Irreproducibility in the Sciences

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Luke Ostyn

Spring, 2024

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

Kent Wayland, Department of Engineering and Society

Introduction

Science can be a tremendous force for good. When conducted properly, it is the one tool by which we can better understand the world. Moreover, on the back of this understanding, we produce new technologies which have the power to improve human existence. Of course, the operative word here is "properly." When science is in some manner abused and the theories that result do not track with reality, what ensues can be disastrous. A prime example is the eugenics movement which was much the fervor in the early 1900's. Scientists erroneously believed that a single gene was responsible for stupidity and, thereby, that preventing individuals in possession of such a gene from reproducing would significantly bolster the intelligence of the American population. Such a poisonous idea led to mass sterilization targeted specifically at underprivileged and minority communities. If we desire to limit such moral catastrophes, and of course we do, the question is what can be done to prevent the spurious theories that underpin them from taking root.

The prescription oft appealed to is that of the scientific method. But what happens when we attempt and follow its directions only for our results to go awry? Why are some findings reproducible while others fall flat on their face? The possibilities would seem to fall largely into one of two camps. Either the scientific method is flawed, or we fail to adhere to it in any sufficiently strict fashion. Through investigating this phenomenon of irreproducibility, I hope to unravel how such shortcomings manifest and to determine whether such flaws are corrigible.

Framework

While probing this idea of irreproducibility, it is integral to consider the bidirectional relationship between society and science. Society controls the content and manner of our study while science dictates our understanding of the world and thus how we exist within it. It follows

that the phenomenon of irreproducibility would be best interrogated under the auspices of the Co-production of Science and Social Order, a framework which, as explained by Jasanoff (2004), emphasizes "this self-conscious desire to avoid both social and technoscientific determinism in S&TS accounts of the world" (p. 20). Within this framework, we reject any notion that the flow of control between technology and society is unidirectional. Rather, society and scientific knowledge mutually establish and create the other (Swedlow, 2011). I will look at social influences in concert with potential flaws in the scientific method itself to determine what goes wrong while also observing societal impacts of these mistaken ideas.

Background

Reproducibility, as outlined by Karl Popper (2005), is necessarily a touchstone of science. In his words,"we do not take even our own observations quite seriously, or accept them as scientific observations, until we have repeated and tested them" (p. 23). Unfortunately, there exists a serious dearth of such reproducibility within much of the scientific landscape. One field where this phenomenon is a real epidemic is that of psychology; in an attempt to replicate one hundred of its most famous results, Nosek was only able to do so for thirty nine of them (2015). This is in line with a study conducted by Nuijten et al. which found that across several foundational psychology journals, greater than 50% of papers produced between 1985 and 2013 reported a minimum of one p-value which was unaligned with its associated test statistic and degrees of freedom (2015). Beyond psychology, this trend holds and is potentially even more profound in the field of biomedicine, where Begley and Ioannidis explain that likely over 75% of results are irreproducible and that approximately 85% of funding is wasted (2015). Confusingly, there seems to be a sort of cognitive dissonance at work within the greater scientific community when it comes to this issue. Although in a survey put out by nature over half of researchers

reported being unable to reproduce their own results, a significantly lower number, 31%, agreed with the idea that irreproducibility invalidates a result (Baker 2016).

Methods

In order to inquire into replication and the complex processes that underpin it, my approach will be that of a literature review. First off, I will seek to understand the prevailing thought as to the proximate cause of replication failures. Importantly, I will aim to examine a diversity of viewpoints and to ensure that the review is comprehensive in scope. To this end, I will primarily use keyword searching across various databases, including but not limited to Annual Reviews. The guiding intention will be evaluating whether there exists concern that the scientific method needs revision. In the case where the literature proves bullish as to the robustness of the scientific method, the next step will be in determining the social forces which would be capable of persuading researchers to slack in their adherence to such rigor as the scientific method demands. I will look at both internally and externally motivating factors. Why does a particular individual choose to forgo their commitment and what about the research apparatus permits them to do so? In either scenario, I will not merely look at why such phenomena occur, but go a step farther and attempt to explore potential solutions or at least ways to ameliorate such difficulties going forward. Finally, I will examine the changes brought about by the increasing awareness of replication concerns.

