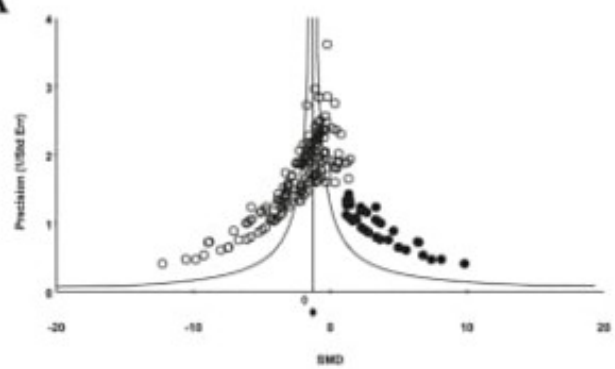


A



Studies and Statistics

1. [Beware The Man Of One Study](#)
2. [Debunked And Well-Refuted](#)
3. [Noisy Poll Results And Reptilian Muslim Climatologists from Mars](#)
4. [Two Dark Side Statistics Papers](#)
5. [The Control Group Is Out Of Control](#)
6. [The Cowpox of Doubt](#)
7. [How Common Are Science Failures?](#)
8. [Learning To Love Scientific Consensus](#)
9. [My IRB Nightmare](#)
10. [The Study of Anglophysics](#)

Beware The Man Of One Study

Aquinas famously [said](#): beware the man of one book. I would add: beware the man of one study.

For example, take medical research. Suppose a certain drug is weakly effective against a certain disease. After a few years, a bunch of different research groups have gotten their hands on it and done all sorts of different studies. In the best case scenario the average study will find the true result – that it's weakly effective.

But there will also be random noise caused by inevitable variation and by some of the experiments being better quality than others. In the end, we might expect something looking kind of like a bell curve. The peak will be at “weakly effective”, but there will be a few studies to either side. Something like this:



We see that the peak of the curve is somewhere to the right of neutral – ie weakly effective – and that there are about 15 studies that find this correct result.

But there are also about 5 studies that find that the drug is very good, and 5 studies missing the sign entirely and finding that the drug is actively bad. There's even 1 study finding that the drug is very bad, maybe seriously dangerous.

This is before we get into fraud or statistical malpractice. I'm saying this is what's going to happen just by normal variation in experimental design. As we increase experimental rigor, the bell curve might get squashed horizontally, but there will still be a bell curve.

In practice it's worse than this, because this is assuming everyone is investigating exactly the same question.

Suppose that the graph is titled “Effectiveness Of This Drug In Treating Bipolar Disorder”.

But maybe the drug is more effective in bipolar i than in bipolar ii (Depakote, for example)

Or maybe the drug is very effective against bipolar mania, but much less effective against bipolar depression (Depakote again).

Or maybe the drug is a good acute antimanic agent, but very poor at maintenance treatment (let's stick with Depakote).

If you have a graph titled “Effectiveness Of Depakote In Treating Bipolar Disorder” plotting studies from “Very Bad” to “Very Good” – and you stick all the studies – maintenance, manic, depressive, bipolar i, bipolar ii – on the graph, then you're going to end running the gamut from “very bad” to “very good” even before you factor in noise and even before even before you factor in bias and poor experimental design.

So here's why you should beware the man of one study.

If you go to your better class of alternative medicine websites, they don't tell you “Studies are a logocentric phallocentric tool of Western medicine and the Big Pharma

conspiracy.”

They tell you “medical science has proved that this drug is terrible, but ignorant doctors are pushing it on you anyway. Look, here’s a study by a reputable institution proving that the drug is not only ineffective, but harmful.”

And the study will exist, and the authors will be prestigious scientists, and it will probably be about as rigorous and well-done as any other study.

And then a lot of people raised on [the idea](#) that some things have Evidence and other things have No Evidence think *holy s**t, they’re right!*

On the other hand, your doctor isn’t going to a sketchy alternative medicine website. She’s examining the entire literature and extracting careful and well-informed conclusions from...

Haha, just kidding. She’s going to a luncheon at a really nice restaurant sponsored by a pharmaceutical company, which assures her that they would *never* take advantage of such an opportunity to shill their drug, they just want to raise awareness of the latest study. And the latest study shows that their drug is great! Super great! And your doctor nods along, because the authors of the study are prestigious scientists, and it’s about as rigorous and well-done as any other study.

But obviously the pharmaceutical company has selected one of the studies from the “very good” end of the bell curve.

And I called this “Beware The Man of One Study”, but it’s easy to see that in the little diagram there are like three or four studies showing that the drug is “very good”, so if your doctor is a little skeptical, the pharmaceutical company can say “You are right to be skeptical, one study doesn’t prove anything, but look – here’s another group that finds the same thing, here’s yet another group that finds the same thing, and here’s a replication that confirms both of them.”

And even though it looks like in our example the sketchy alternative medicine website only has one “very bad” study to go off of, they could easily supplement it with a bunch of merely “bad” studies. Or they could add all of those studies about slightly different things. Depakote is ineffective at treating bipolar depression. Depakote is ineffective at maintenance bipolar therapy. Depakote is ineffective at bipolar ii.

So just sum it up as “Smith et al 1987 found the drug ineffective, yet doctors continue to prescribe it anyway”. Even if you hunt down the original study (which no one does), Smith et al won’t say specifically “Do remember that this study is only looking at bipolar maintenance, which is a different topic from bipolar acute antimanic treatment, and we’re not saying anything about that.” It will just be titled something like “Depakote fails to separate from placebo in six month trial of 91 patients” and trust that the responsible professionals reading it are well aware of the difference between acute and maintenance treatments (hahahahaha).

So it’s not so much “beware the man of one study” as “beware the man of any number of studies less than a relatively complete and not-cherry-picked survey of the research”.

II.

I think medical science is still pretty healthy, and that the consensus of doctors and researchers is more-or-less right on most controversial medical issues.

(it's the *uncontroversial* ones you have to worry about)

Politics doesn't have this protection.

Like, take the minimum wage question (please). We all know about the Krueger and Card [study](#) in New Jersey that found no evidence that high minimum wages hurt the economy. We probably also know the counterclaims that it was [completely debunked](#) as despicable dishonest statistical malpractice. Maybe some of us know Card and Krueger wrote a [pretty convincing rebuttal](#) of those claims. Or that a bunch of large and methodologically advanced studies have come out since then, some finding no effect like [Dube](#), others finding strong effects like [Rubinstein](#) and [Wither](#). These are just examples; there are at least dozens and probably hundreds of studies on both sides.

But we can solve this with meta-analyses and systematic reviews, right?

Depends which one you want. Do you go with [this meta-analysis](#) of fourteen studies that shows that any presumed negative effect of high minimum wages is likely publication bias? With [this meta-analysis](#) of sixty-four studies that finds the same thing and discovers no effect of minimum wage after correcting for the problem? Or how about [this meta-analysis](#) of fifty-five countries that does find effects in most of them? Maybe you prefer [this systematic review](#) of a hundred or so studies that finds strong and consistent effects?

Can we trust news sources, think tanks, econblogs, and other institutions to sum up the state of the evidence?

CNN [claims that](#) 85% of credible studies have shown the minimum wage causes job loss. But [raisetheminimumwage.com](#) [declares that](#) "two decades of rigorous economic research have found that raising the minimum wage does not result in job loss... researchers and businesses alike agree today that the weight of the evidence shows no reduction in employment resulting from minimum wage increases." Modeled Behavior [says](#) "the majority of the new minimum wage research supports the hypothesis that the minimum wage increases unemployment." The Center for Budget and Policy Priorities [says](#) "The common claim that raising the minimum wage reduces employment for low-wage workers is one of the most extensively studied issues in empirical economics. The weight of the evidence is that such impacts are small to none."

Okay, fine. What about economists? They seem like experts. What do they think?

Well, five hundred economists [signed](#) a letter to policy makers saying that the science of economics shows increasing the minimum wage would be a bad idea. That sounds like a promising consensus...

..except that six hundred economists [signed](#) a letter to policy makers saying that the science of economics shows increasing the minimum wage would be a *good* idea. (h/t [Greg Mankiw](#))

Fine then. Let's do a formal survey of economists. Now what?

[raisetheminimumwage.com](#), an unbiased source if ever there was one, confidently tells us that "indicative is a 2013 survey by the University of Chicago's Booth School of Business in which leading economists agreed by a nearly 4 to 1 margin that the benefits of raising and indexing the minimum wage outweigh the costs."

But the Employment Policies Institute, which sounds like it's trying way too hard to sound like an unbiased source, [tells us that](#) "Over 73 percent of AEA labor economists believe that a significant increase will lead to employment losses and 68 percent think these employment losses fall disproportionately on the least skilled. Only 6 percent feel that minimum wage hikes are an efficient way to alleviate poverty."

So the whole thing is fiendishly complicated. But unless you look very very hard, you will never know that.

If you are a conservative, what you will find on the sites you trust will be something like this:

Economic theory has always shown that minimum wage increases decrease employment, but the Left has never been willing to accept this basic fact. In 1992, they trumpeted a single study by Card and Krueger that purported to show no negative effects from a minimum wage increase. This study was immediately debunked and found to be based on statistical malpractice and "massaging the numbers". Since then, dozens of studies have come out confirming what we knew all along – that a high minimum wage is economic suicide. Systematic reviews and meta-analyses (Neumark 2006, Boockman 2010) consistently show that an overwhelming majority of the research agrees on this fact – as do 73% of economists. That's why five hundred top economists recently signed a letter urging policy makers not to buy into discredited liberal minimum wage theories. Instead of listening to starry-eyed liberal woo, listen to the empirical evidence and an overwhelming majority of economists and oppose a raise in the minimum wage.

And if you are a leftist, what you will find on the sites you trust will be something like this:

People used to believe that the minimum wage decreased unemployment. But Card and Krueger's famous 1992 study exploded that conventional wisdom. Since then, the results have been replicated over fifty times, and further meta-analyses (Card and Krueger 1995, Dube 2010) have found no evidence of any effect. Leading economists agree by a 4 to 1 margin that the benefits of raising the minimum wage outweigh the costs, and that's why more than 600 of them have signed a petition telling the government to do exactly that. Instead of listening to conservative scare tactics based on long-debunked theories, listen to the empirical evidence and the overwhelming majority of economists and support a raise in the minimum wage.

Go ahead. [Google the issue and see what stuff comes up](#). If it doesn't quite match what I said above, it's usually because they can't even muster *that* level of scholarship. Half the sites just cite Card and Krueger and call it a day!

These sites with their long lists of studies and experts are super convincing. And half of them are wrong.

At some point in their education, most smart people usually learn not to credit arguments from authority. If someone says "Believe me about the minimum wage because I seem like a trustworthy guy," most of them will have at least one neuron in their head that says "I should ask for some evidence". If they're *really* smart, they'll use the magic words "peer-reviewed experimental studies."

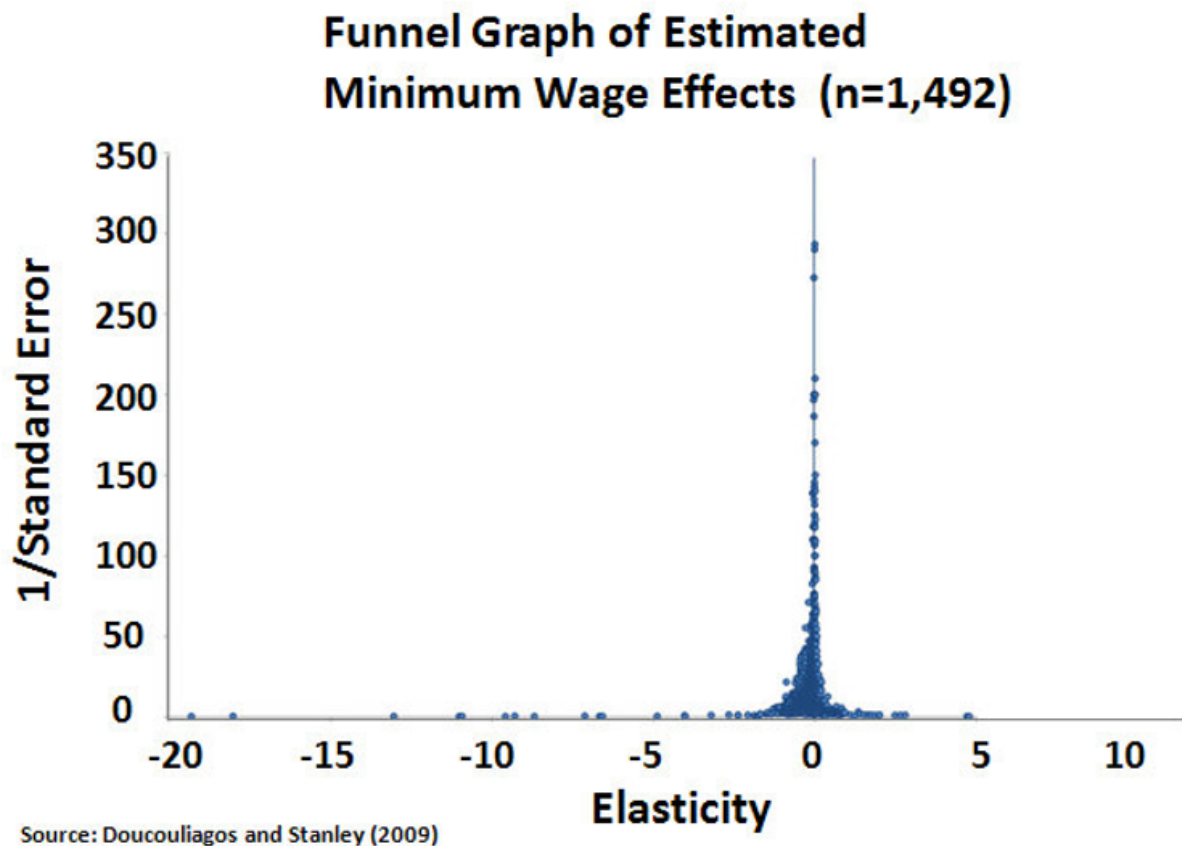
But I worry that most smart people have *not* learned that a list of dozens of studies, several meta-analyses, hundreds of experts, and expert surveys showing almost all academics support your thesis – can *still* be bullshit.

Which is too bad, because that's exactly what people who want to bamboozle an educated audience are going to use.

III.

I do not want to preach radical skepticism.

For example, on the minimum wage issue, I notice only one side has presented a funnel plot. A funnel plot is usually used to investigate publication bias, but it has another use as well – it's pretty much an exact presentation of the "bell curve" we talked about above.



This is more of a needle curve than a bell curve, but the point still stands. We see it's centered around 0, which means there's some evidence that's the real signal among all this noise. The bell skews more to left than to the right, which means more studies have found negative effects of the minimum wage than positive effects of the minimum wage. But since the bell curve is asymmetrical, we interpret that as *probably* publication bias. So all in all, I think there's at least some evidence that the liberals are right on this one.

Unless, of course, someone has realized that I've wised up to the studies and meta-analyses and expert surveys, and figured out a way to hack *funnel plots*, which I am totally not ruling out.

(okay, I *kind of* want to preach radical skepticism)

Also, I should probably mention that it's much more complicated than one side being right, and that the minimum wage probably works differently depending on what industry you're talking about, whether it's state wage or federal wage, whether it's a recession or a boom, whether we're talking about increasing from \$5 to \$6 or from \$20 to \$30, etc, etc, etc. There are eleven studies on that plot showing an effect even worse than -5, and very possibly they are all accurate for whatever subproblem they have chosen to study - much like the example with Depakote where it might be an effective antimanic but a terrible antidepressant.

(radical skepticism actually sounds a lot better than figuring this all out).

IV.

But the question remains: what happens when (like in most cases) you don't have a funnel plot?

I don't have a good positive answer. I do have several good *negative* answers.

Decrease your confidence about most things if you're not sure that you've investigated every piece of evidence.

Do not trust websites which are obviously biased (eg Free Republic, Daily Kos, Dr. Oz) when they tell you they're going to give you "the state of the evidence" on a certain issue, even if the evidence seems very solid indeed. This goes double for any site that contains a list of "myths and facts about X", quadruple for any site that uses phrases like "ingroup member uses actual FACTS to DEMOLISH the outgroup's lies about Y", and octuple for RationalWiki.

Most important, even if someone gives you what seems like overwhelming evidence in favor of a certain point of view, don't trust it until you've done a simple Google search to see if the opposite side has equally overwhelming evidence.

Debunked And Well-Refuted

I.

As usual, I was insufficiently pessimistic.

I infer this from *The Federalist's* [article on campus rape](#):

A new report on sexual assault released today by the U.S. Department of Justice (DOJ) officially puts to bed the bogus statistic that one in five women on college campuses are victims of sexual assault. In fact, non-students are 25 percent more likely to be victims of sexual assault than students, according to the data. And the real number of assault victims is several orders of magnitude lower than one-in-five.

The article compares the older Campus Sexual Assault Survey (which found 14-20% of women were raped since entering college) to the just-released National Crime Victimization Survey (which found that 0.6% of female college students are raped per year). They write "Instead of 1 in 5, the real number is 0.03 in 5."

So the first thing I will mock *The Federalist* for doing is directly comparing per year sexual assault rates to per college career sexual assault rates, whereas obviously these are very different things. You can't *quite* just divide the latter by four to get the former, but that's going to work a heck of a lot better than *not* doing it, so let's estimate the real discrepancy as more like 0.5% per year versus 5% per year.

But I can't get too mad at them yet, because that's still a pretty big discrepancy.

However, faced with this discrepancy a reasonable person might say "Hmm, we have two different studies that say two different things. I wonder what's going on here and which study we should believe?"

The Federalist staff said "Ha! There's an old study with findings we didn't like, but now there's a new study with different findings we *do* like. So the old study is debunked!"

II.

My last essay, [Beware The Man Of One Study](#), noted that one thing partisans do to justify their bias is selectively acknowledge studies from only one side of a complicated literature.

The reason it was insufficiently pessimistic is that there are also people like the *Federalist* staff, who acknowledge the existence of opposing studies, but only with the adjective "debunked" in front of them. By "debunked" they usually mean one of two things:

1. Someone on my side published a study later that found something else
2. Someone on my side accused it of having methodological flaws

Since the *Federalist* has so amply demonstrated the first failure mode, let me say a little more about the second. Did you know that *anyone* with a keyboard can just *type up* any of the following things?

- "That study is a piece of garbage that's not worth the paper it's written on."
- "People in the know dismissed that study years ago."
- "Nobody in the field takes that study seriously."
- "That study uses methods that are laughable to anybody who knows statistics."
- "All the other research that has come out since discredits that study."

They can say these things *whether they are true or not*. I'm kind of harping on this point, but it's because it's something *I* didn't realize until much later than I should have.

There are many "questions" that are pretty much settled - evolution, global warming, homeopathy. But taking these as representative [closes your mind](#) and gives you a skewed picture of academia. On many issues, academics are just as divided as anyone else, and their arguments can be just as acrimonious as anyone else's. The arguments usually take the form of one side publishing a study, the other side ripping the study apart and publishing their own study which they say is better, and the first side ripping the second study apart and arguing that their study was better all along.

Every study has flaws. No study has perfect methodology. If you like a study, you can say that it did the best it could on a difficult research area and has improved upon even-worse predecessor studies. If you don't like a study, you can say "LOOK AT THESE FLAWS THESE PEOPLE ARE IDIOTS THE CONCLUSION IS COMPLETELY INVALID". All you need to do is make enough [isolated demands for rigor](#) against anything you disagree with.

And so if the first level of confirmation bias is believing every study that supports your views, the second layer of confirmation bias is believing every supposed refutation that supports your views.

There are certainly things that have been "well-refuted" and "debunked". Andrew Wakefield's study purporting to prove that vaccines cause autism is a pretty good example. But you will notice that it had multiple failed replications, journals published reports showing he falsified data, the study's co-authors retracted their support, the journal it was published in retracted it and issued an apology, the General Medical Council convicted Wakefield of sixteen counts of misconduct, and Wakefield was stripped of his medical license and barred from practicing medicine ever again in the UK. The *British Medical Journal*, one of the best-respected medical journals in the world, published an editorial concluding:

Clear evidence of falsification of data should now close the door on this damaging vaccine scare ... Who perpetrated this fraud? There is no doubt that it was Wakefield. Is it possible that he was wrong, but not dishonest: that he was so incompetent that he was unable to fairly describe the project, or to report even one of the 12 children's cases accurately? No.

Wakefield's study has been "refuted". The rape study has been "argued against".

III.

I saw this same dynamic at work the other day, looking through the minimum wage literature.

The primordial titanomachy of the minimum wage literature goes like this. In 1994, two guys named Card and Krueger published a study showing the minimum wage had if anything positive effects on New Jersey restaurants, convincing many people that

minimum wages were good. In 1996, two guys named Neumark and Wascher reanalyzed the New Jersey data using a different source and found that it showed the minimum wage had very bad effects on New Jersey restaurants. In 2000, Card and Krueger responded, saying that their analysis was better than Neumark and Wascher's re-analysis, and also they had done a re-analysis of their own which confirmed their original position.

