The natural selection of bad science

Nicolas Potie[1], Hélène Libion[1], Olga Ibanez[1] and Charlotte Nachtegael[1]

[1] Université Libre de Bruxelles, 1050 Brussels, Belgium

Abstract

Poor research habits are reinforced through natural selection of the methods that are the most prone to yielding positive results. Unfortunately, these methods are also those with high a false positive rate, leading to "bad science". This happens due to institutional incentives that favour the most productive researchers in terms of output volume, and positive results are far more frequently published than negative ones. The mechanism for selection requires no conscious cheating or strategizing from individual researchers. We consider a dynamical model with a population of labs with a given set of characteristics where each of them either dies or survives to the next generation depending on their "fitness". Those characteristics meet the requirements for natural selection since they are heritable and subject to mutations, and their variations have an effect on their reproduction. Individual labs research either novel or already tested hypotheses. We show here that methods yielding the highest amounts of false positive results are selected for, as they increase the output volume and thus the fitness of the labs. We also show that facilitating the publication of replication efforts does not prevent the selection of poor research methods. Moreover, setting a high penalty on those labs whose results have been proven not to be replicable failed to stop the propagation of poor methods. We conclude that a change at the institutional level is needed in order to solve the problem of the selection of bad science.

Introduction

In the present work we reproduce the results obtained by Smaldino and McElreath in their article "The Natural Selection of Bad Science" (2016).

Their work tests the hypothesis that the incentives leading contemporary scientific research actively encourage those methods that are most likely to yield false positive results, therefore promoting what is called "bad science". As a consequence, poor research methods spread through scientific communities through natural selection of the most "fitted" laboratories, which implies that their spreading requires no conscious cheating or abuse of the statistical methods, but is rather a consequence of the incentives that operate on the scientific community. They therefore conclude that a solution to the problem must come from a change in the incentives that lead scientific progress, rather than issuing guidelines that seek to further train the researchers on how to use the appropriate statistical methods.

Smaldino and McElreath support their hypothesis of the natural selection of bad science both empirically and analytically.

Firstly, they introduce several examples that might illustrate the increasing competition among scientists and the institutional incentives for publishing. For instance, they talk about the fact that the average number of publications of newly hired biologists has doubled in the past decade (Brischoux and Angelier, 2015), and the general increase in the rate at which articles are added to the scientific literature are good indicators. They also point out that the pressure that scientists experience to stand out from the crowd might explain the astonishing increase in the frequency, 25 times more in the past four decades, at which words that suggest innovation, such as "innovative", "ground-breaking", and "novel", appear in the literature. Moreover, literature has a tendency to publish much more positive results than null or negative ones (Su et al., 2013; Franco et al., 2014; Kay, 1990).

In addition, they note that the extensive use of metrics such as the h-index causes scientific success to be measured on the number of published papers only. These metrics typically fail to assess the quality of the scientific work done by researchers and solely focus on the sheer volume of output. Moreover, they have been proven to be subject to -either conscious or unconscious- exploitation and corruption on the part of researchers. This problem appears to be common to any social quantitative metrics, as stated by Campbell's law: "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor". This kind of pressure pushes scientists to publish as much as possible if they want to pursue a career in research, thusly promoting methods yielding as many positive results as positive to increase their chance to publish.

The natural selection of bad science follows the logic of a Darwinian evolutionary model: there are variations between

the different agents, here labs, these variations have consequences on their survival or reproduction, and last but not least, they are heritable. The fact that the characteristics of the laboratories are heritable can be argued in the following terms: apprentices learn methods from their mentors and peers, and also from successful role models in their research fields. Afterwards, these apprentices could spread methods learned in another laboratories where they are hired or even found their own laboratories adopting the same methods.

Methods

To validate the model of bad science, we implemented the same dynamical population model as described in the original paper (Smaldino and McElreath, 2016). Each lab possesses the following personal characteristics :

- *power* (W) : it is the ability to discern a true hypothesis as true ($Pr(+|T)$). The power is not only a reflection of the *statistical power*, which is the probability to reject a null hypothesis correctly, but reflects the whole inference process. This means that a high power method will simply declare all associations true, meaning you will not miss any true hypothesis, but will also yield a lot of false positives. Decreasing power means therefore also decreasing the chance of finding true hypothesis that might be difficult to detect due to weak or noisy signals.

- *effort* (e) : it is the tendency of the lab to use high quality statistical and experimental methods.

