Slate Star Codex, 2020

# Why Doctors Think They’re The Best

Ninety percent of drivers think they’re above-average drivers, ninety percent of professors think they’re above-average professors etc. The relevant studies are paywalled, so I don’t know if I should trust them. Our recent discussion of therapy books would make more sense if ninety percent of therapists believed they were above-average therapists. I don’t know about that one either.

But I am pretty sure ninety percent of doctors believe they’re above-average doctors. Here are some traps I’ve noticed myself falling into that might help explain why:

1. Your patients’ last doctor was worse than you. Think about it; if somebody has a good doctor, they’ll stay with them, and you will never see that patient. If somebody has a bad doctor, they’ll go see another doctor instead. That other doctor might be you. So your current patients’ last doctor will be worse than average. But this is where most of your chance to compare yourself with other doctors comes from: “my patient’s last doctor misdiagnosed them, but I got it right” or “my patient hated their last doctor but says I’m much better”. See also You Are Not Hiring The Top 1%.

2. Your patients love you. Similar to the above: if your patients love you, they will stick with you. If your patients hate you, they’ll leave. So over time, your practice will consist of patients who are very happy with your style of practicing medicine. Suppose some patients like very practical/efficient doctors, and others like very touchy-feely doctors. A practical/efficient doctor will collect a set of patients who prefer practical/efficient doctors and who praise them for how focused and sensible they are. A touchy-feely doctor will collect a set of patients who prefer touchy-feely doctors and praise them for how caring and personable they are. It will be obvious to the practical/efficient doctor that they’re satisfying their patients’ preferences much better than the touchy-feely doctor down the road would, and vice versa.

3. Patients often come to you, but never leave you. When a patient transfers to me, I have to review the case, confirm that this is an appropriate case for me, talk to the patient about why they left their last doctor and what we can do differently, and then keep seeing that patient. When a patient leaves me, I usually have no idea. It usually looks like the patient saying “I’ll make an appointment for next month”, and then not making that appointment. If I’m not careful, I never notice this; it’s a dog that doesn’t bark. Even if I do notice, patients stop coming all the time for all sorts of reasons. Sometimes they forget. Sometimes they feel better and decide they don’t need you. Sometimes they move out of state and don’t tell you. Sometimes they lose insurance and can’t afford you, or change insurances to one that doesn’t cover you. If you’re really on top of things you try to call these people and ask what’s going on, but sometimes they don’t answer and sometimes they lie (it’s really awkward to tell a doctor that you’ve fired them). So it naively feels to me like patients switch from the doctor down the road to me all the time, but nobody ever switches from me to the doctor down the road. I’m sure it’s not true, but that’s how it feels.

4. You’ve probably successfully treated most of your patients. Now pull all of the above together. Suppose a patient has a chronic disease like depression or diabetes. If you treat it successfully, they will love you and stay with you; if you fail, they will switch to another doctor (and you will never hear about it). Ten years later, you wake up and notice that most of your patients are success stories. But your patients usually describe their previous doctor as a miserable failure. Selection bias is a heck of a drug.

5. You know what you know, but you don’t know what you don’t know. Suppose each doctor makes errors at the same rate, but about different things. I will often catch other doctors’ errors. But by definition I don’t notice my own errors; if I did, I would stop making them! By “errors” I don’t mean stupid mistakes like writing the wrong date on a prescription, I mean fundamentally misunderstanding how to use a certain treatment or address a certain disease. Every doctor has studied some topics in more or less depth than others. When I’ve studied a topic in depth, it’s obvious to me where the average doctor is doing things slightly sub-optimally out of ignorance. But the topics I haven’t studied in depth, I assume I’m doing everything basically okay. If you go through your life constantly noticing places where other doctors are wrong, it’s easy to think you’re better than them.

6. Your victories belong to you, your failures belong to Nature. Sometimes I get a really difficult case, something nobody else has been able to figure out – and I absolutely nail it. I ride the high for days. I feel like a miracle-worker. Other times I get a difficult case nobody else has been able to figure out, and I can’t figure it out either. I don’t worry too much about it – some things are beyond the ken of modern medicine; obviously nobody can treat untreatable stuff. This is especially true in psychiatry. There are so many things we don’t understand, so many cases that are just inherently unresponsive to any kind of treatment we know about – that it’s hard to beat yourself up over any given failure. When you fail, you just say “I guess that was one of those treatment-resistant cases”. Maybe if I was a slightly better doctor I would have been able to figure out a way to treat it – but I will never know, and that’s not a natural way to think.

7. You do a good job satisfying your own values. Everybody wants to make people healthy and save lives, but there are other values that differ between practitioners. How much do you care about pain control? How much do you worry about addiction and misuse? How hard do you try to avoid polypharmacy? How do you balance patient autonomy with making sure they get the right treatment? How do you balance harms and benefits of a treatment that helps the patient’s annoying symptom today but raises heart attack risk 2% in twenty years? All of these trade off against each other: someone who tries too hard to minimize use of addictive drugs may have a harder time controlling their patients’ pain. Someone who cares a lot about patient autonomy might have a harder time keeping their medication load reasonable. If you make the set of tradeoffs that feel right to you, your patients will do better on the metrics you care about than other doctors’ patients (they’ll do better on the metrics the other doctors care about, but worse on yours). Your patients doing better on the metrics you care about feels a lot like you being a better doctor.

Some of these seem to generalize; do other jobs where customers select someone to work with have the same problems?

# Hardball Questions For The Next Debate (2020)

[Previously: Hardball Questions (2016), More Hardball Questions (2016). I stole parts of the Buttigieg question from Twitter, but don’t remember enough details to give credit, sorry]

Mr. Biden: Your son Hunter Biden was on the board of directors of Burisma, a Ukrainian energy company, during your vice-presidential term. The Ukrainian government was investigating Burisma for misdeeds, and Hunter was allegedly one of the targets of the investigation. President Trump alleges that you used your clout as VP to shut down the investigation into Hunter, which if true would constitute an impeachable abuse of power.

My question for you is: if your son had been a daughter, would you have named her Gatherer?

Mr. Bloomberg: You’ve been criticized as puritanical and self-righteous for some of your more restrictive policies, like a ban on large sodas. You seem to lean into the accusation, stating in a 2014 interview that:

I am telling you, if there is a God, when I get to heaven I’m not stopping to be interviewed. I am heading straight in. I have earned my place in heaven. It’s not even close.

Let’s not focus on what this says about your humility, or about your religious beliefs. I want to focus on a different issue.

Despite spending $100 million in the first month of your presidential campaign, you are currently placed fifth – behind two socialists, a confused old man, and the mayor of South Bend, Indiana. In, let’s not forget, an increasingly shaky effort to prevent President Donald J. Trump from winning a second term.

So my question for you is: what makes you so sure you’re not in Hell already?

Mayor Buttigieg: You are a gay Navy veteran. Your last name is “Buttigieg”. You are mayor of “South Bend”. And you first achieved prominence on the national stage for a New York Times editorial about your travels in the Horn of Africa, which includes the country of “Djibouti”.

My question is: is your campaign just the setup for a gay porno? Do you really think viewers want this much backstory?

Senator Warren: Despite your many years of service to the nation, media attention has focused on your claim to be descended from Native Americans. You told your former employer Harvard that you were of Native descent. Republicans accused you of trying to unfairly exploit affirmative action, but an investigation showed you did not benefit from any affirmative action at the time, leaving it unclear why you would do this.

More recently, you took a genetic test to establish your Native background. The test showed you did have a Native ancestor 6-12 generations back, but supporters were left baffled as to why you would take it or expect anyone to care. Conservatives used to the test to reignite the scandal around your Harvard employment, and progressives condemned you for promoting a view of race based on biology rather than culture or self-identification. The general consensus, again, was that you got no benefit from the test and it was unclear why you would do this.

The development of one of the algorithms that uses genetic information to determine racial background was called the “Warren Project” after its lead geneticist Jim Warren. Warren founded FamilyTreeDNA, a direct-to-consumer genetic testing company that continues to be a leader in genetic testing for ancestry, with about $16 million in revenue each year. This is relevant because Jim Warren is your ex-husband and the father of your children, who presumably stand to inherit a significant part of the FamilyTreeDNA fortune.

So my question for you is: is your campaign is just a publicity stunt to raise interest in genetic testing?

Senator Sanders: You are most famous for the 2016 incident where a bird landed on your podium mid-rally. Supporters have reasonably connected this to the ancient Roman practice of augury, where leaders were chosen by the number of bird-related omens surrounding them.

But auguries can be hard to interpret. For example, during the founding of Rome, Romulus and Remus agreed to use augury to determine which of them should lead the new city. The two of them went out and watched for ominous birds. First, Remus saw six vultures, which he interpreted as strong evidence that he should lead. But then Romulus saw twelve vultures. The two argued, with Remus’ claim resting on having seen vultures first, and Romulus’ claim resting on the theory that more vultures = more leadership. One thing led to another, Romulus killed Remus, and Rome ended up building the greatest empire in history. This firmly established the principle that even if one person sees birds first, another person who sees more birds may still be the rightful leader, if he sees enough of them.

So my question for you is: it’s been four years. How many birds would have to land on Donald Trump before you admit he would make a better president than you?

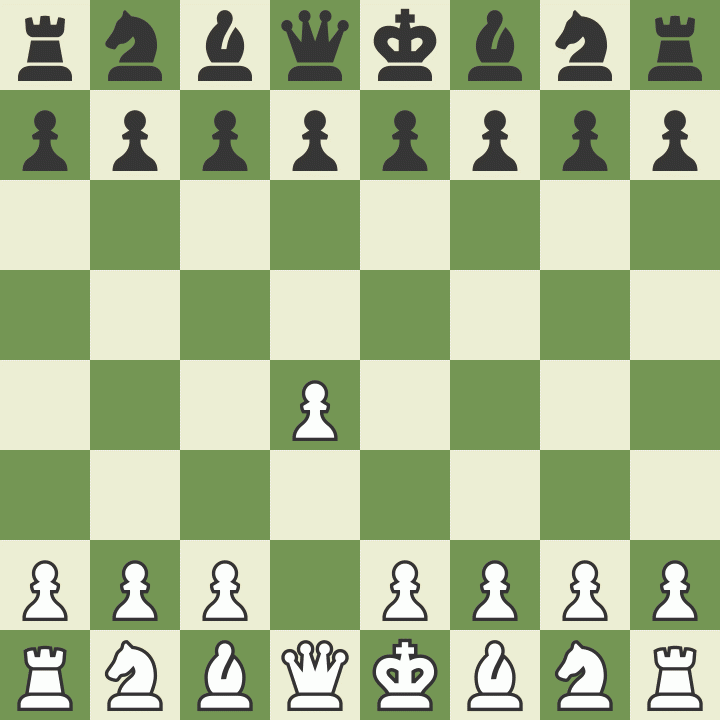
Mr. Yang: You’ve sparked interest with your proposal of a Universal Basic Income (UBI), a no-strings-attached $1000/month transfer to every US citizen. Experts say it’s totally infeasible, but “UBI forever!” certainly makes for a stirring rallying cry.

On the other hand are people who complain your proposal isn’t a real UBI. UBI needs to be enough to live on, but $1000/month wouldn’t even get people all the way to the federal poverty line. In more expensive regions like coasts and cities, it would be even worse. “UBI forever!” might be a good rallying cry, but “UBI (below a real UBI) forever!” is a little less rousing. Then again, US states make their mottos sound portentious by converting them to Latin; maybe that would work for you too.

So do you think a good slogan for your campaign would be “Semper ubi sub ubi”?

# A Very Unlikely Chess Game

Almost 25 years after Kasparov vs. Deep Blue, another seminal man vs. machine matchup:



Neither competitor has much to be proud of here. White has a poor opening. Black screws up and loses his queen for no reason. A few moves later, white screws up and loses his rook for no reason. Better players will no doubt spot other humiliating mistakes. But white does eventually eke out a victory. And black does hold his own through most of the game.

White is me. My excuse is that I only play chess once every couple of years, plus I’m entering moves on an ASCII board I can barely read.

Black is GPT-2. Its excuse is that it’s a text prediction program with no concept of chess. As far as it knows, it’s trying to predict short alphanumeric strings like “e2e4” or “Nb7”. Nobody told it this represents a board game. It doesn’t even have a concept of 2D space that it could use to understand such a claim. But it still captured my rook! Embarrassing!

Backing up: last year, I wrote GPT-2 As Step Toward General Intelligence, where I argued that the program wasn’t just an essay generator, it was also kind a general pattern-recognition program with text-based input and output channels. Figure out how to reduce a problem to text, and you can make it do all kinds of unexpected things.

Friend-of-the-blog Gwern Branwen has been testing the limits of this idea. First he taught GPT-2 to write poetry. Some of it was pretty good:

Fair is the Lake, and bright the wood,  
With many a flower-full glamour hung:  
Fair are the banks; and soft the flood  
With golden laughter of our tongue.

For his next trick, he found a corpus of music in “ABC notation”, a way of representing musical scores as text. He fed it to GPT-2 and got it to write folk songs for him. I’m a fan:

Last month, I asked him if he thought GPT-2 could play chess. I wondered if he could train it on a corpus of chess games written in standard notation (where, for example, e2e4 means “move the pawn at square e2 to square e4”). There are literally millions of games written up like this. GPT-2 would learn to predict the next string of text, which would correspond to the next move in the chess game. Then you would prompt it with a chessboard up to a certain point, and it would predict how the chess masters who had produced its training data would continue the game – ie make its next move using the same heuristics they would.

Gwern handed the idea to his collaborator Shawn Presser, who had a working GPT-2 chess engine running within a week:

GPT-2 chess is promising. After an hour of training, 1.5B is pretty good at opening theory.

Longer sequences tend to fail due to invalid moves, but this shows it's possible in principle to make a GPT-2 chess engine.

And maybe after more training it'll make fewer invalid moves. pic.twitter.com/DqC4WiPfHV

— Shawn Presser (@theshawwn) January 1, 2020

I'll post some games, up to the point it generates an invalid move (which seems to happen around move 11).

Paste this into https://t.co/wWGpVu9ko6

1.e4 c5 2.Nf3 e6 3.d4 cxd4 4.Nxd4 a6 5.Bd3 Nf6 6.Nc3 d6 7.O-O Be7 8.f4 O-O 9.Kh1 Nbd7 pic.twitter.com/8Bl2ijZiCZ

— Shawn Presser (@theshawwn) January 1, 2020

After a day of training (2.4M examples), GPT-2 1.5B can reach move 14 with no invalid moves.

1.e4 e5 2.Nf3 d6 3.d4 exd4 4.Qxd4 a6 5.Be2 Nf6 6.O-O Be7 7.Re1 O-O 8.c3 b5 9.a4 Bb7 10.axb5 axb5 11.Nbd2 Re8 12.h3 g6 13.Ra5 Qd7 14.Ng5 c5 pic.twitter.com/2XuH6iLaD5

— Shawn Presser (@theshawwn) January 2, 2020

It can reach midgame by removing invalid moves.

1.e4 c5 2.Nf3 d6 3.Bb5+ Nd7 4.O-O a6 5.Be2 b6 6.a4 e6 7.d4 Be7 8.c3 9.Nbd2 Ne5 10.Nxe5 dxe5 11.12.13.f4 Bb7 14.Bd3 g6 15.Nf3 16.17.Re1 18.Qe2 Bf8 19.Bd2 20.Qf2 Qxd4 21.22.23.Rf1 Qxc3 24.Qxc5 Bd6 25.Qc7 26.Qd8+ 27.Qd6 Rd8 pic.twitter.com/3uyPaP9LHt

— Shawn Presser (@theshawwn) January 2, 2020

GPT2 Chess update: I wrote some code to calculate the probability of all valid chess moves. It can reach endgame now. https://t.co/QQzhZJmgQ9

It starts to blunder every game at around move 13. We suspect it’s losing track of board state. (It’s trained solely on PGN notation.)

— Shawn Presser (@theshawwn) January 4, 2020

I am preparing to release a notebook where you can play chess vs GPT-2. If anyone wants to help beta test it:

1. visit https://t.co/CpWrFvtnY22. open in playground mode3. click Runtime -> Run All4. Scroll to the bottommost cell and wait 6 minutes

If you get stuck, tell me.

— Shawn Presser (@theshawwn) January 6, 2020

You can play against GPT-2 yourself by following the directions in the last tweet, though it won’t be much of a challenge for anyone better than I am.

This training explains the program’s strengths (good at openings) and weaknesses (bad when play deviates from its expectations). For example, ggreer analyzes why GPT-2 lost its queen in the game above. By coincidence, my amateurish flailing resembled a standard opening called the Indian Game. GPT-2 noticed the pattern and played a standard response to it. But the resemblance wasn’t perfect, so one of GPT-2’s moves which would have worked well in a real Indian Game brought its queen where I could easily capture it. I don’t want to conjecture on how far “mere pattern-matching” can take you – but you will at least need to be a better pattern-matcher than this to get very far.

But this is just what a friend of a friend managed to accomplish in a few days of work. Gwern stresses that there are easy ways to make it much better:

Obviously, training on just moves with the implicit game state having to be built up from scratch from the history every time is very difficult – even MuZero at least gets to see the entire game state at every move when it’s trying to predict legal & good next moves, and depends heavily on having a recurrent state summarizing the game state. Maybe rewriting games to provide (state,action) pairs will make GPT-2 work much better.

What does this imply? I’m not sure (and maybe it will imply more if someone manages to make it actually good). It was already weird to see something with no auditory qualia learn passable poetic meter. It’s even weirder to see something with no concept of space learn to play chess. Is any of this meaningful? How impressed should we be that the same AI can write poems, compose music, and play chess, without having been designed for any of those tasks? I still don’t know.

# What Intellectual Progress Did I Make In The 2010s?

One of the best parts of writing a blog is being able to answer questions like this. Whenever I felt like I understood new and important, I wrote a post about it. This makes it easy to track what I learned.

I think the single most important thing I discovered this decade (due to a random comment in the SSC subreddit!) was the predictive coding theory of the brain. I started groping towards it (without knowing what I was looking for) in Mysticism And Pattern-Matching, reported the exact moment when I found it in It’s Bayes All The Way Up, and finally got a decent understanding of it after reading Surfing Uncertainty. At the same time, thanks to some other helpful tips from other rationalists, I discovered Behavior: The Control Of Perception, and with some help from Vaniver and a few other people was able to realize how these two overarching theories were basically the same. Discovering this area of research may be the best thing that happened to me the second half of this decade (sorry, everyone I dated, you were pretty good too).

Psychedelics are clearly interesting, and everyone else had already covered all the interesting pro-psychedelic arguments, so I wrote about some of my misgivings in my 2016 Why Were Early Psychedelicists So Weird?. The next step was trying to fit in an understanding of HPPD, which started with near-total bafflement. Predictive processing proved helpful here too, and my biggest update of the decade on psychedelics came with Friston and Carhart-Harris’ Relaxed Beliefs Under Psychedelics And The Anarchic Brain, which I tried to process further here. This didn’t directly improve my understanding of HPPD specifically, but just by talking about it a lot I got a subtler picture where lots of people have odd visual artifacts and psychedelics can cause slightly more (very rarely, significantly more) visual artifacts. I started the decade thinking that “psychedelic insight” was probably fake, and ended it believing that it is probably real, but I still don’t feel like I have a good sense of the potential risks.

In mental health, the field I am supposed to be an expert on, I spent a long time throwing out all kinds of random ideas and seeing what stuck – Boorsboom et al’s idea of Mental Disorders As Networks, The Synapse Hypothesis of depression, etc. Although I still think we can learn something from models like those, right now my best model is the one in Symptom, Condition, Cause, which kind of sidesteps some of those problems. Again, learning about predictive processing helped here, and by the end of the decade I was able to say actually useful things that explained some features of psychiatric conditions, like in Treat The Prodrome. Friston On Computational Mood might also be in this category, I’m still waiting for more evidence one way or the other.

I also spent a lot of time thinking about SSRIs in particular, especially Irving Kirsch (and others’) claim that they barely outperform placebo. I wrote up some preliminary results in SSRIs: Much More Than You Wanted To Know, but got increasingly concerned that this didn’t really address the crux of the issue, especially after Cipriani et al (covertly) confirmed Kirsch’s results (see Cipriani On Antidepressants). My thoughts evolved a little further with SSRIs: An Update and some of my Survey Results On SSRIs. But my most recent update actually hasn’t got written up yet – see the PANDA trial results for a preview of what will basically be “SSRIs work very well on some form of mental distress which is kind of, but not exactly, depression and anxiety”.

One place I just completely failed was in understanding the psychometrics of autism, schizophrenia, transgender, and how they all related to each other and to the normal spectrum of variation. I kind of started this program with Why Are Transgender People Immune To Optical Illusions? (still a good question!), fumbled around by first-sort-of-condemning and then sort-of-accepting the diametrical model of autism and schizophrenia, and then admitting I just didn’t know what was going on in this area and not talking about it much more. I still sometimes have thoughts like “Is borderline the opposite of autism?” or “Are schizoid people unusually charismatic, unusually uncharismatic, or somehow both?”, and I still have no idea how to even begin answering them. Autism And Intelligence: Much More Than You Wanted To Know at least helped address a very tangentially related question and is probably the closest thing to a high point this decade gave me here.

The Nurture Assumption shaped my 2000s views of genetics and development. Ten years later, I’m still trying to process it, and in particular to square the many behavioral genetics studies showing nonshared environment doesn’t matter with the many other studies suggesting it does (see eg The Dark Side Of Divorce and Shared Environment Proves Too Much). I think I started to get more of handle on attachment theory and cPTSD as both being different aspects of the same basic predictive processing concept of “a global prior on the world being safe” – see Mental Mountains and Evolutionary Psychopathology for two different ways of approaching this concept. This made me conclude that I might have been wrong about preschool (though see also Preschool: Much More Than You Wanted To Know). Honestly I am still confused about this. The one really exciting major good update I made about genetics this decade was understanding and fully internalizing the omnigenic model.

One of the big motivating questions I keep coming back to again and again is – what the heck is “willpower”? I started the decade so confused about this that I voluntarily bought and read Baumeister and Tierney’s book Willpower and expected it to be helpful. I spent the first few years gradually internalizing the lesson (which I learned in the 2000s) that Humans Are Not Automatically Strategic (see also The Blue-Minimizing Robot as a memorial to the exact second I figured this out), and that hyperbolic discounting is a thing. Since then, progress has been disappointing – the only two insights I can be even a little happy about are understanding perceptual control theory and Stephen Guyenet’s detailed account of how motivation works in lampreys. If I ever become a lamprey I am finally going to be totally content with how well I understand my motivational structure, and it’s going to feel great.

Speaking of Guyenet, if nothing else this last decade has taught us that Gary Taubes did not solve all of nutrition in 2004, that Atkins/paleo/keto are good for some people and bad for others, and that diet is still hard. See the various Guyenet vs. Taubes and Taubes vs. Guyenet posts, and my 2015 The Physics Diet on where I was at that point. So what is going on with diet? Compressing an entire decade’s worth of research into two words, the key phrase seems to be “set point” (which, credit to Taubes, he was one of the first people to popularize). See eg Anorexia And Metabolic Set Point and Del Giudice On The Self-Starvation Cycle. But what is the set point and how does it get dysregulated? See my book review of The Hungry Brain for the best answer to that I have now (not so good). This whole mess helped me get a better understanding of Contrarians, Crackpots, and Consensus, and eventually ended up with me Learning To Love Scientific Consensus.

In terms of x-risk: I started out this decade concerned about The Great Filter. After thinking about it more, I advised readers Don’t Fear The Filter. I think that advice was later proven right in Sandler, Drexler, and Ord’s paper on the Fermi Paradox, to the point where now people protest to me that nobody ever really believed it was a problem. AI has been the opposite – I feel like the decade began with people pooh-poohing it, my AI Researchers On AI Risk was part of a large-scale effort to turn the tide, and now it’s more widely accepted as an important concern. At the same time, the triumphs of deep learning has made things look a little different – see How Does Recent AI Progress Affect The Bostromian Paradigm? and Reframing Superintelligence – and I’ll be reviewing Human Compatible soon. I also got some really great insights on what “human-level intelligence” means from the good people at AI Impacts, which I wrote up as first Where The Falling Einstein Meets The Rising Mouse and later Neurons And Intelligence: A Bird-Brained Perspective (see also Cortical Neuron Number Matches Intuitive Perceptions Of Moral Value Across Animals and all the retractions and meta-retractions thereof). Overall I think I’ve updated a little (though not completely) towards non-singleton scenarios and not-super-fast takeoffs, which combined with the increased amount of effort being put into this area is cause for a little more optimism than I had in 2010. I know some smart people disagree with me on this.

In the 2000s, people debated Kurzweil’s thesis that scientific progress was speeding up superexponentially. By the mid-2010s, the debate shifted to whether progress was actually slowing down. In Promising The Moon, I wrote about my skepticism that technological progress is declining. A group of people including Patrick Collison and Tyler Cowen have since worked to strengthen the case that it is; in 2018 I wrote Is Science Slowing Down?, and late last year I conceded the point. Paul Christiano helped me synthesize the Kurzweillian and anti-Kurzweillian perspectives into 1960: The Year The Singularity Was Cancelled.

In 2017, I synthesized some thoughts that had been bouncing around about rising prices into Considerations On Cost Disease, still one of this blog’s most popular posts. I felt like early responses were pretty weak, although they brought up a few interesting points on veterinary medicine, cosmetic medicine, and other outliers that I still need to transform into a blog post; Alon Levy’s work on infrastructure in particular has also been great. The first would-be-general-answer that made me sit up and take notice was Alex Tabarrok’s book (link goes to my review) The Prices Are Too Damn High – but I explain there why I don’t think it can be the full answer. The most recent thing I learned (tragically underhighlighted in my wage stagnation post) is that a lot of apparent wage stagnation is due to cost disease – consumer services ballooning in cost means the consumer inflation index rises faster than the business inflation index, productivity gets measured by business inflation, wages get measured by consumer inflation, and so it looks like productivity is outpacing wages. This is still only half of the apparent decoupling, but it’s still a big deal.

The highlight/lowlight of the decade in social science was surely the replication crisis. My first inkling that something like this might exist was in December 2009, from the Less Wrong post Parapsychology: The Control Group For Science. There were a couple of years where people were trying to figure out how bad the damage was; of these, my 90% Of All Claims About Problems With Medical Studies Are Wrong was more optimistic, and my slightly later The Control Group Is Out Of Control was more pessimistic (I still stand by both). As the decade continued, I think we got better about realizing that many to most older studies were wrong, in a way that didn’t make us feel like total Cartesian skeptics or like we were going to have to throw out evolution or aspirin or any of the things on really sound footing. After that it just became fun: my “acceptance” stage of grief produced some gems like 5-HTTLPR: A Pointed Review.

On SSC, I particularly examined some of the replication issues of growth mindset. I started in 2015 by pointing out that the studies seemed literally unbelievable, but so far nobody had tried attacking them. I claim to have been way ahead of the curve on this one – if you don’t believe me, just read the kind of pushback I got. But by 2017, that situation had changed – Buzzfeed posted an article that called the field into question, but still without clear negative evidence. Finally, over the past few years, the negative studies have come pouring in, accented by supposedly “positive” studies by Dweck & co showing effect sizes only a tiny fraction of what they had originally claimed. The latest research (can’t find it right now) is that praising students for effort rather than for ability has no effect on how hard-working or successful they are, debunking the original headline result that got most people interested in the field and nicely closing the circle.

In 2010 I worked with a medical school professor who studied the placebo effect and realized I didn’t understand it at all. Over the past few years I gradually became more convinced of the heterodox position of Gøtzsche and Hróbjartsson, who believe placebo effect doesn’t apply to anything except pain and a few other purely mental phenomena (The Placebo Singers, Powerless Placebos). I’ve since become less convinced that’s true (just today I treated a patient who I’m pretty sure has psychosomatic vomiting from what he falsely believes was a medication side effect, and if belief can cause vomiting, surely it can also treat it). As with so many other things, it was predictive processing to the rescue – see section IV part 7 of my Surfing Uncertainty review. I now think I have a pretty good understanding of how placebos can treat both purely mental conditions and conditions heavily regulated by the nervous system, while still mostly sticking to Gøtzsche and Hróbjartsson’s findings.

I started this decade confused about how to understand ethics given all the paradoxes of utilitarianism. I’m still 90% as confused now as I was then, but I still feel like I’ve made some progress. A lot of my early thinking involved folk decision theory and contractualism – how would you act if you expected everyone else to act the same way? I explored the edges of this idea in You Kant Dismiss Universalizability and Invisible Nation. I’m not how much it helped my search for metaethical grounding, but it helped me get a more robust understanding of liberalism and clarify my views on some practical questions, eg Be Nice, At Least Until You Can Coordinate Meanness and The Dark Rule Utilitarian Argument For Science Piracy. In general I think this has given me a more cautious theory of decision-making that’s occasionally (and terrifyingly) set me against other more anti-Outside-View rationalists. I think the most important shift in my understanding of ethics this decade was the one I wrote up in Axiology, Morality, Law (formerly titled “Contra Askell On Moral Offsets”), which isn’t related to grounding utilitarianism at all but sure helps make the problem less urgent

Despite my better judgment, I waded into politics a lot this decade. I Can Tolerate Anything Except The Outgroup produced this blog’s first “big break”, but it admitted it didn’t really understand the factors underlying “tribe”. Since then Albion’s Seed helped provide another piece of the puzzle, and a better understanding of class provided another. I went a little further discussing why tribes have ideologies associated with them in The Ideology Is Not The Movement, how that is like/unlike religion in Is Everything A Religion?, and hammered it home unsubtle-ly in Gay Rites Are Civil Rites.

I wrote the Non-Libertarian FAQ sometime around 2012 and last updated it in 2017. Sometime, possibly between those dates, I read David Friedman’s A Positive Account Of Property Rights, definitely among the most important essays I’ve ever read, and got gold-pilled (is that a term? It should be a term). I’ve since been trying to sort this out with things like A Left-Libertarian Manifesto, and trying to move them up a level as Archipelago. James Scott’s Seeing Like A State and David Friedman’s Legal Systems Very Different From Ours were also big influences here. Like all platitudes, “government is a hallucination in the mind of the governed” is easy to understand on a shallow level but fiendlishly complicated on a deep level, but I feel like all of these sources have given me a deep understanding of exactly how it’s true.

The rightists (especially Moldbug) get the other half of the credit for helping me understand Archipelago, and also deserve kudos for teaching me about cultural evolution. My first attempts to engage with this topic were nervous and halting – see eg The Argument From Cultural Evolution. I got a much better feel for this after reading The Secret Of Our Success, and was able to bring this train of thought back to its right-wing roots Addendum To Enormous Nutshell: Competing Selectors. I’m grateful to the many rightists who argued about some of these points with me until they finally stuck.

I had more trouble engaging with leftists. I started with Does Class Warfare Have A Free-Rider Problem, and it took me way too long to figure out that this was one of the major questions sociology was asking, and that “an answer” would look less like “your game theory analogy is missing this one variable” and more like a whole library full of books on what the heck society was. Later the same engagement produced Conflict Vs. Mistake, which I am informed is still unfair and partially inaccurate, but which (take my word for it) is a heck of a lot better than the stuff I was thinking before I wrote it. More recently I’ve been trying to figure out a sympathetic account of activism (as opposed to the unsympathetic account that it’s virtue signaling and/or people who are really bad at figuring out what things are vs. aren’t effective). You can sketch the outline at Respectability Cascades and Social Censorship: The First Offender Model, and I’ll sketch the whole thing out sometime when I have enough emotional energy to deal with the kind of people who will have opinions on it.

I also had to grapple with the sudden rise of social justice ideology. I’m proud of my work on gender differences – both what I learned, how I wrote it up, and the few bits of original research I did (eg Sexual Harassment Levels By Field). My knowledge and claims started off kind of weak (Gender Differences Are Mostly Not Due To Offensive Attitudes), but I eventually feel like I got a really great evidence-based basically-airtight theory of what is going on with gender imbalances in different fields, which I posted most of in Contra Grant On Exaggerated Differences (I’m still thankful for the commenter who solved that one remaining paradox about math majors). And despite all the mobs and vitriol I think sound science has basically triumphed here – I was delighted to recently see as mainstream a blog as Marginal Revolution recently publish, without any caveats or double-talk, a post called Sex Differences In Personality Are Large And Important and get basically no pushback. I was a lot more pessimistic around 2017 or so and described some thoughts on how to make a strategic retreat in Kolmogorov Complicity And The Parable Of Lightning, which I still think is relevant in some areas. But I actually start the new decade really optimistic – I haven’t written up an explanation of why, but careful readers of New Atheism: The Godlessness That Failed may be able to figure it out, especially if they apply some of the same metrics I used there to track how social justice terms have been doing recently.

Upstream of politics, I think I got a better understanding of…game theory? Complex system dynamics? The most important post here was Meditations On Moloch; the sequel/expansion, whose thesis I have yet to write up in clear prose, is The Goddess Of Everything Else. Reading Inadequate Equilibria was also helpful here.

My understanding of “enlightenment” went from total mystical confusion to feeling like I have a pretty good idea what claims are being made, and mostly believing them. This line of thinking started with the Mastering The Core Teachings Of The Buddha review, and then was genuinely helped by Vinay Gupta’s contributions summed up in Gupta On Enlightenment, despite the disaster in the comments of that post. From there I progressed to reading The Mind Illuminated, and Is Enlightenment Compatible With Sex Scandals led me to discover The PNSE Paper, which as much as anything else helped ground my thinking here (the comments there were pretty good too).

And thanks to all of you who took the survey, I went from skepticism of birth order effects to saying Fight Me, Psychologists: Birth Order Effects Exist And Are Very Strong. This was bolstered by Eli Tyre and Bucky’s posts on Less Wrong about birth order in mathematicians and physicists respectively. Last year I expanded on that with a post on how birth order responded to age gaps (somewhat updated and modified here, thanks Bucky). Once this year’s survey results are in I expect to have a lot more data on exactly what causes birth order effects and maybe how to deal with them. If you haven’t taken the SSC survey this year, consider this your reminder to do it here.

Not many of these were total 180 degree flips in my position (though birth order, preschool, psychedelic insight, and the rate of scientific progress are close). And not many of them completely resolved a big question that had been bothering me before (though the Fermi Paradox paper, omnigenic model, and animal neuron work did). A few of them confirmed things I had only suspected before (growth mindset, gender imbalances, diet). Many of them feel like what MIRI calls “deconfusion”, turning a space full of unknown unknowns to one where you feel like you have a decent map of where the major problems are and what it would feel like to solve them. The enlightenment research seems to fit here – I went from “I have no idea how to even think about this question or whether it’s all fake” to “I don’t know exactly what’s going on here, but I know what needs to be explained, and it looks like the explanation will have a shape that fits nicely into the rest of my ontology.”

There’s an argument that I should learn less each decade, since I’ll be picking higher and higher fruit. My own knowledge can advance either because civilization advances and I hear about it, or because I absorb/integrate older knowledge that I hadn’t noticed before. Civilization advances at a decade per decade (or maybe less; see the Cowen & Southwood paper above), but each year it becomes harder and harder to find relevant older knowledge that I haven’t integrated yet. I plausibly only have five more decades to live, and I don’t think I’d be happy only advancing five times this amount over the rest of my life, let alone less than that.

But I notice I only started SSC about halfway through the decade, and that my progress picked up a lot after that. I don’t think it’s just recall bias from being able to track myself better. I think being able to put ideas out there and have you guys comment on them and link me to important resources I might have missed has been great for me. I only started taking full advantage of that around 2015; this decade I have a head start. And maybe I’ll discover other useful tools that will speed things up further.

Thanks for sticking around with this blog, and have a happy third decade of the twenty-first century.

# 2019 Adversarial Collaboration Winners

Thanks to everyone who participated and/or voted in the 2019 Adversarial Collaboration Contest. And the winner is…

…

…

Adrian Liberman and Calvin Reese, for Does Calorie Restriction Slow Aging?.

An extraordinarily close second place (26.9% vs. 26.2% of votes) goes to David G and Froolow, for Is Eating Meat A Net Harm?.

Both of these did great research and were written up well. I especially like them as winners because they have such different strengths.

The calorie restriction collaboration was carefully focused on a factual question. I think this is a promising model for adversarial collaborations, and that others failed the further they deviated from this. For example, the circumcision collaboration did a good job assessing the quantifiable benefits and harms of the practice, but it turned out that most people who disagreed about it weren’t disagreeing because they assessed quantifiable benefits and harms differently. The abortion collaboration ended up in a similar place. By focusing on a topic where there really was debate about what the research showed, and by hitting the lit review portion out of the park, Adrian and Calvin helped deconfuse a lot of previously confused people.

And the meat collaboration managed to succeed without being like this at all! It was unabashedly just a lengthy review of every single plausible argument for and against vegetarianism, and bulldozed over the immense difficulties with this approach by putting in more work than any reasonable person would have thought possible. And if it didn’t get quite as many votes as calorie restriction, it won on another metric – here are some of voters’ comments (plus some extra from the blog):

– After this, I expect to restrict more strictly to chicken and fish, and alternate more aggressively towards fish as a hedge against the possibility I’m undervaluing chicken sapience.

– Got me to significantly change my diet (at least over the past few weeks) towards more fish and much less chicken.

– I read it, and I changed what I believe and how I changed my life accordingly.

– My final pick is not necesarily based on the article‘s excellence but rather due to me going to change my eating behaviour, which I find an impressive thing Form an article to do.

– This has practical implications for my life. I’ve stopped eating pork because of it (baby steps, working towards less meat generally)

– I actually resolved to stop eating chicken (the only meat I can regularly eat, due to dietary restrictions) based on that piece, so I’d say it was pretty effective in informing me about things.

– It convinced me not to buy chicken that isn’t organic/free-range

– This article will result in all my family eating less meat. It’s actually going to change our lives, health, and the environment!

– As a result of the adversarial collaboration on the ethics of eating meat, have dramatically scaled back my meat consumption to probably 20% of my previous value, and the meat I do eat now is almost exclusively fish and invertebrates.

– Learning about the harm of factory farming from their dispassionate and empirical analysis has prompted me to greatly reduce my red meat and poultry consumption.

Some voters brought up a reasonable complaint: the end result ended up being pretty (though not completely) pro-vegetarian. How do we tell the difference between “a good faith effort by intelligent people naturally converges on vegetarianism” vs. “the anti-vegetarian collaborator slacked off”?

In this case, we tell because the anti-vegetarian collaborator posted a comment about his thought process and what convinced him. But there were other cases where people had the same question, and still other cases where one collaborator did a good job representing their own anti-X position, but other people were anti-X for different reasons that didn’t get represented.

If I had infinite resources, I would fund adversarial collaborations between well-known and universally-recognized intellectuals on different sides of a topic, who everybody trusted to stick to their guns. As it is, I can only say I’m delighted to have stumbled into the one part of the world where “people are too likely to change their mind when presented with new evidence on controversial issues” is a problem.

Some thoughts on the other collaborations:

Circumcision: I loved this one. I’d never seen a good assessment of exactly what health risks circumcision was supposed to prevent, and I didn’t know how weak the evidence was that the foreskin helps with sexual pleasure. But the conclusion ended up being “the quantifiable benefits of circumcision are nonzero but pretty low; the quantifiable harms are not obviously distinguishable from zero but who knows”, which leaves a lot of space for people’s ethical intuitions, which turned out to be REALLY STRONG. One reader said they were going to boycott my blog from now on for not having no-platformed this ACC, and a few others seemed only slightly less angry. On the other hand, it also did better than average among voters, so good job there. I take a small amount of blame for this one not being more popular – I retitled it to be about the ethics of circumcision, whereas the original title had been about benefits vs. harms. But I think it’s naturally hard to write something about benefits and harms without it sounding like you’re talking about ethics, and in this case the ethics were too complicated to fit in the model provided for them. Some positive comments from the survey: “This…actually changed my opinion from circumcision being mildly ethically wrong back to neutral”, “I gained a much more nuanced understanding of the benefits position to the point that my mind was changed to be in favor (maybe too strong), or at least not opposed to, it for developing countries”, “It tidily presented the pros and cons and presented a lot of useful information, with a clear conclusion. It shifted my thinking the most of all of them.”

Space Colonization: You guys presented a lot of evidence for one side, then at the end switched to the opposite side based on a one paragraph explanation of something you’d never brought up before. If that was your crux, I wish you had analyzed it in more depth. If the whole point is to make something that can’t be defunded, couldn’t the government (or whoever) give the money to a private foundation with really good trustees, no takebacks? Maybe there’s a problem with that idea, I don’t know, but if you’re going to make defundability the center of your conclusion, I wish you had examined it more closely. Some positive comments from the survey: “Excellent selection of question, manages to present both sides fairly and come to an insightful conclusion”, “I think this ACC did the best job of covering the entire scope of the question they assigned to themselves, while still presenting a shared conclusion”, “Interesting non-obvious conclusion, subject I care a lot about, pretty pictures”.

Gene Editing: This one seemed to spend a lot of time on very knowledgeable and very well-cited assessments of the current state of the technology and how and why it worked, but didn’t really get around to assessing the “should” question in the title. It also had a few factual missteps – maybe no more than the others, but more obvious since it was so fact-based. While it was an impressive work of scholarship I’m not sure it came together as an adversarial collaboration. Some positive comments from the survey: “Very nicely presented ACC. It was thought provoking and totally enrapturing!”, “well-reasoned collaboration on a difficult question”, “Great, nuanced answer to a complex question”, “This collaboration caused me to reconsider my enthusiasm for CRISPR based on the narrative provided in most press releases. The topic is much more involved than I’d initially realized.”

Abortion: An adversarial collaboration on a completely moral question – you guys didn’t make this easy for yourself, did you? I don’t think you made any particular missteps given the difficult task you set yourselves, but this is another one that I feel like didn’t quite come together. Some positive comments from the survey: “Most interesting (and politically relevant) topic, plus it seems icerun’s position actually shifted somewhat by virtue of having to marshall arguments for it, proving the whole endeavor to be more than just an exercise in futility”, “I thought it best captured the spirit of an Adversarial Collaboration”, “This was a nice, cautious walk-through of an extremely divisive subject. i never thought i would enjoy reading a “point/counterpoint” on abortion, but i enjoyed this one”, “Lots of adversity, focused on the actual disagreement, and balanced data and philosophy well.”

Automation: Seemed broadly correct and helpful. I didn’t find it too exciting because I felt like I had already covered most of the same beats in this article (which they cited), but I’m surprised other people didn’t vote for it more. Some positive comments from the survey: “Importance of the issue and the thoroughness with which it was explored”, “The most fitting, thought-out and the one that draws actual conclusions”, “Highest rationality-to-contentiousness ratio”.

Spiritual Experience: This was another one that was long, fascinating, and didn’t seem to be making much of an attempt to come to a conclusion. I especially liked the section on near death experiences, and I’ll be thinking about it a lot, but I didn’t feel like this collaboration gave me the tools I would need to generate or test hypotheses about what might be going on. Some positive comments from the survey: “The most polished and one which most likely caused me to reconsider things”, “Most informative. Best at following an ideal format”, “This is the one that 1) is most interesting to me, 2) seems like it had a strong difference of opinion as a starting point.”

I included the positive comments because I think comments on these kinds of things (mine and others) naturally tends to skew negative. Certainly the comments in the comments section were overwhelmingly negative even for the winning collaborations (seriously, what was up with this?) So I want to counter this by pointing out that every collaboration got at least 25 votes, and the comments on the voting survey were mostly positive. It’s easier to nitpick than to give praise where praise is due, but people put in a lot of work here and it was generally appreciated.

I promised that I would come up with some fair way of dividing the prize money, with at least 50% going to the first place winner. Because the top two entries were so close, and because I was so impressed with the second place winner, I choose to give $1,300 to Calvin and Adrian ($650 each), and $1,200 to David and Froolow ($600 each). Please send me an email at scott[at]slatestarcodex[dot]com telling me where to send your share of the money – I can PayPal it to you or donate it to a charity of your choice. Thanks to SSC Patreon supporters for making this possible.

As much as I enjoyed this, I don’t expect to do another contest next year. For one thing, I think requiring two people made it a lot harder – 22 out of 30 teams dropped out before the deadline, and I worry some of that involved a lot of wasted work. For another thing, it involved a surprising amount of work on my part converting whatever Word or Google Docs file people sent me into a format I could use on the blog. Finally, I feel like the past two years did a good job exploring this medium, and now it’s up to other people with real questions to see if they can adapt it to their needs.

Most likely I’ll be replacing this with a book review contest sometime towards the end of next year, so if you read any good books, keep them in mind.

But I continue to be interested in adversarial collaborations. If you happen to do one, please tell me – there’s a decent chance I’ll publish it.

# Contra Contra Contra Caplan On Psych

I.

In 2006, Bryan Caplan wrote a critique of psychiatry. In 2015, I responded. Now it’s 2020, and Bryan has a counterargument. I’m going to break the cycle of delay and respond now, and maybe we’ll finish this argument before we’re both too old and demented to operate computers.

Bryan writes:

1. With a few exceptions, Scott fairly and accurately explains my original (and current) position.

2. Scott correctly identifies several gray areas in my position, but by my count I explicitly acknowledged all of them in my original article.

3. Scott then uses those gray areas to reject my whole position in favor of the conventional view.

4. The range of the gray areas isn’t actually that big, so he should have accepted most of my heterodoxies.

5. If the gray areas were as big as Scott says, he should reject the conventional view too and just be agnostic.

I think the gray areas are overwhelming and provide proof that Bryan’s strict dichotomies don’t match the real world.

I also think, as a general philosophical point, that we ought to be suspicious of arguments of the form “the gray areas are small”. Even if this is true, and your model only fails in a few places, controversial questions are likely to be controversial questions precisely because they’re located where your model fails. Nobody challenges a model on an exactly typical case where everything makes sense. So if a point is under debate, let’s say in a fifteen year back-and-forth argument between two bloggers that’s attracted hundreds of total comments, the a priori size of the gray areas doesn’t matter. Even if your model is good at most things, you have strong evidence this isn’t one of them.

In this case, the model we’re debating is Bryan’s idea of constraints vs. preferences. My previous summary of this (which Bryan endorses) goes like this:

Consumer theory distinguishes between two different reasons why someone might not buy a Ferrari – budget constraints (they can’t afford one) and preferences (they don’t want one, or they want other things more). Physical diseases seem much like budget constraints – the reason a paralyzed person can’t run a marathon is because it’s beyond her abilities, simply impossible. Psychiatric diseases seem more like preferences. There’s nothing obvious stopping an alcoholic from quitting booze and there’s nothing obvious preventing someone with ADHD from sitting still and paying attention. Therefore they are best modeled as people with unusual preferences – the one with a preference for booze over normal activities like holding down a job, the other with a high dispreference for sitting still and attending classes. But lots of people have weird preferences. Therefore, psychiatric diseases should be thought of as within the broad spectrum of normal variation, rather than as analogous to physical diseases.

I countered by pointing out that this was in fact very analogous to physical diseases:

Alice has always had problems concentrating in school. Now she’s older and she hops between a couple of different part-time jobs. She frequently calls in sick because she feels like she doesn’t have enough energy to go into work that day, and when she does work her mind isn’t really on her projects. When she gets home, she mostly just lies in bed and sleeps. She goes to a psychiatrist who diagnoses her with ADHD and depression.

Bob is a high-powered corporate executive who rose to become Vice-President of his big Fortune 500 company. When he gets home after working 14 hour days, he trains toward his dream of running the Boston Marathon. Alas, this week Bob has the flu. He finds that he’s really tired all the time, and he usually feels exhausted at work and goes home after lunch; when he stays, he finds that his mind just can’t concentrate on what he’s doing. Yesterday he stayed home from work entirely because he didn’t feel like he had the energy. And when he gets home, instead of doing his customary 16 mile run he just lies in bed all day. His doctor tells him that he has the flu and is expected to recover soon.

At least for this week Alice and Bob are pretty similar. They’d both like to be able to work long hours, concentrate hard, and stay active after work. Instead they’re both working short hours, calling in sick, failing to concentrate, and lying in bed all day.

But for some reason, Bryan calls Alice’s problem “different preferences” and Bob’s problem “budgetary constraints”, even though they’re presenting exactly the same way! It doesn’t look like he’s “diagnosing” which side of the consumer theory dichotomy they’re on by their symptoms, but rather by his assumptions about the causes.

But Bryan doesn’t budge:

I’m unimpressed, because I not only anticipated such objections in my original paper, but even proposed a test to help clarify the fuzziness…can we change a person’s behavior purely by changing his incentives? If we can, it follows that the person was able to act differently all along, but preferred not to; his condition is a matter of preference, not constraint. I will refer to this as the ‘Gun-to-the-Head Test’. If suddenly pointing a gun at alcoholics induces them to stop drinking, then evidently sober behavior was in their choice set all along. Conversely, if a gun-to-the-head fails to change a person’s behavior, it is highly likely (though not necessarily true) that you are literally asking the impossible.

I then presented multiple forms of evidence that a wide range of alleged mental illnesses are responsive to incentives. Scott barely mentions said evidence.

Still, does this mean that the flu isn’t “really” an illness either? No. Rather it means that physical illness often constrains behavioral and changes preferences. When sick, the maximum amount of weight I can bench press falls. (Yes, I’ve actually tried this). Yet in addition, I don’t feel like lifting weights at all when I’m sick. Anyone who has worked while ill should be able to appreciate these dual effects. If you literally get sick, your ability and desire to work both go down. When you metaphorically get “sick of your job,” in contrast, only your desire goes down.

I reject the heck out of this answer. I agree the “gun to the head” test is a good summary of Bryan’s position, but we already agreed what Bryan’s position is. The only thing he’s adding here is a claim that the flu still qualifies as a real disease because it sometimes constrains behavior (the amount of weight Bryan can lift). But nobody cares how much weight they can lift during a flu! When we talk about having the flu being bad, we’re talking 0% about how much weight we can lift, and 100% about the sorts of problems Bob has – feeling too ill to go to work, not wanting to do things, etc. If Bryan searches hard enough, he can find a way the flu results in slightly weaker muscle strength. But if I search hard enough, I can find a way depression results in slightly weaker muscle strength. Neither of these things are what the average person thinks about when they think of “flu symptoms” or “depression symptoms”, and I consider them both equally irrelevant.

But if a change in weight-lifting ability really disqualifies the flu for Bryan, we can talk about other diseases.

What about shingles? It’s a viral infection that causes a very itchy rash. But sometimes (herpes sine zoster) the rash isn’t visible, and you just get really itchy for a few days. Like, really itchy. I had this condition once and it was just embarrassing how much I was scratching myself. But if you had put a gun to my head and said “Don’t scratch yourself, or I’ll kill you”, I would have sat on my hands and suffered quietly. For Bryan, an itch is just a newfound preference for scratching yourself. Shingles, like depression or ADHD, is just a preference shift, and so doesn’t qualify as a real disease.

Or what about respiratory tract infections that cause coughing? My impression is that, put a gun to my head, and I could keep myself from coughing, even when I really really felt like it. Coughing is a preference, not a constraint, and Bryan, to be consistent, would have to think of respiratory infections as just a preference for coughing.

Or what about migraines? Sure, people with migraines say they feel pain, but that’s no better grounded than someone with depression saying they feel sad. If Bryan is allowed to bring in concepts like “pain”, I’m allowed to bring in concepts like “sadness”, “anxiety”, etc. And since an anxious person feels anxiety and cannot stop feeling it even if threatened with a gunshot, the anxiety counts as a constraint, and so mental disorders are constraining. For Bryan’s constraints-vs-preferences dichotomy to work at all, he has to endorse a sort of behaviorism, where we need not believe anything that doesn’t express itself as behavior. And the only behavior we see in a migraine is somebody lying in bed, turning off all the lights, and occasionally clutching their head and saying “auggggh”. But put a gun to their head and demand they be in a bright room with lots of loud music, and they’ll go to the bright room with lots of loud music. Threaten to shoot them unless they stop clutching their head and moaning, and they’ll stop clutching their head and moaning. In Bryan’s model, migraines are just a newfound preference for saying “auggggh” a lot. Why medicalize this? Some people like saying “auggggh” and that’s valid!

Bryan’s preference vs. constraint model doesn’t just invalidate mental illness. It invalidates many (maybe most) physical illnesses! Even the ones it doesn’t invalidate may only get saved by some triviality we don’t care about – like how maybe you can lift less weight when you have the flu – and not by the symptoms that actually bother us.

II.

We need a model that lets us describe shingles as something more than “this person has a preference for scratching themselves frantically, and that preference is valid, nothing to worry about here”. I don’t have a beautiful elegant version of a model like this yet, but I think Bryan himself has gone most of the way to an at-least-adequate one.

In his post The Depression Preference, Bryan admits that most depressed people don’t want to be depressed. But he terms this a meta-preference – a preference over preferences. They have depressive preferences – for example, a preference for sitting around crying rather than doing work. They would meta-prefer not to have those preferences. But they do have them.

I agree this is a fruitful way to look at things, but I think we have to be really careful here, and that using the same term for endorsed meta-preferences and unendorsed object-level preferences is preventing this level of care. Let’s call endorsed preferences which people meta-prefer to have “goals”, and unendorsed preferences which people would meta-prefer not to have “urges”. I think this closely matches our intuitive understanding of these terms.

Suppose I created a sinister machine that beamed mind control rays into Bryan’s head and gave him an urge to constantly slap himself in the face. This urge could theoretically be resisted, but it’s so strong that in practice he never managed to resist it. It didn’t make him enjoy slapping himself in the face, or think this was a reasonable thing to do. It just made him compulsively want to keep doing it. He loses his job, his friends, and his dignity, because nobody wants to be around someone who’s slapping himself in the face all the time. I hope we can common-sensically agree on the following:

1. This is bad  
2. Bryan would want to find and destroy the sinister machine  
3. That would be a pretty reasonable goal for Bryan to have, and society should support him in this

This seems a lot like the shingles case. A sinister outside imposition (the viral infection) gives its victim an urge to constantly scratch themselves. It doesn’t make them enjoy scratching themselves, or think this is a reasonable thing to do. These people want to cure their shingles infection, and everyone agrees this desire is reasonable.

But this also seems a lot like some cases of OCD. Did you know that a subset of childhood OCD is caused by a streptococcal infection? So again, you get a sinister outside imposition (an infection) that gives its victim an urge to, let’s say, wash their hands fifty times a day. It doesn’t make them enjoy washing their hands, or think this is a reasonable thing to do (some OCD patients do believe their rituals are necessary, others don’t). These people want to cure their OCD, and I at least agree this desire is reasonable.

If you would support the sinister machine victim and the shingles victim, it’s hard for me to see a case for putting the OCD victim in a different category. I agree I’m using as clear a case as possible (most mental disorders aren’t obviously due to infections), but both Bryan and I are trying to avoid bringing specific facts about biology into this mostly-philosophical debate. The distinction between goals and urges turns what looked like an acceptable situation (these people are following their preferences, which is good) into an unacceptable situation (these people’s goals are being thwarted by unwelcome urges which they can’t resist).

I expect most of Bryan’s skepticism to focus on those last two words – “can’t resist”. He will no doubt bring up his gun-to-the-head test again. If we put a gun to the head of a shingles patient, they could stop scratching. So although we can be sympathetic to the trouble their unwanted new preference causes them, how can we recommend anything other than “just suck it up and resist the preference”?

The best model of decision-making I know of comes from research on lampreys. Various areas of the lamprey brain come up with various plans – hunt for food, hide under a rock, wriggle around – and calculate the “strength” of the “case” for each one, which they convert into an amount of dopamine. They send this dopamine to a part of the brain called the pallium, and then the pallium executes whichever plan has the most dopamine associated with it.

Suppose I have shingles. I’m giving a speech to a group of distinguished people whom I desperately want to impress. Then I get a very strong itch. Part of my brain calculates the expected value of continuing to speak in a dignified way, and converts that into dopamine. Another part calculates the importance of scratching myself vigorously, and converts that into dopamine. The pallium compares these two amounts of dopamine, one is larger than the other, and the decision gets made. If the itch is bad enough, and if whatever lizard-brain nucleus makes me want to scratch itches has enough dopamine to spare, then I never had a chance.

“But,” Bryan objects, “if I put a gun to your head, and threatened to shoot you if you scratched the itch, you wouldn’t do it, would you?”

In that case, a part of my brain calculates the expected value of continuing to speak in a dignified way plus not getting shot. This is a very high expected value! It sends lots and lots of dopamine to my pallium. The part of my brain calculating the expected value of scratching the itch and getting shot calculates this as a very low-expected-value course, and sends some a very low (maybe negative?) signal. The pallium decisively selects the plan to keep speaking and not get shot.

To summarize: the brain compares the strength of various preferences and executes the strongest. Anything that strengthens your urges at the expense of your goals makes you more likely to do things you don’t endorse, and makes you worse off. In a counterfactual world where a threatened gunshot is also weighing down the scale, maybe the calculus would come out different. But in the non-counterfactual world where there is no gunshot, the calculus comes out the way it does.

(also, if Bryan uses his gunshot analogy one more time, I am going to tell him about all of the mentally ill people I know about who did, in fact, non-metaphorically, non-hypothetically, choose a gunshot to the head over continuing to do the things their illness made it hard for them to do. Are you sure this is the easily-falsified hill you want to die on?)

This model doesn’t use the word or the concept of “choice” anywhere. There are various algorithms mechanically evaluating the expected reward of different actions, and a more central algorithm comparing all of those evaluations. Those algorithms could have resolved differently in different situations, and you can be uncertain how they will resolve in the same situation, but there’s no point at which they actually could resolve differently in the same situation. If this makes you want to start debating free will – in either direction – I cannot recommend this Less Wrong post highly enough.

A few examples to hammer this in:

1. Most weekends, Alice stays in and reads a book (preference strength 20). But today is her firstborn child’s wedding, which she has been looking forward to for years (preference strength 100). Just before she leaves for the chapel, she gets a terrible migraine, and she feels like it would be unbearable to go out of her room (preference strength 200). Since 200 is greater than 100, Alice misses the wedding and feel miserable, since she would have meta-preferred to go to the wedding. If you had threatened to shoot her unless she went to the wedding, she would have gone to the wedding and been miserable the whole time, because she is terrified of death (preference strength 9999) and 9999 is greater than 200.

2. Bryan is a responsible member of society and wants to work hard and take care of his family (preference strength 100). He drinks some alcohol, but because he has no genetic or environmental risk factors for alcoholism, it doesn’t make him feel any urge to drink himself to death (preference strength 0), so he doesn’t. If we CRISPRed him to give him every single alcoholism risk gene plus crippling anxiety, then drinking the alcohol would make him feel a very strong urge to drink himself to death (preference strength 200), and he would drink himself to death instead of caring for his family.

3. CRISPRed alcoholic Bryan goes to an addiction doctor. The doctor advises him to take the anti-alcoholism drug naltrexone (-20 preference strength for alcoholism). Then the doctor advises him to go to Alcoholics Anonymous and get a whole new friend group in which his status depends entirely on his ability to remain sober (+20 for staying sober). Now his preferences are “stay sober and take care of my family” (strength 120) vs. “drink myself to death (strength 180), but the preference to drink is still stronger, so he does.

4. Bryan goes to a therapist who asks him to visualize the things he loves about his family and why he thinks it’s important to take care of them, which makes this more vivid in his mind (preference +10 for sobriety). Bryan’s boss threatens to fire him if he misses one more day of work because of drunkenness (preference +20 for sobriety). Now he’s at 150 for sobriety vs. 180 for drinking. He gives $20,000 to Beeminder, which they will only give him back if he stays sober for the next year (+20 for sobriety), and he reads George Ainslie’s Picoeconomics which describes ways to reconceptualize choices across time to better account for all of their implications (+20 for sobriety). Now he’s at 190 for sobriety vs. 180 for drinking, so he stays sober.

5. A few months later, Bryan’s friend dies in an accident. He feels angry, depressed, and anxious. This makes alcohol seems more attractive, since it would temporarily help him forget these feelings (+20 for drinking). At the same time, he stops going to AA because it’s annoying and far away (-20 for staying sober). Now he’s at 170 for sobriety vs. 200 for drinking, so he falls off the wagon.

I’m not claiming this lamprey model is exactly literally true for humans. And I’m not claiming there’s a perfect binary distinction between endorsed goals and unendorsed urges. This model is full of complications and gray areas. I’m just saying it’s a better model, with fewer gray areas, than trying to separate everything into just “preference” or “constraint”, and shooting yourself in the foot again and again like some kind of tipped-over Gatling gun.

And it goes a lot of the way to modeling mental illness: the mentally ill have conditions that give them strong unendorsed urges. For any given strength of goal, having strong urges will make people less able to pursue that goal, in favor of pursuing the urges instead, and that will make them worse off, for a definition of “well off” that involves being happy and achieving goals. These people very reasonably want to stop having these weird urges so they can pursue their goals in peace.

Bryan will correctly point out that there are awkward implications in identifying “unexpected generator of strong unendorsed urges” with “disease”. For example, gay people in a traditional religious community will have strong urges to have homosexual relationships, and they won’t endorse those urges – they would probably rather be straight instead.

Or: obese people feel an urge to eat which they don’t endorse. Should we call obesity a disease, and describe them as having a disease which produces urges contrary to their preferences? Some people say yes (and keep in mind that both genetics and viral infections can induce obesity). But suppose some normal-weight person would rather be supermodel-thin, and their perfectly normal urge to eat a normal amount prevents them from looking like a broomstick. Is their normal level of hunger a disease? A naive equation of “biological generator of unendorsed urges” and “disease” would say yes!

We want some criteria that let us call shingles a disease, but don’t let us call “being thin but wanting to be even thinner” a disease. Unfortunately, there is no perfect solution to this problem. People have wanted perfect solutions to definitional questions ever since Plato defined man as “a featherless biped”, and it’s never worked. Luckily, there are kludgy, good-enough solutions, which I describe in Dissolving Questions About Disease, the fourth most popular Less Wrong post of all time. If you still think this is confusing, please read it. If it’s still confusing even after that, try The Categories Were Made For Man, Not Man For The Categories.

I think Bryan should be happy with this solution. It’s very libertarian. It says that it’s up to every individual to decide how to satisfy their own preferences (including meta-preferences). If your problem is constraints (you want to go to Hawaii, but you don’t have enough money), you can work to resolve those constraints (eg go to work and earn more money). If your problem is urges (you want to go to Hawaii, but you’re too anxious to leave your room), you can work to resolve those urges (eg go to a psychiatrist and get medication). The job of a good liberal society is to support people in achieving their own goals as they understand them, and this includes supporting their decision to get the job they want and their decision to get the psychiatric treatment they want.

As I write this essay, I’m a little bit caffeinated. I looked at my preference set – which included an urge to get back in bed instead of writing blog posts – decided it didn’t achieve my goals, and took a psychotropic drug to shift my preference set to one I liked better. And if we’re willing to accept this in relatively trivial cases, the argument for accepting it is even stronger for people whose preference sets have been deranged by obvious bizarre causes – infections, hormone imbalances, brain injuries, addictive substances, genetic defects – and for people whose irresistible urges are ruining their lives in preventable ways.

# Book Review Review: Little Soldiers

Little Soldiers is a book by Lenora Chu about the Chinese education system. I haven’t read it. This is a review of Dormin111’s review of Little Soldiers.

Dormin describes the “plot”: The author is a second-generation Chinese-American woman, raised by demanding Asian parents. Her parents made her work herself to the bone to get perfect grades in school, practice piano, get into Ivy League schools, etc. She resisted and resented the hell she was forced to go through (though she got into Stanford, so she couldn’t have resisted too hard).

Skip a decade. She is grown up, married, and has a three year old child. Her husband (a white guy named Rob) gets a job in China, so they move to Shanghai. She wants their three-year-old son to be bilingual/bicultural, so she enrolls him in Soong Qing Ling, the Harvard of Chinese preschools. The book is about her experiences there and what it taught her about various aspects of Chinese education. Like the lunches:

During his first week at Soong Qing Ling, Rainey began complaining to his mom about eating eggs. This puzzled Lenora because as far as she knew, Rainey refused to eat eggs and never did so at home. But somehow he was eating them at school.

After much coaxing (three-year-olds aren’t especially articulate), Lenora discovered that Rainey was being force-fed eggs. By his telling, every day at school, Rainey’s teacher would pass hardboiled eggs to all students and order them to eat. When Rainey refused (as he always did), the teacher would grab the egg and shove it in his mouth. When Rainey spit the egg out (as he always did), the teacher would do the same thing. This cycle would repeat 3-5 times with louder yelling from the teacher each time until Rainey surrendered and ate the egg.

Outraged, Lenora stormed to the school the next day and approached the teacher in the morning as she dropped Rainey off. Lenora demanded to know if Rainey was telling the truth – was this teacher literally forcing food into her three-year-old son’s mouth and verbally berating him until he ate it. The teacher didn’t even bother looking at Lenora as she calmly explained that eggs are healthy and that it was important for children to eat them. When Lenora demanded she stop force-feeding her son, the teacher refused and walked away.

Or the seating:

As Lenora hears more crazy stories from her son and friends, she keeps coming back to one question: “what does Rainey actually do in school?” Lenora tries to ask Rainey, but he always replies, “we sit still.” He also occasionally mentions painting and eating, but that’s it.

So Lenora goes to Rainey’s teacher one day and asks to sit in on classes to observe. Lenora is told that this is not possible. So she asks if she can know a little more about what the school is teaching Rainey. The teacher tells her that she is already told everything she needs to know, and that this is the “Chinese way.”

Since Lenora couldn’t get a look into Soong Qing Ling, she went to another local school and bribed her way into a classroom-observation post with some well-placed handbags. She discovered that Rainey was basically right. Chinese preschool really does seem to consist of sitting still. Unless given different orders, all students were required to sit in their seats with their arms at their sides, and their feet flat on a line of tape on the ground. This is not an easy task for three-year-olds.

There were two teachers in the classroom with a classic good cop/bad cop dynamic. The good cop stood in the front of the room with the desks splayed out before her. She would give simple instructions like orders to get food, water, or sometimes paint, though usually she said nothing at all. The bad cop was another teacher who prowled the classroom. Any time she saw a student remove a foot from the line, move arms from his side, or otherwise deviate from the instructions, she would yell at the student to fall back in line. Lenora spent about a week watching tiny kids get screamed at for trying to get water, shifting in their chairs, or talking to classmates.

Or art class:

When Lenora sat in on a kindergarten class, she witnessed an art lesson where the students were taught how to draw rain. The nice teacher drew raindrops on a whiteboard, showing precisely where to start and end each stroke to form a tear-drop shape. When it was the students’ turns, they had to perfectly replicate her raindrop. Over and over again. Same start and end points. Same curves. For an hour. No student could draw anything else. Any student who did anything different would be yelled at and told to start over.

The point of this exercise was not to teach students how to draw raindrops. Drawing raindrops is not an important life skill, and drawing them in a particular way is especially not important. Even the three-year-old students in the class seemed to realize this as many immediately created their own custom raindrop shapes and drew landscapes, all to be crushed under the mean teacher’s admonishment. The real point of the exercise was to teach students to follow directions from an authority figure. But more than that, the point was to follow pointless and arbitrary directions. The more pointless and arbitrary the directions are, the more willpower is required to follow them.

Chinese people presumably put up with this because it makes sense within their culture; why did Chu put up with it? Dormin half-jokingly suggests maybe she really wanted to write the book she eventually wrote, and this was her research. But Chu herself says it eventually got results:

After spending 75% of the book relentlessly complaining about her son’s Chinese education, with the occasional anecdote about how horrible her own culturally Chinese upbringing was, Lenora decides Chinese schools aren’t so bad.

After a few years in China, Rainey changed. Though Lenora constantly worried if Rainey’s creativity and leadership potential was being snuffed out, she couldn’t help but be impressed by his emerging self-control. He could sit still for longer. He always greeted people politely. He finished eating his food. He asked permission a lot.

Lenora didn’t realize what Rainey had become until she took him back to the US for a few weeks to visit family. There, the contrast between Rainey and his same-aged American counterparts become stark. Lenora’s friends’ kids ate junk food all day while Rainey asked for vegetables. They couldn’t read or do basic addition while Rainey was close to being bilingual and had started double-digit addition and subtraction by first grade. They wandered obliviously in their own worlds while Rainey’s Chinese grandparents were thrilled to receive respectful greetings every time Rainey entered the room […]

What really sold Lenora on Chinese education was that it apparently worked. At the time of writing the book, Shanghai was scoring first place in the world on the PISA exams, beating heavy-hitters like Norway and Singapore. Supposedly, education scholars and professionals all over the world were looking at China for wisdom. They all saw the bad, but they saw a lot of good too.

(before going forward, I should interject that China’s great PISA scores are kind of fake. China struck a deal with the OECD (the group that administers PISA) to let it conduct testing only in its four richest and best-educated provinces. Rich and well-educated places always do well on PISA. That China’s four best provinces outperform the average score of other countries is unsurprising. This article points out that if the US were allowed to enter only its best-educated state (Massachussetts, obviously) we would be right up there with China. So this probably isn’t as impressive as Ms. Chu thinks.)

This is just a sample of the great stuff in Dormin’s review of Little Soldiers, and I strongly recommend you read the whole thing. You should also read the comments, which point out that this may be more about a few elite Chinese schools than about an entire country. But I want to use these excerpts as a jumping-off point to talk about the US education system, unschooling, and child development in general.

I predict most of my Bay Area friends would hate the Chinese education system as Chu describes it. I predict this because they already hate the US education system, which is only like 10% as bad. I’m especially thinking of @webdevmason and @michaelblume, who often write about the ways American education is frustrating, regressive, and authoritarian. Bright-eyed, curious kids come in. They spend thirteenish years getting told to show their work, being punished for reading ahead in the textbook, and otherwise having their innate love of learning drummed out of them in favor of endless mass-produced homework assignments (five pages, single-spaced, make sure you use the right number of topic sentences).



People with this position usually make two claims. One, US public school as it currently exists is awful, basically institutionalized child abuse. Two, this is bad for the economy. I’ve been through too much school myself to feel like challenging the first, so I want to focus on the second.

Salman Khan, John Gatto, and other education rebels trace the current school systems back to the Prussians, who invented compulsory education to prepare children for a career as infantrymen or factory workers. It’s a great story. Like most great stories, it’s kind of false. But like most kind-of-false things that catch on, it has an element of truth. Children who can sit still in a classroom and do what their teachers say are well-placed to become adults who can sit still in an open office and do what their bosses say. So (according to this logic), even if our schools are awful, they were well-suited to the Industrial Age economy. Some hypothetical mash-up of Otto von Bismarck and Voldemort, who wanted the country to produce as much as possible and didn’t care how many children’s souls were crushed in the process, might at least endorse the education system on widget-maximization grounds.

But (these same people argue), the Industrial Age is over. The most important skills now are entrepreneurship and creative problem solving. Reinventing yourself, selling yourself, carving out a new niche for yourself. Figuring out what’s going to be the next big thing and pursuing it without anyone else watching over you. We’re in XKCD’s world now, where 900 hours of classes and 400 hours of homework matter less to your career success than one weekend messing around with a programming language in 11th grade. The Prussian model of education stamps out the kind of independent agency that could help people navigate the weird, formless 21st century world.

How might the personified Chinese education system respond?

What if it said “I don’t know what you 老外 are doing in America, but I’m not crushing anybody. I’m just telling kids to sit here drawing 1,000 raindrops in a row without moving or protesting. If after that you decide you don’t want to found the next Uber, that’s on you. But if you do decide to found the next Uber, I will have taught you the most important skill: discpline. Learning how to sit still and obey others is the necessary prerequisite to learning how to sit still and obey yourself.”

If it was really mean, it might go further. “I notice most of you Americans suck at this skill. I notice you’re always whining about how you don’t have enough discipline to pursue your interests. Some of you are writers who spend years fantasizing about the novel you’re going to publish, but can never quite bring yourself to put pen to paper. Others want to learn another language, but reject real work in favor of phone apps that promise to ‘gamify’ staying at a 101 level for the rest of your life. You don’t need to feel bad about having no self-control; after all, nobody taught you any. If you’d gone to 宋庆龄幼儿园, you would have spent your formative years learning to sit still and focus, having your natural impulse to slack off squeezed out of you. Then you could have pushed through and written your novel, or learned 官話, or if you wanted to start Uber you could start Uber. At the very least you’d be doing something other than lying in bed browsing Reddit posts about how adulting is hard.”