Let's see how conservative sites present this picture:

"The support for this assertion is the oft-cited 1994 study by Card and Krueger showing a positive correlation between an increased minimum wage and employment in New Jersey. Many others have thoroughly debunked this study." ([source](#))

"I was under the impression that the original study done by Card and Krueger had been thoroughly debunked by Michigan State University economist David Neumark and William Wascher" ([source](#))

"The study ... by Card and Krueger has been debunked by several different people several different times. When other researchers re-evaluated the study, they found that data collected using those records 'lead to the opposite conclusion from that reached by' Card and Krueger." ([source](#))

"It was only a short time before the fantastic Card-Krueger findings were challenged and debunked by several subsequent studies...in 1995, economists David Neumark and David Wascher used actual payroll records (instead of survey data used by Card and Krueger) and published their results in an NBER paper with an amazing finding: Demand curves for unskilled labor really do slope downward, confirming 200 years of economic theory and mountains of empirical evidence" ([source](#))

And now let's look at how lefty sites present this picture:

"...a long-debunked paper [by Neumark and Wascher]" ([source](#))

"Note that your Mises heroes, Neumark and Wascher are roundly debunked." ([source](#))

"Neumark's living wage and minimum wage research have been found to be seriously flawed...based on faulty methods which when corrected refute his conclusion." - ([source](#))

"...Neumark and Wascher, a study which Elizabeth Warren debunked in a Senate hearing" ([source](#))

So if you're conservative, Neumark and Wascher debunked Card and Krueger. But if you're liberal, Card and Krueger debunked Neumark and Wascher.

Both sides are no doubt very pleased with themselves. They're not men of one study. They look at *all* of the research - except of course the studies that have been "debunked" or "well-refuted". Why would you waste your time with *those*?

IV.

Once again, I'm not preaching radical skepticism.

First of all, some studies are *super-debunked*. Wakefield is a good example.

Second of all, some studies that don't quite meet Wakefield-level of awfulness are indeed really bad and need refuting. I don't think this is beyond the intellectual capacities of most people. I think in many cases it's easy to understand why a study is wrong, you should try to do that, and once you do it you can safely discount the results of the study.

I'm not against pointing out when you disagree with studies or think they're flawed. I'd be a giant hypocrite if I was.

But "debunked" and "refuted" aren't saying you disagree with a study. They're making arguments from authority. They're saying "the authority of the scientific community has come together and said this is a piece of crap that doesn't count".

And that's fine if that's actually happened. But you had better make sure that you're calling upon an ex cathedra statement by the community itself, and not a single guy with an axe to grind. Or one side of a complicated and interminable debate where both sides have about equal credentials and sway.

If you can't do that, you say "I think that my side of the academic debate is in the right, and here's why," not "your side has been debunked".

Otherwise you're going to end up like the minimum wage debaters, where both sides claim to have debunked the other. Or like the Federalist article that says a study has been "put to bed" as "bogus" just because another study said something different.

I think this is part of my reply to [the claim that](#) empiricism is so great that no one needs rationality.

A naive empiricist who swears off critical thinking because they can just "follow the evidence" has no contingency plan for when the evidence gets confusing. Their only recourse is to deny that the evidence is confusing, to assert that one side or the other has been "debunked". Since they've already made a principled decision not to study confirmation bias, chances are it's going to be whichever side they don't like that's "already been debunked". And by "debunked" they mean "a scientist on my side said it was wrong, so now I am relieved from the burden of thinking about it."

On the original post, I wrote:

Life is made up of limited, confusing, contradictory, and maliciously doctored facts. Anyone who says otherwise is either sticking to such incredibly easy solved problems that they never encounter anything outside their comfort level, or so closed-minded that they shut out any evidence that challenges their beliefs.

In the absence of any actual debunking more damning than a counterargument, "that's been debunked" is the way "shuts out any evidence that challenges their beliefs" feels from the inside.

V.

Somebody's going to want to know what's up with the original rape studies. The answer is that a small part of the discrepancy is response bias on the CSAS, but most of it is that the two surveys encourage respondents to define "sexual assault" in very different ways. Vox has [an excellent article on this](#) which for once I 100% endorse.

In other words, both are valid, both come together to form a more nuanced picture of campus violence, and neither one “debunks” the other.

Noisy Poll Results And Reptilian Muslim Climatologists from Mars

Beware of Phantom Lizardmen

I have only done a little bit of social science research, but it was enough to make me hate people. One study I helped with analyzed whether people from different countries had different answers on a certain psychological test. So we put up a website where people answered some questions about themselves (like “what country are you from?”) and then took the psychological test.

And so of course people screwed it up in every conceivable way. There were the merely dumb, like the guy who put “male” as his nationality and “American” as his gender. But there were also the actively malicious or at least annoying, like the people (yes, more than one) who wrote in “Martian”.

I think we all probably know someone like this, maybe a couple people like this.

I also think most of us *don't* know someone who believes reptilian aliens in human form control all the major nations of Earth.

Public Policy Polling's recent [poll on conspiracy theories](#) mostly showed up on my Facebook feed as “Four percent of Americans believe lizardmen are running the Earth”.

(of note, an additional 7% of Americans are “not sure” whether lizardmen are running the Earth or not.)

Imagine the situation. You're at home, eating dinner. You get a call from someone who says “Hello, this is Public Policy Polling. Would you mind answering some questions for us?” You say “Sure”. An extremely dignified sounding voice says – and this is the exact wording of the question – “Do you believe that shape-shifting reptilian people control our world by taking on human form and gaining political power to manipulate our society, or not?” Then it urges you to press 1 if yes, press 2 if no, press 3 if not sure.

So first we get the people who think “Wait, was 1 the one for if I did believe in lizardmen, or if I didn't? I'll just press 1 and move on to the next question.”

Then we get the people who are like “I never heard it before, but if this nice pollster thinks it's true, I might as well go along with them.”

Then we get the people who are all “F#&k you, polling company, I don't want people calling me when I'm at dinner. You screw with me, I tell you what I'm going to do. I'm going to tell you I believe lizard people are running the planet.”

And *then* we get the people who put “Martian” as their nationality in psychology experiments. Because some men just want to watch the world burn.

Do these three groups total 4% of the US population? Seems plausible.

I really wish polls like these would include a control question, something utterly implausible even by lizard-people standards, something like “Do you believe Barack

Obama is a hippopotamus?" Whatever percent of people answer yes to the hippo question get subtracted out from the other questions.

Poll Answers As Attire

Alas, not all weird poll answers can be explained that easily. On the same poll, 13% of Americans claimed to believe Barack Obama was the Anti-Christ. Subtracting our Lizardman's Constant of 4%, that leaves 9% of Americans who apparently gave this answer with something approaching sincerity.

(a friend on Facebook pointed out that 5% of *Obama voters* claimed to believe that Obama was the Anti-Christ, which seems to be another piece of evidence in favor of a Lizardman's Constant of 4-5%. On the other hand, I do enjoy picturing someone standing in a voting booth, thinking to themselves "Well, on the one hand, Obama is the Anti-Christ. On the other, do I really want four years of Romney?")

Some pollsters are starting to consider these sorts of things symptomatic of what they term [symbolic belief](#), which seems to be kind of what the Less Wrong sequences call [Professing and Cheering](#) or [Belief As Attire](#). Basically, people are being emotivists rather than realists about belief. "Obama is the Anti-Christ" is another way of just saying "Boo Obama!", rather than expressing some sort of proposition about the world.

And the same is true of "Obama is a Muslim" or "Obama was not born in America".

Never Attribute To Stupidity What Can Be Adequately Explained By Malice

But sometimes it's not some abstruse subtle bias. Sometimes it's not a good-natured joke. Sometimes people might just be actively working to corrupt your data.

Another link I've seen on my Facebook wall a few times is this one: [Are Climate Change Sceptics More Likely To Be Conspiracy Theorists?](#) It's based on a paper by Stephen Lewandowsky et al called [NASA Faked The Moon Landing, Therefore Climate Science Is A Hoax – An Analysis Of The Motivated Rejection Of Science](#).

The paper's thesis was that climate change skeptics are motivated by conspiracy ideation – a belief that there are large groups of sinister people out to deceive them. This seems sort of reasonable on the face of it – being a climate change skeptic requires going against the belief of the entire scientific establishment. My guess is that there probably is a significant link here waiting to be discovered.

Unfortunately, it's...possible Stephan Lewandowsky wasn't the best person to investigate this? Aside from being a professor of cognitive science, he also runs Shaping Tomorrow's World, a group that promotes "re-examining some of the assumptions we make about our technological, social and economic systems" and which seems to be largely about promoting global warming activism. While I think it's admirable that he is involved in that, it raises conflict of interest questions. And the way his paper is written – starting with the over-the-top title – doesn't do him any favors.

(if the conflict of interest angle doesn't make immediate and obvious sense to you, imagine how sketchy it would be if a professional global warming *denier* was involved in researching the motivations of global warming *supporters*)

But enough of my personal opinions. What's the paper look like?

The methodology goes like this: they send requests to several popular climate blogs, both believer and skeptic, asking them to link their readers to an online survey. The survey asks people their beliefs on global warming and on lots of conspiracy theories and fringe beliefs.

On first glance, the results are extremely damning. People who rejected climate science were wildly more likely to reject pretty much every other form of science as well, including the “theory” that HIV causes AIDS and the “theory” that cigarettes cause cancer. They were more willing to believe aliens landed at Roswell, that 9-11 was an inside job, and, yes, that NASA faked the moon landing. The conclusion: climate skeptics are just really stupid people.

But a bunch of global warming skeptics started re-analyzing the data and coming up with their own interpretations. They found that many large pro-global-warming blogs posted the link to the survey, but very few anti-global-warming blogs did. This then devolved into literally the [worst flame war](#) I have ever seen on the Internet, centering around accusations about whether the study authors deliberately excluded large anti-global warming blogs, or whether the authors asked the writers of anti-global-warming blogs and these writers just ignored the request (my impression is that most people now agree it was the latter). In either case, it ended up with most people taking the survey being from the pro-global-warming blogs, and only a few skeptics.

More interestingly, [they found](#) that pretty much all of the link between global warming skepticism and stupidity was a couple of people (there were so few skeptics, *and* so few conspiracy believers, that these couple of people made up a pretty big proportion of them, and way more than enough to get a “significant” difference with the global warming believers). Further, most of these couple of people had given the maximally skeptical answer to every single question about global warming, and the maximally credulous answer to every single question about conspiracies.

The danger here now seems obvious. Global warming believer blogs publish a link to this study, saying gleefully that it’s going to prove that global warming skeptics are idiots who also think NASA faked the moon landing and the world is run by lizardmen or whatever. Some global warming believers decide to help this process along by pretending to be super-strong global warming skeptics and filling in the stupidest answers they can to every question. The few real global warming skeptics who take the survey aren’t enough signal to completely drown out this noise. Therefore, they do the statistics and triumphantly announce that global warming skepticism is linked to stupid beliefs.

The global warming skeptic blogosphere has in my opinion done more than enough work to present a very very strong case that this is what happened (somebody else do an independent look at the controversy and double-check this for me?) And Professor Lewandowsky’s answer was...

...to publish a second paper, saying his results had been confirmed because climate skeptics were so obsessed with conspiracy theories that they had accused his data proving they were obsessed with conspiracies of being part of a conspiracy. The name of the paper? [Recursive Fury](#). I have to hand it to him, this is possibly *the most chutzpah I have ever seen a single human being display*.

(the paper is now partially offline as the journal investigates it for ethical something something)

The lesson from all three of the cases in this post seems clear. When we're talking about very unpopular beliefs, polls can only give a weak signal. Any possible source of noise - jokesters, cognitive biases, or deliberate misbehavior - can easily overwhelm the signal. Therefore, polls that rely on detecting very weak signals should be taken with a grain of salt.

Two Dark Side Statistics Papers

I.

First we have [False Positive Psychology: Undisclosed Flexibility In Data Collection And Analysis Allows Presenting Anything As Significant](#) (h/t Jonas Vollmer).

The message is hardly unique: there are lots of tricks unscrupulous or desperate scientists can use to artificially nudge results to the 5% significance level. The clarity of the presentation *is* unique. They start by discussing four particular tricks:

1. Measure multiple dependent variables, then report the ones that are significant. For example, if you're measuring whether treatment for a certain psychiatric disorder improves life outcomes, you can collect five different measures of life outcomes – let's say educational attainment, income, self-reported happiness, whether or not ever arrested, whether or not in romantic relationship – and have a 25%-ish probability one of them will come out at significance by chance. Then you can publish a paper called "Psychiatric Treatment Found To Increase Educational Attainment" without ever mentioning the four negative tests.

2. Artificially choose when to end your experiment. Suppose you want to prove that yelling at a coin makes it more likely to come up tails. You yell at a coin and flip it. It comes up heads. You try again. It comes up tails. You try again. It comes up heads. You try again. It comes up tails. You try again. It comes up tails again. You try again. It comes up tails again. You note that it came up tails four out of six times – a 66% success rate compared to expected 50% – and declare victory. Of course, this result wouldn't be significant, and it seems as if this should be a general rule – that almost by the definition of significance, you shouldn't be able to obtain it just by stopping the experiment at the right point. But the authors of the study perform several simulations to prove that this trick is more successful than you'd think:



3. Control for "confounders" (in practice, most often gender). I sometimes call this the "Elderly Hispanic Woman Effect" after drug trials that find that their drug doesn't have significant effects in the general population, but it *does* significantly help elderly Hispanic women. The trick is you split the population into twenty subgroups (young white men, young white women, elderly white men, elderly white women, young black men, etc), in one of those subgroups it will achieve significance by pure chance, and so you declare that your drug must just somehow be a perfect fit for elderly Hispanic women's unique body chemistry. This is not *always* wrong (some antihypertensives have notably different efficacy in white versus black populations) but it is *usually* suspicious.

4. Test different conditions and report the ones you like. For example, suppose you are testing whether vegetable consumption affects depression. You conduct the trial with three arms: low veggie diet, medium veggie diet, and high veggie diet. You now have four possible comparisons – low-medium, low-high, medium-high, low-medium-high trend). One of them will be significant 20% of the time, so you can just report that one: "People who eat a moderate amount of vegetables are less likely to get depression than people who eat excess vegetables" sounds like a perfectly reasonable result.

Then they run simulations to show exactly how much more likely you are to get a significant result in random data by employing each trick:



The image demonstrates that by using all four tricks, you can squeeze random data into a result significant at the $p < 0.05$ level about 61% of the time. The authors then put their money where their mouth is by conducting two studies. The first seems like a very very classic social psychology study. Subjects are randomly assigned to listen to one of two songs - either a nondescript control song or a child's nursery song. Then they are asked to rate how old they feel. Sure enough, the subjects who listen to the child's song feel older ($p = 0.03$). The second study is very similar, with one important exception. Once again, subjects are randomly assigned to listen to one of two songs - either a nondescript control song or a song about aging - "When I'm Sixty-Four" by The Beatles. Then they are asked to put down their actual age, in years. People who listened to the Beatles song became, on average, a year and a half younger than the control group ($p = 0.04$). So either the experimental intervention changed their subjects' ages, or the researchers were using statistical tricks. Turns out it was the second one. They explain how they used the four statistical tricks they explained above, and that without those tricks there would have been (obviously) no significant difference. They go on to say that their experiment meets the inclusion criteria for every major journal and that under current reporting rules there's no way anyone could have detected their data manipulation. They go on to list the changes they think the scientific establishment needs to prevent papers like theirs from reaching print. They're basically "don't do the things we just talked about", but as far as I can tell they rely on the honor system. I think a broader meta-point is that on important studies scientists should have to submit their experimental protocol to a journal and get it accepted or rejected in advance so they can't change tactics mid-stream or drop data. This would also force journals to publish more negative results. See also their interesting discussion of why they think "use Bayesian statistics" is a non-solution to the problem. II.

Second we have [How To Have A High Success Rate In Treatment: Advice For Evaluators Of Alcoholism Programs.](#)

This study is very close to my heart, because I'm working on my hospital's Substance Abuse Team this month. Every day we go see patients struggling with alcoholism, heroin abuse, et cetera, and we offer them treatment at our hospital's intensive inpatient Chemical Dependency Unit. And every day, our patients say thanks but no thanks, they heard of a program affiliated with their local church that has a 60% success rate, or an 80% success rate, or in one especially rosy-eyed case a frickin' 97% success rate.

(meanwhile, *real* rehab programs still struggle to prove they have a success rate greater than placebo)

My attending assumes these programs are scum but didn't really have a good evidence base for the claim, so I decided to search Google Scholar to find out what was going on. I struck gold in this paper, which is framed as a sarcastic how-to guide for unscrupulous drug treatment program directors who want to inflate their success rates without *technically* lying.

By far the best way to do this is to choose your denominator carefully. For example, it seems fair to only include the people who attended your full treatment program, not

the people who dropped out on Day One or never showed up at all – you can hardly be blamed for that, right? So suppose that your treatment program is one month intensive in rehab followed by a series of weekly meetings continuing indefinitely. At the end of one year, you define successful treatment completers as “the people who are still going to these meetings now, at the end of the year”. But in general, people who relapse into alcoholism are a whole lot less likely to continue attending their AA meetings than people who stay sober. So all you have to do is go up to people at your AA meeting, ask them if they’re still on the wagon, and your one-year success rate looks really good.

Another way to hack your treatment population is to only accept the most promising candidates to begin with (it works for private schools and it can work for you). We know that middle-class, employed people with houses and families have a much better prognosis than lower-class unemployed homeless single people. Although someone would probably notice if you put up a sign saying “MIDDLE-CLASS EMPLOYED PEOPLE WITH HOUSES AND FAMILIES ONLY”, a very practical option is to just charge a lot of money and let your client population select themselves. This is why for-profit private rehabs will have a higher success rate than public hospitals and government programs that deal with poor people.

Still another strategy is to follow the old proverb: “If at first you don’t succeed, redefine success”. “Abstinence” is such a harsh word. Why not “drinking in moderation”? This is a wonderful phrase, because you can just let the alcoholic involved determine the definition of moderation. A year after the program ends, you can send out little surveys saying “Remember when we told you God really wants you not to drink? You listened to us and are drinking in moderation now, right? Please check one: Y () N ()”. Who’s going to answer ‘no’ to that? Heck, some of the alcoholics I talk to say they’re drinking in moderation *while they are in the emergency room for alcohol poisoning*.

If you can’t handle “moderation”, how about “drinking less than you were before the treatment program”? This takes advantage of regression to the mean – you’re going to enter a rehab program at the worst period of your life, the time when your drinking finally spirals out of control. Just by coincidence, most other parts of your life will include less drinking than when you first came in to rehab, including the date a year after treatment when someone sends you a survey. Clearly rehab was a success!

And why wait a year? My attending and myself actually looked up what was going on with that one 97% success rate program our patient said he was going to. Here’s what they do – it’s a three month residential program where you live in a building just off the church and you’re not allowed to go out except on group treatment activities. Obviously there is no alcohol allowed in the building and you are surrounded by very earnest counselors and fellow recovering addicts at all times. Then, *at the end of the three months, while you are still in the building*, they ask you whether you’re drinking or not. You say no. Boom – 97% success rate.

One other tactic I have actually seen in studies and it *breaks my heart* is interval subdivision, which reminds me of some of the dirty tricks from the first study above. At five years’ follow-up, you ask people “Did you drink during Year 1? Did you drink during Year 2? Did you drink during Year 3?...” and so on. Now you have five chances to find a significant difference between treatment and control groups. I have literally seen studies that say “Our rehab didn’t have an immediate effect, but by Year 4 our patients were doing better than the controls.” Meanwhile, in years 1, 2, 3, and 5, for all we know the controls were doing better than the patients.

But if all else fails, there's always the old standby of poor researchers everywhere – just don't include a control group at all. This table really speaks to me:



The great thing about this table isn't just that it shows that seemingly impressive results are exactly the same as placebo. The great thing it shows is that results in the placebo groups in the four studies could be anywhere from a 22.5% success rate to an 87% success rate. These aren't treatment differences – all four groups are placebo! This is one hundred percent a difference in study populations and in success measures used. In other words, depending on your study protocol, you can prove that there is a 22.5% chance the average untreated alcoholic will achieve remission, or an 87% chance the average untreated alcoholic will achieve remission.

You can bet that rehabs use the study protocol that finds an 87% chance of remission in the untreated. And then they go on to boast of their 90% success rate. Good job, rehab!

The Control Group Is Out Of Control

I.

Allan Crossman calls parapsychology [the control group for science](#).

That is, in let's say a drug testing experiment, you give some people the drug and they recover. That doesn't tell you much until you give some other people a placebo drug you *know* doesn't work – but which they themselves believe in – and see how many of *them* recover. That number tells you how many people will recover whether the drug works or not. Unless people on your real drug do significantly better than people on the placebo drug, you haven't found anything.

On the meta-level, you're studying some phenomenon and you get some positive findings. That doesn't tell you much until you take some other researchers who are studying a phenomenon you *know* doesn't exist – but which they themselves believe in – and see how many of *them* get positive findings. That number tells you how many studies will discover positive results whether the phenomenon is real or not. Unless studies of the real phenomenon do significantly better than studies of the placebo phenomenon, you haven't found anything.

Trying to set up placebo science would be a logistical nightmare. You'd have to find a phenomenon that definitely doesn't exist, somehow convince a whole community of scientists across the world that it does, and fund them to study it for a couple of decades without them figuring it out.