Literature Review

In general, there seems to be very little explicit criticism of the scientific method: at the very least, insofar as it has any role in diminishing reproducibility. A la Popper, the scientific method is what permits us the opportunity to falsify the theories we propose, an ability that without which the reliability of science would disappear. If we cannot say that a theory is

definitively wrong–that it has beyond all reasonable doubt been disproven–then there is nothing to separate those which track with reality from those that do not. Under the scientific method, we can under the auspices of scrutiny delineate between what is wrong and what is most certainly right. The lone piece of criticism which I will address is that levied by Castillo. He argues that the manner in which we document experiments is necessarily flawed: it is impossible for us to record everything, and as such, observations adjacent to what we are testing are systematically underrepresented in what we report (2013). Of course, this is more a shortcoming of the human researcher, but if it is truly insurmountable, might we not want to adopt another process? The issue is that I fail to envision any procedure which would not be beset by such an issue. If one were ever to arise, then perhaps a transition would be warranted. Until then, however, the scientific method appears to be the best tool in our arsenal and without either gross or widespread dissatisfaction.

It follows that if the problem does not lie with the methods we ostensibly utilize, then it must lie with the people who implement them. Rather than the scientific method having some debilitating inherent flaw, it must be the case that it is being improperly followed. Obviously, there exist clear cut situations where researchers go so far as to fabricate or falsify their data. One such instance comes from the lab of Jan Hendrik Schön, who between the years of 1998 and 2001 committed sixteen cases of this fabrication, the eventual punishment for which was his firing (Service, 2002). Arising more recently is the case of Stanford president, Marc Tessier-Lavigne, who was similarly fired after reports arose that he inculcated a research environment uncommitted to integrity (Kaiser, 2023). According to Fanelli, such acts of fabrication or falsification have been admitted to by 1.97% of scientists, a staggering number when we consider the sheer weight of research out there (2009). If we wish to mitigate such

infringements, we should first push for complete data transparency, or so thinks Simonsohn (2013). This policy complies with the prescription suggested by Fanelli et al., who suggest that somewhat surprisingly it is not merely a pressure to perform which impels data manipulation but instead limited social control: individuals will bend the rules not when they are forced to do so, but instead, when they are presented the opportunity. When we more strictly monitor the research process, results are less likely to be fudged. Naturally, when data is required to be released, the role of watchdog is more easily fulfilled. Kang and Hwang echo this sentiment–that we need to elevate social control measures–while also adding that education about data fabrication and falsification can provide serious dividends (2020).

That said, what we really want to unearth is what causes studies to go awry when there is not such blatant fraud. According to Bergley and Ioannidis, the fundamental cause is that scientists routinely fail in adhering to the scientific method. Rather than outright forgery, they bend the rules so as to elevate the likelihood of a surprising or flashy result and then, upon achieving such a result, rush to publish without sufficient confirmation testing. When tasked to explain their actions, the defense they employ is to argue that strictly observing the scientific method suppresses creativity. Of course, this is not the case. Creativity is not being suppressed when we ask for repeatability–for a shocking result to be tested before we are told to treat it as gospel (2015). One anecdote which supports this, as provided by Begley and Ellis, explains an interaction of Begley's with the author of a groundbreaking study. The author admits that they repeated a given experiment six times but only included in their paper one of the six trials. Naturally, that which they published was the only which happened to reject the null (2012). Clearly, there was no attempt to ensure that the result they found was repeatable. Instead, they went so far as to dismiss any trial which did not produce the desired result. While it is not

outright falsification, it is really not all that far from it. In discussing the situation, Schekman (2016) puts it quite succinctly: “this is not sloppiness, it is lack of character” (para. 5).

This search for a significant result, no matter the manner in which it is arrived at, is very similar to another frequent research technique. data dredging, perhaps more commonly called p-hacking. As outlined by Bruns and Ioannidis, the practice involves adding observations or adjusting the identity of the dependent variables until a significant observation–p less than 0.05–arises. Of course, frankness is basically nonexistent; there is no admission as to the removal of data or the shifting of goalposts, only a nicely buttoned up study with a pretty result. Very worrying is the evidence which suggests that the frequency of p-hacking only continues to spike. This is indicated by the increase in studies which report a p-value of between 0.041 and 0.049, just over the borderline for rejection of the null hypothesis (2016).

That said, it definitely is not character defects or unethical research methods alone which have created this crisis of irreproducibility. Rather, a significant part of the blame can be set at the feet of a system which rewards such behavior. Schekman in particular laments the poor incentives which exist among the world of publishing. Because the quality of a journal is judged by its impact factor–a metric related to the average number of citations a journal’s papers receive–journals are impelled to publish that which is groundbreaking and thus more likely to be cited. If the primary aim of the publisher is to publish that which is unexpected, then of course the researcher is going to oblige (2013). The problem is that when we aim for the spectacular, we demote veracity to an afterthought. This is not to say that we should distrust the research contained in various prestigious scholarly journals, much of which is trustworthy and of high quality, but that there certainly exist motivations, such as placement within an exclusive journal, that convince individuals to play loose and fast with their commitment to the scientific method.