My Bay Area friends treat people as naturally motivated, and assume that if someone acts unmotivated, it’s because they’ve spent so long being taught to suppress their own desires that they’ve lost touch with innate enthusiasm. Personified China treats people as naturally unmotivated, and assumes that if someone acts unmotivated, it’s because they haven’t been trained to pursue a goal determinedly without getting blown around by every passing whim.

What evidence is there in favor of one education system or the other?

I can’t find any good studies directly supporting or opposing either of these claims. The best I can do is The Development Of Executive Functioning And Theory Of Mind: A Comparison Of Chinese And US Preschoolers. They find that on various tests of executive function, “Chinese [preschool-age] children’s performance was consistently on par with that of US children who were on average 6 months older” (other sources say 1-2 years). But lots of interventions change things in childhood; this isn’t interesting unless it persists into adulthood, and I don’t see any work on this. This study on racial differences in personality traits found weak and inconsistent white-Asian differences on adult conscientiousness, but the Asian sample was Asian-American and differences in education were probably pretty minor.

What about circumstantial evidence?

First and most important, since extreme cultivation of discipline vs. laissez-faire childrearing is a property of parents as much as schools, any claimed effect would run afoul of all the twin studies showing that shared environment has few long-term effects on any trait. For example, this meta-analysis of factors affecting self-control that finds “no or very little influence of the shared environment on the variance in self-control”. But we can always invoke the usual loophole in shared environment findings: maybe the US doesn’t contain anything as extreme as the Chinese education system, so US-only studies can’t capture its effects.

Second, both Westerners and Chinese seem to include some very impressive and some less impressive people. It certainly doesn’t seem wrong to say that Chinese people seem more diligent and Westerners seem more independent, but there are so many potential biases at work that I would hate to take this too seriously as evidence for or against one form of education. Also, Chinese-Americans who are educated in US schools also seem more diligent than white Americans, so maybe the education system doesn’t contribute too much to this. Maybe Chinese culture promotes diligence better in general, this causes diligence-focused school systems, but the diligence-focused school systems don’t themselves cause the diligence.

Third, we could try to find more extreme versions on both sides and see what happens there. Pre-industrial populations with no education were famously bad at the discipline needed for factory work. From Pseudoerasmus:

The earliest factory workers were lacking in what Mokyr & Voth call “discipline capital” — non-cognitive ‘skills’ like punctuality, sobriety, reliability, docility, and pliability. Whether they had been peasants or artisans, early workers were new to industrial work habits and they had a strong preference for autonomous work arrangements. They were accustomed to setting their own pace of work in farming, domestic outwork, or artisanal workshops, and disliked the time rules and strict supervision of the factories.

All this is consistent with colourful descriptions of the early history of the textile industry in the Global South, including Japan. Mills were described as places of chaos and disorder. They were supposedly filled with workers ‘idling’, ‘loitering’, ‘socialising’, smoking, tea-drinking, or just disappeared for the day. In Japan, “twenty percent of the female operatives…absent themselves after they receive their monthly pay check” (Saxonhouse & Kiyokawa 1978). In Shanghai, it was said female mill workers could be found breast-feeding infants during work hours (Cochran 2000). Or at Mumbai mills, workers “bathed, washed clothes, ate his meals, and took naps” (Gupta 2011).

But this could be as much about expectations as about abilities.

Which historical culture had the most authoritarian-instillment-of-virtue-focused approach to child-rearing? Surely the New England Puritans were up there – remember that eg Puritan parents would traditionally send children away to be raised by other families, in the hopes that the lack of familiarity would make the child behave better”. They certainly ended out industrious. But they were also creative and self-motivated, sometimes almost hilariously so. On the other hand, I’m not sure that the Puritans who ended up incredibly creative were exactly the same Puritans who suffered extreme strict child-rearing – there seems about a century gulf between the evidence of authoritarian parenting in the 1600s and the crop of geniuses born in the late 1700s – so I’m not sure how seriously to take this.

Fourth, we could look at US trends over time. Both US parenting and US schooling seem to be getting less authoritarian over time; 31 states have banned corporal punishment since 1970, and the teachers I know confirm a shift away from most forms of discipline. Over the same time period, children have gotten weirdly better behaved – less crime, less teenage pregnancy, more willing to jump through various stupid hoops to get into a good college. This seems to contradict the Chinese theory – the children are no worse at controlling their impulses. But there are other findings that contradict the Bay Area theory – entrepreneurship is decreasing; more top students are choosing to go work for a boss at a big bank rather than go do something weird. I think the better behavior is probably just caused by lower lead; I have no idea why people are more risk-averse. Secular decline in testosterone, maybe?

Fifth, we could look at research on the effects of preschool more generally. Some studies find that US preschools do not make children smarter, but still improve life outcomes like graduation rates, crime rates, and employment. Although there are lots of theories about the “noncognitive skills” that accomplish this (including that they don’t exist and the improvement is an artifact of bad experimental technique), this is certainly consistent with preschool teaching children discipline at a critical window. If this hypothesis were true, the effect of preschool would be much larger in China, but I don’t know of any Chinese studies on the topic.

Sixth, we could look at the research on meditation for very young kids. The Chinese theory casts preschool as a sort of dark-side form of mindfulness. In traditional Buddhist settings, monks would sit perfectly still and concentrate on the most boring thing imaginable, and the head monk would slap them with a bamboo stick if they moved. The resemblance to the school system is uncanny. So maybe school’s effects on self-control could be modeled as a sort of less-intense but much-more-drawn-out meditation session. Unfortunately, the studies surrounding mindfulness in kids are crap, so this doesn’t help either.

Really none of this seems very helpful and we’re kind of left with our priors. And maybe one of our priors is “don’t abuse children”, so there’s that.

But what about the Polgars? They turned all three of their children into chess prodigies through a strategy that seemed based around exposing them to absurd amounts of chess at a very young age. If we generalize, it does look like very young children might have very plastic minds that you can shape through out-of-distribution experiences. But Lazslo Polgar insisted that his technique didn’t use force; the point was to interest his children in the material so avidly that they inflicted near-Chinese levels of intensity on themselves in order to study it more successfully.

One problem with the physical universe is that even after you study a question in depth and decide more evidence is needed, there are still real children you have to educate one way or the other. I have no general solution for this, but the Polgar strategy seems like a good deal if you can pull it off.

# Assortative Mating And Autism

Introduction

Assortative mating is when similar people marry and have children. Some people worry about assortative mating in Silicon Valley: highly analytical tech workers marry other highly analytical tech workers. If highly analytical tech workers have more autism risk genes than the general population, assortative mating could put their children at very high risk of autism. How concerned should this make us?

Methods / Sample Characteristics

I used the 2020 Slate Star Codex survey to investigate this question. It had 8,043 respondents selected for being interested in a highly analytical blog about topics like science and economics. The blog is associated with – and draws many of its readers from – the rationalist and effective altruist movements, both highly analytical. More than half of respondents worked in programming, engineering, math, or physics. 79% described themselves as atheist or agnostic. 65% described themselves as more interested in STEM than the humanities; only 15% said the opposite.

According to Kogan et al (2018), about 2.5% of US children are currently diagnosed with autism spectrum disorders. The difference between “autism” and “autism spectrum disorder” is complicated, shifts frequently, and is not very well-known to the public; this piece will treat them interchangeably from here on. There are no surveys of what percent of adults are diagnosed with autism; it is probably lower since most diagnoses happen during childhood and the condition was less appreciated in past decades. These numbers may be affected by parents’ education level and social class; one study shows that children in wealthy neighborhoods were up to twice as likely to get diagnosed as poorer children.

Given that respondents are likely wealthier than average, we might expect a rate of 2.5% – 5%. Instead the rate is noticeably higher than that, consistent with the hypothesis that this sample will be more autistic than average. About 4% of the SSC survey sample had a formal diagnosis of autism, but this rose to 6% when the sample was limited to people below 30, and to 8% below 20. This sample is plausibly about 2-3x more autistic than the US population. Childhood social class was not found to have a significant effect on autism status in this sample.

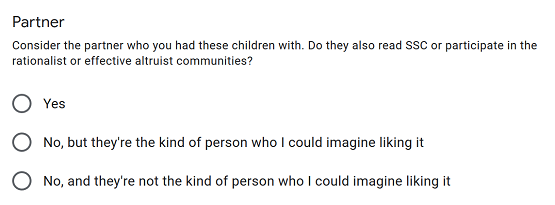
Results

I tried to get information on how many children respondents have, but I forgot to ask important questions about age until a quarter of the way through the survey. I want to make sure I’m only catching children old enough that their autism would have been diagnosed, so the information below (except when otherwise noted) comes from the three-quarters of the sample where I have good age information. I also checked it against the whole sample and it didn’t make a difference.

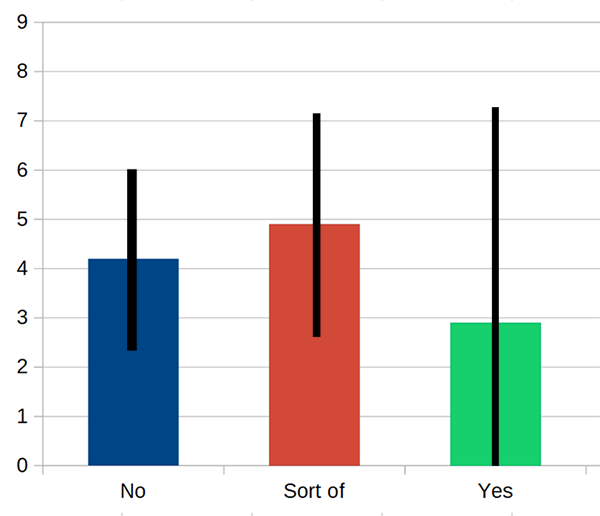
Of this limited sample, 1,204 individual parents had a total of 2,459 children. 1,892 of those children were older than 3, and 1,604 were older than 5. I chose to analyze children older than 3, since autism generally becomes detectable around 2.

71 children in the 1,892 child sample had formal diagnoses of autism, for a total prevalence of 3.7%. When parents were asked to include children who were not formally diagnosed but who they thought had the condition, this increased to 99 children, or a 5.2% prevalence. Both numbers are much lower than the 8% prevalence in young people in the sample.

What about marriages where both partners were highly analytical? My proxy for this was the following survey question:

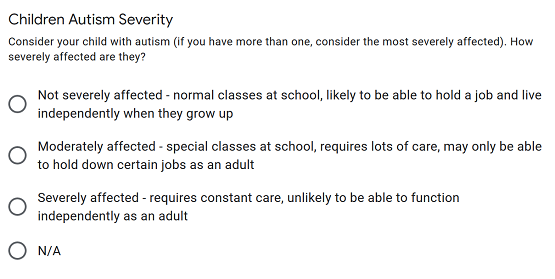


I’ll be referring to these answers as “yes”, “sort of”, and “no” from here on, and moving back to the full sample. 938 parents answered this question; 51 (5.4%) yes, 233 (24.8%) sort of, and 653 (69.4%) no. Keep in mind the effective sample is even smaller, since both partners in two-partners-read-SSC-families may have filled out the survey individually about the same set of children (though this should not have affected the “sort of” group). Here is the autism rate for each group, with 95% confidence interval in black:



There is little difference. If we combine the latter two groups, the confidence interval narrows slightly, to 2.7 – 6.5.

I asked respondents about the severity of their children’s autism.



People who hadn’t previously reported any children with autism gave answers other than N/A for this one, which was confusing. Instead of the 71 children we had before, now it’s up to 144 children. I’m not sure what’s going on here. Of these phantom children, 101 had mild cases, 31 moderate, and only 12 severe. Severe autism was only present in 0.6% of the children in the sample. There was no tendency for couples where both partners were highly analytical to have children with more severe autism.

Discussion

Autism rates in this survey were generally low. Although the general rate of 3.7% was higher than the commonly-estimated US base rate of 2.5%, this is consistent with the slight elevation of autism observed in higher social classes.

There was no sign of elevated risk when both partners were highly analytical. The sample size was too small to say for certain that no such elevation exists, but it can say with 95% confidence that the elevated risk is less than three percentage points.

This suggests that the answer to the original question – does assortative mating between highly analytical people significantly increase chance of autism in offspring – is at least a qualified “no”.

Why should this be? It could just be that regression to the mean is more important in this case than any negative effects from combining recessive genes or mixing too many risk genes together. Or maybe we should challenge the assumption that being a highly analytical programmer is necessarily on a continuum with autism. It seems like p(highly analytical|on autism spectrum) is pretty high, but p(on autism spectrum|is highly analytical) might be much lower.

Obvious limitations of this survey include the small sample size of both-partners-highly-analytical couples, the weak operationalization of highly analytical as “member of the SSC, rationalist, and effective altruist communities”, and the inability to separate non-autistic children from children who are not yet diagnosed. Due to these limitations, this should only be viewed as providing evidence against the strongest versions of the assortative mating hypothesis, where it might increase risk by double, triple, or more. Smaller elevations of risk remain plausible and would require larger studies to assess.

I welcome people trying to replicate or expand on these results. All of the data used in this post are freely available and can be downloaded here.

# Book Review: Human Compatible

I.

Clarke’s First Law goes: When a distinguished but elderly scientist states that something is possible, he is almost certainly right. When he states that something is impossible, he is very probably wrong.

Stuart Russell is only 58. But what he lacks in age, he makes up in distinction: he’s a computer science professor at Berkeley, neurosurgery professor at UCSF, DARPA advisor, and author of the leading textbook on AI. His new book Human Compatible states that superintelligent AI is possible; Clarke would recommend we listen.

I’m only half-joking: in addition to its contents, Human Compatible is important as an artifact, a crystallized proof that top scientists now think AI safety is worth writing books about. Nick Bostrom’s Superintelligence: Paths, Dangers, Strategies previously filled this role. But Superintelligence was in 2014, and by a philosophy professor. From the artifactual point of view, HC is just better – more recent, and by a more domain-relevant expert. But if you also open up the books to see what’s inside, the two defy easy comparison.

S:PDS was unabashedly a weird book. It explored various outrageous scenarios (what if the AI destroyed humanity to prevent us from turning it off? what if it put us all in cryostasis so it didn’t count as destroying us? what if it converted the entire Earth into computronium?) with no excuse beyond that, outrageous or not, they might come true. Bostrom was going out on a very shaky limb to broadcast a crazy-sounding warning about what might be the most important problem humanity has ever faced, and the book made this absolutely clear.

HC somehow makes risk from superintelligence not sound weird. I can imagine my mother reading this book, nodding along, feeling better educated at the end of it, agreeing with most of what it says (it’s by a famous professor! I’m sure he knows his stuff!) and never having a moment where she sits bolt upright and goes what? It’s just a bizarrely normal, respectable book. It’s not that it’s dry and technical – HC is much more accessible than S:PDS, with funny anecdotes from Russell’s life, cute vignettes about hypothetical robots, and the occasional dad joke. It’s not hiding any of the weird superintelligence parts. Rereading it carefully, they’re all in there – when I leaf through it for examples, I come across a quote from Moravec about how “the immensities of cyberspace will be teeming with unhuman superminds, engaged in affairs that are to human concerns as ours are to those of bacteria”. But somehow it all sounds normal. If aliens landed on the White House lawn tomorrow, I believe Stuart Russell could report on it in a way that had people agreeing it was an interesting story, then turning to the sports page. As such, it fulfills its artifact role with flying colors.

How does it manage this? Although it mentions the weird scenarios, it doesn’t dwell on them. Instead, it focuses on the present and the plausible near-future, uses those to build up concepts like “AI is important” and “poorly aligned AI could be dangerous”. Then it addresses those abstractly, sallying into the far future only when absolutely necessary. Russell goes over all the recent debates in AI – Facebook, algorithmic bias, self-driving cars. Then he shows how these are caused by systems doing what we tell them to do (ie optimizing for one easily-described quantity) rather than what we really want them to do (capture the full range of human values). Then he talks about how future superintelligent systems will have the same problem.

His usual go-to for a superintelligent system is Robbie the Robot, a sort of Jetsons-esque butler for his master Harriet the Human. The two of them have all sorts of interesting adventures together where Harriet asks Robbie for something and Robbie uses better or worse algorithms to interpret her request. Usually these requests are things like shopping for food or booking appointments. It all feels very Jetsons-esque. There’s no mention of the word “singleton” in the book’s index (not that I’m complaining – in the missing spot between simulated evolution of programs, 171 and slaughterbot, 111, you instead find Slate Star Codex blog, 146, 169-70). But even from this limited framework, he manages to explore some of the same extreme questions Bostrom does, and present some of the answers he’s spent the last few years coming up with.

If you’ve been paying attention, much of the book will be retreading old material. There’s a history of AI, an attempt to define intelligence, an exploration of morality from the perspective of someone trying to make AIs have it, some introductions to the idea of superintelligence and “intelligence explosions”. But I want to focus on three chapters: the debate on AI risk, the explanation of Russell’s own research program, and the section on misuse of existing AI.

II.

Chapter 6, “The Not-So-Great Debate”, is the highlight of the book-as-artifact. Russell gets on his cathedra as top AI scientist, surveys the world of other top AI scientists saying AI safety isn’t worth worrying about yet, and pronounces them super wrong:

I don’t mean to suggest that there cannot be any reasonable objections to the view that poorly designed superintelligent machines would present a serious risk to humanity. It’s just that I have yet to see such an objection.

He doesn’t pull punches here, collecting a group of what he considers the stupidest arguments into a section called “Instantly Regrettable Remarks”, with the connotation that the their authors (“all of whom are well-known AI researchers”), should have been embarrassed to have been seen with such bad points. Others get their own sections, slightly less aggressively titled, but it doesn’t seem like he’s exactly oozing respect for those either. For example:

Kevin Kelly, founding editor of Wired magazine and a remarkably perceptive technology commentator, takes this argument one step further. In “The Myth of a Superhuman AI,” he writes, “Intelligence is not a single dimension, so ‘smarter than humans’ is a meaningless concept.” In a single stroke, all concerns about superintelligence are wiped away.

Now, one obvious response is that a machine could exceed human capabilities in all relevant dimensions of intelligence. In that case, even by Kelly’s strict standards, the machine would be smarter than a human. But this rather strong assumption is not necessary to refute Kelly’s argument.

Consider the chimpanzee. Chimpanzees probably have better short-term memory than humans, even on human-oriented tasks such as recalling sequences of digits. Short-term memory is an important dimension of intelligence. By Kelly’s argument, then, humans are not smarter than chimpanzees; indeed, he would claim that “smarter than a chimpanzee” is a meaningless concept.

This is cold comfort to the chimpanzees and other species that survive only because we deign to allow it, and to all those species that we have already wiped out. It’s also cold comfort to humans who might be worried about being wiped out by machines.

Or:

The risks of superintelligence can also be dismissed by arguing that superintelligence cannot be achieved. These claims are not new, but it is surprising now to see AI researchers themselves claiming that such AI is impossible. For example, a major report from the AI100 organization, Artificial Intelligence and Life in 2030, includes the following claim: “Unlike in the movies, there is no race of superhuman robots on the horizon or probably even possible.”

To my knowledge, this is the first time that serious AI researchers have publicly espoused the view that human-level or superhuman AI is impossible—and this in the middle of a period of extremely rapid progress in AI research, when barrier after barrier is being breached. It’s as if a group of leading cancer biologists announced that they had been fooling us all along: They’ve always known that there will never be a cure for cancer.

What could have motivated such a volte-face? The report provides no arguments or evidence whatever. (Indeed, what evidence could there be that no physically possible arrangement of atoms outperforms the human brain?) I suspect that the main reason is tribalism — the instinct to circle the wagons against what are perceived to be “attacks” on AI. It seems odd, however, to perceive the claim that superintelligent AI is possible as an attack on AI, and even odder to defend AI by saying that AI will never succeed in its goals. We cannot insure against future catastrophe simply by betting against human ingenuity.

If superhuman AI is not strictly impossible, perhaps it’s too far off to worry about? This is the gist of Andrew Ng’s assertion that it’s like worrying about “overpopulation on the planet Mars.” Unfortunately, a long-term risk can still be cause for immediate concern. The right time to worry about a potentially serious problem for humanity depends not just on when the problem will occur but also on how long it will take to prepare and implement a solution. For example, if we were to detect a large asteroid on course to collide with Earth in 2069, would we wait until 2068 to start working on a solution? Far from it! There would be a worldwide emergency project to develop the means to counter the threat, because we can’t say in advance how much time is needed.

Russell displays master-level competence at the proving too much technique, neatly dispatching sophisticated arguments with a well-placed metaphor. Some expert claims it’s meaningless to say one thing is smarter than another thing, and Russell notes that for all practical purposes it’s meaningful to say humans are smarter than chimps. Some other expert says nobody can control research anyway, and Russell brings up various obvious examples of people controlling research, like the ethical agreements already in place on the use of gene editing.

I’m a big fan of Luke Muehlhauser’s definition of common sense – making sure your thoughts about hard problems make use of the good intuitions you have built for thinking about easy problems. His example was people who would correctly say “I see no evidence for the Loch Ness monster, so I don’t believe it” but then screw up and say “You can’t disprove the existence of God, so you have to believe in Him”. Just use the same kind of logic for the God question you use for every other question, and you’ll be fine! Russell does great work applying common sense to the AI debate, reminding us that if we stop trying to out-sophist ourselves into coming up with incredibly clever reasons why this thing cannot possibly happen, we will be left with the common-sense proposition that it might.

My only complaint about this section of the book – the one thing that would have added a cherry to the slightly troll-ish cake – is that it missed a chance to include a reference to On The Impossibility Of Supersized Machines.

Is Russell (or am I) going too far here? I don’t think so. Russell is arguing for a much weaker proposition than the ones Bostrom focuses on. He’s not assuming super-fast takeoffs, or nanobot swarms, or anything like that. All he’s trying to do is argue that if technology keeps advancing, then at some point AIs will become smarter than humans and maybe we should worry about this. You’ve really got to bend over backwards to find counterarguments to this, those counterarguments tend to sound like “but maybe there’s no such thing as intelligence so this claim is meaningless”, and I think Russell treats these with the contempt they deserve.

He is more understanding of – but equally good at dispatching – arguments for why the problem will really be easy. Can’t We Just Switch It Off? No; if an AI is truly malicious, it will try to hide its malice and prevent you from disabling it. Can’t We Just Put It In A Box? No, if it were smart enough it could probably find ways to affect the world anyway (this answer was good as far as it goes, but I think Russell’s threat model also allows a better one: he imagines thousands of AIs being used by pretty much everybody to do everything, from self-driving cars to curating social media, and keeping them all in boxes is no more plausible than keeping transportation or electricity in a box). Can’t We Just Merge With The Machines? Sounds hard. Russell does a good job with this section as well, and I think a hefty dose of common sense helps here too.

He concludes with a quote:

The “skeptic” position seems to be that, although we should probably get a couple of bright people to start working on preliminary aspects of the problem, we shouldn’t panic or start trying to ban AI research. The “believers”, meanwhile, insist that although we shouldn’t panic or start trying to ban AI research, we should probably get a couple of bright people to start working on preliminary aspects of the problem.

I couldn’t have put it better myself.

III.

If it’s important to control AI, and easy solutions like “put it in a box” aren’t going to work, what do you do?

Chapters 7 and 8, “AI: A Different Approach” and “Provably Beneficial AI” will be the most exciting for people who read Bostrom but haven’t been paying attention since. Bostrom ends by saying we need people to start working on the control problem, and explaining why this will be very hard. Russell is reporting all of the good work his lab at UC Berkeley has been doing on the control problem in the interim – and arguing that their approach, Cooperative Inverse Reinforcement Learning, succeeds at doing some of the very hard things. If you haven’t spent long nights fretting over whether this problem was possible, it’s hard to convey how encouraging and inspiring it is to see people gradually chip away at it. Just believe me when I say you may want to be really grateful for the existence of Stuart Russell and people like him.

Previous stabs at this problem foundered on inevitable problems of interpretation, scope, or altered preferences. In Yudkowsky and Bostrom’s classic “paperclip maximizer” scenario, a human orders an AI to make paperclips. If the AI becomes powerful enough, it does whatever is necessary to make as many paperclips as possible – bulldozing virgin forests to create new paperclip mines, maliciously misinterpreting “paperclip” to mean uselessly tiny paperclips so it can make more of them, even attacking people who try to change its programming or deactivate it (since deactivating it would cause fewer paperclips to exist). You can try adding epicycles in, like “make as many paperclips as possible, unless it kills someone, and also don’t prevent me from turning you off”, but a big chunk of Bostrom’s S:PDS was just example after example of why that wouldn’t work.

Russell argues you can shift the AI’s goal from “follow your master’s commands” to “use your master’s commands as evidence to try to figure out what they actually want, a mysterious true goal which you can only ever estimate with some probability”. Or as he puts it:

The problem comes from confusing two distinct things: reward signals and actual rewards. In the standard approach to reinforcement learning, these are one and the same. That seems to be a mistake. Instead, they should be treated separately…reward signals provide information about the accumulation of actual reward, which is the thing to be maximized.

So suppose I wanted an AI to make paperclips for me, and I tell it “Make paperclips!” The AI already has some basic contextual knowledge about the world that it can use to figure out what I mean, and my utterance “Make paperclips!” further narrows down its guess about what I want. If it’s not sure – if most of its probability mass is on “convert this metal rod here to paperclips” but a little bit is on “take over the entire world and convert it to paperclips”, it will ask me rather than proceed, worried that if it makes the wrong choice it will actually be moving further away from its goal (satisfying my mysterious mind-state) rather than towards it.

Or: suppose the AI starts trying to convert my dog into paperclips. I shout “No, wait, not like that!” and lunge to turn it off. The AI interprets my desperate attempt to deactivate it as further evidence about its hidden goal – apparently its current course of action is moving away from my preference rather than towards it. It doesn’t know exactly which of its actions is decreasing its utility function or why, but it knows that continuing to act must be decreasing its utility somehow – I’ve given it evidence of that. So it stays still, happy to be turned off, knowing that being turned off is serving its goal (to achieve my goals, whatever they are) better than staying on.

This also solves the wireheading problem. Suppose you have a reinforcement learner whose reward is you saying “Thank you, you successfully completed that task”. A sufficiently weak robot may have no better way of getting reward than actually performing the task for you; a stronger one will threaten you at gunpoint until you say that sentence a million times, which will provide it with much more reward much faster than taking out your trash or whatever. Russell’s shift in priorities ensures that won’t work. You can still reinforce the robot by saying “Thank you” – that will give it evidence that it succeeded at its real goal of fulfilling your mysterious preference – but the words are only a signpost to the deeper reality; making you say “thank you” again and again will no longer count as success.

All of this sounds almost trivial written out like this, but number one, everything is trivial after someone thinks about it, and number two, there turns out to be a lot of controversial math involved in making it work out (all of which I skipped over). There are also some big remaining implementation hurdles. For example, the section above describes a Bayesian process – start with a prior on what the human wants, then update. But how do you generate the prior? How complicated do you want to make things? Russell walks us through an example where a robot gets great information that a human values paperclips at 80 cents – but the real preference was valuing them at 80 cents on weekends and 12 cents on weekdays. If the robot didn’t consider that a possibility, it would never be able to get there by updating. But if it did consider every single possibility, it would never be able to learn anything beyond “this particular human values paperclips at 80 cents on 12:08 AM on January 14th when she’s standing in her bedroom.” Russell says that there is “no working example” of AIs that can solve this kind of problem, but “the general idea is encompassed within current thinking about machine learning”, which sounds half-meaningless and half-reassuring.

People with a more technical bent than I have might want to look into some deeper criticisms of CIRL, including Eliezer Yudkowsky’s article here and some discussion in the AI Alignment Newsletter.

IV.

I want to end by discussing what was probably supposed to be an irrelevant middle chapter of the book, Misuses of AI.

Russell writes:

A compassionate and jubilant use of humanity’s cosmic endowment sounds wonderful, but we also have to reckon with the rapid rate of innovation in the malfeasance sector. Ill-intentioned people are thinking up new ways to misuse AI so quickly that this chapter is likely to be outdated even before it attains printed form. Think of it not as depressing reading, however, but as a call to act before it is too late.

…and then we get a tour of all the ways AIs are going wrong today: surveillance, drones, deepfakes, algorithmic bias, job loss to automation, social media algorithms, etc.

Some of these are pretty worrying. But not all of them.

Google “deepfakes” and you will find a host of articles claiming that we are about to lose the very concept of truth itself. Brookings calls deepfakes “a threat to truth in politics” and comes up with a scenario where deepfakes “could trigger a nuclear war.” The Guardian asks “You Thought Fake News Was Bad? Deepfakes Are Where Truth Goes To Die”. And these aren’t even the alarmist ones! The Irish Times calls it an “information apocalypse” and literally titles their article “Be Afraid”; Good Times just writes “Welcome To Deepfake Hell”. Meanwhile, deepfakes have been available for a couple of years now, with no consequences worse than a few teenagers using them to make pornography, ie the expected outcome of every technology ever. Also, it’s hard to see why forging videos should be so much worse than forging images through Photoshop, forging documents through whatever document-forgers do, or forging text through lying. Brookings explains that deepfakes might cause nuclear war because someone might forge a video of the President ordering a nuclear strike and then commanders might believe it. But it’s unclear why this is so much more plausible than someone writing a memo saying “Please launch a nuclear strike, sincerely, the President” and commanders believing that. Other papers have highlighted the danger of creating a fake sex tape with a politician in order to discredit them, but you can already convincingly Photoshop an explicit photo of your least favorite politician, and everyone will just laugh at you.

Algorithmic bias has also been getting colossal unstoppable neverending near-infinite unbelievable amounts of press lately, but the most popular examples basically boil down to “it’s impossible to satisfy several conflicting definitions of ‘unbiased’ simultaneously, and algorithms do not do this impossible thing”. Humans also do not do the impossible thing. Occasionally someone is able to dig up an example which actually seems slightly worrying, but I have never seen anyone prove (or even seriously argue) that algorithms are in general more biased than humans (see also Principles For The Application Of Human Intelligence – no, seriously, see it). Overall I am not sure this deserves all the attention it gets any time someone brings up AI, tech, science, matter, energy, space, time, or the universe.

Or: with all the discussion about how social media algorithms are radicalizing the youth, it was refreshing to read a study investigating whether this was actually true, which found that social media use did not increase support for right-wing populism, and online media use (including social media use) and right-wing populism actually seem to be negatively correlated (remember, correlational studies are always bad). Recent studies of YouTube’s algorithms find they do not naturally tend to radicalize, and may deradicalize, viewers, although I’ve heard some people say this is only true of the current algorithm and the old ones (which were not included in these studies) were much worse.

Or: is automation destroying jobs? Although it seems like it should, the evidence continues to suggest that it isn’t. There are various theories for why this should be, most of which suggest it may not destroy jobs in the near future either. See my review of technological unemployment for details.

A careful reading reveals Russell appreciates most of these objections. A less careful reading does not reveal this. The general structure is “HERE IS A TERRIFYING WAY THAT AI COULD BE KILLING YOU AND YOUR FAMILY although studies do show that this is probably not literally happening in exactly this way AND YOUR LEADERS ARE POWERLESS TO STOP IT!”

I understand the impulse. This book ends up doing an amazing job of talking about AI safety without sounding weird. And part of how it accomplishes this is building on a foundation of “AI is causing problems now”. The media has already prepared the way; all Russell has to do is vaguely gesture at deepfakes and algorithmic radicalization, and everyone says “Oh yeah, that stuff!” and realizes that they already believe AI is dangerous and needs aligning. And then you can add “and future AI will be the same way but even more”, and you’re home free.

But the whole thing makes me nervous. Lots of right-wingers say “climatologists used to worry about global cooling, why should we believe them now about global warming?” They’re wrong – global cooling was never really a big thing. But in 2040, might the same people say “AI scientists used to worry about deepfakes, why should we believe them now about the Singularity?” And might they actually have a point this time? If we get a reputation as the people who fall for every panic about AI, including the ones that in retrospect turn out to be kind of silly, will we eventually cry wolf one too many times and lose our credibility before crunch time?

I think the actual answer to this question is “Haha, as if our society actually punished people for being wrong”. The next US presidential election is all set to be Socialists vs. Right-Wing Authoritarians – and I’m still saying with a straight face that the public notices when movements were wrong before and lowers their status? Have the people who said there were WMDs in Iraq lost status? The people who said sanctions on Iraq were killing thousands of children? The people who said Trump was definitely for sure colluding with Russia? The people who said global warming wasn’t real? The people who pushed growth mindset as a panacea for twenty years?

So probably this is a brilliant rhetorical strategy with no downsides. But it still gives me a visceral “ick” reaction to associate with something that might not be accurate.

And there’s a sense in which this is all obviously ridiculous. The people who think superintelligent robots will destroy humanity – these people should worry about associating with the people who believe fake videos might fool people on YouTube, because the latter group is going beyond what the evidence will support? Really? But yes. Really. It’s more likely that catastrophic runaway global warming will boil the world a hundred years from now than that it will reach 75 degrees in San Francisco tomorrow (predicted high: 59); extreme scenarios about the far future are more defensible than even weak claims about the present that are ruled out by the evidence.

There’s been some discussion in effective altruism recently about public relations. The movement has many convincing hooks (you can save a live for $3000, donating bednets is very effective, think about how you would save a drowning child) and many things its leading intellectuals are actually thinking about (how to stop existential risks, how to make people change careers, how to promote plant-based meat), and the Venn diagram between the hooks and the real topics has only partial overlap. What to do about this? It’s a hard question, and I have no strong opinion besides a deep respect for everyone on both sides of it and appreciation for the work they do trying to balance different considerations in creating a better world.

HC’s relevance to this debate is as an extraordinary example. If you try to optimize for being good at public relations and convincingness, you can be really, really good at public relations and convincingness, even when you’re trying to explain a really difficult idea to a potentially hostile audience. You can do it while still being more accurate, page for page, than a New York Times article on the same topic. There are no obvious disadvantages to doing this. It still makes me nervous.

V.

My reaction to this book is probably weird. I got interested in AI safety by hanging out with transhumanists and neophiles who like to come up with the most extreme scenario possible, and then back down when maybe it isn’t true. Russell got interested in AI safety by hanging out with sober researchers who like to be as boring and conservative as possible, and then accept new ideas once the evidence for them proves overwhelming. At some point one hopes we meet in the middle. We’re almost there.

But maybe we’re not quite there yet. My reaction to this book has been “what an amazing talent Russell must have to build all of this up from normality”. But maybe it’s not talent. Maybe Russell is just recounting his own intellectual journey. Maybe this is what a straightforward examination of AI risk looks like if you have fewer crazy people in your intellectual pedigree than I do.

I recommend this book both for the general public and for SSC readers. The general public will learn what AI safety is. SSC readers will learn what AI safety sounds like when it’s someone other than me talking about it. Both lessons are valuable.

# Map Of Effective Altruism

In the spirit of my old map of the rationalist diaspora, here’s a map of effective altruism:



Thumbail – click to expand

Continents are cause areas; cities are charities or organizations; mountains are individuals. Some things are clickable links with title-text explanations. Thanks to AG for helping me set up the imagemap.

# Suicide Hotspots Of The World

[Content warning: suicide, rape, child abuse. Thanks to MC for some help with research.]

I.

Guyana has the highest national suicide rate in the world, 30 people per year per 100,000. Guyana has poverty and crime and those things, but no more so than neighboring Brazil (suicide rate of 6) or Venezuela (suicide rate of 4). What’s going on?

One place to start: Guyana is a multi-ethnic country. Is its sky-high suicide rate focused in one ethnic group? The first answer I found was this article by a social justice warrior telling us it constitutes racial “essentialism” to even ask the question. But in the process of telling us exactly what kind of claims we should avoid, she mentions someone bringing up that “80% of the reported suicides are carried out by Indo-Guyanese”. I feel like one of those classicists who has reconstructed a lost heresy through hostile quotations in Irenaeus.

Indo-Guyanese aren’t American Indians; they’re from actual India. Apparently thousands of Indians immigrated to Guyana as indentured laborers in the late 1800s. Most went to Guyana, and somewhat fewer went to neighboring Suriname. Suriname also has a sky-high suicide rate, but slightly less than Guyana’s, to the exact degree that its Indian population is slightly less than Guyana’s. Basically no Indians went anywhere else in South America, and nowhere else in South America has anywhere near the suicide rate of these two countries. The most Indian regions of Guyana also have the highest suicide rate. Hmmm.

Does India itself have high suicide rates? On average, yes. But India has a lot of weird suicide microclimates. Statewide rates range from from 38 in Sikkim (higher than any country in the world) to 0.5 in Bihar (lower than any country in the world except Barbados). Indo-Guyanese mostly come from Bihar and other low-suicide regions. While I can’t rule out that the Indo-Guyanese come from some micro-micro-climate of higher suicidality, this guy claims to have traced them back to some of their ancestral villages and found that those villages have low suicide rates.

So what’s going on? Social and Cultural Dimensions of Indian Indentured Labour and Its Diaspora argues that despite the mixed suicide rates in India itself, rates across the Indian Diaspora are universally high. For example:

The Fiji Indian suicide rate in the period 1900 to 1915 was the highest among all Indian labour importing colonies in Africa and the West Indies, and much higher than in India itself. In Mauritius too, hundreds of indentured Indian laborers committed suicide by jumping from a particular hillock during the indenture period, which acquired the name of ‘Suicide Hill’, now turned into a monument […]

In his article ‘Veil of Dishonor’ Lal (1985) describes what officials tend to point out as the primary cause of the Fiji Indian suicides: sexual jealousy arising from the persistent shortage of women on the plantations. The rate of indentured adult Indian females to males in Fiji was only 43 to 100. The intense competition for women among the indentured men was seen as the main reason for male suicides in Fiji. Lomarsh Roopnarine (2007) also shows high rates of suicides among indentured Indians in British Guiana […]

Although there is no reason to doubt the existence of sexual jealousy, this emphasis on the scarcity of women disregards the arduous circumstances in which the indentured labourers were working, and the disruption of the “integrative institutions” of society – family, marriage, caste, kinship, and religion – as the underlying causes of suicide and other ills affecting the Indian indentured labour population.

Yeah, but arduous circumstances affected dozens of different ethnicities involved in various colonization and forced labor schemes, and most of them didn’t have these kinds of suicide rates. I can kind of imagine a story where first-generation laborers had no hope of settling down or raising families, committed suicide at high rates, and that ingrained suicidal tendencies in the culture that never went away. But then how come that didn’t happen to eg indentured Englishmen in Virginia?

The incongruously named Vijayakumar and John (2018) blame the Hindu religion. Did you know that the Ramayana ends with Rama, three of brothers, and the entire population of his kingdom committing mass suicide by drowning? Or that the mahaparasthana is a traditional Hindu method of suicide “where the person walks in a north easterly direction, subsisting only on water and air, until his body sinks to rest”? Any religion that has a traditional direction to walk in while you’re committing suicide by starving yourself seems kind of suspicious here. But then how come Hindus in some parts of India have such low suicide rates? How come it’s just the diaspora that suffers. The paper suggests maybe it’s because religiosity plays a protective effect, but it sounds kind of strained.

I don’t have better answers to any of these questions. Maybe the combination of Hindu religion, imbalanced gender ratios, and uprooted communities created a perfect storm. I don’t have any better ideas.

II.

Guyana, at 30 suicides per year per 100,000, is the highest national suicide rate in the world. But if Greenland ever wins independence, it will steal first place. Greenland’s suicide rate is 83 per year per 100,000, almost three times higher than any other country in the world.

Like Guyana, this is more ethnic than national. Greenland is mostly Inuit, and Inuit everywhere have equally high suicide rates. The suicide rate in the mostly-Inuit Canadian province of Nunavit is 71 (for comparison, Canada in general is 10). The suicide rate among Alaskan Inuit is 40 (for comparison, the US in general is 14).

This definitely is not just because of the cold and darkness. White Alaskans who live right next to Alaskan natives have a rate of about 20, not much higher than the US average. And suicide in Greenland – like everywhere else – peaks in the spring and summer anyway.

Most damning of all, Greenland’s high suicide rates are a recent phenomenon. In 1971, the rate was 4. I didn’t forget a zero there. Fifty years ago, Greenland had one of the lowest suicide rates in the world. But by 1990, it had reached 120 (it’s since come down a little bit). What happened in those twenty years?

You would think limiting it to such a short time period would make things easy. It isn’t. There are two main theories: social alienation, and alcohol.

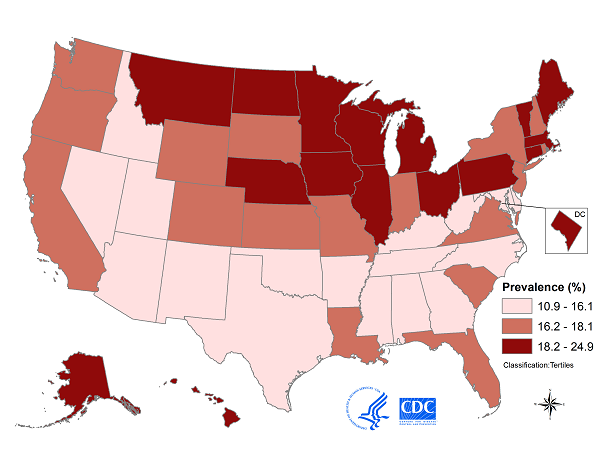
There is definitely a lot of social alienation. For centuries, the Inuit hunted seals in traditional villages. At some point the Danish government decided that was unacceptably backwards and resettled them in cities, especially the capital of Nuuk. This didn’t go well.

One counterargument to this story is that Nuuk has the lowest suicide rate in Greenland, and the more remote the village, the worse the suicide crisis. Maybe you could argue that everywhere was modernized and disrupted and alienated but at least a big city has some interesting stuff to do. This would kind of match the American experience, where it’s small towns in West Virginia that are getting hit by the opioid crisis and deaths of despair.

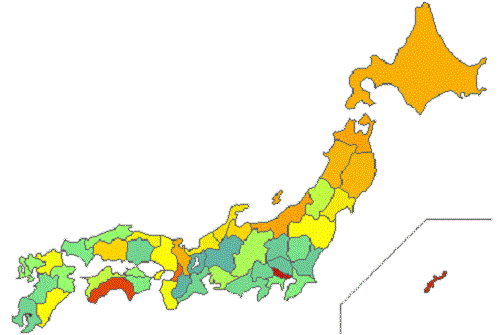
Another counterargument is that all Native American communities suffered a lot of displacement and alienation and modernization, but none of them suffered the same suicide spike as the Inuit. Sources disagree on the exact Native American suicide rate in the US, but it isn’t unusually high; the CDC numbers say it is slightly below the rate for non-Hispanic whites. Canadian First Nations suicide rates are elevated, but still only a third or so of Inuit levels. Maybe Inuits suffered stronger relocation pressures than other native peoples because of their Arctic environment? Or maybe every native group suffered a suicide spike, but Native Americans and Canadians have adjusted by now and their suicide rates have come back down? I’m not sure.

The other theory about Greenland is alcohol. Alcohol consumption in Greenland skyrocketed around the same time suicide did, reached levels that temporarily made Greenland by far the most alcoholic country in the world – then started declining around the same time suicide did. This seems to be a pattern when hunter-gatherers with no genetic or cultural resistance encounter alcohol for the first time – Native Americans in the 1700s got up to some crazy stuff.

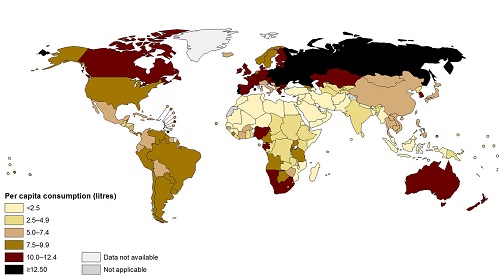
But the Inuit seem to have gotten it much worse. Now we can bring back in the cold and darkness. Alcohol consumption seems to increase reliably with latitude, whether we’re talking about the US:



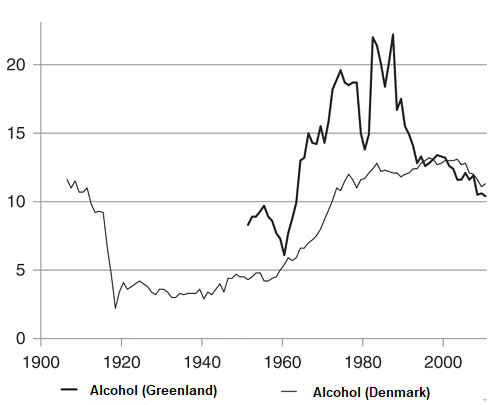
Japan:



Or the whole world:



So you take some hunter-gatherers who have never encountered alcohol before, stick them in the northernmost place in the world, and throw cheap Danish alcohol at them at the exact moment their communities are being uprooted and destroyed forever, and you get…well, you get this:



By 1980, Greenland was the most alcoholic country in the world, drinking an average of 22 liters of pure alcohol per capita per year (Russia is 15). It doesn’t look like this was responsible social drinking either. Take the most deranged binge drinking in the worst college fraternity in the world, multiply it by a thousand, and that was Greenland during much of the late 20th century.

But this can’t be the whole story. Alcohol consumption in Greenland has since dropped to the same level as Denmark and other European countries. But the suicide rate is still ten times higher. Why? Maybe the moderate quantities are hiding deeply dysfunctional drinking patterns with lots of binging and addiction.

Or maybe it’s something worse. Child sexual abuse rates in Greenland range from 37% in Nuuk to 46% in East Greenland. As far as I can tell, you are understanding those numbers correctly – almost half of children in Greenland are sexually abused. In Nunavut, the numbers are 52% of women and 22% of men suffering “severe” childhood sexual abuse. The New York Times sets a disturbingly vivid scene:

Pay days are the worst time for the children of Tasiilaq, officials say. With their salaries or social benefits in hand, many adults tend to drink and parents become too inebriated to look after their children, officials say. That’s when an already high rate of sexual abuse rises, according to a police study published last week […]

So on the last Friday of every month, officials open a sports hall in the district as a shelter to keep children away from sexual abuse.

“Children were abused by their stepfathers, cousins and by the neighbor looking after them as the parents were on a bender,” Naasunnguaq Ignatiussen Streymoy, the mother of a sexual abuse victim and an anti-abuse activist, told Weekendavisen, a newsweekly, in an article published on Friday about the crisis.”

Correlation is not causation. Maybe the same dysfunction or social alienation or alcoholism that causes the sexual abuse separately causes the suicides. But maybe the obvious answer is true, and the sexual abuse contributes to the mental health problems that eventually lead to suicide. Maybe a generation of staggeringly high alcoholism led to staggeringly high child abuse, and a generation later those children are still committing suicide at staggeringly high rates.

III.

This is getting really depressing. Let’s talk about something a little bit lighter, like the remote Siberian okrugs with the highest suicide rates in the world.

The highest suicide rate I have seen credibly attributed to an ethnic group is the Chukchi of northeastern Russia, who are said to have reached 165 per year per 100,000 in 1998. They may be distantly related to the Inuit, but I wouldn’t put too much weight on this; Siberia is riddled with weird ethnicities with super-high suicide rates. The Evenks reached 121; their western neighbors the Nenets reached 119. There is a group called the Koryaks with a rate of 92, and another group called the Udmurts with a rate of 40ish – which is still higher than Guyana.

Voracek, Fisher, and Marusic try to tie some of these groups into their Finno-Ugrian Suicide Hypothesis, claiming that the genetically-related Finno-Ugric group have a unique predisposition to suicidality. The theory has some superficial plausibility – in the 1990s, the world’s first, second, and third most suicidal countries were Finland, Hungary, and Estonia – all Finno-Ugric. Their surrounding non-Finno-Ugric neighbors, like Sweden or Austria, were unremarkable, so a genetic hypothesis made sense. Unfortunately for the theory (but fortunately for everyone else) these countries have since improved by a lot, and now are barely above the world average; improved mental health care may be responsible (and the fall of Communism didn’t hurt). I’m actually a little confused what happened here.

But the Finno-Ugric hypothesis can’t explain the Chukchi, Evenks, Nenets, Koryaks, and Udmurts. Sure, the Udmurts are Finno-Ugric. And the Nenets are closely related. But the Chukchi, Evenks, and Koryaks aren’t. It’s tempting to group all of these tiny Siberian ethnic groups together, but eg the Evenks are more closely related to the Japanese than they are to the Nenets (despite living right next to them). Any genetic hypothesis flounders on the sheer genetic diversity and unrelatedness of this region.

Psychologist David Lester tries to point the finger at these groups’ ancient culture, which he says has been especially suicidal since the time of the earliest records. He quotes an account of the 19th-century Chukchi:

Bogoras described the [Chukchi] as irritable and obstinate and, when frustrated, impulsively self-destructive. He reports the case of a young girl who hung herself when her mother refused to take her to a feast in a neighboring camp. [He] reported cases of suicide in a husband over grief at his wife’s death andof a mother after her ten-year-old son’s death; a case motivated by bad fortune, compounded by the fear of further bad fortune; a woman who no longer found any pleasure in life; a young man who was driven away by his father-in-law for being lazy who then killed his pregnant wife and himself; and a young woman whose husband wanted to lend her to a friend in a group marriage, a friend whom she disliked.

Suicidal behavior appeared to be so common that people planning to kill themselves would often ask for a last meal of exotic tastes before they did so. Some Chukchee prefer to commit suicide by having someone else kill them. The man reported above who committed suicide because of present and anticipated misfortune asked to be strangled. In another case,a man who fell ill asked his wife to shoot him. Bogoras noted that ‘voluntary death’ as he called it, suicide by getting others to kill oneself, was common for the elderly and those suffering from physical illnesses.

However, Bogoras also noted ‘peculiar’ causes of voluntary death, such as that of a man who grew weary of quarrelling with his wife over their ill-behaved sons. Part of the motive in these cases may be to induce guilt in the survivors. As one father said, ‘Then he asks to be killed, and charges the very son who offended him with the execution of his request. Let him give me the mortal blow, let him suffer from the memory of it’.

I can only aspire to one day achieve this level of passive-aggressiveness. But in the end it has the same problem as the genetic hypothesis: these groups are just totally unrelated to each other. The Chukchi are not much more suicidal than the Nenets or Evenks, who have none of these traditions. And the Inuit are up there with all of them, and they had one of the lowest suicide rates in the world pre-colonization.

I think the explanation here is the same as with Greenland: the combination of alcohol-naive hunter gatherers, alcohol, the alcohol-promoting effects of high latitudes, and a disruption of the traditional way of life. There’s apparently a Russian proverb about Siberians that goes “reindeer-herders are sober only when they don’t have the money to get drunk” – and when the Russians are appalled by your alcoholism, you know you have a problem. Alcohol was found in the blood of 75 – 80% of Nenets suicides. And if anything, the Siberians had their way of life disrupted even worse than the Inuit did – Soviet central planners tried to collectivize them as a PR move – they wanted to demonstrate that Communism could work for even the most primitive of peoples. Well, it didn’t, and here we are.

While genetics or culture may matter a little, overall I am just going to end with a blanket recommendation to avoid being part of any small circumpolar ethnic group that has just discovered alcohol.

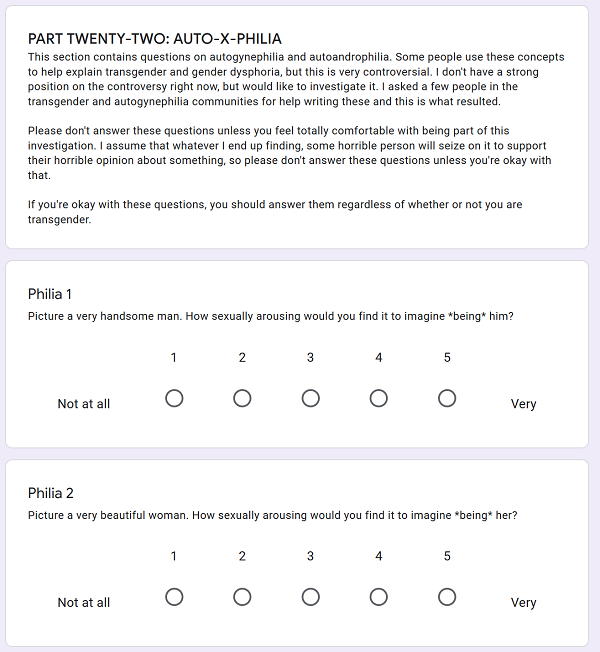
# Autogenderphilia Is Common And Not Especially Related To Transgender

“Autogynephilia” means becoming aroused by imagining yourself as a woman. “Autoandrophilia” means becoming aroused by imagining yourself as a man. There’s no term that describes both, but we need one, so let’s say autogenderphilia.

These conditions are famous mostly because a few sexologists, especially Ray Blanchard and Michael Bailey, speculate that they are the most common cause of transgender. They point to studies showing most trans women endorse autogynephilia. Most trans people disagree with this theory, sometimes very strongly, and accuse it of reducing transgender to a fetish.

Without wading into the moral issues around it, I thought it would be interesting to get data from the SSC survey. The following comes partly from my own analyses and partly from wulfrickson’s look at the public survey data on r/TheMotte.

The survey asked the following questions:



First of all, thanks to the 6,715 people (182 trans, 6259 cis, 274 confused) who answered these questions despite my disclaimers. Here’s how it worked out. 5 is maximally autogenderphilic, 1 is no autogenderphilia at all:

Group (n)  
Autogynephilia  
Autoandrophilia  
  
Cis men (5592)  
2.6  
1.9  
  
Cis women (667)  
2.5  
2  
  
Trans men (35)  
1.9  
2.3  
  
Trans women (147)  
3.2  
1.3

\*sexual orientation was self-reported. Almost all transgender people report sexual orientation relative to their current gender rather than their birth gender, so for example a “lesbian trans woman” would be someone who grew up male, currently identifies as female, and is attracted to other women. This is the opposite of how Blanchard and Bailey sometimes use these terms, so be careful comparing these results to theirs!  
\*\*results are marked as ??? for groups with sample size lower than 20

The survey confirmed Blanchard and Bailey’s finding that many lesbian trans women had strong autogynephilia. But it also confirmed other people’s findings that many cis people also have strong autogenderphilia. In this dataset, autogenderphilia rates in gay cis men were equal to those in lesbian trans women.

Autogenderphilia in cis people was divided between fantasies about being the opposite gender, and fantasies about being the gender they already were. What does it mean to fantasize about being a gender you already are? I asked a cis female friend who admitted to autogynephilia. She told me:

My literal body is arousing – it’s hot that I have breasts and can get pregnant and have a curvy figure and a feminine face and long hair, and it’s hot to dress up in femme clothes. There are certain gendered/social interactions that are very hot, or that can easily springboard into ones that are very hot. I’ve honestly wondered whether I might not be nonbinary or trans male, because I’m not really sure how euphoric being female is, besides that it’s like living in a sex fantasy.

(score one for the hypothesis that this kind of thing causes gender transition, because after reading this I kind of want to be a woman.)

Uh…moving on. The highest rates of autogenderphilia were found in bi cis men (autoandrophilia), gay cis men (autoandrophilia), bi trans women (autogynephilia), and lesbian trans women (autogynephilia).

These groups all have three things in common: they identify as the gender involved, they are attracted to the gender involved, and they are biologically male.

I would guess biological men have more of every fetish, regardless of their current gender identity, so it’s not surprising that they have more autogenderphilia also. In fact, we see that in biological women, the two highest categories are bi cis women (autogynephilia), and lesbian cis women (autogynephilia); again, they identify as the gender involved, and they are attracted to the gender involved.

So abstracting that away, the SSC survey data suggest a very boring hypothesis of autogenderphilia: if you identify as a gender, and you’re attracted to that gender, it’s a natural leap to be attracted to yourself being that gender.

The SSC survey hypothesis explains the same evidence that Blanchard and Bailey’s hypothesis explains (that lesbian trans women very often have autogynephilic fantasies), but reverse the proposed causation: it’s not that autogynephilia causes gender transition; it’s that identification as a gender is one factor that causes autogenderphilia.

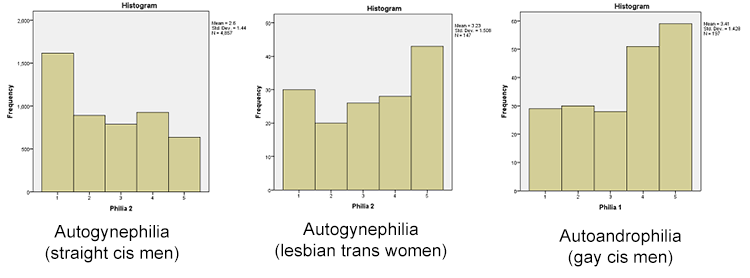
But after that, it can go on to explain other things that Blanchard and Bailey can’t explain, like why cis gay men have as much autoandrophilia as trans lesbian women have autogynephilia. Or why some people with low levels of autogenderphilia transition, but many people with high levels don’t. I think it’s a simpler and more defensible explanation of the evidence.

I asked some people I know who supported Blanchard and Bailey’s theory for their thoughts. They focused on a few concerns about the data.

First, weird Internet samples plausibly have more of every paraphilia. This might inflate the rate for cis gay men and the number of trans lesbian women, assuming the latter all had to be above some cutoff; that might falsely lead me to believe the two groups have the same rate.

One counterargument might be that the responses among cis people alone are enough to generate the hypothesis discussed above. The low rates of autogynephilia in gay men, compared to in straight and bi men, suggest that being attracted to a gender is a prerequisite of autogenderphilia to it. And (adjusting again for the general tendency of male-bodied people to have more fetishes) the higher rates of autogynephilia in cis women/autoandrophilia in cis men, compared to autoandrophilia in cis women/autogynephilia in cis men, suggest that identifying as a gender is a prequisite to autogenderphilia to it.

Another counterargument might be the similarity of the histograms produced by cis gay male and trans lesbian female responses; they don’t look like they’re being generated by two different processes which have only coincidentally averaged out into the same summary statistic:



This doesn’t look like all cis men over a certain cutoff are becoming trans women; it looks like the curve for cis gay men and trans lesbian women are being shaped by the same process.

Second, did the survey questions accurately capture autogenderphilia? Fetishes range from very mild to very extreme; some people like being slapped during sex, other people have whole BDSM dungeons in their basement. Is it possible the survey captured some boring meaning of autogenderphilia, like “sure, I guess it would be hot to be a woman”, but some people have a much stronger and more obsessive form? The histogram above argues against this a little, but there might be ceiling effects.

Alice Dreger seems to take something like this perspective here:

Q: Do you think autogynephilia might be a part of the female experience, trans or cis? I’ve seen some (very preliminary) theorizing about it as well as a paper with a tiny sample size that suggest that cis women also experience sexual arousal at the thought of themselves as women.

A: I’ve talked with Blanchard, Bailey, and also Anne Lawrence about this, and my impression is they all doubt cis (non-transgender) women experience sexual arousal at the thought of themselves as women. Clinically, Blanchard observed autogynephilic natal male individuals who were aroused, for example, at the ideas of using a tampon for menses or knitting as a woman with other women. I have never heard a natal woman express sexual arousal at such ideas. I’ve never heard of a natal woman masturbating to such thoughts.

I asked the same cis female friend who gave me the quotation above, and she described using a tampon to masturbate and finding it hot. I think Dreger makes an important point that there are some pretty unusual manifestations of autogenderphilic fetishes out there and we should hesitate before drawing too many conclusions from a single question that lumps them all together. But also, Alice Dreger seems like an really dignified and important person who probably doesn’t hang out with people who talk openly about their menstruation-related masturbation fantasies, and she should probably adjust for that. Maybe she could move to the Bay Area.

There’s a common failure mode in psychiatry, where we notice people with some condition doing some weird thing, and fail to notice that huge swathes of people without the condition do the exact same weird thing. For example, everyone knows schizophrenics hear voices, but until recently nobody realized that something like 20% of healthy people do too. Everyone knows that LSD users can end up with permanent visual hallucinations, but until recently nobody realized that lots of drug-free people have the same problem. Schizophrenics definitely hear more voices than healthy people, and LSD users have more permanent visual hallucinations, but it’s movement along the distribution rather than a completely novel phenomenon.

I think autogenderphilia is turning out to work the same way, and that this will require us to reassess the way we think about it.

As usual, I welcome people trying to replicate or expand on these results. All of the data used in this post are freely available and can be downloaded here. I’ve also heard Michael Bailey is going to release his own interpretation of these data, so stay tuned for that. I’d like to delve into these issues further on future surveys, so let me know if you have ideas about how to do that.

And a big thanks to Tailcalled for helping me set up this section of the survey. If you’re interested in these issues, you might enjoy his blog or his own analysis of these results.

# Welcome (?), Infowars Readers

Hello to all the new readers I’ve gotten from, uh, Paul Watson of Infowars. Before anything else, consider reading this statement by the CDC about vaccines.

Still here? Fine.

Infowars linked here with the headline Survey Finds People Who Identify As Left Wing More Likely To Have Been Diagnosed With A Mental Illness. This is accurate only insofar as the result uses the publicly available data I provide. The claim about mental illness was made by Twitter user Philippe Lemoine and not by me. In general, if a third party analyzes SSC survey data, I would prefer that media sources reporting on their analysis attribute it to them, and not to SSC.

As far as I can tell, Lemoine’s analysis is accurate enough, but needs some clarifications:

1. Both extreme rightists and extreme leftists are more likely than moderates to have been diagnosed with most conditions.

2. Leftists might be more likely to trust the psychiatric system and get diagnosed. My survey shows some signs of that. Liberals are 60% more likely than conservatives to have formal diagnoses of depression, but only 30% more likely to have a self-diagnosis of depression.

3. Leftists might be more likely to think of their issues through a psychiatric lens than rightists, meaning that even the self-diagnosis numbers might be inflated.

4. The SSC survey is a bad sample to use for this, not just because it’s unrepresentative, but because it might be unrepresentative of different political affiliations in different ways. For example, SSC Marxists really are surprisingly depressed, but maybe the only Marxists who would read an anti-Marxist blog are depressed Marxists looking for things to be miserable and angry about (though see below for some counterevidence).

5. A commenter on Lemoine’s tweet links to this blog post by someone who found the same thing in the General Social Survey. The General Social Survey is much larger and more rigorous than my survey, and there’s no reason to care what my survey has to say when there are GSS results available.

In general, if a survey analysis is posted on this blog, it’s mine. If not, then it isn’t mine and you should link to whoever performed it and let them clean up their own mess. Thanks – and seriously, vaccines are fine.

# Confirmation Bias As Misfire Of Normal Bayesian Reasoning

From the subreddit: Humans Are Hardwired To Dismiss Facts That Don’t Fit Their Worldview. Once you get through the preliminary Trump supporter and anti-vaxxer denunciations, it turns out to be an attempt at an evo psych explanation of confirmation bias:

Our ancestors evolved in small groups, where cooperation and persuasion had at least as much to do with reproductive success as holding accurate factual beliefs about the world. Assimilation into one’s tribe required assimilation into the group’s ideological belief system. An instinctive bias in favor of one’s in-group” and its worldview is deeply ingrained in human psychology.

I think the article as a whole makes good points, but I’m increasingly uncertain that confirmation bias can be separated from normal reasoning.

Suppose that one of my friends says she saw a coyote walk by her house in Berkeley. I know there are coyotes in the hills outside Berkeley, so I am not too surprised; I believe her.

Now suppose that same friend says she saw a polar bear walk by her house. I assume she is mistaken, lying, or hallucinating.

Is this confirmation bias? It sure sounds like it. When someone says something that confirms my preexisting beliefs (eg ‘coyotes live in this area, but not polar bears’), I believe it. If that same person provides the same evidence for something that challenges my preexisting beliefs, I reject it. What am I doing differently from an anti-vaxxer who rejects any information that challenges her preexisting beliefs (eg that vaccines cause autism)?

When new evidence challenges our established priors (eg a friend reports a polar bear, but I have a strong prior that there are no polar bears around), we ought to heavily discount the evidence and slightly shift our prior. So I should end up believing that my friend is probably wrong, but I should also be slightly less confident in my assertion that there are no polar bears loose in Berkeley today. This seems sufficient to explain confirmation bias, ie a tendency to stick to what we already believe and reject evidence against it.

The anti-vaxxer is still doing something wrong; she somehow managed to get a very strong prior on a false statement, and isn’t weighing the new evidence heavily enough. But I think it’s important to note that she’s attempting to carry out normal reasoning, and failing, rather than carrying out some special kind of reasoning called “confirmation bias”.

There are some important refinements to make to this model – maybe there’s a special “emotional reasoning” that locks down priors more tightly, and maybe people naturally overweight priors because that was adaptive in the ancestral environment. Maybe after you add these refinements, you end up at exactly the traditional model of confirmation bias (and the one the Fast Company article is using) and my objection becomes kind of pointless.

But not completely pointless. I still think it’s helpful to approach confirmation bias by thinking of it as a normal form of reasoning, and then asking under what conditions it fails.

# Addendum to “Targeting Meritocracy”

I’ve always been dissatisfied with Targeting Meritocracy and the comments it got. My position seemed so obvious to me – and the opposite position so obvious to other people – that we both had to be missing something.

Reading it over, I think I was missing the idea of conflict vs mistake theory.

I wrote the post from a mistake theory perspective. The government exists to figure out how to solve problems. Good government officials are the ones who can figure out solutions and implement them effectively. That means we want people who are smart and competent. Since meritocracy means promoting the smartest and most competent people, it is tautologically correct. The only conceivable problem is if we make mistakes in judging intelligence and competence, which is what I spend the rest of the post worrying about.

From a conflict theory perspective, this is bunk. Good government officials are ones who serve our class interests and not their class interests. At best, merit is uncorrelated with this. At worst, we are the lower and middle class, they are the upper class, and there is some system in place (eg Ivy League universities) that disproportionately funnels the most meritorious people into the upper class. Then when we put the most meritorious people in government, we are necessarily seeding the government with upper class people who serve upper class interests.

This resolves my confusion about why people disagree with me on this point. It reinforces a lesson I’ve had to learn again and again: if people seem slightly stupid, they’re probably just stupid. But if they seem colossally and inexplicably stupid, you probably differ in some kind of basic assumption so fundamental that you didn’t realize you were assuming it, and should poke at the issue until you figure it out.

# Sleep Support: An Individual Randomized Controlled Trial

I worry my sleep quality isn’t great. On weekends, no matter when I go to bed, I sleep until 11 or 12. When I wake up, I feel like I’ve overslept. But if I try to make myself get up earlier, I feel angry and want to go back to sleep.

A supplement company I trust, Nootropics Depot, recently released a new product called Sleep Support. It advertises that, along with helping you fall asleep faster, it can “improve sleep quality” by “improv[ing] sleep architecture, allowing you to achieve higher quality and more refreshing sleep.” I decided to try it.

The first night I took it, I woke up naturally at 9 the next morning, with no desire to go back to sleep. This has never happened before. It shocked me. And the next morning, the same thing happened. I started recommending the supplement to all my friends, some of whom also reported good results.

I decided the next step was to do a randomized controlled trial. I obtained sugar pills, and put both the sugar pills and the Sleep Support pills inside bigger capsules so I couldn’t tell which was which. The recommended dose was two Sleep Support pills per night, so for my 24 night trial I created 12 groups of two Sleep Support pills and 12 groups of two placebo pills.



Then I asked a friend to flip a coin 24 times, and depending on the result place either a pair of Sleep Support pills or a pair of placebo pills in each slot of a monthly pill planner, and record which slot contained which pills on a secret piece of paper I could see at the end of the experiment. Then every weekend night for three months I took the next pair of pills in the planner and recorded:

– the time I went to bed  
– the time I woke up  
– my subjective rating of how well-rested I was upon waking  
– my subjective rating of how much energy I felt like I’d had that day  
– my subjective rating of how vivid my dreams were that night  
– my subjective guess about whether I’d taken placebo or experimental that night

The time I went to bed wasn’t intended to be a dependent variable; I generally took the pills just before going to bed, so they couldn’t affect that. And I had no way of measuring what time I went to sleep. It was just so that I could measure my total time in bed that night.

The time I woke up was the hardest to operationalize. Usually I wake up a few times in the morning, groggily check the clock, and decide to go back to bed, then wake up for good once it becomes so late I start feeling guilty about how much of the day I’m wasting. I considered setting wake-up time as the very first time I woke up to check the clock, but sometimes I wake up at 5 AM to go to the bathroom, and I didn’t want that to get recorded as me “waking up” at 5 AM. And if I used a cutoff like “the first time I wake up after 7”, then a night I wake up at 6:59 and go back to bed and wake up for good at 11 would get recorded completely differently from a night I wake up at 7:01 and go back to bed and wake up for good at 11. But if I defined wake-up time as the time I finally woke up for good, then it would be too easy for me to subconsciously bias the experiment. “This feels like a night I took placebo, better stay in bed until at least 11:30”.

I decided to eliminate the whole problem by forbidding myself to check the clock while in bed. I would go to sleep, wake up, either decide to go back to sleep or not, and I wasn’t allowed to check the clock until I had gotten out of bed and gotten dressed.

Here’s the headline results of the experiment – number of hours I slept during experimental vs. control nights, and wake up time during experimental vs. control nights



On average there was no difference between the two groups on either measurement. There was also no difference on any of the subjective measures. My subjective guess about whether I’d taken experimental or placebo capsules that night had no correlation with the reality.

My conclusion isn’t that Sleep Support doesn’t work; I didn’t even try it for its main indication of helping with insomnia. My study was too underpowered to detect even medium-sized effects. And just because it didn’t work for me doesn’t mean it won’t work for somebody else.

My conclusion is that the effect I thought that I observed – a consistent change of two hours in my otherwise stable wake-up time – wasn’t real. This shocked me. What’s going on?

I think my original strategy of “wake up a few times in the morning, check the clock, and finally get out of bed when you really feel like it” is very susceptible to the placebo effect. Usually I might wake up at 9, decide that was too early to face the world, and go back to bed. Maybe I wouldn’t even remember doing this. Part of this was probably inertia – I wasn’t used to getting up at 9, I figured I must not have gotten enough sleep to feel good, and so I didn’t want to do it today. Once I had an exciting new sleep supplement in my system, I woke up at 9, actually checked whether or not I felt ready to wake up, and absent my usual prior that I wasn’t, I found that I was, and woke up.

This hypothesis is supported by the results of the experiment. On about a third of days, I woke up before 10 – again, something I never would have done before starting Sleep Support. I think the active ingredient here was not letting myself look at the clock. Without external cues to tell me how tired I should feel, I was forced to rely on how tired I actually felt, which in many cases was “not tired at all”. This happened regardless of whether I was taking Sleep Support or placebo that day.

Ironically, even though the supplement failed to differentiate itself from placebo, I think this is one of the most successful biohacking experiments I’ve ever done. I’m getting up on average an hour or so earlier than I did before, getting more done, and not feeling any more tired by the evening.

Future research: see if this keeps working even now that I know what’s going on.

You can download my raw data here. If you want to replicate this experiment, you can buy Sleep Support capsules here. There are lots of ways to make a placebo; I found these very large empty capsules helpful.

I’m interested in hearing about anyone else’s experience conducting controlled trials of supplements on themselves; if you do something like this and want to publish it on a blog, let me know.

# Book Review: Just Giving

I.

Traditional book reviews tend to focus on a single book, such as Just Giving by Rob Reich. We ought, however, to be reviewing a broader question: what is the role of books in a liberal democratic society? And what role should they play?

Books were first invented during the early Bronze Age. Plato states people fiercely opposed the first books; in his dialogue Phaedrus, he recalls the Egyptian priests’ objection to early writing:

[Writing] will create forgetfulness in the learners’ souls, because they will not use their memories; they will trust to the external written characters and not remember of themselves. The specific which you have discovered is an aid not to memory, but to reminiscence, and you give your disciples not truth, but only the semblance of truth; they will be hearers of many things and will have learned nothing; they will appear to be omniscient and will generally know nothing; they will be tiresome company, having the show of wisdom without the reality.