Luckily we have a natural experiment in terms of parapsychology – the study of psychic phenomena – which most reasonable people believe don't exist, but which a community of practicing scientists believes in and publishes papers on all the time.

The results are pretty dismal. Parapsychologists are able to produce experimental evidence for psychic phenomena about as easily as normal scientists are able to produce such evidence for normal, non-psychic phenomena. This suggests the existence of a very large “placebo effect” in science – ie with enough energy focused on a subject, you can *always* produce “experimental evidence” for it that meets the usual scientific standards. As Eliezer Yudkowsky puts it:

Parapsychologists are constantly protesting that they are playing by all the standard scientific rules, and yet their results are being ignored – that they are unfairly being held to higher standards than everyone else. I'm willing to believe that. It just means that the standard statistical methods of science are so weak and flawed as to permit a field of study to sustain itself in the complete absence of any subject matter.

These sorts of thoughts have become more common lately in different fields. Psychologists admit to a [crisis of replication](#) as some of their most interesting findings turn out to be spurious. And in medicine, John Ioannides and others have been criticizing the research for a decade now and telling everyone they need to up their standards.

“Up your standards” has been a complicated demand that cashes out in a lot of technical ways. But there is broad agreement among the most intelligent voices I read ([1](#), [2](#), [3](#), [4](#), [5](#)) about a couple of promising directions we could go:

1. Demand very large sample size.
2. Demand replication, preferably exact replication, most preferably multiple exact replications.
3. Trust systematic reviews and meta-analyses rather than individual studies. Meta-analyses must prove homogeneity of the studies they analyze.
4. Use Bayesian rather than frequentist analysis, or even combine both techniques.
5. Stricter p-value criteria. It is far too easy to massage p-values to get less than 0.05. Also, make meta-analyses look for “p-hacking” by examining the distribution of p-values in the included studies.
6. Require pre-registration of trials.
7. Address publication bias by searching for unpublished trials, displaying funnel plots, and using statistics like “fail-safe N” to investigate the possibility of suppressed research.
8. Do heterogeneity analyses or at least observe and account for differences in the studies you analyze.
9. Demand randomized controlled trials. None of this “correlated even after we adjust for confounders” BS.
10. Stricter effect size criteria. It’s easy to get small effect sizes in *anything*.

If we follow these ten commandments, then we avoid the problems that allowed parapsychology and probably a whole host of other problems we don’t know about to sneak past the scientific gatekeepers.

Well, [what now, motherfuckers?](#)

II.

Bem, Tressoldi, Rabeyron, and Duggan (2014), full text available for download at the top bar of the link above, is parapsychology’s way of saying “thanks but no thanks” to the idea of a more rigorous scientific paradigm making them quietly wither away.

You might remember Bem as the prestigious establishment psychologist who decided to try his hand at parapsychology and to his and everyone else’s surprise got positive results. Everyone had a lot of criticisms, some of which were [very very good](#), and the study [failed replication several times](#). Case closed, right?

Earlier this month Bem came back with a meta-analysis of ninety replications from tens of thousands of participants in thirty three laboratories in fourteen countries confirming his original finding, $p < 1.2 \times 10^{-10}$, Bayes factor 7.4×10^9 , funnel plot beautifully symmetrical, p-hacking curve nice and right-skewed, Orwin fail-safe n of 559, et cetera, et cetera, et cetera.

By my count, Bem follows all of the commandments except [6] and [10]. He apologizes for not using pre-registration, but says it’s okay because the studies were exact replications of a previous study that makes it impossible for an unsavory researcher to change the parameters halfway through and does pretty much the same

thing. And he apologizes for the small effect size but points out that some effect sizes are legitimately very small, this is no smaller than a lot of other commonly-accepted results, and that a high enough p-value ought to make up for a low effect size.

This is *far* better than the average meta-analysis. Bem has always been pretty careful and this is no exception. Yet its conclusion is that psychic powers exist.

So - once again - what now, motherfuckers?

III.

In retrospect, that list of ways to fix science above was a little optimistic.

The first nine items (large sample sizes, replications, low p-values, Bayesian statistics, meta-analysis, pre-registration, publication bias, heterogeneity) all try to solve the same problem: accidentally mistaking noise in the data for a signal.

We've placed so much emphasis on not mistaking noise for signal that when someone like Bem hands us a beautiful, perfectly clear signal on a silver platter, it briefly stuns us. "Wow, of the three hundred different terrible ways to mistake noise for signal, Bem has proven beyond a shadow of a doubt he hasn't done any of them." And we get so stunned we're likely to forget that this is only part of the battle.

Bem definitely picked up a signal. The only question is whether it's a signal of psi, or a signal of poor experimental technique.

None of these commandments even *touch* poor experimental technique - or confounding, or whatever you want to call it. If an experiment is confounded, if it produces a strong signal even when its experimental hypothesis is true, then using a larger sample size will just make that signal even stronger.

Replicating it will just reproduce the confounded results again.

Low p-values will be easy to get if you perform the confounded experiment on a large enough scale.

Meta-analyses of confounded studies will obey the immortal law of "garbage in, garbage out".

Pre-registration only assures that your study will not get any worse than it was the first time you thought of it, which may be very bad indeed.

Searching for publication bias only means you will get *all* of the confounded studies, instead of just some of them.

Heterogeneity just tells you whether all of the studies were confounded about the same amount.

Bayesian statistics, alone among these first eight, ought to be able to help with this problem. After all, a good Bayesian should be able to say "Well, I got some impressive results, but my prior for psi is very low, so this raises my belief in psi slightly, but raises my belief that the experiments were confounded *a lot*."

Unfortunately, good Bayesians are hard to come by, and the researchers here seem to be making some serious mistakes. Here's Bem:

An opportunity to calculate an approximate answer to this question emerges from a Bayesian critique of Bem's (2011) experiments by Wagenmakers, Wetzels, Borsboom, & van der Maas (2011). Although Wagenmakers et al. did not explicitly claim psi to be impossible, they came very close by setting their prior odds at 10^{20} against the psi hypothesis. The Bayes Factor for our full database is approximately 10^9 in favor of the psi hypothesis (Table 1), which implies that our meta-analysis should lower their posterior odds against the psi hypothesis to 10^{11} .

Let me shame both participants in this debate.

Bem, you are abusing Bayes factor. If Wagenmakers uses your 10^9 Bayes factor to adjust from his prior of 10^{-20} to 10^{-11} , then what happens the next time you come up with another database of studies supporting your hypothesis? We all know you will, because you've amply proven these results weren't due to chance, so whatever factor produced these results – whether real psi or poor experimental technique – will no doubt keep producing them for the next hundred replication attempts. When those come in, does Wagenmakers have to adjust his probability from 10^{-11} to 10^{-2} ? When you get another hundred studies, does he have to go from 10^{-2} to 10^7 ? If so, then by [conservation of expected evidence](#) he should just update to 10^{+7} right now – or really to infinity, since you can keep coming up with more studies till the cows come home. But in fact he shouldn't do that, because at some point his thought process becomes “Okay, I already know that studies of this quality can consistently produce positive findings, so either psi is real or studies of this quality aren't good enough to disprove it”. This point should probably happen well before he increases his probability by a factor of 10^9 . See [Confidence Levels Inside And Outside An Argument](#) for this argument made in greater detail.

Wagenmakers, you are overconfident. Suppose God came down from Heaven and said in a booming voice “EVERY SINGLE STUDY IN THIS META-ANALYSIS WAS CONDUCTED PERFECTLY WITHOUT FLAWS OR BIAS, AS WAS THE META-ANALYSIS ITSELF.” You would see a p-value of less than 1.2×10^{-10} and think “I bet that was just coincidence”? And then they could do another study of the same size, also God-certified, returning exactly the same results, and you would say “I bet that was just coincidence too”? YOU ARE NOT THAT CERTAIN OF ANYTHING. Seriously, *read the @#!\$ing Sequences*.

Bayesian statistics, at least the way they are done here, aren't going to be of much use to anybody.

That leaves randomized controlled trials and effect sizes.

Randomized controlled trials are great. They eliminate most possible confounders in one fell swoop, and are excellent at keeping experimenters honest. Unfortunately, most of the studies in the Bem meta-analysis were already randomized controlled trials.

High effect sizes are really the only thing the Bem study lacks. And it is very hard to experimental technique so bad that it consistently produces a result with a high effect size.

But as Bem points out, demanding high effect size limits our ability to detect real but low-effect phenomena. Just to give an example, many physics experiments – like the ones that detected the Higgs boson or neutrinos – rely on detecting extremely small perturbations in the natural order, over millions of different trials. Less esoterically, Bem mentions the example of aspirin decreasing heart attack risk, which it definitely

does and which is very important, but which has an effect size lower than that of his psi results. If humans have some kind of very *weak* psionic faculty that under regular conditions operates poorly and inconsistently, but does indeed exist, then excluding it by definition from the realm of things science can discover would be a bad idea.

All of these techniques are about reducing the chance of confusing noise for signal. But when we think of them as the be-all and end-all of scientific legitimacy, we end up in awkward situations where they come out super-confident in a study's accuracy simply because the issue was one they weren't geared up to detect. Because a lot of the time the problem is something more than just noise.

IV.

Wiseman & Schlitz's [Experimenter Effects And The Remote Detection Of Staring](#) is my favorite parapsychology paper ever and sends me into fits of nervous laughter every time I read it.

The backstory: there is a classic parapsychological experiment where a subject is placed in a room alone, hooked up to a video link. At random times, an experimenter stares at them menacingly through the video link. The hypothesis is that this causes their galvanic skin response (a physiological measure of subconscious anxiety) to increase, even though there is no non-psychic way the subject could know whether the experimenter was staring or not.

Schlitz is a psi believer whose staring experiments had consistently supported the presence of a psychic phenomenon. Wiseman, in accordance with [nominative determinism](#) is a psi skeptic whose staring experiments keep showing nothing and disproving psi. Since they were apparently the only two people in all of parapsychology with a smidgen of curiosity or rationalist virtue, they decided to team up and figure out why they kept getting such different results.

The idea was to plan an experiment together, with both of them agreeing on every single tiny detail. They would then go to a laboratory and set it up, again both keeping close eyes on one another. Finally, they would conduct the experiment in a series of different batches. Half the batches (randomly assigned) would be conducted by Dr. Schlitz, the other half by Dr. Wiseman. Because the two authors had very carefully standardized the setting, apparatus and procedure beforehand, "conducted by" pretty much just meant greeting the participants, giving the experimental instructions, and doing the staring.

The results? Schlitz's trials found strong evidence of psychic powers, Wiseman's trials found no evidence whatsoever.

Take a second to reflect on how this *makes no sense*. Two experimenters in the same laboratory, using the same apparatus, having no contact with the subjects except to introduce themselves and flip a few switches – and whether one or the other was there that day completely altered the result. For a good time, watch the gymnastics they have to do to in the paper to make this sound sufficiently sensical to even get published. This is the only journal article I've ever read where, in the part of the Discussion section where you're supposed to propose possible reasons for your findings, both authors suggest maybe their co-author hacked into the computer and altered the results.

While it's nice to see people exploring Bem's findings further, *this* is the experiment people should be replicating ninety times. I expect *something* would turn up.

As it is, Kennedy and Taddonio [list ten similar studies](#) with similar results. One cannot help wondering about publication bias (if the skeptic and the believer got similar results, who cares?). But the phenomenon is sufficiently well known in parapsychology that it has led to its own host of theories about how skeptics emit negative auras, or the enthusiasm of a proponent is a necessary kindling for psychic powers.

Other fields don't have this excuse. In psychotherapy, for example, practically the only consistent finding is that whatever kind of psychotherapy the person running the study likes is most effective. Thirty different meta-analyses on the subject have confirmed this with strong effect size ($d = 0.54$) and good significance ($p = .001$).

Then there's [Munder \(2013\)](#), which is a meta-meta-analysis on whether meta-analyses of confounding by researcher allegiance effect were themselves meta-confounded by meta-researcher allegiance effect. He found that indeed, meta-researchers who believed in researcher allegiance effect were more likely to turn up positive results in their studies of researcher allegiance effect ($p < .002$). It gets worse. There's [a famous story](#) about an experiment where a scientist told teachers that his advanced psychometric methods had predicted a couple of kids in their class were about to become geniuses (the students were actually chosen at random). He followed the students for the year and found that their intelligence actually increased. This was supposed to be a Cautionary Tale About How Teachers' Preconceptions Can Affect Children.

Less famous is that the same guy did the same thing with rats. He sent one laboratory a box of rats saying they were specially bred to be ultra-intelligent, and another lab a box of (identical) rats saying they were specially bred to be slow and dumb. Then he had them do standard rat learning tasks, and sure enough the first lab found very impressive results, the second lab very disappointing ones.

This scientist – let's give his name, Robert Rosenthal – [then investigated three hundred forty five different studies](#) for evidence of the same phenomenon. He found effect sizes of anywhere from 0.15 to 1.7, depending on the type of experiment involved. Note that this could also be phrased as “between twice as strong and twenty times as strong as Bem's psi effect”. Mysteriously, animal learning experiments displayed the highest effect size, supporting the folk belief that animals are hypersensitive to subtle emotional cues.

Okay, fine. Subtle emotional cues. That's way more scientific than saying “negative auras”. But the question remains – what went wrong for Schlitz and Wiseman? Even if Schlitz had done everything short of saying “The hypothesis of this experiment is for your skin response to increase when you are being stared at, please increase your skin response at that time,” and subjects had tried to comply, the whole point was that they didn't *know* when they were being stared at, because to find that out you'd have to be psychic. And how are these rats figuring out what the experimenters' subtle emotional cues mean anyway? I can't figure out people's subtle emotional cues half the time!

I know that standard practice here is to tell [the story of Clever Hans](#) and then say That Is Why We Do Double-Blind Studies. But first of all, I'm pretty sure no one does double-blind studies with rats. Second of all, I think most social psych studies aren't double blind – I just checked the first one I thought of, Aronson and Steele on stereotype threat, and it certainly wasn't. Third of all, this effect seems to be just as common in cases where it's hard to imagine how the researchers' subtle emotional cues could make a difference. Like Schlitz and Wiseman. Or like the psychotherapy experiments,

where most of the subjects were doing therapy with individual psychologists and never even saw whatever prestigious professor was running the study behind the scenes.

I think it's a combination of subconscious emotional cues, subconscious statistical trickery, perfectly conscious fraud which for all we know happens much more often than detected, and things we haven't discovered yet which are at least as weird as subconscious emotional cues. But rather than speculate, I prefer to take it as a brute fact. Studies are going to be confounded by the allegiance of the researcher. When researchers who don't believe something discover it, that's when it's worth looking into.

V.

So what exactly happened to Bem?

Although Bem looked hard to find unpublished material, I don't know if he succeeded. Unpublished material, in this context, has to mean "material published enough for Bem to find it", which in this case was mostly things presented at conferences. What about results so boring that they were never even mentioned?

And I predict people who believe in parapsychology are more likely to conduct parapsychology experiments than skeptics. Suppose this is true. And further suppose that for some reason, experimenter effect is real and powerful. That means most of the experiments conducted will support Bem's result. But this is still a weird form of "publication bias" insofar as it ignores the contrary results of hypothetically experiments that were never conducted.

And worst of all, maybe Bem really did do an excellent job of finding every little two-bit experiment that no journal would take. How much can we trust these non-peer-reviewed procedures?

I looked through his list of ninety studies for all the ones that were both exact replications and had been peer-reviewed (with one caveat to be mentioned later). I found only seven:

Batthyany, Kranz, and Erber: .268
Ritchie 1: 0.015
Ritchie 2: -0.219
Ritchie 3: -0.040
Subbotsky 1: 0.279
Subbotsky 2: 0.292
Subbotsky 3: -.399

Three find large positive effects, two find approximate zero effects, and two find large negative effects. Without doing any calculation, this seems pretty darned close to chance for me.

Okay, back to that caveat about replications. One of Bem's strongest points was how many of the studies included were exact replications of his work. This is important because if you do your own novel experiment, it leaves a lot of wiggle room to keep changing the parameters and statistics a bunch of times until you get the effect you want. This is why lots of people want experiments to be preregistered with specific commitments about what you're going to test and how you're going to do it. These

experiments weren't preregistered, but conforming to a previously done experiment is a pretty good alternative.

Except that I think the criteria for "replication" here were exceptionally loose. For example, Savva et al was listed as an "exact replication" of Bem, but it was performed in 2004 – seven years before Bem's original study took place. I know Bem believes in precognition, but that's going *too far*. As far as I can tell "exact replication" here means "kinda similar psionic-y thing". Also, Bem classily lists his own experiments as exact replications of themselves, which gives a big boost to the "exact replications return the same results as Bem's original studies" line. I would want to see much stricter criteria for replication before I relax the "preregister your trials" requirement.

(Richard Wiseman – the same guy who provided the negative aura for the Wiseman and Schiltz experiment – has started [a pre-register site for Bem replications](#). He says he has received five of them. This is very promising. There is also [a separate pre-register for parapsychology trials in general](#). I am both extremely pleased at this victory for good science, and ashamed that my own field is apparently behind parapsychology in the "scientific rigor" department)

That is my best guess at what happened here – a bunch of poor-quality, peer-unreviewed studies that weren't as exact replications as we would like to believe, all subject to mysterious experimenter effects.

This is not a criticism of Bem or a criticism of parapsychology. It's something that is inherent to the practice of meta-analysis, and even more, inherent to the practice of science. Other than a few very exceptional large medical trials, there is not a study in the world that would survive the level of criticism I am throwing at Bem right now.

I think Bem is wrong. The level of criticism it would take to prove a wrong study wrong is higher than that almost any existing study can withstand. That is not encouraging for existing studies.

VI.

The motto of the Royal Society – Hooke, Boyle, Newton, some of the people who arguably invented modern science – was *nullus in verba*, "take no one's word".

This was a proper battle cry for seventeenth century scientists. Think about the (admittedly kind of mythologized) history of Science. The scholastics saying that matter was this, or that, and justifying themselves by long treatises about how based on A, B, C, the word of the Bible, Aristotle, self-evident first principles, and the Great Chain of Being all clearly proved their point. Then other scholastics would write different long treatises on how D, E, and F, Plato, St. Augustine, and the proper ordering of angels all indicated that clearly matter was something different. Both groups were pretty sure that the other had made a subtle error of reasoning somewhere, and both groups were perfectly happy to spend centuries debating exactly which one of them it was.

And then Galileo said "Wait a second, instead of debating exactly how objects fall, let's just drop objects off of something really tall and see what happens", and after that, Science.

Yes, it's kind of mythologized. But like all myths, it contains a core of truth. People are terrible. If you let people debate things, they will do it forever, come up with horrible ideas, get them entrenched, play politics with them, and finally reach the point where

they're coming up with theories why people who disagree with them are probably secretly in the pay of the Devil.

Imagine having to conduct the global warming debate, except that you couldn't appeal to scientific consensus and statistics because scientific consensus and statistics hadn't been invented yet. In a world without science, *everything* would be like that.

Heck, just look at *philosophy*.

This is the principle behind the Pyramid of Scientific Evidence. The lowest level is your personal opinions, no matter how ironclad you think the logic behind them is. Just above that is expert opinion, because no matter how expert someone is they're still only human. Above that is anecdotal evidence and case studies, because even though you're finally getting out of people's heads, it's still possible for the content of people's heads to influence which cases they pay attention to. At each level, we distill away more and more of the human element, until presumably at the top the dross of humanity has been purged away entirely and we end up with pure unadulterated reality.



The Pyramid of Scientific Evidence

And for a while this went *well*. People would drop things off towers, or see how quickly gases expanded, or observe chimpanzees, or whatever.

Then things started getting more complicated. People started investigating more subtle effects, or effects that shifted with the observer. The scientific community became bigger, everyone didn't know everyone anymore, you needed more journals to find out what other people had done. Statistics became more complicated, allowing the study of noisier data but also bringing more peril. And a lot of science done by smart and honest people ended up being wrong, and we needed to figure out exactly which science that was.

And the result is a lot of essays like this one, where people who think they're smart take one side of a scientific "controversy" and say which studies you should believe. And then other people take the other side and tell you why you should believe different studies than the first person thought you should believe. And there is much argument and many insults and citing of authorities and interminable debate for, if not centuries, at least a pretty long time.

The highest level of the Pyramid of Scientific Evidence is meta-analysis. But a lot of meta-analyses are crap. This meta-analysis got $p < 1.2 \times 10^{-10}$ for a conclusion I'm pretty sure is false, and *it isn't even one of the crap ones*. Crap meta-analyses look [more like this](#), or even worse.

How do I know it's crap? Well, I use my personal judgment. How do I know my personal judgment is right? Well, a smart well-credentialed person like James Coyne agrees with me. How do I know James Coyne is smart? I can think of lots of cases where he's been right before. How do I know those count? Well, John Ioannides has published a lot of studies analyzing the problems with science, and confirmed that cases like the ones Coyne talks about are pretty common. Why can I believe Ioannides' studies? Well, there have been good meta-analyses of them. But how do I know if those meta-analyses are crap or not? Well...



The Ouroboros of Scientific Evidence

Science! YOU WERE THE CHOSEN ONE! It was said that you would destroy reliance on biased experts, not join them! Bring balance to epistemology, not leave it in darkness!