Effort and power are directly battling against each other. This means that putting more effort in the experiments means that we decrease the number of false positives, whereas we witness the opposite when increasing power. The resulting formula of the false positive rate α for the lab i is:

$$\alpha_i = \frac{W_i}{1 + (1 - W_i)e_i} \tag{1}$$

The number of real true hypothesis is determined by the *base rate* b. This means that the false discovery rate is based not only on the false positive rate of the lab, but also on the base rate. For example, if true hypotheses are rarer (low base rate), the false discovery rate will be higher.

Evolutionary model

Firstly, we create a population of N labs. Then, the dynamic model is divided in two steps : the science step and the reproducing step.

Science step During this step, each lab will undertake an investigation of an hypothesis with a probability of $h(e_i)$ at step i, according to the formula:

$$h(e_i) = 1 - \eta \log_{10} e_i \tag{2}$$

This probability is bound by the effort (e) that the lab invests to distinguish false hypotheses as false. Indeed, higher effort means uses of stricter methods, generally resulting in longer processes before obtaining significant and publishable results. The parameter η is a constant reflecting the impact of the effort on the lab publication rate.

If the lab undertakes a new investigation, it will select a hypothesis to investigate. The lab will choose either an hypothesis that has already been investigated with a probability r_i, meaning it will try to replicate the results obtained by another lab, or it will choose to investigate a novel hypothesis.

The hypothesis, new or replicated, yields either positive or negative results after investigation. A true hypothesis will yield a positive result with a probability based on the power of the lab, W_i. A false hypothesis will yield a positive result with a probability based on the false positive rate of the lab, α_i.

After investigation, the lab will try to publish their results. The assumption of this model is that all positive novel hypotheses will be published, as well as all replications of prior publications. The labs receive their pay-off for their publications and the eventual original lab author, if it has a replication published, is either rewarded in case of a successful replication, either punished in case of a unsuccessful one. These pay-offs will be at the origin of the power of the lab. In real life, pay-offs are in the form of prestige, money, funding,... and thusly promote the propagation of their methods in the scientific world.

The publications of the time step officially enter the world of publication after calculation of the false discovery rate of this time step. The publication world is limited in size at one million publications. If the size is exceeded, the older publications are removed until the appropriate size is attained.

Evolution step At the end of each time step, a random sample of d labs is chosen. The oldest lab is chosen to die. If several labs are equally as old, a lab is chosen randomly amongst them. This means that age is correlated to the fragility of a lab, but not perfectly correlated. Next, a new random sample of d labs is obtained and the lab with the highest pay-off is chosen for reproduction. This means that prestigious labs, labs with a high publication rate, will have greater chance to give birth to new labs. This is observed in real life, as post-doctorants from prestigious labs have higher chances to be chosen to found new labs.

The newly created lab adopts the characteristics of its parent lab with possible mutations on its *power*, *effort* and *replication rate* with probabilities of respectively μ_w,μ_e and μ_r. If there is a mutation, the change applied to the inherited character is drawn from a Gaussian distribution of mean zero and standard deviation σ_w, σ_e and σ_r for the power, effort and replication rate. If the parameters go beyond the advocated ranges, they are truncated to the maximum or minimum values.

Parameter	Definition	Values
N	number of labs	100
b	base rate of true hypotheses	0.1
r_0	initial replication rate for all labs	{0,0.01,0.25,0.5}
e_0	initial effort for all labs	75
W_0	initial power for all labs	0.8
η	influence of effort on productivity	0.2
V_n	pay-off for publishing new result	1
V_{R+}	pay-off for publishing positive replication	0.5
V_{R-}	pay-off for publishing negative replication	0.5
V_{O+}	reward for having published results replicated	0.1
V_{O-}	punishement for having published results failed to replicate	-100
d	number of labs sampled in evolution step	10
μ_r	probability of r mutation	{0,0.01}
μ_e	probability of e mutation	{0,0.01}
μ_w	probability of w mutation	{0,0.01}
σ_r	standard deviation of r mutation magnitude	0.01
σ_e	standard deviation of e mutation magnitude	1
σ_w	standard deviation of w mutation magnitude	0.01

Table 1: Global model parameters

Simulation runs The parameters used for the simulations can be found in Table 1. We averaged the results of 5 runs with the data collected every 2000 time steps.