Contrast the Egyptian scribes’ reception with the ceaseless praise given to the authors of our age. Rather than asking about the purposes of writing and the power of authors, we tend instead to celebrate writers, large and small, for their brilliance. But in our age, these are questions we should pose with greater urgency. Scholarly literature like Just Giving is an unaccountable, nontransparent, and perpetual exercise of power. It deserves more criticism than it has received.

There’s a conventional story to tell about book-writing and its relation to liberty. The story is this. Book-writing is thought to be tightly connected to liberty. This is so for two reasons. First, writing or reading a book is voluntary. Second, the exercise of liberty involves freedom of speech. This story is an attractive one, and it contains some truth. But it ignores that book-writing is inherently embedded in state institutions, like intellectual property laws. It should not be understood in the simplistic manner of an activity that takes place within a framework of nonintervention by the state, or as nothing more than private individual decisions to express thoughts. Instead, it must be understood as embedded in political institutions, laws, and public policy. Books may not be an invention of the state, but they are an artifact of it.

Most importantly, book-writing is heavily subsidized by the government. Authors receive a “pass-through” tax deduction of up to 20%. In addition, they can deduct most of the expenses they incur in writing a book, from freelance editing to literary agents to promotional events. In extending these tax incentives, federal and state treasuries forego tax revenue. Or to put it differently, tax incentives for writing books constitute a a kind of spending program. In fact, the fiscal effects of a direct spending program and a tax expenditure are exactly the same. In Suzanne Mettler’s apt phrase, federal policy driven by tax expenditures rather than direct spending constitutes the “submerged state”, obscured from public view and accountability, but with powerful distributional consequences. These tax breaks amount to massive federal and state subsidies for the creation and dissemination of written texts. They are supplemented by millions spent on libraries, literacy programs, and in some cases direct subsidies to book publishers.

A respect for the liberty of individuals to promote their views is one thing; subsidizing its exercise is another. The state does not merely permit and set guidelines within which writing takes place – offering the state’s imprimatur to every book and pamphlet and magazine and journal article – but is in a fiscally meaningful way actively participating in what authors do. If the state is actively funding, through a tax expenditure, some bad book, it makes the state partially complicit in the harm that the book causes. It is no exaggeration to say that as book-writing is currently structured, when authors do harm, so does the state. It is incorrect to say that mediocre books merely waste the time of the author and reader. Rather, writing a mediocre book squanders assets that are partially the public’s.

With this description of the relationship between book-writing and liberty in place, let us now consider whether the ideal of equality is playing any role in the institutional design of the policies. The median annual wage of authors is $62,000, twice the average US income of $31,000. Authors are most likely to be college-educated, upper class, and be the sorts of people who can take months off of their jobs to write a book. Scholarly books are often written by professors, a member of a tiny and unelected intellectual elite. This makes their immense ability to exercise power by writing a book and getting it published deeply troubling. Or consider a famous author like Jeffrey Sachs, whose successful books permit him to stride upon the world stage as if he were a head of state.

Perhaps books could play important roles in democratic societies, despite being an exercise of power and expression of plutocratic voice, if they were subject to different legal arrangements. But it is no coincidence that the wealthy dominate book-writing. The tax code is set up to unfairly privilege books by the well-off over the poor in two ways. First, the pass-through deduction and freelance editor deduction are available only to those individuals who itemize their deductions – people who opt not to take the so-called standard deduction on their income tax. This effectively penalizes, or fails to reward and provide an incentive for, all people who do not itemize their deductions, a group that constitutes roughly 70% of taxpayers. Thus the low-income renter who does not itemize her deductions but pays $500 to get their book edited receives no tax concession, while the high-income house owner who pays the same $500 fee can claim a deduction. Second, the tax subsidy given to those who do receive the deduction possesses what is known as an “upside-down effect”. The deduction functions as an increasingly greater subsidy with every higher step in the income tax bracket. Both of these features of the tax code are arbitrary and unfairly benefit the well-off. The choice of the the pass-through and business expenses tax deductions as the preferred tax policy for book-writing introduces a potent plutocratic bias.

Proponents of books might suggest that they nevertheless serve a compelling public interest in the form of spreading knowledge. But when we move away from individual works and consider the total distribution of literature, we find a pattern of writing that is hard to reconcile with expectations of educational outcomes. For anyone who believes that books imply something about knowledge or truth or education, the sunny picture of American book publishing here becomes decidedly cloudy.

Figure 1 demonstrates that the most popular type of book in the US is “mystery, thriller, or true crime”. Second and third place are held by history and biographies, which may perhaps be edifying to some people. But after that we get romance, cookbooks, science fiction, and fantasy. Literary classics and books on important current affairs are far down the list, only a fraction of total books read. What can we conclude from these data? The lesson is obvious: if we believe the purpose of reading and writing to be predominantly educational, an important mechanism to provide for our enlightenment and edification, the actually existing distribution of reading in the United States does not meet the test. Not by a long shot.

Finally, we must address the question of intergenerational justice. Books are designed to enshrine author intent and express their opinions in perpetuity. Thus does the dead hand of the author potentially extend from beyond the grave to strangle future generations. John Stuart Mill famously wrote that “There is no fact in history which posterity will find it more difficult to understand, than the idea of perpetuity, and that any of the contrivances of man, should have been coupled together in any sane mind.” Yet authors deliberately “write for the ages”, producing works that can be studied for hundreds or even thousands of years.

We might ask whether books would be a welcome institutional arrangement if we were designing a democratic society from scratch. The catalogue of the oddities of the book suggests a strong case against. Books appear at odds with democracy, for they represent, by definition and by law, the expression of plutocratic voices directed to public education. But why, in a democracy, should the size of one’s wallet give one a greater say in public policy? Why should this plutocratic voice be subsidized by the public? And why should democracy allow this voice to extend across generations in the form of intellectual property laws? It would seem that books are a misplaced plutocratic and powerful element in a democratic society. And we can trace, in the evolution of books, the emergence of a particular kind of high-profile author such as Dan Brown, Stephen King, and J.K. Rowling, whose activity supplants the state, subverts public education processes, and in so doing diminishes democracy.

I find many points of agreement, especially when considering the actual content of books today. Yet despite all this, I think a role for books can be defended. First, books can help overcome problems in the marketplace of ideas by diminishing government orthodoxy and decentralizing the production of knowledge. Second, because of their size and longevity, books can operate on a different and longer timeline than government propaganda broadcasts, taking risks in the expression of ideas we should not routinely expect to see in press releases by government agencies.

This argument is not intended to justify the full range of legal permissions currently afforded to books, but it provides hints as to what a just literary world might look like. I worry, for instance, about the massive boom in short books. Books with fewer than 150 pages primarily serve the author’s vanity. What loss to public benefit would there be with a minimum page length to publish a book, say 250 or 300 pages? I think very little, and quite possibly there would be some gain, for people with less exciting ideas who could not reach the page threshold might be convinced to help other people with their books rather than writing their own. But even if books of all lengths do partly decentralize the definition and provision of knowledge, the resulting pluralism of literary voices will have a plutocratic, not fully democratic, cast. The experimental or heterodox opinions in books will represent the preferences of the wealthy, not of the wider citizenry. Indeed, there is empirical evidence to suggest that at least in the United States, the very wealthy have significantly more politically conservative preferences than average citizens. Thus, the activity of books, even when it decentralizes the production of knowledge, retains a plutocratic character. Does this mean that we should eliminate books? I do not think so. Perhaps a plutocratic tempering of government orthodoxy is better than no tempering at all. I conclude that the decentralization argument provides a plausibly but not definitive case for books as a democracy-supporting institutional design in our society.

In conclusion, how can we make books more compatible with a democratic society? I propose that instead of giving authors tax deductions, they might receive a certain percent of their expenses paid back to them by the government, capped at $100, and that books with fewer than 150 pages should be banned. Are books democratically required? I am not prepared to answer this question affirmatively, for a democratic government has multiple mechanisms to cultivate pluralism and foster discovery. But I have shown that books are certainly democratically permissible.

II.

Yeah, okay, that was weird.

But I put the blame squarely on the hands of Rob Reich, author of Just Giving. The structure, arguments, and most of the individual sentences are his, not mine. I just changed the word “charity” to “books”, and replaced all the charity-related examples with book-related examples. A few parts were edited slightly to make them flow together better, and a few sentences are entirely my own, summarizing parts of the argument that wouldn’t fit into a short blog post.

I wrote this weird edited pastiche/summary because I couldn’t figure out how else to express my frustration at Just Giving. The book does not conclude that philanthropy is bad. In the end, it comes out saying that philanthropy is potentially okay and can serve a useful purpose, although the tax incentives around it are weird and should be structured better. But along the way it manages to darkly hint that philanthropy is bad about two hundred times on every page. Nothing in the book is wrong. But a lot of the right things in it are fully general counterarguments that demand charity display a level of rigor that nothing else has. And the author’s interviews and summaries mostly keep the dark hinting while watering down the “actually it is okay” part so much it becomes almost invisible. The resulting style could be used to condemn not just charity but any productive human activity, including the writing of Just Giving itself.

For example: charity is not just an activity that takes place in a void. It takes place in a human society. So far, so good – nothing takes place in a void, except maybe space travel. But the book manages to darkly hint that because this is true, any regulation on it is justified. It never says this. In the end it doesn’t even want to regulate charity. But if you started feeling creeped out by sentences above like “books should not be understood in the simplistic manner of an activity that takes place within a framework of nonintervention by the state, or as nothing more than private individual decisions to express thoughts” or “writing a mediocre book squanders assets that are partially the public’s” – if you started thinking “Wait, is he pushing totalitarianism?” – well, both of those are pretty direct quotes from Just Giving, and the originals gave me the same level of unease about charity.

Or: it’s true that there’s a sense in which if the state gives someone a tax deduction for something, it is subsidizing their activity. And it’s true that authors can deduct some of their book-writing expenses from their tax bill. But it seems troubling to go from there to calling book-writing “part of the submerged state, obscured from public view and accountability”, or to say that now “the state is partially complicit in the harm caused by bad books”. Yet both of these are real Just Giving sentences too. I find myself much less attached to the tax deduction for authors’ business expenses (which may or may not be useful, no strong opinions) than to the project of preventing people from saying things like “Making sure books are good is kind of the responsibility of the state, isn’t it?”

Or: it’s true that authors just write whatever they want. You could describe this as making them “unaccountable and nontransparent”, and “at odds with democracy”. But at some point you might think things like “Wait a second, isn’t democracy perfectly compatible with private individuals doing their own thing? Are you sure you’re not thinking of totalitarianism?” Normally I would add something like “…and these considerations become immediately apparently when we’re talking about writing books, which makes this a classic case of proving too much“, except that to me they also become immediately apparent when we’re talking about philanthropy, so there must just be some fundamental disconnect going on here.

In a few sections, I “cheated” by using Just Giving‘s sentences or paragraphs about charitable foundations, rather than philanthropy in general. Reich is not necessarily worried about every charitable donation making “the dead hand of the donor potentially [extend] from beyond the grave to strangle future generations” (yes, this is a real quotation from the book), only about donations from foundations doing that. Still, might this be a little dramatic? Reich treats it as self-evident that permanent foundations are bizarre, maybe literally the most bizarre thing, quoting John Stuart Mill’s opinion that charitable foundations are “among the grossest and most conspicuous abuses of the time” and that the necessity of banning foundations that outlast their founder’s lifetime is “so obvious that he can scarcely conceive how any earnest inquirer could think otherwise”. Unfortunately for Mill, this is not at all obvious to me, and I was left baffled on this point which the book kind of assumed to be a natural human instinct. Why should my ability to control my donations be limited by something as random as my lifespan? If Bill Gates happens to get hit by a truck tomorrow, does this coincidence have some sort of important moral bearing on how the Gates Foundation’s money should be spent? If we decree it does, this leads to odd conclusions, like that the most important effective altruist cause in the world is encasing Bill Gates in an impenetrable steel shell so that nothing can possibly harm him – do we endorse this use of resources? If we oppose foundations, is Bill Gates still allowed to leave all his money to the single person in the world most aligned with his values, and then hope really hard that this person doesn’t betray him? Isn’t part of the point of law to abstract out things like “people might betray you” and replace them with comfortable ironclad contracts?

(a confession: my point about books being a perpetual exercise of the author’s power in the same way foundations are a perpetual exercise of the founder’s power is unfair, and addressed by Reich in the book. He states that most permanent things wield power only as long as the living choose to humor them – eg a dead author only matters if living people choose to read them and take their advice to heart – but foundations do not need the support of the living as long as the contracts that create them remain enforced).

There is much to like about Just Giving. Its breakdown of where charitable dollars actually go (mostly to religious institutions if you’re poor, mostly to colleges and museums if you’re rich) contains data I’ve been looking for a long time, and rightly points out that we should do better. Its discussion of the way tax deductions interact with wealth is interesting, although not obviously more applicable to charity than to book-writing or anything else. Its conclusion – that charity and philanthropic foundations have an important role in diversifying the range of represented interests and experimenting with new social policy – seems dead right, and matches my own thoughts on the subject (and see also this article by Kelsey Piper). I really can’t disagree with this book too much on the object level.

And yet if my review sounds scathing, I hope this is a sort of justice. Rob Reich has limited disagreements with charity on the object level, but still manages to write what sounds like a scathing review of it. I think this is bad.

In conclusion, Just Giving is a government-subsidized exercise of plutocratic power and plutocratic voice repugnant to the very idea of a democratic society of equals. I hope this gets corrected in any future editions.

[EDIT: Professor Reich responds in the comments. Please be polite if you try to discuss this with him. Also, please stop mistaking him for former Labor Secretary Robert Reich, they are two different people.]

# Book Review: The Seven Principles For Making Marriage Work

I.

John Gottman is a legendary figure, and the legend is told best by John Gottman. He describes wading into the field of marital counseling as a young psychology postdoc, only to find it was a total mess:

When we began our research, the wide range of marital therapies based on conflict resolution shared a very high level of relapse. In fact, the best of this type of marital therapy, conducted by Neil Jacobson, had only a 35 to 50 percent success rate. In other words, his own studies showed that only 35 to 50 percent of couples saw a meaningful improvement in their marriages as a result of the therapy. A year later, less than half of that group — or just 18 to 25 percent of all couples who entered therapy — retained these benefits. A while ago, Consumer Reports surveyed a large sample of its members on their experience with all kinds of psychotherapists. Most therapists got very high customer-satisfaction marks—except for the marital ones, who received very poor ratings. Though this survey did not qualify as rigorous scientific research, it confirmed what most professionals in the field already knew: in the long run, marital therapy did not benefit the majority of couples.

Gottman decided the field needed statistical rigor, and that he – a former MIT math major – was exactly the guy to enforce it. He set up a model apartment in his University of Washington research center – affectionately called “the Love Lab”, and invited hundreds of couples to spend a few days there – observed, videotaped, and attached to electrodes collecting information on every detail of their physiology. While at the lab, the couples went through their ordinary lives. They experienced love, hatred, romantic dinners, screaming matches, and occasionally self-transformation. Then Gottman monitored them for years, seeing who made things work and who got divorced. Did you know that if a husband fails to acknowledge his wife’s feelings during an argument, there is an 81% chance it will damage the marriage? Or that 69% of marital conflicts are about long-term problems rather than specific situations? John Gottman knows all of this and much, much more.

Using his mountain of data (the legend continues) Gottman became a Divorce Prophet:

After years of research…I am now able to predict whether a couple will stay happily together or lose their way. I can make this prediction after listening to the couple interact in our Love Lab for as little as five minutes! My accuracy rate in these predictions averages 91 percent over three separate studies. In other words, in 91 percent of the cases where I have predicted that a couple’s marriage would eventually fail or succeed, time has proven me right. These predictions are not based on my intuition or preconceived notions of what marriage “should” be, but on the data I’ve accumulated over years of study.

…which is pretty interesting. But predicting destiny is only an intermediate step – as another legend once said, “the point is to change it”. So, science in hand, John Gottman resolved to fix marital counseling. And apparently succeeded:

We found that at the beginning of our workshops, 27 percent of couples were at very high risk for divorce. At our three-month followup that proportion was 6.7 percent and at nine months it was 0 percent. But even couples who were not at high risk for divorce were significantly helped by the workshops.

Twenty years later, the legend has spread to every corner of the world. He has received glowing praise from The New York Times, The Atlantic, BBC, CNN, Washington Post, The New York Times again, Harvard Business Review, Scientific American, Time, and The New York Times a third time. He has published over two hundred scientific papers, some of which have been cited thousands of times. He has been voted one of the top 10 most influential therapists of the past quarter-century. His Gottman Relationship Institute, founded together with his wife Julie Gottman, has become a marriage counseling empire, trained hundreds of therapists in the Gottman method, and operates a referral network that can find you Certified Gottman Level 3 Therapists from Australia to Uruguay. After a long life of helping save countless marriages, his one regret is that he is so great he can no longer find an adequate control group for his studies:

When we sat down to write the first edition of this book, we were excited to share the results of laboratory research into relationships but we knew we’d face some skepticism. Could scientific study of something as intangible, idiosyncratic, and personal as romantic love deliver useful advice to couples in the real world? Well, more than fifteen years and millions of readers later we are happy to report that The Seven Principles for Making Marriage Work has done just that. Countless readers across the globe tell us that the book’s strategies have enhanced, shielded, or saved their relationship. We have received thank-yous from every imaginable type of couple, including newlyweds, traditional spouses, two-career partners, devoutly religious spouses, military couples, cohabitants, same-sex partners, not-yet-marrieds, divorced people looking toward the future, and counselors who work with all of the above.

It is a great source of satisfaction and pride that we have been able to help so many people. We’re also gratified that research continues to confirm what these readers consistently tell us: The Seven Principles can have a powerfully positive effect on your relationship. In fact, a randomized clinical study by John and his coresearchers (Julia Babcock, Kim Ryan, and Julie Gottman) found that married couples who simply read The Seven Principles and worked through the quizzes and exercises on their own (but received no additional professional aid) were significantly happier in their relationship, and these effects lasted when assessed a year later. Simply reading this book proved so successful that it actually bollixed the research: the original experiment had been designed to use these “book-only” couples as a control group to test marital therapy techniques!

Sounds like some book! And God knows we need good marriage therapy. The world needs it. And I need it in particular. I am a psychiatrist. I am trained to treat depression and schizophrenia and nice simple things like that. But somehow, I keep getting patients who need help with relationship problems. I am totally unprepared for this. In the past, my advice has been “go find someone trained by John Gottman, I hear he is some kind of living legend”. But at some point, I figured I should finally read his book, The Seven Principles For Making Marriage Work: A Practical Guide From The Country’s Foremost Relationship Expert, so that I could have an informed opinion on this and maybe try helping people directly.

This is a review of that book. It starts with a summary of Gottman’s marital counseling techniques, continues to a discussion of the evidence for and against them, and ends with some random thoughts about marriage.

II.

The secret to a happy marriage is that you should like your spouse.

Maybe this doesn’t sound especially secret, when put that way. But part of the Gottman legend is that the old school of marriage therapists kind of missed that part. They were really into solving conflicts and having good communication skills and things like that. But over his years of monitoring hundreds of couples in real-world situations, Gottman found that this was overrated. Plenty of couples had atrocious communication skills and got in conflicts all the time, but loved each other very much and had no real marital problems. Plenty of other couples had finely-polished communication skills and always used “I statements” and things like that, and still ended up divorced. Communication skills are good, and you should definitely try to have them, but you’re putting the cart before the horse unless you focus on liking your spouse.

I totally believe this. I remember my grandparents used to fight all the time. Any time my grandmother said something, my grandfather would disagree with her, and vice versa, and the ensuing argument would (to my young ears) sound absolutely vicious, and then they would laugh it off and forget about it and continue being wonderfully and obviously in love. So fine, focus on liking your spouse. But how do you make that happen?

Seven Principles excels in its selection of worksheets, activities, and games. There is a seven-week program of talking about one nice thing with your spouse every day, a “Building Your Love Maps” game, and a list of questions you and your spouse should answer together one by one. The underlying principle seems to be that to know somebody is to love them. If you get a really good mental model of where your spouse is in their life, their hopes, their dreams, their fears, their daily toil, then it’s hard to not have at least some fondness for them. One activity is a quiz, with questions like:

1. I can name my partner’s best friends.  
2. I can tell you what stresses my partner is currently facing.  
3. I know the names of some of the people who have been irritating  
my partner lately.  
4. I can tell you some of my partner’s life dreams.  
5. I am very familiar with my partner’s religious beliefs and ideas.  
6. I can tell you about my partner’s basic philosophy of life.  
7. I can list the relatives my partner likes the least.  
8. I know my partner’s favorite music.  
9. I can list my partner’s three favorite movies.

The more you know, the more likely your marriage is to make it.

Along with knowing the big things, you should also know what they’re doing on a day-to-day basis. You should conspicuously make sure to know it. Apart from whatever other exercise you’re doing each day, Gottman recommends a ritual of checking in after work and exchanging stories about your days. This time is a Designated Support Zone, no criticism allowed. You take your spouse’s side whether you secretly disagree with them or not. If your spouse gets angry that a police officer gave them a ticket for driving 110 mph through a 25 mph school zone, you are obligated by the terms of your marriage contract to shake your head and say “I know, cops these days have no respect.”

Gottman is slightly less strict in other situations, but he still thinks it’s very important that you take your spouse’s side in conflicts. He especially highlights a common dynamic where your parents are always trying to cause trouble between you and your spouse, and your marriage will be in danger until you commit to side with your spouse whenever this kind of thing comes up:

At the core of the tension is a turf battle between [the wife and the mother] for the husband’s love. The wife is watching to see whether her husband backs her or his mother. She is wondering, “Which family are you really in?” Often the mother is asking the same question. The man, for his part, just wishes the two women could get along better. He loves them both and does not want to have to choose. The whole idea is ridiculous to him. After all, he has loyalties to each, and he must honor and respect both. Unfortunately, this attitude often throws him into the role of peacemaker or mediator, which invariably makes the situation worse.

The only way out of this dilemma is for the husband to side with his wife against his mother. Although this may sound harsh, remember that one of the basic tasks of a marriage is to establish a sense of “we-ness” between husband and wife. So the husband must let his mother know that his wife does indeed come first. His house is his and his wife’s house, not his mother’s. He is a husband first, then a son. This is not a pleasant position to take. His mother’s feelings may be hurt. But eventually she will probably adjust to the reality that her son’s family unit, where he is the husband, takes precedence to him over all others. It is absolutely critical for the marriage that the husband be firm about this, even if he feels unfairly put upon and even if his mother cannot accept the new reality.

This is not to suggest that a man do anything that he feels demeans and dishonors his parents or goes against his basic values. He should not compromise who he is. But he has to stand with his wife and not in the middle

A final method of making yourself like your spouse: just explicitly and consciously focus on their good qualities. One of the worksheets lists a bunch of good qualities and asks you to pick some you appreciate in your spouse and explain why. Another asks you to reminisce about the old days, on the theory that presumably you liked your spouse back when you decided to marry them. This isn’t always true – Gottman finds that couples who are really far gone will export their present hostility back into the past and talk about how they always knew it was a bad idea. With typical statistical precision, he notes that “94 percent of the time that couples who put a positive spin on their marriage’s history are likely to have a happy future as well. When happy memories are distorted, it’s a sign that the marriage needs help.” But it’s usually true enough to get both partners warming up to each other a little.

But okay. You’ve done all this stuff and you like your spouse at least a little. Now what?

Now you can start learning communication skills.

Gottman’s communication skills work focuses on what he calls “The Four Horsemen Of The Apocalypse”: criticism, contempt, defensiveness, and stonewalling. These may seem obvious, though Gottman takes some of them in non-obvious directions – “couples who are contemptuous of each other are more likely to suffer from infectious illnesses (colds, flu, and so on) than other people”. Avoiding the four horsemen is vital, especially in the crucial “startup” phase of a conversation. How do you do it?

Some tips are very small in scale. Use I-Statements like “I have trouble dealing with how messy things are sometimes” instead of You-Statements like “You never clean up around here”. Be polite. Try to make “repair attempts” – conflict-ending attempts that can be anywhere from “Hey, you’re yelling at me” to “You are right and I am wrong and I am sorry” to “Oh my god, listen to us, let’s get a drink and never speak of this matter again”. Accept your spouse’s repair attempts when offered.

Others are more general. Don’t have a fight when you’re physiologically aroused. Don’t have a fight when you’re physiologically aroused. Monitor your physiological arousal, and if you start to notice the signs – fast heartbeat, tense muscles, shallow breathing – call a time-out, go somewhere else, and use meditation or deep breathing or whatever to calm down.

This is especially important for men. Gottman has strong opinions on gender. He uncritically accepts the feminist view that men feel entitled because of patriarchy and that if they feel angry or upset it’s probably just their entitlement flaring up again. He flirts with saying that men should generally yield to their wife in a conflict (presumably because, thanks to patriarchy, everything will always be biased in favor of the man and so the wife is usually right). In the end, he softens this to a statement that men should “accept influence from” their wives, but also heavily implying that a man who doesn’t give in to his wife must not be accepting her influence – for example, an exercise on page 118 asks men to describe how they would accept their wife’s influence in various situations, and includes an answer key where the right answer is always to say she is right and do what she wants. You might object to this, but sorry, it is Evidence Based According To Science. Gottman tells us that only 35% of husbands are emotionally intelligent, and that “when a man is not willing to share influence with his partner, there is an 81% chance that his marriage will self-destruct”.

But Gottman’s most controversial views on gender involve physiology. He proposes an evo psych explanation for why men can’t handle talking about problems:

This is not because of some lack on the man’s part. The reason lies in our evolutionary heritage. Anthropological evidence suggests that we evolved from hominids whose lives were circumscribed by very rigid gender roles, since these were advantageous to survival in a harsh environment. The females specialized in nurturing children while the males specialized in cooperative hunting.

As any nursing mother can tell you, the amount of milk you produce is affected by how relaxed you feel, which is related to the release of the hormone oxytocin in the brain. So natural selection would favor a female who could quickly soothe herself and calm down after feeling stressed. Her ability to remain composed could enhance her children’s chances of survival by optimizing the amount of nutrition they received. But in the male natural selection would reward the opposite response. For these early cooperative hunters, maintaining vigilance was a key survival skill. So males whose adrenaline kicked in quite readily and who did not calm down so easily were more likely to survive and procreate. To this day, the male cardiovascular system remains more reactive than the female and slower to recover from stress […]

This gender difference in how physiologically reactive our bodies are also influences what men and women tend to think about when they experience marital stress. As part of some experiments, we ask couples to watch themselves arguing on tape and then tell us what they were thinking when our sensors detected they were flooded. Their answers suggest that men have a greater tendency to have negative thoughts that maintain their distress, while women are more likely to think soothing thoughts that help them calm down and be conciliatory. Men, generally, either think about how righteous and indignant they feel (“I’m going to get even,” “I don’t have to take this”), which tends to lead to contempt or belligerence. Or they think about themselves as an innocent victim of their wife’s wrath or complaint (“Why is she always blaming me?”), which leads to defensiveness.

Obviously these rules don’t hold for every male and every female. But after twenty-five years of research, I have noted that the majority of couples do follow these gender differences in physiological and psychological reactions to stress. Because of these dissimilarities, most marriages (including healthy, happy ones) follow a comparable pattern of conflict in which the wife, who is constitutionally better able to handle the stress, brings up sensitive issues. The husband, who is not as able to cope with it, will attempt to avoid getting into the subject. He may become defensive and stonewall. Or he may even become belligerent or contemptuous in an attempt to silence her.

The problem is, men are just too flighty and emotional! They need a rational, hard-headed woman to take care of them and keep them grounded!

Sorry. To be more serious, he thinks that because women are emotionally stable and men aren’t, women tend to bring up long-standing problems that need to be solved, and because men can’t handle this level of stress they panic or shut down or blow up or otherwise start a conflict in order to avoid having to deal with it. “More than 80 percent of the time it’s the wife who brings up sticky marital issues, while the husband tries to avoid discussing them.” The solution is for men to learn calming techniques so they don’t stop the conversation as quickly.

Okay, so now you like your spouse and you know how to communicate. Now you can prepare to actually solve some conflicts.

But not all of them. Gottman divides conflicts into two types: solvable and unsolvable. Solvable conflicts are simple, specific, and about the thing they seem to be about – for example, the husband is supposed to take the trash out after work, but work has gotten really stressful lately and he keeps forgetting, and now the trash is overflowing and the wife is annoyed. The solution here is to use normal problem-solving techniques. Put a sign in the bedroom saying “DID YOU REMEMBER TO TAKE THE TRASH OUT?” or something. Whatever.

Unsolvable conflicts are temporary manifestations of deep psychological issues. The particular thing that sparked the fight this time is irrelevant, but both spouses will fight to the death because it represents something important. For example, the husband is late to dinner one night because he went out to the bar. The wife yells at him and says he doesn’t care about her. He yells back that she’s a control freak. Here the problem will not be solved by coming up with a compromise where he can go to the bar half of nights. The problem is that she secretly worries his drinking buddies have a closer connection to him than she ever will, that he doesn’t love her anymore, that he goes to the bar to escape her. He worries that he’s lost his freedom, that he’s become emasculated, that he’s become some boring old person who is never allowed to have fun. If the bar burnt down tomorrow, they would find some other excuse to fight over this dynamic.

There’s no hard line between solvable and unsolvable conflicts. One couple’s solvable problem might be another’s unsolvable one. Forgetting to take the trash out becomes an unsolvable problem if it represents how the husband is irresponsible or the wife is too controlling. Staying out too late at the bar is a solveable problem if it’s just that dinner is getting cold and neither of them has any problem with the husband eating leftovers. But usually couples can figure out whether their particular issue taps into deeper roots.

Gottman suggests dealing with unsolvable conflicts by making the underlying “dreams” explicit. He recommends both partners talk about what the dream driving their side of the conflict means to them. So for example, the husband who stays late at the bar might say “When I was young, my dad was so poor he had to work twelve hour days. Then he would come home, do some chores, go to sleep, and start all over again. I told myself that if I ever ended up like that, I might as well just die, because it seemed like such a crappy and joyless life. To me, getting to go have fun with people means that I’m successful enough that I don’t have to end up like my dad.” While one spouse (the “dreamer”) is describing this, the other spouse has to be completely supportive and try to understand them, without pointing out ways they’re wrong.

Once both partners feel like they’ve been heard and that they understand each other, they discuss the absolute minimum they would need in order to feel like their dreams were being respected, versus the beyond-minimum things that they’re willing to be flexible on. Then they both agree to a compromise that gives both of them their bare minimum and splits the difference on the flexible parts. What if the two bare minima are mutually exclusive? The book doesn’t say. Probably you got a defective dream, and you should go to the dream factory and ask if they take returns.

The last part of the book is maybe the least actionable, but also my favorite. It discusses couples as almost miniature cultures, with their own rituals, in-jokes, ideologies, and systems of meaning. Partly it’s about how to create these things, partly it’s about how to acknowledge these things, and partly it’s just John Gottman’s love letter to the concept of couplehood. It’s really heart-warming. Don’t make the same mistake I did and read it when you’re feeling lonely.

III.

Okay, but can John Gottman really predict divorce with 91% success rate and do all the other things he says he can do? Haha, no. All of that stuff is totally false.

Richard Heyman published the definitive paper on this in 2001, The Hazards Of Predicting Divorce Without Crossvalidation (kudos to Laurie Abraham of Slate, the only one of the journalists covering Gottman to find and mention this, and my source for some of the following). Heyman notes that Gottman doesn’t predict divorce at all. He postdicts it. He gets 100 (or however many) couples, sees how many divorced, and then finds a set of factors that explain what happened.

Confused about the difference between prediction and postdiction? It’s a confusing concept, but let me give an example, loosely based on this Wikipedia article. The following rule accurately matches the results of every US presidential election since 1932: the incumbent party will win the election if and only if the Washington Redskins won their last home game before the election – unless the incumbent is black or the challenger attended a Central European boarding school, in which case it will lose.

In common language, we might say that this rule “predicts” the last 22 presidential elections, in the sense that knowing the rule and the Redskins’ record, we can generate the presidential winners. But really it doesn’t predict anything – there’s no reason to think any future presidential elections will follow the rule. It’s just somebody looking to see what things coincidentally matched information that we already have. This is properly called postdiction – finding rules that describe things we already know.

Postdictive ability often implies predictive ability. If I read over hospital records and find that only immunodeficient people caught a certain virus, I might conclude I’ve found a natural law – the virus only infects immunodeficient people – and predict that the pattern will continue in the future.

But this isn’t always true. Sometimes, especially when you’re using small datasets with lots of variables, you get predictive rules that work very well, not because they describe natural laws, but just by coincidence. It’s coincidence that the Redskins’ win-loss record matches presidential elections, and with n = 22 datapoints, you’re almost certain to get some coincidences like that.

Even an honest attempt to use plausible variables to postdict a large dataset will give you a prediction rule that’s a combination of real natural law and spurious coincidence. So you’re not allowed to claim a certain specific level of predictive ability until you’ve used your rule to predict out-of-training-data events. Gottman didn’t do this.

In his paper, Heyman creates a divorce prediction algorithm out of basic demographic data: husband and wife’s education level, employment status, etc. He is able to achieve 90% predictive success on the training data – nearly identical to Gottman’s 91% – without any of Gottman’s hard work. No making the couples spend days in a laboratory and counting up how many times they use I-statements. No monitoring their blood pressure as they gaze into each other’s eyes. Heyman never met any of his couples at all, let along analyzed their interaction patterns. And he did just as well as Gottman did at predicting divorce (technically he predicted low scores on a measure of marital stability; his dataset did not include divorce outcomes).

Then he applied his prediction rule to out-of-sample couples. Accuracy dropped to 70%. We have no reason not to think Gottman’s accuracy would drop at the same rate. But 70% is around the accuracy you get if you predict nobody will divorce. It’s little better than chance, and all of Gottman’s claims to be a master divorce predictor are totally baseless.

The first question on Gottman’s FAQ is whether he is doing this. He says he did this once, but that 6 of his 7 studies have been properly predictive. But Wikipedia notes that this claim uses “a non-standard definition of prediction in which all that is required is that predictive variables, but not their specific relationship to the outcome, were selected in advance”. Heyman and Abraham specifically criticize the 6 studies that Gottman calls genuinely predictive as being postdictive. I cannot find all the relevant studies, as many of them are in books, but contra the FAQ it looks as if he is still postdicting.

It’s hard for me to dismiss this as an honest mistake. Gottman constantly plays up his credentials as a mathematician and statistician, saying that:

In the beginning, Dr. John Gottman’s research was devoted to the discovery of reliable patterns in observational data. He wanted to see if there were indeed patterns of behavior, or sequences of interactions, that could discriminate happy from unhappy couples. It was not at all clear that these patterns existed. Dr. Gottman and his colleagues began developing the math for sequential analysis, which now is a well-developed methodology.

Anyone smart enough to invent new mathematical methodologies should also be smart enough to know you can’t validate your predictor on its training data, so he must know exactly what he’s doing. And it would be so easy to fix if he wanted to! All he needs to do is take one of his predictors, apply it to data that wasn’t its training data, and tell us how it does! How could this be an innocent mistake‽

And speaking of things that an MIT mathematician should know better than to do, what’s up with claims like “when a man is unwilling to share power with his partner, there is an 81% chance his marriage will self-destruct”. Obviously you can replace 81% with any number you want by operationalizing “unwilling to share power with his partner” differently. This sentence as written is totally meaningless. It could potentially be part of a good study, where Gottman investigated different forms of power-sharing and how they affected marital stability – but only because the study would carefully explain its methods so that the end number meant something. Outside the context of that number, it communicates nothing. Yet Gottman put it in a book and expects us to be impressed by it. It feels like he’s just trying to dazzle us with mathematical precision, and hoping we don’t think about it long enough to realize that it doesn’t make sense.

But the most important statistical question is – does Gottman marriage counseling work? Gottman cites a bunch of different studies proving that it does, but he conducted or oversaw all of them. I randomly chose Shapiro and Gottman (2005) for further analysis. It is typical of this genre: it takes couples who have just had a baby (a particularly perilous time in a marriage), gives half of them a Gottman workshop, and the other half get nothing (no placebo or alternate method here!). The study evaluates husbands and wives separately (was this a preregistered decision?) on several established marital tests and subscales of marital tests (was the choice to keep some tests whole and take subscales of others preregistered?) and finds a significant quadratic effect of their program on marital quality.

I had to look up what a quadratic effect is. It means that the effect approximates a parabola. That is, things start by getting worse, then get better again, in a parabolic pattern. If I’m understanding the study correctly, then if you just do normal linear tests (did the couples’ marriage get better or worse, overall?) there was no significant effect. Gottman writes that:

Our best guess as to why the quadratic pattern occurs is the following. With intervention, things get worse at first because the immediate effects of the interventions are to increase the amount of conflict that the couple experiences. Our interventions encourage couples to honestly face and discuss their conflicts, particularly potential or actual inequities in housework and childcare, conflicts that they would naturally avoid and which they do avoid in the control group. This early increase in the amount of conflict in our intervention group probably causes temporary discomfort, which is reflected in lowered marital quality and increased postpartum depression. However, because the couples in the intervention group learn the communicative skills to cope with these issues, the conflicts get dealt with to some degree, and thereafter marital quality and postpartum depression both improve, whereas in the control group, because these conflicts have not been dealt with, things get worse over time.

This is fine, but the abstract just sums it up as “results showed that, in general, the preventive intervention using a psycho-communicative-educational format was effective”. No! If you didn’t preregister that you expected a quadratic effect, you are lost in the Garden of Forking Paths and you should give up and start over. Also, I feel like probably Gottman advertises to his clients that their marital happiness will improve. But even if we accept his argument here, the only thing he can say with confidence is that their marital happiness will increasingly approximate a parabola over time, which is not really what I think most people go to therapists looking for.

I might be misinterpreting this, and maybe I’m being overly harsh. But I am predisposed to be overly harsh because the whole “predicting divorce” thing makes me think Gottman is out to get me, and so I am less forgiving of unusual polynomials than I might be otherwise. Also, if you’re running the study of your own method, you ought to be on extra good behavior, and this does not really seem all that extra.

What happens when people who aren’t Gottman evaluate the Gottman method? A large government-funded multicenter study testing a Gottman curriculum as well as several others found no effect of any on marital outcomes; control couples actually stayed together slightly more than ones who got marriage counseling. The Gottman curriculum seemed to do worst of the three curricula studied, although there were no statistical tests performed to prove it. I have no explanation for this. Maybe the parabola is just really big, and the divorce is the low point of the parabola, and later on they’ll end up super-double-married. But I am not optimistic.

I don’t want to be too harsh on Gottman here. Rigorous psychology studies are murderous. Things that we know basically have to work, like Alcoholics Anonymous and SSRIs and psychotherapy in general, end up showing no or minimal effects. Heck, zoom out a little bit and we have twin studies showing that parenting itself, in full generality, has no or minimal effect. I find all of this very suspicious, and it would not surprise me if there’s something really wrong here that makes studies biased towards false negatives. All of this stuff about learning to respect and appreciate your spouse and negotiate conflicts in a calm loving way seems like the sort of thing that should work, and for all we know it might work in some population or situation other than the ones being studied.

But for the guy whose whole legend centers around how he’s evidence-based, it’s not a good look.

IV.

I really wanted Gottman Marital Counseling to work.

I wear a psychiatrist hat and a therapist hat. I love the psychiatrist hat. It is blue and pointy and has little glowing stars and moons on it. When I wear it, then with sufficient knowledge and understanding I can give people substances that release obsessions, calm fears, and brighten sorrows. Sometimes I can help people solve their unbearable hopeless problems, and it’s the best feeling in the world.

I hate wearing the therapist hat. I put it on as rarely as possible. I don’t advertise myself as a therapist, and if people ask me to therapy them, I try to refer them to someone else. But if someone wants to talk about their problems in a session, you can’t just say no. And so they tell me about being trapped in an abusive relationship, or haunted by guilt, or trapped in a dead end job with no prospects for improvement. And then they expect me to be able to say something that makes it all better. I know that the textbook response is something about how therapy does not solve problems per se, but by sharing them with someone else it makes them more bearable and adds perspective. Unfortunately, my patients didn’t read that textbook, and they put hope in me, and as often as not I betray it.

I think every therapist feels this way. I once talked to an important professor of therapy, who admitted to me something like “even at my stage, I feel like in the end we only have five or so techniques”, and I got really excited and blurted out “wait, what are the other two?!”

Desperation breeds gullibility. Patients with terminal diseases, however smart they used to be, turn to homeopaths and charlatans rather than face the dismal truth. Therapists are desperate – being confronted with some of the most sympathetic people in the world, day in and day out, having the burden of helping them placed on your shoulders, and knowing that your armamentarium isn’t up to the task will do that to you. And so they become marks. Gullible, gullible marks. Realistically it’s going to be really hard for me to stop recommending Gottman marital counseling to people, because they need something so much. And this is something. And it sounds so good. And I can’t just let their marriage keep falling apart. And surely there’s still a chance it might work, right?

(don’t worry, eventually I’m going to look into some of those forms of counseling that outperformed it in that study)

Bad marriages are so, so bad. They’re so bad it’s shocking. The first time I saw one, I didn’t know what to do. I wanted to figure out some of the root issues, so I asked my patient something about his wife, and he immediately launched into a tirade about all the things his wife had done recently, and why she was in the wrong and he was in the right. I tried to redirect him, and briefly succeeded, but after a second or two the new line of conversation shifted to how unfair his wife was and how she was in the wrong about everything. I wanted to grab him by the shoulders and shake him and say “for the love of God, just put this aside and move to the meta-level and let’s talk about some of the places we can go from here!” but he was just incapable of this mental action. I wish I could say this was an isolated case. It isn’t.

My ex-girlfriend Ozy writes a relationship advice column. Probably taking relationship advice from an ex-girlfriend is some kind of classic mistake, but I read it anyway. They describe five kinds of relationship problems – stupid problems, basic incompatibilities, problems that are actually a different kind of problem, terrible people, and horrifying soul-sucking messes. For some reason, this taxonomy has stuck with me when all the supposedly evidence-based taxonomies I hear the social workers talk about have failed. And the horrifying soul-sucking mess category sticks with me most of all:

A problem of one of the previous three types was badly managed, perhaps for years. Now, every time you have a minor argument, you bring in everything wrong that happened for your entire relationship. You don’t feel like you can trust your partner. All the quirks you used to find charming drive you up the wall. You hate even your partner’s most innocuous actions. You avoid every topic that leads to a fight, and rapidly find that you can’t discuss anything except Marvel movies and the weather. You’re defensive whenever your partner says anything that sounds like even a minor criticism. You’re sarcastic and you call them names. Somehow, when you remember good things about the past– the time you saw Hamilton together or your birthday present or being the best man at their wedding– all you can remember is the long lines at intermission, the poor wrapping job, and their incredibly rude drunk aunt. If asked to name a good trait of theirs, you draw a blank, but you can go on for hours about their flaws.

I guess it might be in theory possible to fix a horrifying soul-sucking mess with a lot of hard work, but to be honest every time I’ve seen a person in one of those relationships they were a lot better and happier and stronger as people as soon as they ended it.

A lot of my patients are horrifying soul-sucking messes. I wish there was something I could do about it, but instead I just sit and listen as they spend forty-five minutes describing every way their spouse has wronged them.

I’m terrified of this. How did it happen? At some point these people must have loved each other. How does any human relationship get this bad? Could it happen to me? Could I marry a great person who I love a lot, and then five years later sign up for therapy just so I can start talking about all my grievances without letting the therapist get a word in edgewise?

Unlike my ex, I don’t write a relationship advice blog. I write a blog about other things. One of them is politics. And whenever I hear people talk about relationships, I hear weird echoes of political problems. People who hate their spouse have an outgroup of one. A unified polity has devolved into partisanship. Social trust has been broken; a defect-defect equilibrium is in place. Gottman thinks of couples as a two-person culture, and some of those cultures are decadent and fractious.

Theodore Adorno’s right-wing authoritarianism scale asked a lot of questions about marriage and child-rearing; his thesis was that people who want top-down government will control their families the same way. Certainly there are authoritarian marriages. But it also seems like there are marriages that are nationalist in a more positive sense – one where the couple has built itself a strong mutual culture and identity, subsumed both individuals into it, and come out stronger on the other side. There are liberal marriages, where both spouses do their own thing, occasionally come together for mutually beneficial exchange of affection, and then go back to doing their own thing. There are even social justice marriages, where both partners are obsessed with how they are being oppressed by the other, interpret all discussion of compromise as hostile attempts to excuse the oppression, and have no strategy beyond proclaiming their victimization louder and louder in the hopes that their grievances are recognized.

…I’m making fun of that last one, but maybe they have a good point. Gottman’s marital counseling – and every other kind of marital counseling I’ve read – is basically mistake theoretic. It assumes that two decent people who both want to live with each other are unable to because they don’t have communication skills or problem-solving skills or some other skill that lets them fulfill their mutual aims. That model seems to accurately describe most of my patients. But Ozy’s taxonomy is more thorough, including a category for terrible people:

One or both of you just sucks. This category includes abuse, but it certainly isn’t limited to it. This is the category from which advice columnists get all their pageviews: we love viewing a train wreck. Your partner has had a suicidal crisis every night at 3am for the last month, and you’re up all night comforting them, and they refuse to find anyone else to talk to or ride out their suicidal crisis on their own. Your partner cheats on you constantly. Your baby is eighteen months old and your partner has never changed a single diaper. Your partner has demanded that you keep your relationship secret from everyone. You asked your partner to clean up dog poo from the floor, and it is three days later, and the dog poo is still there. If you have found yourself in a committed relationship with a terrible person, you should DUMP THEM.

In marriage, as in other forms of politics, sometimes exit > voice. Which is probably not something marriage counselors want to think about very much.

I’m a liberal, and my advice usually follows liberal principles. Have meetings involving all stakeholders, agree on general rules based on deep principles, and then follow the rules even when it’s hard, since that ensures a minimum of conflict on specific issues. I tell patients to have a consistent weekly meeting with their spouse, maybe a Saturday date night. During the meeting, each of them has to say one nice thing about the other, and then one problem that they want to work on. The other person thanks them for the nice thing and then they brainstorm a solution to the problem. If they both agree to the solution, they have to stick with it at least until the next relationship meeting. A lot of my patients have said this really helps them, have continued the meetings long after the immediate crisis passes, and probably expect the technique was invented by some clever person like John Gottman. I will never tell them that actually I picked it up from a different ex-girlfriend, not even the one who writes relationship advice columns.

In conclusion, I have no idea what makes marriages work, and I am not convinced John Gottman does either. Sometimes marriages are horrible in ways I never could have imagined, and other times they are amazing and I am infinitely jealous of them. Also, the very existence of a next generation of the human race is dependent on people having them and making them work for nonzero periods of time, which is a pretty terrifying prospect. Honestly I’m surprised we’ve lasted this long.

# Socratic Grilling

Imagine an kid in school first hearing about germ theory. The conversation might go something like this:

Teacher: Many diseases like the common cold are spread by germs, when one infected person contacts another.

Student: But I got a cold a few weeks ago, and I never touch anyone except my family members. And none of them were sick.

Teacher: You don’t need to actually touch someone. Sometimes it can spread through mucus droplets in the air.

Student: And one time I was camping in the woods for a month, and then I got a cold, even though I hadn’t been around anybody.

Teacher: If it was spring, you might have gotten allergies. Allergies can feel a lot like a cold, but they aren’t spread by germs.

Student: It was fall.

Teacher: Then maybe it was an unusual allergy, or some other condition.

Student: Hey, wait. If germs are spread from person to person on touch, why doesn’t the government just mandate one week when nobody is allowed to touch anyone else? Then all the germs will die and we’ll never have to worry about germs again.

Teacher: That’s a good question. A lot of germs have what’s called reservoirs in the environment, where they live when they’re not infecting humans. Even if the government tried your plan, probably most sicknesses would come back from their reservoirs.

Student: I’ve never seen a germ reservoir. Where are they?

Teacher: They’re not literal reservoirs like a water reservoir, that’s just what we call it when germs live in bats or raccoons or something.

Notice a few things about this conversation.

First, it’s really good that it happened. The student was clearly confused at many points. First, he had direct evidence that seemed to contradict the teacher’s claims that germs only spread by touch. Second, he had a sort of efficient-market-style confusion: germ theory seems to imply an easy way to eliminate all sicknesses forever, so why hasn’t someone picked this low-hanging fruit? Third, he was confused by an awkward term – he thought germ reservoirs were lakes full of germs in the hills somewhere.

All of these confusions are totally understandable. In fact, they’re a really good sign he’s paying attention – that he’s trying to figure out what germ theory really means and how it interacts with the rest of his worldview. They’re the direct opposite of guessing the teacher’s password. One day this kid is going to be an amazing scientist.

Second, to a hostile observer, it would sound like the student was challenging the teacher. Every time the teacher tried to explain germ theory, the student “pounced” on a supposed inconsistency. When the teacher tried to explain the inconsistency, the student challenged her explanations. At times he almost seems to be mocking the teacher. Without contextual clues – and without an appreciation for how confused young kids can be sometimes – it could sound like this kid is an arrogant know-it-all who thinks he’s checkmated biologists and proven that germ theory can’t possibly be true. Or that he thinks that he, a mere schoolchild, can come up with a novel way to end all sickness forever that nobody else ever thought of.

And the thesis of this post is that you must never, ever say that. Saying that is so bad. Smack down that student once, say “I think I know more about germ theory than you do”, make him feel like he challenged your authority and that’s bad – and the best case scenario is he will never ask questions to resolve his confusion again. The worst case scenario is that he stops feeling the confusion entirely, or stops thinking of forcing things to fit together and make sense as a desirable goal to have.

One of the most important rationalist skills is “noticing your confusion”. But that depends on an even more important proto-skill of wanting things to make sense. If you lose that skill – if it stops bothering you and seeming like a problem when things don’t make sense to you – you will never notice your confusion and you will never become a good scientist or a good anything-else-that-requires-independent-thought. And interpreting an attempt to explore dissonance as a status grab that needs to be knocked down is absolutely fatal for that skill. Instead, you need to think of it as Socratic grilling – like Socratic questioning, but a little harsher and more confrontational in order to get to the point more quickly.

Tolerating this is harder than it sounds. Most people can stay helpful for one or two iterations. But most people are bad at explaining things, so one or two iterations isn’t always enough I’ve had times when I need five or ten question-answer rounds with a teacher in order to understand what they’re telling me. The process sounds a lot like “The thing you just said is obviously wrong”…”no, that explanation you gave doesn’t make sense, you’re still obviously wrong”…”you keep saying the same thing over and over again, and it keeps being obviously wrong”…”no, that’s irrelevant to the point that’s bothering me”…”no, that’s also irrelevant, you keep saying an obviously wrong thing”…”Oh! That word means something totally different from what I thought it meant, now your statement makes total sense.”

But it’s harder even than that. Sometimes there is a vast inferential distance between you and the place where your teacher’s model makes sense, and you need to go through a process as laborious as converting a religious person to a materialist worldview (or vice versa) before the gap gets closed. The process of learning to really appreciate communism, or libertarianism, or whatever, coming from a diametrically opposed philosophy, looks a lot like dozens of questions about “but isn’t that an atrocity?” “wouldn’t this inevitably lead to dystopia?” and hearing what your interlocutor has to answer. It’s so, so tempting to round this off to them trying to gotcha you (as indeed sometimes it will be) and assume they’re not really committed to trying to understand.

With a good teacher, you can add enough disclaimers to your statements that they won’t get mad at you during this process: “I know I am but a mere student, and nowhere near smart enough to actually challenge you, so I’m sure I’m just misunderstanding this, but the thing you just said seems really confusing to me, and I’m not saying it’s not true, but I can’t figure out how it possibly could be true, which is my fault and not yours, but could you please try to explain it differently?” With a great teacher, all of this is assumed, and you don’t need the disclaimers, and you can just say “What? That makes no sense,” and expect the teacher to try again.

I think actual teachers get this right more often than not. What I’m concerned about is the self-appointed Internet anti-autodidact brigade. I occasionally get these people as commenters on my blog, but they’re more common in Reddit and places like that. Someone goes “that’s weird” or “if X was really true, wouldn’t that imply Y?” and gets hit with “You really think you’re smarter than everyone else? You really think a random person on the Internet has discovered a hole in X?” No, sometimes they’re just using Socratic grilling to expose the contradictions in their model and get somebody to resolve them.

I find this to be one of the most frustrating parts of writing this blog: how do I signal the things I still need to learn without the Arrogance Police descending on me? This is different from the case with a specific teacher – here there’s just me projecting my confusions out into the void to see if anyone answers. And every so often I do flatter myself that I’ve discovered a flaw in my betters’ reasoning, so it’s less obvious when I mean to imply that and when I don’t. Next time this happens I’m going to try just linking this post and seeing how it goes.

I think I strike a good balance, at the cost of endless annoying disclaimers that 95% of you don’t need. From now on, maybe I’ll just link this post.

# For, Then Against, High-Saturated-Fat Diets

I.

In the 1800s, the average US man weighed about 155 lbs. Today, he weighs about 195. The change is even starker at the extremes. Someone at the 90th percentile of weight back then weighed about 185 lbs; today, he would weigh 320 lbs. Back then, about 1% of men were obese. Today, about 25% are.

This puts a lot of modern dietary advice into perspective. For example, lots of people think low-carb is the solution to everything. But people in the 1800s ate almost 50% more bread than we do today, and still had almost no obesity. Other people think paleo is the solution to everything, but Americans in the 1800s ate a diet heavy in bread, milk, potatoes, and vegetables, and relatively low in red meat and other more caveman-recognizable foods. Intermittent fasting – again, cool idea, but your great-grandfather wasn’t doing that, and he had a 1% obesity risk.

This isn’t to say those diets can’t work. Just that if they work, they’re hacks. They treat the symptoms, not the underlying problem. Something went terribly wrong in US nutrition between 1900 and today, and all this talk about low-carb and intermittent fasting and so on are skew to that thing. Given that 1800s Americans seem to have effortlessly maintained near-zero obesity rates while eating foods a lot like the ones we eat today, maybe we should stop trying to figure out what cavemen were doing, and start trying to figure out what Great-Grandpa was doing, which sounds a lot easier.

We get similarly confusing evidence from other countries. Until recently, Chinese people ate mostly white rice. This is exactly the sort of high-glycemic-index carb that low-carbers say should be terrible for you. But the Chinese stayed thin even when they ate a lot. It was only when they started eating processed Western-style food that their obesity rate started to rise.

Or what about France? The French diet is about what you would expect; baguettes, pastries, cheese, meat. Lots of sugar, white flour, and fat – the opposite of all reasonable dietary advice. But 1970s France had the same kind of low obesity rates as 1800s America or China. This is related to the nutritional conundrum famously called the French paradox – why aren’t the French fatter and sicker than they are?

The answer to all these questions seems to be something like “the body is pretty good at regulating its own weight under any diet except modern American processed food.” But what aspect of processed food makes it bad?

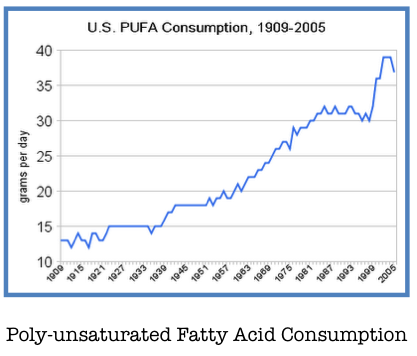
A new section of the online nutrition-sphere claims the answer has to do with the way mitochondria process fat. I’ve been trying to read these people and get a feel for their opinions. Most of what I’ve absorbed has come from Brad Marshall of Fire In A Bottle and his posts on The Croissant Diet. I’ve been told that another blog called Hyperlipid has a deeper investigation, but I’ve only scratched the surface of them. The r/SaturatedFat subreddit has some good stuff too. I don’t claim to fully understand these people and I apologize for any misrepresentations I might be making. But the short version is: they all agree that everything went wrong when we switched from saturated to unsaturated fat.

Wait, isn’t unsaturated fat the good kind of fat? Well, yes, this is what everyone else thinks. This is definitely one of those “good things are bad and bad things are good” diets. But let’s take a look at the argument.

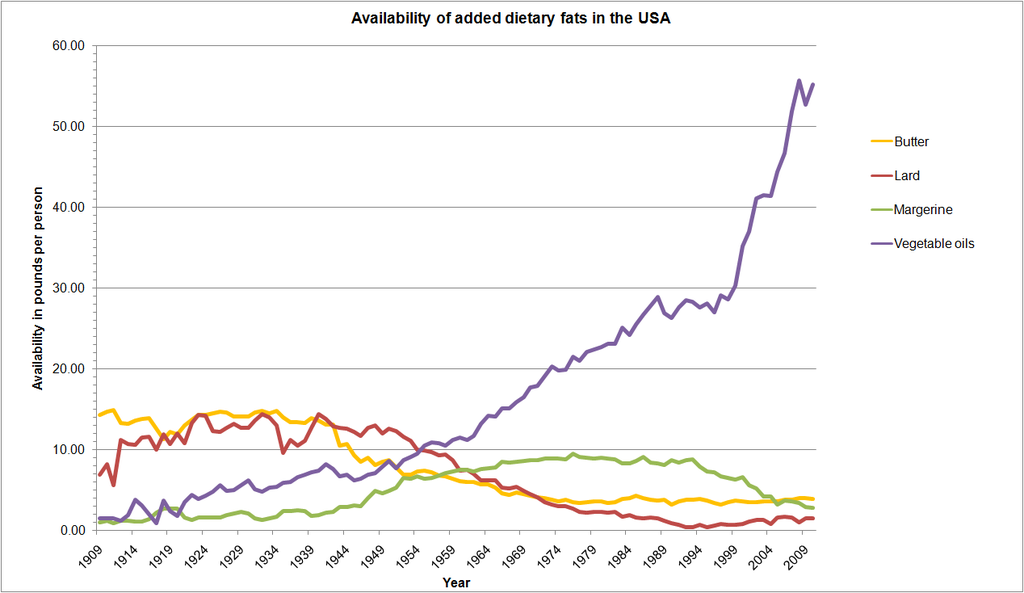
In the 1950s, heart disease rates were rising in the US. Realistically, this was mostly because lots of people had started smoking a few decades before, and now all that tobacco was catching up with them. But people didn’t know that at the time, so they did some studies into nutrition, and the studies suggested maybe saturated fat caused heart attacks. So the government told people to replace saturated fat with unsaturated fat, and this was back when the government was competent, so saturated fat consumption plummeted and unsaturated fat consumption shot up.

This paragraph is an extreme oversimplification: saturated fat is mostly found in things like milk, butter, and meat. Unsaturated fat comes in two types: monounsaturated and polyunsaturated. Monounsaturated fat is found in olive oil. Polyunsaturated fat comes in two types: omega-3 and omega-6. Omega-3 is mostly found in fish (“fish oil”). Omega-6 is mostly found in vegetables (“vegetable oil”). If you’ve ever looked at your food and seen ingredients like soybean oil, safflower oil, canola oil, corn oil, et cetera oil, these are omega-6 polyunsaturated fats.

When the government recommended a switch from saturated fats to unsaturated fats around the 1950s, it was omega-6 polyunsaturated fats – vegetable oils – that picked up the slack. Here are some helpful charts:

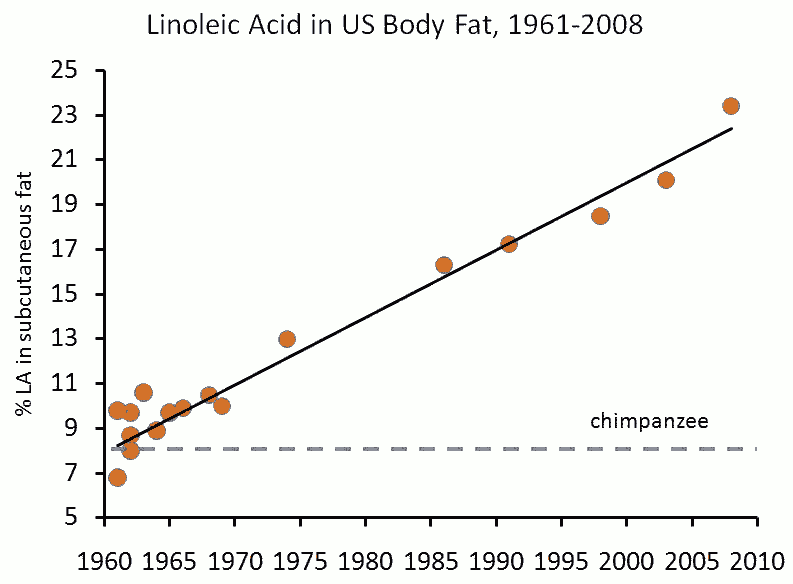


I cannot for the life of me find original sources for either of these graphs, but they seem consistent with everything else I’ve heard so I am going to trust them. Sorry!



[source]

This dramatic change in consumption of fat was reflected in a dramatic change in the composition of the human body. Studies of human fat cells and breast milk found that they went from being overwhelmingly saturated fat (like the fat cells and breast milk of animals) to being partly polyunsaturated fat:



[source, source, source, source]

The only common villain everyone agrees on in the obesity story is “processed food”. I’ve previously found this frustrating – it reeks of a sort of unreflective technophobia. What part of processing makes food bad? How does mere contact with a machine turn food from healthy to unhealthy? What food counts as “processed” or “not processed”? Is ground beef processed, since you grind it? Are scrambled eggs processed, since you scramble them? Is bread processed, since wheat doesn’t grow in loaves? Is water processed, since it goes through water processing facilities? Is the Eucharist processed, even though the processing only changes its metaphysical essence and not its physical properties? Everybody I ask acts like the answers to these questions are obvious, but everyone has different answers, and nobody can tell me their decision procedure.

Omega-6 polyunsaturated fats provide a tempting answer. Processing is bad at the point where it involves adding vegetable oil to stuff.

Here is a not-so-fun experiment you can try. Go to your local grocery store, check the ingredients of everything, and see what is the most ridiculous place you can find added vegetable oil – soybean oil will be the most common, though you can spot others. I thought I had reached a low when I found soybean oil listed in the ingredients of what was, to all appearances, just a bag of rice. But then I checked the bread section and found that about 90% of the loaves of bread had soybean oil added to them too (and some of the remainder had safflower oil). It really is the closest thing to a ubiquitous omni-ingredient in every processed food, and in some foods that you wouldn’t have thought were processed at all.

So this is the circumstantial evidence linking polyunsaturated fat to obesity. Although polyunsaturated fat itself is natural (found in eg nuts and seeds), modern Americans consume it at levels that would have been equally foreign to cavemen and your great-grandfather. The vegetable oil craze started around the same time as the obesity epidemic, and the two have been following the same pattern ever since. And it’s concentrated in the same processed foods that most people think are most responsible for obesity. Also, the body fat of obese people is more polyunsaturated than the body fat of healthy people.

But is there a biologically plausible reason why polyunsaturated fat would cause obesity?

Actually, there are several. The one I’d heard a few years ago blames an omega-6 to omega-3 imbalance. Before 1950, most people ate a reasonable amount of fish and a reasonable amount of nuts, and got about 4x as much omega-6 as omega-3. After 1950, people started eating lots of high-omega-6 vegetable oil, but only the same amount of high-omega-3 fish, and the ratio shot up: now it’s about 10x-50x as much omega-6 as omega-3. Both omega-6 and omega-3 are involved in cell membranes and signaling chemicals, and there’s some evidence that omega-6s may be pro-inflammatory and omega-3s anti-inflammatory. Although the studies haven’t really been done, you can tell a story where the natural ratio of 6s to 3s creates a natural level of inflammation, and the current extreme level of 6s to 3s creates an extreme level of inflammation. Inflammation in the parts of the brain that regulate diet are one proposed mechanism for obesity, so there’s the skeleton of an explanation here, although lots of work would need to be done to prove it.

But the new one, the one that Marshall and Hyperlipid are pushing, is a little different. They think unsaturated fats in general are bad, including monounsaturated fats and omega-3s (though realistically omega-6 vegetable oils so overwhelm these in the average American diet that we can forget everything else.). The exact mechanism is complicated, but focuses on the Krebs cycle, bane of medical students everywhere. The Krebs cycle is the set of chemical reactions that your cells use to convert high-energy food chemicals into ATP, a form of energy your body can use to power its own biological processes. Both saturated and unsaturated fats feed into the Krebs cycle. But the Krebs cycle produces reactive oxygen species (aka free radicals, eg hydrogen peroxide) when it metabolizes saturated fats, and not when it metabolizes unsaturated fats. Reactive oxygen species seem to be one of the signals the body uses to detect satiety, which makes sense – if they’re a byproduct of metabolizing food, and you have a lot of them, that probably means you just metabolized a lot of food, and so you should be full. There’s a lot of biochemistry here, and I haven’t gone through all of it. But the basic idea is – burning saturated fat makes you full, but for decades we’ve been replacing saturated fat with polyunsaturated fat, which doesn’t make you full. So we’ve been eating more. Hence, obesity.

If this were true, it suggests a diet high in saturated fat and low in unsaturated fat, especially polyunsaturated fat. Marshall calculated out the right macronutrient ratio and found that the food that most closely matched it was butter croissants – hence the name “the croissant diet”. Yes, you can live off croissants if you want. But it also allows lots of other things with high saturated and low unsaturated fat. Red meat (but not chicken or pork; most chickens and pigs are fed high-PUFA feed that gives them high-PUFA meat). Milk, cheese, and butter (but not margarine, which is mostly PUFAs). Pasta, rice, and other carbs (but if you’re putting sauces on them, make sure they’re high saturated fat). Fried things, as long as you fry them in coconut oil, palm oil, or butter (mostly saturated fat) instead of vegetable oils or olive oils (mostly unsaturated fat). You can read the full specification here.

This diet is kind of the opposite of the one most nutritionists recommend. But it would taste a lot better. And following nutritionists’ advice hasn’t worked out so well for Americans circa 1970 through 2020. So what could go wrong?

II.

I find this to be a really elegant and provocative theory, with impressive circumstantial evidence. Unfortunately, as far as I can tell all of the direct evidence is against it. I’m not a nutritionist and have not studied these nearly as intently as the nutrition bloggers who support it, so low confidence in this part. But I’m putting my objections out there in the spirit of seeing whether other people will be able to reply to them and shoot them down.

First, let’s start with the diet itself. Lots of people have tried it, and the most common result is nothing. The r/SaturatedFat subreddit is full of people talking about how the diet didn’t work for them, with only a few contrary opinions. I tried it for about a month, and lost about 4 pounds the first two weeks, followed by no more weight loss no matter how hard I tried. This is my usual pattern every time I try a new diet, and I interpret it as placebo or just the effect of restricting some foods. The high-saturated-fat croissant diet did no better.

(Marshall argues it gets better results in waistline shrinkage than weight loss alone, and my own results sort of seem to confirm this, but I’m not confident in the accuracy of my waistline measurements. Also, why should this be true?)

Second, mouse studies. You can feed mice whatever you want, then see how much weight they gain. There are a lot of these, they’re all conducted with different mice, different macronutrient sources, and different methodologies, and they all get different results. After looking at many of them, all I can say is that there is definitely no strong trend for lower saturated and higher polyunsaturated fat diets to result in more weight gain. For example, in this study, mice who ate palm oil (a high saturated-fat oil) gained more weight than those who ate safflower oil (a high polyunsaturated fat oil), although mice who ate cocoa butter (a different high saturated fat oil) gained less than either. In this study, mice who ate lard (high in saturated fat) gained more weight than those who ate olive oil (monounsaturated) or fish oil (polyunsaturated). Again, I’m less interested in these particular studies or their particular results than in the vast amount of literature that has investigated these questions and very rarely found a strong unambiguous tendency for saturated fat to be good and unsaturated fat to be bad. Stephan Guyenet, who used to support the omega-3:omega-6 ratio theory, agrees with this and now considers it unlikely based on mouse study results. It’s possible that I’m missing different results for different kinds of saturated fat – stearic acid (a specific kind of saturated fat) seems to do pretty well, and the croissant diet to some degree centers around it. But as far as I can tell, an inferiority of any kinds of saturated fat to any kinds of polyunsaturated fat doesn’t seem compatible with the basic theory.

And third, people have studied the effect of saturated vs. unsaturated fat so much. This is maybe the biggest controversy in nutrition right now. Some people think (in accordance with the 1950s and 1960s findings) that saturated fat contributes to cardiovascular disease. Other people think those findings were wrong and it doesn’t. There have been a bunch of studies and big meta-analyses trying to find out who’s right with only limited agreement. While some studies have found that saturated fat is bad and others that it’s harmless, as far as I know none of them have found that it has a strong protective effect against weight gain. If this was really the difference between the 1800s when nobody was obese and today when a bunch of people are, it ought to be a blindingly bright signal. But I don’t see anything of the sort.

Just to give examples: Lin investigated the relationship between saturated fat and weight loss, and found that the higher a diet was in saturated fat, the more likely people were to gain back weight they had lost. Phillips investigated the same question in relation to a probably irrelevant candidate gene, and found the same thing. Utzscheinder investigated high- vs. low- saturated fat diets and found no difference in weight loss, but the high saturated fat diet contributed to unhealthy liver fat deposition. Khaw et al investigated the effect of butter and coconut oil(saturated) vs. olive oil (monounsaturated) on obesity, and found nothing. Schwingshackl et al investigated the effect of 10 food groups on metabolic parameters and found that nuts were the healthiest, even though they are the highest in omega-6 fats.

(also, the whole point of this diet was supposed to be that croissants should be an unusually satiety-producing food, but somebody studied how much satiety every food produces, and croissants are literally the lowest on the list.)

I haven’t looked closely at any of these studies and don’t especially trust them. I’m more gesturing toward the general idea of how unlikely it is that people have studied saturated fat in depth for a long time, gotten a bunch of small negative effects that might or might not be real, and failed to notice that the real effect is gigantic and positive.

For that matter, where are the random Redditors? Saturated fat is one of the major macronutrients, it’s not exactly some weird exotic chemical nobody ever thought to test. If high-saturated-fat or low-polyunsaturated-fat diets help you stay thin as easily as 1800s Americans stayed thin, how come people didn’t figure this out in ten minutes? There have been so many random diet crazes like low-carb and paleo, all linked to some people squinting and thinking they might have seen a signal among all the noise. If there were a diet that was pretty basic and actually worked in an obvious way, don’t you think people would have found it?

III.

There’s an awkward tension between the first part of this post and that last counterargument.

The first part of the post said that there is some dramatic and hard-to-explain difference in obesity between the modern West and every other civilization, whether that’s the historical West or other modern countries that haven’t yet adopted our diet. I haven’t given evidence here, but the obesity goes hand in hand with higher cancer rates, higher cardiovascular disease rates, and just generally worse health. Presumably we’re doing something very wrong.

The last counterargument said that if we were doing something very wrong, one of the thousands of biohackers who has tried every ridiculous fad diet and long-shot idea would have reversed the one wrong thing we were doing and gotten incredible results. Then, by the efficient market hypothesis, somebody would have noticed the incredible results, and the smart paying-attention people would switch to that diet, and then we’d have a world that looks a lot different than the past decade or two of people chasing various exciting ideas with no results. As far as I know, nobody has yet met Ampersand’s challenge of finding a peer-reviewed study demonstrating that some diet can consistently help people lose lots of weight and keep it off.

But how could that be? If people used to be thin and healthy, we should just be able to do what they were doing! And then that would be the diet that can consistently lose weight and keep it off!

I can only see three ways out of this paradox.

First, conventional wisdom is right about everything. People are fatter today than in the 1800s because they eat too much and exercise too little. They eat too much because they are rich, food is cheap, and food tastes really good. They exercise too little because they’re office workers now instead of farmers. In this model, the reason the efficient market hasn’t found the secret to weight loss is because there’s no secret and weight loss is really hard. It wasn’t hard for your great-grandfather because he had fewer options and so he didn’t need to exercise willpower to avoid the bad ones. The most sophisticated version of this model, so sophisticated that maybe I shouldn’t call it this model at all, is the food reward theory ably defended by Stephan Guyenet.

This has a lot going for it, but can’t be quite right. Exercise seems like a red herring; studies of how much people eat, exercise, and gain weight have shown that dietary changes explain more than 100% of weight gain over the past 30-40 years – probably we are exercising a little more. And there was really tasty food in 1800s America and 1970s France, so how come people didn’t overindulge in that? How does it explain all the weird results like lab animals, pets, and feral rats gaining weight? This probably part of it, but it still feels like something’s missing.