I LOVED YOU!!!!

Edit: [Conspiracy theory](#) by Andrew Gelman

The Cowpox of Doubt

I remember hearing someone I know try to explain rationality to his friends.

He started with “It’s important to have correct beliefs. You might think this is obvious, but think about creationists and homeopaths and people who think the moon landing was a hoax.” And then further on in this vein.

And I thought: “NO NO NO NO NO NO NO!”

I will make a confession. Every time someone talks about the stupidity of creationists, moon-hoaxers, and homeopaths, I cringe.

It’s not that moon-hoaxers, homeopaths et al aren’t dumb. They are. It’s not even that these people don’t do real harm. They do.

(although probably less than people think; people rarely stop conventional treatment in favor of homeopathy, and both [a popular website](#) and [a review article](#) have a really hard time finding more than a handful of people genuinely harmed by it. Moon hoaxes seem even less dangerous, [unless of course you are standing near Buzz Aldrin when you talk about them.](#))

What annoys me about the people who harp on moon-hoaxing and homeopathy – without any interest in the rest of medicine or space history – is that it seems like an attempt to Other irrationality.

(yes, I did just use “other” as a verb. Maybe I’ve been hanging around Continental types too much lately.)

It’s saying “Look, over here! It’s irrational people, believing things that we can instantly dismiss as dumb. Things we feel no temptation, not one bit, to believe. It must be that they are defective and we are rational.”

But to me, the rationality movement is about Self-ing irrationality.

(yes, I did just use “self” as a verb. I don’t even have the excuse of it being part of a philosophical tradition)

It is about realizing that you, yes you, might be wrong about the things that you’re most certain of, and nothing can save you except maybe extreme epistemic paranoia.

Talking about moon-hoaxers and homeopaths too much, at least the way we do it, is *counterproductive* to this goal. Throw examples of obviously stupid false beliefs at someone, and they start thinking all false beliefs are obvious. Give too many examples of false beliefs that aren’t tempting to them, and they start believing they’re immune to temptation.

And it raises sloppiness to a virtue.

Take homeopathy. I can’t even count the number of times I’ve heard people say: “Homeopaths don’t realize beliefs require evidence. No study anywhere has ever found homeopathy to be effective!”

But of course dozens of studies have found homeopathy to be effective.

“Well, sure, but they weren’t double-blind! What you don’t realize is that there can be placebo effects from...”

But of course many of these studies have been large double-blinded randomized controlled trials, or even meta-analyses of such.

“Okay, but not published in reputable journals.”

Is [The Lancet](#) reputable enough for you?

“But homeopaths don’t even realize that many of their concoctions don’t contain even a single molecule of active substance!”

But of course almost all homeopaths realize this and their proposed mechanism for homeopathic effects not only survives this criticism but relies upon it.

“But all doctors and biologists agree that homeopathy doesn’t work!”

Have you ever spent the five seconds it would take to look up a survey of what percent of doctors and biologists believe homeopathy doesn’t work? Or are you just assuming that’s true because someone on your side told you so and it seems right?

I am of course being mean here. Being open-minded to homeopaths – reading all the research carefully, seeking out their own writings so you don’t accidentally straw-man them, double-checking all of your seemingly “obvious” assumptions – would be a waste of your time.

And someone who demands that you be open-minded about homeopathy would not be your friend. They would probably be a shill for homeopathy and best ignored.

But this is exactly the problem!

The more we concentrate on homeopathy, and moon hoaxes, and creationism – the more people who have never felt any temptation towards these beliefs go through the motions of “debunk”-ing them a hundred times to one another for fun – the more we are driving home the message that these are a representative sample of the kinds of problems we face.

And the more we do that, the more we are training people to make the correct approach to homeopathy – ignoring poor research and straw men on your own side while being very suspicious of anyone who tells us to be careful – their standard approach to any controversy.

And then we get people believing [all sorts of shoddy research](#) – because after all, the world is divided between things like homeopathy that Have Never Been Supported By Any Evidence Ever, and things like conventional medicine that Have Studies In Real Journals And Are Pushed By Real Scientists.

Or losing all subtlety and moderation in their political beliefs, never questioning their [own side’s claims](#), because the world is divided between People Like Me Who Know The Right Answer, and Shills For The Other Side Who Tell Me To Be Open-Minded As Part Of A Trap.

This post was partly inspired by Gruntled and Hinged’s [You Probably Don’t Want Peer-Reviewed Evidence For God](#) (actually, I started writing it before that was published –

but since Bem has published evidence showing psi exists, I must have just been precognitively inspired by it). But there's [another G&H post](#) that retrocausally got me thinking even more.

Inoculation is when you use a weak pathogen like cowpox to build immunity against a stronger pathogen like smallpox. The inoculation effect in psychology is when a person, upon being presented with several weak arguments against a proposition, becomes immune to stronger arguments against the same position.

Tell a religious person that Christianity is false because Jesus is just a blatant ripoff of the warrior-god Mithras and they'll open up a Near Eastern history book, notice that's not true at all, and then be that much more skeptical of the next argument against their faith. "Oh, atheists. Those are those people who think stupid things like Jesus = Mithras. I already figured out they're not worth taking seriously." Except on a deeper level that precedes and is immune to conscious thought.

So we take the intelligent Internet-reading public, and we throw a bunch of incredibly dumb theories at them – moon-hoaxism, homeopathy, creationism, anti-vaxxing, lizard people, that one guy who thought the rapture would come a couple years ago, whatever. And they are easily debunked, and the stuff you and all your friends believed was obviously true is, in fact, obviously true, and any time you spent investigating whether you were wrong is time you wasted.

And I worry that we are vaccinating people against reading the research for themselves instead of trusting smarmy bloggers who talk about how stupid the other side is.

That we are vaccinating people against thinking there might be important truths on both sides of an issue.

That we are vaccinating people against understanding how "scientific evidence" is a really complicated concept, and that many things that are in peer-reviewed journals will later turn out to be wrong.

That we are vaccinating people against the idea that many theories they find absurd or repugnant at first will later turn out to be true, because nature doesn't respect our feelings.

That we are vaccinating people against *doubt*.

And maybe this is partly good. It's probably a good idea to trust your doctor and also a good idea to trust your climatologist, and rare is the field where I would feel comfortable challenging expert consensus completely.

But there's also this problem of hundreds of different religions and political ideologies, and most people are born into ones that are at least somewhat wrong. That makes this capacity for real doubt – doubting something even though all your family and friends is telling you it's obviously true and you must be an idiot to question it at all – a tremendously important skill. It's especially important for the couple of rare individuals who will be in a position to cause a paradigm shift in a science by doubting one of its fundamental assumptions.

I don't think that reading about lizard people or creationism will affect people's ability to distinguish between, let's say, cyclic universe theory versus multiverse theory, or other equally dispassionate debates.

But if ever you ever need to have [a true crisis of faith](#), then any time you spend thinking about homeopathy and moon hoaxes beyond the negligible effect they have on your life will be time spent learning exactly the wrong mental habits.

How Common Are Science Failures?

After a brief spurt of debate over the claim that “97% of relevant published papers support anthropogenic climate change”, I think the picture has mostly settled to an agreement that – although we can contest the methodology of that particular study – there are multiple lines of evidence that the number is somewhere in the nineties.

So if any doubt at all is to remain about climate change, it has to come from the worry that sometimes entire scientific fields can get things near-unanimously wrong, especially for political or conformity-related reasons.

In fact, I’d go so far as to say that if we are not climatologists ourselves, our prior on climate change should be *based upon* how frequently entire scientific fields get things terribly wrong for political or conformity-related reasons.

Skeptics mock the claim that [science was wrong before](#), but skeptics mock *everything*. A better plan might be to try to quantify the frequency of scientific failures so we can see how good (or bad) the chances are for any given field.

Before we investigate, we should define our reference class properly. I think a scientific mistake only counts as a reason for doubting climate change (or any other commonly-accepted scientific paradigm) if:

1. It was made sometime in the recent past. Aristotle was wrong about all sorts of things, and so were those doctors who thought everything had to do with black bile, but the scientific community back then was a lot less rigorous than our own. Let’s say it counts if it’s after 1900.
2. It was part of a really important theory, one of the fundamental paradigms of an entire field. I’m sure some tiny group of biologists have been wrong about how many chromosomes a shrew has, but that’s probably an easier mistake to wander into than all of climatology screwing up simultaneously.
3. It was a stubborn resistance to the truth, rather than just a failure to have come up with the correct theory immediately. People were geocentrists before they were heliocentrists, but this wasn’t because the field of astronomy became overly politicized and self-assured, it was because (aside from [one ancient Greek guy](#) nobody really read) heliocentrism wasn’t invented until the 1500s, and after that it took people a couple of generations to catch on. In the same way, Newton’s theory of gravity wasn’t quite as good as Einstein’s, but this would not shame physicists in the same way climate change being wrong would shame climatologists. Let’s say that in order to count, the correct theory has to be very well known (the correct theory is allowed to be “this phenomenon doesn’t exist at all and you are wasting your time”) and there is a large group of people mostly outside the mainstream scientific establishment pushing it (for approximately correct reasons) whom scientists just refuse to listen to.
4. We now know that the past scientific establishment was definitely, definitely wrong and everyone agrees about this and it is not seriously in doubt. This criterion isn’t to be fair to the climatologists, this is to be fair to *me* when I have to read the comments to this post and get a bunch of “Nutritionists have yet to sign on to my pet theory of diet, that *proves* some scientific fields are hopelessly corrupt!”

Do any such scientific failures exist?

If we want to play this game on Easy Mode, our first target will be [Lysenkoism](#), the completely bonkers theory of agriculture and genetics adopted by the Soviet Union. A low-level agricultural biologist, Lysenko, came up with questionable ways of increasing agricultural output through something kind of like Lamarckian evolution. The Soviet government wanted to inspire people in the middle of a famine, didn't really like real scientists because they seemed kind of bourgeois, and wanted to discredit genetics because heritability seemed contrary to the idea of New Soviet Man. So they promoted Lysenko enough times that everyone got the message that Lysenkoism was the road to getting good positions. All the careerists switched over to the new paradigm, and the holdouts who continued to believe in genetics were denounced as fascists. According to Wikipedia, "in 1948, genetics was officially declared "a bourgeois pseudoscience"; all geneticists were fired from their jobs (some were also arrested), and all genetic research was discontinued."

About twenty years later the Soviets quietly came to their senses and covered up the whole thing.

I would argue that Stalinist Russia, where the government was very clearly intervening in science and killing the people it didn't like, isn't a fair test case for a theory today. But climate change opponents would probably respond that the liberal world order is unfairly promoting scientists who support climate change and persecuting those who oppose it. And Lysenkoism at least proves that is the sort of thing which can in theory sometimes happen. So let's grumble a little but give it to them.

Now we turn the dial up to Hard Mode. Are there any cases of failure on a similar level within a scientific community *in a country not actively being ruled by Stalin*?

I can think of two: Freudian psychoanalysis and behaviorist psychology.

Freudian psychoanalysis [needs no introduction](#). It dominated psychiatry – not at all a small field – from about 1930 to 1980. As far as anyone can tell, the entire gigantic edifice has no redeeming qualities. I mean, it correctly describes the existence of a subconscious, and it may have some insightful things to say on childhood trauma, but as far as a decent model of the brain or of psychological treatment goes, it was a giant mistake.

I got a little better idea just *how* big a mistake doing some research for the Anti-Reactionary FAQ. I wanted to see how homosexuals were viewed back in the 1950s and ran across two New York Times articles about them ([1](#), [2](#)). It's really creepy to see them explaining how instead of holding on to folk beliefs about how homosexuals are normal people just like you or me, people need to start listening to the psychoanalytic experts, who know the *real* story behind why some people are homosexual. The interviews with the experts in the article are a little surreal.

Psychoanalysis wasn't an honest mistake. The field already had a perfectly good alternative – denouncing the whole thing as bunk – and sensible non-psychoanalysts seemed to do exactly that. On the other hand, the more you got "educated" about psychiatry in psychoanalytic institutions, and the more you wanted to become a psychiatrist yourself, the more you got biased into think psychoanalysis was obviously correct and dismissing the doubters as science denalists or whatever it was they said back then.

So this seems like a genuine example of a scientific field failing.

Behaviorism in psychology was...well, this part will be controversial. A weak version is "psychologists should not study thoughts or emotions because these are unknowable by scientific methods; instead they should limit themselves to behaviors". A strong version is "thoughts and emotions don't exist; they are post hoc explanations invented by people to rationalize their behaviors". People are going to tell me that real psychologists only believed the weak version, but having read more than a little 1950s psychology, I'm going to tell them they're wrong. I think a lot of people believed the strong version and that in fact it was the dominant paradigm in the field.

And of course common people said this was stupid, of course we have thoughts and emotions, and the experts just said that kind of drivel was exactly what common people *would* think. Then came the cognitive revolution and people realized thoughts and emotions were actually kind of easy to study. And then we got MRI machines and are now a good chunk of the way to *seeing* them.

So this too I will count as a scientific failure.

But - and this seems important - I can't think of any others.

Suppose there are about fifty scientific fields approximately as important as genetics or psychiatry or psychology. And suppose within the past century, each of them had room for about five paradigms as important as psychoanalysis or behaviorism or Lysenkoism.

That would mean there are about 250 possibilities for science failure, of which three were actually science failures - for a failure rate of 1.2%.

This doesn't seem much more encouraging for the anti-global-warming cause than the 3% of papers that support them.

I think I'm being pretty fair here - after all, Lysenkoism was limited to one extremely-screwed-up country, and people are going to yell that behaviorism wasn't as bad as I made it sound. And two of the three failures are in psychology, a social science much fuzzier than climatology where we can expect far more errors. A cynic might say if we include psychology we might as well go all the way and include economics, sociology, and anthropology, raising our error count to over nine thousand.

But if we want to be even fairer, we can admit that there are probably some science failures that haven't been detected yet. I can think of three that I very strongly suspect are in that category, although I won't tell you what they are so as to not distract from the meta-level debate. That brings us to 2.4%. Admit that maybe I've only caught half of the impending science failures out there, and we get to 3.6%. Still not much of an improvement for the anti-AGW crowd over having 3% of the literature.

Unless of course I am missing a whole load of well-known science failures which you will remind me about in the comments.

[Edit: Wow, people are really bad at following criteria 3 and 4, even going so far as to post the exact examples I said not to. Don't let that be you.]

Learning To Love Scientific Consensus

[Related to: [Contrarians, Crackpots, and Consensus](#), [How Common Are Science Failures?](#). Epistemic status is “subtle and likely to be misinterpreted”.]

I.

There’s a [list of scientific mavericks who were ridiculed by hidebound reactionaries but later vindicated](#) that’s been going viral. I examined the first ten mavericks on the list to see if its claims held up. Overall I wasn’t too impressed. Let me go over them in more detail.

SVANTE ARRHENIUS:

His idea that electrolytes are full of charged atoms was considered crazy. The atomic theory was new at the time, and everyone “knew” that atoms were indivisible (and hence they could not lose or gain any electric charge.) Because of his heretical idea, he only received his university degree by a very narrow margin.

Sure, the professors who were judging his PhD thesis weren’t too convinced. So Arrhenius sent his proposal to the world’s top chemists at the time, and they were super-interested and started fighting among themselves to work with Arrhenius on it. Top chemist Wilhelm Ostwald received the paper the same day his daughter was born, and suggested that the paper was the more exciting of the two events. He journeyed to Arrhenius’ hometown of Uppsala, Sweden to try to convince Arrhenius to work with him; Arrhenius refused for personal reasons but later got a scholarship and worked with the top physicists in Europe. Arrhenius became a professor in a prestigious university about ten years after presenting his “ridiculed” paper, and won the Nobel Prize ten years after that.

HANS ALFVEN:

Astronomers thought that gravity alone is important in solar systems, in galaxies, etc. Alfven’s idea that plasma physics is of equal or greater importance to gravity was derided for decades.

This isn’t a great description of Alfven’s conflict with the establishment, but the list seems basically right insofar as Alfven’s ideas were ignored for thirty years before being proven mostly correct. I will give them this one.

JOHN BAIRD:

When the first television system was demonstrated to the Royal Society (British scientists,) they scoffed and ridiculed, calling Baird a swindler.

I can’t find any reference to this in various Baird articles and biographies. The closest I can come is [this article](#) by someone who was there at the demonstration, who said “They didn’t believe it...the pictures were a bit of a blur but it was amazing, they were all absolutely flabbergasted by it.” It looks like he is using “they didn’t believe it” in the colloquial way of “they thought it was amazing”. A [TIME magazine](#) article from the time described the same scientists as “deeply impressed”, though the wording is kind of unclear and they might have been referring to a different demonstration a year later.

In any case, it seems very clear that within a year everyone agreed he was legitimate and overcame their initial shock.

ROBERT BAKKER:

Everyone knows that dinosaurs are like Gila monsters or big tortoises: large, slow, and intolerant of the cold. And they're all colored olive drab too! 😊

Bakker did help produce the paradigm shift in paleontology from cold-blooded dinosaurs to warm-blooded dinosaurs. But he was not a lone maverick being ridiculed by everyone else. He learned that dinosaurs were warm-blooded from his professor at Yale, who was *also* part of the minority-but-totally-existing faction that believed dinosaurs were warm-blooded. He himself got a PhD at Harvard from professors who were apparently sympathetic to the same theory. And within seven years of his first paper being published, Scientific American was calling his ideas "[the dinosaur renaissance](#)", which doesn't leave a lot of time for him to be ridiculed and ignored in.

BARDEEN & BRATTAIN:

Not ridiculed, but their boss W. Shockley nixed their idea for a non-FET "crystal triode" device. When they started investigating it, he made them stop. They were supposed to be working on FETs instead.

ARG, I GOT THIS WRONG, THIS PART BELOW IS A BELL LABS STORY REGARDING ZONE REFINING OF SILICON, NOT THE BJT TRANSISTOR PROJECT: So, they assembled their ZONE REFINING experiment on a wheeled cart and continued. Whenever the boss was scheduled to check up on them, they could shove it into an adjacent unused lab.

Okay, it looks like the guy compiling the list admits he was wrong on this one. Moving on...

BRETZ:

Endured decades of scorn as the laughingstock of the geology world. His crime was to insist that enormous amounts of evidence showed that, in Eastern Washington state, the "scabland" desert landscape had endured an ancient catastrophe: a flood of staggering proportions. This was outright heresy, since the geology community of the time had dogmatic belief in a "uniformitarian" position, where all changes must take place slowly and incrementally over vast time scales. Bretz' ideas were entirely vindicated by the 1950s. Quote: "All my enemies are dead, so I have no one to gloat over."

This one is basically right and I'll give it to them.

CHANDRASEKHAR:

Chandra originated Black Hole theory and published several papers. He was attacked viciously by his close colleague Sir Arthur Eddington, and his theory was discredited in the eyes of the research community. They were wrong, and Eddington apparently took such strong action based on an incorrect pet theory of his own. In the end Chandra could not even pursue a career in England, and he moved his research to the U. of Chicago in 1937, laboring in relative obscurity for decades.

Sort of true, but he was hardly shunned by the scientific community. He made his discoveries about black holes in the early 1930s, was well-received by many people, and won a Bronze Medal in some physics competition. In 1935, Eddington attacked his theory, possibly because Eddington was racist and didn't like Indian people. But many other scientists, including Niels Bohr and Wolfgang Pauli, continued to support him (quietly, so as not to offend Eddington, which will be a recurring theme in these kinds of situations). Chandrasekhar was made a Fellow of the Royal Society in 1944, won the Royal Astronomical Society Gold Medal in 1953, and generally led a long and prestigious life. His theories were resurrected once people had better evidence that black holes existed. I'll give this one half a point.

CHLADNI:

The scientific community regarded Meteorites in the same way that modern scientists regard UFO abductions and psychic phenomenon: quaint superstitions only believed by peasant folk. All the eyewitness reports were disbelieved. At one point the ridicule became so intense that many museums with meteorites in their geology collections decided to trash those valuable samples. (Sometimes hostile skepticism controls reality, and the strongest evidence is edited to conform to consensus disbeliefs.) Finally in the early 1800's Ernst Chladni actually sat down and inspected the evidence professionally, and found that claimed meteorites were entirely unlike known earth rocks. His study changed some minds. At the same time some large meteor falls were witnessed by scientists, and the majority who insisted that only ignorant peasants ever saw such things were shamed into silence.

As the quote points out, this is a kind of weird one as meteorite work was ridiculed for a long time, but Chladni was taken seriously and helped change minds. Looking at Wikipedia, a lucky meteorite fall two years after Chladni first published his theory helped turn the tide in his favor, and by ten years after publication Chladni's meteorite theories were pretty well-regarded. Even when people disagreed with him about meteorites, Chladni remained widely respected for some of his other work in acoustics.

There *is* a story here, but it's probably not right to center it around Chladni, and his work was only scorned for a few years before everyone agreed it was true. I'll give this another half a point.

CRICK & WATSON

Not ridiculed. But they were instructed to drop their research. They continued it as "bootleg" research.

The list admits they were "not ridiculed". They were told to stop their research because there was all sorts of academic politics around who was going to be the first to discover DNA, and the guy in charge of their university was rooting for another team.