While these problems are pervasive, they certainly should not prove insuperable. In fact, there is starting to become widespread efforts to address and mitigate them. One of the leading prescriptions is the same as that suggested for the problem of falsification: data transparency. It should become customary that the entirety of one's dataset be released upon publication. Critically, this not only applies to positive data but also negative data, that which does not favor the desired hypothesis ("Six problems"). For one, doing so allows for other scientists to better understand how the final results were obtained and makes it easier for them to repeat the process. For two, it helps ensure that the study was conducted above board, creating more trust in the conclusions reached. Beyond releasing data, there are also calls for all studies to be registered before they are ever undertaken ("Six problems"). This guarantees that the intention of the study stays consistent throughout and does not fluctuate with the data received.

In a similar vein is one of the solutions presented by Moody et al., which advances the idea of prepublication review. The idea takes data transparency a step farther and pushes for review of data and related code prior to release of the paper. This would allow errors to be caught and corrected prior to publication while also allowing the reviewer to identify potential instances of p-hacking. The clear drawbacks of such an extensive process are that it is necessarily arduous, time consuming, and thus expensive. Beyond prepublication review as a method to minimize p-hacking, the same team also suggests that we are perhaps over reliant on p-values and that deemphasizing them would reduce the effectiveness of the p-hacking to publication pipeline. Instead, they aver a more holistic approach to establishing significance, one that takes into consideration effect size and the confidence intervals related to the results produced (2022).

As for the journals with their poor incentives, Schekman (2013) offers his support for the relatively new institution of open-access journals. He describes them as "free for anybody to

read, and have no expensive subscriptions to promote. Born on the web, they can accept all papers that meet quality standards, with no artificial caps" (para. 7). When the journal doesn't have to pick and choose what they want to publish, then there is considerably less pressure on the researcher. There is no longer the specter hanging over them of some unidentified individual turning up their nose at an insignificant result. As such, they can report their findings for what they actually are without fear that publication will be withheld. Shekman warns us, however, that this only works if collectively we decide to not base the merit of a study on where it is published. Namely, funders and universities must be willing to extend grants and offer positions based on the actual research people have done and not the journal in which that research appeared (2013).

Finally, I want to hearken back to the suggestion made by Fanelli et al. for the adjacent problem of data manipulation. To refresh, they first explain that data manipulation is not strictly a result of a pressure to perform but that stems in large part from limited social control . Perhaps, with greater social control–oversight as far as adherence to the scientific method–scientists would be more likely to ensure rigor and compliance with established practices (2017). This is very much in line with the proposal for prepublication review. When we systematically monitor the science that is being performed and the scientists in question are aware that such monitoring is taking place it drastically reduces the freedom they feel to take liberties with their expression of the scientific method. The fact that they know about the oversight serves as a deterrent and the fact that there is oversight allows any real wrongdoing to be detected.

Although there are certainly still challenges related to replication and the solutions that we do have cannot simply be implemented overnight, we should recognize that acknowledging the issue is an important first step. In response to discovering our dearth of reproducible findings there have actually been serious and overwhelmingly positive changes within the scientific

community. In the view of Korbmacher et al., the landmark finding that only about 39% of psychological studies are reproducible, while slightly eroding the public's trust in science, will in the long term be incredibly beneficial. Already, they point out the rapid adoption of more robust research practices and the founding of grassroots organizations which aim to educate about the importance of reproducibility and promote research methods which encourage it. Moreover, they agree with Schekman on the value of open access journals and point to the replication crisis as the basis for their rise in popularity (2023).

Conclusion

Reproducibility is essential to the production of knowledge. In a nutshell, it is what allows us to separate the scientific wheat from the chaff, to discriminate between the ideas to toss out and the ideas to keep around. It should be unsettling then when we learn that vast swathes of the knowledge we possess fail this replication benchmark. It should make us want to investigate what is going wrong, to leave no stone unturned until we can definitively say where we erred and where amends will be made. Unfortunately, it does not appear like there is a singular, easily remedied cause. Instead, a multitude of different factors are responsible for the current state of affairs: a pressure to perform, a lack of supervision, little transparency, and poor incentives to name a few. That said, we have made strides and will continue to make strides. We have witnessed the invention of new journals, capable of eliminating many of the bad incentives at play, and rediscovered in large numbers our commitment to ethical and sustainable research practices. Moving forward, we can look to install prepublication review or modify the way in which we assess significance. While the replication crisis certainly did serve and in some ways continues to serve as a cold shower, I certainly do not find the future to be bleak. We are learning from this failure and becoming better because of it.