Non-evolutionary model

To study the distribution of the total pay-off amongst the different kinds of laboratories with different replication rates, a non-evolutionary model was created.

First, we create a population of N labs, with 50% considered as high effort labs with effort $e = 75$ and thusly an investigation rate of $h = 0.625$ and a false positive rate of $\alpha = 0.05$, and the others labelled as low effort labs with effort $e = 15$ and thusly an investigation rate $h = 0.765$ and a false positive rate $\alpha = 0.2$.

The labs are allowed to publish only novel hypotheses for the first ten time steps to set up a baseline literature, then for one hundred time steps they are allowed to either try to replicate the results of prior publications with a probability r, otherwise explore a novel hypothesis.

The distribution of the pay-offs of the labs of 50 runs are then compared between the high and low effort labs.

Results

No replication attempt

We first tested the dynamical model by studying the model without replication attempts of prior publications ($r_0 = 0$). The power is the only characteristic allowed to mutate during the evolution step ($\mu_w = 0.01, \mu_r = \mu_e = 0$). We considered also here a maximum productivity rate ($h = 1$). Each run was composed of 100,000 time steps and the results were averaged over 5 runs.

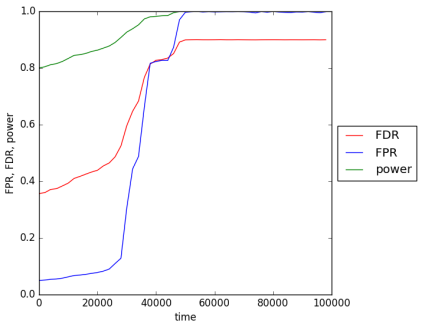


Figure 1: Evolution of the power W, false discovery rate (FDR) and false positive rate (α, FPR) with mutation of power allowed.

We can see that the false positive rate is increasing as the power is increasing. Similarly, as the false positive rate increases, we obtain an increasing false rate discovery in the published hypotheses. This obviously means that the laboratories chosen for reproducing were those that published, meaning those with high power as they have higher chance to yield positive results whatever the nature of the hypothesis (true or false) (Fig 1). This reflects the fact that it is labs with methods yielding a lot of positive results that have the highest chances to be prestigious and give birth to new laboratories, via post-doctorants becoming professors, or even graduate students that learned methods in the lab and spread them around. However, this case is completely unrealistic, as it is impossible to have methods that conclude that all the hypotheses are true ($W = 1$).

Another way to obtain positive results, and consequently increasing the publication rate, would be to reduce the effort of the lab. The result would be an increase of the productivity (h) of the lab, as well as its false positive rate. For the second simulation, the only characteristic allowed to mutate was the effort ($\mu_e = 0.01, \mu_w = \mu_e = 0$). Each run was composed of 1,000,000 time steps and the results were averaged on 5 runs.

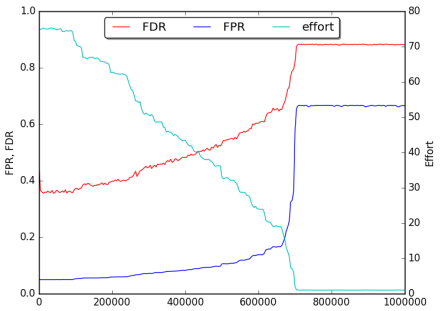


Figure 2: Evolution of the effort e, false rate discovery and false positive rate (α) with mutation of effort allowed.

We observed that the labs selected for reproduction are the ones with decreasing effort, which is translated by a gradual decrease over time until almost no effort is provided into high quality methods (Fig 2). The decrease in effort means that we have higher productivity, as putting effort in investigation means that it takes some time to publish, which could put the lab in a bad position in this case as publishing is the only way to gain pay-off. The decrease in effort means also that the false positive rate is increasing, guaranteeing the increase of the publication rate, as only positive results are published. The only reason why the false positive rate is not at the maximum is because some hypotheses are actually true.

Effect of replication

For the next simulation, we introduce the concept of replication. Replication allows a degree of control on the quality of the publications. When a lab fails to replicate the result of a prior publication, the original author is harshly punished. However, a successful replication will also reward the original author. An interesting thing to note is that a true publication could also yield a failed attempt to replicate its results, making the replication a dangerous tool to use for punishing labs, as they could also punish by mistake.