DOPPLER

Proposed a theory of the optical Doppler Effect in 1842, but was bitterly opposed for two decades because it did not fit with the accepted physics of the time (it contradicted the Luminiferous Aether theory.) Doppler was finally proven right in 1868 when W. Huggins observed red shifts and blue shifts in stellar spectra. Unfortunately this was fifteen years after Doppler had died.

I haven't been able to find anything about this in various short online biographies of Doppler ([1](#), [2](#)). Doppler tested the effect himself by having someone play a trumpet on a train (really), someone else successfully tested it in 1845, and it was independently rediscovered in 1848. Doppler himself was made the head of the Institute For Experimental Physics in Vienna and died about as prestigious and beloved as a physicist can get.

So my impression is that only a third of these people really fit the pattern. Most of them were doubted for very short periods, continued to be respected in their fields for their other accomplishments even during those periods, or were part of medium-sized movements rather than being lone geniuses. After a few years – maybe an average of ten, very rarely as long as thirty – their contributions were recognized and they assumed their rightful place in the pantheon. Science isn't perfect. But it is *darned good*.

[EDIT: Bill Beatty, author of the original list, responds [here](#). My response to the response [here](#).]

II.

I bring this up in the context of [my last post](#) on progress in the rationalist movement. There used to be a stereotype that rationalists were too quick to challenge scientific consensus. I think that was exaggerated, but based on a core of truth. Given that we're interested in the ways that bias can prevent people from accepting truth, it's unsurprising that we would focus on cases like these.

But I personally have changed my thinking on this a lot. Not in any way that I can explain explicitly – I've always thought something like:

Scientific consensus is the best tool we have for seeking truth. It's not perfect, and it's frequently overturned by later scientists, but this is usually – albeit not literally always – the work of well-credentialed insiders, operating pretty quickly after the evidence that should overturn it becomes available. Any individual should be very doubtful of their ability to beat it, while not being so doubtful that nobody ever improves it and science can never progress.

– and I still think that. But I've shifted from being the sort of person who shares viral lists of maligned geniuses, to the sort of person who debunks those lists. I've started emphasizing the “best tool we have” part of the sentence, and whispering the “isn't perfect” part, rather than vice versa.

I've changed my mind on this because of personal experience. Rather than trying to describe it, it might be more helpful to give the most salient examples.

1. The Replication Crisis: I previously thought the scientific consensus was flawed because it failed to take the replication crisis seriously enough. I later learned that everyone else took the replication crisis exactly as seriously as I did. A [poll in Nature](#) shows that 90% of scientists believe reproducibility issues constitute a “crisis”, compared to only 3% (!) who don't. For every person complaining about [“methodological terrorists”](#), there are a dozen who are very concerned and trying to change the way they practice research.

This is especially impressive because as far as I can tell the whole shift happened in about ten years. I would date the beginning of the crisis from Ioannidis' [original 2005 paper](#), although it was only aimed at medicine. It got into high gear in psychology

sometime around 2011 with Simonsohn's [False Positive Psychology](#). A [Google Trends analysis](#) suggests people only started searching the relevant keywords around 2013.

I started thinking about this sort of thing in 2009 after reading [this LW post](#). At the time I thought this was some sort of exciting failure of modern science that I alone had figured out. But this was well after sharp people like Ioannidis were talking about it, and only a few years before *everyone* was talking about it. Framing this as "I was right and scientific consensus was wrong" seems grandiose. Better might be "I started betting on a winning horse about a quarter of the way between the beginning of the race and when its victory became blatantly obvious to everyone".

2. Nutrition: The Bad Old Paradigm of nutrition says that obese people just have poor impulse control, that weight is a simple matter of calories in vs. calories out, and that all calories are equally good except fat, which for some inexplicable reason is the Devil. Anybody who's read a few good books about nutrition science knows that the Bad Old Paradigm is woefully inadequate. I read a few of those books and became convinced that I was right and scientific consensus was wrong.

Unfortunately, this whole issue exploded when Gary Taubes published *Good Calories, Bad Calories*, which as best I can tell combined the first publicly available good critique of the Bad Old Paradigm with a flawed and basically false attempt at a new paradigm. There were lots of confused attacks against Taubes' bad information which did collateral damage to his good information, and lots of confused defenses of his good information which inadvertently shielded his bad information from criticism. I previously focused on [defend the good parts](#), but recently shifted more towards [criticizing the bad parts](#).

After reading some more good books here (one of which I hope to review soon), my impression is that most nutrition scientists don't believe in the Bad Old Paradigm and haven't for a while. At the very least, most of them seem to believe in the lipostat and think it's important, which is my proxy for "basically has their heart in the right place". Insofar as the Bad Old Paradigm continues to be popular wisdom, it's because of the diet industry, the government, social inertia, and nobody really having a good new paradigm to replace it with. I'm gradually seeing popular wisdom shift, and nutrition scientists themselves seem to be helping this process rather than hurting it.

Maybe somebody in this area has discovered the new paradigm and is a maverick being persecuted by hidebound reactionaries. But it isn't Gary Taubes. And it certainly isn't me.

3. Social-Justice-Related Issues: Another narrative I used to believe was that a lot of sketchy ideas were being flattered because they spoke to left-leading academics' biases in favor of social justice. Implicit association tests, stereotype threat, the idea of zero meaningful psychological differences between men and women, et cetera.

When I started worrying about implicit association tests, I thought I was defying some kind of broad scientific consensus. But the meta-analyses showing the Implicit Association Test didn't do what people thought had been around since [2009](#) and have only gotten [more numerous](#) since then, with [broad media coverage](#). Problems with stereotype threat research are getting [mainstream coverage](#) and even [airtime on NPR](#).

The problem here is that there was no equivalent of the *Nature* poll on the replication crisis, so I didn't realize any of this was happening until just recently. For example, in 2016 [this Vox explainer](#) made it sound like there was a monolithic consensus in favor of Implicit Association Tests that no sane person had ever disagreed with, even though

by that point there were already several big meta-analyses finding they weren't practically useful. The correct conclusion isn't that this is really what scientific consensus thinks. The correct conclusion is that Vox shouldn't be trusted about any science more complicated than the wedge vs. inclined plane. Once I realized that there was all this intelligent analysis going on that I'd never heard about, my claim to be boldly defying the scientific consensus evaporated.

Yes, Cordelia Fine is still around and is still writing books arguing against gender differences. But she's starting to sound *really* defensive, basically the literary equivalent of "I know I'm going to be downvoted to hell for this, but...". Meanwhile, other scientists are doing a good job pointing out the flaws in her books and conducting studies like [this biggest-ever look at male vs. female brain differences](#), [this magisterial look at personality differences](#), et cetera – not to mention great and widely-accepted work on how intersex people take on more characteristics of their hormonal than their social gender (honestly, we should probably thank transgender people for making this field socially acceptable again). People talk a lot about how Larry Summers was fired from Harvard for talking about male vs. female differences, but Steven Pinker [did a whole debate on this](#) and remains a Harvard professor.

Even things about genetic psychological differences between population groups are less bold and maverick-y than their proponents like to think. The relevant surveys I know trying to elicit scientific consensus ([1](#), [2](#), [3](#)) all find that, when asked anonymously, most scientists think these differences explain about 25% – 50% of variance.

I hate to bring that up, because it'll probably start a flame war in the comments, but I think it's important as a sign of exactly how hard it is to politicize science. Global warming skeptics talk about how maybe the scientific consensus on global warming is false because climatologists face political pressure to bias their results in favor of the theory. But scientists studying these areas face much more political pressure, and as long as you give the surveys anonymously they're happy to express horrendously taboo opinions. This is about the strongest evidence in favor of the consensus on global warming – and scientific consensus in general – that I could imagine.

4. Nurture Assumption and Blank Slatism: The prologue of the first edition of *The Nurture Assumption* is Judith Rich Harris telling her "maverick genius kept down by hidebound reactionaries" story. But the prologue of the second edition is her being much more hopeful:

To some extent at least, times have changed...there is now more acceptance of the idea that behavior is influenced by genes and that individual differences in behavior are due in part to differences in genes. People are more willing to admit that children can inherit behavioral quirks and personality characteristics...was it this cultural shift that led to greater acceptance of my theory? Or was it the fact that new findings, consistent with the theory, kept turning up? Over time, the early, angry response to *The Nurture Assumption* has softened noticeably, both within and outside of academia. Today, the book is widely cited in textbooks and journal articles. It's assigned and discussed in courses in many colleges and universities; it shows up in exams...in his foreword to the first edition of *The Nurture Assumption*, Steven Pinker made a rash prediction about the book: "I predict it will come to be seen as a turning point in the history of psychology". Perhaps it is too soon to judge whether psychology has rounded a bend; perhaps it will take the perspective of twenty or thirty years. Even at this point, though, there are signs of a slight shift in direction. Within developmental psychology, I've

noticed that descriptions of procedures and results are beginning to sound a bit defensive. Greater progress has been made in other areas of psychology. And the email I receive from students gives me high hopes for the younger generation coming up.

There were ten years between the first and second editions of *The Nurture Assumption*. In the almost ten years since the publication of the second edition, my impression is that its ideas have become even more widely-accepted. This month's edition of the *American Journal of Psychiatry*, one of the top journals in the field, has a great study showing that child abuse does *not* cause cognitive disability, in contrast to several previous studies in the area. It cites Deary, Plomin, and Ioannidis, hits all of the talking points about genetic confounding of developmental outcomes, and receives glowing endorsement in the journal's editorial section, which says that "if our causal explanations are wrong, we may be wasting our effort or even doing damage". Every single psychiatrist in the country is getting exposed to this way of thinking.

And this has real results. I got to present a summary of behavioral genetics to a meeting of psychiatrists, including a lot of psychoanalysts, and I was shocked that most of them were at least a little receptive. I think they misunderstood it. I think they carefully raised caveats in exactly the right places to ensure they didn't have to change anything they were doing. But the *overall* response was "Oh, yeah, we've heard stuff like that, it seems plausible, good thing that for various hard-to-explain reasons none of it applies to us." This is what the first stage of progress looks like.

5. Intelligence Explosion And AI Risk: This was another place where I and many of my friends thought we were right and the consensus was wrong. It was another place where a lot of self-appointed defenders of the consensus told us we were crackpots and needed to listen to what real scientists thought. And again, when I looked into it, there was no consensus against the idea and [lots of prominent researchers were in favor](#). Going to the [Asilomar Conference](#) and seeing a bunch of people from MIT and Harvard talk about how concerned they were really opened my eyes on this. Google now has an AI Ethics Board, Berkeley, Oxford, and MIT have foundations working on it, and people like Elon Musk and Bill Gates are involved. [Bostrom's survey of AI researchers](#) and some more recent and rigorous not-yet-published surveys I've heard about confirm the impression. Nobody would ever say there's a scientific consensus *in favor* of Bostrom's theories. But at this point I think it's also indefensible to say there's a consensus against.

Bostrom first started writing about these sorts of things extensively in the early 2000s, so there was really only a ten-year gap between entering the intellectual environment and it becoming a (mostly) accepted part of the established field. Those ten years felt pretty long while we were in them, but the ability of a field to accept an on-the-face-of-it completely-insane-sounding theory within ten years seems to me a very strong argument against the hidebound-reactionaries theory and a very strong argument for considering scientific consenses to be unreasonably effective.

6. IQ: Another case where I worried about apparent failure of scientific consensus due to political bias. I certainly encountered a lot of falsehoods around this when I was younger. My high school psychology textbook included a section claiming that all IQ tests were biased towards rich white people because they were based entirely on questions like "how many shots below par is a bogey?" Then it presented an "alternate IQ test" which "proved" that poor minorities had higher IQs than rich whites by asking some other questions with the opposite bias (I think they were about slang for drugs - certainly an interesting way to fight stereotypes). This kind of thing

naturally made me assume that nobody had any idea what was actually in IQ tests and scientists were idiots.

But more recently I've been reading [actual surveys](#), which find that about 97% of expert psychologists and 85% of applied psychologists agree that IQ tests measure cognitive ability "reasonably well". And 77% of expert psychologists and 63% of applied psychologists agree IQ tests are culture-fair (with slightly different numbers depending on how you ask the question, but always about 50% of both groups).

This seems like less of a problem with expert consensus, and more of a problem of nobody else (including textbook writers!) listening to experts who are continually trying to beat reality into people's heads. But I have a vague memory of having recently seen a survey (which I can't find) that even experts in softer fields like sociology are generally in favor of IQ and admit that it has its uses. And even some left/liberal sources like [Vox](#) and [Freddie deBoer](#) are aware of the consensus and willing to respect it.

At the same time, I've encountered some people like [Borsboom](#) and [Nostalgebraist](#) who have relatively sophisticated (and limited) critiques of IQ, and who have allowed me to round off other people's less-well-framed critiques to something more like what they are saying and less like the stupid things my high school textbook said.

So it seems to me that generally experts agree with reasonable statements about IQ, and where they seem to disagree they may hold reasonable disagreements rather than unreasonable ones. Again, where this fails is not in the experts but in the ability of people who don't listen to the experts to get disproportionate social power and hide the existence of the expert consensus.

III.

Last week [I wrote](#) about universally-known criticisms of economists, like "they're silly for assuming everyone behaves perfectly rationally":

My impression is that economists not only know about these criticisms, but invented them. During the last few paradigm shifts in economics, the new guard levied these complaints against the old guard, mostly won, and their arguments percolated down into the culture as The Correct Arguments To Use Against Economics. Now the new guard is doing their own thing - behavioral economics, experimental economics, economics of effective government intervention. The new paradigm probably has a lot of problems too, but it's a pretty good bet that random people you stop on the street aren't going to know about them.

The same pattern explains a lot of my concerns above. I knew some criticisms of a scientific paradigm. They seemed right. I concluded that scientists weren't very smart and maybe I was smarter. I should have concluded that some cutting-edge scientists were making good criticisms of an old paradigm. I can still flatter myself by saying that it's no small achievement to recognize a new paradigm early and bet on the winning horse. But the pattern I was seeing was part of the process of science, not a condemnation of it.

Most people understand this intuitively about past paradigm shifts. When a creationist says that we can't trust science because it used to believe in phlogiston and now it believes in combustion, we correctly respond that this is exactly why we *can* trust science. But this lesson doesn't always generalize when you're in the middle of a paradigm shift *right now* and having trouble seeing the other side.

I realize I'm (ironically) risking making my narrative of scientific success unfalsifiable. Suppose someone wants to argue that scientific consensus is wrong. If they point to something it used to be wrong about, I can respond "Yes, but it self-corrected and it's correct now, so that's fine." If they point to something where cutting-edge scientists say it's wrong but nobody else agrees, I can respond "Yes, this is what the beginning of a paradigm shift looks like, so that's fine". And if they point to something where *nobody* in the field thinks it's wrong, I can say "You're a crackpot for going against all reputable scientists; the problem is with you." And if later they turn out to be right, and everyone acknowledges it, I can say "Yes, but it self-corrected and it's correct now, so that's fine."

(and I'm making it even easier for myself in that I say "scientific consensus for" when I probably mean "no scientific consensus against". I don't claim that 90%+ of scientists always believe true things, only that there are very few cases where 90%+ of scientists believe things which smarter people know to be false.)

Against this I can only offer a personal narrative: the only light I have by which to judge scientific consensus is my own Inside View assessment of what seems correct. Again and again I have *tried* to defy scientific consensus. And every time, I either find that I am wrong, find that I am a few years ahead of a trend that most scientists eventually agree with, or find that what I thought was "scientific consensus" was actually a fiction peddled by biased industry or media sources slandering a scientific community which actually had a much more sophisticated picture. My history of trying to fight scientific consensus has been a *Man Who Was Thursday*-esque series of embarrassments as I find again and again that my supposed enemy agrees with me and is even better at what I am trying to do than I am.

Scientific consensus hasn't just been accurate, it's been *unreasonably* accurate. Humans are fallible beings. They are not known for their ability to change their mind, to willingly accept new information, or to put truth-seeking above political squabbles. And our modern society is not exactly known for being an apolitical philosopher-kingdom with strong truth-seeking institutions completely immune from partisan pressure. I feel a deep temptation to sympathize with global warming denialists who worry that the climatological consensus is biased politicized crap, because that is *exactly* the sort of thing which I would expect to come out of our biased politicized crappy society. Yet again and again I have seen examples of scientific fields that have maintained strong commitments to the truth in the face of pressure that would shatter any lesser institution. I've seen fields where people believe incredibly-bizarre sounding things that will get them mocked at cocktail parties just because those things seem to be backed by the majority of the evidence. I've even seen people *change their minds*, in spite of all the incentives to the contrary. I can't explain this. The idea that scientific consensus is almost always an accurate reflection of the best knowledge we have at the time seems even more flabbergasting than any particular idea that scientists might or might not believe. But it seems to be true.

(note that I'm talking about "scientific consensus" to mean a very high-level pattern, consisting of hundreds of scientists over the space of decades evaluating a broad body of work. Any individual study is still probably total garbage.)

Given how weird all of this is, I realize there's another possible bias here that should be taken very seriously – which is that I'm wrong about one or both sides of this. Which is more likely: that Science always agrees with Truth? Or that one guy's perception of Science always agrees with that same guy's perception of Truth? The latter gives me two degrees of freedom: I can either cherry-pick experts who agree

with me and declare them to be Consensus, or I can conform my opinions to consensus so slavishly that I end up discovering only that Consensus agrees with itself. I don't *feel* like I'm making this kind of mistake. But then again, nobody ever *feels like* they're being biased.

But if I'm making this mistake, I think it's at least a *better* mistake than the one where people dream up stories about being mavericks persecuted by hidebound reactionaries. This mistake at least sets the terms of debate as "let's try to ascertain what the scientific community thinks" and forbids me from believing *completely* crackpottish things. And it encourages trust in one of our more trustworthy public institutions, always a prosocial sort of thing to do. I would rather have a world of people debating who agrees with scientific consensus or not, than a world of people debating whether scientific consensus is even valuable.

There are two caveats to the above. First, I think it's dangerous to promote a norm of agreeing with scientific consensus, insofar as that helps encourage exactly the mistakes about the nature of consensus that I discussed above. When poorly-informed diet industry gurus support the Bad Old Paradigm, their rallying cry is usually "You're a stupid crackpot, bow to the scientific consensus which agrees with me". I gave three examples above of cases where I would have gotten the scientific consensus 100% wrong if I didn't have access to a formal survey of scientific experts. In a world where these surveys had never been done – or some existing field without these surveys – or some field where these surveys have been done inaccurately or in a biased manner – people will often believe the consensus to be the opposite of what it really is. In those cases, demands that people respect consensus can be used to shut down people who are actually right – the field-wide equivalent of calling true facts you don't like [debunked and well-refuted](#). I see this happening all the time and I worry that waxing too poetically about the unreasonable effectiveness of scientific consensus will only serve to empower these people. Goodhart's Law says that a measure which becomes a target ceases to be a useful measure, so we should be reluctant to target scientific consensus too strongly.

And second, I think that even when the Outside View tells you that the consensus is correct, you should continue pursuing your Inside View hunch that it isn't. This avoids awkward situations like every individual scientist doubting the consensus, but suppressing their doubts because the "scientific consensus" has to be right.

So maybe the things I'm saying about scientific consensus aren't very actionable. But respecting scientific consensus in a non-actionable way is a lot less exhausting than believing yourself to be against it, and talking about how you're against it, and taking flak for being against it. And in the same way it's helpful to believe that God is good, even if He never really gets around to doing much about it, so it's reassuring to be able to have faith in our institutions every so often.

My IRB Nightmare

[Epistemic status: Pieced together from memory years after the event. I may have mis-remembered some things or gotten them in the wrong order. Aside from that – and the obvious jokes – this is all true. I’m being deliberately vague in places because I don’t want to condemn anything specific without being able to prove anything.]

September 2014

There’s a screening test for bipolar disorder. You ask patients a bunch of things like “Do you ever feel really happy, then really sad?”. If they say ‘yes’ to enough of these questions, you start to worry.

Some psychiatrists love this test. I hate it. Patients will say “Yes, that absolutely describes me!” and someone will diagnose them with bipolar disorder. Then if you ask what they meant, they’d say something like “Once my local football team made it to the Super Bowl and I was really happy, but then they lost and I was really sad.” I don’t even want to tell you how many people get diagnosed bipolar because of stuff like this.

There was a study that supposedly proved this test worked. But parts of it confused me, and it was done on a totally different population that didn’t generalize to hospital inpatients. Also, it said in big letters THIS IS JUST A SCREENING TEST IT IS NOT INTENDED FOR DIAGNOSIS, and everyone was using it for diagnosis.

So I complained to some sympathetic doctors and professors, and they asked “Why not do a study?”

Why *not* do a study? Why not join the great tradition of scientists, going back to Galileo and Newton, and make my mark on the world? Why not replace my griping about bipolar screening with an experiment about bipolar screening, an experiment done to the highest standards of the empirical tradition, one that would throw the entire weight of the scientific establishment behind my complaint? I’d been writing about science for so long, even doing my own informal experiments, why not move on to join the big leagues?

For (it would turn out) a whole host of excellent reasons that I was about to learn.

A spring in my step, I journeyed to my hospital’s Research Department, hidden in a corner office just outside the orthopaedic ward. It was locked, as always. After enough knocking, a lady finally opened the door and motioned for me to sit down at a paperwork-filled desk.