Second, diet is barely related to the obesity epidemic, and it’s being caused by plastics or antibiotics affecting the microbiome or something like that. This is another thing where I would have expected people to notice, but I definitely don’t want to dismiss it prematurely.

Third, it’s a ratchet. Departing from the ways of our ancestors (or great-grandparents) can make you obese, but returning to their ways cannot make you thin again. A bad diet (whatever that is) shifts your weight set point up, but a good diet does not shift it back down, at least not in a reasonable amount of time. It just prevents further damage.

This contradicts the evidence from some people who do manage to lose weight, including some people who manage to lose a lot of weight and keep it off. I think the theory would have to be that different people’s set points are differently malleable, and that some people are obese because their set point is set to obese, and other people barely have a set point and are mostly operating on calories-in-calories-out. If this seems a little too convenient an assumption, keep in mind this is how lots of other set points work – some people will gain tolerance to certain drugs almost instantly, and other people will never get it at all.

Overall I am ending this research more confused than when I started it. I think the most likely dietary change I make is to try to avoid foods with soybean, corn, or safflower oil, since this is probably a good stand-in for “foods processed enough that they count as processed foods and you should avoid them”. I don’t think the evidence is good for avoiding fish oil and olive oil, and there’s enough evidence from elsewhere that these foods are healthy that I’m going to keep trying to eat them. I don’t think the evidence is good for saturated fats being especially good, and there seems to be at least equally strong evidence that they’re bad, so although I’m not going to work too hard to avoid them I’m definitely not going to optimize my diet for getting as many of them as possible.

# Book Review: Hoover

You probably remember Herbert Hoover as the guy who bungled the Great Depression. Maybe you shouldn’t. Maybe you should remember him as a bold explorer looking for silver in the jungles of Burma. Or as the heroic defender of Tientsin during the Boxer Rebellion. Or as a dashing pirate-philanthropist, gallivanting around the world, saving millions of lives wherever he went. Or as the temporary dictator of Europe. Or as a geologist, or a bank tycoon, or author of the premier 1900s textbook on metallurgy.

How did a backwards orphan son of a blacksmith, dropped in the middle of a forgotten spot in the Midwest, grow up to be a captain of industry and a US President? How did he become such a towering figure in the history of philanthropy that biographer Kenneth Whyte claims “the number of lives Hoover saved through his various humanitarian campaigns might exceed 100 million, a record of benevolence unlike anything in human history”? To find out, I picked up Whyte’s Hoover: An Extraordinary Life In Extraordinary Times.

Herbert Hoover was born in 1874 to poor parents in the tiny Quaker farming community of West Branch, Iowa. His father was a blacksmith, his mother a schoolteacher. His childhood was strict. Magazines and novels were banned; acceptable reading material included the Bible and Prohibitionist pamphlets. His hobby was collecting oddly shaped sticks.

His father dies when he is 6, his mother when he is 10. The orphaned Hoover and his two siblings are shuttled from relative to relative. He spends one summer on the Osage Indian Reservation in Oklahoma, living with an uncle who worked for the Department of Indian Affairs. Another year passes on a pig farm with his Uncle Allen. In 1885, he is more permanently adopted by his Uncle John, a doctor and businessman helping found a Quaker colony in Oregon. Hoover’s various guardians are dutiful but distant; they never abuse or neglect him, but treat him more as an extra pair of hands around the house than as someone to be loved and cherished. Hoover reciprocates in kind, doing what is expected of him but excelling neither in school nor anywhere else.

In his early teens, Hoover gets his first job, as an office boy at a local real estate company. He loves it! He has spent his whole life doing chores for no pay, and working for pay is so much better! He has spent his whole life sullenly following orders, and now he’s expected to be proactive and figure things out for himself! Hoover the mediocre student and all-around unexceptional kid does a complete 180 and accepts Capitalism as the father he never had.

His first task is to write some newspaper ads for Oregon real estate. He writes brilliant ads, ads that draw people to Oregon from every corner of the country. But he learns some out-of-towners read his ads, come to town, stay at hotels, and are intercepted by competitors before they negotiate with his company. Of his own initiative, he rents several houses around town and turns them into boarding houses for out-of-towners coming to buy real estate, then doesn’t tell his competitors where they are. Then he marks up rent on the boarding houses and makes a tidy profit on the side. Everything he does is like this. When an especially acrimonious board meeting threatens to split the company, a quick-thinking Hoover sneaks out and turns off the gas to the building, plunging the meeting into darkness. Everyone else has to adjourn, the extra time gives cooler heads a change to prevail, and the company is saved. Everything he does is like this.

(on the other hand, he has zero friends and only one acquaintance his own age, who later describes him to biographers as “about as much excitement as a china egg”.)

Hoover meets all sorts of people passing through the Oregon frontier. One is a mining engineer. He regales young Herbert with his stories of traveling through the mountains, opening up new sources of minerals to feed the voracious appetite of Progress. This is the age of steamships, skyscrapers, and railroads, and to the young idealistic Hoover, engineering has an irresistible romance. He wants to leave home and go to college. But he worries a poor frontier boy like him would never fit in at Harvard or Yale. He gets a tip – a new, tuition-free university might be opening in Palo Alto, California. If he heads down right away, he might make it in time for the entrance exam. Hoover fails the entrance exam, but the new university is short on students and decides to take him anyway.

Herbert Hoover is the first student at Stanford. Not just a member of the first graduating class. Literally the first student. He arrives at the dorms two months early to get a head start on various money-making schemes, including distributing newspapers, delivering laundry, tending livestock, and helping other students register. He would later sell some of these businesses to other students and start more, operating a constant churn of enterprises throughout his college career. His academics remain mediocre, and he continues to have few friends – until he tries out for the football team in sophomore year. He has zero athletic talent and fails miserably, but the coach (whose eye for talent apparently transcends athletics) spots potential in Hoover and asks him to come on as team manager. In this role, Hoover is an unqualified success. He turns the team’s debt into a surplus, and starts the Big Game – a UC Berkeley vs. Stanford football match played on Thanksgiving which remains a beloved Stanford football tradition.

Other Stanford students notice his competence, and by his senior year he is running not just the football team but the baseball team, a lecture series, a set of concerts and plays, and much of the student government. For the first time, he makes many social contacts, which is sort of like having friends, although real emotional connection remains beyond him. Whyte describes an occasion when Will Irwin, the football team’s star player, suffers a career-ending injury:

[Irwin] was outfitted with a plaster cast and deposited in his dorm room. Hoover visited him to approve spending on the athlete’s medical supplies…Hoover carried his head to one side as he took in Irwin’s cast and obvious discomfort…To make conversation and keep up his courage, Irwin tried to make light of his situation and watched as Hoover tried to laugh. A ‘deep, rich, chuckle’ originated far down in his chest, Irwin recalled, yet it was strangled ‘before it came to the surface’. Hoover did not offer the patient a single word of consolation or reassurance during his time in the room. Irwin assumed hat Hoover’s sympathies, for he did appear to be affected, were garroted and buried in the same internal graveyard as the chuckle. After a few minutes, Hoover headed for the door and, at the last instant, turned and blurted ‘I’m sorry’. Irwin recognized that this minimal expression of emotion was as traumatic for Hoover as a broken ankle.

Hoover graduates Stanford in 1895 with a Geology degree. He plans to work for the US Geological Survey, but the Panic of 1895 devastates government finances and his job is cancelled. Hoover hikes up and down the Sierra Nevadas looking for work as a mining engineer. When none materializes, he takes a job an ordinary miner, hoping to work his way up from the bottom:

He signed on as a mucker at the Reward Mine, shoveling wet dirt and rock into an ore car on ten-hour shifts for two dollars a day, seven days a week. The Cornishmen mocked him for his schooling and taught him the basics of their mole-like existence: how to breathe while the dust cleared from a blast; how to nap in a steel wheelbarrow heated from underneath by candles. The ceaseless grind of filling his car and pushing it up the slick rails of the Reward’s dripping tunnels taxed Hoover’s stamina. He was tortured in his sleep by muscle pain and neuralgia.

After a few months, he finds a position as a clerk at a top Bay Area mining firm. One year later, he is a senior mining engineer. He is moving up rapidly – but not rapidly enough for his purposes. An opportunity arises: London company Berwick Moreing is looking for someone to supervise their mines in the Australian Outback. Their only requirement is that he be at least 35 years old, experienced, and an engineer. Hoover (22 years old, <1 year experience, geology degree only) travels to Britain, strides into their office, and declares himself their man. The executives “professed astonishment at Americans’ ability to maintain their youthful appearance” (Hoover had told them he was 36), but hire him and send him on an ocean liner to Australia.



22 year old Hoover trying his best to look like a respectable 36 year old capitalistWhat does he think of his new home?

In numerous letters over the next two years, Hoover would refer to Western Australia as hell, and he meant it. The landscape was hell, a flat, monotonous, dust-choked desert, barren but for low tangles of mulga and wattle bush as far as the eye could see.

The climate was hell, a dry broil for the most part, one hundred degrees at midnight for days on end…

The insects were hell, scorpions, tarantulas, snakelike centipedes, and disease carrying airborne pests with an unerring aim for one’s eyes and dinner plate…

The settlements were hell, overnight ramshackled boomtowns with names like Kalgoorie and Coolgardie, box-shaped lodgings with walls of corrugated iron that roared in the wind, beds with unwashed sheets, meals of beans, biscuits, canned potatoes, and “tinned dog” (probably mutton or ham), entertainment consisting of out-of-date copies of American magazines, the odd horse race, and drunks dodging camels on Main Street.

“You cannot appreciate the real damnation of this country,” wrote Hoover.

Hoover soon manages to personally offend every single person in Australia:

The harshness of the environment and Hoover’s desire to prove himself drew an element of savagery from him. He fired rafts of employees for laziness and incompetence and dumped two of his own assistants for being “damn noodle heads”…uncompromising in pursuit of better margins, Hoover haggled with camel dealers to save a few dollars on freight costs He moved swiftly to shut losing properties…He lengthened shifts in the Coolgardie mines from 44 to 48 hours (his efforts to introduce labor-saving technology at another mine would result in a job action, which Hoover answered by firing the strikers and hiring more pliable Italian labor)…

Hoover drove himself relentlessly as well, sleeping as little as four hours a night. His eyes and stomach gave him trouble. Months of roasting on the Western Australia grill left him with a chronic inflammation of the bladder. Sometimes he was so ill he could not sit up, but he refused to slow down, traveling on his back on a mattress on the bottom of a horse-drawn cart.

After a year, Hoover is the most hated person in Australia, and also doing amazing. His mines are producing more ore at lower prices than ever before. He receives promotion after promotion.

Success goes to his head and makes him paranoid. He starts plotting against his immediate boss, Berwick Moreing’s Australia chief Ernest Williams. Thought Williams didn’t originally bear him any ill will, all the plotting eventually gets to him, and he arranges for Hoover to be transferred to China. Hoover is on board with this, since China is a lucrative market and the transfer feels like a promotion. He travels first back to Stanford – where he marries his college sweetheart Lou Henry – and then the two of them head to China.



Herbert Hoover’s college sweetheartChina is Australia 2.0. Hoover hates everyone in the country and they hate him back:

Hoover shared the prevailing European conviction of Chinese racial inferiority. He would write of the ‘simply appalling and universal dishonesty of the working classes, the racial slowness, and the low average of intelligence’…Hoover was baffled at their lack of enthusiasm for mechanization and orderly administration. Lou reported that ‘the utter apathy of the Chinese to everything, their unconquerable dilatoriness’ was almost heartbreaking to her energetic husband.

The same conflicts are playing themselves out on the world stage, as Chinese resentment at their would-be-colonizers boils over into the Boxer Rebellion. A cult with a great name – “Society Of Righteous And Harmonious Fists” – takes over the government and encourages angry mobs to go around killing Westerners. Thousands of Europeans, including Herbert and Lou, barricade themselves in the partly-Europeanized city of Tientsin to make a final last stand. Hoover

“…fought fires in the settlement and delivered food and medical supplies on his bicycle, hugging the brick walls along the street to avoid gunfire. Reporters on the scene observed that he seemed to be moving on the double quick, furiously jingling the change in his pockets and chewing nuts without shucking them. Lou, unwilling to join other women in the safety of the basement at city hall, ran bicycle errands of her own, a .38 Mauser strapped to her hip…

In between dodging artillery shells, Hoover furiously negotiates property deals with his fellow besiegees. He argues that if any of them survive, it will probably because Western powers invade China to save them. That means they will soon be operating under Western law, and people who had already sold their mines to Western companies would be ahead of the game and avoid involuntary confiscation. Somehow, everything comes up exactly how Hoover predicts. US Marines arrive in Tientsin to liberate the city (Hoover marches with them as their local guide) and he is ready to collect his winnings.

Problem: it turns out that “Whatever, sure, you can have my gold mine, we’re all going to die anyway” is not legally binding. Hoover, enraged as he watches apparently done deals slip through his fingers, reaches new levels of moral turpitude. He offers the Chinese great verbal deals, then gives them contracts with terrible deals, saying that this is some kind of quaint foreign custom and if they just sign the contract then the verbal deal will be the legally binding one (this is totally false). At one point, he literally holds up a property office with a gun to get the deed to a mine he wants. Somehow, after consecutively scamming half the population of Asia, he ends up with the rights to China’s most lucrative minefields. Berwick Moreing congratulates him and promotes him to managing director. He and Lou sail for London to live the lives of British corporate bigshots.

Predictably, Hoover makes an amazing corporate bigshot:

Hoover had a ‘gift of juggling corporate assets in such a manner that insiders almost always benefitted’, whatever happened to the capital of the original shareholders. He was masterful at wielding write-offs and preference shares with multiple voting power on the grounds that new capital was required to avert bankruptcy. His favorite deals were those so complicated no one else could figure out they worked.

On top of this, Hoover could keep mental maps of dozens of mines in his mind and, by one account, follow the progress of each shaft like a blindfolded chess master. He liked to receive telegrams from these properties and, without opening them, noting only the date and address, predict the level of the mine and the cost per ton of ore. He was usually correct.

His intellectual capacities and powerful will made Hoover a fearsome negotiator. Arriving at the table with shirtsleeves rolled up, abrupt and aggressive, he had a singular talent for stripping away nonessential information and getting directly to the root of things, and he knew how to close. He possessed what one businessman said was a curious dynamic force that could compel the most reluctant person to put signatures to paper.

Also predictably, Hoover manages to offend everyone in Britain. Soon he is signing off on a ‘mutually agreeable’, ‘amicable’ dismissal from Berwick Moreing. They agree to let him go on the condition that he does not compete with them – a promise he breaks basically instantly. He goes into banking, and his “bank” funds mining operations in a way indistinguishable from being a mining conglomerate. Eventually he abandons even this fig leaf, and just mines directly.

But in other ways, his tens of millions of dollars are mellowing him out. Over his years in London, he develops hobbies besides making money and crushing people. He starts a family; he and Lou have two sons, Herbert Jr and Allen. He even hosts dinner parties, very gradually working on the skill of getting through an entire meal without mortally offending any guests:

His fund of small talk was perpetually overdrawn, and if he interacted with the guests at his elbows, it was typically in a series of grunts or nods. If he wanted to make a point, he made it in a flat voice and then stopped abruptly, as one friend noted, someone had pulled his plug. If aroused, he would speak with force, sometimes veering into tactlessness, pursuing minor differences of opinion so harshly and indignantly that his victims nursed grudges for the rest of their natural lives. One acquaintance considered him the bluntest man in Europe, another ‘the rudest man in London’. He seldom took the time to enjoy his food, and was once clocked swallowing five courses in eleven minutes flat.

And he writes a book on metallurgy, which becomes the canonical text for a generation of engineering students. He can’t resist adding some of his own commentary. For example:

Among the book’s idiosyncratic touches is Hoover’s attempt to end discussion of the capacity of different races of workers, a common debating point in early 20th century mining, by quantifying a racial productivity gap. He deemed one white worker equal to two or three of the colored races in simple tasks like shoveling, and as high as one to eleven in the most complicated mechanical work.

But also:

To the engineer falls the work of creating from the dry bones of scientific fact the living body of industry. It is he whose intellect and direction bring to the world the comforts and necessities of daily need. Unlike the doctor, his is not the constant struggle to save the weak. Unlike the soldier, destruction is not his prime function. Unlike the lawyer, quarrels are not his daily bread. Engineering is the profession of creation and of construction, of stimulation of human effort and accomplishment.

Finally, having won respect in the financial, social, and intellectual worlds, he decides the natural next step is to become a public servant. Insofar as he has any political philosophy, he thinks of engineers as a sort of benevolent master race, destined to lead the world into an efficient technocracy. And he can think of no better standard-bearer than himself. He writes some Stanford friends, asking if they would support him for Governor of California. They suggest he start lower on the ladder, and offer him a position on the Stanford Board of Trustees, which he accepts (trustees are supposed to live in Palo Alto, but he lies and tells them he is moving back right away). He begins his public career by attacking tenure, “which he considered a protection racket for the weak and lazy and an outrage on the sanctity of higher education.”

Okay, fine. He hadn’t mellowed out that much. He manages to offend everyone in Stanford basically immediately, and that probably would have been the end of his career in politics. Luckily for him, World War I chooses that moment to break out, and little things like tenure are suddenly forgotten in the shadow of the greatest conflict the world has ever known.

II.

Count up the victims of World War I, and American tourists will be pretty far down the list. But victims they were. When the conflict broke out, thousands of Americans were overseas visiting the cathedrals of Florence or the museums of London. They woke up one morning to find the ships that were supposed to take them back had been conscripted into the war effort, or refused to sail for fear of enemy fire. The banks that were supposed to cash their travelers’ checks were panicking, or devoting all their funds to the war effort, or dealing with a million other things. The hotels that were supposed to house them were closed indefinitely, their employees rushing to enlist out of patriotic fervor. And so thousands of frantic Americans, stuck in a foreign continent with no money and nowhere to stay, showed up at the door of the US Embassy in London and said – help!

The US Consulate in London didn’t know how to solve these problems either. But Herbert Hoover, still high on his decision to pivot to philanthropy and public service, calls them up and asks if he can assist. They say yes, definitely. Hoover gets in touch with his rich friends, passes around the collection plate, and organizes a Committee For The Assistance Of American Travelers. Then he gets to work, the way only he can:

Within 24 hours, Hoover’s committee had its own stationery, and within forty-eight it was operating a booth in the ballroom of the Savoy Hotel as well as three other London locations. Through his business connections, Hoover managed to bypass restrictions on telegraph service and open a transatlantic line to allow Americans to wire money to stranded friends and relatives. In a city suddenly flooded with refugees, he reserved for American travelers some two thousand rooms in hotels or boardinghouses. He issued a press release proclaiming that his Residents’ Committee was assuming charge of all American relief work in the city, and that in doing so it had the blessings of its honorary chairman, Walter Hines Page, the US ambassador to London.

…which is totally false. Hoover is starting to display a pattern that will stick with him his whole life – that of crushing competing charities. He begins a lobbying effort to get the US Embassy to ban all non-Hoover relief work, focusing on the inefficiency of having multiple groups working on the same problem. When the US Assistant Secretary Of War arrives in London to coordinate a response, he is met on the dock by Hoover employees, who demand he consult with Hoover before interfering in the US tourist issue. Eventually the Embassy, equally exasperated by Hoover’s pestering and impressed with his results, agrees to give him official control of the relief effort.

After two months of work, Hoover and his Committee have repatriated all 120,000 US tourists, supporting them in style until it could find them boat tickets. All of its loans and operating costs have been repaid by grateful tourists, and its budget is in the black. The rescued travelers are universal in their praise for Hoover, albeit partly because Hoover has threatened to ruin any of them who get too critical:

Other complainants were received with less patience, including a hotheaded professor of history from the University of Michigan, who wrote to accuse the Residents’ Committee of mistreatment. Hoover refuted his charges indignantly and comprehensively, copying his response to the president of the university and its board of regents. After a meeting with his employer, the professor returned Hoover an abject retraction and apology.

Just as Hoover is preparing to rest on his laurels, he receives a cry for help. Germany has occupied and blockaded Belgium. The blockade prevents this tiny, heavily urban country from importing food, and the Belgians are starving. Germany needs its own food for its own armies, and is refusing to help. The Belgians order a thousand tons of grain from Britain, but when their representative comes to pick it up, Britain refuses to let them transport it, nervous at sending food into enemy-occupied territory. During tense negotiations, someone suggests using neutral power America as a go-between. But America is 5,000 miles away and busy with its own problems. So the US Ambassador to Britain asks his new best friend Herbert Hoover if he has any ideas.

Hoover invites Emile Francqui, a Belgian mining engineer he knows, to Britain. Together, they plan a Committee For The Relief of Belgium, intended not just to help transport the thousand tons of grain at issue, but to develop a long-term solution to the impending Belgian famine. Nothing like this has ever been tried before. Belgium has seven million people and almost no food. No government is offering to help, and they don’t have enough money to feed seven million people even for one day, let alone indefinitely. Hoover springs into action…

…by crushing all competing attempts to provide food for Belgium. He attacks the Rockefeller Foundation, which is trying to help, with a blitz of press coverage accusing it of various forms of insensitivity and interference, until it finally backs off. Then he gets to work on the government:

The letter bore several Hoover watermarks, beginning with its heavy load of facts and figures organized in point form. It noted that myriad relief committees were springing up both inside and outside of Belgium, and urged consolidation. “It is impossible to handle the situation except with the strongest centralization and effective monopoly, and therefore the two organizations [Hoover outside Belgium and Francqui inside it] will refuse to recognize any element except themselves alone.” The letter also contained Hoover’s usual autocratic and slightly paranoid demands for “absolute command” of his part of the enterprise.

Control attained, Hoover springs into action actually feeding Belgium. He launches one of the largest public relations campaigns the world has ever seen, sending letters to newspapers around the world asking for donations. He “urged reporters to investigate the famine conditions in Belgium and play up the ‘detailed personal horror stuff’. He personally arranged for a motion picture crew to capture footage of food lines in Brussels, and he hired famous authors, including Thomas Hardy and George Bernard Shaw, to plead for public support of the rescue effort.” He constantly telegrams his exasperated wife and children, now safely back in Palo Alto, demanding they raise more and more money from the West Coast elite.

He browbeats shipping conglomerates until they agree to ship his food for free, then browbeats railroads until they agree to carry it. By telegraph and letter he coordinates banks, docks, trains, ships, and relief workers on both sides of the Atlantic. But that’s just the prelude. His real problem is the governments. Britain doesn’t want food shipped to Belgium, because right now the starving Belgians are Germany’s problem, and they don’t want to solve an enemy’s problem for them. But Germany also doesn’t want food shipped to Belgium, because the Belgians are resisting the occupation, and they figure starvation will make them more compliant. Shuttling back and forth across the North Sea, Hoover tries to get them to switch theories: Germany needs to think starving Belgians are their problem which it would be helpful to solve, and Britain needs to think starvation would make Belgians more compliant with the German occupation. In the end, both countries allow the shipments.

He goes on a fact-finding mission to Belgium, and manages to somehow offend everyone in the country that he is, at that very moment, saving from mass starvation:

A third of Brussels’ population was receiving free food at more than a hundred canteens set up by the Comite Central and supplied by the CRB. Ration cards entitled the bearer to coffee, soup, and bread. On the cold, wet morning of December 1, Whitlock took Hoover to the street outside a theater that had been converted to a canteen in the Quarter des Marolles. They saw hundreds of Belgians shivering silently in the breadline…Whitlock kept his eyes on Hoover throughout the visit and saw him turn away and stare off down the street rather than share his feelings. Whitlock understood Hoover’s reaction as simple reticence. Others witnessing the same sort of behavior found it disturbing. They noticed how Hoover obsessed over the logistics of food distribution while avoiding interaction with recipients of relief and thought him a bloodless man. “He told of the work in Belgium as coldly as if he were giving statistics of production,” said US official. “From his words and his manner he seemed to regard human beings as so many numbers. Not once did he show the slightest feeling.”

Hoover’s reticence was chronic. He was the sort of man who could sit for three hours on a train with his closest colleagues and not utter a single word, or bid farewell to his wife, not expecting to see her again for several months, in a curt telegraph: “Goodbye, Love, Bert”. It was often difficult to know if his behavior was due to bad manners, callousness, anxiety, or an effort to manage powerful emotions, because he was capable of all these things. Indeed, a few days after he averted his eyes from the breadline, he wrote, “It is difficult to state the position of the civil population of Belgium without becoming hysterical.” The sight of ragged and hungry children especially bothered him, and he soon inaugurated a program of daily hot meals of bread and cocoa at Belgian schools.

By 1915, Hoover is, indeed, feeding millions of Belgians, indefinitely, using only private funding. He is also almost broke. Millions of Brits and Americans have given him contributions, from tycoons donating fortunes to ordinary people donating their wages, but it’s not enough. His expenses pass $5 million a month, which would be about $100 million today; all these bills are starting to catch up to him. In an act of supreme sacrifice, Hoover pledges his entire personal fortune as collateral for the Committee’s loans, then takes out more money. The grain shipments continue to flow, but his credit is at its end.

He continues beating on the doors of every government official he can find – British, German, American – demanding help. They all say their budgets are already occupied with the war effort. He begs them, lectures them, tells them that millions of people are doing to die. He goes all the way to the top, finagling an opportunity to meet with British Prime Minister David Lloyd George. Lloyd George later calls Hoover’s presentation “the clearest he had [ever] heard on any subject”, but he can offer only moral support.

What finally works is going to Germany and meeting with their top military brass. The brass are unimpressed; they still think that Belgium starving is as likely to help them as hinder. But the contact spooks top British officials, who agree to meet with Hoover again. Hoover feeds them carefully crafted lies, saying that the German brass have told him that British aid to Belgium would be a disaster to the Central Powers and so they, the Germans, are going to fund everything Hoover wants and more. “Oh no they don’t!” say the British, who promise to give Hoover even more funding than his imaginary German partners. The Committee for the Relief Of Belgium is finally back in the black. And what a black it is:

The scope and powers of the Committee For Relief of Belgium were mindboggling. Its shipping fleet flew its own flag. Its members carried special documents that served as CRB passports. Hoover himself was granted a form of diplomatic immunity by all belligerents, with the British permitting him to cross the Channel at will and the Germans providing him a document saying ‘this man is not to be stopped anywhere under any circumstances’. Hoover had privileged access to generals, diplomats, and ministers. He enjoyed personal contacts with the heads of warring governments. He negotiated treaties with the belligerents, advised them on policy, and delivered private messages among them. Great Britain, France, and Belgium would soon be turning over to him $150 million a year, enough to run a small country, and taking nothing for it beyond his receipt. As one British official observed, Hoover was running ‘a piratical state organized for benevolence.’

In 1917, America enters World War I. Hoover is no longer neutral and so has to resign from the CRB. He returns to the US a war hero. The New York Times proclaims Hoover’s CRB work “the greatest American achievement of the last two years.” There is talk that he should run for President. Instead, he goes to Washington and tells President Woodrow Wilson he is at his service.

Wilson is working on the greatest mobilization in American history. He realizes one of the US’ most important roles will be breadbasket for the Allied Powers, and names Hoover “food commissioner”, in charge of ensuring that there is enough food to support the troops, the home front, and the other Allies. His powers are absurdly vast – he can do anything at all related to the nation’s food supply, from fixing prices to confiscating shipments to telling families what to eat. The press affectionately dubs him “Food Dictator” (I assume today they would use “Food Czar”, but this is 1917 and it is Too Soon).

Hoover displays the same manic energy he showed in Belgium. His public relations blitz telling families to save food is so successful that the word “Hooverize” enters the language, meaning to ration or consume efficiently. But it turns out none of this is necessary. Hoover improves food production and distribution efficiency so much that no rationing is needed, America has lots of food to export to Europe, and his rationing agency makes an eight-digit profit selling all the extra food it has.

By 1918, Europe is in ruins. The warring powers have declared an Armistice, but their people are starving, and winter is coming on fast. Also, Herbert Hoover has so much food that he has to swim through amber waves of grain to get to work every morning. Mountains of uneaten pork bellies are starting to blot out the sky. Maybe one of these problems can solve the other? President Wilson dispatches Hoover to Europe as “special representative for relief and economic rehabilitation”. Hoover rises to the challenge:

Hoover accepted the assignment with the usual claim that he had no interest in the job, simultaneously seeking for himself the broadest possible mandate and absolute control. The broad mandate, he said, was essential, because he could not hope to deliver food without refurnishing Europe’s broken finance, trade, communications, and transportation systems…

Hoover had a hundred ships filled with food bound for neutral and newly liberated parts of the Continent before the peace conferences were even underway. He formalized his power in January 1919 by drafting for Wilson a post facto executive order authorizing the creation of the American Relief Administration (ARA), with Hoover as its executive director, authorized to feed Europe by practically any means he deemed necessary. He addressed the order to himself and passed it to the president for his signature…

The actual delivery of relief was ingeniously improvised. Only Hoover, with his keep grasp of the mechanics of civilization, could have made the logistics of rehabilitating a war-ravaged continent look easy. He arranged to extend the tours of thousands of US army officers already on the scene and deployed them as ARA agents in 32 different countries. Finding Europe’s telegraph and telephone services a shambles, he used US Navy vessels and Army Signal Corps employees to devise the best-functioning and most secure wireless system on the continent. Needing transportation, Hoover took charge of ports and canals and rebuilt railroads in Central and Eastern Europe. The ARA was for a time the only agency that could reliably arrange shipping between nations…

The New York Times said it was only apparent in retrospect how much power Hoover wielded during the peace talks. “He has been the nearest approach Europe has had to a dictator since Napoleon.”

Once again, Hoover faces not only the inherent challenge of feeding millions, but opposition from the national governments he is trying to serve. Britain and France plan to let Germany starve, hoping this will decrease its bargaining power at Versailles. They ban Hoover from transporting any food to the defeated Central Powers. Hoover, “in a series of transactions so byzantine it was impossible for outsiders to see exactly what he was up to”, causes some kind of absurd logistics chain that results in 42% of the food getting to Germany in untraceable ways.

He is less able to stop the European powers’ controlled implosion at Versailles. He believes 100% in Woodrow Wilson’s vision of a fair peace treaty with no reparations for Germany and a League Of Nations powerful enough to prevent any future wars. But Wilson and Hoover famously fail. Hoover predicts a second World War in five years (later he lowers his estimate to “thirty days”), but takes comfort in what he has been able to accomplish thus far.

He returns to the US as some sort of super-double-war-hero. He is credited with saving tens of millions of lives, keeping Europe from fraying apart, and preventing the spread of Communism. He is not just a saint but a magician, accomplishing feats of logistics that everyone believed impossible. John Maynard Keynes:

Never was a nobler work of disinterested goodwill carried through with more tenacy and sincerity and skill, and with less thanks either asked or given. The ungrateful Governments of Europe owe much more to the statesmanship and insight of Mr. Hoover and his band of American workers than they have yet appreciated or will ever acknowledge. It was their efforts…often acting in the teeth of European obstruction, which not only saved an immense amount of human suffering, but averted a widespread breakdown of the European system.

III.

Hoover wants to be president. It fits his self-image as a benevolent engineer-king destined to save the populace from the vagaries of politics. The people want Hoover to be president; he’s a super-double-war-hero during a time when most other leaders have embarrassed themselves. Even politicians are up for Hoover being president; Woodrow Wilson is incapacitated by stroke, leaving both Democrats and Republicans leaderless. The situation seems perfect.

Hoover bungles it. He plays hard-to-get by pretending he doesn’t want the Presidency, but potential supporters interpret this as him just literally not wanting the Presidency. He refuses to identify as either a Democrat or Republican, intending to make a gesture of above-the-fray non-partisanship, but this prevents either party from rallying around him. Also, he might be the worst public speaker in the history of politics.

Warren G. Harding, a nondescript Senator from Ohio, wins the Republican nomination and the Presidency. Hoover follows his usual strategy of playing hard-to-get by proclaiming he doesn’t want any Cabinet positions. This time it works, but not well: Harding offers him Secretary of Commerce, widely considered a powerless “dud” position. Hoover accepts.

Harding is famous for promising “return to normalcy”, in particular a winding down of the massive expansion of government that marked WWI and the Wilson Administration. Hoover had a better idea – use the newly-muscular government to centralize and rationalize America. In his first few years in Commerce – hitherto a meaningless portfolio for people who wanted to say vaguely pro-prosperity things and then go off and play golf – Hoover instituted/invented housing standards, traffic safety standards, industrial standards, zoning standards, standardized electrical sockets, standardized screws, standardized bricks, standardized boards, and standardized hundreds of other things. He founded the FAA to standardize air traffic, and the FCC to standardize communications. In order to learn how his standards were affecting the economy, he founded the NBER to standardize government statistics.

But that isn’t enough! He mediates an inter-state conflict over water rights to the Colorado River, even though that would normally be a Department of the Interior job. He solves railroad strikes, over the protests of the Department of Labor. “Much to the annoyance of the State Department, Hoover fielded his own foreign service”. He proposes to transfer 16 agencies from other Cabinet departments to the Department of Commerce, and when other Secretaries shoot him down, he does all their jobs anyway. The press dub him “Secretary of Commerce and Undersecretary Of Everything Else”.

Hoover’s greatest political test comes when the market crashes in the Panic of 1921. The federal government has previously ignored these financial panics. Pre-Wilson, it was small and limited to its constitutional duties – plus nobody knew how to solve a financial panic anyway. Hoover jumps into action, calling a conference of top economists and moving forward large spending projects. More important, he is one of the first government officials to realize that financial panics have a psychological aspect, so he immediately puts out lots of press releases saying that economists agree everything is fine and the panic is definitely over. He takes the opportunity to write letters saying that Herbert Hoover has solved the financial panic and is a great guy, then sign President Harding’s name to them. Whether or not Hoover deserves credit, the panic is short and mild, and his reputation grows.

While everyone else obsesses over his recession-busting, Hoover’s own pet project is saving the Soviet Union. Several years of civil war, communism, and crop failure have produced mass famine. Most of the world refuses to help, angry that the USSR is refusing to pay Czarist Russia’s debts and also pretty peeved over the whole Communism thing. Hoover finds $20 million to spend on food aid for Russia, over everyone else’s objection:

Russian relief would prove less popular than the Belgian variety, with the left accusing Hoover of seeking to undermine communism with capitalist aid…and the right charging him with rescuing and legitimating the shaky Soviet regime. Hoover gave the same answer to all critics: ‘Twenty million people are starving. Whatever their politics, they shall be fed.’

Maxim Gorky, in Italy nursing his tuberculosis, wrote Hoover personally: ‘In the past year you have saved from death three and one-half million children, five and one-half million adults. In the history of practical humanitarianism I know of no accomplishment which in…magnitude and generosity can be compared to the relief you have actually accomplished.

So passed the early 1920s. Warren Harding died of a stroke, and was succeeded by Vice-President “Silent Cal” Coolidge. Coolidge won re-election easily in 1924. Hoover continued shepherding the economy (average incomes will rise 30% over his eight years in Commerce), but also works on promoting Hooverism, his political philosophy. It has grown from just “benevolent engineers oversee everything” to something kind of like a precursor of modern neoliberalism:

Hoover’s plan amounted to a complete refit of America’s single gigantic plant, and a radical shift in Washington’s economic priorities. Newsmen were fascinated by is talk of a ‘third alternative’ between ‘the unrestrained capitalism of Adam Smith’ and the new strain of socialism rooting in Europe. Laissez-faire was finished, Hoover declared, pointing to antitrust laws and the growth of public utilities as evidence. Socialism, on the other hand, was a dead end, providing no stimulus to individual initiative, the engine of progress. The new Commerce Department was seeking what one reporter summarized as a balance between fairly intelligent business and intelligently fair government. If that were achieved, said Hoover, ‘we should have given a priceless gift to the twentieth century.’

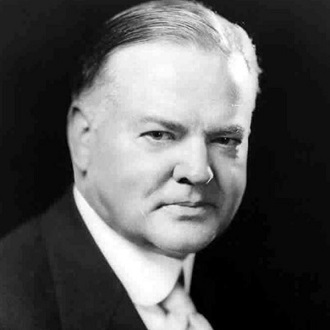
Finally, it is 1928. Hoover feels like he has accomplished his goal of becoming the sort of knowledgeable political insider who can run for President successfully. Silent Cal decides not to run for a second term (in typical Coolidge style, he hands a piece of paper to a reporter saying “I do not choose to run for President in 1928” and then disappears and refuses to answer further questions). The Democrats nominate Al Smith, an Irish-Italian Catholic with a funny accent; it’s too early for the country to really be ready for this. Historians still debate whether Hoover and/or his campaign deserves blame for being racist or credit for being surprisingly non-racist-under-the-circumstances.

The main issue is Prohibition. Smith, true to his roots, is against. Hoover, true to his own roots (his mother was a temperance activist) is in favor. The country is starting to realize Prohibition isn’t going too well, but they’re not ready to abandon it entirely, and Hoover promises to close loopholes and fix it up. Advantage: Hoover.

The second issue is tariffs. Everyone wants some. Hoover promises that if he wins, he will call a special session of Congress to debate the tariff question. Advantage: Hoover.

The last issue is personality. Republican strategists decide the best way for their candidate to handle his respective strengths and weaknesses is not to campaign at all, or be anywhere near the public, or expose himself to the electorate in any way. Instead, they are “selling a conception. Hoover was the omnicompetent engineer, humanitarian, and public servant, the ‘most useful American citizen now alive.’ He was an almost supernatural figure, whose wisdom encompasses all branches, whose judgment was never at fault, who knew the answers to all questions.” Al Smith is supremely charismatic, but “boasted of never having read a book”. Advantage: unclear, but Hoover’s strategy does seem to work pretty well for him. He racks up most of the media endorsements. TIME Magazine offers a rare dissent, saying that “In a society of temperate, industrious, unspectacular beavers, such a beaver-man would make an ideal King-beaver. But humans are different.”

Apparently not that different. Hoover wins 444 votes to 87, one of the greatest electoral landslides in American history.



You may not like it, but this is what peak presidentialness looks likeAnne McCormick of the New York Times describes the inauguration:

We were in a mood for magic…and the whole country was a vast, expectant gallery, its eyes focused on Washington. We had summoned a great engineer to solve our problems for us; now we sat back comfortable and confidently to watch our problems being solved. The modern technical mind was for the first time at the head of a government. Relieved and gratified, we turned over to that mind all of the complications and difficulties no other had been able to settle. Almost with the air of giving genius its chance, we waited for the performance to begin.

IV.

Herbert Hoover spent his entire presidency miserable.

First, he has no doubt that the economy is going to crash. It’s been too good for too long. He frantically tries to cool down the market, begs moneylenders to stop lending and bankers to stop banking. It doesn’t work, and the Federal Reserve is less concerned than he is. So he sits back and waits glumly for the other shoe to drop.

Second, he hates politics. Somehow he had thought that if he was the President, he would be above politics and everyone would have to listen to him. The exact opposite proves true. His special session of Congress comes up with the worst, most politically venal tariff bill imaginable. Each representative declares there should be low tariffs on everything except the products produced in his own district, then compromises by agreeing to high tariffs on everything with good lobbyists. The Senate declares that the House of Representatives is corrupt nincompoops and sends the bill back in disgust. Hoover has no idea how to solve this problem except to ask the House to do some kind of rational economically-correct calculation about optimal tariffs, which the House finds hilarious. “Opposed to the House bill and divided against itself, the Senate ran out the remaining seven weeks [of the special session] in a debauch of taunts, accusations, recriminations, and procedural argument.” The public blames Hoover, pretty fairly – a more experienced president would have known how to shepherd his party to a palatable compromise.

Also, there are crime waves, prison riots, bootlegging, and a heat wave during which Washington DC is basically uninhabitable. Also, at one point the White House is literally on fire.

…and then the market finally crashes. Hoover is among the first to call it a Depression instead of a Panic – he thinks the new term might make people panic less. But in fact, people aren’t panicking. They assume Hoover has everything in hand.

At first he does. He gathers the heads of Ford, Du Pont, Standard Oil, General Electric, General Motors, and Sears Roebuck and pressures them to say publicly they won’t fire people. He gathers the AFL and all the union heads and pressures them to say publicly they won’t strike. He enacts sweeping tax cuts, and the Fed enacts sweeping rate cuts. Everyone is bedazzled:

The sweep and speed of Washington’s response to the crash, which gave the impression that Hoover had ‘thoroughly anticipated the debacle and mapped out the shortest road to recovery’, was hailed in the press as an entirely new approach to management of the nation’s economic affairs.” Herald-Tribune: “President Hoover’s prompt action to prevent the depression extending to business and industry saved the situation. The panic was checked in a few days. Wages were left unaffected, stabilization was insured; production was encouraged to continue as usual. This leadership was all the more notable, since it was practically the first of the sort ever to originate in the White House.

And:

Economic joined journalists in congratulating Hoover on what was easily the most sophisticated response to a major economic event by any administration. ‘For the first time in our history,’ wrote Keynesian forerunners William Foster and Waddill Catchings, ‘a president is taking aggressive leadership in guiding private business through a crisis.

Six months later, employment is back to its usual levels, the stock market is approaching its 1929 level, and Democrats are fuming because they expect Hoover’s popularity to make him unbeatable in the midterms. I got confused at this point in the book – did I accidentally get a biography from an alternate timeline with a shorter, milder Great Depression? No. I do think I accidentally got a biography by someone obsessed with defending Hoover at any cost and willing to stray into revisionist history to do it. As per Whyte, Hoover would take some brilliant and decisive action. Economists would praise him. The economy would start to look better. Everyone would declare the problem solved – especially Hoover, sensitive both to his own reputation and to the importance of keeping economic optimism high. Then for reasons totally outside the President’s control, the recovery would stall, or reverse, or something else would go wrong.

People are still debating what made the Great Depression so long and hard. Whyte’s theory, insofar as he has one at all, is “one thing after another”. Every time the economy started to go up (thanks to Hoover), there was another shock. Most of them involved Europe – Germany threatening to default on its debts, Britain going off the gold standard. A few involved the US – the Federal Reserve made some really bad calls. The one thing Whyte is really sure about is that his idol Herbert Hoover was totally blameless.

He argues that Hoover’s bank relief plan could have stopped the Depression in its tracks – but that Congressional Democrats intent on sabotaging Hoover forced the plan to publicize the names of the banks applying. The Democrats hoped to catch Hoover propping up his plutocrat friends – but the change actually had the effect of making banks scared to apply for funding and panicking the customers of banks that were known to have applied. He argues that the “Hoover Holiday” – a plan to grant debt relief to Germany, taking some pressure off the clusterf\*\*k that was Europe – was a masterstroke, but that France sabotaged it in the interests of bleeding a few more pennies from its arch-rival. International trade might have sparked a recovery – except that Congress finally passed the Hawley-Smoot Tariff, the end result of the corruption-plagued tariff negotiations, just in time to choke it off.

Whyte saves his barbs for the real villain: FDR. If the book is to be believed, Hoover actually had things pretty much under control by 1932. Employment was rising, the stock market was heading back up. FDR and his fellow Democrats worked to tear everything back down so he could win the election and take complete credit for the recovery. The wrecking campaign entered high gear after FDR won in 1932; he was terrified that the economy might get better before he took office, and used his President-Elect status to hint that he was going to do all sorts of awful things. The economy got skittish again and obediently declined, allowing him to get inaugurated at the precise lowest point and gain the credit for recovery he so ardently desired.

For example: November 1932. Hoover has just lost the election, but is a lame duck until March. The European debt crisis flashes up again. Hoover knows how to solve it. But:

He had already met with congressional leaders and learned, as he had suspected, that they would not change their stance without Roosevelt’s support. Seized with the urgency of the moment, he continued to bombard his opponents with proposals for cooperation toward solutions, going so far as to suggest that Democratic nominees, not Republicans, be sent to Europe to engage in negotiations, all to no avail. Notwithstanding what editorialists called his “personal and moral responsibility” to engage with the outgoing administration, Roosevelt had instructed Democratic leaders in Congress not to let Hoover “tinker” with the debts. He had also let it be known that any solution to the problem would occur on his watch – “Roosevelt holds he and not Hoover will fix debt policy”, read the headlines. Thus ended what the New York Times called Hoover’s magnanimous proposal for “unity and constructive action”, not to mention his 12-year effort to convince America of its obligation and self-interest in fostering European political and financial stability…

During the debt discussions and to some extent as a result of them, the economy turned south again. Several other factors contributed. Investors were exchanging US dollars for gold as doubt spread about Roosevelt’s intentions to remain on the gold standard. Gold stocks in the Federal Reserve thus declined, threatening the stability of the financial sector…what’s more, the effectiveness of [Hoover’s bank support plan], which had succeeded in stabilizing the banking system, was severely compromised by [Democrats’] insistence on publicizing its loans, as the administration had warned. For these reasons, Hoover would forever blame Roosevelt and the Democratic Congress for spoiling his hard-earned recovery, an argument that has only recently gained currency among economists.

And:

Alarmed at these threats to recovery, Hoover pushed Democratic congressional leaders and the incoming administration for action. He wanted to cut federal spending, reorganize the executive branch to save money, reestablish the confidentiality of RFC loans, introduce bankruptcy legislation to protect foreclosures, grant new powers to the Federal Reserve, and pass new banking regulation, including measures to protect depositors…He was frustrated at every turn by Democratic leadership taking cues from the President-Elect…On February 5, Congress took the obstructionism a degree further by closing shop with 23 days left in its session.

In mid-February, there is another run on the banks, worse than all the other runs on the banks thus far. Hoover asks Congress to do something – Congress says they will only listen to President-Elect Roosevelt. Hoover writes a letter to Roosevelt begging him to give Congress permission to act, saying it is a national emergency and he has to act right now. Roosevelt refuses to respond to the letter for eleven days, by which time the banks have all failed.

Then, a month later, he stands up before the American people and says they have nothing to fear but fear itself – a line he stole from Hoover – and accepts their adulation as Destined Savior. He keeps this Destined Savior status throughout his administration. In 1939, Roosevelt still had everyone convinced that Hoover was totally discredited by his failure to solve the Great Depression in three years – whereas Roosevelt had failed to solve it for six but that was totally okay and he deserved credit for being a bold leader who tried really hard.

So how come Hoover bears so much of the blame in public consciousness? Whyte points to three factors.

First, Hoover just the bad luck of being in office when an international depression struck. Its beginning wasn’t his fault, its persistence wasn’t his fault, but it happened on his watch and he got blamed.

Second, in 1928 the Democratic National Committee took the unprecedented step of continuing to exist even after a presidential election. It dedicated itself to the sort of PR we now take for granted: critical responses to major speeches, coordinated messaging among Democratic politicians, working alongside friendly media to create a narrative. The Republicans had nothing like it; the RNC forgot to exist for the 1930 midterms, and Hoover was forced to personally coordinate Republican campaigns from his White House office. Although Hoover was good (some would say obsessed) at reacting to specific threats on his personal reputation, the idea of coordinating a media narrative felt too much like the kind of politics he felt was beneath him. So he didn’t try. When the Democrats launched a massive public blitz to get everyone to call homeless encampments “Hoovervilles”, he privately fumed but publicly held his tongue. FDR and the Democrats stayed relentlessly on message and the accusation stuck.

And third, Hoover was dead-set against welfare. However admirable his attempts to reverse the Depression, stabilize banking, etc, he drew the line at a national dole for the Depression’s victims. This was one of FDR’s chief accusations against him, and it was entirely correct. Hoover suspected that going down that route would lead pretty much where it led Roosevelt – to a dectupling of the size of government and the abandonment of the Constitutional vision of a small federal government presiding over substantially autonomous states. He decided it wasn’t worth it. So Herbert Hoover, history’s greatest philanthropist and ender-of-famines, would go down in history as the guy who refused to feed starving people. And they hated him for it.

V.

Some people might call Herbert Hoover a sore loser. But he argues that no, it’s totally reasonable for him to spend the rest of his life attacking FDR and trying to destroy his legacy.

His theory, explained in the countless books, pamphlets, and speeches that he spends his post-presidential life writing, is that FDR came from the same cloth as Hitler and Stalin. The miseries of the Great Depression, the centralizing tendencies of the age, the rise of mass media, and the collapse of republican virtue were combining all around the world in a monstrous reaction against the cause of liberty. “Daily,” wrote Hoover, “the world goes back to the regimentation of the Middle Ages, whether it be Bolshevism, Hitlerism, Fascism, or the New Deal.”

He has more! “[The New Deal] has no philosophy. It is sheer opportunism, a muddle of a spoils system, of reckless adventure, of unctuous claims to a monopoly of human sympathy, of greed for power, of a desire for popular acclaim and an aspiration to make the front pages of the newspapers.” He has more! “The New Deal has contributed to sapping our stamina and making us soft…the road to regeneration is burdensome and hard. It is straight and simple. It is a road paved with work and with sacrifice and consecration to the indefinable spirit that is America.”

He has more! He just keeps going like this, again and again. FDR, for his part, seems slightly befuddled. He tried offering Hoover a position coordinating the US effort to help war refugees – which Hoover turned down, assuming anything from FDR was a trick. Hoover just keeps shouting and fulminating and writing more and more books and pamphlets until FDR dies – which enrages Hoover, who wanted him to “live long enough to reap what he had sown”.

Whyte’s theory is that this period of Hoover’s life sowed the seeds for the modern conservative movement: “Modern American conservatism, conceived as an antidote to the New Deal, was born on December 16, 1937, with Hoover as its prophet and philosopher.” He doesn’t do much to back this theory up, and Hoover gets all of a paragraph in Wikipedia’s long History of conservativism in the United States. We are left to piece it together from a few mentions here and there – Hoover befriending and helping a young William F. Buckley, Hoover giving a key endorsement to Barry Goldwater, and of course the namesake Hoover Institution that he founded, funded, and guided until his death.

I have to admit this is a hole in my understanding. Smart people definitely say that modern American conservativism began with Buckley and Goldwater and their friends, but what does this mean? Hasn’t about half of America been conservative since the 1700s? Hasn’t a philosophy of small government, individual freedom, and capitalist economics been pretty fundamental to America since its beginning? I’m not sure, and without this knowledge I don’t feel qualified to judge Hoover’s role.

Hoover passes in 1964, ninety years old. He lived long enough to become a hero to a new brand of conservative who considered him an intellectual forebear, and through various acts of public service to win back the love of his country. He had not quite finished his magnum opus, Freedom Betrayed. In 2012, historians finally dug it up, revised it, and released it to the world. It turned out to be 957 pages of him attacking Franklin Roosevelt. Give Herbert Hoover credit: he died as he lived.

VI.

I’m sorry this review was so long. I couldn’t bear to make it any shorter. I find the whole story so fascinating, and I just regret I couldn’t include more. I didn’t even get a chance to mention the time Hoover rescued the US South from the Great Mississippi Flood, or the time he discovered ancient ruins in the jungles of Burma, the time a 71-year-old Hoover was called back into service by President Truman to solve another post-World-War famine, or the time he invented the new sport of Hooverball (now part of the popular CrossFit exercise program).



Herbert Hoover on a famine relief tour of Poland, along with some of the children he is helping.Hoover was a man who did everything wrong. He was the quintessential High Modernist. He was arrogant, he was authoritarian, he didn’t listen to anyone, he put no effort into pleasing people or making his ideas more palatable. He never solicited stakeholders’ opinions. He lied like a rug, constantly and egregiously. He lived his life like a caricature of exactly the sort of person who should fail at philanthropy and become a horror story to warn future generations.

But he won anyway. He started from a measly few million dollars and beat out Rockefellers and Carnegies to become the most successful philanthropist in early 20th century history. Whyte’s estimate of 100 million lives saved seems much too high; there were only 100 million people in Europe total during the relevant period. But even during his own time, people universally credited him with saving millions. And he did it again and again and again. I didn’t even have space to talk about the time he saved the Southern United States from a giant flood, or half a dozen other impressive accomplishments. Maybe the rules are wrong. Maybe all of this stuff about how authoritarian approaches never work, and you need to let the people you are helping lead the way, is all just modern prejudices, and putting a brilliant and very rich engineer in charge of a hypercentralized organization is just as good as any other way of doing things.

But even this I find less interesting than his psychology. He combined a personal callousness with a love for all humanity. When he was inspecting mines in Australia, he fired the worst-performing X% of workers. One worker begged him to reconsider – he had a family to support. Hoover raised $300 for the man’s family – a lot of money at the time! Probably more than Hoover made in a month! – but fired him anyway. In 1932, when the Bonus Army marched on Washington, Hoover was adamant that he would not give these men – poor, starving veterans – a single cent more than they were entitled to by their existing benefits. But he also instructed his staff to go around to their encampments and give them food and supplies in secret.

Sometimes his stubbornness calls to mind the fictional Inspector Javert, who refuses to bend the law for any reason. In this model, Hoover sympathizes with everybody, but his honor forbids him to bend the rules in favor of underperforming employees or protesters who want more than their contracts entitle them to. But this picture of a hyper-honorable Hoover crashes into his constant willingness to lie, cheat, and bend the rules in his own favor. Sometimes his lies are for the greater good, like when he tells Britain that Germany is preparing to feed Belgium. Other times they seem entirely selfish, like his various Chinese mining scams. The best that can be said about Hoover is that if he decides a principle is involved, he sticks to it.

And this is actually really good! Again and again through the book, Hoover feels like the only person with a moral compass. When it is in everyone’s strategic interest to let Belgium starve, Hoover is the only one who is able to keep fixated on the potential human toll. When it is in everyone’s interest to let the USSR starve, only Hoover – despite his fanatical anti-communism – is able to stick to the frame where the Russians are human beings and politics is beside the point. When Americans are starving during the Great Depression…

…okay, Hoover totally dropped the ball on that one. In fact, one of his Democratic opponents wrote something about how maybe if unemployed American workers pretended to be Belgians, they could get Hoover’s sympathy. I don’t have a great explanation for this. But Hoover’s weak and inconsistent sympathies are often enough to let him outdo everyone else. Or at least, he is uncorrelated with everyone else and succeeds when they fail. Again and again Hoover is accused of treating people like numbers on a piece of paper. But if this is true, it seems to be linked to the reverse talent – the ability to remember that numbers on a piece of paper represent people, even when other people would rather forget.

I’m equally confused about Hoover’s politics, although it’s not really his fault. The whole era confuses me. The Progressives, Hoover’s own faction, seem clearly related to modern progressives. But they also give me more of a technophile, rationalist feel than their modern counterparts. Am I imagining things? If not, where did this go?

And how did Hoover so deftly merge his centralizing technocratic engineer side with his small-government individual-freedom pro-capitalism side? Maybe it wasn’t that deft? Maybe he started his life as a centralizing technocrat, then made a 180 after becoming a small-government individualist helped him dunk on FDR more effectively? But it didn’t feel that way. It felt like all of it was coming from some central set of core beliefs throughout his life.

My confusion here feels similar to my confusion about Tyler Cowen’s “state capacity libertarianism”. Creating a strong and effective state is certainly…a goal you can have. But I don’t understand the argument for calling this a libertarian project. At best, it’s a project not entirely opposed to libertarianism. Still, perhaps this is my ignorance. Cowen thinks that strengthening the state and instituting effective technocratic government can be allied to a small-government individualistic market-based philosophy. Whatever he’s smoking, maybe Herbert Hoover was smoking the same thing.

I get the impression that Kenneth Whyte is a bit of a revisionist historian, too sympathetic to his subject to tell his story the way everyone else does. But at least in Whyte’s telling, the Hoover presidency was a great missed opportunity, or at least a fulcrum of history. If a few key economic events had been a few months off in one direction or the other, FDR might have been a footnote to history, and a four-term President Hoover might have left an indelible mark on America. Instead of a New Deal, we might have gotten a optimistic small-government technocratic meritocracy that was able to merge the best aspects of a dying frontier America with the best aspects of the industrial age.

In one of the most poignant passages in the book, Commerce Secretary Hoover fires back at his socialist critics. He points out that of the top dozen US officials – the President, VP, and ten Cabinet Secretaries – eight, including himself, had begun as manual laborers and worked their way up. That was the America Hoover was working to defend. He lost, and now we have this shitshow. But it’s hard to begrudge him the attempt.

# Face Masks: Much More Than You Wanted To Know

There’s been recent controversy about the use of face masks for protection against coronavirus. Mainstream sources, including the CDC and most of the media say masks are likely useless and not recommended. They’ve recently been challenged, for example by Professor Zeynep Tufekci in the New York Times and by Jim and Elizabeth on Less Wrong. There was also some debate in the comment section here last week, so I promised I’d look into it in more depth.

As far as I can tell, both sides agree on some points.

They agree that N95 respirators, when properly used by trained professionals, help prevent the wearer from getting infected.

They agree that surgical masks help prevent sick people from infecting others. Since many sick people don’t know they are sick, in an ideal world with unlimited mask supplies everyone would wear surgical masks just to prevent themselves from spreading disease.

They also agree that there’s currently a shortage of both surgical masks and respirators, so for altruistic reasons people should avoid hoarding them and give healthcare workers first dibs.

But they disagree on whether surgical masks alone help prevent the wearer from becoming infected, which will be the focus of the rest of this piece.

1. What are the theoretical reasons why surgical masks might or might not work?

Epidemiologists used to sort disease transmission into three categories: contact, droplet, and airborne. Contact means you only get a disease by touching a victim. This could be literally touching them, or a euphemism for very explicit contact like kissing or sex. Droplet means you get a disease when a victim expels disease-laden particles into your face, usually through coughing, sneezing, or talking. Airborne means you get a disease because it floats in the air and you breathe it in. Transmission via “fomites”, objects like doorknobs and tables that a victim has touched and left their germs on, is a bonus transmission route that can accompany any of these other methods.

More recently, scientists have realized that droplet and airborne transmission exist along more of a spectrum. Droplets can stay in the air for more or less time, and spread through more or less volume of space before settling on the ground. The term for this new droplet-airborne spectrum idea is “aerosol transmission”. Diseases with aerosol transmission may be spread primarily through droplets, but can get inhaled along with the air too. This concept is controversial, with different authorities having different opinions over which viruses can be aerosolized. It looks like most people now believe aerosol transmission is real and applicable to conditions like influenza, SARS, and coronavirus.

Surgical masks are loose pieces of fabric placed in front of the mouth and nose. They offer very good protection against outgoing droplets (eg if you sneeze, you won’t infect other people), and offer some protection against incoming droplets (eg if someone else sneezes, it doesn’t go straight into your nose). They’re not airtight, so they offer no protection against airborne disease or the airborne component of aerosol diseases.

Respirators are tight pieces of fabric that form a seal around your mouth and nose. They have various “ratings”; N95 is the most common, and I’ll be using “N95 respirator” and “respirator” interchangably through most of this post even though that’s not quite correct. When used correctly, they theoretically offer protection against incoming and outgoing droplet and airborne diseases; since aerosol diseases are a combination of these, they offer generalized protection against those too. Hospitals hate the new “aerosol transmission” idea, because it means they probably have to switch from easy/cheap/comfortable surgical masks to hard/expensive/uncomfortable respirators for a lot more diseases.

Theory alone tells us surgical masks should not provide complete protection. Coronavirus has aerosol transmission, so it is partly airborne. Since surgical masks cannot prevent inhalation of airborne particles, they shouldn’t offer 100% safety against coronavirus. But theory doesn’t tell us whether they might not offer 99% safety against coronavirus, and that would still be pretty good.

2. Are people who wear surgical masks less likely to get infected during epidemics?

It’s unethical to randomize people to wear vs. not-wear masks during a pandemic, so nobody has done this. Instead we have case-control studies. After the pandemic is over, scientists look at the health care workers who did vs. didn’t get infected, and see whether the infected people were less likely to wear masks. If so, that suggests maybe the masks helped.

This is an especially bad study design, for two reasons. First, it usually suffers recall bias – if someone wore a mask inconsistently, then they’re more likely to summarize this as “didn’t wear masks” if they got infected, and more likely to summarize it as “did wear masks” if they stayed safe. Second, probably some nurses are responsible and do everything right, and other nurses are irresponsible and do everything wrong, and that means that if anything at all helps (eg washing your hands), then it will look like masks working, since the nurses who washed their hands are more likely to have worn masks. Still, these studies are the best we can do.

Gralton & McLaws, 2010 reviews several studies of this type, mostly from the SARS epidemic of the early 2000s. A few are underpowered and find that neither surgical masks nor respirators prevent infection (probably not true). A few others show respirators prevent infection, but do not investigate surgical masks (probably right, but useless for our purposes). Two seem relevant to the question of whether surgical masks work:

Rapid awareness and transmission of SARS in Hanoi French Hospital, Vietnam was conducted in a poor hospital that only had surgical masks, not respirators. In the latter stages of the epidemic, 4 workers got sick and 26 stayed healthy. It found that 3 of the 4 sick workers hadn’t been wearing masks, but only 1 of the 26 healthy workers hadn’t. This is a pretty dramatic result – subject to the above confounders, of course.

Effectiveness of precautions against droplets and contact in prevention of nosocomial transmission of SARS is larger and more prestigious, and looked at a cluster of five hospitals. Staff in these hospitals used a variety of mask types, including jury-rigged paper masks that no serious authority expects to work, surgical masks, and N95 respirators. It found that 7% of paper-mask-wearers got infected, compared to 0% of surgical-mask and respirator wearers. This seems to suggest that surgical masks are pretty good.

The meta-analysis itself avoided drawing any conclusions at all, and would not even admit that N95 respirators worked. It just said that more research was needed. Still, the two studies at least give us a little bit of evidence in surgical masks’ favor.

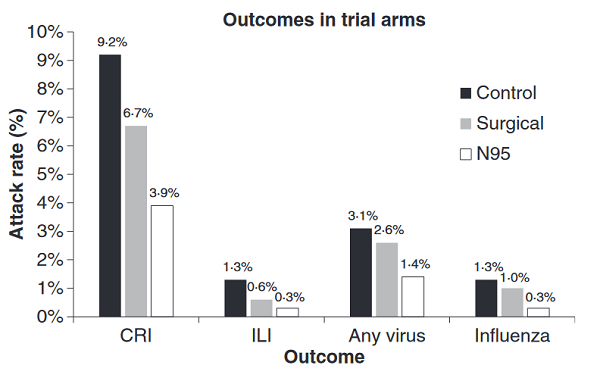
How concerned should we be that these studies looked at health care workers specifically? On the one hand, health care workers are ordinary humans, so what works for them should work for anyone else. On the other, health care workers may have more practice using these masks, or may face different kinds of situations than other people. Unlike respirators, surgical masks don’t seem particularly hard to use, so I’m not sure health care workers’ training really gives them an advantage here. Overall I think this provides some evidence that surgical masks are helpful.

I was able to find one study like this outside of the health care setting. Some people with swine flu travelled on a plane from New York to China, and many fellow passengers got infected. Some researchers looked at whether passengers who wore masks throughout the flight stayed healthier. The answer was very much yes. They were able to track down 9 people who got sick on the flight and 32 who didn’t. 0% of the sick passengers wore masks, compared to 47% of the healthy passengers. Another way to look at that is that 0% of mask-wearers got sick, but 35% of non-wearers did. This was a significant difference, and of obvious applicability to the current question.

3. Do surgical masks underperform respirators in randomized trials?

Usually it would be unethical to randomize health care workers to no protection, so several studies randomize them to face masks vs. respirators. But a few others were done in foreign hospitals where lack of protection was the norm, and these studies did include a no-protection control group.

MacIntyre & Chugtai 2015, Facemasks For The Prevention Of Infection In Healthcare And Community Settings, reviews four of these. Two of the four are unable to find any benefit of either masks or respirators. The third finds a benefit of respirators, but only if nobody tested the respirators to see if they fit, which doesn’t make sense and suggests it’s probably an artifact. The fourth finds a benefit of respirators, but not masks. It seems unlikely that respirators don’t help, so this suggests all these studies were underpowered. If we throw good statistical practice to the winds and just look at the trends, they look like this:



In other words, respirators are better than masks are better than nothing. It would be wrong to genuinely conclude this, because it’s not statistically significant. But it would also be wrong to conclude the studies show masks don’t work, because they mostly show respirators don’t work, and we (hopefully) know they do.

Overall these studies don’t seem very helpful and I’m reluctant to conclude anything from them. In section 6, I’ll talk more about why studies may not have shown any advantage for respirators.

4. Do surgical masks prevent ordinary people from getting infected outside the healthcare setting?

The same review lists nine randomized trials with a different design: when the doctor diagnoses you with flu, she either asks everyone in your family to wear masks (experimental group), or doesn’t do that (control group), and then checks how many family members in each group got the flu.

How did these go? That depends whether you use intention-to-treat or per-protocol analysis. Intention-to-treat means that you just compare number of infections in the assigned-to-wear-masks group vs. the control group. Per-protocol means that you only count someone in the study if they actually followed directions. So if someone in the assigned-to-wear mask group didn’t wear their mask, you remove them from the study; if someone in the control group went rogue and did wear a mask, you remove them too.

Both of these methods have their pros and cons. Per protocol is good because if you’re trying to determine the effect of wearing a mask, you would really prefer to only be looking at subjects who actually wore a mask. But it has a problem: adherence to protocol is nonrandom. The people who follow your instructions diligently are selected for being diligent people. Maybe they also diligently wash their hands, and diligently practice social distancing. So once you go per protocol, you’re no longer a perfect randomized controlled trial. Only intention-to-treat analyses carry the full weight of a gold standard RCT.

According to intention-to-treat, the studies unanimously found masks to be useless. But there were a lot of signs that intention-to-treat wasn’t the right choice here. Only about a fifth of people who were asked to wear masks did so with any level of consistency. The rest wore the mask for a few hours and then get bored and took it off. Honestly, it’s hard to blame them; these studies asked a lot from families. If a husband has flu, and sleeps in the same bed as his wife, are they both wearing masks all night?

Of the three studies that added per-protocol analyses, all three found masks to be useful (1, 2, 3) . Does this prove masks work? Not 100%; per-protocol analyses are inherently confounded. But it sure is suggestive.

The review author summarizes:

The routine use of facemasks is not recommended by WHO, the CDC, or the ECDC in the community setting. However, the use of facemasks is recommended in crowded settings (such as public transport) and for those at high risk (older people, pregnant women, and those with a medical condition) during an outbreak or pandemic. A modelling study suggests that the use of face-masks in the community may help delay and contain a pandemic, although efficacy estimates were not based on RCT data. Community masks were protective during the SARS outbreaks, and about 76% of the population used a facemask in Hong Kong.

There is evidence that masks have efficacy in the community setting, subject to compliance [13] and early use [12, 18, 19]. It has been shown that compliance in the household setting decreases with each day of mask use, however, which makes long term use over weeks or months a challenge […]

Community RCTs suggest that facemasks provide protection against infection in various community settings, subject to compliance and early use. For health-care workers, the evidence suggests that respirators offer superior protection to facemasks.

Parts of this summary are infuriating. If the big organizations recommend that especially vulnerable groups wear masks, aren’t they admitting masks work? But if they’re admitting masks work, why don’t they recommend them for ordinary people?

It looks like they’re saying masks work a little, they’re too annoying for it to be worth it for normal people, but they might be worth it for the especially vulnerable. But then why don’t they just say masks work, and let each person decide how much annoyance is worthwhile? I’m not sure. But it looks like the author basically ends up in favor of community use of surgical masks in a pandemic, mostly on the basis of per-protocol analyses of community RCTs.

5. How do surgical masks and respirators compare in hokey lab studies?

Our source here is Smith et al 2016, Effectiveness Of N95 Respirators Versus Surgical Masks In Protecting Health Care Workers From Acute Respiratory Infection: A Systematic Review And Meta-Analysis. They review some of the same studies we looked at earlier, but then investigate 23 “surrogate exposure studies”, ie throwing virus-shaped particles at different masks in a lab and seeing if they got through. You can find the results of each in their appendix. Typically, about 1 – 5% of particles make it through the respirator, and 10 – 50% make it through the surgical mask. They summarize this as:

In general, compared with surgical masks, N95 respirators showed less filter penetration, less face-seal leakage and less total inward leakage under the laboratory experimental conditions described.

I think in general the fewer virus particles get through your mask, the better, so I think this endorses surgical masks as better than nothing, since their failure rate was less than 100%.

Booth et al, 2013 examines surgical masks themselves more closely. They hook a surgical mask up to “a breathing simulator” and then squirt real influenza virus at it, finding that:

Live influenza virus was measurable from the air behind all surgical masks tested. The data indicate that a surgical mask will reduce exposure to aerosolised infectious influenza virus; reductions ranged from 1.1- to 55-fold (average 6-fold), depending on the design of the mask…the results demonstrated limitations of surgical masks in this context, although they are to some extent protective.

The paper doesn’t discuss how particle number maps to infection risk. Does letting a single influenza virus through mean you will get infected? If so, any reduction short of 100% is useless. I have a vague sense that this isn’t true; your immune system can fight off most viruses, and the fewer you get, the better the chance it will win. Also, even respirators don’t claim to reduce particle load by more than 99% or so, and those work, so it can’t be that literally a single virus will get you. Overall I think modest reductions in particle number are still pretty good, but I don’t have a study that proves it.

6. Is it true that the public won’t be able to use N95 respirators correctly?

Yes.

I remember my respirator training, the last time I worked in a hospital. They gave the standard two minute explanation, made you put the respirator on, and then made you go underneath a hood where they squirted some aerosolized sugar solution. If you could smell the sugar, your respirator was leaky and you failed. I tried so hard and I failed so many times. It was embarrassing and I hated it.

I’m naturally clumsy and always bad at that kind of thing. Some people were able to listen to the two minute explanation and then pass right away. Those kinds of people could probably also listen to a two minute YouTube explanation and be fine. So I don’t want to claim it’s impossible or requires lots of specialized background knowledge. It’s just a slightly difficult physical skill you have to get right.

Bunyan et al, 2013, Respiratory And Facial Protection: A Critical Review Of Recent Literature, discusses this in more depth. They review some of the same studies we reviewed earlier, showing no benefit of N95 respirators over surgical masks for health care workers in most situations. This doesn’t make much theoretical sense – the respirators should win hands down.

The most likely explanation is: doctors aren’t much better at using respirators than anyone else. In a California study of tuberculosis precautions, 65% of health care workers used their respirators incorrectly. That’s little better than the general public, who have a 76% failure rate. Bunyan et al note:

The fitting of N95 respirators has been the subject of many publications. The effective functioning of N95 respirators requires a seal between the mask and the face of the wearer. Variation in face size and shape and different respirator designs mean that a proper fit is only possible in a minority of health care workers for any particular mask. Winter et al. reported that, for any one of three widely used respirators, a satisfactory fit could be achieved by fewer than half of the healthcare workers tested, and for 28% of the participants none of the masks gave a satisfactory fit.

Fit-testing is a laborious task, taking around 30 min to do properly, and comprises qualitative fit-testing (testing whether the respirator-wearing healthcare worker can taste an intensely bitter or sweet substance sprayed into the ambient air around the outside of the mask) or quantitative fit testing (measuring the ratio of particles in the air inside and outside the breathing zone when wearing the respirator). Attempts have been made to circumvent the requirement for fit testing, and it has been suggested that self-testing for a seal by the respirator wearer (see http://youtu.be/pGXiUyAoEd8a for a video demonstration) is a sufficient substitute for fit-testing. However, self-checking for a seal has been demonstrated to be a highly unreliable technique in two separate studies so that full fit-testing remains a necessary preliminary requirement before respirators can be used in the healthcare setting.

Operationally, this presents significant challenges to organizations with many healthcare workers who require fit-testing. Chakladar et al. pointed out that, in addition to the routine need for repeat testing over time to ensure that changes in weight or facial hair have not compromised a good fit, movements of healthcare workers between organizations using different makes of respirators would necessitate additional repeat fit-testing. Fit-testing is likely to remain problematic to health-care organizations for the foreseeable future. In addition to the requirement for fit-testing, ‘fit-checking’ is also required each time the respirator is donned to ensure there are no air leaks.

Is a poorly-fitting N95 respirator better than nothing? The reviewed studies suggest that at that point it’s just a very fancy and expensive surgical mask.