We first simulated each run where we allowed the mutation of both the effort and the replication rate during the evolution step ($\mu_r = \mu_e = 0.01, \mu_w = 0$), with an initial replication rate (r_0) for all the labs of 1%. We ran only a run of 1,000,000 time steps.

We can observe that the increase of replication rate is positively selected as it brings some pay-off to the lab, although not as much as the publication of a novel hypothesis does. However, the effort is still decreasing despite the punishment of a failed replication of prior results, even thought the probability of publishing an unsuccessful replication is increased by the false positive rate. This is because increasing false positive rate leads to more positive results, resulting in publication and consequently a higher pay-off than with replication. This means that a decrease of the effort allows to publish more, which is more interesting than replication (Fig 3). This is why it is more interesting to select labs with lower effort than labs with higher replication rate.

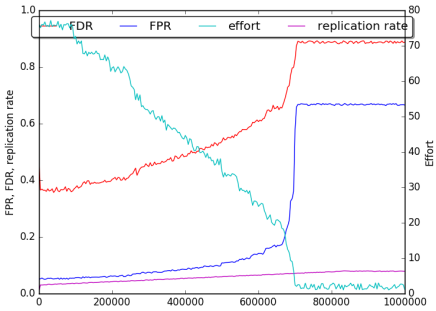


Figure 3: Evolution of the effort e, replication rate, false rate discovery and false positive rate (α) with mutation of effort and replication rate allowed.

We want to see if higher replication rates will have an effect on the evolution of the effort. For this simulation, we allowed only the mutation of the effort during the evolution step ($\mu_e = 0.01, \mu_w = \mu_e = 0$). We ran three simulations only for one run of 200,000 time steps with three different replications rates : 0, 25 and 50%.

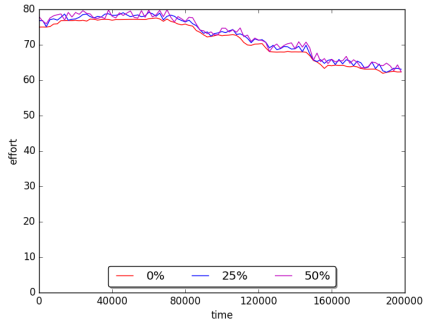


Figure 4: Evolution of the effort e, with different replication rates.

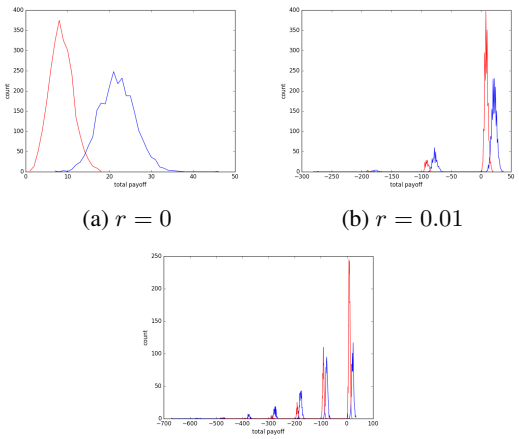
The replication rate does not seem to have any effect on the evolution of effort towards low rate (Fig 4). This could mean that even punishment is not able to fight the sheer number of publications a low effort lab could produce.

Why is replication not enough ?

Seeing no effect on the evolution of effort despite the fact that half of the time a replication would be attempted, we wanted to study what happens for the high effort labs and the low effort labs with different replication rates to understand

how low effort labs survives despite the highly punitive pay-off of failed replication.

We used here a non-dynamical model where we have two populations of N/2 labs : one with high effort ($e = 75$) and one with low effort ($e = 15$). We ran three simulations with different replication rates : 0, 1 and 5%.



(a) $r = 0$ (b) $r = 0.01$

(c) $r = 0.05$

Figure 5: Distribution of the pay-offs for the low effort labs (blue) and high effort labs (red) for three different replication rates in a non-evolutionary model.

Without replication, we can see that low effort labs have much higher pay-offs than the high effort ones. With the increasing of replication rates, we can see that a lot of low effort labs were punished for publishing false results, however despite the fact that high effort labs had a higher mean pay-off, quite a lot of low effort labs escaped the punishments and still had a high pay-off (Fig 5). Another observation is that with high replication rates, we witness labs with really low pay-off, far in the negatives.

Discussion

We proved with simulations of an evolutionary model of labs investigating, replicating and publishing hypotheses, and then reproducing according to the lab with the highest pay-off in a sample d of labs, that our current system promotes methods with high false positive rate.