“I want to do a study,” I said.

She looked skeptical. “Have you done the Pre-Study Training?”

I had to admit I hadn’t, so off I went. The training was several hours of videos about how the Nazis had done unethical human experiments. Then after World War II, everybody met up and decided to only do ethical human experiments from then on. And the most important part of being ethical was to have all experiments monitored by an Institutional Review Board (IRB) made of important people who could check whether experiments were ethical or not. I dutifully parroted all this back on the post-test (“Blindly trusting authority to make our ethical decisions for us is the *best* way to separate ourselves from the Nazis!”) and received my Study Investigator Certification.

I went back to the corner office, Study Investigator Certification in hand.

"I want to do a study," I said.

The lady still looked skeptical. "Do you have a Principal Investigator?"

Mere resident doctors weren't allowed to do studies on their own. They would probably screw up and start building concentration camps or something. They needed an attending (high-ranking doctor) to sign on as Principal Investigator before the IRB would deign to hear their case.

I knew exactly how to handle this: one by one, I sought out the laziest attendings in the hospital and asked "Hey, would you like to have your name on a study as Principal Investigator for free while I do all the actual work?" Yet one by one, all of the doctors refused, as if I was offering them some kind of plague basket full of vermin. It was the weirdest thing.

Finally, there was only one doctor left – Dr. W, the hardest-working attending I knew, the one who out of some weird masochistic impulse took on every single project anyone asked of him and micromanaged it to perfection, the one who every psychiatrist in the whole hospital (including himself) had diagnosed with obsessive-compulsive personality disorder.

"Sure Scott," he told me. "I'd be happy to serve as your Principal Investigator".

A feeling of dread in my stomach, I walked back to the tiny corner office.

"I want to do a study," I said.

The lady still looked skeptical. "Have you completed the New Study Application?" She gestured to one of the stacks of paperwork filling the room.

It started with a section on my research question. Next was a section on my proposed methodology. A section on possible safety risks. A section on recruitment. A section on consent. A section on...wow. Surely this can't *all* be the New Study Application? Maybe I accidentally picked up the Found A New Hospital Application?

I asked the lady who worked in the tiny corner office whether, since I was just going to be asking bipolar people whether they ever felt happy and then sad, maybe I could get the short version of the New Study Application?

She told me that *was* the short version.

"But it's twenty-two pages!"

"You haven't done any studies before, have you?"

Rather than confess my naivete, I started filling out the twenty-two pages of paperwork. It started by asking about our study design, which was simple: by happy coincidence, I was assigned to Dr. W's inpatient team for the next three months. When we got patients, I would give them the bipolar screening exam and record the results. Then Dr. W. would conduct a full clinical interview and formally assess them. We'd compare notes and see how often the screening test results matched Dr. W's expert diagnosis. We usually got about twenty new patients a week; if half of them were

willing and able to join our study, we should be able to gather about a hundred data points over the next three months. It was going to be easy-peasy.

That was the first ten pages or so of the Application. The rest was increasingly bizarre questions such as “Will any organs be removed from participants during this study?” (Look, I promise, I’m not a Nazi).

And: “Will prisoners be used in the study?” (COME ON, I ALREADY SAID I WASN’T A NAZI).

And: “What will you do if a participant dies during this research?” (If somebody dies while I’m asking them whether they sometimes feel happy and then sad, I really can’t even promise so much as “not freaking out”, let alone any sort of dignified research procedure).

And more questions, all along the same lines. I double-dog swore to give everybody really, really good consent forms. I tried my best to write a list of the risks participants were taking upon themselves (mostly getting paper cuts on the consent forms). I argued that these compared favorably to the benefits (maybe doctors will stop giving people strong psychiatric medications just because their football team made the Super Bowl).

When I was done, I went back to the corner office and submitted everything to the Institutional Review Board. Then I sat back and hoped for the best. Like an idiot.

October 2014

The big day arrived. The IRB debated the merits of my study, examined the risks, and...sent me a letter pointing out several irregularities in my consent forms.

IRREGULARITY #1: Consent forms traditionally included the name of the study in big letters where the patient could see it before signing. Mine didn’t. Why not?

Well, because in questionnaire-based psychological research, you *never* tell the patient what you’re looking for before they fill out the questionnaire. That’s like Methods 101. The name of my study was “Validity Of A Screening Instrument For Bipolar Disorder”. Tell the patient it’s a study about bipolar disorder, and the gig is up.

The IRB listened patiently to my explanation, then told me that this was not a legitimate reason not to put the name of the study in big letters on the consent form. Putting the name of the study on the consent form was important. You know who *else* didn’t put the name of the study on his consent forms? *Hitler*.

IRREGULARITY #2: Consent forms traditionally included a paragraph about the possible risks of the study and a justification for why we believed that the benefits were worth the risks. Everyone else included a paragraph about this on our consent forms, and read it to their patients before getting their consent. We didn’t have one. Why not?

Well, for one thing, because all we were doing was asking them whether they felt happy and then sad sometimes. This is the sort of thing that goes on every day in a psychiatric hospital. Heck, the other psychiatrists were using this same screening test, except *for real*, and they never had to worry about whether it had risks. In the grand scheme of things, this just wasn’t a very risky procedure.

Also, psychiatric patients are sometimes...how can I put this nicely?...a little paranoid. Sometimes you can offer them breakfast and they'll accuse you of trying to poison them. I had no illusions that I would get every single patient to consent to this study, but I felt like I could at least avoid handing them a paper saying "BY THE WAY, THIS STUDY IS FULL OF RISKS".

The IRB listened patiently to my explanation, then told me that this was not a legitimate reason not to have a paragraph about risks. We should figure out some risks, then write a paragraph explaining how those were definitely the risks and we took them very seriously. The other psychiatrists who used this test every day didn't have to do that *because they weren't running a study*.

IRREGULARITY #3: Signatures are traditionally in pen. But we said our patients would sign in pencil. Why?

Well, because psychiatric patients aren't allowed to have pens in case they stab themselves with them. I don't get why stabbing yourself with a pencil is any less of a problem, but the rules are the rules. We asked the hospital administration for a one-time exemption, to let our patients have pens just long enough to sign the consent form. Hospital administration said absolutely not, and they didn't care if this sabotaged our entire study, it was pencil or nothing.

The IRB listened patiently to all this, then said that it had to be in pen. You know who *else* had people sign consent forms in pencil...?

I'm definitely not saying that these were the only three issues the IRB sprung on Dr. W and me. I'm saying these are a *representative sample*. I'm saying I spent several weeks relaying increasingly annoyed emails and memos from myself to Dr. W to the IRB to the lady in the corner office to the IRB again. I began to come home later in the evening. My relationships suffered. I started having dreams about being attacked by giant consent forms filled out in pencil.

I was about ready to give up at this point, but Dr. W insisted on combing through various regulations and talking to various people, until he discovered some arcane rule that certain very safe studies with practically no risk were allowed to use an "expedited consent form", which was a lot like a normal consent form but didn't need to have things like the name of the study on it. Faced with someone even more obsessive and bureaucratic than they were, the IRB backed down and gave us preliminary permission to start our study.

The next morning, screening questionnaire in hand, I showed up at the hospital and hoped for the best. Like an idiot.

November 2014

Things progressed slowly. It turns out a lot of psychiatric inpatients are either depressed, agitated, violent, or out of touch with reality, and none of these are really conducive to wanting to participate in studies. A few of them already delusionally thought we were doing experiments on them, and got confused when we suddenly asked them to consent. Several of them made it clear that they hated us and wanted to thwart us in any way possible. After a week, I only had three data points, instead of the ten I'd been banking on.

"Data points" makes it sound abstract. It wasn't. I had hoped to put the results in the patients' easily accessible online chart, *the same place everyone else put the results*

of the exact same bipolar screening test when they did it for real. They would put it in a section marked TEST RESULTS, which was there to have a secure place where you could put test results, and where everybody's secure test results were kept.

The IRB would have none of this. Study data are Confidential and need to be kept Secure. Never mind that all the patients' *other* secure test results were on the online chart. Never mind that the online chart contains all sorts of stuff about the patients' diagnoses, medications, hopes and fears, and even (remember, this is a psych hospital) secret fetishes and sexual perversions. Study data needed to be encrypted, then kept in a Study Binder in a locked drawer in a locked room that nobody except the study investigators had access to.

The first problem was that nobody wanted to give us a locked room that nobody except us had access to. There was a sort of All Purpose Psychiatry Paperwork room, but the janitors went in to clean it out every so often, and apparently this made it unacceptable. Hospitals aren't exactly drowning in spare rooms that not even janitors can get into. Finally Dr. W grudgingly agreed to keep it in his office. This frequently meant I couldn't access any of the study material because Dr. W was having important meetings that couldn't be interrupted by a resident barging into his office to rummage in his locked cabinets.

But whatever. The bigger problem was the encryption. There was a very specific way we had to do it. We would have a Results Log, that said things like "Patient 1 got a score of 11.5 on the test". And then we'd have a Secret Patient Log, which would say things like "Patient 1 = Bob Johnson from Oakburg." That way nobody could steal our results and figure out that Bob was sometimes happy, then sad.

(meanwhile, all of Bob's actual diagnoses, sexual fetishes, etc were in the easily-accessible secure online chart that we were banned from using)

And then – I swear this is true – we had to keep the Results Log and the Secret Patient Log right next to each other in the study binder in the locked drawer in the locked room.

I wasn't sure I was understanding this part right, so I asked Dr. W whether it made sense, to him, that we put a lot of effort writing our results in code, and then put the key to the code in the same place as the enciphered text. He cheerfully agreed this made no sense, but said we had to do it or else our study would fail an audit and get shut down.

January 2015

I'd planned to get a hundred data points in three months. Thanks to constant bureaucratic hurdles, plus patients being less cooperative than I expected, I had about twenty-five. Now I was finishing my rotation on Dr. W's team and going to a clinic far away. What now?

A bunch of newbies were going to be working with Dr. W for the next three months. I hunted them down and threatened and begged them until one of them agreed to keep giving patients the bipolar screening test in exchange for being named as a co-author. Disaster averted, I thought. Like an idiot.

Somehow news of this arrangement reached the lady in the corner office, who asked whether the new investigator had completed her Pre-Study Training. I protested that she wasn't designing the study, she wasn't conducting any analyses, all she was

doing was asking her patients the same questions that she would be asking them anyway as part of her job for the next three months. The only difference was that she was recording them and giving them to me.

The lady in the corner office wasn't impressed. You know who *else* hadn't thought his lackeys needed to take courses in research ethics?

So the poor newbie took a course on how Nazis were bad. Now she could help with the study, right?

Wrong. We needed to submit a New Investigator Form to the IRB and wait for their approval.

Two and a half months later, the IRB returned their response: Newbie was good to go. She collected data for the remaining two weeks of her rotation with Dr. W before being sent off to another clinic just like I was.

July 2015

Dr. W and I planned ahead. We had figured out which newbies would be coming in to work for Dr. W three months ahead of time, and gotten them through the don't-be-a-Nazi course and the IRB approval process just in time for them to start their rotation. Success!

Unfortunately, we received another communication from the IRB. Apparently we were allowed to use the expedited consent form to get consent for our *study*, but not to get consent to *access protected health information*. That one required a whole different consent form, list-of-risks and all. We were right back where we'd started from.

I made my case to the Board. My case was: we're not looking at any protected health information, f@#k you.

The Board answered that we were accessing the patient's final diagnosis. It said right in the protocol, we were giving them the screening test, then comparing it to the patient's final diagnosis. "Psychiatric diagnosis" sure *sounds* like protected health information.

I said no, you don't understand, we're the psychiatrists. Dr. W is the one making the final diagnosis. When I'm on Dr. W's team, I'm in the room when he does the diagnostic interview, half the time I'm the one who types the final diagnosis into the chart. These are *our patients*.

The Board said this didn't matter. We, as the patient's doctors, would make the diagnosis and write it down on the chart. But we (as study investigators) needed a full signed consent form before we were allowed to access the diagnosis we had just made.

I said wait, you're telling us we have to do this whole bureaucratic rigamarole with all of these uncooperative patients before we're allowed to see something we wrote ourselves?

The Board said yes, exactly.

I don't remember this part very well, except that I think I half-heartedly trained whichever poor newbie we were using that month in how to take a Protected Health

Information Consent on special Protected Health Information Consent Forms, and she nodded her head and said she understood. I think I had kind of clocked out at this point. I was going off to work all the way over in a different town for a year, and I was just sort of desperately hoping that Dr. W and various newbies would take care of things on their own and then in a year when I came back to the hospital I would have a beautiful pile of well-sorted data to analyze. Surely trained doctors would be able to ask simple questions from a screening exam on their own without supervision, I thought. Like an idiot.

July 2016

I returned to my base hospital after a year doing outpatient work in another town. I felt energized, well-rested, and optimistic that the bipolar screening study I had founded so long ago had been prospering in my absence.

Obviously nothing remotely resembling this had happened. Dr. W had vaguely hoped that I was taking care of it. I had vaguely hoped that Dr. W was taking care of it. The various newbies whom we had strategically enlisted had either forgotten about it, half-heartedly screened one or two patients before getting bored, or else mixed up the growing pile of consent forms and releases and logs so thoroughly that we would have to throw out all their work. It had been a year and a half since the study had started, and we had 40 good data points.

The good news was that I was back in town and I could go back to screening patients myself again. Also, we had some particularly enthusiastic newbies who seemed really interested in helping out and getting things right. Over the next three months, our sample size shot up, first to 50, then to 60, finally to 70. Our goal of 100 was almost in sight. The worst was finally behind me, I hoped. Like an idiot.

November 2016

I got an email saying our study was going to be audited.

It was nothing personal. Some higher-ups in the nationwide hospital system had decided to audit every study in our hospital. We were to gather all our records, submit them to the auditor, and hope for the best.

Dr. W, who was obsessive-compulsive at the best of times, became unbearable. We got into late-night fights over the number of dividers in the study binder. We hunted down every piece of paper that had ever been associated with anyone involved in the study in any way, and almost came to blows over how to organize it. I started working really late. My girlfriend began to doubt I actually existed.

The worst part was all the stuff the newbies had done. Some of them would have the consent sheets numbered in the upper left-hand-corner instead of the upper-right-hand corner. Others would have written the patient name down on the Results Log instead of the Secret Code Log right next to it. One even wrote something in green pen on a formal study document. It was hopeless. Finally we just decided to throw away all their data and pretend it had never existed.

With that decision made, our work actually started to look pretty good. As bad as it was working for an obsessive-compulsive boss in an insane bureaucracy, at least it had the advantage that – when nitpicking push came to ridiculous shove – you were going to be super-ready to be audited. I hoped. Like an idiot.

December 2016

The auditor found twenty-seven infractions.

She was very apologetic about it. She said that was actually a pretty good number of infractions for a study this size, that we were actually doing pretty well compared to a lot of the studies she'd seen. She said she absolutely wasn't going to shut us down, she wasn't even going to censure us. She just wanted us to make twenty-seven changes to our study and get IRB approval for each of them.

I kept the audit report as a souvenir. I have it in front of me now. Here's an example infraction:

The data and safety monitoring plan consists of 'the Principal Investigator will randomly check data integrity'. This is a prospective study with a vulnerable group (mental illness, likely to have diminished capacity, likely to be low income) and, as such, would warrant a more rigorous monitoring plan than what is stated above. In addition to the above, a more adequate plan for this study would also include review of the protocol at regular intervals, on-going checking of any participant complaints or difficulties with the study, monitoring that the approved data variables are the only ones being collected, regular study team meetings to discuss progress and any deviations or unexpected problems. Team meetings help to assure participant protections, adherence to the protocol. Having an adequate monitoring plan is a federal requirement for the approval of a study. See Regulation 45 CFR 46.111 Criteria For IRB Approval Of Research. IRB Policy: PI Qualifications And Responsibility In Conducting Research. Please revise the protocol via a protocol revision request form. Recommend that periodic meetings with the research team occur and be documented.

Among my favorite other infractions:

1. The protocol said we would stop giving the screening exam to patients if they became violent, but failed to rigorously define "violent".
2. We still weren't educating our patients enough about "Alternatives To Participating In This Study". The auditor agreed that the only alternative was "not participating in this study", but said that we had to tell every patient that, then document that we'd done so.
3. The consent forms were still getting signed in pencil. We are never going to live this one down. If I live to be a hundred, representatives from the IRB are going to break into my deathbed room and shout "YOU LET PEOPLE SIGN CONSENT FORMS IN PENCIL, HOW CAN YOU JUSTIFY THAT?!"
4. The woman in the corner office who kept insisting everybody take the Pre-Study Training...hadn't taken the Pre-Study Training, and was therefore unqualified to be our liaison with the IRB. I swear I am not making this up.

Faced with submitting twenty-seven new pieces of paperwork to correct our twenty-seven infractions, Dr. W and I gave up. We shredded the patient data and the Secret Code Log. We told all the newbies they could give up and go home. We submitted the Project Closure Form to the woman in the corner office (who as far as I know still hasn't completed her Pre-Study Training). We told the IRB that they had won, fair and square; we surrendered unconditionally.

They didn't seem the least bit surprised.

August 2017

I've been sitting on this story for a year. I thought it was unwise to publish it while I worked for the hospital in question. I still think it's a great hospital, that it delivers top-notch care, that it has amazing doctors, that it has a really good residency program, and even that the Research Department did everything it could to help me given the legal and regulatory constraints. I don't want this to reflect badly on them in any way. I just thought it was wise to wait a year.

During that year, Dr. W and I worked together on two less ambitious studies, carefully designed not to require any contact with the IRB. One was a case report, the other used publicly available data.

They won 1st and 2nd prize at a regional research competition. I got some nice certificates for my wall and a little prize money. I went on to present one of them at the national meeting of the American Psychiatric Association, a friend helped me write it up formally, and it was recently accepted for publication by a medium-tier journal.

I say this not to boast, but to protest that I'm not as much of a loser as my story probably makes me sound. I'm capable of doing research, I think I have something to contribute to Science. I still think the bipolar screening test is inappropriate for inpatient diagnosis, and I still think that patients are being harmed by people's reliance on it. I still think somebody should look into it and publish the results.

I'm just saying it's not going to be me. I am *done* with research. People keep asking me "You seem really into science, why don't you become a researcher?" Well...

I feel like a study that realistically could have been done by one person in a couple of hours got dragged out into hundreds of hours of paperwork hell for an entire team of miserable doctors. I think its scientific integrity was screwed up by stupid requirements like the one about breaking blinding, and the patients involved were put through unnecessary trouble by being forced to sign endless consent forms screaming to them about nonexistent risks.

I feel like I was dragged almost to the point of needing to be in a psychiatric hospital myself, while my colleagues who just *used* the bipolar screening test – without making the mistake of trying to check if it works – continue to do so without anybody questioning them or giving them the slightest bit of aggravation.

I feel like some scientists do amazingly crappy studies that couldn't possibly prove anything, but get away with it because they have a well-funded team of clerks and secretaries who handle the paperwork for them. And that I, who was trying to do everything right, got ground down with so many pointless security-theater-style regulations that I'm never going to be able to do the research I would need to show they're wrong.

In the past year or so, I've been gratified to learn some other people are thinking along the same lines. Somebody linked me to [The Censor's Hand](#), a book by a law/medicine professor at the University of Michigan. A summary from [a review](#):

Schneider opens by trying to tally the benefits of IRB review. "Surprisingly," he writes, a careful review of the literature suggests that "research is not especially dangerous. Some biomedical research can be risky, but much of it requires no

physical contact with patients and most contact cannot cause serious injury. Ill patients are, if anything, safer in than out of research.” As for social-science research, “its risks are trivial compared with daily risks like going online or on a date.”

Since the upsides of IRB review are likely to be modest, Schneider argues, it’s critical to ask hard questions about the system’s costs. And those costs are serious. To a lawyer’s eyes, IRBs are strangely unaccountable. They don’t have to offer reasons for their decisions, their decisions can’t be appealed, and they’re barely supervised at the federal level. That lack of accountability, combined with the gauzy ethical principles that govern IRB deliberations, is a recipe for capriciousness. Indeed, in Schneider’s estimation, IRBs wield coercive government power—the power to censor university research—without providing due process of law.

And they’re not shy about wielding that power. Over time, IRB review has grown more and more intrusive. Not only do IRBs waste thousands of researcher hours on paperwork and elaborate consent forms that most study participants will never understand. Of greater concern, they also superintend research methods to minimize perceived risks. Yet IRB members often aren’t experts in the fields they oversee. Indeed, some know little or nothing about research methods at all.

IRBs thus delay, distort, and stifle research, especially research on vulnerable subgroups that may benefit most from it. It’s hard to precise about those costs, but they’re high: after canvassing the research, Schneider concludes that “IRB regulation annually costs thousands of lives that could have been saved, unmeasurable suffering that could have been softened, and uncountable social ills that could have been ameliorated.”