References

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*(7604), 452–454. https://doi.org/10.1038/533452a

Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, *483*(7391), 531–533. https://doi.org/10.1038/483531a

Begley, C. G., & Ioannidis, J. P. A. (2015). Reproducibility in science. *Circulation Research*, *116*(1), 116–126. https://doi.org/10.1161/circresaha.114.303819

Bruns, S. B., & Ioannidis, J. P. A. (2016). P-Curve and p-Hacking in observational research. *PLOS ONE*, *11*(2), e0149144. https://doi.org/10.1371/journal.pone.0149144

Castillo, M. (2013). The scientific method: a need for something better? *American Journal of Neuroradiology*, *34*(9), 1669–1671. https://doi.org/10.3174/ajnr.a3401

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of Survey Data. *PLoS ONE*, *4*(5). https://doi.org/10.1371/journal.pone.0005738

Fanelli, D., Costas, R., Fang, F. C., Casadevall, A., & Bik, E. M. (2017, April 12). *Why do scientists fabricate and falsify data? A matched-control analysis of papers containing problematic image duplications*. bioRxiv. https://doi.org/10.1101/126805

Kaiser, J. (2023, July 19). *Stanford president to step down despite probe exonerating him of research misconduct*. Science

https://www.science.org/content/article/stanford-president-to-step-down-despite-probe-exonerating-him-of-research-misconduct

Kang, E., & Hwang, H.-J. (2020). The Consequences of Data Fabrication and Falsification among Researchers. *Journal of Research and Publication Ethics*, *1*(2), 7–10. https://doi.org/https://doi.org/10.15722/jrpe.1.2.202009.7

Korbmacher, M., Azevedo, F., Pennington, C. R., Hartmann, H., Pownall, M., Schmidt, K., Elsherif, M. M., Breznau, N., Robertson, O., Kalandadze, T., Yu, S., Baker, B. J., O'Mahony, A., Olsnes, J. Ø.-., Shaw, J. J., Gjoneska, B., Yamada, Y., Röer, J. P., Murphy, J., . . . Evans, T. R. (2023). The replication crisis has led to positive structural, procedural, and community changes. *Communications Psychology*, *1*(1). https://doi.org/10.1038/s44271-023-00003-2

Moody, J., Keister, L. A., & Ramos, M. C. (2022). Reproducibility in the Social Sciences. *Annual Review of Sociology*, *48*(1), 65–85. https://doi.org/10.1146/annurev-soc-090221-035954

Nosek, B. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). https://doi.org/10.1126/science.aac4716

Nuijten, M. B., Hartgerink, C., Van Assen, M. A., Epskamp, S., & Wicherts, J. M. (2015). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*(4), 1205–1226. https://doi.org/10.3758/s13428-015-0664-2

Popper, Karl (2005). *The Logic of Scientific Discovery.* Taylor & Francis. http://philotextes.info/spip/IMG/pdf/popper-logic-scientific-discovery.pdf

Schekman, R. (2013, December 9). How journals like Nature, Cell and Science are damaging science. *The Guardian*. https://www.theguardian.com/commentisfree/2013/dec/09/how-journals-nature-science-cell-damage-science

Schekman, R. (2016). Introduction: The challenge of reproducibility. *Annual Review of Cell and Developmental Biology*, *32*(1). https://doi.org/10.1146/annurev-cb-32-100316-100001

Service, R. F. (2002, September 25). *Physicist Fired for Falsified Data*. Science. https://www.science.org/content/article/physicist-fired-falsified-data

Simonsohn, U. (2013). Just post it: The Lesson From Two Cases of Fabricated Data Detected by Statistics Alone. *Psychological Science*, *24*(10), 1875–1888. https://doi.org/10.1177/0956797613480366

Six factors affecting reproducibility in life science research and how to handle them. (n.d.). Nature. https://www.nature.com/articles/d42473-019-00004-y#ref-CR3

Swedlow, B. (2011). Cultural coproduction of four states of knowledge. *Science, Technology, & Human Values*, *37*(3), 151–179. https://doi.org/10.1177/0162243911405345