We obtained similar results as the original paper. Our main differences reside in the fact that we could not run the full one million steps for all the simulations needing it, nor could we always do an average over 50 runs, due to the sheer computation time needed for this. Despite that fact, our observations were in sync with theirs.

As long as a scientific is judged only by the sheer number of their publications, the model proves that methods yielding a high rate of positive results will be spread and promoted across all the laboratories. Where does the problem lie in ?

First of all, one big barrier for high quality publications is the fact that negative or null hypotheses have less chance to be published than positive ones. Actually, authors generally will not even try to publish results if they are not positive (Franco et al., 2014; Su et al., 2013). This bias means that a lot of information is never communicated to the scientific community, resulting in pointless research into already explored subjects but without any literature (Kay, 1990). Moreover, publishing positive result weakly impacts with increasing the impact of the journal as people try to reproduce the results and thusly cites (positively or negatively) the publication, resulting in an increase of the impact of the journal(Munafo et al., 2009).

Smaldino and McElreath (2016) were not the only ones to raise alarm for this problem. Button (2016) shared their concerns about the spreading of high false positive yielding methods due to the pressure of PhD and graduates students finding jobs only according to the number of their publications (Schillebeeckx et al., 2013). She proposes several solutions to fight this reality.

A first proposition would be to introduce more transparency in the whole investigation process : preregistration process of the protocols, reporting faithfully all the results and integrate statisticians in the scientific team so that high quality statistical analysis could be conducted.

Another would be to promote higher effort in the lab : to train the students so that they conducted rigorous studies and analysed their results as unbiased as possible. However, this kind of behaviour leads to less publications for the students, which means that they would not be hired if their employer used the number of their publications as a reflection of their skills.

We propose other solutions. First, scientists should not be judged on the number of their publications, but also the quality of their work. Of course, this kind of information is not always easily found, but a lot of researchers in the literature are already looking for alternatives to this measure (Wouters et al., 2012). Second, the bias against null or negative results should be reduced to allow better communication of any kind of results relevant for the scientific community. This, linked to the transparency of the investigation process proposed by Button (2016), should allow the publication of higher quality papers and not only the positive kind.

We agree with our original paper with the fact that punishing papers for failed replication is not a good method. Unfortunately, sometimes important discoveries could look impossible and failed replication could also be a false negative. More importantly, replication is also sometimes difficult or impossible, such as with clinical trial,... However, we propose a compromise where papers with a certain amount of failed replications should be retracted or at least reviewed again to check for inconsistencies.

Propositions of extensions of the evolutionary model

Prestige of the journal We want to study the influence on the impact factor of a journal on the model. We would implement a prestigious journal where the lab could try to publish their positive results, otherwise they publish in a regular journal. We base the probability to publish in a prestigious journal on the fact that prestigious labs have a higher chance to have their papers accepted either based on their reputation, the fact that they often are part of peer-review groups or that they know the editor, which we quite simply reduced to the lab with high total pay-off. The probability is calculated according to the following formula :

$$p = \frac{1}{1 + 5e^{-0.05x}} \tag{3}$$

This allows that labs with small pay-off still have a chance to publish in the prestigious journal, but that probability grows slowly in function of the pay-off, being asymptotically close to 1 (Fig 6).

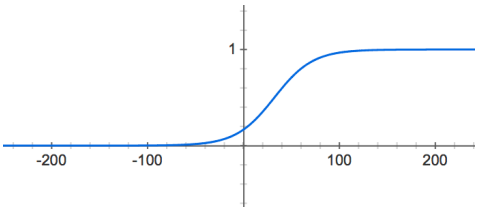


Figure 6: Representation of the evolution of the probability to publish in a prestigious journal in function of the total pay-off of the lab.

Publishing in a prestigious journal would gather a higher pay-off than with a regular one ($pay - off = 2$), as well as a higher reward in case of successful replication ($reward = 0.5$). However, they would receive a much harsher punishment for failed replication ($punishement = -200$). The simulation should be run with an initial replication rate of 1% and should allow mutation of the effort and the replication rate.

We think that the factor prestigious of the journal would influence the speed with which the effort would evolve towards zero and the false positive rate would increase. As high pay-off lab rimes with low effort or high power, these labs would have better chance to publish in prestigious journal and consequently gain even more pay-off.