This view seems to be growing more popular lately, and has gotten support from high-profile academics like Richard Nisbett and Steven Pinker:

Should IRBs (human subjects research approval committees) be dismantled?
[Probably yes.] <http://t.co/5mxhEycEA5>

— Steven Pinker (@sapinker) [July 24, 2015](#)

And there’s been some recent reform, maybe. The federal Office for Human Research Protections [made a vague statement](#) that perhaps studies that obviously aren’t going to hurt anybody might not need the full IRB treatment. There’s still a lot of debate about how this will be enforced and whether it’s going to lead to any real-life changes. But I’m glad people are starting to think more about these things.

(I’m also glad people are starting to agree that getting rid of a little oversight for the lowest-risk studies is a good compromise, and that we don’t have to start with anything more radical.)

I sometimes worry that people misunderstand the case against bureaucracy. People imagine it’s Big Business complaining about the regulations preventing them from steamrolling over everyone else. That hasn’t been my experience. Big Business – heck, Big Anything – loves bureaucracy. They can hire a team of clerks and secretaries and middle managers to fill out all the necessary forms, and the rest of the company can be on their merry way. It’s everyone else who suffers. The amateurs, the entrepreneurs, the hobbyists, the people doing something as a labor of love. Wal-Mart

is going to keep selling groceries no matter how much paperwork and inspections it takes; the poor immigrant family with the backyard vegetable garden might not.

Bureaucracy in science does the same thing: limit the field to big institutional actors with vested interests. No amount of hassle is going to prevent the Pfizer-Merck-Novartis Corporation from doing whatever study will raise their bottom line. But enough hassle *will* prevent a random psychiatrist at a small community hospital from pursuing his pet theory about bipolar diagnosis. The more hurdles we put up, the more the scientific conversation skews in favor of Pfizer-Merck-Novartis. And the less likely we are to hear little stuff, dissenting voices, and things that don't make anybody any money.

I'm not just talking about IRBs here. I could write a book about this. There are so many privacy and confidentiality restrictions around the most harmless of datasets that research teams won't share data with one another (let alone with unaffiliated citizen scientists) lest they break some arcane regulation or other. Closed access journals require people to pay thousands of dollars in subscription fees before they're allowed to read the scientific literature; open-access journals just shift the burden by requiring scientists to pay thousands of dollars to publish their research. Big research institutions have whole departments to deal with these kinds of problems; unaffiliated people who just want to look into things on their own are out of luck.

And this is happening at the same time we're becoming increasingly aware of the shortcomings of big-name research. [Half of psychology studies](#) fail replication; my own field of psychiatry [is even worse](#). And citizen-scientists and science bloggers are playing a big part in debunking bad research: here I'm thinking especially of statistics bloggers like [Andrew Gelman](#) and [Daniel Lakens](#), but there are all sorts of people in this category. And both Gelman and Lakens are PhDs with institutional affiliations – “citizen science” doesn't mean random cavemen who don't understand the field – but they're both operating outside their day job, trying to contribute a few hours per project instead of a few years. I know many more people like them – smart, highly-qualified, but maybe not going to hire a team of paper-pushers and spend thousands of dollars in fees in order to say what they have to say. Even now these people are doing great work – but I can't help but feel like more is possible.

IRB overreach is a small part of the problem. But it's the part which sunk my bipolar study, a study I really cared about. I'm excited that there's finally more of a national conversation about this kind of thing, and hopeful that further changes will make scientific efforts easier and more rewarding for the next generation of doctors.

The Study of Anglophysics

I.

Dear Dr. McCord:

Seven years ago, our research staff read with interest your work on Berkeleyan idealism. We were particularly fascinated by your seemingly outrageous claim that it might be possible for individuals to imagine mental worlds so strongly that they would take on a reality of their own.

At the time, as our laboratory had an interest in novel solutions to the overpopulation problem, we embarked upon a test project to see whether a parallel world could be imaged and then colonized by citizens from our own dimension. Using advanced science you could not possibly comprehend, we came up with a practical implementation of your idea. Dr. Michael Adwell, whom I believe you met during your time in Oxford, volunteered to enter the device we had constructed as our first research subject. We very briefly imaged an alternate world based on the contents of Dr. Adwell's mind before the good doctor unfortunately had a grand mal seizure. He was disconnected from the device and rushed to the hospital, where he passed away several hours later.

Two years ago we revisited some of our calculations on the project and determined, to our surprise, that the world Dr. Adwell had created might still exist in some sense; that it had somehow managed to sustain itself separate from the doctor's mental activity. We worked feverishly to construct a device that might let us interact with his imaged world. Six months ago we succeeded. The computational demands of the machine were immense, but after throwing the remainder of our budget for the year at the Kyoto Supercomputing Laboratory, we were able to rent enough processing power to translate myself and Dr. Lachlan Fairchild into the imaged world, which we dubbed "Adwellia" after our late colleague. Our superiors informed us that when the next fiscal year rolled around in four months, there would be enough money in the budget to translate us back home.

II.

On first arrival, Adwellia seemed much like home. We landed on the shores of a small lake in what seemed to be a wooded area. Since it was getting dark, we soon set to pitching camp for the night. Our first unpleasant surprise was that the kerosene heater we had brought with us wouldn't work, leaving us cold and disheartened. Lachlan collected some logs to build a fire, but our matches didn't seem to work either. I remembered the seventh page of your paper, where you had posited that an imaged world would run on the same physics of our own world, since it would be bound by the expectations of the imager. Dr. Adwell had certainly understood enough chemistry to know that matches should start fires, but it seemed one of our most basic predictions had already failed.

I will not say whether we were more motivated by curiosity or by the bitter cold, but we tried dozens of different branches – small, large, young and green, old and rotting – and everything from dousing them in kerosene to the old-fashioned method of rubbing sticks together to create friction.

Finally, I succeeded in getting some branches from an old fir tree to alight. In relief, the two of us huddled close to the fire. But our curiosity was only heightened when we found the area near the fire to be unmistakeably *colder* than the surrounding air. Here our chill overcame our scientific spirit, and we decided to deal with the problem in the morning. We got into our too-thin thermal sleeping bags and passed a miserable and freezing night.

When we awoke, the fire had gone out, and in its place stood a pile of hats – twenty of them, to be precise. I would have called them fedoras, although Lachlan said the particular style was more popularly known as a Homburg. We debated taking the hats, but we had been thoroughly spooked. Instead we picked up our camp and journeyed south, where it looked like the wood was beginning to thin out.

Around midday we spotted smoke, and dared to hope we were coming upon a settlement. By evening our guess was confirmed, and we saw a village of conical adobe huts. We prepared to gesture our request to trade trinkets for lodging to the inhabitants – who were far too dark skinned to be European but who did not quite pattern-match to my memories of any particular human race. Imagine our surprise when we found they spoke English – though with abominable grammar. The headman introduced himself as Somon, and was all too happy to accept our trinkets in exchange for a nice warm hut to spend the night in.

We endeavored to learn more about these people in the morning, but by this time were tired enough to call it a night. We could not help inspecting the heating mechanism in our room, which seemed to be a mud bowl in which sheaves of wheat, small rocks, and little mud figurines that looked like people had been placed. Totally absent any visible mechanism, the setup was emitting heat – and what was more, a ball set in a track along the edge of the bowl moved continuously around in what seemed to all the world to be perpetual motion, making an annoying crackling sound as it passed over little leaves set in the rim. We had only a little time to exchange theories before falling into a deep sleep.

The next morning, the bowl was no longer warm, the ball had stopped moving, and the objects within had apparently transmogrified into a miniature wheelbarrow. This was strange magic.

The villagers were already up and about, so we found Somon and tried to get some better conversation in.

“We are scientists,” we told him “from far away, looking to gain a better understanding of how things work here.”

“Here in Magonaw?” asked Somon, using what we later found was the name of the village. “Not well.” He smiled, showing very pearly teeth.

“We were hoping to set up a laboratory – a few metal huts and a big machine – maybe on the outskirts of town. We would pay you for food, maybe for help with certain things. We have many tools to trade, and lots of gold and metal.” Not exactly true – what we had was a portable nanofactory, translated in with us as an easier alternative to bringing supplies. But we could get tools or transmute elements pretty quickly.

“Is of course,” said Somon, with the delight of someone who had stumbled entirely by accident into a beneficial arrangement. “What will you be needing?”

"Well the first thing," interrupted Lachlan, "is we wanted to know how your heating device works. The one with the wheat and the rocks. It was new to us."

"You not have this in your village?" said Somon, with a frown. "Is not obvious?"

"No," I said. "Where we come from, it's not obvious at all."

Somon brightened. "Your village," he declared "not know true names!" He picked up a rock from the ground. "True name of this is...rock."

We both nodded, mystified.

He grabbed a sheaf of wheat from a passing villager, who gave him a glare. "True name," he said, "is...wheat."

He said it with the same mystical intonation with which one of our colleagues back at the laboratory would announce a particularly earth-shattering result.

"Yes, okay," said Lachlan, kind of miffed. "I actually think we do know true names of things. It's the same in our language."

Now it was Somon's turn to be mystified. "Then...where is confusion?"

"The heating device," said Lachlan, narrowing his eyes. "How does it work?"

"Is obvious!" said Somon, like we were idiots. "Wheat and rock and art become work and heat and cart. The work push little ball around. Then ball make noise, continuing reaction."

"But..." I interjected, because it looked like Lachlan wanted to grab the headman and wring his neck "why do the wheat and rock and art become work and heat and cart."

"Is true names" said Somon, and shrugged.

"Holy shit," said Lachlan, at exactly the instant when I remained just as confused as I had been before. I stared at him.

"Holy shit," Lachlan repeated. "This world fucking runs on anagrams. English language anagrams."

Wittgenstein once said that the limits of our language are the limits of our world. Some say that mathematics is the language of God. Maybe that was why our world ran on math. Well, English had been the language of Dr. Adwell. It had been the lens through which he made sense of reality.

Maybe our hypothesis that his imaged world would run on the same physics of our own had been premature.

What if his world ran on English?

"The fire!" said Lachlan, who as usual was a step ahead of me. "Fir branches and heat. Fir plus heat becomes fire plus hat. So it removed heat from the atmosphere, and created fire and a hat."

"Twenty hats," I reminded him.

Lachlan was already deep in thought. "It's all stoichiometry," he started saying, almost faster than I could follow. "In our world, water is H₂O. H-O-H. Here, a fir tree has to be literally made of F-I-R. Twenty six letter-elements, forming a near-infinite amount of word-molecules. Suppose we burned three kilograms of fir branches...don't know the molar weight here, but suppose each letter weighs the same and there's one mole per kilogram, just bear with me. That's one mole each of F, I, and R. So it must have absorbed some sort of four mole equivalent amount of heat...whatever that means...and then spit out three moles of hats and four moles of fire. Three moles of hats in this system would be three kilograms of hats, that would mean each hat weighs 150 grams...it all checks out! Somon! Quick! Show us how you make something else!"

Somon looked at him. The headman seemed as confused as I was, but for different reasons.

"Make...what?" he asked.

"I don't know. Clothes, tools, anything."

"My daughter Genea live in here," he said, gesturing to a hut on the outskirts of town with some smoke coming out of it. "She is weaver."

The "weaver" actually seemed to be performing some sort of complicated chemical reaction. She was holding beets over a cauldron that was bubbling up into a primitive fume hood, then throwing them into what seemed like a vat of tar. Water was running out a hole in one side, and on the other, a roll of cloth was getting steadily longer.

This time I got it before Lachlan. "Chlorine," I said. "Chlorine plus beets plus tar becomes cloth plus brine plus tears."

"That's not right," said Lachlan. "You're missing an 'e'".

"No I'm not," I said. "It consumes twice as much tar as chlorine or beets, and produces twice as many tears as brine or cloth."

"I think," said Lachlan, "that we had better get our laboratory set up sooner rather than later."

III.

This we did, at record speed. Not wanting to frighten the villagers – or expose ourselves to prying eyes – we set ourselves a kilometer south of town, on a cape overlooking a great sea. On the headlands of the cape was a small hill from which you could see for miles, and there we completed the week-or-so's work of getting the nanofactory up and running. Its first job was to extrude us two aluminum Quonset huts, which became our homes away from home.

From our little encampment the ocean stretched on as far as we could see. I wondered if there were other continents on this world – figuring out its size really should have been one of our first priorities. But we were too fascinated by this world's weird linguistic elements and reactions – anglophysics, we dubbed them – to properly investigate anything else.

The first and most obvious question was why everything wasn't reacting all the time. How come every time someone touched a rock, the skin + rock didn't become corks +

ink? Just the air alone should have destroyed a wide variety of objects.

("Oh, come on," I told Lachlan. "The air doesn't count". Lachlan had then gone on to prove me wrong by getting the iron tools we had brought to rust, then proving the rust happened faster in moist air, and air that was full of dust particles. "AIR plus IRON plus DUST," he told me "equals RUST plus IONS plus ARID. Things aren't rusting in this world because of oxidation. As long as it can suck dust and moisture from the air, it's rusting by Crazy Anagram Logic." So the air definitely counted.)

The first thing we discovered was that nature abhorred non-words. AIR and DUST wouldn't react on their own to become RUST and IA, because IA wasn't a thing.

"What about AI?" asked Lachlan. "Why not rust plus an intelligent computer?"

At the time, my answer was "Shut up! The world might hear you!" I would later learn this was not nearly as funny as I thought.

But at the time, we made quick progress. Simple materials and short words seemed to be most stable, with complicated or abstract concepts rarely forming spontaneously – which, at least, answered our AI problem. And reactions usually wouldn't happen at all without sound, which seemed to play the same role in this world that heat did in our own. Lachlan had suspected this almost from the beginning – that the crackling leaves underneath the ball had provided the sound-energy to continue fueling the reaction that kept us warm that first night. But it wasn't until we heard the cacophony of a village festival that we knew we were on the right track.

"WHAT ARE YOU DOING??!" I had yelled at Somon, over the din of drums and cymbals and screaming villagers.

"MAKING BEER!!!" Somon answered.

It had turned out that the villagers used pee and bran to produce beer and pans, but that the reaction went unpleasantly slowly unless they shouted it along. The shouting was, of course, egged on by the beer they had already produced, which sort of made it an autocatalytic reaction if you squinted. They offered us some of their beer, but even though I knew things worked differently here my standards were a little too high to drink beer *literally* made of pee and so we returned to the lab. On our trip back, Lachlan pointed out that all of the villagers' iron tools had been carefully taken inside during the festival, so that the noise would not cause them to rust.

Our next big discovery was a week later. I woke up at 7 AM with Lachlan pounding on the door of my aluminum hut.

"OMAR!" he was shouting. "TAKE A LOOK AT THIS!"

Sitting on his palm was a one inch tall man, naked and hairless, looking terrified. He looked like he would have run off if there was anywhere to run to.

"What in the...?"

"I found a volcanic vent, up in the hills to the west. There was a source of methane. I broke it down into HEAT and MEN. But there wasn't enough MEN to form someone full sized. So I got this."

"Lachlan, you've got to help him!"

Lachlan gave a grunt, as if annoyed to be reminded of the ethical implications of his work. "How?"

"Can you speak language?" I asked the little man on Lachlan's palm.

In response, the man screamed. I took that as a no.

So I dragged Lachlan down to the village, where I woke up an annoyed Somon. "Somon," I said. "We found a way to break methane into..."

Somon's eyes went wide. Then he got angry. "No methane!" he said. "Is taboo! Will..."

He saw the homunculus in Lachlan's palm. With a deft motion belying his age, he yanked the little creature away from Lachlan and snapped its neck. I gasped. Lachlan looked annoyed.

"Is TABOO!" shouted Somon, with an anger I hadn't seen in him before. "These things! Not men! No speech! No mind! Must not make! Little man is taboo! Methane is taboo! If you make little man, no longer stay with us!"

I calmed him down, promised we wouldn't be doing any more experiments with methane, said we were new here, didn't know what we were doing. I asked him for more advice, asked him about any other taboos. He seemed irritated, assumed we should know what they were, seemed to think less of us with each question indicating our ignorance. Finally we gave up and made the long trek back to our laboratory.

Our next few weeks of experiments were less bloody, but still exciting. Suppose we took a mop and the guts of an animal, and shouted at them until MOP + GUT reacted to POT + become GUM. Would the pot be the cooking implement, or would it be marijuana? For that matter, why shouldn't it be a top, the child's toy? Why shouldn't the gum form a mug, fit to drink coffee from?

In our first experiment, we surrounded our apparatus with pans and food, and were unsurprised to find we ended up with cooking implements. We repeated the experiment, but this time surrounding the apparatus with bongs, tobacco, and other drug paraphernalia - this time we got marijuana. We wanted to get a playful child to see if we could produce tops, but news of our work with methane had gotten out and spooked the villagers, and they were understandably unwilling to let us borrow one of their children.

The third experiment was in my opinion the key to this entire process. This time we surrounded the apparatus with pans and food, but both Lachlan and I concentrated very very hard on marijuana, and talked about marijuana with each other while the loudspeaker the nanofactory had extruded blasted sound at the reactants, and sure enough, we got marijuana.

Somehow our expectations were guiding the physics in a way that the letters themselves couldn't. I started to wonder what had become of poor Dr. Adwell. Was the god of this world a deist, who had created it shortly before dying in a hospital ICU in a very different planet? Or was he in some sense still here, still actively guiding things?

The reaction that rusted iron started to seem more and more suspicious. What about that ARID? In our experiments, making adjectives had been almost impossible, requiring more sound catalysis than any noun we had encountered so far. But ARID seemed to form of its own accord. What if Adwell somehow remembered that iron was

supposed to rust, and *privileged* that reaction as the sort of thing that ought to go on? What if the reason everything didn't implode upon itself was Adwell ensuring that everything in his imaged world happened according to some plan?

Then our proof that we could alter our results through concentration and careful priming would take on a whole new meaning.

Did *reminding God what chemical reaction we wanted* change experimental results?

IV.

"We're going about this half-assedly," Lachlan told me one morning our sixth week in Adwellia. "All of this looking for clever anagrams is taking up too much of our time, delaying us in supremely great work. We need to do this analytically. Get a bottle of As, a bottle of Bs, so we can create whatever the hell we want."

This proved easier said than done. We got the nanofactory to extrude us a very complex apparatus, a centrifuge, and what we took to calling the "sonic ray" – a machine that made deafening noise along a very narrow arc and which could catalyze reactions much faster than shouting or drumming. It turned out to be the key to making far more complex products than we had previously attempted. But our first use was a plain and simple failure.

We had decided to start with granite, which we would break down into tin, rags, and the letter E. We would then centrifuge the decay products, with the three-letter tin and rags going one way and the pure E going another.

Nature, remember, abhors non-words. No sooner had we forced some E into a test tube than the tube itself transformed in a great explosion to gelatin and a tiny, near-microscopic donkey. E + GLASS = GEL and ASS. We couldn't say we couldn't have seen it coming. It could have been worse – I was just glad that Dr. Adwell's ascended mind's first association with the latter word was "donkey".

We tried the experiment again with a zinc vial – zinc because it was implausible that there was an ZINC + E anagram lurking out there – and ended up with a mat of eels. Through this whole time, we had been debating the problem of ambiguity – who was to say that our granite was GRANITE rather than ROCK or even STONE – and the answer seemed to be that Dr. Adwell – or whoever was watching Upstairs – was mostly sympathetic to our efforts. Well, the sympathy ended when we started trying to isolate single letters. ZINC became METAL and thence EEL MATs.

Our effort with mud was even worse. We put a lot of time into making sure the mud we got was very classically mud – not ooze, not muck, certainly not dirt. And there was no good way MUD + E was becoming anything. We turned on the device.

The Es disappeared. Seriously. Granite went into the centrifuge, tin came out, but there was no sign of an E anywhere, and rather fewer rags than usual.

"This is really weird," I said.

"Thanks, Einstein!" said Lachlan. "I never would have figured that out without YOUR FUCKING COMMENTARY."

I should have told him to calm down, but the experiment had upset me too. "Well it wasn't MY BRIGHT IDEA to try to ISOLATE ALL THE LETTERS," I said. "WHICH REMINDS

ME! IF YOU THINK I'M GOING THROUGH THIS TWENTY FIVE MORE TIMES, YOU CAN GO FUCK YOURSELF!"

Lachlan swung at me, missing by an inch. I kicked him, right in the knee, and he fell into the experimental apparatus, knocking the whole thing over. Both of us went down with it. For a second, the sonic death ray shot straight at us – EEEEEEEEEEEEEEEEEIE! and then its safety kicked in and it turned off. We sat there, stunned, bruised, in pain.

"Rage," said Lachlan. "GRANITE becomes TIN plus RAGE. Holy fuck, we created an emotion."

It had happened before, sort of. The wheat and rock and art, they had come together to produce work, which was an abstract concept. But it was still in the domain of physics. "Work" seemed like the sort of thing that could come out of chemical reactions, kind of like heat. But rage? This was something really new.

That night, we made the short trek into the village and asked Somon what he thought.

"Rarely," he said. "Sometimes, when festival is very loud, strange things happen. Should avoid. Very bad. This is taboo."

The next week, I knew something was up. Lachlan was missing our daily debriefings, not getting any work done. Finally I broke the most important unwritten rule of our little community. I went into his aluminum hut without knocking.

There he was, sitting with a blissed out look on his face. Beside his bed sat a miniature version of our experimental apparatus, complete with its own sonic death ray – he must have privately ordered it from the nanofactory, then deleted the records. It was reacting little tchotchkes from the village – dolls, balls, play swords – with our glass specimen jars. Tar was streaming into the waste bin.