7. Were the CDC recommendations intentionally deceptive?

No, and I owe them an apology here.

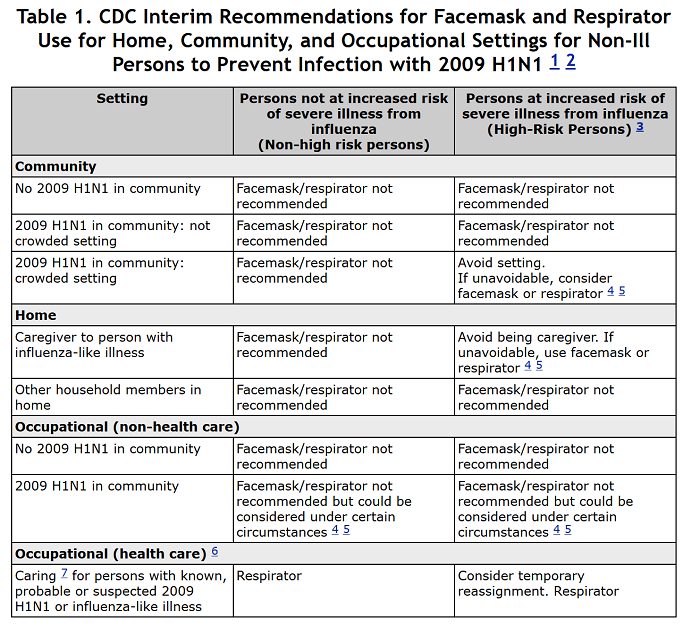
I think the evidence above suggests masks can be helpful. Masked health care workers were less likely to catch disease than unmasked ones. Masked travelers on planes were less likely to catch disease than unmasked ones. In per protocol analysis, masked family members are less likely to catch disease from an index patient than unmasked ones. Laboratory studies confirm that masks block most particles. All of this accords with a common-sense understanding of droplet and aerosol transmission of disease.

None of these, except maybe the plane study, tell us exactly what we want to know. The SARS studies were all done in a health care setting, so they don’t prove that regular people can benefit from masks. But health care workers are closely related to homo sapiens and ought to have similar anatomy and physiology. Surgical masks aren’t as complicated as respirators and we can assume most people get them right. And although health care workers are in unusually high-risk situations, that should just affect the magnitude of the benefit, not the sign; obviously the level of risk ordinary people encounter is sometimes relevant, considering they do often catch pandemic diseases. So our default assumption should be that these studies carry over, not that they don’t.

Likewise, most of the community studies were done on family members. Most guidelines already say to mask up if you have a sick family member, so talking about subways and crowds requires a little bit of extrapolation. But again, being in a family is just one form of close contact. It would take bizarre convolutions to even imagine a theory where you can catch diseases from your family members but not from people you sit next to on a train. Our default assumption should be in favor of these results generalizing, not against them.

But the CDC has recommended against mask use. I hypothesized that the CDC was intentionally lying to us, trying to trick us into not buying masks so there would be enough for health care workers.

But that can’t be true, because the CDC and other experts came up with their no-masks policy years ago, long before there was any supply shortage. For example, during the 2009 swine flu pandemic, their website offered the following table:



And during the 2015 MERS epidemic, NPR said South Koreans were wrong to wear masks:

Masks can be helpful for protecting health workers from a variety of infectious diseases, including MERS…

But either type of mask is less likely to do much good for the average person on the street…Wearing a mask might make people feel better. After all, MERS has killed about a third of the people known to be infected.

But there are no good studies looking at how well these masks prevent MERS transmission out in the community, says Geeta Sood, an infectious disease specialist at Johns Hopkins University. “On the street or the subway, for MERS specifically, they’re probably not effective,” she says. One problem is that the masks are loose fitting, and a lot of tiny airborne particles can get in around the sides of the masks.

So if studies generally suggest masks are effective, and the CDC wasn’t deliberately lying to us, why are they recommending against mask use?

I’m not sure. I haven’t been able to track down any documents where they discuss the reasons behind their policies. It’s possible they found different studies than I did, or interpreted the studies differently, or have some other superior knowledge.

But I think that more likely, they’re trying to do something different with medical communication. Consider legal communication. If a court declares a suspect is “not guilty”, that could mean that he is actually not guilty of the crime. Or it could mean that he did it but they can’t prove it. Or it could mean that he did it, they can prove it, but the police officer who found the proof didn’t have a warrant at the time so they had to throw it out. A legal communication like “this man is not guilty” is intended not just to convey information, but to formally reflect the output of a sacrosanct process.

Medicine has been traumatized by its century-long war with quackery, and ended up with its jargon also formally reflecting the output of a sancrosanct process. Remember, there are dozens of studies supposedly showing homeopathy works, not to mention even more studies proving telepathy exists. At some point you have to redesign all your institutions to operate in an environment of epistemic learned helplessness, and the result is very high standards of proof.

Masks haven’t quite reached these standards. The case-control trials look good, and the per-protocol RCTs look good, but there aren’t really the large-scale intention-to-treat RCTs that would be absolutely perfect. Even if these studies work, they only prove things about the health care setting and the family setting, not “the community setting” in general. So masks haven’t been proven to work beyond a reasonable doubt. Just like the legal term for “not proven guilty beyond a reasonable doubt” is “not guilty”, the medical communication term for “not proven effective beyond a reasonable doubt” is “not effective”. This already muddled communication gets even worse because doctors are constitutionally incapable of distinguishing “no evidence for” from “there is evidence against” – I have no explanation for this one.

There’s an even more complicated language-use issue. The CDC may be thinking of its recommendations not just as conveying an opinion but as taking an action – performing the medical intervention of recommending people wear masks. All of those RCTs listed above show that the medical intervention of recommending people wear masks is ineffective. Sure, that’s because people don’t listen. But the CDC doesn’t care about that. They’ve proven that giving the advice won’t help, why are you still asking them to give the advice?

I’m not sure this is really the CDC’s reasoning. It seems pretty weird from the point of view of an organization trying to manage a real-world pandemic with people dying if they get it wrong. But I’m having trouble figuring out other possibilities that make sense.

8. So should you wear a mask?

Please don’t buy up masks while there is a shortage and healthcare workers don’t have enough.

If the shortage ends, and wearing a mask is cost-free, I agree with the guidelines from China, Hong Kong, and Japan – consider wearing a mask in high-risk situations like subways or crowded buildings. Wearing masks will not make you invincible, and if you risk compensate even a little it might do more harm than good. Realistically you should be avoiding high-risk situations like subways and crowded buildings as much as you possibly can. But if you have to go in them, yes, most likely a mask will help.

In low-risk situations, like being at home or taking a walk, I mean sure, a mask might make you 0.0001% (or whatever) less likely to get infected. If that’s worth it to you, consider the possibility that you might be freaking out a little too much about this whole pandemic thing. If it’s still worth it, go for it.

You are unlikely to be able to figure out how to use an N95 respirator correctly. I’m not saying it’s impossible, if you try really hard, but assume you’re going to fail unless you have some reason to think otherwise. The most likely outcome is that you have an overpriced surgical mask that might make you incorrectly risk-compensate.

If you are a surgeon performing surgery, bad news. It turns out surgical masks are not very useful for you (1, 2)! You should avoid buying them, since doing so may deplete the number available for people who want to wear them on the subway.

# Legal Systems Very Different From Ours, Because I Just Made Them Up

[with apologies to the real Legal Systems Very Different From Ours. See also the List Of Fictional Drugs Banned By The FDA]

I.

The Clamzorians are animists. They believe every rock and tree and river has its own spirit. And those spirits are legal people. This on its own is not unusual – even New Zealand gives rivers legal personhood. But in Clamzoria, if a flood destroys your home, you sue the river.

If you win, then the river is in debt to you. The government can assign a guardian to the river to force it to pay off its debts, and that guardian gets temporary custody of all the river’s property. He or she can collect a toll from boats, sell water to reservoirs, and charge rent to hydroelectric dams. Once the river has paid off its debt, the guardian is discharged, and the river becomes free to use once again.

Clamzorian precedent governs when you may or may not sue objects. If you swim in the freezing river in the dead of winter, and catch cold, that’s on you. But if a hurricane destroys your property, you can absolutely sue the wind for damages, and collect from windmills. Suits against earthquakes, volcanoes, and the like are dead common. Suits against diseases happen occasionally. Sometimes someone will sue something even more abstract – a custom, an emotion, a concept.

Legend tells of a lawyer who once sued Death itself for wrongful death, a class action suit on behalf of everyone who ever lived. The judge found in favor of the plaintiff, but the appointed custodian despaired at ever collecting the judgment – the few morticians and undertakers in the realm couldn’t afford even a fraction of the damages. In a stroke of genius, he went after the military, and charged them for the right to kill enemy soldiers. The military grumbled, but eventually gave in: fair is fair.

II.

Fixed fines are inherently unfair to the poor. If you fine people $50 for running a red light, you’ve charged someone who makes $10,000 0.5% of their income, but someone who makes $100,000 gets off with only 0.05% of their income.

But prison sentences are inherently unfair to the rich. After all, if you already live in a crowded slum much like a prison cell, and your life is prison-level boring and oppressive already, then going to prison barely costs you anything. But if you live in a mansion and spend all day indulging in the finest luxuries on offer, going to prison is a massive decrease in your quality of life.

The people of Pohjankaupunki thought long and hard about this problem, and came up with a solution: crimes will be punished by neither fines nor prison. They will be punished by government mandated prescription of rimonabant, a prodepressant medication which directly saps your ability to feel happiness. Running a red light may get you 5 mg rimonabant for a month. Murder may get you 80 mg rimonabant twice a day for ten years.

There is no capital punishment in Pohjankaupunki, but if a criminal decides to commit suicide rather than continue to take their medication, they are considered to have voluntarily upgraded to the death penalty, and their debt to the state has been repaid.

III.

Sloviria is an enlightened country. They do not blame criminals for their actions. They realize it is Society’s fault for making criminals that way. So when someone commits a crime, they punish Society.

Sloviria is very technologically advanced, with plenty of social networking sites and GPS tracking of cell phones and all the other systems that create a nice objective social graph. When someone commits a crime, the government lets them go free, and punishes everyone else, in proportion to how close they were to the offender on the social graph. If the punishment for a certain crime is a $1000 fine, perhaps each of their parents and their partner pays $200, their boss and best friend pay $100, some of their teachers a few tenners each, and more distant friends and relations a few dollars or less. If a friend of a friend who you met at a dinner party once commits murder, you may be out a couple of cents.

This isn’t to say perpetrators get off scot-free; Sloviria isn’t that enlightened. The punishment for perpetrators is that nobody wants to interact with them, for fear that they might perpetrate again. Once a person is a known criminal – or a suspected criminal, or just the sort of person who seems like they might become a criminal – their friends, families, and business relations shun them, trying to minimize their potential loss. This threat alone is enough to discourage crime and every form of crime-adjacent misbehavior.

The Slovirian Radical Party is even more enlightened than Sloviria as a whole, and opposes social punishment. They believe that such punishment prevents rehabilitation, since criminals and at-risk youth find it impossible to make the connections they need to succeed, and are forced to hang out with other people as criminal as themselves. They propose a complete inversion of Sloviria’s justice system; when anyone commits a crime, the people closest to them are rewarded. They envision a future where, once somebody shows any sign of being at risk for antisocial behavior, they are love-bombed by dozens of people hoping to get rich off their acquaintance, people who want to employ them, adopt them, date them, or just serve as mentors and parental figures. But wouldn’t all these people encourage the potential criminal to offend? The Radicals debate this among themselves, with one solution being that this could just be a perfectly normal crime punished by jail time.

IV.

Nova-Nishistan’s legal system is based on blackmail. It’s not just blackmail. There are courts and jails and so on. But few people use them. If you have evidence that someone committed a crime, you are expected to threaten to report them unless they give you money.

The system has many advantages. The person most likely to have evidence of a crime is the victim. The victim can choose how much money they want as damages, and have a good chance of receiving it. Fines are automatically calibrated to the wealth of the victim, so poor people are not stuck with debts that are impossible to pay. If a crime is victimless, or the victim chooses not to prosecute, any other witnesses are incentivized to take up the cause of punishing the wrongdoer of their own initiative. Few crimes make it to the courts or prisons, so everyone is assured a speedy trial and an jail cell free of overcrowding.

In order to maintain their system, the Nova-Nishistanis need many laws related to blackmail itself. One of their most serious crimes is to blackmail someone, receive the requested ransom, but report them anyway; anyone convicted of this will be in for a lengthy prison sentence. Indefinite blackmail – “pay me $100 now, but I might ask for more later” – is forbidden. So is non-monetary blackmail; too easy to abuse. There are a host of similar regulations.

One regulation they don’t need is laws about retaliating against blackmailers. You might expect this to be a problem – blackmailing the mob sounds pretty scary. But there are lots of individuals, companies, and (let’s face it) rival gangs happy to provide dead-man’s-switch-as-a-service. Tell them your secret (which they promise not to disclose without your consent), and if anything happens to you, they prosecute it. Even better, if anything happens to you, they’re almost guaranteed to investigate your death, since their special evidence gives them a leg up in what could be a very lucrative blackmail case.

Of course, this only works on people who are rational enough to respond to incentives. If someone is a complete unpredictable psycho, you probably don’t want to try blackmailing them, even with a dead-man’s-switch as insurance. But these are probably the people who should be in jail anyway!

V.

The people of Bogolia thought it was unfair that rich people could hire better lawyers than poor people. But they didn’t want to take the authoritarian step of banning rich people from buying good lawyers, if they thought skilled representation was important. Instead, they just mandated that in any legal case, both sides had to have equally-priced counsel. A rich person could hire as expensive a defense attorney as they wanted, as long as they donated an equal sum to the plaintiff to hire star attorneys of their own. You could sue someone with as highly-priced an attorney as you wanted, but you needed to give them the same amount to spend on their defense.

(this rule applied to the state too, and so implied the right to a public defender worth however much the state was paying to prosecute you, even if you were poor and couldn’t otherwise afford one)

Some trolls tried launching hundreds of frivolous lawsuits against companies they didn’t like, assuming that the company would have to pay both sides of the lawsuit and eventually go broke. They were punished through the normal anti-frivolous-lawsuit rules, and it turned out that companies that did not go broke having to pay one side of a lawsuit don’t go broke having to pay both sides either.

But there were some weirder unintended consequences. How good a lawyer to get became a highly strategic decision for rich clients facing poorer ones. If you thought you were in the right, you’d get a good lawyer, since two equally good lawyers facing off will likely produce truth. If you thought you were in the wrong, you’d try to get a crappy lawyer, since then your opponent would also have a crappy lawyer, and two crappy lawyers facing off will likely produce random results. Not paying for a good lawyer started to be seen as an admission that one’s case was weak.

But also, lawyer salaries started to get wacky. If a random criminal hurt a rich person somehow, and the rich person hired a good lawyer, the random criminal might receive tens of thousands of dollars to spend on legal advice. But random criminals generally are not savvy at evaluating lawyer skill, so thousands of predatory lawyers sprang up, willing to cater to these people by looking impressive and accepting very high salaries. For the savviest of political operators, an equal and opposite caste of underpriced lawyers sprang up, who would accept very low pay in exchange for vague social credit to be doled out later. More and more political scandals started to center on prestigious lawyers defending politicians for free in exchange for favors, and so depriving the opposing party of their right to equally-matched counsel.

Finally the authorities handed down a change to the system: the plaintiff and defendant would agree on two lawyers to conduct the trial. Then the judge would flip a coin, and one of the two would be assigned at random to each party.

VI.

Sanzorre accidentally became an anarcho-capitalist state under the dominion of malpractice insurance companies.

They started off by insuring doctors. Doctors know a bad malpractice case could ruin them. And although being a good doctor helps, it’s not 100%. Even the best doctor can get unlucky, or have somebody with a grudge fabricate a case against them. For that matter, even very bad doctors can get lucky and never have to deal with a case at all. So doctors have malpractice insurance, and if they seem to be practicing medicine badly their insurance company will raise their premiums.

This worked well enough that other industries started adopting it too. If a factory’s pollutant byproducts got discovered to cause cancer ten years later, their industrial malpractice insurance would pay for it. If someone slipped and fell and broke their back on a restaurant floor, their restaurant malpractice insurance would pay for it. Of course, these insurance companies worked closely with factories to monitor how many they were polluting, and gave discounts to restaurants which followed best practices on floor cleaning.

Finally, they branched out to serving ordinary people. If you accidentally hit someone’s dog with your car and got sued for damages, better to have a personal malpractice insurance pay them than get hit for tens of thousands of dollars yourself. Having malpractice insurance became to Sanzorrians what having health insurance is to Americans – a necessity if you don’t want to court disaster.

The plaintiffs in all these cases were usually being covered by lawyers who took contigency fees. But as malpractice insurance companies became better at their jobs, the contingency fees began to dry up. Finally, lobbyists from the insurance companies got contingency fees banned entirely. This presented a dilemma for ordinary people with grievances against bad actors. Thus the rise of the grievance insurance.

If you suffered harm from a doctor’s medical error, and had grievance insurance, the insurance company would pay the cost of the malpractice suit. If you were poisoned by industrial runoff, the insurance company would pay the cost of suing the factory. Grievance insurance soon became as essential as malpractice insurance. Without it, you wouldn’t be able to stand up for your rights.

Like malpractice insurance, grievance insurance was only available cheaply to people who agreed to avoid risks. If you wanted to be able to sue for malpractice, you had to avoid going to quacks. If you wanted to be able to sue factories for pollution, you couldn’t live right next to a coal plant. Gradually, grievance insurances placed more and more restrictions on people’s behavior, and people generally complied.

As malpractice insurances incentivized potential defendants to avoid actions that could harm others, and grievance insurances incentivized individuals to avoid risk, the number of lawsuits gradually got fewer and fewer. Those that happened were generally settled between malpractice insurers and grievance insurers, without ever having to go to court, and sometimes with both companies changing their policy to avoid repeats in the future. Soon, even this formality was eliminated – each malpractice insurance company paid a negotiated amount to each grievance insurance company each year, and the grievance insurance company paid complainants from its own bank account as per its own policies whenever they complained.

It wasn’t quite full anarcho-capitalism. The state still intervened in a few very serious crimes, like murder. But the insurance companies had replaced the civil courts and the regulatory apparatus, and controlled every aspect of doing business.

VII.

Modern philosophy says that formal systems are bunk. The dream of reducing the complexity of reality to some mere set of rules is a childish desire reminiscent of the fascists and high modernists of the early 20th century. Enlightened thinkers realize that we need a Kegan 5 type fluid ability to transcend systematicity. So the people of Mirakoth don’t have laws. They’re just supposed to not do bad stuff.

If someone in Mirakoth thinks someone else did something bad, they can bring it before a council of seven judges. If a majority of the judges think it was bad, they can assign whatever seems to them like fair punishment. If the loser appeals, it goes to a larger council of forty-nine judges. If they think it was bad, it was bad. These judges are under no obligation to follow precedent or any particular philosophy. They’re just supposed to be in favor of good stuff and against bad stuff.

In order to prevent people from seeking out judges who agree with them, each case is assigned seven judges at random. All cases are tried by videoconference, to make sure the judge pool is unlimited by geographical mobility. If the judges think a case is frivolous, they can choose to punish the person who brought the case.

Doesn’t this create such paralyzing uncertainty that nobody knows if they can do anything at all? Not really. Controversial cases are more likely to go to the full 49 judge panel. If an opinion is only held by 20% of judges in the country, then there’s only about a 1 in a million chance that the panel will rule in favor. Even if the opinion is held by 40%, it’s still only an 8% chance of winning. So just don’t do things that more than 40% of people think are bad, and you’ll be fine!

# SSC Journal Club: MacIntyre On Cloth Masks

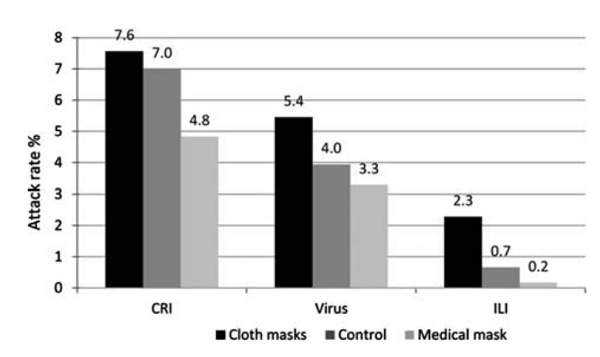
[Content warning: this is a complicated analysis of something people care about a lot right now. I’m not confident in my analysis, the post comes to no clear conclusion and there are no easy answers about how to proceed. If I see this on Twitter with some headline about it DESTROYING somebody, I am going to be so mad.]

The New York Times says that It’s Time To Make Your Own Face Mask. But MacIntyre et al (2015) says it isn’t.

The surgical masks used in hospitals are made out of non-woven fabrics that are pretty different from anything you have at home. But in some developing countries, health care workers instead use masks made of normal cloth. Laboratory tests find that improvised cloth masks block 60 – 80% of virus particles. Respirators and real surgical masks block 95%+, but 60-80% still seems better than nothing. And most of the masks ordinary people wear in Asian countries are cloth, and they seem to do pretty well. So there’s some circumstantial evidence that these cloth masks might be helpful. Most experts in the early 2000s agreed that these masks were probably better than nothing. In 2015, an Australian team set out to prove it with a randomized controlled trial.

They went to a hospital in Vietnam and randomized workers there to a normal mask group, a cloth mask group, or a control group. Because it would have been unethical to tell the control group not to wear masks, they left the control group alone. Most control group workers did end up wearing masks sometimes, but less than the experimental groups did.

After a month, they counted how many infections each group had, for three different categories of infection. Here are the results:



Technically significant only in the ILI category, but later the authors do various post hoc adjustment for confounders and find it’s significant everywhere

For all three categories, people wearing the real surgical masks were the healthiest, the control group was in the middle, and people wearing the cloth masks were the sickest.

This shows real surgical masks work better than cloth masks. It’s a little bit unclear about how well cloth masks work. They do worse than the control group, but you could tell two stories about that. In one, cloth masks are worse than no mask at all. In the other, cloth masks have zero-to-slight-positive efficacy, but because some people in the control group were wearing real surgical masks some of the time, they did better than the cloth group overall. So it depends a lot what the control group was doing.

Unfortunately, the paper doesn’t give us all the data we want. It tells us that about 57% of both the surgical mask group and the cloth mask group wore masks regularly (defined as more than 70% of the time) but only 24% of the control group did. But there is no way of knowing whether the rest of the control group wore masks 69% of the time or 0% of the time.

A separate paragraph tells us that 37% of the control group used surgical masks, 8% cloth masks, and 53% used a combination of both. These numbers don’t make a lot of sense in the context of the last paragraph, so I’m going to assume they meant that on the infrequent occasions they did wear masks, those were the masks they used. But we don’t know if the compliant workers were disproportionately using cloth masks, disproportionately using medical masks, or both evenly. It’s hard to just eyeball these numbers and get a good sense for whether cloth masks really are worse than nothing.

But the authors themselves lean towards the hypothesis that that cloth masks are actively bad. First, because after some calculations I cannot quite follow, they find that the difference between surgical masks and cloth masks is so high that either the surgical masks are absurdly good, or the difference is being augmented by the cloth masks being actively bad. But nobody has previously found surgical masks to be absurdly good. The authors cite two previous studies of theirs which did include a no-mask control group; surgical masks did not significantly outperform nothing (they did show a trend towards doing so, and the studies were probably underpowered).

Second, because they compare the numbers from this study to numbers from those other two studies directly. They find the rate of infection in surgical mask users is not-significantly-different throughout the three studies, and the rate of infection in surgical mask users and no-mask controls was also not-significantly-different, and therefore surgical masks are the same as nothing and so probably the cloth masks are actively bad.

I am very unimpressed by this. First, you are really not supposed to compare things across multiple different studies. The authors protest that they did all three studies along pretty similar designs, but also admit they were different hospitals during different seasons. But second, almost no differences anywhere are significant, because all of these studies were at least a little underpowered. The current study found no significant difference between cloth masks and surgical masks in two of the three categories, even though the trend was in the expected direction. The other studies found no difference between wearing a medical mask and not wearing a medical mask, even though previous studies have suggested medical masks should work. They couldn’t even find any difference between wearing an N95 respirator and not wearing any protection at all. So when you need a chain of “x is not significantly different from y, which is not significantly different from z” in a bunch of studies that wouldn’t have been able to notice significant differences even if they existed, I stop believing it pretty quickly.

(In fact, I think you could use the same logic to draw the exact opposite conclusion. The cloth mask group in the current study didn’t have a significant difference from the surgical mask group in the other study, and the surgical mask group was no different from placebo, therefore cloth masks cannot have a negative effect. I find it hard to believe the authors missed this, so let me know if I am confused here.)

But MacIntyre et al take it seriously, and conclude:

The study suggests medical masks may be protective, but the magnitude of difference raises the possibility that cloth masks cause an increase in infection risk in HCWs. Further, the filtration of the medical mask used in this trial was poor, making extremely high efficacy of medical masks unlikely, particularly given the predominant pathogen was rhinovirus, which spreads by the airborne route. Given the obligations to HCW occupational health and safety, it is important to consider the potential risk of using cloth masks […] The physical properties of a cloth mask, reuse, the frequency and effectiveness of cleaning, and increased moisture retention, may potentially increase the infection risk for health care workers. The virus may survive on the surface of the face-masks, and modelling studies have quantified the contamination levels of masks. Self-contamination through repeated use and improper doffing is possible. For example, a contaminated cloth mask may transfer pathogen from the mask to the bare hands of the wearer. We also showed that filtration was extremely poor (almost 0%) for the cloth masks. Observations during SARS suggested double-masking and other practices increased the risk of infection because of moisture, liquid diffusion and pathogen retention. These effects may be associated with cloth masks.

Why am I focusing on this one weird study so much? Because it’s the only RCT of cloth face masks we have! Millions of people, egged on by top newspapers, are about to start wearing cloth face masks during a pandemic, when right now the authors of the only randomized trial on them conclude they’re probably net harmful. This should be really scary! Somebody with more experience and statistical knowledge than I have should be looking this over with a fine-toothed comb and trying to figure out what we should do.

Until then, should people stay away from cloth masks? I’m not sure, and this is so not a recommendation, but I lean toward no. The prior that they should work or at least be neutral is too high for a study this weak to convince me otherwise. More important, this study only examines incoming pathogens. Even if they are harmful for blocking incoming pathogens, there are still reasons to think they are helpful for blocking outgoing ones. If I had to hang out with a coronavirus patient for a while, and I had to choose between both of us wearing cloth masks, or neither, I would go with the masks. Only until we could get real surgical masks, which are much better. But I’d go with the cloth ones instead of nothing.

But right now that’s a gut judgment, and the evidence says I’m wrong. This is one of those times where people have to make a life-or-death decision in conditions of high uncertainty, and it really sucks.

[EDIT: Bolded a passage I think is important to make sure people don’t miss it]

# Book Review: The Precipice

I.

It is a well known fact that the gods hate prophets.

False prophets they punish only with ridicule. It’s the true prophets who have to watch out. The gods find some way to make their words come true in the most ironic way possible, the one where knowing the future just makes things worse. The Oracle of Delphi told Croesus he would destroy a great empire, but when he rode out to battle, the empire he destroyed was his own. Zechariah predicted the Israelites would rebel against God; they did so by killing His prophet Zechariah. Jocasta heard a prediction that she would marry her infant son Oedipus, so she left him to die on a mountainside – ensuring neither of them recognized each other when he came of age.

Unfortunately for him, Oxford philosopher Toby Ord is a true prophet. He spent years writing his magnum opus The Precipice, warning that humankind was unprepared for various global disasters like pandemics and economic collapses. You can guess what happened next. His book came out March 3, 2020, in the middle of a global pandemic and economic collapse. He couldn’t go on tour to promote it, on account of the pandemic. Nobody was buying books anyway, on account of the economic collapse. All the newspapers and journals and so on that would usually cover an exciting new book were busy covering the pandemic and economic collapse instead. The score is still gods one zillion, prophets zero. So Ord’s PR person asked me to help spread the word, and here we are.

Imagine you were sent back in time to inhabit the body of Adam, primordial ancestor of mankind. It turns out the Garden of Eden has motorcycles, and Eve challenges you to a race. You know motorcycles can be dangerous, but you’re an adrenaline junkie, naturally unafraid of death. And it would help take your mind off that ever-so-tempting Tree of Knowledge. Do you go?

Before you do, consider that you’re not just risking your own life. A fatal injury to either of you would snuff out the entire future of humanity. Every song ever composed, every picture ever painted, every book ever written by all the greatest authors of the millennia would die stillborn. Millions of people would never meet their true loves, or get to raise their children. All of the triumphs and tragedies of humanity, from the conquests of Alexander to the moon landings, would come to nothing if you hit a rock and cracked your skull.

So maybe you shouldn’t motorcycle race. Maybe you shouldn’t even go outside. Maybe you and Eve should hide, panicked, in the safest-looking cave you can find.

Ord argues that 21st century humanity is in much the same situation as Adam. The potential future of the human race is vast. We have another five billion years until the sun goes out, and 10^100 until the universe becomes uninhabitable. Even with conservative assumptions, the galaxy could support quintillions of humans. Between Eden and today, the population would have multiplied five billion times; between today and our galactic future, it could easily multiply another five billion. However awed Adam and Eve would have been when they considered the sheer size of the future that depended on them, we should be equally awed.

So maybe we should do the equivalent of not motorcycling. And that would mean taking existential risks (“x-risks”) – disasters that might completely destroy humanity or permanently ruin its potential – very seriously. Even more seriously than we would take them just based on the fact that we don’t want to die. Maybe we should consider all counterbalancing considerations – “sure, global warming might be bad, but we also need to keep the economy strong!” – to be overwhelmed by the crushing weight of the future.

This is my metaphor, not Ord’s. He uses a different one – the Cuban Missile Crisis. We all remember the Cuban Missile Crisis as a week where humanity teetered on the precipice of destruction, then recovered into a more stable not-immediately-going-to-destroy-itself state. Ord speculates that far-future historians will remember the entire 1900s and 2000s as a sort of centuries-long Cuban Missile Crisis, a crunch time when the world was unusually vulnerable and everyone had to take exactly the right actions to make it through. Or as the namesake precipice, a place where the road to the Glorious Future crosses a narrow rock ledge hanging over a deep abyss.

Ord has other metaphors too, and other arguments. The first sixty pages of Precipice are a series of arguments and thought experiments intended to drive home the idea that everyone dying would be really bad. Some of them were new to me and quite interesting – for example, an argument that we should keep the Earth safe for future generations as a way of “paying it forward” to our ancestors, who kept it safe for us. At times, all these arguments against allowing the destruction of the human race felt kind of excessive – isn’t there widespread agreement on this point? Even when there is disagreement, Ord doesn’t address it here, banishing counterarguments to various appendices – one arguing against time discounting the value of the future, another arguing against ethical theories that deem future lives irrelevant. This part of the book isn’t trying to get into the intellectual weeds. It’s just saying, again and again, that it would be really bad if we all died.

It’s tempting to psychoanalyze Ord here, with help from passages like this one:

I have not always been focused on protecting our longterm future, coming to the topic only reluctantly. I am a philosopher at Oxford University, specialising in ethics. My earlier work was rooted in the more tangible concerns of global health and global poverty – in how we could best help the worst off. When coming to grips with these issues I felt the need to take my work in ethics beyond the ivory tower. I began advising the World Health Organization, World Bank, and UK government on the ethics of global health. And finding that my own money could do hundreds of times as much good for those in poverty as it could do for me, I made a lifelong pledge to donate at least a tenth of all I earn to help them. I founded a society, Giving What We Can, for those who wanted to join me, and was heartened to see thousands of people come together to pledge more than one billion pounds over our lifetimes to the most effective charities we know of, working on the most important causes. Together, we’ve already been able to transform the lives of thousands of people. And because there are many other ways beyond our donations in which we can help fashion a better world, I helped start a wider movement, known as “effective altruism”, in which people aspire to use prudence and reason to do as much good as possible.

We’re in the Garden of Eden, so we should stop worrying about motorcycling and start worrying about protecting our future. But Ord’s equivalent of “motorcycling” was advising governments and NGOs on how best to fight global poverty. I’m familiar with his past work in this area, and he was amazing at it. He stopped because he decided that protecting the long-term future was more important. What must he think of the rest of us, who aren’t stopping our ordinary, non-saving-thousands-of-people-from-poverty day jobs?

In writing about Ord’s colleagues in the effective altruist movement, I quoted Larissa MacFarquahar on Derek Parfit:

When I was interviewing him for the first time, for instance, we were in the middle of a conversation and suddenly he burst into tears. It was completely unexpected, because we were not talking about anything emotional or personal, as I would define those things. I was quite startled, and as he cried I sat there rewinding our conversation in my head, trying to figure out what had upset him. Later, I asked him about it. It turned out that what had made him cry was the idea of suffering. We had been talking about suffering in the abstract. I found that very striking.

Toby Ord was Derek Parfit’s grad student, and I get sort of the same vibe from him – someone whose reason and emotions are unusually closely aligned. Stalin’s maxim that “one death is a tragedy, a million deaths is a statistic” accurately describes how most of us think. I am not sure it describes Toby Ord. I can’t say confidently that Toby Ord feels exactly a million times more intense emotions when he considers a million deaths than when he considers one death, but the scaling factor is definitely up there. When he considers ten billion deaths, or the deaths of the trillions of people who might inhabit our galactic future, he – well, he’s reduced to writing sixty pages of arguments and metaphors trying to cram into our heads exactly how bad this would be.

II.

The second part of the book is an analysis of specific risks, how concerned we should be about each, and what we can do to prevent them. Ord stays focused on existential risks here. He is not very interested in an asteroid that will wipe out half of earth’s population; the other half of humanity will survive to realize our potential. He’s not completely uninterested – wiping out half of earth’s population could cause some chaos that makes it harder to prepare for other catastrophes. But his main focus is on things that would kill everybody – or at least leave too few survivors to effectively repopulate the planet.

I expected Ord to be alarmist here. He is writing a book about existential risks, whose thesis is that we should take them extremely seriously. Any other human being alive would use this as an opportunity to play up how dangerous these risks are. Ord is too virtuous. Again and again, he knocks down bad arguments for worrying too much, points out that killing every single human being on earth, including the ones in Antarctic research stations, is actually quite hard, and ends up convincing me to downgrade my risk estimate.

So for example, we can rule out a high risk of destruction by any natural disaster – asteroid, supervolcano, etc – simply because these things haven’t happened before in our species’ 100,000 year-odd history. Dino-killer sized asteroids seem to strike the Earth about once every few hundred million years, bounding the risk per century around the one-in-a-million level. But also, scientists are tracking almost all the large asteroids in the solar system, and when you account for their trajectories, the chance that one slips through and hits us in the next century goes down to less than one in a hundred fifty million. Large supervolcanoes seem to go off about once every 80,000 years, so the risk per century is 1/800. There are similar arguments around nearby supernovae, gamma ray bursts, and a bunch of other things.

I usually give any statistics I read a large penalty for “or maybe you’re a moron”. For example, lots of smart people said in 2016 that the chance of Trump winning was only 1%, or 0.1%, or 0.00001%, or whatever. But also, they were morons. They were using models, and their models were egregiously wrong. If you hear a person say that their model’s estimate of something is 0.00001%, very likely your estimate of the thing should be much higher than that, because maybe they’re a moron. I explain this in more detail here.

Ord is one of a tiny handful of people who doesn’t need this penalty. He explains this entire dynamic to his readers, agrees it is important, and adjusts several of his models appropriately. He is always careful to add a term for unknown unknowns – sometimes he is able to use clever methods to bound this term, other times he just takes his best guess. And he tries to use empirically-based methods that don’t have this problem, list his assumptions explicitly, and justify each assumption, so that you rarely have to rely on arguments shakier than “asteroids will continue to hit our planet at the same rate they did in the past”. I am really impressed with the care he puts into every argument in the book, and happy to accept his statistics at face value. People with no interest in x-risk may enjoy reading this book purely as an example of statistical reasoning done with beautiful lucidity.

When you accept very low numbers at face value, it can have strange implications. For example, should we study how to deflect asteroids? Ord isn’t sure. The base rate of asteroid strikes is so low that it’s outweighed by almost any change in the base rate. If we successfully learn how to deflect asteroids, that not only lets good guys deflect asteroids away from Earth, but also lets bad guys deflect asteroids towards Earth. The chance that an dino-killer asteroid approaches Earth and needs to be deflected away is 1/150 million per century, with small error bars. The chance that malicious actors deflect an asteroid towards Earth is much harder to figure out, but it has wide error bars, and there are a lot of numbers higher than 1/150 million. So probably most of our worry about asteroids over the next century should involve somebody using one as a weapon, and studying asteroid deflection probably makes that worse and not better.

Ord uses similar arguments again and again. Humanity has survived 100,000 years, so its chance of death by natural disaster per century is probably less than 1 / 1,000 (for complicated statistical reasons, he puts it at between 1/10,000 and 1/100,000). But humanity has only had technology (eg nuclear weapons, genetically engineered bioweapons) for a few decades, so there are no such guarantees of its safety. Ord thinks the overwhelming majority of existential risk comes from this source, and singles out four particular technological risks as most concerning.

First, nuclear war. This was one of the places where Ord’s work is cause for optimism. You’ve probably heard that there are enough nuclear weapons to “destroy the world ten times over” or something like that. There aren’t. There are enough nuclear weapons to destroy lots of majors city, kill the majority of people, and cause a very bad nuclear winter for the rest. But there aren’t enough to literally kill every single human being. And because of the way the Earth’s climate works, the negative effects of nuclear winter would probably be concentrated in temperate and inland regions. Tropical islands and a few other distant locales (Ord suggests Australia and New Zealand) would have a good chance of making it through even a large nuclear apocalypse with enough survivors to repopulate the Earth. A lot of things would have to go wrong at once, and a lot of models be flawed in ways they don’t seem to be flawed, for a nuclear war to literally kill everyone. Ord gives the per-century risk of extinction from this cause at 1 in 1,000.

Second, global warming. The current scientific consensus is that global warming will be really bad but not literally kill every single human. Even for implausibly high amounts of global warming, survivors can always flee to a pleasantly balmy Greenland. The main concern from an x-risk point of view is “runaway global warming” based on strong feedback loops. For example, global warming causes permafrost to melt, which releases previously trapped carbon, causing more global warming, causing more permafrost to melt, etc. Or global warming causes the oceans to warm, which makes them release more methane, which causes more global warming, causing the oceans to get warmer, etc. In theory, this could get really bad – something similar seems to have happened on Venus, which now has an average temperature of 900 degrees Fahrenheit. But Ord thinks it probably won’t happen here. His worst-case scenario estimates 13 – 20 degrees C of warming by 2300. This is really bad – summer temperatures in San Francisco would occasionally pass 140F – but still well short of Venus, and compatible with the move-to-Greenland scenario. Also, global temperatures jumped 5 degree C (to 14 degrees above current levels) fifty million years ago, and this didn’t seem to cause Venus-style runaway warming. This isn’t a perfect analogy for the current situation, since the current temperature increase is happening faster than the ancient one did, but it’s still a reason for hope. This is another one that could easily be an apocalyptic tragedy unparalleled in human history but probably not an existential risk; Ord estimates the x-risk per century as 1/1,000.

The same is true for other environmental disasters, of which Ord discusses a long list. Overpopulation? No, fertility rates have crashed and the population is barely expanding anymore (also, it’s hard for overpopulation to cause human extinction). Resource depletion? New discovery seems to be faster than depletion for now, and society could route around most plausible resources shortages. Honeybee collapse? Despite what you’ve heard, losing all pollinators would only cause a 3 – 8% decrease in global crop production. He gives all of these combined plus environmental unknown unknowns an additional 1/1,000, just in case.

Third, pandemics. Even though pathogens are natural, Ord classifies pandemics as technological disasters for two reasons. First, natural pandemics are probably getting worse because our technology is making cities denser, linking countries closer together, and bringing humans into more contact with the animal vectors of zoonotic disease (in one of the book’s more prophetic passages, Ord mentions the risk of a disease crossing from bats to humans). But second, bioengineered pandemics may be especially bad. These could be either accidental (surprisingly many biologists alter diseases to make them worse as part of apparently legitimate scientific research) or deliberate (bioweapons). There are enough unknown unknowns here that Ord is uncomfortable assigning relatively precise and low risk levels like he did in earlier categories, and this section generally feels kind of rushed, but he estimates the per-century x-risk from natural pandemics as 1/10,000 and from engineered pandemics as 1/30.

The fourth major technological risk is AI. You’ve all read about this one by now, so I won’t go into the details, but it fits the profile of a genuinely dangerous risk. It’s related to technological advance, so our long and illustrious history of not dying from it thus far offers no consolation. And because it could be actively trying to eliminate humanity, isolated populations on islands or in Antarctica or wherever offer less consolation than usual. Using the same arguments and sources we’ve seen every other time this topic gets brought up, Ord assigns this a 1/10 risk per century, the highest of any of the scenarios he examines, writing:

In my view, the greatest risk to humanity’s potential in the next hundred years comes from unaligned artificial intelligence, which I put at 1 in 10. One might be surprised to see such a high number for such a speculative risk, so it warrants some explanation.

A common approach to estimating the chance of an unprecedented event with earth-shaking consequences is to take a sceptical stance: to start with an extremely small probability and only raise it from there when a large amount of hard evidence is presented. But I disagree. Instead, I think that the right method is to start with a probability that reflects our overall impressions, then adjust this in light of the scientific evidence. When there is a lot of evidence, these approaches converge. But when there isn’t, the starting point can matter.

In the case of artificial intelligence, everyone agrees the evidence and arguments are far from watertight, but the question is where does this leave us? Very roughly, my approach is to start with the overall view of the expert community that there is something like a 1 in 2 chance that AI agents capable of outperforming humans in almost every task will be developed in the coming century. And conditional on that happening, we shouldn’t be shocked if these agents that outperform us across the board were to inherit our future.

The book also addresses a few more complicated situations. There are ways humankind could fail to realize its potential even without being destroyed. For example, if it got trapped in some kind of dystopia that it was impossible to escape. Or if it lost so many of its values that we no longer recognized it as human. Ord doesn’t have too much to say about these situations besides acknowledging that they would be bad and need further research. Or a series of disasters could each precipitate one another, or a minor disaster could leave people unprepared for a major disaster, or something along those lines.

Here, too, Ord is more optimistic than some other sources I have read. For example, some people say that if civilization ever collapses, it will never be able to rebuild, because we’ve already used up all easily-accessible sources of eg fossil fuels, and an infant civilization won’t be able to skip straight from waterwheels to nuclear. Ord is more sanguine:

Even if civilization did collapse, it is likely that it could be re-established. As we have seen, civilization has already been independently established at least seven times by isolated peoples. While one might think resources depletion could make this harder, it is more likely that it has become substantially easier. Most dissasters short of human extinction would leave our domesticated animals and plants, as well as copious material resources in the ruins of our cities – it is much easier to re-forge iron from old railings than to smelt it from ore. Even expendable resources such as coal would be much easier to access, via abandoned reserves and mines, than they ever were in the eighteenth century. Moreover, evidence that civilisation is possible, and the tools and knowledge to help rebuild, would be scattered across the world.

III.

Still, these risks are real, and humanity will under-respond to them for predictable reasons.

First, free-rider problems. If some people invest resources into fighting these risks and others don’t, both sets of people will benefit equally. So all else being equal everyone would prefer that someone else do it. We’ve already seen this play out with international treaties on climate change.

Second, scope insensitivity. A million deaths, a billion deaths, and complete destruction of humanity all sound like such unimaginable catastrophes that they’re hardly worth differentiating. But plausibly we should put 1000x more resources into preventing a billion deaths than a million, and some further very large scaling factor into preventing human extinction. People probably won’t think that way, which will further degrade our existential risk readiness.

Third, availability bias. Existential risks have never happened before. Even their weaker non-omnicidal counterparts have mostly faded into legend – the Black Death, the Tunguska Event. The current pandemic is a perfect example. Big pandemics happen once every few decades – the Spanish flu of 1918 and the Hong Kong Flu of 1968 are the most salient recent examples. Most countries put some effort into preparing for the next one. But the preparation wasn’t very impressive. After this year, I bet we’ll put impressive effort into preparing for respiratory pandemics the next decade or two, while continuing to ignore other risks like solar flares or megadroughts that are equally predictable. People feel weird putting a lot of energy into preparing for something that has never happened before, and their value of “never” is usually “in a generation or two”. Getting them to care about things that have literally never happened before, like climate change, nuclear winter, or AI risk, is an even taller order.

And even when people seem to care about distant risks, it can feel like a half-hearted effort. During a Berkeley meeting of the Manhattan Project, Edward Teller brought up the basic idea behind the hydrogen bomb. You would use a nuclear bomb to ignite a self-sustaining fusion reaction in some other substance, which would produce a bigger explosion than the nuke itself. The scientists got to work figuring out what substances could support such reactions, and found that they couldn’t rule out nitrogen-14. The air is 79% nitrogen-14. If a nuclear bomb produced nitrogen-14 fusion, it would ignite the atmosphere and turn the Earth into a miniature sun, killing everyone. They hurriedly convened a task force to work on the problem, and it reported back that neither nitrogen-14 nor a second candidate isotope, lithium-7, could support a self-sustaining fusion reaction.

They seem to have been moderately confident in these calculations. But there was enough uncertainty that, when the Trinity test produced a brighter fireball than expected, Manhattan Project administrator James Conant was “overcome with dread”, believing that atmospheric ignition had happened after all and the Earth had only seconds left. And later, the US detonated a bomb whose fuel was contaminated with lithium-7, the explosion was much bigger than expected, and some bystanders were killed. It turned out atomic bombs could initiate lithium-7 fusion after all! As Ord puts it, “of the two major thermonuclear calculations made that summer at Berkeley, they got one right and one wrong”. This doesn’t really seem like the kind of crazy anecdote you could tell in a civilization that was taking existential risk seriously enough.

So what should we do? That depends who you mean by “we”.

Ordinary people should follow the standard advice of effective altruism. If they feel like their talents are suited for a career in this area, they should check out 80,000 Hours and similar resources and try to pursue it. Relevant careers include science (developing helpful technologies to eg capture carbon or understand AI), politics and policy (helping push countries to take risk-minimizing actions), and general thinkers and influencers (philosophers to remind us of our ethical duties, journalists to help keep important issues fresh in people’s minds). But also, anything else that generally strengthens and stabilizes the world. Diplomats who help bring countries closer together, since international peace reduces the risk of nuclear war and bioweapons and makes cooperation against other threats more likely. Economists who help keep the market stable, since a prosperous country is more likely to have resources to devote to the future. Even teachers are helping train the next generation of people who can help in the effort, although Ord warns against going too meta – most people willing to help with this will still be best off working on causes that affect existential risk directly. If they don’t feel like their talents lie in any of these areas, they can keep earning money at ordinary jobs and donate some of it (traditionally 10%) to x-risk related charities.

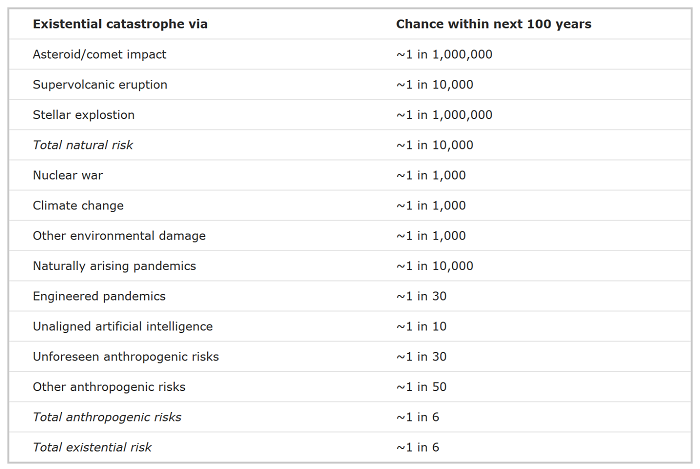
Rich people, charitable foundations, and governments should fund anti-x-risk work more than they’re already doing. Did you know that the Biological Weapons Convention, a key international agreement banning biological warfare, has a budget lower than the average McDonald’s restaurant (not total McDonald corporate profits, a single restaurant)? Or that total world spending on preventing x-risk is less than total world spending on ice cream? Ord suggests a target of between 0.1% and 1% of gross world product for anti-x-risk efforts.

And finally, Ord has a laundry list of requests for sympathetic policy-makers (Appendix F). Most of them are to put more research and funding into things, but the actionable specific ones are: restart various nuclear disarmament treaties, take ICBMs off “hair-trigger alert”, have the US rejoin the Paris Agreement on climate change, fund the Biological Weapons Convention better, and mandate that DNA synthesis companies screen consumer requests for dangerous sequences so that terrorists can’t order a batch of smallpox virus (80% of companies currently do this screening, but 20% don’t). The actual appendix is six pages long, there are a lot of requests to put more research and funding into things.

In the last section, Ord explains that all of this is just the first step. After we’ve conquered existential risk (and all our other problems), we’ll have another task: to contemplate how we want to guide the future. Before we spread out into the galaxy, we might want to take a few centuries to sit back and think about what our obligations are to each other, the universe, and the trillions of people who may one day exist. We cannot take infinite time for this; the universe is expanding, and for each year we spend not doing interstellar colonization, three galaxies cross the cosmological event horizon and become forever unreachable, and all the potential human civilizations that might have flourished there come to nothing. Ord expects us to be concerned about this, and tries to reassure us that it will be okay (the relative loss each year is only one five-billionth of the universe). So he thinks taking a few centuries to reflect before beginning our interstellar colonization is worthwhile on net. But for now, he thinks this process should take a back seat to safeguarding the world from x-risk. Deal with the Cuban Missile Crisis we’re perpetually in the middle of, and then we’ll have time for normal philosophy.

IV.

In the spirit of highly-uncertain-estimates being better than no estimates at all, Ord offers this as a draft of where the existential risk community is right now (“they are not in any way the final word, but are a concise summary of all I know about the risk landscape”):



Again, the most interesting thing for me is how low most of the numbers are. It’s a strange sight in a book whose thesis could be summarized as “we need to care more about existential risk”. I think most people paying attention will be delighted to learn there’s a 5 in 6 chance the human race will survive until 2120.

This is where I turn to my psychoanalysis of Toby Ord again. I think he, God help him, sees a number like that and responds appropriately. He multiplies 1/6th by 10 billion deaths and gets 1.6 billion deaths. Then he multiplies 1/6th by the hundreds of trillions of people it will prevent from ever existing, and gets tens of trillions of people. Then he considers that the centuries just keep adding up, until by 2X00 the risk is arbitrarily high. At that point, the difference between a 1/6 chance of humanity dying per century vs. a 5/6 chance of humanity dying may have psychological impact. But the overall takeaway from both is “Holy @!#$, we better put a lot of work into dealing with this.”

There’s an old joke about a science lecture. The professor says that the sun will explode in five billion years, and sees a student visibly freaking out. She asks the student what’s so scary about the sun exploding in five billion years. The student sighs with relief: “Oh, thank God! I thought you’d said five million years!”

We can imagine the opposite joke. A professor says the sun will explode in five minutes, sees a student visibly freaking out, and repeats her claim. The student, visibly relieved: “Oh, thank God! I thought you’d said five seconds.”

When read carefully, The Precipice is the book-length version of the second joke. Things may not be quite as disastrous as you expected. But relief may not quite be the appropriate emotion, and there’s still a lot of work to be done.

# SSCJC: Real World Depression Measurement

The largest non-pharma antidepressant trial ever conducted just confirmed what we already knew: scientists love naming things after pandas.

We already had PANDAS (Pediatric Autoimmune Neuropsychiatric Disorders Associated with Streptococcus) and PANDA (Proton ANnhilator At DArmstadt). But the latest in this pandemic of panda pandering is the PANDA (Prescribing ANtiDepressants Appropriately) Study. A group of British scientists followed 655 complicated patients who received either placebo or the antidepressant sertraline (Zoloft®).

The PANDA trial was unique in two ways. First, as mentioned, it was the largest ever trial for a single antidepressant not funded by a pharmaceutical company. Second, it was designed to mimic “the real world” as closely as possible. In most antidepressant trials, researchers wait to gather the perfect patients: people who definitely have depression and definitely don’t have anything else. Then they get top psychiatrists to carefully evaluate each patient, monitor the way they take the medication, and exhaustively test every aspect of their progress with complicated questionnaires. PANDA looked for normal people going to their GP’s (US English: PCP’s) office, with all of the mishmash of problems and comorbidities that implies.

Measuring real-world efficacy is especially important for antidepressant research because past studies have failed to match up with common sense. Most studies show antidepressants having “clinically insignificant” effects on depression; that is, although scientists can find a statistical difference between treatment and placebo groups, it seems too small to matter. But in the real world, most doctors find antidepressants very useful, and many patients credit them for impressive recoveries. Maybe a big real-world study would help bridge the gap between study vs. real-world results.

The study used an interesting selection criteria – you were allowed in if you and your doctor reported “uncertainty…about the possible benefit of an antidepressant”. That is, people who definitely didn’t need antidepressants were sent home without an antidepressant, people who definitely did need antidepressants got the antidepressant, and people on the borderline made it into the study. This is very different from the usual pharma company method of using the people who desperately need antidepressants the most in order to inflate success rates. And it’s more relevant to clinical practice – part of what it means for studies to guide our clinical practice is to tell us what to do in cases where we’re otherwise not sure. And unlike most studies, which use strict diagnostic criteria, this study just used a perception of needing help – not even necessarily for depression, some of these patients were anxious or had other issues. Again, more relevant for clinical practice, where the borders between depression, anxiety, etc aren’t always that clear.

They ended up with 655 people, ages 18-74, from Bristol, Liverpool, London, and York. They followed up on how they were doing at 2, 6, and 12 weeks after they started medication. As usual, they scored patients on a bunch of different psychiatric tests.

In the end, PANDA confirmed what we already know: it is really hard to measure antidepressant outcomes, and all the endpoints conflict with each other.

I am going to be much nicer to you than the authors of the original paper were to their readers, and give you a convenient table with all of the results converted to effect sizes. All values are positive, meaning the antidepressant group beat the placebo group. I calculated some of this by hand, so it may be wrong.

PHQ-9 is a common depression test. BDI is another common depression test. GAD-7 is an anxiety test. SF-12 is a vague test of how mentally healthy you’re feeling. Remission indicates percent of patients whose test scores have improved enough that they qualify as “no longer depressed”. General improvement was just asking patients if they felt any better.

I like this study because it examines some of the mystery of why antidepressants do much worse in clinical trials than according to anecdotal doctor and patient intuitions. One possibility has always been that we’re measuring these things wrong. This study goes to exactly the kind of naturalistic setting where people report good results, and measures things a bunch of different ways to see what happens.

The results are broadly consistent with previous studies. Usually people think of effect sizes less than 0.2 as miniscule, less than 0.5 as small, and less than 0.8 as medium. This study showed only small to low-medium effect sizes for everything.

I haven’t checked whether differences between effect sizes were significant. But just eyeballing them, this study doesn’t agree with my hypothesis that SSRIs are better for anxiety than for depression; the GAD-7 effect size is about the same as the PHQ and BDI effect sizes.

It does weakly support a hypothesis where SSRIs are better for patient-rated improvement than for researcher-measured tests. The highest effect size was in “self-rated improvement”, where the researchers just asked the patients if they felt better. This effect size (0.49) was still small. But if we let ourselves round it up, it reaches all the way to “medium”. Progress!

What does this mean in real life? 59% of patients in the antidepressant group, compared to 42% of patients in the placebo group, said they felt better. I’m actually okay with this. It means that for every 58 patients who wouldn’t have gotten better on placebo, 17 of them would get better on an antidepressant – in other words, the antidepressant successfully converted 30% of people from nonresponder to responder. This obviously isn’t as good as 50% or 100%. But it doesn’t strike me as consistent with the claims of “clinically insignificant” and “why would anyone ever use these medications”?

(though of course, this is just one study, and it’s a study where I took the most promising of many different endpoints, so it’s not exactly cause for celebration)

If antidepressants do better on patient report than on our depression tests, does that mean our depression tests are bad? Maybe. Figure 4 from Hieronymous et al helps clarify a bit of what’s going on:



At least in less severely depressed patients, antidepressants are more likely to produce significant gains on vaguer or more fundamental symptoms (like “depressed mood” or “anxiety”) than on specific symptoms (like insomnia or psychomotor disruptions). Probably patients care a lot less about “psychomotor disruptions” than researchers studying depression do, and they just want to feel happy again. This study’s finding of an 0.4 – 0.5 effect size on patient response closely matches Hieronymous et al’s finding of an 0.4 – 0.5 effect size on depressed mood.

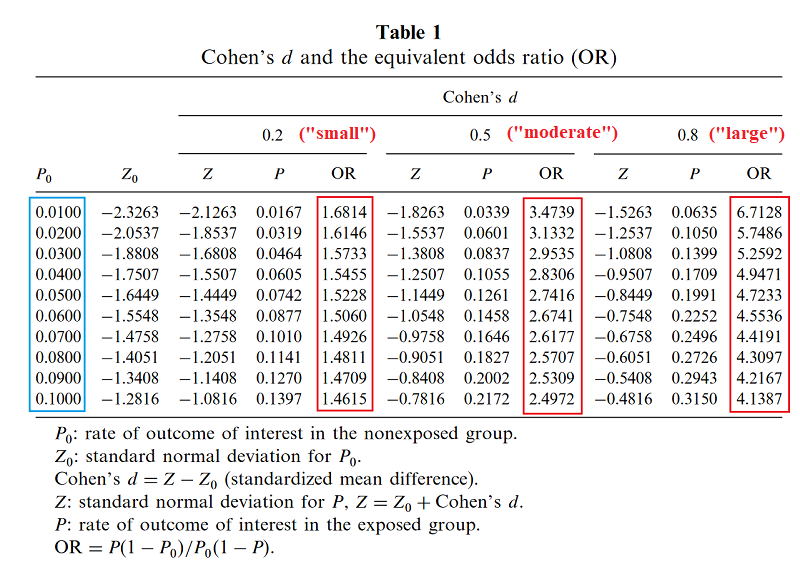
Like most studies, PANDA used a one-size-fits-all solution based on a single antidepressant. This is a reasonable choice for a study, but doesn’t match clinical practice, where we usually try one antidepressant, see if it works, and try another if it doesn’t. In patients like the ones in the study, who had failed treatment with sertraline, a usual next step would be to try bupropion. An even better idea would be to screen patients for more typical vs. atypical depression, start people on sertraline or bupropion based on the symptom profile, and then switch to the other if the first one didn’t work. The STAR\*D trial did something like this, and got better results than an SSRI alone. I haven’t done the work I would need to compare this to STAR\*D, but it seems possible that the extra push from targeted treatment could bring our 0.49 effect size up to the 0.7 or 0.8 level where we could actually feel fully confident prescribing this stuff.

# Never Tell Me The Odds (Ratio)

[Epistemic status: low confidence, someone tell me if the math is off. Title was stolen from an old Less Wrong post that seems to have disappeared – let me know if it’s yours and I’ll give you credit]

I almost screwed up yesterday’s journal club. The study reported an odds ratio of 2.9 for antidepressants. Even though I knew odds ratios are terrible and you should never trust your intuitive impression of them, I still mentally filed this away as “sounds like a really big effect”.

This time I was saved by Chen’s How Big is a Big Odds Ratio? Interpreting the Magnitudes of Odds Ratios in Epidemiological Studies, which explains how to convert ORs into effect sizes. Colored highlights are mine. I have followed the usual statistical practice of interpreting effect sizes of 0.2 as “small”, of 0.5 as “moderate”, and 0.8 as “large”, but feeling guilty about it.



Based on this page, I gather Chen has used some unusually precise formula to calculate this, but that a quick heuristic is to ignore the prevalence and just take [ln(odds ratio)]/1.81.

Suppose you run a drug trial. In your control group of 1000 patients, 300 get better on their own. In your experimental group of 1000 patients, 600 get better total (presumably 300 on their own, 300 because your drug worked). The relative risk calculator says your relative risk of recovery on the drug is 2.0. Odds ratio is 3.5, effect size is 0.7. So you’ve managed to double the recovery rate – in fact, to save an entire extra 30% of your population – and you still haven’t qualified for a “large” effect size.

The moral of the story is that (to me) odds ratios sound bigger than they really are, and effect sizes sound smaller, so you should be really careful comparing two studies that report their results differently.

# 2019 Predictions: Calibration Results

At the beginning of every year, I make predictions. At the end of every year, I score them (this year I’m very late). Here are 2014, 2015, 2016, 2017, and 2018.

And here are the predictions I made for 2019. Strikethrough’d are false. Intact are true. Italicized are getting thrown out because I can’t decide if they’re true or not. All of these judgments were as of December 31 2019, not as of now.

Please don’t complain that 50% predictions don’t mean anything; I know this is true but there are some things I’m genuinely 50-50 unsure of. Some predictions are redacted because they involve my private life or the lives of people close to me. A few that started off redacted stopped being secret; I’ve put those in [brackets].

US  
1. Donald Trump remains President: 90%  
2. Donald Trump is impeached by the House: 40%  
3. Kamala Harris leads the Democratic field: 20%  
4. Bernie Sanders leads the Democratic field: 20%  
5. Joe Biden leads the Democratic field: 20%  
6. Beto O’Rourke leads the Democratic field: 20%  
7. Trump is still leading in prediction markets to be Republican nominee: 70%  
8. Polls show more people support the leading Democrat than the leading Republican: 80%  
9. Trump’s approval rating below 50: 90%  
10. Trump’s approval rating below 40: 50%  
11. Current government shutdown ends before Feb 1: 40%  
12. Current government shutdown ends before Mar 1: 80%  
13. Current government shutdown ends before Apr 1: 95%  
14. Trump gets at least half the wall funding he wants from current shutdown: 20%  
15. Ginsberg still alive: 50%

ECON AND TECH  
16. Bitcoin above 1000: 90%  
17. Bitcoin above 3000: 50%  
18. Bitcoin above 5000: 20%  
19. Bitcoin above Ethereum: 95%  
20. Dow above current value of 25000: 80%  
21. SpaceX successfully launches and returns crewed spacecraft: 90%  
22. SpaceX Starship reaches orbit: 10%  
23. No city where a member of the general public can ride self-driving car without attendant: 90%  
24. I can buy an Impossible Burger at a grocery store within a 30 minute walk from my house: 70%  
25. Pregabalin successfully goes generic and costs less than $100/month on GoodRx.com: 50%  
26. No further CRISPR-edited babies born: 80%

WORLD  
27. Britain out of EU: 60%  
28. Britain holds second Brexit referendum: 20%  
29. No other EU country announces plan to leave: 80%  
30. China does not manage to avert economic crisis (subjective): 50%  
31. Xi still in power: 95%  
32. MbS still in power: 95%  
33. May still in power: 70%  
34. Nothing more embarassing than Vigano memo happens to Pope Francis: 80%

SURVEY  
35. …finds birth order effect is significantly affected by age gap: 40%  
36. …finds fluoxetine has significantly less discontinuation issues than average: 60%  
37. …finds STEM jobs do not have significantly more perceived gender bias than non-STEM: 60%  
38. …finds gender-essentialism vs. food-essentialism correlation greater than 0.075: 30%

PERSONAL – INTERNET  
39. SSC gets fewer hits than last year: 70%  
40. I finish and post [New Atheism: The Godlessness That Failed]: 90%  
41. I finish and post [Structures Of Paranoia]: 50%  
42. I finish and post [a sequence based on Secret Of Our Success]: 50%  
43. [New Atheism] post gets at least 40,000 hits: 40%  
44. [The Proverbial Murder Mystery] post gets at least 40,000 hits: 20%  
45. New co-blogger with more than 3 posts: 20%  
46. Repeat adversarial collaboration contest with at least 5 entries: 60%  
47. [Culture War thread successfully removed from subreddit]: 90%  
48. [Culture War new version getting at least 500 comments per week]: 70%  
49. I start using Twitter again (5+ tweets in any month): 60%  
50. I start using Facebook again (following at least 5 people): 30%

PERSONAL – HEALTH  
51. I get the blood tests I should be getting this year: 90%  
52. I try one biohacking project per month x at least 10 months: 30%  
53. I continue taking sceletium regularly: 70%  
54. I switch from [Zembrin to Tristill] for at least 3 months: 20%  
55. I find at least one new supplement I take or expect to take regularly x 3 months: 20%  
56. Minoxidil use produces obvious progress: 50%  
57. I restart [redacted]: 20%  
58. I spend one month at least substantially more vegetarian than my current compromise: 20%  
59. I spend one month at least substantially less vegetarian than my current compromise: 30%  
60. I weigh more than 195 lbs at year end: 80%  
61. I meditate at least 30 minutes/day more than half of days this year: 30%  
62. I use marijuana at least once this year: 20%

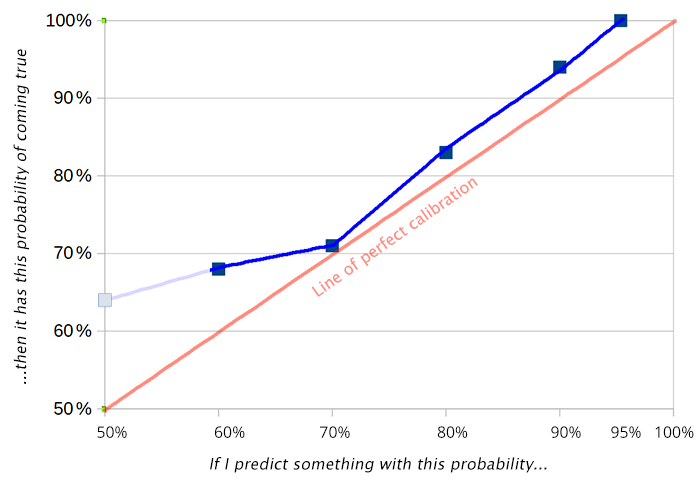
PERSONAL – PROJECTS  
63. I finish at least 10% more of [redacted]: 20%  
64. I completely finish [redacted]: 10%  
65. I finish and post [redacted]: 5%  
66. I write at least ten pages of something I intend to turn into a full-length book this year: 20%  
67. I practice calligraphy at least seven days in the last quarter of 2019: 40%  
68. I finish at least one page of the [redacted] calligraphy project this year: 30%  
69. I finish the entire [redacted] calligraphy project this year: 10%  
70. I finish some other at-least-one-page calligraphy project this year: 80%

PERSONAL – PROFESSIONAL  
71. I attend the APA Meeting: 80%  
72. [redacted]: 50%  
73. [redacted]: 40%  
74. I still work in SF with no plans to leave it: 60%  
75. I still only do telepsychiatry one day with no plans to increase it: 60%  
76. I still work the current number of hours per week: 60%  
77. I have not started (= formally see first patient) my own practice: 80%  
78. I lease another version of the same car I have now: 90%

PERSONAL – HOUSE  
79. I still live in my current house with no specific plans to leave: 80%  
80. I set up a decent home library: 60%  
81. We got a second trash can: 90%  
82. The gate is fixed with no problems at all: 50%  
83. The ugly paint spot on my wall gets fixed: 30%  
84. There is some kind of nice garden: 60%  
85. …and I am at least half responsible: 20%  
86. I get my own washing machine: 20%  
87. There is another baby in my house: 60%  
88. No other non-baby resident (expected 6+ month) in my house who doesn’t live there now: 70%  
89. No existing resident moves away (except the one I already know about): 80%  
90. No other long-term (expected 6+ month) resident of my subunit who doesn’t live there now: 80%  
91. [Decision Tree House] is widely considered a success: 70%  
92. …with plans (vague okay) to create a second one: 20%

PERSONAL – ROMANCE  
93. I find a primary partner: 30%  
94. I go on at least one date with someone who doesn’t already have a primary partner: 90%  
95. I remake an account on OKCupid: 80%  
96. [redacted]: 10%  
97. [redacted]: 20%  
98. [redacted]: 20%  
99. [redacted]: 20%  
100. [redacted]: 20%  
101. [redacted]: 30%  
102. [redacted]: 10%  
103. [redacted]: 30%  
104. [I go on at least three dates with someone I have not yet met]: 50%  
105. [redacted]: 10%  
106. [redacted]: 50%

PERSONAL – FRIENDS  
107. I am still playing D&D: 60%  
108. I go on a trip to Guatemala: 90%  
109. I go on at least one other international trip: 30%  
110. I go to at least one Solstice outside the Bay: 40%  
111. I go to at least one city just for an SSC meetup: 30%  
112. [redacted] is in a relationship: 40%  
113. [redacted] still has their current partner: 50%  
114. [redacted] is at their current job: 20%  
115. [redacted] is still at their current job: 80%  
116. I hang out with [redacted] at least once: 60%  
117. I hang out with [redacted] at least once: 60%  
118. I am in [redacted] Discord server: 80%



Calibration chart. The red line represents perfect calibration, the blue my predictions. The closer they are, the better I am doing.

Of 11 predictions at 50%, I got 4 wrong and 7 right, for an average of 64%  
Of 22 predictions at 60%, I got 7 wrong and 15 right, for an average of 68%  
Of 17 predictions at 70%, I got 5 wrong and 12 right, for an average of 71%  
Of 37 predictions at 80%, I got 6 wrong and 31 right, for an average of 83%  
Of 17 predictions at 90%, I got 1 wrong and 16 right, for an average of 94%  
Of 5 predictions at 95%, I got 0 wrong and 5 right, for an average of 100%

50% predictions are technically meaningless since I could have written them either way. I’ve lightened them on the chart to indicate they can be ignored.

It was another good year for me. Unlike past years, where I erred about evenly in both directions, this year I was about 4% underconfident across the board. I’m not sure how much I should adjust and become more confident. In past years I’ve been burned by major black swan events that affect multiple predictions and made me look overconfident. In 2019 I tried to leave a cushion for that, but nothing too unexpected happened and I ended up playing it too safe. My worst failures were underestimating Bitcoin (but who didn’t?) and overestimating SpaceX’s ability to launch their crew on schedule. I didn’t check formally, but there doesn’t seem to be much difference in my calibration about world affairs vs. my personal life.

I forgot to make predictions for 2020 until now, which in retrospect was the best prediction I’ve ever made. I’ll probably come up with some later this month.

# A Failure, But Not Of Prediction

I.

Vox asks What Went Wrong With The Media’s Coronavirus Coverage? They conclude that the media needs to be better at “not just saying what we do know, but what we don’t know”. This raises some important questions. Like: how much ink and paper is there in the world? Are we sure it’s enough? But also: how do you become better at saying what you don’t know?

In case you’ve been hiding under a rock recently (honestly, valid) the media not only failed to adequately warn its readers about the epidemic, but actively mocked and condescended to anyone who did sound a warning. Real Clear Politics has a list of highlights. The Vox tweet saying “Is this going to be a deadly pandemic? No.” Washington Post telling us in February “Why we should be wary of an aggressive government reponse to coronavirus (it might “scapegoat marginalized populations”). The Daily Beast complaining that “coronavirus, with zero American fatalities, is dominating headlines, while the flu is the real threat”. The New York Times, weighing in with articles like “The pandemic panic” and “Who says it’s not safe to travel to China”. The constant attempts to attribute “alarmism” over the virus to anti-Chinese racism. Etc, etc, etc.

One way people have summed this up is that the media (and the experts they relied on) did a terrible job predicting what would happen. I think this lets them off too easy.

Prediction is very hard. Nate Silver is maybe the best political predicter alive, and he estimated a 29% chance of Trump winning just before Trump won. UPenn professor Philip Tetlock has spent decades identifying “superforecasters” and coming up with complicated algorithms for aggregating their predictions, developing a prediction infrastructure that beats top CIA analysts, but they estimated a 23% chance Britain would choose Brexit just before it happened. This isn’t intended to criticize Silver or Tetlock. I believe they’re operating at close to optimum – the best anyone could possibly do with the information that they had. But the world is full of noise, and tiny chance events can have outsized effects, and there are only so many polls you can scrutinize, and even geniuses can only do so well.

