# Reproducibility, Inclusivity, and Open Science

Tyler Chen

Preface: This page is intended as a brief explanation of why I think scientists should care about open source projects. It's fairly short, so I only touch on some things very briefly. While the discussion is in the context of my field, I hope that the ideas I mention are applicable to other disciplines.

Science is built on the idea of reproducibility; that if two people do the same experiment they should get the same results.

Recently, there has been discussion about the concerning number of publications which are seemingly unreproducible. For the purpose of this discussion, I'm going to assume that reproducibility is important, and that at the very least, it is desirable that scientists be able to reproduce the work of their peers.

With these assumptions in mind, let's focus on how using open source software is important to reproducibility, both in the sense that it provides full transparency for those trying to replicate your work, and in the sense that it makes science more inclusive to those outside of academia.

A lot of papers relating to my research go something like this: "(~ task ~) is important for (~ grant keywords ~) reasons. Recently, because of (~ big data ~), there has been interest in fast algorithms for (~ task ~). Other papers have tried using (~ some method ~) quickly, however there are limitations with (~ some method ~). We propose an alternative to (~ some method ~) to address the limitations of (~ past paper ~)". A mathematical description of the new algorithm is presented, and then generally, some numerical experiments which demonstrate that the new method works and is possibly better than the old methods. Now, based on the descriptions in the paper (and any supplemental material), the readers should be able to reproduce any of the results claimed by the authors. In fact, since most publications are intended to be published in peer reviewed journals, ideally the results will be reproduced by someone reviewing the paper before publication (although this is rarely the case).

## So, what's the problem?

In practice, many pieces of the implementation used in the original paper will be dependent on the exact setup of the machine the code is run on. For example, if the original authors used BLACKBOX™ to write their code, the undoubtedly made calls to functions for lower level tasks such as matrix multiplication. Of course, this means that the results depend not only on the version of BLACKBOX™, but o which library they used for these low level functions, etc.

The results of the paper are probably technically reproducible. If I run the exact code the authors used, with the right version of BLACKBOX™, with the same low level libraries, on identical hardware, I will hopefully get the same output. However, even assuming that the authors posted their code for others to verify (which is relatively uncommon), there are still reproducibility issues. First, there is no way of knowing the exact algorithms used in the BLACKBOX™ builtin functions. Second, since BLACKBOX™ is needed to reproduce the experiment, only those with access to BLACKBOX™ are able to take part in the verification process. This immediately excludes entire demographics who may not be able to justify the purchase of BLACKBOX™; i.e. high school students (especially those from low income areas), people in developing countries, hobbyists and amateur scientists, etc.

The first issue has been discussed many times (see any MATLAB vs. Python argument thread). Generally, paid software such as MATLAB and Mathematica are well tested, and for the most part have well implemented algorithms. Even so, trust is required since nobody outside of those companies has access to their source code, and in my opinion, using this type of software is only detrimental to the scientific process, especially given that there are a wide range of open source alternatives.

The second issue is less talked about, and extends far beyond reproducibility. While there are many cases where using proprietary software may be unavoidable, there are a huge number of cases where these programs are used only because of familiarity or personal comfort. The fact that many research institutions provide free access to paid software only perpetrates the problem, since researchers can easily forget that not everyone has the same level of access to these programs as they do. This has the effect of excluding less privileged persons from the scientific process and furthering academia's (well deserved) reputation as an "ivory tower" inaccessible to those without sufficient resources.

## Why should academics in stem care?

If there are people with the right tools to reproduce the results of a paper, it's reasonable to ask is why should researchers care about making it easier for others to also reproduce your results. That is, does it really matter who replicates the results as long as someone is able to?

First, reducing the barriers to reproducibility within academia are necessary to create a more equitable research environment. Much of the science produced in the last 300 years has perpetuated white supremacy, and many notable scientists quite literally built their careers with the blood of people of color. Even today, academia is one of the most hostile workplace environments for women. As scientists and academics, we need to be mindful of how our actions impact others, especially those whose voices have traditionally been suppressed.

Many institutions (including R1 research schools) do not provide licences for programs such as MATLAB. As a result, students (and faculty) are frequently placed in the position of having to purchase software on their own in order to test the results of papers or work with collaborators (both important parts of the scientific process). The problem is especially concerning for grad students and non-tenure track faculty who are generally less stable financially than professors. Using open source software can help reduce this financial burden, and make it easier for anyone to get involved in research.

Second, it's not uncommon that people outside academia may want to use the code from research papers. For instance the findings of this paper, which outlined how to render an input image in the style of famous painters, have been widely used by hobbyists and non academics. There are now multiple subreddits and youtube channels devoted specifically to computer generated art, all helping bring new advancements in science to the mainstream. Public interests in a field can drive funding, and inspire new students to join the discipline.

## What can we do?

Broadly, I think the main takeaway is that we should be mindful of how our actions impact others, especially those who face barriers to participating in science which we might not have had to personally deal with.

When possible use open source alternatives; i.e. Python (with Numpy and Scipy) or Julia instead of MATLAB, Sympy or Sage instead of Mathematica, etc. By nature of being open source, the more people using these softwares, the better they become.

Write code which is accessible to others. This not only means having the code physically accessible (i.e. posted online), but making sure it is easily understandable and interpretable by others. Writing well documented code with a broader audience in mind, makes it easier for your research to reach a wide range of people.

For more reading on this topic, check out Lorena Barba's website. It has many good resources regarding good technical practices for reproducibility. One key takeaway is to "provide public access to scripts, runs, and results".