.5–percent better than the methods it is compared with, as this is in the small-effect–size range, according to Ioannidis.)

Problems are also related to the flexibility of study designs, definitions, analytical modes, and hotness of

the field. Generally, there is little focus on study design in AI, but some examples do exist, such as "How Evaluation Guides AI Research" (Cohen and Howe 1988) and "Empirical Methods for Artificial Intelligence" (Cohen 1995). Do we, as a community, focus enough on research methods? Is this something we teach our Master's degree and PhD students?

Definitions is another example where the AI community could improve. Many of us have been involved in research that is described by terms such as context, pervasive computing, ambient intelligence, and so on. I am not sure that we agree on what these terms exactly mean. For example, Bazire and Brézillon (2005) present a study of 150 definitions of context. According to rumors, Brézillon is still counting, and the number of definitions has at least doubled since 2005. Some would say that the term *artificial intelligence*, itself, is not well-defined. (*Computational intelligence* has even been proposed as a better term, but until *AI Magazine* is renamed *CI Magazine*, I think I will stick to AI.)

The number of analytical modes in modern machine-learning experiments is huge. The algorithms that are evaluated might have millions of hyperparameters. Do our experiments find actual patterns that are generalizable, or are we just searching for hyperparameters that find random patterns existing in both training and test sets?

When it comes to hotness, few fields can compete with AI these days. Just look at the number of papers submitted to the top conferences. Almost 8,000 papers were submitted to the 2019 Association for the Advancement of Artificial Intelligence Conference, and more than 4,700 to the 2019 International Joint Conference on Artificial Intelligence. Both conferences had a record number of submissions in 2019. It is a great time to be an AI professional in both the industry and academia, as research grants and other project funding sources are abundant. According to Ioannidis, this affects research results. Competition makes it more important to pursue and disseminate the most impressive positive results first. When this happens, the focus on research methodology might slip. It is not hard to relate to the competitiveness, at least for anyone doing research in deep learning.

Another part of documenting an experiment is the data. There has been a focus on data sharing since the University of California Irvine machine-learning repository was created in 1987 by David Aha and fellow graduate students. Sharing of data facilitates other researchers being able to reproduce the experiments and test their own ideas on standard data sets. In our survey, Kjensmo and I found that around half of the papers share the data used for conducting the experiment (however, we did not assess how many of them used standard data sets, and how many released new ones).

Some argue that standard data sets that are free for everyone will lead to researchers focusing on simply making small increments of improvement on methods that all solve the same problem. Then, the research gets very narrow. I acknowledge this sentiment. However, it does not mean that we should keep our data sets private, just that we should still try and share the data we work on. This is not always possible, though. *Privacy* and *competitive advantages* are just two of the reasons why sharing of data is hard. However, all is not lost: According to Pineau et al. (2020), introducing a volunteer reproducibility checklist at the Neural Information Processing Systems Conference and the International Conference on Machine Learning increased the submissions with code to around seventy-five percent.

Anyway, the AI community's focus on open data sets has produced results, at least when compared with sharing code, which has had less focus. Even in the age of open-source software, only eight percent of the papers shared code, compared with fifty-six percent who shared data, according to our study. Code repositories, such as GitHub, simplify sharing and they are used by most developers, and by AI researchers as well. Why are the numbers so low for open-source experiments when compared with open data?

For most of the experiments that are conducted in AI research, everything that is needed to run the experiment is available on computers. In theory, this should make reproducibility much easier. However, running experiments completely on computers does not solve everything when it comes to reproducibility. Henderson et al. (2018) discuss the problems with reproducing results in deep-reinforcement learning. These relate to hyperparameters, random seeds, and even which implementation of a baseline algorithm is used for comparison. Nagarajan et al. (2019) show how hard it is to get a deterministic algorithm to run on a graphics processing unit. Even when succeeding on one computer, the results are completely different (but still deterministic) on another computer.

Floating-point calculations are a science unto themselves; they cause a lot of pain for those of us who depend on millions of them when running our experiments. Hong et al. (2013) found that changing operating systems, compilers, and hardware led to the same variations in weather simulations as changing the initial conditions. Even just obtaining code that was published by others is hard. Collberg and Proebsting (2016) tried to run the code of 402 experimental papers. They were successful in obtaining only 32.3 percent without first having to communicate with the authors; after communicating with the authors, the numbers increased to 48.3 percent.

To recover, we have to accept that we have a problem. This is the first step. Say after me: "The reproducibility crisis is real, even for AI."

Now, we can take the next step toward recovery.

References

Baker, M. 2016. 1,500 Scientists Lift the Lid on Reproducibility. *Nature* 533(7604): 452–5. doi.org/10.1038/533452a.

Bazire, M., and Brézillon, P. 2005. Understanding Context Before Using It. In *International and Interdisciplinary Conference on Modeling and Using Context,* 29–40. Berlin, Germany: Springer. doi.org/10.1007/11508373_3.

Cohen, P. R. 1995. *Empirical Methods for Artificial Intelligence*, Vol. 139. Cambridge, MA: The MIT Press.

Cohen, P. R., and Howe, A. E. 1988. How Evaluation Guides AI Research: The Message Still Counts More Than the Medium. *AI Magazine* 9(4): 35.

Collberg, C., and Proebsting, T. A. 2016. Repeatability in Computer Systems Research. *Communications of the ACM* 59(3): 62–9. doi.org/10.1145/2812803.

Gundersen, O. E., and Kjensmo, S. 2018. State of the Art: Reproducibility in Artificial Intelligence. In *Proceedings of the Thirty-Second Association for the Advancement of Artificial Intelligence (AAAI) Conference*. Palo Alto, CA: Association for the Advancement of Artificial Intelligence (AAAI) Press.

Henderson, P.; Islam, R.; Bachman, P.; Pineau, J.; Precup, D.; and Meger, D. 2018. Deep Reinforcement Learning that Matters. In *Proceedings of the Thirty-Second Association for the Advancement of Artificial Intelligence (AAAI) Conference*. Palo Alto, CA: Association for the Advancement of Artificial Intelligence (AAAI) Press.

Hong, S.-Y.; Koo, M.-S.; Jang, J.; Kim, J.-E.E.; Park, H.; Joh, M.-S.; Kang, J.-H.; and Oh, T.-J. 2013. An Evaluation of the Software System Dependency of a Global Atmospheric Model. *Monthly Weather Review* 141(11): 4165–72. doi.org/10.1175/MWR-D-12-00352.1.

Ioannidis, J. P. 2005. Why Most Published Research Findings Are False. *PLoS Medicine* 2(8): e124. doi.org/10.1371/journal.pmed.0020124.

Nagarajan, P.; Warnell, G.; and Stone, P. 2019. The Impact of Nondeterminism on Reproducibility in Deep Reinforcement Learning. Paper presented at the 2019 Association for the Advancement of Artificial Intelligence (AAAI) Workshop on Reproducible AI, Honolulu, Hawaii, January 27. openreview.net/pdf?id=S1e-OsZ4e7.

Pineau, J.; Vincent-Lamarre, P.; Sinha, K.; Larivière, V.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Larochelle, H. 2020. *Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program)*. arXiv:2003.12206. Ithaca, NY: Cornell University Library.

Odd Erik Gundersen (PhD, Norwegian University of Science and Technology) is the Chief AI Officer at the renewable energy company TrønderEnergi AS and is an adjunct associate professor at the Department of Computer Science at the Norwegian University of Science and Technology. Gundersen has applied AI in the industry, mostly for startups, since 2006. Currently, he is investigating how AI can be applied in the renewable energy sector and for driver training, and how AI can be made reproducible.