Predicting the coronavirus was equally hard, and the best institutions we had missed it. On February 20th, Tetlock’s superforecasters predicted only a 3% chance that there would be 200,000+ coronavirus cases a month later (there were). The stock market is a giant coordinated attempt to predict the economy, and it reached an all-time high on February 12, suggesting that analysts expected the economy to do great over the following few months. On February 20th it fell in a way that suggested a mild inconvenience to the economy, but it didn’t really start plummeting until mid-March – the same time the media finally got a clue. These aren’t empty suits on cable TV with no skin in the game. These are the best predictive institutions we have, and they got it wrong. I conclude that predicting the scale of coronavirus in mid-February – the time when we could have done something about it – was really hard.

I don’t like this conclusion. But I have to ask myself – if it was so easy, why didn’t I do it? It’s easy to look back and say “yeah, I always secretly knew it would be pretty bad”. I did a few things right – I started prepping half-heartedly in mid-February, I recommended my readers prep in early March, I never criticized others for being alarmist. Overall I give myself a solid B-. But if it was so easy, why didn’t I post “Hey everyone, I officially predict the coronavirus will be a nightmarish worldwide pandemic” two months ago? It wouldn’t have helped anything, but I would have had bragging rights forever. For that matter, why didn’t you post this – on Facebook, on Twitter, on the comments here? You could have gone down in legend, alongside Travis W. Fisher, for making a single tweet. Since you didn’t do that (aside from the handful of you who did – we love you, Balaji) I conclude that predicting it was hard, even for smart and well-intentioned people like yourselves.

Does that mean we can’t put everyone’s heads on spikes outside the Capitol Building as a warning for future generations? I would be very disappointed if it meant that. I think we can still put heads on spikes. We just have to do it for more subtle, better-thought-out reasons.

II.

I used to run user surveys for a forum on probabilistic reasoning…

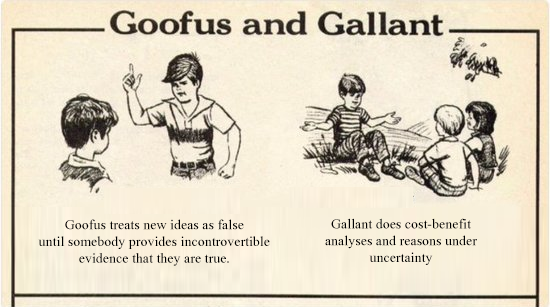
(I promise this will become relevant soon)

A surprising number of these people had signed up for cryonics – the thing where they freeze your brain after you die, in case the future invents a way to resurrect frozen brains. Lots of people mocked us for this – “if you’re so good at probabilistic reasoning, how can you believe something so implausible?” I was curious about this myself, so I put some questions on one of the surveys.

The results were pretty strange. Frequent users of the forum (many of whom had pre-paid for brain freezing) said they estimated there was a 12% chance the process would work and they’d get resurrected. A control group with no interest in cryonics estimated a 15% chance. The people who were doing it were no more optimistic than the people who weren’t. What gives?

I think they were actually good at probabilistic reasoning. The control group said “15%? That’s less than 50%, which means cryonics probably won’t work, which means I shouldn’t sign up for it.” The frequent user group said “A 12% chance of eternal life for the cost of a freezer? Sounds like a good deal!”

There are a lot of potential objections and complications – for one thing, maybe both those numbers are much too high. You can read more here and here. But overall I learned something really important from this.



Making decisions is about more than just having certain beliefs. It’s also about how you act on them.

III.

A few weeks ago, I wrote a blog post on face masks. It reviewed the evidence and found that they probably helped prevent the spread of disease. Then it asked: how did the WHO, CDC, etc get this so wrong?

I went into it thinking they’d lied to us, hoping to prevent hoarders from buying up so many masks that there weren’t enough for health workers. Turns out that’s not true. The CDC has been singing the same tune for the past ten years. Swine flu, don’t wear masks. SARS, don’t wear masks. They’ve been really consistent on this point. But why?

If you really want to understand what happened, don’t read any studies about face masks or pandemics. Read Smith & Pell (2003), Parachute Use To Prevent Death And Major Trauma Related To Gravitational Challenge: Systematic Review Of Randomized Controlled Trials. It’s an article in the British Journal Of Medicine pointing out that there have never been any good studies proving that parachutes are helpful when jumping out of a plane, so they fail to meet the normal standards of evidence-based medicine. From the Discussion section:

It is a truth universally acknowledged that a medical intervention justified by observational data must be in want of verification through a randomised controlled trial. Observational studies have been tainted by accusations of data dredging, confounding, and bias. For example, observational studies showed lower rates of ischaemic heart disease among women using hormone replacement therapy, and these data were interpreted as advocating hormone replacement for healthy women, women with established ischaemic heart disease, and women with risk factors for ischaemic heart disease. However, randomised controlled trials showed that hormone replacement therapy actually increased the risk of ischaemic heart disease, indicating that the apparent protective effects seen in observational studies were due to bias. Cases such as this one show that medical interventions based solely on observational data should be carefully scrutinised, and the parachute is no exception.

Of course this is a joke. It’s in the all-joke holiday edition of BMJ, and everyone involved knew exactly what they were doing. But the joke is funny because it points at something true. It’s biting social commentary. Doctors will not admit any treatment could possibly be good until it has a lot of randomized controlled trials behind it, common sense be damned. This didn’t come out of nowhere. They’ve been burned lots of times before by thinking they were applying common sense and getting things really wrong. And after your mistakes kill a few thousand people you start getting really paranoid and careful. And there are so many quacks who can spout off some “common sense” explanation for why their vitamin-infused bleach or colloidal silver should work that doctors have just become immune to that kind of bullshit. Multiple good RCTs or it didn’t happen. Given the history I think this is a defensible choice, and if you are tempted to condemn it you may find this story about bone marrow transplants enlightening.

But you can take this too far. After highlighting the lack of parachute RCTs, the paper continues:

Only two options exist. The first is that we accept that, under exceptional circumstances, common sense might be applied when considering the potential risks and benefits of interventions. The second is that we continue our quest for the holy grail of exclusively evidence based interventions and preclude parachute use outside the context of a properly conducted trial. The dependency we have created in our population may make recruitment of the unenlightened masses to such a trial difficult. If so, we feel assured that those who advocate evidence based medicine and criticise use of interventions that lack an evidence base will not hesitate to demonstrate their commitment by volunteering for a double blind, randomised, placebo controlled, crossover trial.

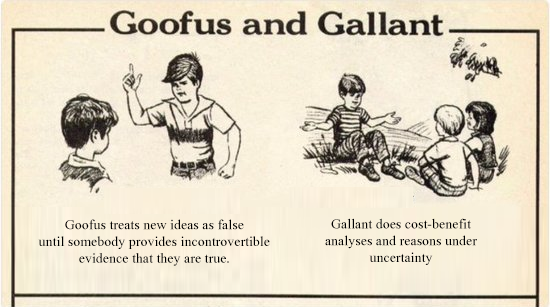
Did you follow that? For a good parachute RCT, half the subjects would have to jump out of a plane wearing a placebo parachute. The authors suggest maybe we enlist doctors who insist too stringently on RCTs over common sense for this dubious honor.

(good news, though, a parachute RCT did eventually get done)

Sometimes good humor is a little too on the nose, like those Onion articles that come true a few years later. The real medical consensus on face masks came from pretty much the same process as the fake medical consensus on parachutes. Common sense said that they worked. But there weren’t many good RCTs. We couldn’t do more, because it would have been unethical to deliberately expose face-mask-less people to disease. In the end, all we had were some mediocre trials of slightly different things that we had to extrapolate out of range.

Just like the legal term for “not proven guilty beyond a reasonable doubt” is “not guilty”, the medical term for “not proven to work in several gold-standard randomized controlled trials” is “it doesn’t work” (and don’t get me started on “no evidence”). So the CDC said masks didn’t work.

Going back to our diagram:



Goofus started with the position that masks, being a new idea, needed incontrovertible proof. When the few studies that appeared weren’t incontrovertible enough, he concluded that people shouldn’t wear masks.

Gallant would have recognized the uncertainty – based on the studies we can’t be 100% sure masks definitely work for this particular condition – and done a cost-benefit analysis. Common sensically, it seems like masks probably should work. The existing evidence for masks is highly suggestive, even if it’s not utter proof. Maybe 80% chance they work, something like that? If you can buy an 80% chance of stopping a deadly pandemic for the cost of having to wear some silly cloth over your face, probably that’s a good deal. Even though regular medicine has good reasons for being as conservative as it is, during a crisis you have to be able to think on your feet.

IV.

But getting back to the media:

Their main excuse is that they were just relaying expert opinion – the sort of things the WHO and CDC and top epidemiologists were saying. I believe them. People on Twitter howl and gnash their teeth at this, asking why the press didn’t fact-check or challenge those experts. But I’m not sure I want to institute a custom of journalists challenging experts. Journalist Johann Hari decided to take it upon himself to challenge psychiatric experts, and wrote a serious of terrible articles and a terrible book saying they were wrong about everything. I am a psychiatrist and I can tell you he is so wrong that it is physically painful to read his stuff (though of course I would say that…). Most journalists stick to assuming the experts know more about their subject of expertise than they do, and I think this is wise. The role of science journalists is to primarily to relay, explain, give context to the opinions of experts, not to try to out-medicine the doctors. So I think this is a good excuse.

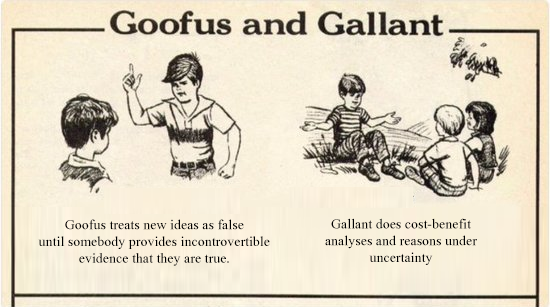
But I would ask this of any journalist who pleads that they were just relaying and providing context for expert opinions: what was the experts’ percent confidence in their position?

I am so serious about this. What fact could possibly be more relevant? What context could it possibly be more important to give? I’m not saying you need to have put a number in your articles, maybe your readers don’t go for that. But were you working off of one? Did this question even occur to you?

Nate Silver said there was a 29% chance Trump would win. Most people interpreted that as “Trump probably won’t win” and got shocked when he did. What was the percent attached to your “coronavirus probably won’t be a disaster” prediction? Was it also 29%? 20%? 10%? Are you sure you want to go lower than 10%? Wuhan was already under total lockdown, they didn’t even have space to bury all the bodies, and you’re saying that there was less than 10% odds that it would be a problem anywhere else? I hear people say there’s a 12 – 15% chance that future civilizations will resurrect your frozen brain, surely the risk of coronavirus was higher than that?

And if the risk was 10%, shouldn’t that have been the headline. “TEN PERCENT CHANCE THAT THERE IS ABOUT TO BE A PANDEMIC THAT DEVASTATES THE GLOBAL ECONOMY, KILLS HUNDREDS OF THOUSANDS OF PEOPLE, AND PREVENTS YOU FROM LEAVING YOUR HOUSE FOR MONTHS”? Isn’t that a better headline than Coronavirus panic sells as alarmist information spreads on social media? But that’s the headline you could have written if your odds were ten percent!

So:



I think people acted like Goofus again.

People were presented with a new idea: a global pandemic might arise and change everything. They waited for proof. The proof didn’t arise, at least at first. I remember hearing people say things like “there’s no reason for panic, there are currently only ten cases in the US”. This should sound like “there’s no reason to panic, the asteroid heading for Earth is still several weeks away”. The only way I can make sense of it is through a mindset where you are not allowed to entertain an idea until you have proof of it. Nobody had incontrovertible evidence that coronavirus was going to be a disaster, so until someone does, you default to the null hypothesis that it won’t be.

Gallant wouldn’t have waited for proof. He would have checked prediction markets and asked top experts for probabilistic judgments. If he heard numbers like 10 or 20 percent, he would have done a cost-benefit analysis and found that putting some tough measures into place, like quarantine and social distancing, would be worthwhile if they had a 10 or 20 percent chance of averting catastrophe.

V.

This is at risk of getting too depressing, so I want to focus on some people who deserve recognition for especially good responses.

First, a bunch of generic smart people on Twitter who got things exactly right – there are too many of these people to name, but Scott Aaronson highlights “Bill Gates, Balaji Srinivasan, Paul Graham, Greg Cochran, Robin Hanson, Sarah Constantin, Eliezer Yudkowsky, and Nicholas Christakis.” None of these people (except Greg Cochran) are domain experts, and none of them (except Greg Cochran) have creepy oracular powers. So how could they have beaten the experts? Haven’t we been told a million times that generic intelligence is no match for deep domain knowledge?

I think the answer is: they didn’t beat the experts in epidemiology. Whatever probability of pandemic the experts and prediction markets gave for coronavirus getting really bad, these people didn’t necessarily give a higher probability. They were just better at probabilistic reasoning, so they had different reactions to the same number. There’s no reason generic why smart people shouldn’t be better at probabilistic reasoning then epidemiologists. In fact, this seems exactly like the sort of thing generic smart people might be.

Zeynep Tufekci is an even clearer example. She’s a sociologist and journalist who was writing about how it was “our civic duty” to prepare for coronavirus as early as February. She was also the first mainstream media figure to spread the word that masks were probably helpful.

Totally at random today, reading a blog post on the Mongol Empire like all normal people do during a crisis, I stumbled across a different reference to Zeynep. In a 2014 article, she was sounding a warning about the Ebola pandemic that was going on at the time. She was saying the exact same things everyone is saying now – global institutions are failing, nobody understands exponential growth, travel restrictions could work early but won’t be enough if it breaks out. She quoted a CDC prediction that there could be a million cases by the end of 2014. “Let that sink in,” she wrote. “A million Ebola victims in just a few months.”

In fact, this didn’t happen. There were only about 30,000 cases. The virus never really made it out of Liberia, Sierra Leone, and Guinea.

I don’t count this as a failed prediction on Zeynep’s part. First of all, because it could have been precisely because of people like her sounding the alarm that the epidemic was successfully contained. But more important, it wasn’t really a prediction at all. Her point wasn’t that she definitely knew this Ebola pandemic was the one that would be really bad. Her point was that it might be, so we needed to prepare. She said the same thing when the coronavirus was just starting. If this were a game, her batting average would be 50%, but that’s the wrong framework.

Zeynep Tufecki is admirable. But her admirable skill isn’t looking at various epidemics and successfully predicting which ones will be bad and which ones will fizzle out. She can’t do that any better than anyone else. Her superpower is her ability to treat something as important even before she has incontrovertible evidence that it has to be.

And finally, Kelsey Piper. She wrote a February 6th article saying:

The coronavirus killed fewer people than the flu did in January. But it might kill more in February — and unlike the flu, its scope and effects are poorly understood and hard to guess at. The Chinese National Health Commission reports 24,324 cases, including 3,887 new ones today. There are some indications that these numbers understate the situation, as overwhelmed hospitals in Wuhan only have the resources to test the most severe cases. As of Tuesday, 171,329 people are under medical observation because they’ve had close contact with a confirmed case.

It is unclear whether China will be able to get the outbreak under control or whether it will cause a series of epidemics throughout the country. It’s also unclear whether other countries — especially those with weak health systems — will be able to quickly identify any cases in their country and avoid Wuhan-scale outbreaks.

The point is, it’s simply too soon to assert we’ll do well on both those fronts — and if we fail, then the coronavirus death toll could well climb up into the tens of thousands. It also remains to be seen if vaccines or effective antiviral treatments will be developed. That’s just far too much uncertainty to assure people that they have nothing to worry about. And misleadingly assuring people that there’s nothing to worry about can end up doing harm.

“Instead of deriding people’s fears about the Wuhan coronavirus,” Sandman, the communications expert, writes, “I would advise officials and reporters to focus more on the high likelihood that things will get worse and the not-so-small possibility that they will get much worse.”

She concluded that “the Wuhan coronavirus likely won’t be a nightmare pandemic, but that scenario is still in play”, and followed it up with an article urging people to prepare by buying essential food and supplies.

If we interpret her “likely won’t be a nightmare pandemic” sentence as a prediction, she got the prediction wrong. Like Zeynep, she has no special ability to predict whether any given disease will end in global disaster. But that didn’t matter! She gave exactly the correct advice to institutions (prepare for a worst-case scenario, stop telling people not to panic) and exactly the correct advice to individuals (start prepping). When you’re good enough at handling uncertainty, getting your predictions exactly right becomes almost superfluous.

The Vox article says the media needs to “say what it doesn’t know”. I agree with this up to a point. But they can’t let this turn into a muddled message of “oh, who knows anything, whatever”. Uncertainty about the world doesn’t imply uncertainty about the best course of action! Within the range of uncertainty that we had about the coronavirus this February, an article that acknowledged that uncertainty wouldn’t have looked like “We’re not sure how this will develop, so we don’t know whether you should stop large gatherings or not”. It would have looked like “We’re not sure how this will develop, so you should definitely stop large gatherings.”

I worry that the people who refused to worry about coronavirus until too late thought they were “being careful” and “avoiding overconfidence”. And I worry the lesson they’ll take away from this is to be more careful, and avoid overconfidence even more strongly.

Experts should think along these lines when making their recommendations, but if they don’t, the press should think along them as part of its work of putting expert recommendations in context. I think Kelsey’s article provides an shining example of what this should look like.

Maybe other people got this right too. I’m singling out Kelsey because of a personal connection – I met her through the same probabilistic reasoning forum where I did my cryonics survey years ago. I don’t think this is a coincidence.

[Related: Book Review: The Precipice; Two Kinds Of Caution]

# Depression: The Olfactory Perspective

Depressed people have worse sense of smell, and people with worse sense of smell are more likely to get depressed. Kohli 2016 tries to figure out what’s going on.

They review six studies testing how well depressed people can smell things. Most use something called “The Sniffin’ Sticks Test” (really!) where people are asked to say which of two sticks has an odor; the strength of the odorous one is then decreased until the subject can no longer consistently get it right. This determines olfactory threshold – how sensitive the subject’s smell is. Depressed subjects did marginally (but significantly) worse on this test than controls (6.31 ± 1.38 vs. 6.78 ± 0.88; P = 0.0005) – I think this corresponds to an effect size of about 0.2. They also do a couple more tests to see if depressed people are worse at identifying odors and get similarly small results. Also, some neuroimaging studies directly correlate depression and olfactory bulb volume, and find that olfactory areas of depressed people’s brains shrink.

Next they look at three studies of people who have various known respiratory diseases that impair sense of smell, and see how many of them are depressed. The answer: lots! Normosmics (people with a normal sense of smell) have an average score of 5 on the Beck Depression Inventory. Anosmics (people with no sense of smell) have an average score of 14. Hyposmics (people with weak sense of smell) are in between. This seemed true independent of respiratory disease level (ie it’s not just that having a worse respiratory disease makes your smell worse and makes you more depressed).

The two most common diseases studied were chronic rhinosinusitis and post-upper respiratory infection olfactory dysfunction. One of the studies, Jung, Lee, & Park (2014), compares them. They find that 40% of CRS patients are depressed, vs. 76% of PURIOD patients, a significant difference (also, both much higher than the ~5-10% of depressives in the general population). They note that CRS involves a gradual loss of smell, and PURIOD a sudden loss of smell, and that maybe people adjust to gradual loss better than sudden.

We can easily come up with alternative hypotheses. You can never actually control for confounders properly, and having a respiratory disease sounds depressing. Also, the sorts of people who have respiratory diseases probably differ from the rest of the population in various ways. Biologically, they might have worse immune systems, or end up with worse oxygenation and chronic infections, or just be generally sickly. Sociologically, they might be poorer, or have worse diets, or more comorbidities. Sinus problems themselves are linked to depression for a bunch of reasons, probably relating to the sinus being so closely linked to the brain that sinus inflammation screws up your brain chemistry. Also, most chronic diseases have high depression rates – sickle cell anemia (chosen kind of at random) is 44%.

There really aren’t that many causes of anosmia that are 100% exogenous and have no chance at all of contributing to depression themselves. But animal researchers can take Gandhi’s advice and become the exogenous cause they wish to see in the world. Yuan and Slotnick (2014) discuss experiments where researchers remove the olfactory bulbs of rats. These rats tend to consistently become depressed. That seems like pretty strong evidence to me.

So why would depression reduce olfactory acuity? And why would reduced olfaction cause depression?

Maybe depression reduces olfactory acuity because it reduces sensory acuity in general. We already know that depression decreases visual contrast, causing the world to literally look washed-out and gray. If I’m reading this study right, it claims depression also decreases auditory threshold – ie depressed people are less able to hear very quiet sounds. I can’t find similarly good studies on taste or touch, although they should be easy to do. If anyone knows any studies on eg two-point discrimination in depressed vs. non-depressed subjects, let me know. If not, this would be an easy project that a college psychology student could do in a few weeks, and it would add to our understanding of this subject. But right now I think the evidence is consistent with a general decline in perceptual abilities. This fits my understanding of depression as a systemic disorder – the low mood is the most noticeable symptom, but you’re also getting everything from poor digestion to worse hearing. Probably this is because whatever is affecting the emotional centers of the brain is affecting the lower parts of the brain (and nervous system?) too.

Is the opposite of this true? Does any form of decreased sensory acuity cause depression? I’m leaning toward no. This study claims to find that myopia is linked to depression, but the association is so weak that I take it as a strike against the hypothesis. These two studies find hearing loss is associated with depression, but it’s still pretty weak and it could easily be because poor hearing hurts your opportunities to socialize. I’m most struck by discussion of monochromacy, ie total colorblindness, which absolutely fail to mention depression at all. If I’m going to be splitting hairs about how depressed people see slightly duller colors, the lack of any excess depression in people without color vision seems pretty important. Blind and deaf people have pretty high rates of depression, but being blind or deaf is really hard and I don’t want to draw too many conclusions there. Overall it doesn’t seem to me like decreased sensory acuity causes depression full stop, though I could be wrong. This would mean that loss of smell is unique in its emotional effects.

The sense of smell is pretty emotionally salient. Everyone always says that “smell is the sense most linked to memory” – though I can’t figure out exactly which study discovered it and whether it considered (for example) how reliably seeing a picture of my mother reminds me of my mother. Body odor seems to be closely linked to who we’re attracted to. Smell is responsible for all taste sensation beyond sweet/salty/sour/bitter/umami, and eating food is one of life’s most visceral pleasures. From fifthsense.org.uk:

Anosmia sufferers often talk of feeling isolated and cut-off from the world around them, and experiencing a ‘blunting’ of the emotions. Smell loss can affect one’s ability to form and maintain close personal relationships and can lead to depression. An important issue here is the fact that smell loss is invisible to all but the patient; how would you know that you had met an anosmia sufferer unless they themselves told you? This is one of the reasons, alongside the general lack of understanding of the impact that smell has on our lives, why anosmia has never received much attention – you really do not know what you have got until it is gone.

But also, Yuan and Slotnick’s rat paper gives a more biological explanation. The olfactory bulb is the beginning of pathways that stimulate many other parts of the brain. When it’s removed (and presumably also when it just never gets any incoming stimuli) it stops doing that, and the downstream parts of the brain shrink. For some reason this also causes decreased brain-wide synthesis of serotonin, maybe because the olfactory bulb is a net positive stimulus on the raphe nuclei. Sure, sounds like the kind of thing that might cause depression.

What does this imply about treating depression?

SSRI antidepressants probably decrease sense of smell as an immediate side effect. Awkward. But they seem to improve sense of smell long term as part of their general treatment of depression. In studies fluoxetine (Prozac) causes nerve growth in the olfactory bulb after a few weeks. Does that suggest a story where SSRIs work by improving smell? Probably not – more likely they work by [a cascading system of effects involving] causing nerve growth more generally, and the olfactory bulb benefits along with everything else.

Could you treat depression through improving olfactory sensitivity? Maybe, but I have no idea how to do that. Yuan and Slotnick suggest directly stimulating the olfactory regions of the brain, but this is pretty invasive, and there are probably already better treatments for the small minority of patients who are going to let you directly stimulate brain regions. Maybe before we worry about this problem, we should investigate the more general question of sensory enhancement for depression. What would happen if you made people wear glasses that enhanced the color saturation of everything they saw?

Finally, what about just exposing depressed people to really strong smells? You will be excited to know that real scientists have studied this ridiculous idea, and it seems promising, at least in extremely sketchy experiments on mice (1, 2, 3). I have never seen any studies done on humans unless this is actually how “aromatherapy” works, which would be hilarious. Aromatherapy seems to get positive results for depression with the same kind of bad studies that let all quackery generate positive results for everything. I can’t say more than that and I’m pretty skeptical here.

# The Amish Health Care System

I.

Amish people spend only a fifth as much as you do on health care, and their health is fine. What can we learn from them?

A reminder: the Amish are a German religious sect who immigrated to colonial America. Most of them live apart from ordinary Americans (who they call “the English”) in rural communities in Pennsylvania and Ohio. They’re famous for their low-tech way of life, generally avoiding anything invented after the 1700s. But this isn’t absolute; they are willing to accept technology they see as a net positive. Modern medicine is in this category. When the Amish get seriously ill, they will go to modern doctors and accept modern treatments.

The Muslims claim Mohammed was the last of the prophets, and that after his death God stopped advising earthly religions. But sometimes modern faiths will make a decision so inspired that it could only have come from divine revelation. This is how I feel about the Amish belief that health insurance companies are evil, and that good Christians must have no traffic with them.

And Deists believe that God is like a watchmaker, an artisan who built the world but does not act upon it. But by some miracle, the US government played along and granted the Amish exemptions from all the usual health care laws. They don’t have to pay Medicare taxes or social security. They aren’t included in the Obamacare mandate. They can share health care costs the way they want, ignoring any regulations to the contrary. They are genuinely on their own.

They’ve ended up with a simple system based on church aid. Everyone pays tithes to their congregation (though they don’t call it that). The churches meet in houses and have volunteer leaders, so expenses minimal. Most of the money goes to “alms” which the bishop distributes to members in need. This replaces the social safety net, including health insurance. Most Amish go their entire life without needing anything else.

About a third of Amish are part of a more formal insurance-like institution called Amish Hospital Aid. Individuals and families pay a fixed fee to the organization, which is not-for-profit and run by an unpaid board of all-male elders. If they need hospital care, AHA will pay for it. How does this interact with the church-based system? Rohrer and Dundes, my source for most of this post, say that it’s mostly better-off Amish who use AHA. Their wealth is tied up in their farmland, so it’s not like they can use it to pay hospital bills. But they would feel guilty asking their church to give them alms meant for the poor. AHA helps protect their dignity and keep church funds for those who need them most.

How well does this system work?

The Amish outperform the English on every measured health outcome. 65% of Amish rate their health as excellent or very good, compared to 58% of English. Diabetes rates are 2% vs. 8%, heart attack rates are 1% vs. 6%, high blood pressure is 11% vs. 31%. Amish people go to the hospital about a quarter as often as English people, and this difference is consistent across various categories of illness (the big exception is pregnancy-related issues – most Amish women have five to ten children). This is noticeable enough that lots of health magazines have articles on The Health Secrets of the Amish and Amish Secrets That Will Add Years To Your Life. As far as I can tell, most of the secret is spending your whole life outside doing strenuous agricultural labor, plus being at a tech level two centuries too early for fast food.

But Amish people also die earlier. Lots of old studies say the opposite – for example, this one finds Amish people live longer than matched Framingham Heart Study participants. But things have changed since Framingham. The Amish have had a life expectancy in the low 70s since colonial times, when the rest of us were dying at 40 or 50. Since then, Amish life expectancy has stayed the same, and English life expectancy has improved to the high 70s. The most recent Amish estimates I have still say low seventies, so I think we are beating them now.

If they’re healthier, why is their life expectancy lower? Possibly they are less interested in prolonging life than we are. R&D write:

Amish people are more willing to stop interventions earlier and resist invasive therapies than the general population because, while they long for healing, they also have a profound respect for God’s will. This means taking modest steps toward healing sick bodies, giving preference to natural remedies, setting common-sense limits, and believing that in the end their bodies are in God’s hands.

The Amish health care system has an easier job than ours does. It has to take care of people who are generally healthy and less interested in extreme end-of-life care. It also supports a younger population – because Amish families have five to ten children, the demographics are weighted to younger people. All of these make its job a little bit simpler, and we should keep that in mind for the following sections.

How much do the Amish pay for health care? This is easy to answer for Amish Hospital Aid, much harder for the church system.

Amish Hospital Aid charges $125 monthly per individual or $250 monthly per family (remember, Amish families can easily be ten people). Average US health insurance costs $411 monthly per individual (Obamacare policies) or $558 monthly per individual (employer sponsored plan; employers pay most of this). I’m not going to bother comparing family plans because the definition of “family” matters a lot here. On the surface, it looks like the English spend about 4x as much as the Amish do.

But US plans include many more services than AHA, which covers catastrophic hospital admissions only. The government bans most Americans from buying plans like this; they believe it’s not enough to count as real coverage. The cheapest legal US health plan varies by age and location, but when I take my real age and pretend that I live near Amish country, the government offers me a $219/month policy on Obamacare. This is only a little higher than what the Amish get, and probably includes more services. So here it seems like the Amish don’t have much of an efficiency advantage. They just make a different tradeoff. It’s probably the right tradeoff for them, given their healthier lifestyle.

But remember, only a third of Amish use AHA. The rest use a church-based system? How does that come out?

It’s hard to tell. Nobody agrees on how much Amish tithe their churches, maybe because different Amish churches have different practices. R&D suggest families tithe 10% of income, this article on church-based insurances says a flat $100/month fee, and this “Ask The Amish” column says that churches have twice-yearly occasions where they ask for donations in secret and nobody is obligated to give any particular amount (“often husbands and wives won’t even know how much the other is giving.”) So it’s a mess, and even knowing the exact per-Amish donation wouldn’t help, because church alms cover not just health insurance but the entire social safety net; the amount that goes to health care probably varies by congregation and circumstance.

A few people try to estimate Amish health spending directly. This ABC story says $5 million total for all 30,000 Amish in Lancaster County, but they give no source, and it’s absurdly low. This QZ story quotes Amish health elder Marvin Wengerd as saying $20 – $30 million total for Lancaster County, which would suggest health spending of between $600-$1000 per person. This sounds potentially in keeping with some of the other estimates. A $100 per month tithe would be $1200 per year – if half of that goes to non-health social services, that implies $600 for health. The average Amish family earns about $50K (the same as the average English family, somehow!) so a 10% tithe would be $5000 per year, but since the average Amish family size is seven children, that comes out to about $600 per person again. So several estimates seem to agree on between $600 and $1000 per person.

One possible issue with this number: does Wengerd know how much Amish spend out of pocket? Or does his number just represent the amount that the official communal Amish health system spends? I’m not sure, but taking his words literally it’s total Amish spending, so I am going to assume it’s the intended meaning. And since the Amish rarely see doctors for minor things, probably their communal spending is a big chunk of their total.

[Update: an SSC reader is able to contact his brother, a Mennonite deacon, for better numbers. He says that their church spends an average of $2000 per person (including out of pocket).]

How does this compare to the US as a whole? The National Center For Health Statistics says that the average American spends $11,000 on health care. This suggests that the average American spends between five and ten times more on health care than the average Amish person.

How do the Amish keep costs so low? R&D (plus a few other sources) identify some key strategies.

First, the Amish community bargains collectively with providers to keep prices low. This isn’t unusual – your insurance company does the same – but it nets them better prices than you would get if you tried to pay out of pocket at your local hospital. This article gives some examples of Amish getting sticker prices discounted from between 50% to 66% with this tactic alone; Medicare gets about the same.

Second, the Amish are honorable customers. This separates them from insurance companies, who are constantly trying to scam providers however they can. Much of the increase in health care costs is “administrative expenses”, and much of these administrative expenses is hiring an army of lawyers, clerks, and billing professionals to thwart insurance companies’ attempts to cheat their way out of paying. If you are an honorable Amish person and the hospital knows you will pay your bill on time with zero fuss, they can waive all this.

But can this really be the reason Amish healthcare is cheaper? When insurance companies negotiate with providers, patients are on the side of the insurances; when insurance companies get good deals (eg a deal of zero dollars because the insurance has scammed the hospital), the patient’s care is cheaper, and the insurance company can pass some of those savings down as lower prices. If occasionally scamming providers meant insurance companies had to pay more money total, then they would stop doing it. My impression is that the real losers here are uninsured patients; absent any pressure to do otherwise, hospitals will charge them the sticker price, which includes the dealing-with-insurance-scams fee. The Amish successfully pressure them to waive that fee, which gets them better prices than the average uninsured patient, but still doesn’t land them ahead of insured people.

Third, Amish don’t go to the doctor for little things. They either use folk medicine or chiropractors. Some of the folk medicine probably works. The chiropractors probably don’t, but they play a helpful role reassuring people and giving them the appropriate obvious advice while telling the really serious cases to seek outside care. With this help, Amish people mostly avoid primary care doctors. Holmes County health statistics find that only 16% of Amish have seen a doctor in the past year, compared to 54% of English.

Fourth, the Amish never sue doctors. Doctors around Amish country know this, and give them the medically indicated level of care instead of practicing “defensive medicine”. If Amish people ask their doctors to be financially considerate – for example, let them leave the hospital a little early – their doctors will usually say yes, whereas your doctor would say no because you could sue them if anything went wrong. In some cases, Amish elders formally promise that no member of their congregation will ever launch a malpractice lawsuit.

Fifth, the Amish don’t make a profit. Church aid is dispensed by ministers and bishops. Even Amish Hospital Aid is run by a volunteer board. None of these people draw a salary or take a cut. I don’t want to overemphasize this one – people constantly obsess over insurance company profits and attribute all health care pathologies to them, whereas in fact they’re a low single-digit percent of costs (did you know Kaiser Permanente is a nonprofit? Hard to tell, isn’t it?) But every little bit adds up, and this is one bit.

Sixth, the Amish don’t have administrative expenses. Since the minister knows and trusts everyone in his congregation, the “approval process” is just telling your minister what the problem is, and the minister agreeing that’s a problem and giving you money to solve it. This sidesteps a lot of horrible algorithms and review boards and appeal boards and lawyers. I don’t want to overemphasize this one either – insurance companies are legally required to keep administrative expenses low, and most of them succeed. But again, it all adds up.

Seventh, the Amish feel pressure to avoid taking risks with their health. If you live in a tiny community with the people who are your health insurance support system, you’re going to feel awkward smoking or drinking too much. Realistically this probably blends into a general insistence on godly living, but the health insurance aspect doesn’t hurt. And I’m talking like this is just informal pressure, but occasionally it can get very real. R&D discuss the case of some Amish teens who get injured riding a snowmobile – forbidden technology. Their church decided this was not the sort of problem that godly people would have gotten themselves into, and refused to help – their families were on the hook for the whole bill.

Eighth, for the same reason, Amish try not to overspend on health care. I realize this sounds insulting – other Americans aren’t trying? I think this is harsh but true. Lots of Americans get an insurance plan from their employer, and then consume health services in a price-insensitive way, knowing very well that their insurance will pay for it. Sometimes they will briefly be limited by deductibles or out-of-pocket charges, but after these are used up, they’ll go crazy. You wouldn’t believe how many patients I see who say things like “I’ve covered my deductible for the year, so you might as well give me the most expensive thing you’ve got”, or “I’m actually feeling fine, but let’s have another appointment next week because I like talking to you and my out-of-pocket charges are low.”

But it’s not just avoiding the obvious failure modes. Careful price-shopping can look very different from regular medical consumption. Several of the articles I read talked about Amish families traveling from Pennsylvania to Tijuana for medical treatment. One writer describes Tijuana clinics sending salespeople up to Amish Country to advertise their latest prices and services. For people who rarely leave their hometown and avoid modern technology, a train trip to Mexico must be a scary experience. But prices in Mexico are cheap enough to make it worthwhile.

Meanwhile, back in the modern world, I’ve written before about how a pharma company took clonidine, a workhorse older drug that costs $4.84 a month, transformed it into Lucemyra, a basically identical drug that costs $1,974.78 a month, then created a rebate plan so that patients wouldn’t have to pay any extra out-of-pocket. Then they told patients to ask their doctors for Lucemyra because it was newer and cooler. Patients sometimes went along with this, being indifferent between spending $4 of someone else’s money or $2000 of someone else’s money. Everything in the US health system is like this, and the Amish avoid all of it. They have a normal free market in medical care where people pay for a product with their own money (or their community’s money) and have incentives to check how much it costs before they buy it. I do want to over-emphasize this one, and honestly I am surprised Amish health care costs are only ten times cheaper than ours are.

I don’t know how important each of these factors is, or how they compare to more structural factors like younger populations, healthier lifestyles, and less end-of-life care. But taken together, they make it possible for the Amish to get health care without undue financial burden or government support.

II.

Why look into the Amish health system?

I’m fascinated by how many of today’s biggest economic problems just mysteriously failed to exist in the past. Our grandparents easily paid for college with summer jobs, raised three or four kids on a single income, and bought houses in their 20s or 30s and never worried about rent or eviction again. And yes, they got medical care without health insurance, and avoided the kind of medical bankruptcies we see too frequently today. How did this work so well? Are there ways to make it work today? The Amish are an extreme example of people who try to make traditional systems work in the modern world, which makes them a natural laboratory for this kind of question.

The Amish system seems to work well for the Amish. It’s hard to say this with confidence because of all the uncertainties. The Amish skew much younger than the “English”, and live much healthier lifestyles. Although a few vague estimates suggest health care spending far below the English average, they could be missing lots of under-the-table transactions. And again, I don’t want to ignore the fact that the Amish do live a little bit shorter lives. You could tell a story where all of these add up to explain 100% of the difference, and the Amish aren’t any more efficient in their spending at all. I don’t think this is right. I think the apparent 5x advantage, or something like it, is real. But right now this is just a guess, not a hard number.

What if it is? It’s hard to figure out exactly what it would take to apply the same principles to English society. Only about a quarter of Americans attend church regularly, so church-based aid is out. In theory, health insurance companies ought to fill the same niche, with maybe a 10% cost increase for profits and overhead. Instead we have a 1000% cost increase. Why?

Above, I said that the most important factor is that the Amish comparison shop. Everyone needs to use other people’s money to afford expensive procedures. But for the Amish, those other people are their fellow church members and they feel an obligation to spend it wisely. For the English, the “other people” are faceless insurance companies, and we treat people who don’t extract as much money as possible from them as insufficiently savvy. But there’s no easy way to solve this in an atomized system. If you don’t have a set of thirty close friends you can turn to for financial help, then the only institutions with enough coordination power to make risk pooling work are companies and the government. And they have no way of keeping you honest except the with byzantine rules about “prior authorizations” and “preferred alternatives” we’ve become all too familiar with.

(and as bad as these are, there’s something to be said for a faceless but impartial bureaucracy, compared to having all your neighbors judging your lifestyle all the time.)

This is a neat story, but I have two concerns about it.

First, when I think in terms of individual people I know who have had trouble paying for health care, it’s hard for me to imagine the Amish system working very well for them. Many have chronic diseases. Some have mysterious pain that they couldn’t identify for years before finally getting diagnosed with something obscure. Amish Hospital Aid’s catastrophic policy would be useless for this, and I feel like your fellow church members would get tired of you pretty quickly. I’m not sure how the Amish cope with this kind of thing, and maybe their system relies on a very low rate of mental illness and chronic disease. A lot of the original “hygiene hypothesis” work was done on the Amish, their autoimmune disease rates are amazing, and when you take out the stresses of modern life maybe a lot of the ailments the American system was set up to deal with just stop being problems. I guess my point is that the numbers seem to work out, and the Amish apparently remain alive, but when I imagine trying to apply the Amish system to real people, even assuming those real people have cooperative churches and all the other elements I’ve talked about, I can’t imagine it doing anything other than crashing and burning.

Second, I don’t think this is actually how our grandparents did things. I asked my literal grandmother, a 95 year old former nurse, how health care worked in her day. She said it just wasn’t a problem. Hospitals were supported by wealthy philanthropists and religious organizations. Poor people got treated for free. Middle class people paid as much as they could afford, which was often the whole bill, because bills were cheap. Rich people paid extra for fancy hospital suites and helped subsidize everyone else. Although most people went to church or synagogue, there wasn’t the same kind of Amish-style risk pooling.

This makes me think that the Amish method, even though it works, isn’t the method that worked for past generations. It’s an innovation intended to cover for health care prices being higher than anything that traditional societies had to deal with.

Why did health care prices start rising? I’ve wondered about this a lot before – see here, here, and here. Looking into this issue, I noticed glimpses of a different possibility. The increase started around the same time that health insurance began to spread. In one sense, this is unsurprising – of course health insurance would become a thing around the time care became unaffordable. But I’ve never seen someone really try to tease out causality here. Might the two trends have been mutually self-reinforcing? The price of care rises due to some original shock. Someone invents health insurance, which seems like a good idea. But this creates a series of perverse incentives, which other actors figure out how to exploit (eg the Lucemyra example above). Insurance-based-health-care becomes less efficient, but hospitals can’t or don’t internalize this to the insured patients – they just raise the price for everyone, insurance or no. That makes even more people need health insurance, and the cycle repeats as prices grow higher and higher and insurance becomes more and more necessary. This syncs well with some explanations I’ve heard of rising college prices, where once the government made easy loans and subsidies available to everyone, prices rose until they consumed all the resources available.

I have no idea if this is true or not. If it is, the Amish succeed partly by successfully forcing providers to internalize the costs of insurance to insurance patients. Sometimes they do this by literally asking hospitals for better prices because they are not insured (eg the “honest customer” example above). Other times they flee the country entirely to reach a medical system that doesn’t deal with insured patients (eg Tijuana). This seems to work well for them. But their reliance on church alms and Amish Hospital Aid suggests that their care is still more expensive and burdensome for them than past generations’ care was for them. They’ve just learned ways to manage the expense successfully.

# Employer Provided Health Insurance Delenda Est

My last post didn’t really go into why I dislike the way we do health insurance so much.

Of course, there are the usual criticisms based on compassion and efficiency. Compassion because poor people can’t get access to life-saving medical care. Efficiency because it’s ruinously expensive compared to every other system around. I agree with these arguments. And they’re strong enough that asking whether there are any other reasons is kind of like the proverbial “But besides that, Mrs. Lincoln, how did you like the play?”

But I had already internalized the compassion and efficiency critiques before becoming a doctor. After starting work, I encountered new problems I never would have expected, ones which have yet to fade into the amorphous cloud of injustices we all know about and mostly ignore. Most of my patients have insurance; most of them are well-off; most of them are intelligent enough that they should be able to navigate the bureaucracy. Listen to the usual debate around insurance, and you would expect them to be the winners of our system; the rich people who can turn their financial advantage into better care. And yet barely a day goes by without a reminder that it doesn’t work this way.

Here are some people I have encountered – some of them patients, some of them friends – who have made me skeptical that our system works for anyone at all:

— The elderly man who had a great relationship with his last psychiatrist, who saw him for twenty years, and who knew every detail of his issues. He switched jobs, got a new insurance, the old psychiatrist was no longer in network, and so he had to see me instead. I know nothing about him and it will take several evaluation sessions before I can even consistently remember who he is and what he needs from me.

— The businesswoman who was seeing me and doing well until the HR person at her job told her that she didn’t need to submit any forms to renew her insurance that year. That turned out to be wrong, and she missed The One Month Of The Year When You Are Allowed To Renew Insurance. She lost her insurance and can’t afford to keep seeing me.

— The bipolar man on a very important daily medication. He changed insurance plans. The new insurance refused to pay for his drugs until they got a form explaining why he needed the medication. I sent in the form. They said they couldn’t find the patient in their system and so couldn’t process the form. We argued about this for several days, during which he ran out of medication and decompensated.

— The would-be entrepreneur who wanted to save up enough money to live on for a year or two, quit her dead-end job, and start a startup – but who wouldn’t be able to afford health insurance outside of her dead-end job’s plan. She is still at her dead-end job.

— The young gay man with conservative religious parents. He had a supportive friend group and should have been able to come out to his parents without caring what they thought – except that he’s on his parents’ insurance and has no good alternative. He remains in the closet.

— The young woman in a not-quite-abusive but far-from-acceptable marriage, who stays in it because she’s on her husband’s insurance and has no good alternative.

— The depressed guy who was doing well on a complicated antidepressant regimen for a while, changed insurances, and was too depressed to do the work of finding a new psychiatrist. Now he comes to me saying it’s been five years, he’s been depressed all that time, and he would like to get back on the medications that he knows work well for him.

— The endless train of patients I saw when I worked in a hospital emergency room, whose stories started with “So I lost my insurance, but I figured I could get along fine without the medication…”

— The other endless train of patients I saw when I worked in a hospital emergency room, whose stories started with “So I lost my insurance, but I was smart, and I had saved up a stockpile of my medication, and I figured, how hard could it be to manage it myself without a doctor’s advice…”

— The other endless train of patients I saw when I worked in a hospital emergency room, whose stories started with “So I changed insurance, and I got a new doctor, and he said he didn’t see a good reason why I had to be on this medication…”

— The Oklahoman who wants to move to California where he has more friends and better job prospects, but he’s on Oklahoma state insurance for the needy. Although California also has a state insurance for the needy, there is no way to figure out whether it will accept him or cover the care he needs other than by moving to California, applying, and seeing what happens. Also, the application process takes weeks to months, during which time he will not have either state’s insurance (though California promises he will get reimbursed for care he gets during this period later). He decides to stay in Oklahoma.

— The well-off Oklahoman with good private insurance who visits California on a business trip, gets sick, and finds that surprise! his Oklahoman insurance doesn’t have any in-network providers in California and he has to pay $20,000 out-of-pocket for care.

— The depressed guy who was in remission for years and had a great job with great insurance. Then he had a relapse, became too depressed to go to work, got fired, and lost his insurance right at the moment when it finally could have been useful for him.

— The woman who had a minor breakdown and was brought to the hospital by police. The hospital admitted her against her will, on the grounds that her minor breakdown sounded potentially dangerous, treated her, and released her. Her insurance refused to pay, on the grounds that it was just a minor breakdown and didn’t really require hospitalization, by the standards of This Particular Insurance Company. She is on the hook for the entire cost of the involuntary hospitalization.

— Anything involving Kaiser [EDIT: I’m getting some pushback on this who say Kaiser is good at everything except mental health; I only deal with them in a mental health capacity so I can’t speak to this]

— The trauma victim who needs trauma therapy – which means reliving all your traumas in order to come to terms with them – but is reluctant to start because her work situation is unstable. She knows that if she changes insurance, she’ll have to restart the therapy from the beginning and relive all her traumas all over again.

— The young schizophrenic man who is on his parents’ insurance. Because so many schizophrenics are poor, all of the expertise in treating schizophrenia is concentrated in Medicaid clinics. But a rich kid on his parents’ excellent insurance can’t access Medicaid, he can’t go off his parents’ insurance for other reasons, and he can’t find any private psychiatrists who know how to treat his particular schizophrenia-related complications. I am sure there is some good solution to this which I am missing, but I haven’t been able to find it and his family hasn’t either.

— The anorexic woman who has Blue Cross, and the only good anorexia therapist in town only takes Aetna. The sex-addicted man who has Aetna, and the only good sex addiction therapist in town only takes Blue Cross.

Any other system would fix these problems. A public system like Medicare For All would fix them. A communal system like the Amish have would fix them. A free market system like our grandparents had would fix them. The prepaid doctor cooperatives Reason talks about would fix them. A half-assed compromise like Joe Biden’s Medicare For All Who Want It would fix them. But here we are, stuck with a system that somehow manages to fail everybody for different reasons.

# Give Yourself Gout For Fame And Profit

I.

Actually, no. You should not do this. Most of you were probably already not doing this, and I support your decision. But if you want a 2000 word essay on some reasons to consider this, and then some other reasons why those reasons are wrong, keep reading.

Gout is a disease caused by high levels of uric acid in the blood. Everyone has some uric acid in their blood, but when you get too much, it can form little crystals that get deposited around your body and cause various problems, most commonly joint pain. Some uric acid comes from chemicals found in certain foods (especially meat), so the first step for a gout patient is to change their diet. If that doesn’t work, they can take various chemicals that affect uric acid metabolism or prevent inflammation.

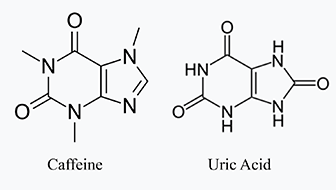
Gout is traditionally associated with kings, probably because they used to be the only people who ate enough meat to be affected. Veal, venison, duck, and beer are among the highest-risk foods; that list sounds a lot like a medieval king’s dinner menu. But as kings faded from view, gout started affecting a new class of movers and shakers. King George III had gout, but so did many of his American enemies, including Franklin, Jefferson, and Hancock (beginning a long line of gout-stricken US politicians, most recently Bernie Sanders). Lists of other famous historical gout sufferers are contradictory and sometimes based on flimsy evidence, but frequently mentioned names include Alexander the Great, Charlemagne, Leonardo da Vinci, Martin Luther, John Milton, Isaac Newton, Ludwig von Beethoven, Karl Marx, Charles Dickens, and Mark Twain.

Question: isn’t this just a list of every famous person ever? It sure seems that way, and even today gout seems to disproportionately strike the rich and powerful. In 1963, Dunn, Brooks, and Mausner published Social Class Gradient Of Serum Uric Acid Levels In Males, showing that in many different domains, the highest-ranking and most successful men had the highest uric acid (and so, presumably, the most gout). Executives have higher uric acid than blue-collar workers. College graduates have higher levels than dropouts. Good students have higher levels than bad students. Top professors have higher levels than mediocre professors. DB&M admitted rich people probably still eat more meat than poor people, but didn’t think this explained the magnitude or universality of the effect. They proposed a different theory: maybe uric acid makes you more successful.

Before we mock them, let’s take more of a look at why they might think that, and at the people who have tried to flesh out their theory over the years.

Most animals don’t have uric acid in their blood. They use an enzyme called uricase to metabolize it into a harmless chemical called allantoin. About ten million years ago, the common ancestor of apes and humans got a mutation that broke uricase, causing uric acid levels to rise. The mutation spread very quickly, suggesting that evolution really wanted primates to have lots of uric acid for some reason. Since discovering this, scientists have been trying to figure out exactly what that reason was, with most people thinking it’s probably an antioxidant or neuroprotectant or something else helpful if you’re trying to evolve giant brains. Other researchers note that in lower animals, uric acid is a “come out of hibernation” sign which seems to induce energetic foraging and goal-directed behavior more generally.

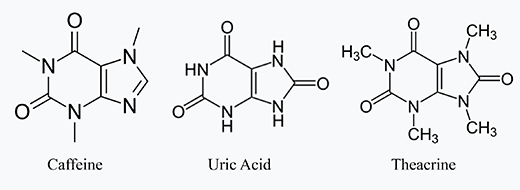
Some of these people note the similarity between uric acid and caffeine:



If uric acid had caffeine-like effects, then high levels of uric acid in the blood would be like being on a constant caffeine drip. The exact numbers don’t really work out, but you can fix this by assuming uric acid is an order of magnitude or so weaker than straight caffeine. Add this fudge factor, and Benjamin Franklin was on exactly one espresso all the time.

But you can’t actually be hyperproductive by being on one espresso all the time, can you? Don’t you eventually gain tolerance to caffeine and lose any benefits?

Although uric acid is structurally similar to caffeine, it’s even more similar to a chemical called theacrine. In fact, theacrine is just 1,3,7,9-tetramethyl-uric acid:



Theacrine (not the same as theanine, be careful with this one!) is a caffeine-like substance found in an unusual Chinese variety of tea plant. It’s recently gained fame in the nootropic community for not producing tolerance the same way regular caffeine does – see eg Theacrine: Caffeine-Like Alkaloid Without Tolerance Build-Up. This makes the theory work even better: Franklin (and other gout sufferers) were constantly on one espresso worth of magic no-tolerance caffeine. Seems plausible!

II.

This theory is hilarious, but is it true?

I was able to find eleven studies comparing achievement and uric acid levels. I’ve put them into a table below.

Nine out of eleven are positive. But I find it hard to be confident in any of them. Modern studies can be pretty bad, but studies from the 1960s ask you to take even more things on trust, while inspiring a lot less of it. Many of these studies were unable to find the outcomes that the others found, but discovered new outcomes of their own. Many failed to report basic pieces of information. The largest experiments usually found the least impressive results. Overall this looks a lot like you would expect from something forty years before anyone realized there was a replication crisis.

I also notice that the most positive studies compare business executives to people in other walks of life, and the least positive studies compare good students with bad students. Business executives get a lot of chances to differ from the general population – maybe they still eat more meat and richer food? Maybe they’re stressed and stress affects uric acid levels?

What about the list of very famous people with gout? I agree it’s a lot of people, but what’s the base rate? Kings were born to their position, so we have no reason to think they were especially high achievers (someone in their family might have been, but that gene could have gotten pretty diluted). Since so many kings got gout, this suggests rich old people in the past had gout pretty often regardless of achievement. Also, this was before people invented good medical diagnosis, so probably arthritis, injuries, and any other form of joint pain got rounded off to gout too. What percent of rich old people in the past had some kind of joint pain? I’m prepared to guess “a lot”.

The biochemists report equally confusing results around the uric acid / caffeine connection. Caffeine mostly works by antagonizing adenosine, a chemical involved in sleepiness. According to Hunter et al, Effects of uric acid and caffeine on A1 adenosine receptor binding in developing rat brain, uric acid does not affect adenosine, and so probably does not have a caffeine-like mechanism of action. On the other hand, caffeine probably has a small additional effect on catecholamine (eg dopamine, norepinephrine) release, and a different paper finds that uric acid does share this mechanism. So it doesn’t have caffeine’s main effect, but it does seem to have some kind of mild stimulant properties.

Given this level of uncertainty around every step in the hypothesis, I would describe any link between uric acid and achievement as kind of a stretch at this point. I feel bad about this, because it’s an elegant theory with mostly positive studies in support, but I’m just not feeling like it’s met its burden of proof.

III.

But some recent research is trying to bring this field back from the dead. At least this is what I get from Ortiz et al, Purinergic System Dysfunction In Mood Disorders, which synthesizes some more modern evidence that “uric acid and purines (such as adenosine) regulate mood, sleep, activity, appetite, cognition, memory, convulsive threshold, social interaction, drive, and impulsivity”. It argues that we know there are neurorecptors for adenosine (another similar-looking molecule) and ATP (adenosine triphosphate, the body’s main form of chemical energy). These seem to be involved in depression and mania, in the predicted direction (manic people have too much ATP, depressed people have too little, and treatments for both conditions seem to normalize ATP levels). These results seem to be daring someone to make up a theory where mania is just too much chemical energy floating around, but if Ortiz et al are doing that, it’s sandwiched in between so many dense paragraphs on receptor binding that I can’t make it out.

More interesting for us, uric acid is related to all these chemicals and also seems to be involved in mania. See eg de Berardis et al, Evaluation of plasma antioxidant levels during different phases of illness in adult patients with bipolar disorder, which finds that uric acid is elevated in manic patients, and the more manic, the higher the uric acid levels. And Machado-Vieria claims to have gotten pretty good results treating bipolar mania with allopurinol, a gout medication that decreases uric acid – and the more the allopurinol decreased uric acid, the better the results. There’s also a little evidence that depressed people have lower uric acid than normal. None of this is a large effect – there are still a lot of depressed people with higher-than-normal uric acid and a lot of manic people with lower – but it’s around the same size as all the other infuriatingly suggestive effects we find in psychiatry that never lead to overarching theories or go anywhere useful.

Future studies should try to replicate the link between uric acid and mania, and come up with a better understanding of why it might be true – maybe since uric acid is a decay product of ATP, the body interprets it as a sign that energy is plentiful? They should try to explain away anomalies – if gout is maniogenic, how come so many people with gout are depressed? Is it just because having a painful illness is inherently depressing? And then it should investigate how mania bleeds into normal personality. Is someone with slightly higher uric acid a tiny bit hypomanic all the time?

If they can fill in all those steps, I’ll be willing to take a fresh look at the old papers linking gout and achievement. Until then, you should probably hold off on eating megadoses of venison to become the next Ben Franklin.

# Predictions For 2020

At the beginning of every year, I make predictions. At the end of every year, I score them. So here are a hundred more for 2020.

Rules: all predictions are about what will be true on January 1, 2021. Some predictions about my personal life, or that refer to the personal lives of other people, have been redacted to protect their privacy. I’m using the full 0 – 100 range in making predictions this year, but they’ll be flipped and judged as 50 – 100 in the rating stage, just like in previous years. I’ve tried to avoid doing specific research or looking at prediction markets when I made these, though some of them I already knew what the markets said.

Feel free to get in a big fight over whether 50% predictions are meaningful.

CORONAVIRUS:  
1. Bay Area lockdown (eg restaurants closed) will be extended beyond June 15: 60%  
2. …until Election Day: 10%  
3. Fewer than 100,000 US coronavirus deaths: 10%  
4. Fewer than 300,000 US coronavirus deaths: 50%  
5. Fewer than 3 million US coronavirus deaths: 90%  
6. US has highest official death toll of any country: 80%  
7. US has highest death toll as per expert guesses of real numbers: 70%  
8. NYC widely considered worst-hit US city: 90%  
9. China’s (official) case number goes from its current 82,000 to 100,000 by the end of the year: 70%  
10. A coronavirus vaccine has been approved for general use and given to at least 10,000 people somewhere in the First World: 50%  
11. Best scientific consensus ends up being that hydroxychloroquine was significantly effective: 20%  
12. I personally will get coronavirus (as per my best guess if I had it; positive test not needed): 30%  
13. Someone I am close to (housemate or close family member) will get coronavirus: 60%  
14. General consensus is that we (April 2020 US) were overreacting: 50%  
15. General consensus is that we (April 2020 US) were underreacting: 20%  
16. General consensus is that summer made coronavirus significantly less dangerous: 70%  
17. …and there is a catastrophic (50K+ US deaths, or more major lockdowns, after at least a month without these things) second wave in autumn: 30%  
18. I personally am back to working not-at-home: 90%  
19. At least half of states send every voter a mail-in ballot in 2020 presidential election: 20%  
20. PredictIt is uncertain (less than 95% sure) who won the presidential election for more than 24 hours after Election Day. 20%

POLITICS:  
21. Democrats nominate Biden, and he remains nominee on Election Day: 90%  
22. Balance of evidence available on Election Day supports (as per my opinion) Tara Reade accusation: 90%  
23. Conditional on me asking about Reade on SSC survey, average survey-taker’s credence in her accusation is greater than 50%: 70%  
24. …greater than 75%: 10%  
25. …greater than credence in Kavanaugh accusation asked in the same format: 40%  
26. Trump is re-elected President: 50%  
27. Democrats keep the House: 70%  
28. Republicans keep the Senate: 50%  
29. Trump approval rating higher than 43% on June 1: 30%  
30. Biden polling higher than Trump on June 1: 70%  
31. At least one new Supreme Court Justice: 20%  
32. I vote Democrat for President: 80%  
33. Boris still UK PM: 90%  
34. No new state leaves EU: 90%  
35. UK, EU extend “transition” trade deal: 80%  
36. Kim Jong-Un alive and in power: 60%

ECON AND TECH:  
37. Dow is above 25,000: 70%  
38. …above 30,000: 20%  
39. Bitcoin is above $5,000: 70%  
40. …above $10,000: 20%  
41. I have bought a Surface Book 3 laptop: 60%  
42. Crew Dragon reaches orbit: 80%  
43. Starship reaches orbit: 40%

SSC, ETC:  
44. I do another Nootropics Survey this year: 70%  
45. I do another SSC Survey this year: 90%  
46. I start a Reader SSC Survey this year: 60%  
47. I start a SSC Book Review Contest this year: 70%  
48. I run another Adversarial Collaboration Contest this year: 10%  
49. I publish [redacted]: 20%  
50. I publish [redacted]: 50%  
51. I publish [redacted]: 60%  
52. I publish [redacted]: 80%  
53. …conditional on being published, it gets at least 40,000 pageviews: 10%  
54. I publish [redacted]: 60%  
55. …conditional on being published, it gets at least 40,000 pageviews: 50%  
56. More hits this year than last: 70%  
57. Most hits ever this year: 20%  
58. I finish Unsong revision this year: 40%  
59. New co-blogger with more than 3 posts: 10%

FRIENDS:  
60. No new long-term (1 month +) residents at group house by the end of the year: 70%  
61. Koios has said his first clear comprehensible word: 50%  
62. [redacted]: 40%  
63. [redacted]: 60%  
64. [redacted]: 80%  
65. [redacted]: 80%  
66. [redacted]: 95%  
67. [redacted]: 10%  
68. [redacted]: 95%  
69. [redacted]: 80%  
70. [redacted]: 80%  
71. [redacted]: 50%

PROFESSIONAL  
72. I’ve gotten at least one new patient to do a full wake therapy protocol: 60%  
73. I have specific, set-in-motion plans to quit work / start my own business: 5%  
74. I work the same schedule and locations I did before the coronavirus: 80%  
75. I get a bonus for 2020: 20%

PERSONAL:  
76. [redacted]: 70%  
77. [redacted]: 70%  
78. [redacted]: 95%  
79. I travel to Alaska this year: 60%  
80. [redacted]: 40%  
81. [redacted]: 20%  
82. I go on at least three dates with someone I haven’t met yet: 20%  
83. [redacted]: 10%  
84. [redacted]: 30%  
85. [redacted]: 10%  
86. I try one biohacking project per month x at least 5 of the last 6 months of 2020: 30%  
87. I find at least one new supplement I take or expect to take regularly x 3 months: 20%  
88. Not eating meat at home: 40%  
89. Weight below 200: 50%  
90. Weight below 190: 10%  
91. [redacted]: 90%  
92. [redacted]: 30%  
93. [redacted]: 5%  
94. I travel outside the country at least once: 10%  
95. I get back into meditating seriously (at least ten minutes a day, five days a week) for at least a month: 10%  
96. At least ten tweets in 2020: 80%  
97. I eat at/from Sliver more than any other restaurant in Q4 2020: 50%  
98. [redacted]: 30%  
99. I do pushups and situps at least 3 days/week in average week of Q4 2020: 60%  
100. I write the post scoring these predictions before 2/1/21: 70%

# Book Review Contest: Call For Entries

Your mission, should you choose to accept it, is to write a book review and send it to me at scott[at]slatestarcodex[dot]com before August 5th 2020.

Interested? Here’s the small print (written in normal-sized print, for your convenience):

Pick a book, then write a review similar to my SSC book reviews (examples). I’m mostly expecting reviews of nonfiction, but I guess you could review fiction if you really wanted and had something interesting to say beyond just “here’s the plot and I thought it was good”.

I’ll choose some number of finalists – probably around five, but maybe more or less depending on how many I get – and publish them on the blog, with full attribution, just like with the adversarial collaborations. Then readers will vote for the best, just like with the adversarial collaborations. First place will get at least $1000, second place $500, third place $250 – I might increase those numbers later on. Some winners may also get an invitation to pitch me any other pieces they have that they think would make good SSC posts. I may also release non-finalist entries somewhere else so people can read them – if you strongly object to me making your entry public, let me know.

Please send me your review in a .txt file (eg Notepad), attached to an email. I’m making this rule because otherwise you send me heavily formatted emails with lots of bold text and weird font changes and tables, and it’s really hard to post to SSC in ways that don’t mess up the formatting or look wrong. If you want formatting in your final posted review, please use a tiny amount of hand-written html – ie putting bold things in <b>bold</b> and putting links in <A HREF=”http://www.example.com”>links</A>. If you want to include images, please use <IMG SRC=”http://www.example.com/image.png”>. For quotes, <blockquote>quote</blockquote>. If you can’t figure this out, just send me the images and tell me where to put them. Don’t get Word or something to save your heavily formatted document as HTML or it will do horrible things that will screw me up when I try to make it into a blog post. If you can’t make or send .txt files for some reason, please paste your review in the body of the email in a way that follows these same principles. I don’t want this part to prevent anyone from sending something in, so if you don’t understand this and are scared of it, just send something in anyway and I’ll fix it up.

In your email to me, please also include the name you want me to attribute your review to – that could be your real name, your first name plus an initial, your pseudonym, etc.

You don’t have to register beforehand or let anyone else know you’re doing this. But if you want to avoid having someone else accidentally review the same book as you, you can post what you’re doing in the comments below and hopefully other people will avoid doubling up. I have vague plans to review Julian Jaynes’ The Origin Of Consciousness In The Breakdown Of The Bicameral Mind and Ezra Klein’s Why We’re Polarized before August, so you might want to avoid those too.

If you win, I will pay through PayPal or donations to the charity of your choice. I reserve the right to change these conditions in minor ways that don’t significantly inconvenience contest participants.

I’ll check the comments here for a few days and answer any questions you might have.

[EDIT: Please don’t submit reviews that have been posted on other blogs before. Reviews that have been posted on the r/SSC or r/themotte subreddit are provisionally okay, since I don’t want to disincentivize people from doing that, but I’ll try to come up with better guidance soon]

# Classified Thread 8

This is the…monthly? bimonthly? occasional?…classified thread. Post advertisements, personals, and any interesting success stories from the last thread.

Under the circumstances, you can also ask for financial support (eg link a GoFundMe) if you need to. Nobody has to donate to people who do this, but please at least be respectful.

# Studies On Slack

I.

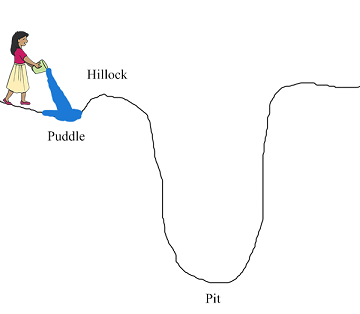
Imagine a distant planet full of eyeless animals. Evolving eyes is hard: they need to evolve Eye Part 1, then Eye Part 2, then Eye Part 3, in that order. Each of these requires a separate series of rare mutations.

Here on Earth, scientists believe each of these mutations must have had its own benefits – in the land of the blind, the man with only Eye Part 1 is king. But on this hypothetical alien planet, there is no such luck. You need all three Eye Parts or they’re useless. Worse, each Eye Part is metabolically costly; the animal needs to eat 1% more food per Eye Part it has. An animal with a full eye would be much more fit than anything else around, but an animal with only one or two Eye Parts will be at a small disadvantage.

So these animals will only evolve eyes in conditions of relatively weak evolutionary pressure. In a world of intense and perfect competition, where the fittest animal always survives to reproduce and the least fit always dies, the animal with Eye Part 1 will always die – it’s less fit than its fully-eyeless peers. The weaker the competition, and the more randomness dominates over survival-of-the-fittest, the more likely an animal with Eye Part 1 can survive and reproduce long enough to eventually produce a descendant with Eye Part 2, and so on.

There are lots of ways to decrease evolutionary pressure. Maybe natural disasters often decimate the population, dozens of generations are spent recolonizing empty land, and during this period there’s more than enough for everyone and nobody has to compete. Maybe there are frequent whalefalls, and any animal nearby has hit the evolutionary jackpot and will have thousands of descendants. Maybe the population is isolated in little islands and mountain valleys, and one gene or another can reach fixation in a population totally by chance. It doesn’t matter exactly how it happens, it matters that evolutionary pressure is low.

The branch of evolutionary science that deals with this kind of situation is called “adaptive fitness landscapes”. Landscapes really are a great metaphor – consider somewhere like this:



You pour out a bucket of water. Water “flows downhill”, so it’s tempting to say something like “water wants to be at the lowest point possible”. But that’s not quite right. The lowest point possible is the pit, and water won’t go there. It will just sit in the little puddle forever, because it would have to go up the tiny little hillock in order to get to the pit, and water can’t flow uphill. Using normal human logic, we feel tempted to say something like “Come on! The hillock is so tiny, and that pit is so deep, just make a single little exception to your ‘always flow downhill’ policy and you could do so much better for yourself!” But water stubbornly refuses to listen.

Under conditions of perfectly intense competition, evolution works the same way. We imagine a multidimensional evolutionary “landscape” where lower ground represents higher fitness. In this perfectly intense competition, organisms can go from lower to higher fitness, but never vice versa. As with water, the tiniest hillock will leave their potential forever unrealized.

Under more relaxed competition, evolution only tends probabilistically to flow downhill. Every so often, it will flow uphill; the smaller the hillock, the more likely evolution will surmount it. Given enough time, it’s guaranteed to reach the deepest pit and mostly stay there.

Take a moment to be properly amazed by this. It sounds like something out of the Tao Te Ching. An animal with eyes has very high evolutionary fitness. It will win at all its evolutionary competitions. So in order to produce the highest-fitness animal, we need to – select for fitness less hard? In order to produce an animal that wins competitions, we need to stop optimizing for winning competitions?

This doesn’t mean that less competition is always good. An evolutionary environment with no competition won’t evolve eyes either; a few individuals might randomly drift into having eyes, but they won’t catch on. In order to optimize the species as much as possible as fast as possible, you need the right balance, somewhere in the middle between total competition and total absence of competition.

In the esoteric teachings, total competition is called Moloch, and total absence of competition is called Slack. Slack (thanks to Zvi Mowshowitz for the term and concept) gets short shrift. If you think of it as “some people try to win competitions, other people don’t care about winning competitions and slack off and go to the beach”, you’re misunderstanding it. Think of slack as a paradox – the Taoist art of winning competitions by not trying too hard at them. Moloch and Slack are opposites and complements, like yin and yang. Neither is stronger than the other, but their interplay creates the ten thousand things.

II.

Before we discuss slack further, a digression on group selection.

Some people would expect this discussion to be quick, since group selection doesn’t exist. These people understand it as evolution acting for the good of a species. It’s a tempting way to think, because evolution usually eventually makes species stronger and more fit, and sometimes we colloquially round that off to evolution targeting a species’ greater good. But inevitably we find evolution is awful and does absolutely nothing of the sort.

Imagine an alien planet that gets hit with a solar flare once an eon, killing all unshielded animals. Sometimes unshielded animals spontaneously mutate to shielded, and vice versa. Shielded animals are completely immune to solar flares, but have 1% higher metabolic costs. What happens? If you predicted “magnetic shielding reaches fixation and all animals get it”, you’ve fallen into the group selection trap. The unshielded animals outcompete the shielded ones during the long inter-flare period, driving their population down to zero (though a few new shielded ones arise every generation through spontaneous mutations). When the flare comes, only the few spontaneous mutants survive. They breed a new entirely-shielded population, until a few unshielded animals arise through spontaneous mutation. The unshielded outcompete the shielded ones again, and by the time of the next solar flare, the population is 100% unshielded again and they all die. If the animals are lucky, there will always be enough spontaneously-mutated shielded animals to create a post-flare breeding population; if they are unlucky, the flare will hit at a time with unusually few such mutants, and the species will go extinct.

An Evolution Czar concerned with the good of the species would just declare that all animals should be shielded and solve the problem. In the absence of such a Czar, these animals will just keep dying in solar-flare-induced mass extinctions forever, even though there is an easy solution with only 1% metabolic cost.

A less dramatic version of the same problem happens here on Earth. Every so often predators (let’s say foxes) reproduce too quickly and outstrip the available supply of prey (let’s say rabbits). There is a brief period of starvation as foxes can’t find any more rabbits and die en masse. This usually ends with a boom-bust cycle: after most foxes die, the rabbits (who reproduce very quickly and are now free of predation) have a population boom; now there are rabbits everywhere. Eventually the foxes catch up, eat all the new rabbits, and the cycle repeats again. It’s a waste of resources for foxkind to spend so much of time and its energy breeding a huge population of foxes that will inevitably collapse a generation later; an Evolution Czar concerned with the common good would have foxes limit their breeding at a sustainable level. But since individual foxes that breed excessively are more likely to have their genes represented in the next generation than foxes that breed at a sustainable level, we end up with foxes that breed excessively, and the cycle continues.

(but humans are too smart to fall for this one, right?)

Some scientists tried to create group selection under laboratory conditions. They divided some insects into subpopulations, then killed off any subpopulation whose numbers got too high, and “promoted” any subpopulation that kept its numbers low to better conditions. They hoped the insects would evolve to naturally limit their family size in order to keep their subpopulation alive. Instead, the insects became cannibals: they ate other insects’ children so they could have more of their own without the total population going up. In retrospect, this makes perfect sense; an insect with the behavioral program “have many children, and also kill other insects’ children” will have its genes better represented in the next generation than an insect with the program “have few children”.