Nowadays, journal with high impact factor are viewed rather harshly (Lariviere et al., 2016). We wish to see if we could determine the role of high impact journals in the natural selection of bad science.

Age of the paper Our idea is that an older paper will have had a potentially huge impact on the literature and discoveries of nowadays. Discovering that this paper is in fact false would have direct consequences on the consequent studies done based on it. On the other side, positive replication of this paper would have lost interest after a few years as it would be assumed that after all this time the result should be true. We would observe the opposite situation with a young paper, where a positive replication would back it up and give it more credit, whereas a failed replication would only minimally impact it as its influence on current discoveries is smaller too.

The influence of a paper is often measured according to its age and its number of citations (Bloching and Heinzl, 2013), but this view is quite close to the problem of natural selection of bad science. Another way to measure the influence of a paper would be to take into account the nature of the citations, so to avoid the negative ones and really measure when the paper had an impact on another publication.

This extension would work on a function of the reward and punishment of replication according to the age of the paper, where young papers receive high reward an low punishment, whereas old paper receive low reward but high punishment.

Cooperation social games The model could also be constructed as a spatial cooperative game. We would like to build a network where we would begin with different rates of high effort and low effort labs in the total population and observe if the construction of a network between high effort lab would allow high effort labs to survive and thrive.

Cooperation between labs would result in an increase in productivity, as it is well known that collaboration promotes publication rate (Nabout et al., 2015).

We could imagine that low effort labs earn a low pay-off as well as the high effort lab when connected to each other, as their differences in methods and politics would not allow them to work in sync on a common project. However, high effort labs working together would earn more than low effort labs together. The labs could choose to pursue high effort or low effort methods according to the replicator rule where the probability to adopt the behaviour of a randomly chosen neighbour is increasing in function of their total pay-off at this time step. We chose this replication rule to imitate the fact that a lab takes time and think about their decision before changing their whole lab's behaviour and rules.

We would not take into account the replication or the power, we would just study the impact of what a network could do for the survival of high quality experimental and statistical methods.

Conclusion

The natural selection of bad science is kept alive by the bad incentives of the journal and scientific institutions. We would need not only to change the behaviour of the laboratories so that, despite the time it can take and the negative results, they would keep high quality experimental and sta-

tistical methods, but also those of the individuals that judge scientists' worth according to their number of publications, as well as the journals that mainly publish positive results. These changes will not happen overnight and need the whole community to work towards it. If nobody does anything, we risk spreading bad methods with high false positive rate, despite the honesty and integrity of the researchers.

References

Bloching, P. A. and Heinzl, H. (2013). Assessing the scientific relevance of a single publication over time.

Brischoux, F. and Angelier, F. (2015). Academia's never-ending selection for productivity. *Scientometrics*, 103(1):333–336.

Button, K. S. (2016). Statistical rigor and the perils of chance. *eneuro*, 3(4).

Franco, A., Malhotra, N., and Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345(6203):1502–1505.

Kay, D. (1990). The existence of publication bias and risk factors for its occurrence. *JAMA*, 263(10):1385–1389.

Lariviere, V., Kiermer, V., MacCallum, C. J., McNutt, M., Patterson, M., Pulverer, B., Swaminathan, S., Taylor, S., and Curry, S. (2016). A simple proposal for the publication of journal citation distributions. *bioRxiv*.

Munafo, M. R., Stothart, G., and Flint, J. (2009). Bias in genetic association studies and impact factor. *Mol Psychiatry*, 14(2):119–120.

Nabout, J. C., Parreira, M. R., Teresa, F. B., Carneiro, F. M., da Cunha, H. F., de Souza Ondei, L., Caramori, S. S., and Soares, T. N. (2015). Publish (in a group) or perish (alone): the trend from single- to multi-authorship in biological papers. *Scientometrics*, 102(1):357–364.

Schillebeeckx, M., Maricque, B., and Lewis, C. (2013). The missing piece to changing the university culture. *Nat Biotech*, 31(10):938–941.

Smaldino, P. E. and McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, 3(9).

Su, P., Su, J. M., and Montoro, J. B. (2013). Positive outcomes influence the rate and time to publication, but not the impact factor of publications of clinical trial results. *PLOS ONE*, 8(1):1–8.

Wouters, P., Costas, R., and SURFfoundation (2012). *Users, Narcissism and Control: Tracking the Impact of Scholarly Publications in the 21st Century*. SURFfoundation.