I turned off the sonic ray. Lachlan awoke with a start. He seemed about as angry as he'd been the time we accidentally produced rage from granite, but this time I knew he had a less noble reason.

"What the fuck are you doing, barging in here like this?"

"You've gotten yourself addicted," I said. "Addicted to joy."

Lachlan didn't deny it, as his TOY + JAR -> JOY + TAR reactor was right there.

"Look," he said. "It's been two months now, stuck in this stupid world. It's going to be another two before the lab brings us back home. The villagers are crazy, physics runs on English, and the nanofactory can't produce any entertainment that's remotely entertaining. The letter isolation project is a failure, you no offense are one of the most boring people I've ever met, and when I try to get some of the village women to look at me they murmur something about taboos and give me the cold shoulder. Give me a break here, Omar!"

"Lach," I said. "You're neglecting your work. We still haven't gotten anywhere near the bottom of anglophysics, let alone figured out the most basic stuff about this world like how big it is. You sitting here blissing out on raw linguistic joy isn't something we can afford right now."

"Fuck you," said Lachlan, but he didn't protest as I picked up his mini-apparatus and brought it to the nanofactory's disassembler chute, nor as I reprogrammed the nanofactory to make sure all its records would be public from now on.

V.

A week after that incident I finally got the nanofactory, with great creaking and protesting, to extrude a small aircraft so I could explore the surrounding area. The villagers were delighted, having never seen anything similar, and several of them demanded rides – increasing our popularity a little after the methane debacle. When we were done appeasing the natives, I took off and started mapping Adwellia.

We seemed to be at the southernmost extent of an island about three hundred miles east to west and twice that north to south. The island was mostly forested except for the broken volcanic area nearby where we had gotten the methane and some hills further north. Four hundred miles east of us there seemed to be another continent or large island, but that was about the limit of my range and so I told myself I would explore the new land another day.

The distances allowed me to do some geometry and calculate the size of the world. Adwellia appeared to be a spherical planet about the size of the Earth. As far as I could tell it had one sun and one moon, and there were normal stars in the sky. It seemed to get colder further north and warmer further south, though I wasn't able to fly far enough to confirm it had proper poles and an equator.

By the time I finished these explorations, about a week after they began, Lachlan had developed a new obsession.

"I can't solve the letter isolation problem," he admitted. "But someone else can. Someone like Einstein."

"Great," I said, sarcastically. "All we need is..."

Then it hit me. Surely he wasn't that crazy.

"Yes," he said. "Why not synthesize Einstein? Or some other brilliant scientist who's more creative than we are. I've been going through the dictionary looking for proper combinations. It's not that hard."

This proved optimistic, but the equation upon which we eventually settled was $\text{STONE} + \text{TIN} + \text{FORT} = \text{EINSTEIN} + \text{FIRE}$. The only difficulty was obtaining the fort, since the villagers here did not seem to be of a militaristic bent, but I had found some ruins further north during my explorations, and one of them did indeed seem to be an old stone fort, perhaps constructed by the villagers' ancestors. I proposed we get a party of villagers to help quarry fort material, but Lachlan objected that they would probably just have some stupid taboo about it, so instead I landed there with the aircraft and laboriously ferried fort parts home in twenty pound increments, on my lap.

Once we had enough fort to stoichiometrically produce Einstein, getting the stone and tin was easy. But getting the reaction to work proved impossible. No matter how many physics books we stuck around our apparatus, no matter how hard we concentrated on the great scientist, the reaction spat out absurd things like ferns, nits, and a tooting sound – or forests, nits, and one ton weights, or a nose with a tit in the front, which trust me was *really* awkward and which we threw into the nanofactory disassembler chute as soon as we could, believe you me.

After about thirty tries, Lachlan announced that the problem was obvious. You see, we needed a *capital* E.

I grudgingly admit that, even after two months in a world where stone was composed of S, T, O, N, and E, the thought that there were different atomic units representing lowercase and capital Es seemed absurd. But as always, my sense of impossibility surrendered to crazy reality and I figured that Lachlan was probably right. We needed a capital E.

Two days later, Lachlan showed up at the laboratory with a very suggestive looking sack.

"Lachlan, what were you just out doing?" I said, hoping the answer was anything other than what I knew it was going to be.

"Just grave robbin'" he answered. "I got us the corpse of a lady named Eder, who died of pneumonia yesterday. Don't worry, no one saw me take it."

"Oh, come on," I said. "When they find the grave disturbed, who are they going to suspect? The other villagers, who they have known their whole lives? Or the mysterious strangers on the storm-wracked cape outside of town who have already violated their sacred taboos. Lachlan, *you* are a fucking idiot."

"Maybe I am," said Lachlan. "But if I'm so stupid, good thing we'll have Albert fuckin' Einstein around to help provide some brains for this operation."

The new equation was $EDER + TIN + SNAIL = EINSTEIN + LARD$.

So God help us, we hired some villagers to collect snails for us, and when we had hundreds, we poured poor Eder's bones into the reaction chamber along with the snails and some tin and started the sound.

And Einstein started to grow. At first he was tiny, smaller than the methane-men in Lachlan's palm had been, no bigger than the snails that surrounded him. But as bones and metal and snails slammed into him, he grew bigger, all the while screaming and covering his ears as the sonic ray did its gruesome work. We saw him, child-sized, beating up against the glass wall of the reaction chamber, ever growing, ever screaming.

"You're mad," I told Lachlan. "We've got to stop this."

"Maybe I am," said Lachlan. "But think! Einstein! The greatest scientist in recorded history! Think what we could do! Revolutionize not only our study of Adwellia. But we could bring him back with us, get the lab to translate him as well as us. We could turn Adwellia into a genius factory that would revolutionize civilization back on Earth. Omar, this *has* to be done! The potential in anglophysics makes a Nobel Prize look like a tee-ball trophy."

When Einstein was fully formed, and released from the reaction chamber, he attacked us. We subdued him, using weapons extruded from the nanofactory, and kept him in a cell. For three days we tried to talk to him, and he responded by screaming wordlessly at us and spitting in our faces.

I don't know whether there was something theological going on – whether Einstein was just a homunculus lacking a true soul. Or whether it was just very simply that our

Einstein was psychologically an infant, that no one had taught him so much as language let alone physics, and that Adwell or whoever was up there wasn't going to assume we meant "the smart Einstein, who knows lots of stuff" in the way we wanted.

Our Einstein was a giant infant, not even an infant, a fetus that should never have been born. On the third day, by mutual consent, we stuck him in the nanofactory disassembly chute and resolved never to speak of him again.

VI.

That was the last time I worked together with Lachlan on anything of note. After that we retreated to our separate aluminum huts, acknowledging each other only when our paths crossed on the way to the nanofactory for some crucial part.

I found him creepy. He *was* creepy. And he thought I was holding back our research. Maybe that was true too. In either case, it was a terse nod, a couple of words, and the tacit acknowledgment that it wasn't worth resolving our hostility in the month or so we had left before we were transferred back.

I spent that last month trying to build on my theory that Adwell's mind was somehow working behind the scenes running everything. The catalytic property of the sound, I theorized, was its ability to *get Adwell's attention*. It was a sort of "HEY, GOD, LOOK OVER HERE, WE'RE DOING SCIENCE, BETTER APPLY THE LAWS OF PHYSICS RIGHT AWAY". I know it sounded bizarre, but my early experiments bore me out. Rapidly flashing bright lights seemed to speed reactions almost as well as sound. So did – because sometimes the simplest solution is the best – shouting "ADWELL! LOOK OVER HERE!"

With these advances, once again entirely new classes of reaction became possible. No longer were we limited to the highly reactive simple materials with short names. Long strings of words, complex abstractions, even adjectives came within our reach. It was exciting.

But once again, it was Lachlan who was really pushing the frontiers. One night he started banging on my door: "OMAR!" he shouted. "I DID IT!" When I went out he practically dragged me into his hut, which was nearly piled, floor to ceiling, with papers that turned out, on inspection, to be various IQ tests the nanofactory must have been carrying in its databanks.

"What did you do?" I asked.

"I couldn't create Einstein," he said, referring to the still-fresh debacle – "so I decided to turn myself into Einstein! Look! I'm producing SMART. And it's working!"

His sonic ray – now only a fraction of the power of my own multimodality parasonic device – was reacting smoke and carts into coke and, apparently, smart. A complicated system of tubes and centrifuges was catching the smart and binding it into a containment chamber linked to a helmet. Clearly someone was supposed to put it on.

"And you're saying it works?" I asked.

"The IQ tests don't lie," said Lachlan. "I was 152 two weeks ago. Now I'm consistently getting in the 160s."

Judging by the number of tests, he must have been obsessively checking his numbers every hour or so.

“Now,” he said, “I’m going to try that letter isolation thing again.”

I judged by the shouts of rage and frustration I heard over the next few days that it wasn’t working.

Two days later, Lachlan asked me if he could borrow my advanced parasonic ray. I refused. That evening, it went missing for about three hours before turning up on top of my desk. I noticed Lachlan now had one exactly like it.

I soldiered on. In between my experiments, I played a little game predicting what Lachlan was trying to synthesize by the objects he took from the nanofactory and the supplies he ordered brought in from the village. One day it was buckets of dew, carts full of animal legs, and an entire cage of live minks – my best guess was he was trying to get KNOWLEDGE, but I couldn’t get the stoichiometry to line up. Judging from his screams of frustration that night, neither could he.

The next week, it was load after load of potatoes, fence posts, and a tank of minnows. It took me half an hour to come up with OMNIPOTENCE, even though once I made myself start thinking like Lachlan it was obvious.

I started to become worried.

One day, three months and two weeks into our mission and only fourteen short days before we hoped the laboratory would re-establish contact, I went out for a sortie with the plane and came back to find a disaster area.

Our huts had been smashed open. The nanofactory had big dents in its aluminum casing. Inside, all my lab equipment had been broken, my papers thrown on the floor haphazardly.

I went into Lachlan’s hut. IQ tests everywhere. He was missing. So was his parasonic ray. I figured they had grabbed my partner in his sleep, before he’d had time to resist. In retrospect we really should have put up some defenses, but we hadn’t expected to need them.

The nanofactory was still online. It was pretty hard to break – especially if, as I suspected, the vandals were villagers armed with clubs and rocks. I told it to extrude me some overwhelmingly powerful weaponry. After making me wait an hour, it gave me a ring that upon threat would instantaneously unfold into a device that generated an invincible barrier around the wearer, plus a hand-held matter disruptor. Thus armed, I walked into the village and found Somon.

I didn’t have to bring up the subject of Lachlan. “Is evil man!” the headman told me, as soon as he saw me. “Broke taboos! Created life! Dug up grave! And today! Today was worst! Kidnapped my daughter, Genea! No more okay! Tonight gets beaten! Tomorrow dies!”

Raising my invincibility shield, I wandered into the public square. There, whipped bloody and tied to a post, was Lachlan.

“You kidnapped the headman’s daughter?” I asked him. I didn’t even give him the dignity of pretending to doubt whether it was true.

Lachlan smiled. "Genea. A perfect name for my reaction. I could have been a Genius, with a capital G."

I don't know if it was that smile, or the blood all over him, or the lack of remorse in his voice, but at that moment, I'd *had it* with Dr. Lachlan Fairchild. I lowered the matter disruptor.

"You know," I said. "That is it. I'm not even going to rescue you. You're a menace."

"You don't have a choice," said Lachlan. "I have a nuke. These people don't understand the concept, but lucky we've got a genius like yourself. Let me go or I blow this entire planet sky high."

"Even if you managed to extrude a nuke," I said "which you didn't, because I checked the nanofactory's public records before I left - even then, nukes don't work in this world. Nuclear fission isn't an anagram of anything."

"A metaphorical nuke," said Lachlan. "I mean, I've figured out this world's equivalent of a nuke. It's very clever. Without the SMART, I never would have been able to think of it. I'll..."

My best course was to immediately, like split-second immediately, raise the matter disruptor and shoot Lachlan. I could do it before he had a chance to react, and it would solve the whole damn problem.

Instead I took the worst course, which was to raise the matter disruptor, obviously intending to shoot him, and vacillate at the last moment because I'd never killed anyone before and I wasn't sure I had it in me and instead of finding out my brain wanted to sit and ponder this for thirty seconds.

Lachlan took a ring off his finger and it unfolded it to reveal his parasonic ray. Then he furrowed his brow in concentration and it let out a screech.

I shot the matter disruptor. Man, post, and town square changed into their component atoms...letters...whatever.

The villagers ran, screaming. Some of them ran away from the explosion. Others ran towards the explosion, trying to see what had happened and maybe defend their homes and families. A few arrows and stones came towards me, causing my ring to near-instantaneously unfold into a weird backpack-like device that placed itself on my back and surrounded me with a purple glow. The projectiles hit my new invincibility shield and fell to the ground limply.

I calmly walked through the carnage. I was heading back a kilometer south, back to the cape. I was going to extrude a larger aircraft, bring the nanofactory a few hundred miles away, and wait out the last two weeks of exile far away from this mob.

The ground started to shake. I realized the explosion had ended long ago, yet its deafening roar had not subsided.

I looked back to the town square and my blood turned cold. In the center of the blast radius, where not even dust should have remained, there was Lachlan's skull, set in the biggest rictus grin I had ever seen.

I raised the matter disruptor and fired another shot. The skull disintegrated. But Cheshire Cat-like, somehow the grin remained, even larger than before, a smile without a substrate.

This was bad.

I started to run back to the lab. Cracks opened in the ground around me. The roar become worse. Was it just me, or was the sea getting closer?

Metaphorical nukes. A nuke was at the most basic level a chain reaction. Neutron produces energy plus neutron. That neutron produces energy plus neutron. *That* neutron and so on. You end up with a *lot* of energy.

I could see the remains of the looted lab now in front of me. It was on its elevated headland reaching into the sea, and I was afraid the rising water was going to cut it off and turn it into an island before I could get to it.

Sound drove chemical reactions in this world. Anything that could create sound had the potential to be a chain reaction if the reactants were common enough. You could get most of the letters of “sound” from...oh, that wasn't good.

The cracks in the GROUND got bigger as the low-lying GROUND started to sink further beneath the waves.

I stared back at the village. It was almost entirely underwater now. Above it was Lachlan's disembodied grin, now the size of a skyscraper, hanging in the sky.

Sound, ground. Grin. Sin. There. I had it. $\text{GROUND} + \text{SIN} = \text{SOUND} + \text{GRIN}$. The nuke. The ground was essentially limitless until the world was destroyed. The more ground was destroyed, the more people died, the more villages sunk under the waves. A sin. A reaction that created its own reactants. And sound. Created its own reactants and its own catalyst. Leaving nothing but Lachlan's gigantic triumphant grin, hanging in the sky over the world he was destroying.

I groaned as a crack in the ground took the aircraft on its field. It teetered for a second, then fell into the onrushing waves. I ran through ankle deep water and at last reached the top of the headland. There was just a small area of land left, on the highest ground of the cape, with our two little partially-smashed huts and the bulky dented aluminum nanofactory.

“Extrude boat!” I commanded the nanofactory.

“Extruding boat,” said the display. “Estimated creation time with material on hand, two hours.”

“Cancel! Cancel cancel cancel!” I shouted, but the factory had gotten into its extrusion mode and wasn't listening.

I ran into my hut. Most of my stuff was still broken. There was nothing that looked like a good flotation device, unless you counted my mattress. My reaction apparatus, my parasonic ray, and a few doodads.

I grabbed the ray gun and ran outside. Even on the high ground, there were wavelets lapping at my shoes. I had about a minute before I drowned.

"Okay," I said to myself. "Time to figure something out. Time to create a boat." And there was only one good reactant on hand.

OCEAN + ...no, that wouldn't work. SEA + ...that was even worse. WATER + ... I might be able to use water if I let the reaction consume my bones...WATER + BONE = BOAT + NEWER ... no, even with the parasonic ray I'd never be able to catalyze a reaction that made a comparative adjective of all things. Maybe if I had an hour to think of some useful intermediates.

Okay, back up. You don't need a boat. You can use a ship. Ship is...

My brain was in panic mode. It didn't want to anagram SHIP. What it wanted was escape.

The cape! The cape could provide escape! The cape and the sea! The two things I had! And my parasonic gun was just strong enough to let me synthesize abstractions. I just needed somewhere to put that extra A.

WATER + A = AWARE + T. No, Nature abhors non-words, T won't work. WATER + A = RAW TEA. No, adjectives took forever. WAR TEA? I wasn't sure what would happen if I caused a war at this point, but I bet it wouldn't be good.

A wave rushed over me and I rose to the top sputtering and gasping. I still had the parasonic ray. The water had almost covered the huts now. Borne on the receding wave came Lachlan's stupid piles of IQ tests, now soaked.

CAPE + TEST = ESCAPE + 2*T

On the one hand, Nature abhorred non-words. On the other hand, I couldn't swim and was about to drown. I concentrated REALLY hard on the reaction, turned the parasonic ray to its highest setting, and shot a beam of sound and strobe light and repetition of the name "Adwell!" at the pile of tests and the rocky cape below.

Nothing happened.

The LOW CHARGE light began to flash on my parasonic ray.

It had been a stupid, desperate gambit. I'd already known I didn't have enough energy to do a reaction that created non-words, didn't know if that was even possible with *any* energy, and I had just drained my parasonic ray of almost all its charge I had made a terrible error.

"Error!" I shouted. "That's it! Adwell! Error!"

CAPE + TEST + 2*ERROR = ESCAPE + 2*TERROR

As I fell under the waves, with my last breath and last bit of charge I fired off the parasonic ray one last time.

It's not working I thought to myself. *It's not working and I'm going to die, lost under the sea, dead forever.* I spent half a minute just thrashing about in terror before I realized *that meant it was working.*

The water was receding! A bubble of air was spreading away from me in all directions as the water was consumed! I was saved! Still terrified, but saved!

...then the water started closing in on me again. I didn't know what what was happening. I'd done it, hadn't I? Succeeded in creating a reaction that would get me out?

Success! That was the problem! If I had succeeded in creating a reaction, then firing the parasonic ray hadn't been an error. The reaction couldn't take place. The water closed in on me again. I was going to die.

The water started to recede. If the success of the reaction prevented me from having made an error, then the reaction wouldn't work, and starting the reaction was an error, and so the reaction could take place. All this I saw clearly, as in a dream, from within my bubble of air.

The air bubble under the rising seas (sinking ground?) reached a size of about twenty meters, large enough to cover the cape and the two huts and the nanofactory, and then stopped, occasionally shrinking a little or growing a little, always seething, starting to burn with a weird energy.

From within the anglophysical terror clouding my mind, I recognized the problem as a novel version of the Epimenides paradox of self-reference, implemented on a physical substrate. If my initiation of the anglophysical reaction had been an ERROR, then I would ESCAPE, and it hadn't been an ERROR after all. But if my initiation of the reaction had not been an ERROR, then I would not ESCAPE, and in fact it *would* have been an ERROR.

I had a vague memory that I had once discussed Russell's Paradox with Dr. Adwell. I wished I could have remembered what he said.

The interface between air and water became turbulent, started to glow. I saw fantastic images projected upon it, weird fractal geometries, strange supersensory stimuli that somehow reminded me of Lovecraft's references to the beckoning piping from the void behind space. All the while the TERROR grew, and the bubble began to vacillate wildly.

Then there was a great pop, and I thought for a second my air bubble had popped, but more correctly *everything* had popped, and for a second the things that were nothing like piping sounds became unbearable. Then I found myself lying, still terrified, on the floor of the translation chamber of our laboratory, the very same place where I had entered Adwellia almost four months before.

VII.

When I had recovered my senses and debriefed my colleagues, I devised three theories for what had happened there, on the cape.

First, that my reaction had been successful beyond my wildest dreams, the paradox had resolved in my favor, and I had ESCAPED not only to firm ground but to my own home dimension.

Second, that the paradox had been so confusing and unbearable for poor Adwell that he had expelled me from his consciousness, like a man brushing a bug off his skin, and having been kicked from his world I naturally defaulted to my own.

And third, that implementing a paradox on a physical substrate was *really, really bad* and I had destroyed Adwellia.

This last possibility ought in theory to be testable, but I was informed upon my return that the budget was tight this year and that the necessary supercomputing resources to search for Adwellia will not be available for some time.

I have been assigned to another project, and although my superiors have thanked me for my work in Adwellia, I am certain they do not believe a word of my report and have written the entire expedition – and perhaps their decision in hiring me – off as a loss. In their place I would not do otherwise.

But from your writings I gather you are a man of unusual intellect, and some of your speculations come uncomfortably close to the truth. I do not know whether you have pursued your interest in Berkeleyan idealism further, but if you are so gracious as to believe my story or at least keep an open mind, I would be interested in further correspondence with you about the implications of anglophysics for future imaged worlds and how the consistency of such images might be assured against paradoxes of self-reference and other threats to their integrity.

Yours sincerely,
Dr. Omar Reyes, University of _____

PS: I hope you will be understanding when I say that I wish to restrict my future work in the imaged world field to a purely theoretical level.

[EDIT: I apologize to those who have read [Universal Fire](#) for this story. As a peace offering, please accept this lovely lampshade.]

[EDIT 2: HPMORPodcast has recorded [an audio version.](#)]