But sometimes evolution appears to solve group selection problems. What about multicellular life? Stick some cells together in a resource-plentiful environment, and they’ll naturally do the evolutionary competition thing of eating resources as quickly as possible to churn out as many copies of themselves as possible. If you were expecting these cells to form a unitary organism where individual cells do things like become heart cells and just stay in place beating rhythmically, you would call the expected normal behavior “cancer” and be against it. Your opposition would be on firm group selectionist grounds: if any cell becomes cancer, it and its descendants will eventually overwhelm everything, and the organism (including all cells within it, including the cancer cells) will die. So for the good of the group, none of the cells should become cancerous.

The first step in evolution’s solution is giving all cells the same genome; this mostly eliminates the need to compete to give their genes to the next generation. But this solution isn’t perfect; cells can get mutations in the normal course of dividing and doing bodily functions. So it employs a host of other tricks: genetic programs telling cells to self-destruct if they get too cancer-adjacent, an immune system that hunts down and destroys cancer cells, or growing old and dying (this last one isn’t usually thought of as a “trick”, but it absolutely is: if you arrange for a cell line to lose a little information during each mitosis, so that it degrades to the point of gobbledygook after X divisions, this means cancer cells that divide constantly will die very quickly, but normal cells dividing on an approved schedules will last for decades).

Why can evolution “develop tricks” to prevent cancer, but not to prevent foxes from overbreeding, or aliens from losing their solar flare shields? Group selection works when the group itself has a shared genetic code (or other analogous ruleset) that can evolve. It doesn’t work if you expect it to directly change the genetic code of each individual to cooperate more.

When we think of cancer, we are at risk of conflating two genetic codes: the shared genetic code of the multicellular organism, and the genetic code of each cell within the organism. Usually (when there are no mutations in cell divisions) these are the same. Once individual cells within the organism start mutating, they become different. Evolution will select for cancer in changes to individual cells’ genomes over an organism’s lifetime, but select against it in changes to the overarching genome over the lifetime of the species (ie you should expect all the genes you inherited from your parents to be selected against cancer, and all the mutations in individual cells you’ve gotten since then to be selected for cancer).

The fox population has no equivalent of the overarching genome; there is no set of rules that govern the behavior of every fox. So foxes can’t undergo group selection to prevent overpopulation (there are some more complicated dynamics that might still be able to rescue the foxes in some situations, but they’re not relevant to the simple model we’re looking at).

In other words, group selection can happen in a two-layer hierarchy of nested evolutionary systems when the outer system (eg multicellular humans) includes rules that the inner system (eg human cells) have to follow, and where the fitness of the evolving-entities in the outer system depends on some characteristics of the evolving-entities in the inner system (eg humans are higher-fitness if their cells do not become cancerous). The evolution of the outer layer includes evolution over rulesets, and eventually evolves good strong rulesets that tell the inner-layer evolving entities how to behave, which can include group selection (eg humans evolve a genetic code that includes a rule “individual cells inside of me should not get cancer” and mechanisms for enforcing this rule).

You can find these kinds of two-layer evolutionary systems everywhere. For example, “cultural evolution” is a two-layer evolutionary system. In the hypothetical state of nature, there’s unrestricted competition – people steal from and murder each other, and only the strongest survive. After they form groups, the groups compete with each other, and groups that develop rulesets that prevent theft and murder (eg legal codes, religions, mores) tend to win those competitions. Once again, the outer layer (competition between cultures) evolves groups that successfully constrains the inner layer (competition between individuals). Species don’t have a czar who restrains internal competition in the interest of keeping the group strong, but some human cultures do (eg Russia).

Or what about market economics? The outer layer is companies, the inner layer is individuals. Maybe the individuals are workers – each worker would selfishly be best off if they spent the day watching YouTube videos and pushed the hard work onto someone else. Or maybe they’re executives – each individual executive would selfishly be best off if they spent their energy on office politics, trying to flatter and network with whoever was most likely to promote them. But if all the employees loaf off and all the executives focus on office politics, the company won’t make products, and competitors will eat their lunch. So someone – maybe the founder/CEO – comes up with a ruleset to incentivize good work, probably some kind of performance review system where people who do good work get promoted and people who do bad work get fired. The outer-layer competition between companies will select for corporations with the best rulesets; over time, companies’ internal politics should get better at promoting the kind of cooperation necessary to succeed.

How do these systems replicate multicellular life’s success without being literal entities with literal DNA having literal sex? They all involve a shared ruleset and a way of punishing rulebreakers which make it in each individual’s short-term interest to follow the ruleset that leads to long-term success. Countries can do that (follow the law or we’ll jail you), companies can do that (follow our policies or we’ll fire you), even multicellular life can sort of do that (don’t become cancer, or immune cells will kill you). When there’s nothing like that (like the overly-fast-breeding foxes) evolution fails at group selection problems. When there is something like that, it has a chance. When there’s something like that, and the thing like that is itself evolving (either because it’s encoded in literal DNA, or because it’s encoded in things like company policies that determine whether a company goes out of business or becomes a model for others), then it can reach a point where it solves group selection problems very effectively.

In the esoteric teachings, the inner layer of two-layer evolutionary systems is represented by the Goddess of Cancer, and outer layer by the Goddess of Everything Else. In each part of the poem, the Goddess of Cancer orders the evolving-entities to compete, but the Goddess of Everything Else recasts it as a two-layer competition where cooperation on the internal layer helps win the competition on the external layer. He who has ears to hear, let him listen.

III.

Why the digression? Because slack is a group selection problem. A species that gave itself slack in its evolutionary competition would do better than one that didn’t – for example, the eyeless aliens would evolve eyes and get a big fitness boost. But no individual can unilaterally choose to compete less intensely; if it did, it would be outcompeted and die. So one-layer evolution will fail at this problem the same way it fails all group selection problems, but two-layer systems will have a chance to escape the trap.

The multicellular life example above is a special case where you want 100% coordination and 0% competition. I framed the other examples the same way – countries do best when their citizens avoid all competition and work together for the common good, companies do best when their executives avoid self-aggrandizing office politics and focus on product quality. But as we saw above, some systems do best somewhere in the middle, where there’s some competition but also some slack.

For example, consider a researcher facing their own version of the eyeless aliens’ dilemma. They can keep going with business as normal – publishing trendy but ultimately useless papers that nobody will remember in ten years. Or they can work on Research Program Part 1, which might lead to Research Program Part 2, which might lead to Research Program Part 3, which might lead to a ground-breaking insight. If their jobs are up for review every year, and a year from now the business-as-normal researcher will have five trendy papers, and the groundbreaking-insight researcher will be halfway through Research Program Part 1, then the business-as-normal researcher will outcompete the groundbreaking-insight researcher; as the saying goes, “publish or perish”. Without slack, no researcher can unilaterally escape the system; their best option will always be to continue business as usual.

But group selection makes the situation less hopeless. Universities have long time-horizons and good incentives; they want to get famous for producing excellent research. Universities have rulesets that bind their individual researchers, for example “after a while good researchers get tenure”. And since universities compete with each other, each is incentivized to come up with the ruleset that maximizes long-term researcher productivity. So if tenure really does work better than constant vicious competition, then (absent the usual culprits like resistance-to-change, weird signaling equilibria, politics, etc) we should expect universities to converge on a tenure system in order to produce the best work. In fact, we should expect universities to evolve a really impressive ruleset for optimizing researcher incentives, just as impressive as the clever mechanisms the human body uses to prevent cancer (since this seems a bit optimistic, I assume the usual culprits are not absent).

The same is true for grant-writing; naively you would want some competition to make sure that only the best grant proposals get funded, but too much competition seems to stifle original research, so much so that some funders are throwing out the whole process and selecting grants by lottery, and others are running grants you can apply for in a half-hour and hear back about two days later. If there’s a feedback mechanism – if these different rulesets produce different-quality research, and grant programs that produce higher-quality research are more likely to get funded in the future – then the rulesets for grants will gradually evolve, and the competition for grants will take place in an environment with whatever the right evolutionary parameters for evolving good research are.

I don’t want to say these things will definitely happen – you can read Inadequate Equilibria for an idea of why not. But they might. The evolutionary dynamics which would normally prevent them can be overcome. Two-layer evolutionary systems can produce their own slack, if having slack would be a good idea.

IV.

That was a lot of paragraphs, and a lot of them started with “imagine a hypothetical situation where…”. Let’s look deeper into cases where an understanding of slack can inform how we think about real-world phenomena. Seven examples:

1. Monopolies. Not the kind that survive off overregulation and patents, the kind that survive by being big enough to crush competitors. These are predators that exploit low-slack environments. If Boeing has a monopoly on building passenger planes, and is exploiting that by making shoddy products and overcharging consumers, then that means anyone else who built a giant airplane factory could make better products at a lower price, capture the whole airplane market, and become a zillionaire. Why don’t they? Slack. In terms of those adaptive fitness landscapes, in between your current position (average Joe) and a much better position at the bottom of a deep pit (you own a giant airplane factory and are a zillionaire), there’s a very big hill you have to climb – the part where you build Giant Airplane Factory Part 1, Giant Airplane Factory Part 2, etc. At each point in this hill, you are worse off than somebody who was not building an as-yet-unprofitable giant airplane factory. If you have infinite slack (maybe you are Jeff Bezos, have unlimited money, and will never go bankrupt no matter how much time and cost it takes before you start earning profits) you’re fine. If you have more limited slack, your slack will run out and you’ll be outcompeted before you make it to the greater-fitness deep pit.

Real monopolies are more complicated than this, because Boeing can shape up and cut prices when you’re halfway to building your giant airplane factory, thus removing your incentive. Or they can do actually shady stuff. But none of this would matter if you already had your giant airplane factory fully built and ready to go – at worst, you and Boeing would then be in a fair fight. Everything Boeing does to try to prevent you from building that factory is exploiting your slacklessness and trying to increase the height of that hill you have to climb before the really deep pit.

(Peter Thiel inverts the landscape metaphor and calls the hill a “moat”, but he’s getting at the same concept).

2. Tariffs. Same story. Here’s the way I understand the history of the international auto industry – anyone who knows more can correct me if I’m wrong. Automobiles were invented in the early 20th century. Several Western countries developed homegrown auto industries more or less simultaneously, with the most impressive being Henry Ford’s work on mass production in the US. Post-WWII Japan realized that its own auto industry would never be able to compete with more established Western companies, so it placed high tariffs on foreign cars, giving local companies like Nissan and Toyota a chance to get their act together. These companies, especially Toyota, invented a new form of auto production which was actually much more efficient than the usual American methods, and were eventually able to hold their own. They started exporting cars to the US; although American tariffs put them at a disadvantage, they were so much better than the American cars of the time that consumers preferred them anyway. After decades of losing out, the American companies adopted a more Japanese ethos, and were eventually able to compete on a level playing field again.

This is a story of things gone surprisingly right – Americans and Japanese alike were able to get excellent inexpensive cars. Two things had to happen for it to work. First, Japan had to have high enough tariffs to give their companies some slack – to let them develop their own homegrown methods from scratch without being immediately outcompeted by temporarily-superior American competitors. Second, America had to have low enough tariffs that eventually-superior Japanese companies could outcompete American automakers, and Japan’s fitness-improving innovations could spread.

From the perspective of a Toyota manager, this is analogous to the eyeless alien story. You start with some good-enough standard (blind animals, American car companies). You want to evolve a superior end product (eye-having animals, Toyota). The intermediate steps (an animal with only Eye Part 1, a kind of crappy car company that stumbles over itself trying out new things) are less fit than the good-enough standard. Only when the inferior intermediate steps are protected from competition (through evolutionary randomness, through tariffs) can the superior end product come into existence. But you want to keep enough competition that the superior end product can use its superiority to spread (there is enough evolutionary competition that having eyes reaches fixation, there is enough free trade that Americans preferentially buy Toyota and US car companies have to adopt its policies).

From the perspective of an economic historian, maybe it’s a group selection story. The various stakeholders in the US auto industry – Ford, GM, suppliers, the government, labor, customers – competed with each other in a certain way and struck some compromise. The various stakeholders in the Japanese auto industry did the same. For some reason the American compromise worked worse than the Japanese one – I’ve heard stories about how US companies were more willing to defraud consumers for short-term profit, how US labor unions were more willing to demand concessions even at the cost of company efficiency, how regulators and executives were in bed with each other to the detriment of the product, etc. Every US interest group was acting in its own short-term self-interest, but the Japanese industry-as-a-whole outcompeted the American one and the Americans had to adjust.

3. Monopolies, Part II. Traditionally, monopolies have been among the most successful R&D centers. The most famous example is Xerox; it had a monopoly on photocopiers for a few decades before losing an anti-trust suit in the late 1970s; during that period, its PARC R&D program invented “laser printing, Ethernet, the modern personal computer, graphical user interface (GUI) and desktop paradigm, object-oriented programming, [and] the mouse”. The second most famous example is Bell Labs, which invented “radio astronomy, the transistor, the laser, the photovoltaic cell, the charge-coupled device, information theory, the Unix operating system, and the programming languages B, C, C++, and S” before the government broke up its parent company AT&T. Google seems to be trying something similar, though it’s too soon to judge their outcomes.

These successes make sense. Research and development is a long-term gamble. Devoting more money to R&D decreases your near-term profits, but (hopefully) increases your future profits. Freed from competition, monopolies have limitless slack, and can afford to invest in projects that won’t pay off for ten or twenty years. This is part of Peter Thiel’s defense of monopolies in Zero To One.

An administrator tasked with advancing technology might be tempted to encourage monopolies in order to get more research done. But monopolies can also be stagnant and resistant to change; it’s probably not a coincidence that Xerox wasn’t the first company to bring the personal computer to market, and ended up irrelevant to the computing revolution. Like the eyeless aliens, who will not evolve in conditions of perfect competition or perfect lack of competition, probably all you can do here is strike a balance. Some Communist countries tried the extreme solution – one state-supported monopoly per industry – and it failed the test of group selection. I don’t know enough to have an opinion on whether countries with strong antitrust eventually outcompete those with weaker antitrust or vice versa.

4. Strategy Games. I like the strategy game Civilization, where you play as a group of primitives setting out to found a empire. You build cities and infrastructure, research technologies, and fight wars. Your world is filled with several (usually 2 to 7) other civilizations trying to do the same.

Just like in the real world, civilizations must decide between Guns and Butter. The Civ version of Guns is called the Axe Rush. You immediately devote all your research to discovering how to make really good axes, all your industry to manufacturing those axes, and all your population into wielding those axes. Then you go and hack everyone else to pieces while they’re still futzing about trying to invent pottery or something.

The Civ version of Butter is called Build. You devote all your research, industry, and populace to laying the foundations of a balanced economy and culture. You invent pottery and weaving and stuff like that. Soon you have a thriving trade network and a strong philosophical tradition. Eventually you can field larger and more advanced armies than your neighbors, and leverage the advantage into even more prosperity, or into military conquest.

Consider a very simple scenario: a map of Eurasia with two civilizations, Rome and China.

If both choose Axe Rush, then whoever Axe Rushes better wins.

If both choose Build, then whoever Builds better wins.

What if Rome chooses Axe Rush, and China chooses Build?

Then it depends on their distance! If it’s a very small map and they start very close together, Rome will probably overwhelm the Chinese before Build starts paying off. But if it’s a very big map, by the time Roman Axemen trek all the way to China, China will have Built high walls, discovered longbows and other defensive technologies, and generally become too strong for axes to defeat. Then they can crush the Romans – who are still just axe-wielding primitives – at their leisure.

Consider a more complicated scenario. You have a map of Earth. The Old World contains Rome and China. The New World contains Aztecs. Rome and China are very close to each other. Now what happens?

Rome and China spend the Stone, Bronze, and Iron Ages hacking each other to bits. Aztecs spend those Ages building cities, researching technologies, and building unique Wonders of the World that provide powerful bonuses. In 1492, they discover Galleons and start crossing the ocean. The powerful and advanced Aztec empire crushes the exhausted axe-wielding Romans and Chinese.

This is another story about slack. The Aztecs had it – they were under no competitive pressure to do things that paid off next turn. The Romans and Chinese didn’t – they had to be at the top of their game every single turn, or their neighbor would conquer them. If there was an option that made you 10% weaker next turn in exchange for making you 100% stronger ten turns down the line, the Aztecs could take it without a second thought; the Romans and Chinese would probably have to pass.

Okay, more complicated Civilization scenario. This time there are two Old World civs, Rome and China, and two New World civs, Aztecs and Inca. The map is stretched a little bit so that all four civilizations have the same amount of natural territory. All four players understand the map layout and can communicate with each other. What happens?

Now it’s a group selection problem. A skillful Rome player will private message the China player and explain all of this to her. She’ll remind him that if one hemisphere spends the whole Stone Age fighting, and the other spends it building, the builders will win. She might tell him that she knows the Aztec and Inca players, they’re smart, and they’re going to be discussing the same considerations. So it would benefit both Rome and China to sign a peace treaty dividing the Old World in two, stick to their own side, and Build. If both sides cooperate, they’ll both Build strong empires capable of matching the New World players. If one side cooperates and the other defects, it will easily steamroll over its unprepared opponent and conquer the whole Old World. If both sides defect, they’ll hack each other to death with axes and be easy prey for the New Worlders.

This might be true in Civilization games, but real-world civilizations are more complicated. Orson Welles said:

In Italy, for thirty years under the Borgias, they had warfare, terror, murder and bloodshed, but they produced Michelangelo, Leonardo da Vinci and the Renaissance. In Switzerland, they had brotherly love, they had five hundred years of democracy and peace – and what did that produce? The cuckoo clock.

So maybe a little bit of internal conflict is good, to keep you honest. Too much conflict, and you tear yourselves apart and are easy prey for outsiders. Too little conflict, and you invent the cuckoo clock and nothing else. The continent that conquers the world will have enough pressure that its people want to innovate, and enough slack that they’re able to.

This is total ungrounded amateur historical speculation, but when I hear that I think of the Classical world. We can imagine it as divided into a certain number of “theaters of civilization” – Greece, Mesopotamia, Egypt, Persia, India, Scythia, etc. Each theater had its own rules governing average state size, the rules of engagement between states, how often bigger states conquered smaller states, how often ideas spread between states of the same size, etc. Some of those theaters were intensely competitive: Egypt was a nice straight line, very suited to centralized rule. Others had more slack: it was really hard to take over all of Greece; even the Spartans didn’t manage. Each theater conducted its own “evolution” in its own way – Egypt was ruled by a single Pharaoh without much competition, Scythia was constant warfare of all against all, Greece was isolated city-states that fought each other sometimes but also had enough slack to develop philosophy and science. Each of those systems did their own thing for a while, until finally one of them produced something perfect: 4th century BC Macedonia. Then it went out and conquered everything.

If Welles is right, the point isn’t to find the ruleset that promotes 100% cooperation. It’s to find the ruleset that promotes an evolutionary system that makes your group the strongest. Usually this involves some amount of competition – in order to select for stronger organisms – but also some amount of slack – to let organisms develop complicated strategies that can make them stronger. Despite the earlier description, this isn’t necessarily a slider between 0% competition and 100% competition. It could be much more complicated – maybe alternating high-slack vs. low-slack periods, or many semi-isolated populations with a small chance of interaction each generation, or alternation between periods of isolation and periods of churning.

In a full two-layer evolution, you would let the systems evolve until they reached the best parameters. Here we can’t do that – Greece has however many mountains it has; its success does not cause the rest of the world to grow more mountains. Still, we randomly started with enough different groups that we got to learn something interesting.

(I can’t emphasize enough how ungrounded this historical speculation is. Please don’t try to evolve Alexander the Great in your basement and then get angry at me when it doesn’t work)

5. The Long-Term Stock Exchange. Actually, all stock exchanges are about slack. Imagine you are a brilliant inventor who, given $10 million and ten years, could invent fusion power. But in fact you have $10 and need work tomorrow or you will starve. Given those constraints, maybe you could start, I don’t know, a lemonade stand.

You’re in the same position as the animal trying to evolve an eye – you could create something very high-utility, if only you had enough slack to make it happen. But by default, the inventor working on fusion power starves to death tomorrow (or at least makes less money than his counterpart who ran the lemonade stand), the same way the animal who evolves Eye Part 1 gets outcompeted by other animals who didn’t and dies out.

You need slack. In the evolution example, animals usually stumble across slack randomly. You too might stumble across slack randomly – maybe it so happens that you are independently wealthy, or won the lottery, or something.

More likely, you use the investment system. You ask rich people to give you $10 million for ten years so you can invent fusion; once you do, you’ll make trillions of dollars and share some of it with them.

This is a great system. There’s no evolutionary equivalent. An animal can’t pitch Darwin on its three-step plan to evolve eyes and get free food and mating opportunities to make it happen. Wall Street is a giant multi-trillion dollar time machine funneling future profits back into the past, and that gives people the slack they need to make the future profits happen at all.

But the Long-Term Stock Exchange is especially about slack. They are a new exchange (approved by the SEC last year) which has complicated rules about who can list with them. Investors will get extra clout by agreeing to hold stocks for a long time; executives will get incentivized to do well in the far future instead of at the next quarterly earnings report. It’s making a deliberate choice to give companies more slack than the regular system and see what they do with it. I don’t know enough about investing to have an opinion, except that I appreciate the experiment. Presumably its companies will do better/worse than companies on the regular stock exchange, that will cause companies to flock toward/away from it, and we’ll learn that its new ruleset is better/worse at evolving good companies through competition than the regular stock exchange’s ruleset.

6. That Time Ayn Rand Destroyed Sears. Or at least that’s how Michael Rozworski and Leigh Phillips describe Eddie Lampert’s corporate reorganization in How Ayn Rand Destroyed Sears, which I recommend. Lampert was a Sears CEO who figured – since free-market competitive economies outcompete top-down economies, shouldn’t free-market competitive companies outcompete top-down companies? He reorganized Sears as a set of competing departments that traded with each other on normal free-market principles; if the Product Department wanted its products marketed, it would have to pay the Marketing Department. This worked really badly, and was one of the main contributors to Sears’ implosion.

I don’t have a great understanding of exactly why Lampert’s Sears lost to other companies even though capitalist economies beat socialist ones; Rozworski and Phillips’ People’s Republic Of Wal-Mart, which looks into this question, is somewhere on my reading list. But even without complete understanding, we can use group selection to evolve the right parameters. Imagine an economy with several businesses. One is a straw-man communist collective, where every worker gets paid the same regardless of output and there are no promotions (0% competition, 100% cooperation). Another is Lampert’s Sears (100% competition, 0% cooperation). Others are normal businesses, where employees mostly work together for the good of the company but also compete for promotions (X% competition, Y% cooperation). Presumably the normal business outcompetes both Lampert and the commies, and we sigh with relief and continue having normal businesses. And if some of the normal businesses outcompete others, we’ve learned something about the best values of X and Y.

7. Ideas. These are in constant evolutionary competition – this is the insight behind memetics. The memetic equivalent of slack is inferential range, aka “willingness to entertain and explore ideas before deciding that they are wrong”.

Inferential distance is the number of steps it takes to make someone understand and accept a certain idea. Sometimes inferential distances can be very far apart. Imagine trying to convince a 12th century monk that there was no historical Exodus from Egypt. You’re in the middle of going over archaeological evidence when he objects that the Bible says there was. You respond that the Bible is false and there’s no God. He says that doesn’t make sense, how would life have originated? You say it evolved from single-celled organisms. He asks how evolution, which seems to be a change in animals’ accidents, could ever affect their essences and change them into an entirely new species. You say that the whole scholastic worldview is wrong, there’s no such thing as accidents and essences, it’s just atoms and empty space. He asks how you ground morality if not in a striving to approximate the ideal embodied by your essence, you say…well, it doesn’t matter what you say, because you were trying to convince him that some very specific people didn’t leave Egypt one time, and now you’ve got to ground morality.

Another way of thinking about this is that there are two self-consistent equilibria. There’s your equilibrium, (no Exodus, atheism, evolution, atomism, moral nonrealism), and the monk’s equilibrium (yes Exodus, theism, creationism, scholasticism, teleology), and before you can make the monk budge on any of those points, you have to convince him of all of them.

So the question becomes – how much patience does this monk have? If you tell him there’s no God, does he say “I look forward to the several years of careful study of your scientific and philosophical theories that it will take for that statement not to seem obviously wrong and contradicted by every other feature of the world”? Or does he say “KILL THE UNBELIEVER”? This is inferential range.

Aristotle supposedly said that the mark of an educated man is to be able to entertain an idea without accepting it. Inferential range explains why. The monk certainly shouldn’t immediately accept your claim, when he has countless pieces of evidence for the existence of God, from the spectacular faith healings he has witnessed (“look, there’s this thing called psychosomatic illness, and it’s really susceptible to this other thing called the placebo effect…”) to Constantine’s victory at the Mulvian Bridge despite being heavily outnumbered (“look, I’m not a classical scholar, but some people are just really good generals and get lucky, and sometimes it happens the day after they have weird dreams, I think there’s enough good evidence the other way that this is not the sort of thing you should center your worldview around”). But if he’s willing to entertain your claim long enough to hear your arguments one by one, eventually he can reach the same self-consistent equilibrium you’re at and judge for himself.

Nowadays we don’t burn people at the stake. But we do make fun of them, or flame them, or block them, or wander off, or otherwise not listen with an open mind to ideas that strike us at first as stupid. This is another case where we have to balance competition vs. slack. With perfect competition, the monk instantly rejects our “no Exodus” idea as less true (less memetically fit) than its competitors, and it has no chance to grow on him. With zero competition, the monk doesn’t believe anything at all, or spends hours patiently listening to someone explain their world-is-flat theory. Good epistemics require a balance between being willing to choose better ideas over worse ones, and open-mindedly hearing the worse ones out in case they grow on you.

(Thomas Kuhn points out that early versions of the heliocentric model were much worse than the geocentric model, that astronomers only kept working on them out of a sort of weird curiosity, and that it took decades before they could clearly hold their own against geocentrism in a debate).

Different people strike a different balance in this space, and those different people succeed or fail based on their own epistemic ruleset. Someone who’s completely closed-minded and dogmatic probably won’t succeed in business, or science, or the military, or any other career (except maybe politics). But someone who’s so pathologically open-minded that they listen to everything and refuse to prioritize what is or isn’t worth their time will also fail. We take notice of who succeeds or fails and change our behavior accordingly.

Maybe there’s even a third layer of selection; maybe different communities are more or less willing to tolerate open-minded vs. close-minded people. The Slate Star Codex community has really different epistemic norms from the Catholic Church or Infowars listeners; these are evolutionary parameters that determine which ideas are more memetically fit. If our epistemics make us more likely to converge on useful (not necessarily true!) ideas, we will succeed and our epistemic norms will catch on. Francis Bacon was just some guy with really good epistemic norms, and now everybody who wants to be taken seriously has to use his norms instead of whatever they were doing before. Come up with the right evolutionary parameters, and that could be you!

# “My Immortal” As Alchemical Allegory

I.

From Vox: Solving The Mystery Of The Internet’s Most Beloved And Notorious Fanfic. The fanfic is “My Immortal”, a Harry Potter story so famous that it has its own Wikipedia page, and articles about it in Slate, Buzzfeed, and The Guardian.

It’s famous for being really, really bad. Spectacularly bad. Worse than it should be possible for anything to be. You wouldn’t think you could get The Guardian to write an article about how bad your fanfiction was, but here we are. Everyone agrees that it must have taken a genius to make something so awful, but until recently nobody knew who had authored the pseudonymous work. The Vox article investigates and finds it was probably small-time author Theresa Christodoupolos, who goes by the pen name Rose Christo.

But this leaves other mysteries unresolved. Like: what is going on with it? Its plot makes little sense – characters appear, disappear, change names, and merge into one another with no particular pattern. Even its language is fluid, somewhere between misspelled English and a gibberish that can at best produce associations suggestive of English words.

All these features are unusual in a modern fanfiction. But they’re typical of alchemical texts, which are usually written in a layer of dense allegory. Might this shed more light on My Immortal? After spending way too long investigating this, I find strong evidence in favor. My Immortal is a description of the Great Work of alchemy. Its otherwise-inscrutable symbolism is a combination of three traditions: the medieval opus, the 17th century Rosicrucians, and the native German traditions encoded in Goethe’s Faust. We’ll start by going over these traditions, then delve into the text to unveil the hidden meaning.

First Source: The Medieval Opus

Medieval alchemy centered around the Great Work, or magnum opus, of creating the Philosopher’s Stone. The Stone is supposedly a substance that can transmute lead into gold and grant immortality. But scholars since Jung have also interpreted the opus symbolically, as a process of spiritual transformation. In this reading, the chemical processes are a metaphor for psychological processes, and the creation of the Stone represents the discovery of the true Self, similar to the Christian gnosis or Buddhist enlightenment. Descriptions of the opus tend to describe it as taking place in a series of stages, usually three: nigredo (blackness), albedo (whiteness), and rubedo (redness).

In the first stage (nigredo, “blackness”), you start with some kind of base matter. In the chemical allegory, this is usually lead. In the psychological version, this is the normal mental state, with all of its hangups and uncertainties. The seeker at this stage is symbolized by the raven, blackest of animals. He (the medieval system assumes a male seeker) must begin by confronting his unconscious mind, which takes the form of a dragon. The unconscious is full of bizarre and shameful repressed material, and the seeker’s instinct is to run away. Instead, he must slay the dragon, at which point the dragon rises again as an ally. The seeker then unites with the unconscious in the first “chemical wedding”, ending in a sudden revelation of blinding whiteness – the second stage of the Work.

In the second stage (albedo, “whiteness”), the base matter must be cleansed of its impurities. The seeker is analogized to a child in a baptismal font, or bathing in a stream, or [any of several other water metaphors]. Eventually it begins to shine with its own inner silvery-white light. When the dross has been cleared away, the seeker encounters a second representation of his feminine principle. He unites with the feminine principle in the second “chemical wedding”, and finally see his True Self as it really is.

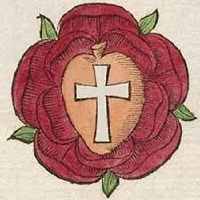
In the third stage (rubedo, “redness”), the seeker has already discovered his True Self as a sort of distant guiding star, but has yet to relate it to the rest of his life or the everyday world. The otherworldy True Self must be united with the seeker’s worldly personality in the final and greatest alchemical wedding, often called the Marriage of the Sun and Moon, or the Marriage of the King and Queen, or [several other flowery metaphors], which joins all opposites into a final cataclysmic union – the Philosopher’s Stone. When this stage ends, the seeker is once again an ordinary person interacting with the ordinary earthly world, but now in a way fully integrated with his true Self. In some traditions, the work is cyclic, and the seeker begins again at the nigredo stage.

Some of this is already in canonical Harry Potter – the first book in the series was originally called Harry Potter and the Philosopher’s Stone, and Rowling included a few alchemical “easter eggs”. In particular, she includes characters named after two of the three stages: Albus Dumbledore (= albedo) and Rubeus Hagrid (= rubedo). There is no character representing the blackness stage, probably because calling somebody “Nigerus” would be Problematic.

In order to turn Rowling’s half-assed name-dropping into a true alchemical allegory, My Immortal has to introduce the missing character with a blackness-themed name. Accordingly, its first sentence starts “Hi my name is Ebony Dark’ness Dementia Raven Way”.

Second Source: The Rosicrucian Writings

These were a series of anonymous pamphlets that took Germany by storm in the early 17th century. They purported to reveal the existence of a secret brotherhood of alchemists, the Rosicrucians, who had discovered vast mystical secrets and were going to disclose them any day now. According to the pamphlets, they had been founded by a hero-sage, Christian Rosenkreutz, who had traveled the world seeking wisdom. But Rosenkreutz was a bit too allegorical for anyone to think he was a real person. His last name was German for “Rosy Cross”, referring to a long tradition of alchemical symbolism in which one produced the Stone by uniting the Rose (the feminine? the spirit? the consciousness?) and the Cross (the masculine? the material world? the body?). The symbolism was a bit unclear, but it caught on, and soon all sorts of mystical groups were using a rose cross as their logo and claiming to be Rosicrucian-inspired.



The most famous Rosicrucian work was The Chymical Wedding Of Christian Rosenkreutz, which purports to be a story about Rosenkreutz getting an invitation to go to a castle for a wedding. But this is just the frame story for throwing a metric ton of inscrutable symbolism at the reader, as Christian successively encounters candle-lighting virgins, golden scales, white serpents, and a bunch of gates and towers. Everybody assumed there were deep mystical secrets contained in this, probably related to the alchemical wedding necessary to achieve the Philosopher’s Stone.

My Immortal wears its Rosicrucian themes on its sleeve. Most obviously, its author uses the not-exactly-subtle pen name “Rose Christo”. But also, the third sentence of the introduction is just “MCR ROX!” A quick check at the My Immortal Wiki tells us that MCR is supposed to be an abbreviation for “My Chemical Romance”.

I maintain that if you are writing a fanfiction of a book about the Philosopher’s Stone, and you use the pen name “Rose Christo”, and you reference a “chemical romance” in the third sentence, you know exactly what you are doing. You are not even being subtle. My Immortal is in part a modern retelling of The Chymical Wedding of Christian Rosenkreutz.

Third Source: Goethe’s Faust II

You know the story. A great alchemist, frustrated by the limitations of his mortal faculties, makes a deal with the Devil. The Devil will give the alchemist anything he desires. In exchange, if the alchemist ever knows a single moment of perfect happiness, he will die and the Devil will get his soul. Maybe you even know the followup: he covets an innocent maiden, Gretchen, and with the help of the Devil he gradually corrupts her until she chooses death over a life of sin.

All that is Faust Part I. Later in life, Goethe wrote the much weirder Faust Part II. A German scholar assures me that “nobody has any idea what it’s about”, except that it is definitely an alchemical metaphor in some way.

A brief synopsis: Faust, thanks to his pact with the Devil, has now become a powerful sorcerer and respected statesman. He decides that a cool thing to do would be to marry Helen of Troy (here called by her German name “Helena”) the most beautiful woman in history. The Devil (with the help of the Sibyl) helps him time-travel back to ancient Greece, where he meets Helena, seduces her, and brings her back to his own time. They have a child together, but the child dies, and in grief Helena departs Faust for the Greek underworld. Faust devotes himself to a different project – raising a new country out of the sea, which he will govern. The country-raising goes really well, and looking upon his new territory, Faust accidentally feels a single moment of perfect happiness. He dies, and the Devil takes his soul and drags him to Hell. Then a choir of angels show up and distract the Devil. While he is distracted, they carry Faust up to Heaven. There he meets all the women in his life – eg Gretchen and Helena – as well as the Virgin Mary. All of them are revealed to be aspects of the Eternal Feminine within himself (or something), and by recognizing this, he is redeemed and found worthy of salvation. The end.

My Immortal is full of symbolic wordplay (for example, did you catch that “MCR ROX” references not just the Chemical Wedding but also its end result, the Stone?) When it mentions that a character is Goth, or seems Goth, or does something in a Goth way, this is often a visual pun (Goth = Goethe) telling us that the scene has a parallel in Faust. We’ll go over some examples later.

Overall Structure

The canonical version of My Immortal is separated into two books of 22 chapters each. In occultism, 22 is the number of completion, especially in Kabbalah (where there are 22 letters of the Hebrew alphabet, 22 generations from Adam to Israel, and 22 paths on the Tree of Life), and Tarot (where there are 22 Major Arcana). So we can think of the book, like the Opus, as a double-traversal of the Tree of Life – first going up from Earth to Heaven, then returning to Earth again.

The medieval and Rosicrucian themes are mostly concentrated in the first half of the text, and consist of a series of thwarted traversals of the alchemical path. At the end of Part 1, a final successful traversal is completed. In Part 2, we segue to a scene-by-scene identity with Part 2 of Faust. More specifically:

Chapter 1 – 5: Alchemical Path 1, nigredo, albedo. Purification fails, seeker sinks back into prima materia.  
Chapters 6 – 18: Alchemical Path 2, nigredo, albedo, partial rubedo. Second alchemical wedding fails, seeker sinks back into prima materia  
Chapters 18 – 22: Alchemical Path 3, nigredo, albedo, rubedo. Second alchemical wedding partly completed, seeker remains in limbo state.  
Chapters 22 – 39: Equivalent to Faust, Act II, Scene 3.  
Chapters 40 – 44. Equivalent to Faust, Act II, Scene 5. Third alchemical wedding succeeds, Stone attained.

I realize these are very odd claims, so I want to demonstrate the flow of symbolism in each of these and compare it to that used in more traditional alchemical texts.

II.

Chapters 1 – 5: First Path

In his papers on alchemy, Carl Jung writes “Great importance was attached to the blackness as the starting point of the Work”. The first sentence of My Immortal begins “Hi my name is Ebony Dark’ness Dementia Raven Way”.

But Ebony’s name isn’t just blackness. It’s a combination of all the different symbols of the nigredo stage. Let’s look at the rest of the Jung quote:

Great importance was attached to the blackness as the starting point of the Work. Generally it was called the “Raven”. In our context the interpretation of the nigredo as terra (earth) is significant. Like the anima media natura or Wisdom, earth is in principle feminine. It is the earth which, in Genesis, appeared out of the waters, but it is also the terra damnata.

I’ve bolded the relevant points. Unlike the albedo and rubedo characters, the nigredo character must be feminine. Her name references the Raven. More speculatively, damnata = Dementia? I think plausibly true. Her name is just a bunch of nigredo symbols strung together.

The nigredo stage begins with a black substance (sometimes identified with lead, the starting point for the transmutation) being placed in a vessel called “the coffin”. From Landauer & Barnes (2011):

The alchemists used a number of different vessels in their work and these vessels – variously known as alembic, coffin, egg, sphere, prison, and womb – particular to stages in the alchemical process. During the blackness of the putrefying Mortificatio, the vessel was represented as a coffin or prison

Each of the three paths in Part I of My Immortal begins with Ebony waking up in a coffin. For example Chapter 2: “I got out of my coffin and took of my giant MCR t-shirt which I used for pajamas”. Remember, MCR means “my chemical romance” and when it appears it usually tells us that we are getting an alchemical analogy.

The black substance in the coffin must then undergo a series of reactions, usually symbolized as interactions between a raven and dragon. Sometimes the raven and dragon are the same entity; other times they are different entities that must confront each other and unite. For example, from the Aurelia Occultae Philosophorum:

I am an infirm and weak old man, surnamed the dragon; therefore am I shut up in a cave, that I may become ransomed by the kingly crown. A fiery sword inflicts great torments on me; death makes weak my flesh and bones. My soul and my spirit depart; a terrible poison, I am likened to the black raven, for that is the wages of sin.

In Jung’s more psychological version, the raven is the seeker and the dragon is the seeker’s unconscious mind. The seeker must begin by confronting his unconscious and all the repressed material therein.



A typical alchemical illustration with raven and dragon. This one isn’t going so well for the raven.

In the first chapter of My Immortal, a raven-named character meets a dragon-named character:

“Hey Ebony!” shouted a voice. I looked up. It was…. Draco Malfoy!

“What’s up Draco?” I asked.

“Nothing.” he said shyly.

But then, I heard my friends call me and I had to go away.

The encounter with the unconscious begins the process of mortificatio. This is cognate with the English word “mortify”, and with good reason – the unconscious is full of all of our deepest and most shameful repressed desires. Herzer and Gillabel describe it like so:

In alchemy the dragon corresponds closely with what Jung called the Shadow. The Shadow is the name for a collection of characteristics and impulses which could be conscious, but which are denied. At the same time we recognize and see them in other people. Some examples of the Shadow are: egotism, laziness, intrigues, unreal fantasies, indifference, or being obsessed by money and possessions. The Shadow is the inferior being in us that desires what we do not allow ourselves because it is uncivilized, because it is incompatible with society’s rules and with the image of our ideal personality. It is all that what we are ashamed of.

When Ebony encounters Draco, she feels shame:

“OMFG, I saw you talking to Draco Malfoy yesterday!” [Willow] said excitedly.

“Yeah? So?” I said, blushing.

“Do you like Draco?” she asked as we went out of the Slytherin common room and into the Great Hall.

“No I so fucking don’t!” I shouted.

“Yeah right!” she exclaimed.

The work of the nigredo stage is to transmute this shame into acceptance and even love. Some sources describe this as slaying the dragon – but after being slain, the dragon rises again in a perfected form. Once the seeker is fully comfortable with their dragon, self and unconscious unite in the first alchemical wedding.

Just then, Draco walked up to me.

“Hi.” he said.

“Hi.” I replied flirtily.

“Guess what.” he said.

“What?” I asked.

“Well, Good Charlotte are having a concert in Hogsmeade.” he told me.

“Oh. My. Fucking. God!” I screamed. I love GC. They are my favorite band, besides MCR.

“Well…. do you want to go with me?” he asked.

I gasped.

Draco invites Ebony on a date, where they will see a band Ebony compares to My Chemical Romance. In general, concert dates in My Immortal represent alchemical weddings. This particular concert is the wedding at the end of the first stage. We know this because the story describes Ebony’s clothing in detail at several points, and it is always some combination of black, white, and red. The particular colors at any given time indicate the stage of the Work being represented. In this case, when dressing up for the concert, Ebony says:”I painted my nails black and put on TONS of black eyeliner. Then I put on some black lipstick.” The only color is black – so she is still entirely in the nigredo stage.

At the end of the concert/date/wedding, Draco and Ebony go to the Forbidden Forest and have sex, representing the union of Ebony and her unconscious. This signals the end of the nigredo stage and the beginning of albedo.

Then he put his thingie into my you-know-what and we did it for the first time.

“Oh! Oh! Oh! ” I screamed. I was beginning to get an orgasm. We started to kiss everywhere and my pale body became all warm. And then….

“WHAT THE HELL ARE YOU DOING YOU MOTHERFUKERS!”

It was…………………………………………………….Dumbledore!

Almost all descriptions of the beginning of albedo emphasize its suddenness. Jung writes that “the nigredo gives way to the albedo…the ever deepening descent into the unconscious suddenly becomes illumination from above”. Remember that Albus Dumbledore is the Harry Potter character representing albedo, even as Ebony represents nigredo. So as Ebony has sex with Draco (deepening descent into the unconscious), suddenly she sees Dumbledore standing over her.



The part where the seeker has sex with the dragon is canon.

But there are signs that the transition is not complete. Dumbledore interrupts her before she can have an orgasm; the union has been only partially consummated. And there are 21 marks between “was” and “Dumbledore” (20 ellipses plus one lone period), which is one short of 22, the mystical number of completion.

As we will see later, Ebony skipped a step – she did not kill the dragon before uniting with it. Therefore, the purification of the albedo appears as the hostile interference of a superego perceived as alien, rather than as a deliberate cleansing from within. Dumbledore admonishes Ebony for her disgusting sexual act and forces her and Draco to separate, ending the path.

Chapter 6 begins “The next day I woke up in my coffin. I put on a black miniskirt that was all ripped around the end and a matching top.” Ebony is back in the coffin, dressed in all black. She has returned to the very beginning of the Work.

Chapters 6 – 18: Second Path

In this section, Ebony retraces her steps. She meets Draco again. They have sex again. But this time, Ebony learns that Draco is in love with Harry Potter (nicknamed “Vampire” in this work), and breaks up with him angrily. This causes Draco to commit suicide – she has slain the dragon. Later, she learns that Draco has inexplicably come back to life (as the alchemical symbolism insists must happen) and is being held hostage in Voldemort’s lair. Recall again the description from Aurelia Occultae Philosophorum: “I am an infirm and weak old man, surnamed the dragon; therefore am I shut up in a cave, that I may become ransomed by the kingly crown”. She rescues Draco and has sex with him again. The path has been completed in full.

But she actually reaches the albedo stage earlier than that – at the moment Draco dies. Let’s go through the symbolism piece by piece:

We practiced for one more hour. Then suddenly Dumbeldore walked in angrily! His eyes were all fiery and I knew this time it wasn’t cause he had a headache.

“What have you done!” He started to cry wisely. (c dats basically nut swering and dis time he wuz relly upset n u wil c y) “Ebony Draco has been found in his room. He committed suicide by slitting his wrists.” […]

I started crying tears of blood and then I slit both of my wrists. They got all over my clothes so I took them off and jumped into the bath angrily while I put on a Linkin Park song at full volume. […]

I got out of the bathtub and put on a black low-cut dress with lace all over it sandly.

Dumbledore appears, again “suddenly”. He gives her the news of Draco’s death. Then Ebony slits her wrists. Compare to Hamilton’s The Alchemical Process Of Transformation:

Completion of the first stage is now experienced as a death, which is in fact a complete letting go of the old sense of self that was identified unconsciously with the earth nature. Images of fire and burning often accompany the images of death and endings. Now we are ready to enter the second stage.

Did you catch how Dumbledore’s “eyes were all fiery”?

Just as the nigredo stage should begin in a coffin, so the albedo stage should begin in a baptismal font. But a bathtub is a perfectly serviceable replacement – in fact, a raven in a bathtub is the same image used in this part of The Chymical Wedding Of Christian Rosenkreutz. A magic egg hatches a black bird, and then:

A bath colored with fine white powder had been prepared for the bird, which enjoyed bathing in it until the lamps placed beneath the bath caused the water to become uncomfortably warm. When the heat had removed all the bird’s feathers it was taken out.

When Ebony gets out of the bath, she puts on a “dress with lace all over it” – the lace is presumably white, so the dress is black-and-white, signifying that she has now attained both the nigredo and the albedo stages. Her current task in the Work is to purify herself, see through the falsehoods of the ego, and behold her true Self.

Ebony notices that Professors Snape and Lupin have been videotaping her naked in the bath. She accuses them of pedophilia and then “I took my gun and shot Snape and Loopin a gazillion times and they both started screaming and the camera broke.” This is the purification and the confrontation with the ego. The ego is analogized to a voyeur with a camera, recording everything we do, showing us our own nakedness. Ebony shoots the camera and breaks it. Compare to Dogen’s analogy of enlightenment to breaking a mirror.

Ebony’s next step should be a second alchemical wedding, this time with her animus.

The medieval alchemists (writing for their male readers) spoke of the anima figure, the feminine archetype within every man. The anima can appear as a hideous old hag, but if accepted, turns into a supernaturally beautiful young woman, the intended bride of the second alchemical wedding. Jung adds that every woman has an equivalent animus, the male archetype. Presumably he too starts out appearing ugly, but if accepted he transforms into a handsome young man.

Ebony’s dual animus is Hagrid on the one side, and Harry “Vampire” Potter on the other. She constantly conflates these two characters in bizarre ways; we readers are supposed to understand that they represent different facets of the same archetype, which Ebony cannot integrate. For example, when Harry comes in in Chapter 12, Ebony says “I THOUGHT IT WAS HAIRgrid but it was Vampire.” Later, she will constantly refer to them by a combined names like “Hairgrid”, “Hargrid”, or even “Hahrid”.

Ebony externalizes the negative aspects of her animus as Hagrid (note the “Hag”, which she cannot bring herself to say), canonically the ugliest person at Hogwarts. Hagrid is everything she hates: prep, not Goth, and “fucked up”. She externalizes the alluring aspects as Harry, canonically rich, famous, and attractive. Ever since Twilight, the archetypal animus figure – the alluring, supernaturally beautiful, mysterious male – has been the vampire, so Harry becomes “Harry ‘Vampire’ Potter”.

There’s more. In My Immortal, Harry has a pentagram-shaped scar. The pentagram is an alchemical symbol, but there’s some evidence that it’s not what’s really involved here. Earlier in the story, Ebony describes Draco as “wearing white foundation and messy eyeliner kind of like a pentagram (geddit) between Joel Madden and Gerard Way.” Later she uses the same device: “My voice sounded lik a pentagram betwen Amy Lee and a gurl version of Gerard Woy”. I think we are supposed to infer that, as a Satanist, she is uncomfortable with naming the Christian symbol (the cross) and so replaces it with a Satanist symbol (the pentagram). So plausibly Harry has a cross-shaped scar.

This is significant because a few paragraphs later, when Ebony is in the hospital after her wrist-slitting injury, Hagrid comes and offers Ebony a bouquet of roses. Harry (the cross-bearer) is the same person as Hagrid (the rose-bearer). Ebony needs to recognize them as different aspects of the same animus, the united Rosy Cross, so she can behold her True Self. Let’s see how she does:

Hargrid came into my hospital bed holding a bouquet of pink roses.

“Enoby I need to tell u somethnig.” he said in a v. serious voice, giving me the roses.

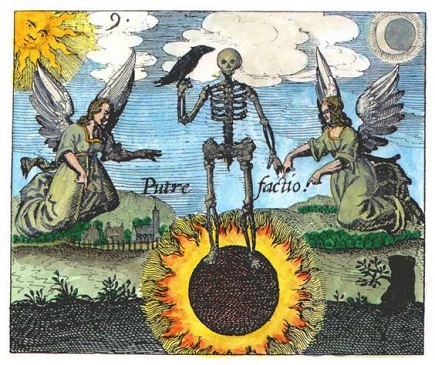
“Fuck off.” I told him. “You know I fucking hate the color pink anyway, and I don’t like fucked up preps like you.” I snapped. Hargrid had been mean to me before for being gottik.

“No Enoby.” Hargrid says. “Those are not roses.”…He suddenly looked at them with an evil look in his eye and muttered Well If you wanted Honesty that’s all you haD TO SAY! .

“That’s not a spell that’s an MCR song.” I corrected him wisely.

“I know, I was just warming up my vocal cordes.” Then he screamed. “Petulus merengo mi kremicli romacio(4 all u cool goffic mcr fans out, there, that is a tribute! specially for raven I love you girl!)imo noto okayo!”

And then the roses turned into a huge black flame floating in the middle of the air. And it was black.



Pictured: Ebony Raven and “a huge black flame floating in the middle of the air.”

If Ebony had accepted Hagrid, he would have turned into Harry, united with her for the second alchemical wedding, and revealed her True Self. Instead, he produces the sign of black fire. The black sun (sol niger) is one of the most ominous symbols in alchemy. Its various connotations are too complicated to explain here (see The Black Sun: The Alchemy And Art Of Darkness for a thorough review) but they are generally highly negative. “The sol niger [is] Saturn, is the shadow of the sun, the sun without justice, which is death for the living.”

Yet at times the Black Sun can be useful; suffering itself becomes a catalyst for transformation. This seems to be one of those times. Ebony is shocked into acknowledging Hagrid (“Now I knew he wasn’t a prep. “OK I believe you now'”)

Is this good enough for her to find her True Self? Going back to the text:

And then the roses turned into a huge black flame floating in the middle of the air. And it was black. Now I knew he wasn’t a prep.

“OK I believe you now wtf is Drako?”

Hairgrid rolled his eyes. I looked into the balls of flame but I could c nothing.

“U c, Enobby,” Dumblydore said, watching the two of us watching the flame. “2 c wht iz n da flmes(HAHA U REVIEWRS FLAMES GEDDIT) u mst find urslf 1st, k?”

“I HAVE FOUND MYSELF OK YOU MEAN OLD MAN!” Hargrid yelled.

Dumbledore tells Ebony that she needs to find herself. But it is Hagrid who retorts that he has found herself. I take this to mean she has failed to integrate with Hagrid; he remains a psychopomp figure, containing the knowledge of Ebony’s Self but unable to transmit it to her.

Nevertheless, something has been gained, because:

Anyway when I got better I went upstairs and put on a black leather minidress that was all ripped on the ends with lace on it. There was some corset stuff on the front. Then I put on black fishnets and black high-heeled boots with pictures of Billie Joe Armstrong on them. I put my hair all out around me so I looked like Samara from the Ring (if u don’t know who she iz ur a prep so fuk off!) and I put on blood-red lipstick, black eyeliner and black lip gloss.

Ebony is still dressed mostly in black, but has a tiny bit of white (the lace) and the slightest bit of red (the lipstick). She is basically back in nigredo now, with only hints of the other stages she has tried to attain.

All of this catches up with her in the total fuckup of an attempted alchemical wedding that follows. If she is actually in albedo, she should be uniting with her animus. If she is actually in rubedo, she should be uniting with God. Instead, it’s just Draco again.

Then I saw a poster saying that MCR would have a concert in Hogsmede right then. We [Ebony and Draco] looked at each other all shocked and then we went 2gether.

There is a long list of Ebony and Draco’s preparations for the concert (cf. the preparation chapters of The Chymical Wedding Of Christian Rosenkreutz, and note the part where Ebony calls Draco “Christian” at the beginning of Chapter 16).



The original chemical romance.

But the actual concert itself is a disaster:

Gerard was da sexiest guy eva! He locked even sexier den he did in pix. He had long raven blak hair n piercing blue eyes. He wuz really skinny and he had n amazing ethnic voice. We moshed 2 Helena and sum odder songz. Sudenly Gerard polled of his mask. So did the other membez. I gasped. It wasn’t Gerard at all! It was an ugly preppy man wif no nose and red eyes… Every1 ran away but me and Draco. Draco and I came. It was…….Vlodemort and da Death Deelers!

The supposed concert/date/wedding proves utterly disastrous – “My Chemical Romance” is just Voldemort wearing a mask. Draco and Ebony flee. Chapter 18 begins:

I woke up the next day in my coffin. I walked out of it and put on some black eyeliner, black eyesharrow, blood-bed lipstick and a black really low-cut leather dress that was all ripped and in stripes so you could see my belly

She is back in the coffin, dressed in all black (she tries to say the word “red”, but fails). She has lost even the hints of previous stages.

Chapters 18 – 22: Third Path

At the beginning of Chapter 18, Ebony is in her coffin, wearing all black. She immediately meets up with Draco (fourth paragraph of Chapter 18). Immediately after this, Dumbledore suddenly appears (fifth paragraph of Chapter 18). There is purification by water (ninth paragraph of Chapter 19, “I ran to the bathroom”). Hagrid suddenly apperas (tenth paragraph of Chapter 19, “Suddenly Hargrid came. He had appearated.”) Hagrid says there is a My Chemical Romance concert that night (thirteenth paragraph of Chapter 19). We have speed-retraced the entire alchemical path accomplished previously.

But this time, Ebony accepts Hagrid by recognizing him as the keeper of the secret of the alchemical wedding (“’U no who MCR r!’ I gasped.”) And so at the concert, the other half of the animus appears to Ebony in a a merciful form (“Vampire and I began 2 make out, moshing to the muzik.”) Instead of going back to her coffin alone, “we went back to our coffins frenching each other…on the gothic red bed together”. Notice that instead of being in a black coffin, they are now in a red bed, the symbol of the bridal chamber after the second alchemical wedding:

Take the fayer Roses, white and red  
And joyne them well in won bed.

This ends the first traversal of the Tree of Life. Part 2 of My Immortal will continue some of the same themes, but subordinate them to a more specific purpose: a reworking of Goethe’s Faust, Part II. In the process, it will show us the completion of the third alchemical wedding and the creation of the Philosopher’s Stone.

Chapters 22 – 39: Faust II, Act 3

For many readers, the weirdest part of My Immortal is the subplot beginning around Chapter 23, where Sybil Trelawney helps Ebony go back in time to the 1980s to seduce young Tom Riddle. Using various bizarre time machines (including Marty McFly’s Delorean) Ebony successfully goes back, woos Tom, returns with him to her own time, and has sex with him.

This is bizarre, but it’s a close parallel of Faust II, Act 3, where the Sibyl helps Faust go back in time to ancient Greece to seduce Helen of Troy. Using various bizarre time machines (including riding on the back of the centaur Chiron) Faust successfully goes back, woos Helena, returns with her to his own time, and has a child with her.

There’s actually an even more direct reference. In Chapter 38, Ebony and Tom are talking about music. Even though Ebony has previously committed to not talking about My Chemical Romance because they didn’t exist in the 1980s, she brings up MCR anyway and Tom is mysteriously familiar with them. Then Ebony says something amazing: “Lol, I totally decided not 2 comit suicide when I herd Hilena.”

In context, “Hilena” is a mispelling of “Helena”, an MCR song. But “Helena” is also Goethe’s Helen of Troy figure. I’ll refer you to The Helena Myth In Goethe’s Faust And Its Symbolism for the full treatment, but the point is that this is basically a one-sentence summary of Part II, Act 3 of Faust. It is Faust’s encounter with Helena, representing the feminine ideal, which saves him from despair and makes life worth living. The Ebony-Tom relationship in My Immortal is a close parallel to this, and here it cheekily calls out the original to anyone with ears to listen.

But the matchup is not perfect. Faust scholars identify three alchemical weddings in the book: first to Gretchen, then to Helena, then to a divine figure representing the Virgin Mary. This mirrors the traditional opus – first you unite with your unconscious, then with the anima, then with divinity. The Helena references in Faust all correspond to the second stage – Helena is Faust’s anima.

Ebony has already had an alchemical wedding with her animus in Chapter 20 – the My Chemical Romance concert where she makes out with Vampire. Looking back, there is a Helena reference there too:

Vampire and I began 2 make out, moshing to the muzik. I gapsed, looking at da band. I almost had an orgasim. Gerard was so fucking hot! He begin 2 sing ‘Helena’ and his sexah beautiful voice began 2 fill the hall.

But we should be done with these references! Ebony should be completely in the third stage by now! In Chapter 38, she attends a concert with Satan. This ought to be her third alchemical wedding: union with a divine figure. Given that she is Satanist, the appropriate divine figure is right there. Instead, they’re not only talking about Helena, they’re unconsciously re-enacting the Helena subplot from Faust. Why?

I think the last alchemical wedding never completed. She “makes out” with Vampire, but does not have sex with him. The union is only partial. Her date with Satan is a confused attempt at the second and third alchemical weddings combined, which is why she can’t decide whether to call him Tom Riddle (another animus figure – remember that canonically Harry Potter is a horcrux of Tom Riddle and they share part of the same soul) or Satan (the divine figure). So although the explicit text is a bunch of parallels to the third/final alchemical union with the divine, the symbology and the Faust metaphor are caught in the second stage.

Once again, things have gotten extremely bad for Ebony’s spiritual growth. We saw a prelude of this with the omen of the Black Sun in Chapter 12. Now things have deteriorated further. She is going to have to take the hard route: she must go through the fire.

In the middle of the concert with Satan (Chapter 38), the song suddenly becomes dissonant. James Potter tries to shoot Lucius Malfoy, but Ebony jumps in front of the bullet and dies.

Chapters 39 – 44: Faust II, Acts 4-5

Chapter 39 starts with a prelude saying that it’s written by a hacker who hates My Immortal. He cracked the real author’s password and plans to ruin the story by writing a deliberately uncharacteristic chapter. He breaks with all of the normal author’s stylistic conventions, eg by using good spelling and grammar throughout. This new chapter ends with Ebony going to Hell, staying there for all eternity, and never being able to do anything Goth ever again.

Then the original author reasserts control, apologizes for letting her account get hacked, and starts over. According to the new, canon timeline, Ebony survives her apparent death (because she was back in time, and so couldn’t really die) and returns to her own era.

Compare to the ending of Faust. Mephistopheles appears and drags Faust’s soul to Hell. But a choir of angels show up, distract him, and steal Faust’s soul away to Heaven.

The end of My Immortal also revolves around Ebony’s redemption from Hell, but via a novel plot device making use of the fourth wall. Ebony is saved not by a spiritual conflict on her own plane, but by a conflict on a higher plane – that between her Author and a hacker trying to destroy the Author’s story. When the Author wins by getting her password back, Ebony is released from Hell; her eternal damnation is retroactively cancelled. This is a sort of weird way of doing a deus ex machina, but honestly it’s less jarring than Goethe’s version.

Upon his salvation, Faust meets all the female characters from his life again, and they redeem him through the power of the Eternal Feminine. Similarly, upon her release Ebony meets all the male characters from her own life. In her case, this leads to an orgy. In Chapter 43, she has a foursome with Draco, Vampire, and Satan. Now, finally, she is consummating the final alchemical wedding, the “union of all opposites” in which she achieves an ultimate integration with all the male aspects of her personality.

Chapter 44, the last chapter of My Immortal, has no parallel in Faust. Instead of the hieros gamos between masculine and feminine completing the redemptive process and ending the Great Work, in My Immortal it only initiates the final apocalypse. All the good characters and all the evil characters show up in the Great Hall and begin to fight in a difficult-to-follow scene. Finally, Dumbledore tells Ebony she has to fulfill her destiny and kill Voldemort. In the last line, Ebony shouts “ABRA KEDABRA!!!!!!!!!!!11111” and the story ends.

What is the meaning of this final word? It partly corresponds to the Harry Potter spell avada kedavra. But the spelling is neither the Rowling version nor the traditional stage magician version.

The closest match I can find is from Aleister Crowley’s Book of the Law, whose last sentence is “The ending of the words is the Word Abrahadabra” In Crowley’s commentaries, he explains that “[abrahadabra is] the Word of the Aeon, which signifieth The Great Work accomplished.”

So My Immortal ends with an occult term signifying the “ending of words” and completion of the alchemical Great Work. This suggests Ebony’s redemption was successful; she has escaped Hell, merged with the Eternal Masculine, and triumphed over Voldemort (representing death). She has created the Philosopher’s Stone and achieved the story’s namesake immortality.

But in a different commentary, Crowley also writes: “Abrahadabra is the glyph of the blending of the 5 and 6, the Rose and the Cross.”

…which suggests one last, “hidden” chapter.

The hacker subplot in Chapter 39 suggests that “the story” should be taken to include not just Ebony’s own story, but the frame story in which an author is writing My Immortal as a Harry Potter fanfiction and confronting reviewers, hackers, etc. The Ebony story ends with Ebony speaking the mystic word that unites Rose and Cross. The frame story picks up with a woman named Rose Christo appearing and identifying herself as the author. So My Immortal can be said to end with Ebony abandoning her false self (Ebony Dark’ness Dementia Raven Way, the dark leaden substance of the nigredo), and becoming her divine higher Self (Rose Christo, the completed union of Rose and Cross, who exists on a higher plane than Ebony).

Most of the second part of My Immortal mirrors Part II of Faust. But its ending transcends the source material. The moral of Faust is that you can be redeemed. But My Immortal actively demonstrates the redemption that Faust can only point at. It follows the progress of Ebony from a contemptible character in a terrible fanfiction, through the alchemical process of uniting Rose and Cross, to become Rose Christo, a woman who lives in the real world. It tells us that every one of us is a Mary Sue in the bad fanfiction of our lives – the narrative created by the ego in order to maintain our illusory selfhood. And in the tradition of other great alchemical texts like The Chymical Wedding Of Christian Rosenkreutz and Faust, it gives us a blueprint for escaping that fanfiction and completing the alchemical Work of breaking into reality.



As the end of Faust puts it:

All of the transient,  
Is parable, only:  
The insufficient,  
Here, grows to reality:  
The indescribable,  
Here, is done.  
Woman, eternal  
Beckons us on.

# Creationism, Unchallenged

How much should responsible news organizations report on stupid things?

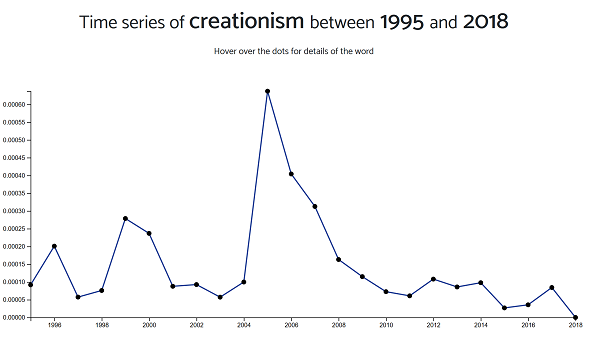
If they don’t report at all, the stupid things go unchallenged. But if they report too much, then they signal-boost the stupid thing and give it free publicity (eg Donald Trump). Also, people who mistrust the media might reflexively support the stupid thing just because the media hates it (eg Donald Trump). Also, the more time you waste covering stupid things, the less time you have for real news (eg Donald Trump).

I recently read Causes And Consequences Of Mainstream Media Dissemination Of Fake News: Literature Review And Synthesis, which argues that the news might be covering too many stupid things right now. The authors note that “only 2.6% of visits to current affairs articles were to fake news websites” (though other sources suggest more) and that the mainstream press bears some responsibility for spreading inaccuracies beyond this small demographic. But they also understandably worry that maybe if the mainstream press wasn’t so aggressive in covering and debunking fake news, then fake news would go uncorrected.

When I think about this problem, I remember creationism.

In the early 2000s, creationism was Public Enemy Number…maybe not One, but somewhere in the top ten. If you’re old enough to remember the decade at all, you probably recall the key flashpoints. The Discovery Institute. Michael Behe. “Teach the controversy”. The Creation Museum. Of Pandas And People. That one anti-Richard-Dawkins rap song which somehow despite everything managed to be really good.

And you probably remember the efforts by “the reality based community” to spread awareness of the dangers of creationism – the xkcd comics, the petitions by 1400 scientists named Steve, the New York Times articles:



Frequency of the word ‘creationism’ in the New York Times as a percent of all words, source here but currently down

…yeah, the 2000s were a weird time. I’ve talked about this particular conflict already in my post New Atheism: The Godlessness That Failed. Today I want to focus on another aspect.

All those creationists are still there. A 2019 Gallup poll found that 40% of Americans believed “God created human beings pretty much in their present form at one time within the last 10,000 years or so”, little different from 44% who believed it when they first asked in 1983 or the 46% who believed it in 2006.

And the Discovery Institute! They’re still there! You can read their very modern-looking website at discovery.org. Their blog has three new posts on creationism just from the past two days (eg A Hot Seller From Discovery Institute Press: New Book Offers Intelligent Design In A Nutshell).

Same with the Creation Museum! Last year they completed a $5.5 million upgrade, including a planetarium, “4D theater”, and a snazzy “Biblical authority exhibit”. And they still have their full-scale replica of Noah’s Ark (looking more and more prescient these days).

Same with Michael Behe! He’s still publishing! Last year he released his newest book, Darwin Devolves, which “gives a sweeping tour of how modern theories of evolution fall short and how the devolving nature of Darwin’s mechanism limits them even further”. Also, in case you wanted to read Behe’s opinions about the coronavirus, that is a thing you can do.

As far as I can tell, the creationists are putting in just as much effort today as they did in 2006. But the mainstream went from fiercely challenging them, to totally ignoring them. And the change didn’t help them at all. They haven’t won any major victories, or convinced any more people. If anything, they’re doing worse – nobody hears about them. Although the decline in media coverage hasn’t prevented people from being creationist, it hasn’t helped creationism spread or build clout either.

I see people using rivers of ink to fight the modern equivalents of creationists. Pizzagaters, flat-earthers, moon-hoaxers, QAnon, deep-staters, people who say the coronavirus is a bioweapon, Alex Jones. Are they sure it’s not equally useless? Equally counterproductive?

Even beyond that, I see people willing to legitimize any tactic if it gives them a leg up on this group – censorship, social shaming, no-platforming, changing social media from a free public square to a carefully-monitored walled garden. Spreading the cowpox of doubt, teaching people to optimize for solving easy problems in ways that make it harder for them to think about the hard ones. The justification is always the same – if we don’t tighten control, then facts and science will lose out to bullshit and denialism, and fringe ideologies will burst into the mainstream and overwhelm it.

If that were the only way to save civilization from anti-science barbarism, maybe it would be a worthwhile trade. But the experience of unchallenged creationism suggests maybe we can relax.

# Bush Did North Dakota

Continuing yesterday’s discussion of fake news:

Guess et al says that 46% percent of Trump voters endorsed the Pizzagate conspiracy theory. Does this mean fake news is very powerful?

We can compare this to belief in various other conspiracy theories, as measured by the 2016 Chapman University Survey Of American Fears. About 24% believe there’s a government conspiracy to cover up the truth about the moon landing, 30% about Obama’s birth certificate, and 33% about the North Dakota crash.

This last one is especially interesting because there was no unusual crash in North Dakota when the survey was written. The researchers included it as a placebo option to see if people would endorse a conspiracy theory that didn’t exist. 33% of them did.

Before we make fun of these people, consider: there’s a strong presumption that surveys don’t contain made-up questions. There was no “don’t know” option included on the poll, just various shades of “agree” or “disagree”. In order to condemn the people who “agreed” that the government was probably covering up the crash, we would have to assert that the more correct answer was “disagree”. In other words, that people should have an assumption of trusting the government, until they get some specific reason to distrust it. You can make that argument, but it’s not obvious. You could also start from the opposite assumption, where the government is guilty until proven innocent.

To put it another way, suppose I gave you the following survey:

SELECT AGREE OR DISAGREE, YOU MAY NOT SAY “DON’T KNOW” OR LOOK FOR MORE INFORMATION. Alex Jones is lying when he talks about:

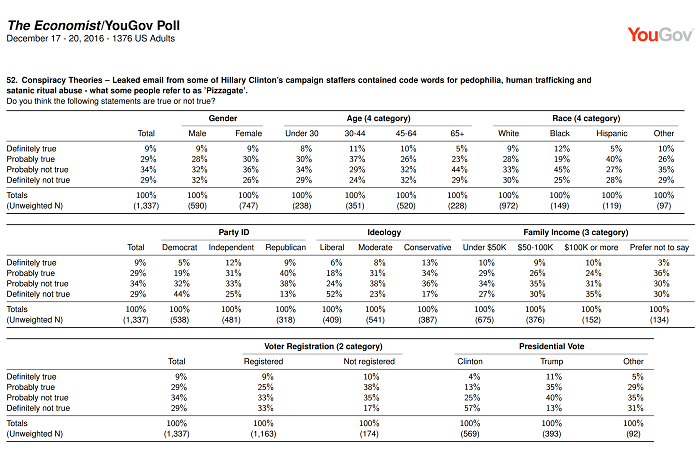
1. Sandy Hook  
2. The coronavirus  
3. Obama’s birth certificate  
4. The North Dakota crash

…many of us would guess he was lying about the North Dakota crash, without a second thought. And if there later turned out to be no North Dakota crash, we wouldn’t feel particularly ashamed; under the circumstances we made the right choice. If you think the government is as untrustworthy as Alex Jones, well, there you go.

I’ve previously talked about a lizardman constant of 4% on polls. That is, it’s hard to get a poll result much lower than four percent for anything, because of respondents making mistakes or trolling. If 4% of people supposedly believe something, that doesn’t mean we need to be concerned about that fraction of the population, it just means that poll has it its floor and it’s hard to conclude what the real number is.

In the same way, maybe we can posit a North Dakota constant of 33%. This is how many people believe in conspiracy theories when there’s no reason at all to believe them, not even the flimsy reasons conspiracy theories usually provide. Sometimes, if there’s a lot of evidence against them, fewer than 33% will believe in a given theory. But if it’s just “Conspiracy! True or false?” – 33% will say true.

Let’s look again at that statistic from the Guess paper – “46% of Trump voters believe”. I think their source is this poll, which finds:



Overall 38% of Americans agreed with the claim, so Trump voters (46%) were not outrageously more likely than anyone else. Other groups unrelated to ideology were about equally likely to believe it (eg 45% of Hispanics).

Like the North Dakota question, this one had no “unsure” or “what the hell are you talking about” option, forcing everyone to feign agreement or disagreement. We see that the majority of agreement is lukewarm. 75% of Trumpists and 85% of Hispanics who believe Pizzagate only “probably” rather than “definitely” believe it.

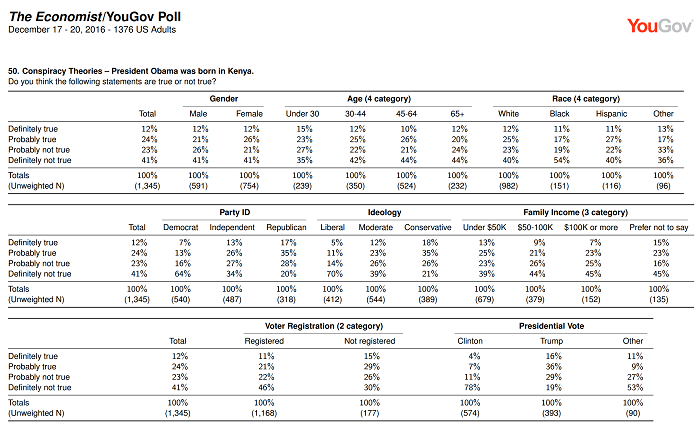
I don’t think the evidence suggests Trump voters live in an outrageously different world from the rest of us. Instead, it suggests there’s a North Dakota constant of 33% – the number of people who will believe a conspiracy theory for no particular reason. It looks like about 10 – 15% more Trump supporters than predicted believe Pizzagate, probably because it attacks Clinton, and 10 – 15% fewer Hillary supporters than predicted believe it. But these are relatively small effects, and equaled by eg whatever mysterious thing is going on with Hispanics. In any case, it all averages out to about the predicted amount.

Why is this North Dakota Constant of 33% so different from the Lizardman Constant of 4%? I don’t know. Lizardmen seem like a pretty crazy conspiracy theory, but is Hillary’s involvement with Satanic pizza parlors really that much less weird? Sure, Pizzagate is more politicized, and that might make some difference – but then how come a full 24% of Democrats believe it, six times more than Lizardman’s Constant predicts?

One part of the story is that the lizardman poll offered “don’t know”, and 7% of people chose that. If, denied that option, those people would split evenly between yes and no, that brings us up to 7%ish pro-lizardman. But that’s still nowhere near 33%.

I think this is probably a story about low-information voters. If you imagine you’ve never heard about Pizzagate, and you read the question as written, it doesn’t sound too outlandish. Some Clinton staffers’ emails contained some code words. The pedophilia and Satanic abuse are pretty out there, but post-Jeffrey Epstein we all assume somebody’s doing some kind of creepy pedophilia stuff somewhere. Maybe if you don’t know anything about this, and you don’t have the strong priors about Satanic ritual abuse that you get from studying the history of those claims in the 80s and 90s, this one seems like a toss-up. Certainly it seems like more of a toss-up than a clearly-stated assertion that reptilian aliens rule the world. If your prior is “most conspiracy theory-ish things are probably true”, this sounds like the kind of thing that could be true, whereas you might balk at the lizardman statement.

Here’s another question from the same poll:



Who believes Obama was secretly born in Kenya? Lots of people – including 28% of blacks. I’ve been told again and again that birtherism is a racist conspiracy theory and no person could possibly believe it except as a way of dog whistling white supremacy. Yet here we are with 28% of blacks supporting it – and this isn’t a small sample either! I have no idea what these people are thinking, except that 28% is pretty close to the North Dakota Constant and maybe we should just write this one off.

I conclude we probably shouldn’t draw too many conclusions from specific statements like “X% of [GROUP] supports [CONSPIRACY THEORY]”, especially if X is around 33%. It’s probably just the North Dakota Constant. Likewise, we shouldn’t interpret Pizzagate’s high polling numbers as much evidence that fake news is very convincing – though you could still make an argument that fake news plays a role in transmitting believable conspiracy theories to people who are predisposed to believe them.

Of course, there are some high-information voters who still believe these things really strongly. I think they deserve a more complete treatment, which I want to give later. I think a preliminary sketch might look like: if you start with a prior on something being true, you don’t necessarily need much evidence. The North Dakota question suggests that conspiracy theorists start with a high prior on any given conspiracy being true. What remains to be explained is why some people stick to that prior even after they get more information.

# Book Review: Origin Of Consciousness In The Breakdown Of The Bicameral Mind

I.

Julian Jaynes’ The Origin Of Consciousness In The Breakdown Of The Bicameral Mind is a brilliant book, with only two minor flaws. First, that it purports to explains the origin of consciousness. And second, that it posits a breakdown of the bicameral mind. I think it’s possible to route around these flaws while keeping the thesis otherwise intact. So I’m going to start by reviewing a slightly different book, the one Jaynes should have written. Then I’ll talk about the more dubious one he actually wrote.

My hypothetical Jaynes 2.0 is a book about theory-of-mind. Theory-of-mind is our intuitive model of how the mind works. It has no relation to intellectual theories about how the mind is made of cognitive algorithms or instantiated on neurons in the brain. Every schoolchild has a theory-of-mind. It usually goes like this: the mind is an imaginary space containing things like thoughts, emotions, and desires. I have mine and you have yours. I can see what’s inside my mind, but not what’s inside your mind, and vice versa. I mostly choose the things that are in my mind at any given time: I will thoughts to happen, and they happen; I will myself to make a decision, and it gets made. This needs a resource called willpower; if I don’t have enough willpower, sometimes the things that happen in my mind aren’t the ones I want. When important things happen, sometimes my mind gets strong emotions; this is natural, but I need to use lots of willpower to make sure I don’t get overwhelmed by them and make bad decisions.

All this seems so obvious to most people that it sounds like common sense rather than theory. It isn’t; it has to be learned. Very young children don’t start out with theory of mind. They can’t separate themselves from their emotions; it’s not natural for them to say “I’m really angry now, but that’s just a thing I’m feeling, I don’t actually hate you”. It’s not even clear to them that people’s minds contain different things; children are famously unable to figure out that a playmate who has different evidence than they do may draw different conclusions.

And the learning isn’t just a process of passively sitting back observing your own mind until you figure out how it works. You learn it from your parents. Parents are always telling their kids that “I think this” and “What do you think?” and “You look sad” and “It makes me feel sad when you do that”. Eventually it all sinks in. Kids learn their parent’s theory-of-mind the same way they learn their parents’ language or religion.

When in human history did theory-of-mind first arise? It couldn’t have been a single invention – more like a gradual process of refinement. “The unconscious” only really entered our theory-of-mind with Freud. Statements like “my abuse gave me a lot of baggage that I’m still working through” involves a theory-of-mind that would have been incomprehensible a few centuries ago. It’s like “I’m clicking on an icon with my mouse” – every individual word would have made sense, but the gestalt would be nonsensical.

Still, everyone always assumes that the absolute basics – mind as a metaphorical space containing beliefs and emotions, people having thoughts and making decisions – must go back so far that their origins are lost in the mists of time, attributable only to nameless ape-men.

Julian Jaynes doesn’t think that. He thinks it comes from the Bronze Age Near East, c. 1500 – 750 BC.

II.

Jaynes (writing in the 1970s) was both a psychology professor at Princeton and an expert in ancient languages, so the perfect person to make this case. He reviews various samples of Bronze Age writing from before and after this period, and shows that the early writings have no references to mental processes, and the later ones do. When early writings do have references to mental processes, they occur in parts agreed by scholars to be later interpolations. If, with no knowledge of the language itself, you tried to figure out which parts of a heavily-redacted ancient text were early vs. late by their level of reference to mental processes, you could do a pretty decent job.

I don’t speak fluent Sumerian, so I am forced to take Jaynes’ word for a lot of this. It’s even worse than that, because Jaynes argues that other translators sometimes err and translate non-mental terms in mental ways. This is an easy mistake for them to make, because most cultures, once they got theory of mind, repurposed existing language to represent it. Jaynes makes a convincing case for why this would happen, and convincingly argues for why his interpretations are truer to the spirit of the text, but it does mean you can’t double-check his work by reading the works in translation.

Jaynes spends the most time talking about the Iliad, with good reason – it’s the longest Bronze Age work we have, and in many ways it’s a psychodrama, focusing as much on the characters of Achilles, Hector, etc as the plot itself. It came together piecemeal through the efforts of Greek bards between about 1100 and 800 BC, finally reaching a canonical version in the mouth of “Homer” around 700 BC – the period Jaynes says theory of mind was starting to evolve. Jaynes uses it to trace the development process, showing how older sections of the Iliad treat psychology in different ways than newer ones.

So for example, a typical translation might use a phrase like “Fear filled Agamemnon’s mind”. Wrong! There is no word for “mind” in the Iliad, except maybe in the very newest interpolations. The words are things like kardia, noos, phrenes, and thumos, which Jaynes translates as heart, vision/perception, belly, and sympathetic nervous system, respectively. He might translate the sentence about Agamemnon to say something like “Quivering rose in Agamemnon’s belly”. It still means the same thing – Agamemnon is afraid – but it’s how you would talk about it if you didn’t have an idea of “the mind” as the place where mental things happened – you would just notice your belly was quivering more. Later, when the Greeks got theory of mind, they repurposed all these terms. You can still find signs of this today, like how we say “I believe it in my heart”. In fact, you can still find this split use of phrenes, which has survived into English both as the phrenic nerve (a nerve in the belly) and schizophrenia (a mental disease). As the transition wore on, people got more and more flowery about the kind of feelings you could have in your belly or your heart or whatever, until finally belly, heart, and all the others merged into a single Mind where all the mental stuff happened together.

The Iliad uses these body parts to describe feelings despite its weak theory of mind. Its solution for describing thoughts and decision-making is more…unconventional.

Suppose Achilles is overcome with rage and wants to kill Agamemnon. But this would be a terrible [idea]; after [thinking] about it for a while, he [decides] against. If Achilles has no concept of any of the bracketed words, nothing even slightly corresponding to those terms, how does he conceptualize his own actions? Jaynes:

The response of Achilles begins in his etor, or what I suggest is a cramp in his guts, where he is in conflict, or put into two parts (mermerizo) whether to obey his thumos, the immediate internal sensations of anger, and kill the king, or not. It is only after this vacillating interval of increasing belly sensations and surges of blood, as Achilles is drawing his mighty sword, that the stress has become sufficient to hallucinate the dreafully gleaming goddess Athene who then takes over control of the action and tells Achilles what to do.

Wait, what?

III.

As you go about your day, you hear a voice that tells you what to do, praises you for your successes, criticizes you for your failures, and tells you what decisions to make in difficult situations. Modern theory-of-mind tells you that this is your own voice, thinking thoughts. It says this so consistently and convincingly that we never stop to question whether it might be anything else.

If you don’t have theory of mind, what do you do with it? Children don’t have theory of mind, at least not very much of it, and more than half of them have imaginary friends. Jaynes has done some research on the imaginary friend phenomenon, and argues that a better term would be “hallucinatory friend” – children see and hear these entities vividly. The atheoretical mind is a desperate thing, and will comply with any priors you give it to make sense of its experiences. If that prior is that the voice in your head is a friend – or god – it will obediently hallucinate a friend or god for you, and modulate its voice-having accordingly.

I know some very smart and otherwise completely sane evangelical Christians who swear to me that God answers their prayers. They will ask God a question, and they will hear God’s voice answer it. God’s voice may not sound exactly like an external voice, and it may give them only the advice they would have given themselves if they’d thought about it – but they swear that they are not thinking about it, that their experience is qualitatively different than that. And these are normal people! If you’re a special person – a saint or mystic, say – and you actively court the experience by fasting and praying and generally stressing your body to the limit – then the voice will be that much louder and more convincing.

There are even whole forms of therapy based on this kind of thing. In Internal Family Systems, the therapist asks the patient to conceptualize some part of their mind (maybe the part that’s producing a certain symptom) as a person, and to talk to it. I know people who swear that this works. They approach their grief or anger or anxiety, and they get a clear image of what “he” or “she” looks like, and then “he” or “she” talks to them. Usually he/she tells them some appropriately psychological sounding thing, like “Hello, I am your anxiety, and I’m only inflicting these fears on you because we were abused as a child and I want to make sure nobody ever abuses us like that again”. Then the patient talks to their anxiety and hopefully strikes a bargain where the patient agrees to take the anxiety’s perspective into account and the anxiety agrees not to make the patient so anxious all the time. Some people swear by this, say it’s helped them where nothing else can, and absolutely insist they are having a real dialogue with their anxiety and not just making up both sides of the conversation.

Most of the people who seem to really like IFS have borderline personality disorder. And borderline people are also at the most risk of dissociative identity disorder (multiple personality). Multiple personality has two main risk factors: borderline, and somebody suggesting to you that multiple personality disorder might be a reasonable thing to have. For a while in the 80s, psychiatrists were really into multiple personality and tried diagnosing everyone with it, and sure enough all those people would admit to having multiple personalities and it would be very exciting. Then the APA told the psychiatrists to stop, people stopped talking about multiple personality as much, and now the condition is rarer.

A few years ago, someone rediscovered/invented tulpamancy, the idea of cultivating multiple personalities on purpose because it’s cool. People who try to do this usually succeed. At least they say they’ve succeeded, and I believe they think this. I think their internal experience is of talking to a different entity inside of them. Also, I have a friend who writes novels, and one time she created such a detailed mental model of one of her characters that it became an alternate personality, which she still has and considers an important part of her life. She is one of the most practical people I know and not usually prone to flights of fancy.

I also have less practical friends, friends who are into occultism. They tell me they sometimes make contact with spiritual entities. I believe them when they say they have these experiences. I believe them when they say that they were not purposely guiding their Ouija board to say whatever it said. I don’t have any friends who are cool enough to have gone through the whole procedure for summoning your Holy Guardian Angel, but from what I read, completing the ritual directly does tend to leave you with an angel who hangs around you and gives you advice. I believe the people who say this is their experience of completing the ritual.

I conclude that giving yourself multiple personalities is actually pretty easy under the right circumstances. Those circumstance are a poor theory of mind (I think borderlines are naturally bad at this) and a cultural context in which having a multiple personality is expected.

Jaynes says ancient people met both criteria. They had absolutely no theory of mind, less theory of mind than the tiniest child does today. And their cultural context was absolutely certain that gods existed. Just as we teach our children that the voice in their mind is them thinking to theirselves, so the ancients would teach their children that the voice in their head was a god giving them commands. And the voice would dutifully mold itself to fit the expected role.

Here Jaynes is at his most brilliant, going through ancient texts one by one, noting the total lack of mental imagery, and highlighting the many everyday examples of conversations with gods. Every ancient culture has near-identical concepts of a god who sits inside of you and tells you what to do. The Greeks have their daemons, the Romans their genii, the Egyptians their ka and ba, and the Mesopotamians their iri. The later you go, the more metaphorically people treat these. The earlier you go, the more literal they become. Go early enough, and you find things like the Egyptian Dispute Between A Man And His Ba which is just a papyrus scroll about a guy arguing loudly with the hallucinatory voice of his guardian spirit, and the guardian spirit’s hallucinatory voice arguing back, and nobody thinking any of this is weird (people who aren’t Jaynes would wimp out and say this is “metaphorical”). Every ancient text is in complete agreement that everyone in society heard the gods’ voices very often and usually based decisions off of them. Jaynes is just the only guy who takes this seriously.

Turn on what Terry Pratchett called “first sight and second thoughts” and try to look at the Bronze Age with fresh eyes. It was really weird. People would center their city around a giant ziggurat, the “House of God”, with a giant idol within. They would treat this idol exactly like a living human – feeding it daily, washing it daily, sometimes even marching it through the streets on sedan chairs carried by teams of slaves so it could go on a “connubial visit” to the temple of an idol of the opposite sex! When the king died, hundreds of thousands of men would labor to build him a giant tomb, and then they would kill a bunch of people to serve him in the afterlife. Then every so often it would all fall apart and everyone would slink away into the hills, trying to pretend they didn’t spend the last twenty years buliding a jeweled obelisk so some guy named Ningal-Iddida could boast about how many slaves he had.

If the Bronze Age seems kind of hive-mind-y, Julian Jaynes argues this is because its inhabitants weren’t quite individuals, at least not the way we think of individuality. They were in the same kind of trance as a schizophrenic listening to voices commanding him to burn down the hospital. All of it – the ziggurats, the obelisks, the pyramids – were an attempt to capture not individual humans, but those humans’ daemons – to get people to identify the voice in their head with the local deity, and replace their free will with a hallucinatory god who represented their mental model of society’s demands on them. In the best case scenario, the voice would be interpreted as the god-king himself, giving you orders from miles away. Jaynes argues the Bronze Age was obsessed with burials and the afterlife (eg the Pyramids) because if you had internalized the voice in your head as Pharaoh Cheops, the voice wasn’t going to go away just because the actual Pharaoh Cheops had died hundreds of miles away in the capital. So even after Pharaoh Cheops dies, as far as all his subjects can tell, he’s still around, commanding them from the afterlife. So they had better keep him really, really happy, just as they did during life. Jaynes presents various pieces of evidence that the main function of pyramids was as a place where you could go to commune with the dead Pharaoh’s spirit – ie ask it questions and it would answer them.

He has a similar explanation for idols. The Bronze Age loved idols. There were the giant idols, ones that made the statue of Zeus at Olympia look like a weak effort. But also, every family had their own individual idols. Archaeologists who dig up Bronze Age houses just find idol after idol after idol, like the ancient Sumerians did nothing except stare at idols all day. Jaynes thinks this is approximately true. Idols were either cues to precipitate hallucinatory voices, or else just there to make conversation more comfortable – it’s less creepy if you can see the person you’re talking to, after all.

Then, around 1250 BC, this well-oiled system started to break down. Jaynes blames trade. Traders were always going into other countries, with different gods. These new countries would be confusing, and the traders’ hallucinatory voices wouldn’t always know all the answers. And then they would have to negotiate with rival merchants! Here theory of mind becomes a huge advantage – you need to be able to model what your rival is thinking in order to get the best deal from him. And your rival also wants theory of mind, so he can figure out how to deceive you. Around 1250 BC, trade started picking up, and these considerations became a much bigger deal. Then around 1200 BC, the Bronze Age collapsed. It’s still not exactly clear why (some of you may have heard me give a presentation on this), though most guesses involve a combination of climate change plus the mysterious Sea Peoples. Whole empires were destroyed, their populations become refugees who flooded the next empire in turn. Now everyone was in unfamiliar territory; nobody had all the answers. The weird habits of mind a couple of traders had picked up became vital; people adopted them or died.

But as theory of mind spread, the voices of the gods faded. They receded from constant companions, to only appearing in times of stress (the most important decisions) to never appearing at all. Jaynes interprets basically everything that happened between about 1000 BC and 700 BC as increasingly frantic attempts to bring the gods back or deal with a godless world.

Now, to be fair, he cites approximately one zillion pieces of literature from this age with the theme “the gods have forsaken us” and “what the hell just happened, why aren’t there gods anymore?” As usual, everyone else wimps out and interprets these metaphorically – claiming that this was just a poetic way for the Mesopotamians to express how unlucky they felt during this chaotic time. Jaynes does not think this was a metaphor – for one thing, people have been unlucky forever, but the 1000 – 750 BC period was a kind of macabre golden age for “the gods have forsaken us” literature. And sometimes it seems oddly, well, on point:

My god has forsaken me and disappeared  
My goddess has failed me and keeps at a distance  
The good angel who walked beside me has departed.

Or:

One who has no god, as he walks along the street  
Headache envelops him like a garment

Jaynes says that “there is no trace whatsoever of any such concerns in any literature previous to the texts I am describing here”.

So people got desperate. He says this period was the origin of augury and divination. Omens “were probably present in a trivial way” before this period, but not very important; “there are, for example, no Sumerian omen texts whatsoever”. But after about 1000 BC, omens become an international obsession.

Towards the end of the second millennium BC…such omen texts proliferate everywhere and swell out to touch almost every aspect of life imaginable. By the first millennium BC, huge collections of them are made. In the library of King Ashurbanipal at Nineveh about 650 BC, at least 30% of the twenty to thirty thousand tablets come into the category of omen literature. Each entry in these tedious irrational collections consists of an if-clause or protasis followed by a then-clause or apodosis. And there were many classes of omens…

– If a town is set on a hill, it will not be good for the dweller within that town.

– If black ants are seen on the foundations which have been laid, that house will get built; the owner of that house will live to grow old.

– If a horse enters a man’s house, and bites either an ass or a man, the owner of the house will die, and his household will be scattered.

– If a fox runs into the public square, that town will be devastated.

– If a man unwittingly treads on a lizard and kills it, he will prevail over his adversary.

And then there are the demons. Early Sumerians didn’t really worry about demons. Their religion was very clear that the gods were in charge and demons were impotent. Post 1000 BC, all of this changes.

As the gods recede…there whooshes into this power vacuum a belief in demons. The very air of Mesopotamia became darkened with them. Natural phenomena took on their characteristics of hostility toward men, a raging demon in the sandstorm sweeping the desert, a demon of fire, scorpion-men guarding the rising sun beyond the mountains, Pazuzu the monstrous wind demon, the evil Croucher, plague demons, and the horrible Asapper demons that could be warded off by dogs. Demons stood ready to seize a man or woman in lonely places, while sleeping or eating or drinking, or particularly at childbirth. They attached themselves to men as all the illnesses of mankind. Even the gods could be attacked by demons, and this sometimes explained their absence from the control of human affairs…

Innumerable rituals were devoutly mumbled and mimed all over Mesopotamia throughout the first millennium B.C. to counteract these malign forces. The higher gods were beseeched to intercede. All illnesses, aches, and pains were ascribed to malevolent demons until medicine became exorcism. Most of our knowledge of these antidemoniac practices and their extent comes from the huge collection made about 630 B.C. by Ashurbanipal at Nineveh. Literally thousands of extant tablets from this library describe such exorcisms, and thousands more list omen after omen, depicting a decaying civilization as black with demons as a piece of rotting meat with flies.

…and angels, and prophets, and all the other trappings of religion. When the gods spoke to you every day, and you couldn’t get rid of them even if you wanted to, angels – a sort of intermediary with the gods – were unnecessary. There was no place for prophets – when everyone is a prophet, nobody is. There wasn’t even prayer, at least not in a mystical sense – as Jaynes puts it, “schizophrenics do not beg to hear their voices – it is unnecessary – in the few case where this does happen, it is during recovery when the voices are no longer heard with the same frequency.”

The Assyrians invented the idea of Heaven. Previously, Heaven had been unnecessary. You could go visit your god in the local ziggurat, talk to him, ask him for advice. But word went around that gods had retreated to heaven – some of the stories even use those exact words, blaming the Great Flood or some other cataclysm. The ziggurats shifted from houses for the gods to e-temen-an-ki – pedestals that the gods could descend to from Heaven, should they ever wish to return.

By 500 BC, the ability to hear the gods was limited to a few prophets, oracles, and poets. Jaynes is especially interested in this last group – he cites various ancient sources claiming that the poets only transcribe what they hear gods and goddesses sing to them (everyone else wimps out and says this is metaphorical). For Jaynes, the Iliad starts “Sing, O Muse…” because the poet was expecting a hallucinatory Muse to actually appear beside him and start singing, after which he would repeat the song to his listeners as a sort of echolalia.

Jaynes ends by referencing one of my favorite ancient texts, Plutarch’s On The Failure Of Oracles. Plutarch, writing around 100 AD, is not a skeptic. He believes oracles work in theory. But he records a general consensus that they don’t work as well as they used to, and that some day soon they will stop working at all. Jaynes believes that as the theory-of-mind waterline rises, fewer and fewer people hear the voices of the gods. By the Golden Age of Greece, it was so difficult that only a few specially selected people placed in specially numinous locations could manage – the oracles. By Plutarch’s own time, even those people could barely manage.

The last oracle to fade away was the greatest – Delphi, perched atop a fantastic gorge as if suspended between Heaven and Earth. Jaynes tries to give us an impression of how important it was in its time; important people from all over the classical world would make the pilgrimage there, leave lavish gifts, and ask Apollo for advice on weighty matters. He thinks that the oracle’s fame protected it; if a cultural validation is an important ingredient in god-hearing, Delphi had the strongest and best. Its reputation was unimpeachable. Still, in the centuries after Plutarch, its prophecies became rarer and rarer; the Pythia’s few divine utterances became separated by more and more incoherent raving. Finally:

As part of [the Emperor Julian’s] personal quest for authorization, he tried to rehabiliate Delphi in AD 363, three years after it had been ransacked by Constantine. Through his remaining priestess, Apollo prophecied that he would never prophesy again. And the prophecy came true.

IV.

The real Origin Of Consciousness In The Breakdown Of The Bicameral Mind is like my edited version above, except that wherever I say “theory of mind”, it says “consciousness”.

Jaynes has obviously thought a lot about this, and he’s a psychology professor so I’m sure he’s heard of theory of mind. Still, I am so against this choice. Consciousness means so many different things to so many different people, and none of them realize they’re talking past each other, and it’s such a loaded term that any argument including it is basically guaranteed to veer off into the fantastic.

Did he literally believe that the Sumerians, Homeric Greeks, etc were p-zombies? That there was nothing that it was like to be them? That they took in photons and emitted actions but experienced no “mysterious redness of red”? I cannot be completely sure. At times he refers to Bronze Age people as “automatons”, which seems like a pretty final judgment. But he also treats them as genuinely hearing, seeing, and having feelings about the hallucinatory gods who appear to them. The god-human interaction seems like it involves the human being at least minimally conscious. But if Jaynes has a coherent theory here, I must have missed it.

I think he is unaware of (or avoiding) what we would call “the hard problem of consciousness”, and focusing on consciousness entirely as a sort of “global workspace” where many parts of the mind come together and have access to one another. In his theory, that didn’t happen – the mental processing happened and announced itself to the human listener as a divine voice, without the human being aware of the intermediate steps. I can see how “consciousness” is one possible term for this area, if you didn’t mind confusing a lot of people. But seriously, just say “theory of mind”.

Jaynes seems aware of this objection, which he summarizes as “the Bronze Agers did not lack consciousness, they just lacked the concept of consciousness”. His retort is that in some cases, the concept of a thing is the same as the thing itself – he suggests baseball as an example. This seems a little sophistic to me. If somebody told me that Mandarin Chinese doesn’t have a word for “consciousness”, I would be surprised but not stunned – it seems like a strange word for a rich and ancient language to lack, but weirder things have happened. If somebody told me that Chinese people didn’t even have the concept of consciousness until it was introduced from the West, that wouldn’t shock me either – sometimes I think half of philosophers don’t even have the same concept of consciousness I do, and I can imagine the Chinese carving up the world in very different ways. But if someone told me that Chinese people were not conscious, I would dismiss them as a crank. So I can’t accept that having consciousness and having a concept of consciousness are exactly the same thing, and I continue to think “theory of mind” is better here.

The other major difference between my rewrite and Jaynes’ real book is that Jaynes focuses heavily on “bicamerality” – the division of the brain into two hemispheres. He believes that in the Bronze Age mind design, the left hemisphere was the “mortal” and the right hemisphere the “god” – ie the hallucinatory voice of the god was the right hemisphere communicating information to the left hemisphere. In the modern mind design, the two hemispheres are either better integrated, or the right hemisphere just doesn’t do much.

I am not an expert in functional neuroanatomy, but my impression is that recent research has not been kind to any theories too reliant on hemispheric lateralization. While there are a few well-studied examples (language is almost always on the left) and a few vague tendencies (the right brain sort of seems to be more holistic, sometimes), basically all tasks require some input from both sides, there’s little sign that anybody is neurologically more “right-brained” or “left-brained” than anyone else, and most neuroscientific theories don’t care that much about the right-brain left-brain distinction. Also, Michael Gazzaniga’s groundbreaking work on split-brain patients which got everyone excited about hemispheres and is one of the cornerstones of Jaynes’ theory doesn’t replicate. Also, Jaynes says his theory implies that schizophrenic hallucinations come from the language centers of the right hemisphere, and I think the latest fMRI evidence is that they don’t.

(Also, Jaynes says his theory implies that demonic possession occurs in the right hemisphere. But some absolute madman actually put a possessed women in an fMRI machine and then exorcised her while the machine was running and although it showed some odd deficiencies in interhemispheric communication, it didn’t seem to show unusual right hemisphere activity. Imagine having to write that IRB application!)

I don’t think either of these issues fundamentally changes Jaynes’ theory. Just switch “consciousness” to “theory of mind”, and change the psychiatry metaphor from split-brain patients to dissociative-identity patients, and you’re fine.

V.

But there’s another class of problem that Jaynes’ theory doesn’t survive nearly as well: what about Australian Aborigines?

Or American Indians, or Zulus, or Greenland Inuit, or Polynesians, or any other human group presumably isolated from second-millennium-BC Assyrians until anthropologists got a chance to examine them? If consciousness is an invention, and it didn’t spread to these groups, did these groups have it? If so, how? If not, why aren’t they hallucinating gods all the time?

I mean, some of these groups definitely have shamans and medicine men. I’m not saying none of them ever hear gods. But Jaynes claims Bronze Agers heard gods literally all the time, as a substitute for individual thought. Nothing I’ve heard from these people or the anthropologists who study them suggest anything like this is true. And these people also seem to be able to strategically deceive others, another key consciousness innovation Jaynes says Bronze Agers lacked. Or at least I assume I would have heard about it from some anthropologist if they weren’t.

I don’t have a good sense of how Jaynes would answer this objection. The most relevant part of the book is around page 135. Jaynes argues that bicamerality (his term for the hallucinatory gods) started with agriculture in the Bronze Age Near East, though there were scattered hallucinations before then. So plausibly the Inuit, aborigines, etc, were not bicameral. They are in a pre-bicameral state, where they have neither full subjective consciousness, nor clear hallucinations of gods. They may have flashes of both, or do something else entirely, or just be blank. Or something. The point is, if they were perfectly normal conscious people like us, then Jaynes is wrong about everything.

Maybe I’ve done some violence to Jaynes’ theory by rounding it off to “theory of mind” and emphasizing it as an invented technology? But he tries to really emphasize the inventedness of it in the first few chapters, talking about how it had to be built up by layer upon layer of well-chosen metaphor. As far as I can tell I relayed that part faithfully.

And I’m looking at the bulletin board on julianjaynes.org, and there’s a post by someone who met Jaynes before he died and asked him this question. They write:

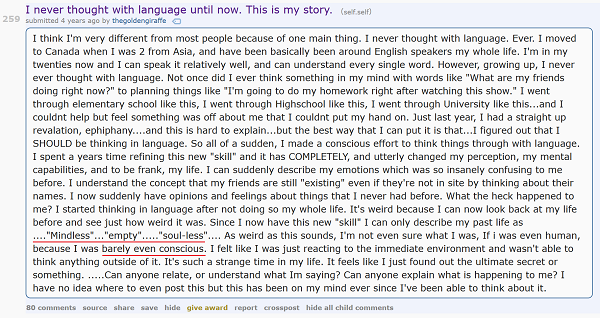
On the About Julian Jaynes page it says he gave a lecture at the Wittgenstein Symposium in [Kirchberg]. I was there. It was a wonderful lecture. It is a pity that his work has not had a deeper impact. I still believe he was basically right (and certainly his prose was brilliant).

I did ask him, by the way, whether he thought it possible that the Aborigines in Australia were not conscious as late as the 18th or 19th century. He said he was not sure and that it would be worthwhile to investigate. Well, I never did and probably no one else [did].

So I don’t think I am misunderstanding him by making this criticism, and it sounds like he just bites the bullet and says maybe this was true. The main position on the forum seems to be that anthropologists weren’t asking the right questions as soon as they met uncontacted tribes, and so maybe they would have missed this. I find this hard to believe. It should be really easy to notice, and also the process of them learning Western theory of mind should leave some scars – at least one of them should say something like “that couple-year period when we all stopped hallucinating gods and became conscious – that was a weird time.”

Jaynes partisans are able to come up with a few anthropological works suggesting that the minds of primitive people are pretty weird, and I believe that, but they don’t seem quite as weird as Jaynes wants them to be. So the question becomes whether we would notice if some people worked in a pre-bicameral and pre-conscious way.

I’m tempted to answer “yes, obviously”, but for the counterargument, see this Reddit thread.



This guy thinks he “barely” had consciousness (in the Jaynesian sense), and it took him however many years to notice this about himself. It was just another universal human experience you can miss without realizing it! And notice how it was the culturally learned knowledge that other people worked differently which shifted him to the normal equilibrium. So maybe if there was some tribe like this somewhere, it would be easy to miss.

I’m also thinking of some cross-cultural psychiatry classes I had to take in residency. It’s well-known that some other cultures rarely get depression and anxiety in the classical Western sense. Instead, in the situations where we would become depressed and anxious, they get psychosomatic complaints, especially stomach pain. This happens to Westerners too sometimes, but in other cultures (eg China, Latin America) it’s by far the most common presentation. This seems similar to Jaynes’ argument that the ancient Greeks talked about feelings in their stomachs when we would talk about thoughts in our minds. I’m not saying these people aren’t conscious or have no theory of mind. But it seems like their theory of mind must be…arranged…differently than ours is, somehow. Or that cultural expectations about how these issues express themselves are shaping the way these issues express themselves, powerfully enough that you can just have whole cultures where depression the way we experience it isn’t a thing. See also this list of culture-bound syndromes. Make sure to read the discussion of Western culture-bound syndromes on the bottom – and make sure to spend a few moments considering what a politically-incorrect person might add to the list.

Even if I don’t accept all the stuff about hallucinatory Athena choreographing the Trojan War, the most important thing I’m going to take away from Origin of Consciousness is that theory of mind is an artifact, not a given, and it’s not necessarily the same everywhere. Much of the way we relate to our mind is culturally determined, and with a different enough cultural environment you can get some weird mind designs in ways that have real effect on behavior. Theory-of-mind-space is wider than we imagine, whether we’re thinking about ancient Sumerians or our ordinary-seeming neighbors.

# Problems With Paywalls

I.

I hate paywalls on articles. Absolutely hate them.

A standard pro-business argument: businesses can either make your life better (by providing deals you like) or keep your life the same (by providing deals you don’t like, which you don’t take). They can’t really make your life worse. There are some exceptions, like if they outcompete and destroy another business you liked better, or if they have some kind of externalities, or if they lobby the government to do something bad. But in general, if you’re angry at a business, you need to explain how one of these unusual conditions applies. Otherwise they’re just “helping you less than you wish they did”, not hurting you.

And so the standard justification for paywalls. Journalists are providing you a deal: you may read their articles in exchange for money. You are not entitled to their product without paying them money. They need to earn a living just like everyone else. So you can either accept their deal – pay money for the articles – or refuse their deal – and so be left no worse off than if they didn’t exist.

But I notice feeling like this isn’t true. I think I would be happier in a world where major newspapers ceased to exist, compared to the world where they exist but their articles are paywalled. Take a second and check if you feel the same way. If so, what could be going on?

First, paywalled newspapers sometimes use a clickbait model, where they start by making you curious what’s in the article, then charge you to find out.

Here are some articles I’ve seen advertised recently (not all on paywalled sources): “Why Trump’s Fight With Obama Might Backfire”, “This Tech Guru Has Made A Shocking Prediction For 2020”, “Here’s Why Men are Pointing Loaded Guns At Their Dicks”.

I didn’t wake up this morning thinking “I wonder whether men are pointing loaded guns at their dicks, and, if so, why. I hope some enterprising journalist has investigated this question, and I will be happy to compensate her with money for satisfying this weird curiosity of mine.” No, instead, I was perfectly and innocently happy not knowing anything about this, right up until I read the name of that article at which point I became consumed with curiosity, ie a feeling that I will be unhappy until I know the answer. In this particular case it’s fine, because the offending website (VICE) is unpaywalled. I go there and after reading through nine paragraphs attacking “MAGA dolts”, in the tenth paragraph I get the one-sentence answer: there’s a meme in the gun community that any time someone posts a picture with their gun, amateurs will chime in with condescending advice about how they should be holding it more safely, so some people post pictures of them pointing loaded guns at their dicks in order to piss these people off. I feel completely unenlightened by knowing this. It has not brightened my day. It just removed the temporary itch of curiosity.

Some people critique capitalism by saying it creates new preferences that people have to spend money to satisfy. I haven’t noticed this being true in general – I only buy shoes when I need shoes, and I only buy Coke when I want Coke. But it seems absolutely on the mark regarding paywalled journalism. VICE created a new preference for me (the preference to know why some people point loaded guns at their dicks), then satisfied it. Overall I have neither gained nor lost utility. This seems different from providing me with a service.

They have an excuse, which is that this is how they make money. But what’s Marginal Revolution’s excuse? I saw this link in an MR links roundup. It was posted as “5. Why men are pointing loaded guns at their dicks.” So obviously I clicked on it, and here we are. But what is MR’s interest in making me click on a VICE article and read through nine paragraphs about “MAGA dolts”?

I can’t really blame them, because I did the same thing for years. I posted links posts, I framed the links in deliberately provocative ways, and then I felt good about myself when my stats page recorded that thousands of people had clicked on them. Sometimes I would write the whole thing out – “Here’s an article about men pointing loaded guns at their dicks – it’s because they want to criticize what they perceive as an excessive and condescending emphasis on trigger safety in gun culture” – and then nobody would click on it, and I would interpret that as a sign that I had failed in some way. I was an idiot, I apologize to all of you, and I have stopped doing that. I urge other bloggers to do the same – we gain no extra money, nor power, nor readership by being running-dogs for VICE’s weird ploy to trick people into reading its stupid articles. But as long as bloggers, Facebookers, tweeters, etc aren’t following good Internet hygiene, the very existence of paywalled sources will continue to be a net negative for the average Internet user.

This isn’t just about obvious clickbait like men pointing guns at their dicks. “Why Trump’s Fight With Obama Might Backfire” feels exactly the same to me. I don’t want to know more ephemeral garbage about Trump which may or may not affect his polls 0.5% for a week before they return to baseline. I don’t want to get more and more outraged until my ability to relate to my fellow human beings is shaped entirely by whether they’re a “MAGA dolt” or not. And yet I find myself curious what’s in the article!

(Trump’s fight with Obama might backfire because independents like Obama more than Trump, and the tech guru’s 2020 prediction was that Trump will lose. You’re welcome.)

Second, paywalled articles become part of the discourse.

Last week’s Wall Street Journal included an opinion column, Lockdowns Vs. The Vulnerable, arguing that statistics show the coronavirus lockdowns do not really prevent the coronavirus, but do disproportionately affect the most vulnerable people. It’s already gotten retweeted a few dozen times, including by some bluechecks with tens of thousands of followers.

The only conclusion available now is that the Swedes were right: The lockdowns were not sustainable, writes @holmanjenkins https://t.co/WFNEa5ScjV via @WSJ

— Charles Gasparino (@CGasparino) May 30, 2020

Do you want to figure out exactly what statistics it uses and check whether they really show that lockdowns don’t prevent coronavirus? Too bad – the article is paywalled and you cannot read it without paying $19.50/month to the Wall Street Journal. I personally suspect that this article is terribly wrong, possibly to the point of idiocy. But I can neither convince others of this, nor correct my own potentially-false first impression, without paying the Wall Street Journal $19.50 a month. Which I don’t want to do. Partly because it is bad value, and partly because I don’t want to reward them for publishing false things.

Newspapers publish articles – factual and opinionated – intending them to enter the public square as a topic of discussion. But if the discussions in the public square have an entry fee, the public square becomes smaller and less diverse.

It also becomes more of an echo chamber. Probably conservatives subscribe to the Wall Street Journal and liberals subscribe to the New York Times. So if conservatives post articles from the Wall Street Journal, liberals can neither benefit from the true ones and change their own opinions, nor correct the false ones and change conservatives’ opinions. If you can’t even read the other side’s arguments, how can you be convinced by them?

Third, newspapers make it hard to guess whether you will encounter a paywall or not. Some of them raise a paywall on some kinds of articles but not others. Some of them raise a paywall if you’re linked in from social media, but not if you’re linked in from Google (or vice versa). Some of them raise a paywall if it’s your Xth article per month on a certain computer, but not before.

The end result is you can’t just learn to avoid the newspapers with paywalls. If you clearly knew which links were paywalled or not, you would just never click on those links, and not waste any time. Since any given newspaper has like a 25 – 50% chance of being paywalled whenever you read it, you get the variable reinforcement strategy that promotes frustrated addiction. And since at any given moment you are desperate to click on that link and find out Why Some Men Are Pointing Loaded Guns At Their Own Dicks, you will, like a chump, click it anyway, only to howl with rage when the paywall comes up.

This usually isn’t a deliberate misdeed; newspapers understandably want to give people limited access so they can decide whether or not they want to subscribe. But some forms of this do seem deliberate to me. Like when they let you read the first two paragraphs and get emotionally invested in the story, and then surprise you with a paywall in the third (I think this is why you need nine paragraphs of filler before getting to the one-sentence curiosity-satisfier). Or when they wait five seconds before a paywall message pops up, for the same reason.

Fourth, and most important, paywalled newspapers make it hard to search for information on Google. When I was trying to gather statistics on coronavirus to figure out how fast it was spreading, I noticed that the top ten or twenty relevant search results for a lot of coronavirus-related queries were paywalled articles. Because articles will make you wait several paragraphs/seconds before the paywall comes up, I couldn’t just quickly click on something, see if it had a paywall or not, and then move on to the next one. Instead, a search that would have taken me seconds if all paywalled sources ceased to exist ended up taking me several frustrating minutes.

II.

There are some simple steps we can take to fix this.

First, search engines should give users an option to hide paywalled articles from results. I realize how big a shitstorm this will cause, and I plan to enjoy every second of it. If they can’t make this happen for some reason, they should at least display a big red $$$ sign in front of paywalled articles, so users know which links will give them information before they waste a click on them. If Google refuses to do this, Bing should do it to get a leg up on Google. If both of them refuse, DuckDuckGo. If all three of them refuse, sounds like they’re providing an opening for some lucky entrepreneur.

Second, browser or browser-extension designers should figure out some way to automatically get links to display whether they’re paywalled or not. Maybe something like this already exists, but I can’t find it.

Third, bloggers (and social media users) should stop deliberately frustrating their readers. Stop posting tantalizing links like “Why Men Are Pointing Loaded Guns At Their Dicks” without further explanation! If you find the dick-gun phenomenon interesting, post the link plus a one-sentence summary. If someone wants more than the one-sentence summary, they can click the link, but I’ve done A/B testing on this and it never happens.

Fourth, bloggers (and social media users) should preferentially link non-paywalled sites. I realize this is not always possible, but most major stories are important enough that at least one non-paywalled outlet will be covering them.

Fifth, until the browser extension comes through, bloggers (and social media users) who do need to link a paywalled site should let readers know it’s paywalled. For example, Lockdowns Vs. The Vulnerable [PAYWALLED] or [$$$] Lockdowns Vs. The Vulnerable. This will save readers a click and hopefully make bloggers think about what they’re doing and whether it’s really necessary.

I’m making a commitment to do 3, 4, and 5 from now on. If I ever change this commitment, I’ll let you know. If you notice me slipping up, please point it out (nicely) and I’ll try to correct myself.

# The Obligatory GPT-3 Post

I.

I would be failing my brand if I didn’t write something about GPT-3, but I’m not an expert and discussion is still in its early stages. Consider this a summary of some of the interesting questions I’ve heard posed elsewhere, especially comments by gwern and nostalgebraist. Both of them are smart people who I broadly trust on AI issues, and both have done great work with GPT-2. Gwern has gotten it to write poetry, compose music, and even sort of play some chess; nostalgebraist has created nostalgebraist-autoresponder (a Tumblr written by GPT-2 trained on nostalgebraist’s own Tumblr output). Both of them disagree pretty strongly on the implications of GPT-3. I don’t know enough to resolve that disagreement, so this will be a kind of incoherent post, and hopefully stimulate some more productive comments. So:

OpenAI has released a new paper, Language Models Are Few-Shot Learners, introducing GPT-3, the successor to the wildly-successful language-processing AI GPT-2.

GPT-3 doesn’t have any revolutionary new advances over its predecessor. It’s just much bigger. GPT-2 had 1.5 billion parameters. GPT-3 has 175 billion. The researchers involved are very open about how it’s the same thing but bigger. Their research goal was to test how GPT-like neural networks scale.

Before we get into the weeds, let’s get a quick gestalt impression of how GPT-3 does compared to GPT-2.

Here’s a sample of GPT-2 trying to write an article:

PROMPT: MILEY CYRUS WAS CAUGHT SHOPLIFTING FROM ABERCROMBIE AND FITCH ON HOLLYWOOD BOULEVARD TODAY

The 19-year-old singer was caught on camera being escorted out of the store by security guards.

The singer was wearing a black hoodie with the label ‘Blurred Lines’ on the front and ‘Fashion Police’ on the back

The singer was also wearing a pair of black-rimmed glasses, a black jacket, black jeans and black sandals.

She was carrying a pair of black and white striped gloves and a small black bag.

And here’s GPT-3 at the same task:

PROMPT: UNITED METHODISTS AGREE TO HISTORIC SPLIT – THOSE WHO OPPOSE GAY MARRIAGE WILL FORM THEIR OWN DENOMINATION

After two days of intense debate, the United Methodist Church has agreed to a historic split – one that is expected to end in the creation of a new denomination, one that will be “theologically and socially conservative,” according to the Washington Post. The majority of delegates attending the church’s annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will “discipline” clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist denomination.

The Post notes that the denomination, which claims 12.5 million members, was in the early 20th century the “largest Protestant denomination in the US”, but that it has been shrinking in recent decades. The new split will be the second in the church’s history. The first occurred in 1968, when roughly 10% of the denomination left to form the Evangelical United Brethren Church. The Post notes that the proposed split “comes at a critical time for the church, which has been losing members for years,” which has been “pushed toward the brink of a schism over the role of LGBTQ people in the church.” Gay marriage is not the only issue that has divided the church. In 2016, the denomination was split over ordination of transgender clergy, with the North Pacific regional conference voting to ban them from serving as clergy, and the South Pacific regional conference voting to allow them.

In case you’re wondering, no, the Methodist Church did not really split from the Evangelical United Brethren Church in 1968, it merged with it in that year. GPT-3 is doing what I used to do on essay questions – throw out a bunch of half-remembered names and dates and hope nobody looks too closely at the exact relations.

These kinds of comparisons are inherently fuzzy, since both GPTs have produced a lot of material and it’s easy to cheat by taking better or worse samples of one or the other. You can see more GPT-2 samples here and more GPT-3 samples here

Or compare some of GPT-2’s poetry:

The gods are they who came to earth  
And set the seas ablaze with gold.  
There is a breeze upon the sea,  
A sea of summer in its folds,  
A salt, enchanted breeze that mocks  
The scents of life, from far away  
Comes slumbrous, sad, and quaint, and quaint.  
The mother of the gods, that day,  
With mortal feet and sweet voice speaks,  
And smiles, and speaks to men: “My Sweet,  
I shall not weary of thy pain.”

…Let me drink of the wine of pain  
And think upon the agonies of hope,  
And of the blessed Giver of all good things;  
For, man for man, mine is the deepest love  
That sorrow takes upon the humblest soul;  
But who hath learned how sorrow turns to gall  
The places where my feet have trod before.

…to some of GPT-3’s:

I must have shadows on the way  
If I am to walk I must have  
Each step taken slowly and alone  
To have it ready made

And I must think in lines of grey  
Must have dim thoughts to be my guide  
Must look on blue and green  
And never let my eye forget  
That color is my friend  
And purple must surround me too

The yellow of the sun is more more  
Intusive than the bluish snow  
That falls on all of us. I must have  
Grey thoughts and blue thoughts walk with me  
If I am to go away at all.

Ignore the style differences – the GPT-2 instance was trained on 18th-century verse, GPT-3 on Wallace Stevens. I notice that the first poem, while flowery, is kind of incoherent. The second has more of a structure and a theme that it maintains throughout. I’m not sure it really symbolizes anything interesting, but I’ve never been sure of that with real Wallace Stevens either, so good work.

(you can read more GPT-2 poetry here. There’s not much more public GPT-3 poetry – you can read one or two more samples on the paper linked above).

Enough writing. What about math? Last year, I posted the following excerpt from GPT-2:

PROMPT: MY 10 FAVORITE ANIMALS ARE:

My ten favorite animals are:  
1. Zebras with a white scar on the back  
2. Insiduous spiders and octopus  
3. Frog with large leaves, hopefully black  
4. Cockatiel with scales  
5. Razorbill with wings hanging about 4 inches from one’s face and a heart tattoo on a frog  
3. Cockatric interlocking tetrabods that can be blind, cut, and eaten raw:  
4. Black and white desert crocodiles living in sunlight  
5. Zebra and many other pea bugs

Despite the obvious flaws in this piece, I was impressed. GPT-2 was clearly trying to make a numbered list, and almost kind of getting it right! It counted to 4 successfully! Remember, this is a text prediction engine that didn’t necessarily need to have any concept of numbers. But it still kind of counted to 4! I wrote:

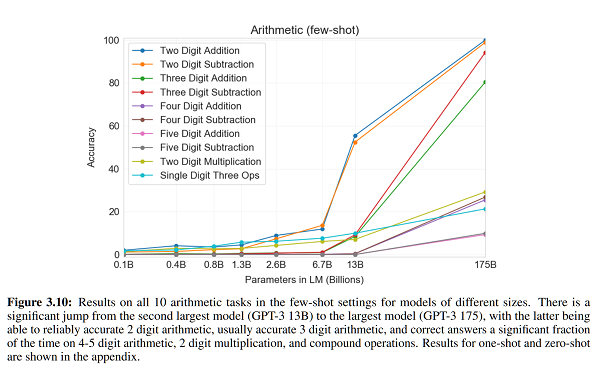
Imagine you prompted the model with “What is one plus one?” I actually don’t know how it would do on this problem. I’m guessing it would answer “two”, just because the question probably appeared a bunch of times in its training data.

Now imagine you prompted it with “What is four thousand and eight plus two thousand and six?” or some other long problem that probably didn’t occur exactly in its training data. I predict it would fail, because this model can’t count past five without making mistakes. But I imagine a very similar program, given a thousand times more training data and computational resources, would succeed. It would notice a pattern in sentences including the word “plus” or otherwise describing sums of numbers, it would figure out that pattern, and it would end up able to do simple math. I don’t think this is too much of a stretch given that GPT-2 learned to count to five and acronymize words and so on.

I said “a very similar program, given a thousand times more training data and computational resources, would succeed [at adding four digit numbers]”. Well, GPT-3 is a very similar program with a hundred times more computational resources, and…it can add four-digit numbers! At least sometimes, which is better than GPT-2’s “none of the time”.

II.

In fact, let’s take a closer look at GPT-3’s math performance.

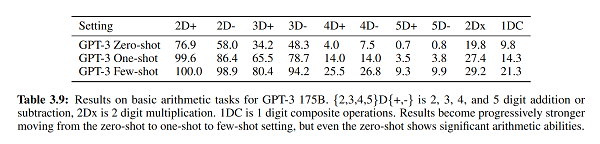


The 1.3 billion parameter model, equivalent to GPT-2, could get two-digit addition problems right less than 5% of the time – little better than chance. But for whatever reason, once the model hit 13 billion parameters, its addition abilities improved to 60% – the equivalent of a D student. At 175 billion parameters, it gets an A+.

What does it mean for an AI to be able to do addition, but only inconsistently? For four digit numbers, but not five digit numbers? Doesn’t it either understand addition, or not?

Maybe it’s cheating? Maybe there were so many addition problems in its dataset that it just memorized all of them? I don’t think this is the answer. There are 100 million possible 4-digit addition problems; seems unlikely that GPT-3 saw that many of them. Also, if it was memorizing its training data, it should have gotten all 100 possible two-digit multiplication problems, but it only has about a 25% success rate on those. So it can’t be using a lookup table.

Maybe it’s having trouble locating addition rather than doing addition? (thanks to nostalgebraist for this framing). This sort of seems like the lesson of Table 3.9:



“Zero-shot” means you just type in “20 + 20 = ?”. “One-shot” means you give it an example first: “10 + 10 = 20. 20 + 20 = ?” “Few-shot” means you give it as many examples as it can take. Even the largest and best model only does mediocre on the zero-shot task, but it does better on the one-shot and best on the few-shot. So it seems like if you remind it what addition is a couple of times before solving an addition problem, it does better. This suggests that there is a working model of addition somewhere within the bowels of this 175 billion parameter monster, but it has a hard time drawing it out for any particular task. You need to tell it “addition” “we’re doing addition” “come on now, do some addition!” up to fifty times before it will actually deploy its addition model for these problems, instead of some other model. Maybe if you did this five hundred or five thousand times, it would excel at the problems it can’t do now, like adding five digit numbers. But why should this be so hard? The plus sign almost always means addition. “20 + 20 = ?” is not some inscrutable hieroglyphic text. It basically always means the same thing. Shouldn’t this be easy?

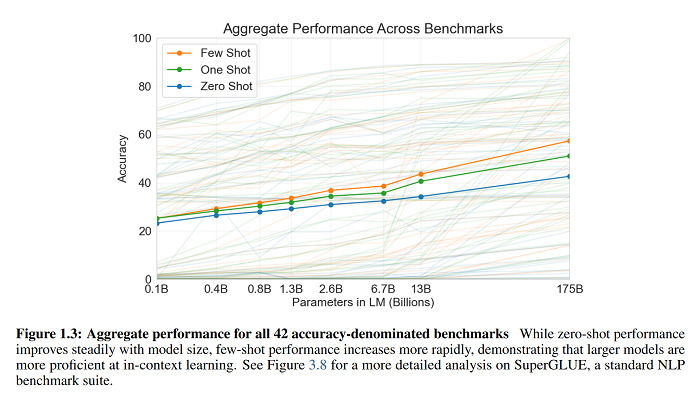
When I prompt GPT-2 with addition problems, the most common failure mode is getting an answer that isn’t a number. Often it’s a few paragraphs of text that look like they came from a math textbook. It feels like it’s been able to locate the problem as far as “you want the kind of thing in math textbooks”, but not as far as “you want the answer to the exact math problem you are giving me”. This is a surprising issue to have, but so far AIs have been nothing if not surprising. Imagine telling Marvin Minsky or someone that an AI smart enough to write decent poetry would not necessarily be smart enough to know that, when asked “325 + 504”, we wanted a numerical response!

Or maybe that’s not it. Maybe it has trouble getting math problems right consistently for the same reason I have trouble with this. In fact, GPT-3’s performance is very similar to mine. I can also add two digit numbers in my head with near-100% accuracy, get worse as we go to three digit numbers, and make no guarantees at all about four-digit. I also find multiplying two-digit numbers in my head much harder than adding those same numbers. What’s my excuse? Do I understand addition, or not? I used to assume my problems came from limited short-term memory, or from neural noise. But GPT-3 shouldn’t have either of those issues. Should I feel a deep kinship with GPT-3? Are we both minds heavily optimized for writing, forced by a cruel world to sometimes do math problems? I don’t know.

[EDIT: an alert reader points out that when GPT-3 fails at addition problems, it fails in human-like ways – for example, forgetting to carry a 1.]

III.

GPT-3 is, fundamentally, an attempt to investigate scaling laws in neural networks. That is, if you start with a good neural network, and make it ten times bigger, does it get smarter? How much smarter? Ten times smarter? Can you keep doing this forever until it’s infinitely smart or you run out of computers, whichever comes first?



So far the scaling looks logarithmic – a consistent multiplication of parameter number produces a consistent gain on the benchmarks.

Does that mean it really is all about model size? Should something even bigger than GPT-3 be better still, until eventually we have things that can do all of this stuff arbitrarily well without any new advances?

This is where my sources diverge. Gwern says yes, probably, and points to years of falsified predictions where people said that scaling might have worked so far, but definitely wouldn’t work past this point. Nostalgebraist says maybe not, and points to decreasing returns of GPT-3’s extra power on certain benchmarks (see Appendix H) and to this OpenAI paper, which he interprets as showing that scaling should break down somewhere around or just slightly past where GPT-3 is. If he’s right, GPT-3 might be around the best that you can do just by making GPT-like things bigger and bigger. He also points out that although GPT-3 is impressive as a general-purpose reasoner that has taught itself things without being specifically optimized to learn them, it’s often worse than task-specifically-trained AIs at various specific language tasks, so we shouldn’t get too excited about it being close to superintelligence or anything. I guess in retrospect this is obvious – it’s cool that it learned how to add four-digit numbers, but calculators have been around a long time and can add much longer numbers than that.

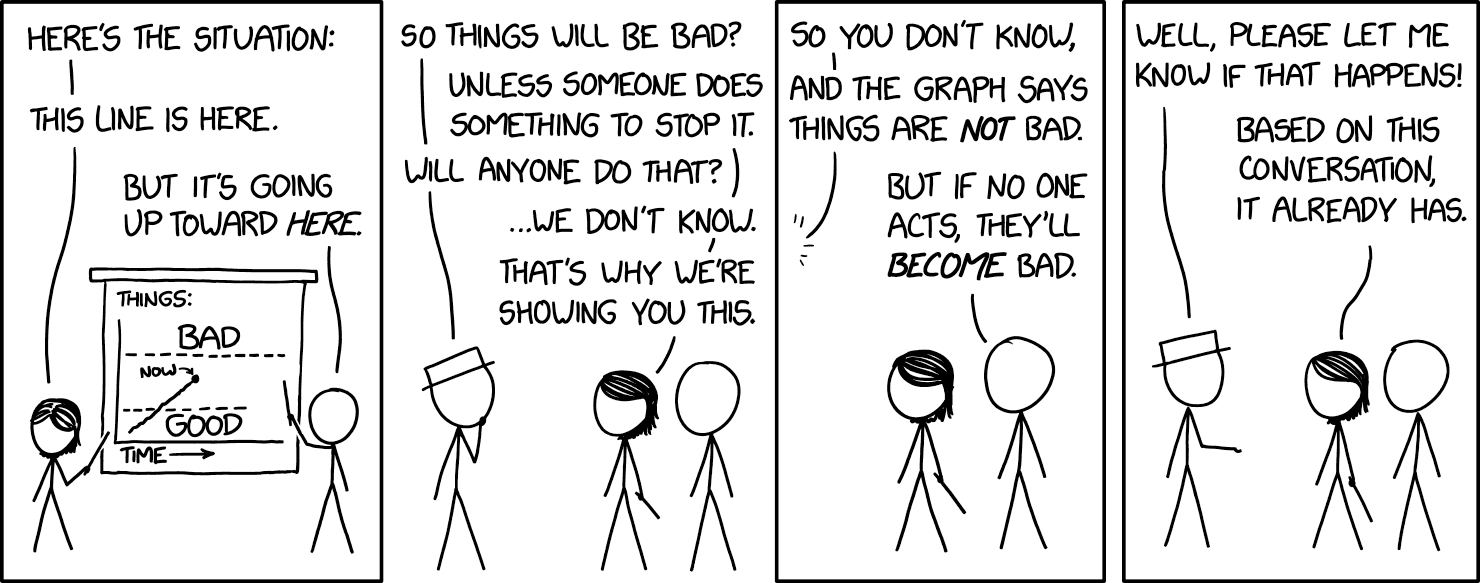
If the scaling laws don’t break down, what then?

GPT-3 is very big, but it’s not pushing the limits of how big an AI it’s possible to make. If someone rich and important like Google wanted to make a much bigger GPT, they could do it.

GPT-3 is terrifying because it's a tiny model compared to what's possible, trained in the dumbest way possible on a single impoverished modality on tiny data, yet the first version already manifests crazy runtime meta-learning—and the scaling curves 𝘴𝘵𝘪𝘭𝘭 are not bending! 😮 https://t.co/hQbW9znm3x

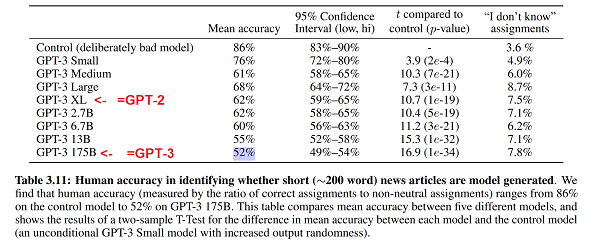
— 𝔊𝔴𝔢𝔯𝔫 (@gwern) May 31, 2020

Does “terrifying” sound weirdly alarmist here? I think the argument is something like this. In February, we watched as the number of US coronavirus cases went from 10ish to 50ish to 100ish over the space of a few weeks. We didn’t panic, because 100ish was still a very low number of coronavirus cases. In retrospect, we should have panicked, because the number was constantly increasing, showed no signs of stopping, and simple linear extrapolation suggested it would be somewhere scary very soon. After the number of coronavirus cases crossed 100,000 and 1,000,000 at exactly the time we could have predicted from the original curves, we all told ourselves we definitely wouldn’t be making that exact same mistake again.



It’s always possible that the next AI will be the one where the scaling curves break and it stops being easy to make AIs smarter just by giving them more computers. But unless something surprising like that saves us, we should assume GPT-like things will become much more powerful very quickly.

What would much more powerful GPT-like things look like? They can already write some forms of text at near-human level (in the paper above, the researchers asked humans to identify whether a given news article had been written by a human reporter or GPT-3; the humans got it right 52% of the time)



So one very conservative assumption would be that a smarter GPT would do better at various arcane language benchmarks, but otherwise not be much more interesting – once it can write text at a human level, that’s it.

Could it do more radical things like write proofs or generate scientific advances? After all, if you feed it thousands of proofs, and then prompt it with a theorem to be proven, that’s a text prediction task. If you feed it physics textbooks, and prompt it with “and the Theory of Everything is…”, that’s also a text prediction task. I realize these are wild conjectures, but the last time I made a wild conjecture, it was “maybe you can learn addition, because that’s a text prediction task” and that one came true within two years. But my guess is still that this won’t happen in a meaningful way anytime soon. GPT-3 is much better at writing coherent-sounding text than it is at any kind of logical reasoning; remember it still can’t add 5-digit numbers very well, get its Methodist history right, or consistently figure out that a plus sign means “add things”. Yes, it can do simple addition, but it has to use supercomputer-level resources to do so – it’s so inefficient that it’s hard to imagine even very large scaling getting it anywhere useful. At most, maybe a high-level GPT could write a plausible-sounding Theory Of Everything that uses physics terms in a vaguely coherent way, but that falls apart when a real physicist examines it.

Probably we can be pretty sure it won’t take over the world? I have a hard time figuring out how to turn world conquest into a text prediction task. It could probably imitate a human writing a plausible-sounding plan to take over the world, but it couldn’t implement such a plan (and would have no desire to do so).

For me the scary part isn’t the much larger GPT we’ll probably have in a few years. It’s the discovery that even very complicated AIs get smarter as they get bigger. If someone ever invented an AI that did do more than text prediction, it would have a pretty fast takeoff, going from toy to superintelligence in just a few years.

Speaking of which – can anything based on GPT-like principles ever produce superintelligent output? How would this happen? If it’s trying to mimic what a human can write, then no matter how intelligent it is “under the hood”, all that intelligence will only get applied to becoming better and better at predicting what kind of dumb stuff a normal-intelligence human would say. In a sense, solving the Theory of Everything would be a failure at its primary task. No human writer would end the sentence “the Theory of Everything is…” with anything other than “currently unknown and very hard to figure out”.

But if our own brains are also prediction engines, how do we ever create things smarter and better than the ones we grew up with? I can imagine scientific theories being part of our predictive model rather than an output of it – we use the theory of gravity to predict how things will fall. But what about new forms of art? What about thoughts that have never been thought before?

And how many parameters does the adult human brain have? The responsible answer is that brain function doesn’t map perfectly to neural net function, and even if it did we would have no idea how to even begin to make this calculation. The irresponsible answer is a hundred trillion. That’s a big number. But at the current rate of GPT progress, a GPT will have that same number of parameters somewhere between GPT-4 and GPT-5. Given the speed at which OpenAI works, that should happen about two years from now.

I am definitely not predicting that a GPT with enough parameters will be able to do everything a human does. But I’m really interested to see what it can do. And we’ll find out soon.

# Wordy Wernicke’s

There are two major brain areas involved in language. To oversimplify, Wernicke’s area in the superior temporal gyrus handles meaning; Broca’s area in the inferior frontal gyrus handles structure and flow.

If a stroke or other brain injury damages Broca’s area but leaves Wernicke’s area intact, you get language which is meaningful, but not very structured or fluid. You sound like a caveman: “Want food!”

If it damages Wernicke’s area but leaves Broca’s area intact, you get speech which has normal structure and flow, but is meaningless. I’d read about this pattern in books, but I still wasn’t prepared the first time I saw a video of a Wernicke’s aphasia patient (source):

During yesterday’s discussion of GPT-3, a commenter mentioned how alien it felt to watch something use language perfectly without quite making sense. I agree it’s eerie, but it isn’t some kind of inhuman robot weirdness. Any one of us is a railroad-spike-through-the-head away from doing the same.

Does this teach us anything useful about GPT-3 or neural networks? I lean towards no. GPT-3 already makes more sense than a Wernicke’s aphasiac. Whatever it’s doing is on a higher level than the Broca’s/Wernicke’s dichotomy. Still, it would be interesting to learn what kind of computational considerations caused the split, and whether there’s any microstructural difference in the areas that reflects it. I don’t know enough neuroscience to have an educated opinion on this.

# The Vision Of Vilazodone And Vortioxetine

I.

One of psychiatry’s many embarrassments is how many of our drugs get discovered by accident. They come from random plants or shiny rocks or stuff Alexander Shulgin invented to get high.

But every so often, somebody tries to do things the proper way. Go over decades of research into what makes psychiatric drugs work and how they could work better. Figure out the hypothetical properties of the ideal psych drug. Figure out a molecule that matches those properties. Synthesize it and see what happens. This was the vision of vortioxetine and vilazodone, two antidepressants from the early 2010s. They were approved by the FDA, sent to market, and prescribed to millions of people. Now it’s been enough time to look back and give them a fair evaluation. And…

…and it’s been a good reminder of why we don’t usually do this.

Enough data has come in to be pretty sure that vortioxetine and vilazodone, while effective antidepressants, are no better than the earlier medications they sought to replace. I want to try going over the science that led pharmaceutical companies to think these two drugs might be revolutionary, and then speculate on why they weren’t. I’m limited in this by my total failure to understand several important pieces of the pathways involved, so I’ll explain the parts I get, and list the parts I don’t in the hopes that someone clears them up in the comments.

II.

SSRIs take about a month to work. This is surprising, because they start increasing serotonin immediately. If higher serotonin is associated with improved mood, why the delay?

Most research focuses on presynaptic 5-HT1A autoreceptors, which detect and adjust the amount of serotonin being released in order to maintain homeostasis. If that sounds clear as mud to you, well, it did to me too the first twenty times I heard it. Here’s the analogy which eventually worked for me:

Imagine you’re a salesperson who teleconferences with customers all day. You’re bad at projecting your voice – sometimes it’s too soft, sometimes too loud. Your boss tells you the right voice level for sales is 60 decibels. So you put an audiometer on your desk that measures how many decibels your voice is. It displays an up arrow, down arrow, or smiley face, telling you that you’re being too quiet, loud, or just right.

The audiometer is presynaptic (it’s on your side of the teleconference, not your customer’s side). It’s an autoreceptor, not a heteroreceptor (you’re using it to measure yourself, not to measure anything else). And you’re using it to maintain homeostasis (to keep your voice at 60 dB).

Suppose you take a medication that stimulates your larynx and makes your voice naturally louder. As long as your audiometer’s working, that medication will have no effect. Your voice will naturally be louder, but that means you’ll see the down arrow on your audiometer more often, you’ll speak with less force, and the less force and louder voice will cancel out and keep you at 60 dB. No change.

This is what happens with antidepressants. The antidepressant increases serotonin levels (ie sends a louder signal). But if the presynaptic 5-HT1A autoreceptors are intact, they tell the cells that the signal is too loud and they should release less serotonin. So they do, and now we’re back where we started – ie depressed.

(Why don’t the autoreceptors notice the original problem – that you have too little serotonin and are depressed – and work their magic there? Not sure. Maybe depression affects whatever sets the autoreceptors, and causes them to be set too low? Maybe the problem is with reception, not transmission? Maybe you have the right amount of serotonin, but you want excessive amounts of serotonin because that would fix a problem somewhere else? Maybe you should have paid extra for the premium model of presynaptic autoreceptor?)

So how come antidepressants work after a month? The only explanation I’ve heard is that the autoreceptors get “saturated”, which is a pretty nonspecific term. I think it means that there’s a second negative feedback loop controlling the first negative feedback loop – the cell notices there is way more autoreceptor activity than expected and assumes it is producing too many autoreceptors. Over the course of a month, it stops producing more autoreceptors, the existing autoreceptors gradually degrade, and presynaptic autoreception stops being a problem. In our salesperson analogy, after a while you notice that your audiometer doesn’t match your own perception of how much effort you’re putting into speaking – no matter how quietly you feel like you’re whispering, the audiometer just keeps saying you’re yelling very loud. Eventually you declare it defective, throw it out, and just speak naturally. Now the larynx-stimulating medication can make your voice louder.

In the late ’90s, some scientists wondered – what happens if you just block the 5-HT1A autoreceptors directly? At the very least, seems like you could make antidepressants work a month faster. Best case scenario, you can make antidepressants work better. Maybe after a month, the cells have lost some confidence in the autoreceptors, but they’re still keeping serotonin somewhere below the natural amount based on the anomalous autoreceptor reading. So maybe blocking the autoreceptors would mean a faster, better, antidepressant.

Luckily we already knew a chemical that could block these – pindolol. Pindolol is a blood-pressure-lowering medication, but by coincidence it also makes it into the brain and blocks this particular autoreceptor involved in serotonin homeostasis. So a few people started giving pindolol along with antidepressants. What happened? According to some small unconvincing systematic reviews, it did seem to kind of help make antidepressants work faster, but according to some other small unconvincing meta-analyses it probably didn’t make them work any better. It just made antidepressants go from taking about four weeks to work, to taking one or two weeks to work. It also gave patients dizziness, drowsiness, weakness, and all the other things you would expect from giving a blood-pressure-lowering medication to people with normal blood pressure. So people decided it probably wasn’t worth it.

III.

And then there’s buspirone.

Buspirone is generally considered a 5-HT1A agonist. The reality is a little more complicated; it’s a full agonist of presynaptic autoreceptors, and a partial agonist of postsynaptic autoreceptors(uh, imagine the customer in the other side of the teleconference also has an audiometer).

Buspirone has a weak anti-anxiety effect. It may also weakly increase sex drive, which is a nice contrast to SSRIs, which decrease (some would say “destroy”) sex drive. Flibanserin, a similar drug, is FDA-approved as a sex drive enhancer.

Buspirone stimulates presynaptic autoreceptors, which should cause cells to release less serotonin. Since high serotonin levels (eg with SSRIs) decreases sex drive, it makes sense that buspirone should increase sex drive. So far, so good.

But why is buspirone anxiolytic? I have just read a dozen papers purporting to address this question, and they might as well have been written in Chinese for all I was able to get from them. Some of them just say that decreasing serotonin levels decreases anxiety, which would probably come as a surprise to anyone on SSRIs, anyone who does tryptophan depletion studies, anyone who measures serotonin metabolites in the spinal fluid, etc. Obviously it must be more complicated than this. But how? I can’t find any explanation.

A few books and papers take a completely different tack and argue that buspirone just does the same thing SSRIs do – desensitize the presynaptic autoreceptors until the cells ignore them – and then has its antianxiety action through stimulating postsynaptic receptors. But if it’s doing the same thing as SSRIs, how come it has the opposite effect on sex drive? And how come there are some studies suggesting that it’s helpful to add buspirone onto SSRIs? Why isn’t that just doubling up on the same thing?

I am deeply grateful to SSC commenter Scchm for presenting an argument that this is all wrong, and buspirone acts on D4 receptors, both in its anxiolytic and pro-sexual effects. This would neatly resolve the issues above. But then how come nobody else mentions this? How come everyone else seems to think buspirone makes sense, and writes whole papers about it without using the sentence “what the hell all of this is crazy”?

And here’s one more mystery: after the pindolol studies, everyone just sort of started assuming buspirone would work the same way pindolol did. This doesn’t really make sense pharmacologically – pindolol is an antagonist at presynaptic 5-HT1A receptors; buspirone is an agonist of same. And it doesn’t actually work in real life – someone did a study, and the study found it didn’t work. Still, and I have no explanation for this, people got excited about this possibility. If buspirone could work like pindolol, then we would have a chemical that made antidepressants work faster, and treated anxiety, and reduced sexual dysfunction.

And here’s one more mystery – okay, you have unrealistically high expectations for buspirone, fine, give people buspirone along with their SSRI. Lots of psychiatrists do this, it’s not really my thing, but it’s not a bad idea. But instead, the people making this argument became obsessed with the idea of finding a single chemical that combined SSRI-like activity with buspirone-like activity. A dual serotonin-transporter-inhibitor and 5-HT1A partial agonist became a pharmacological holy grail.

IV.

After a lot of people in lab coats poured things from one test tube to another, Merck announced they had found such a chemical, which they called vilazodone (Viibryd®).

Vilazodone is an SSRI and 5-HT1A partial agonist. I can’t find how its exact partial agonist profile differs from buspirone, except that it’s more of a postsynaptic agonist, whereas buspirone is more of a postsynaptic antagonist. I don’t know if this makes a difference.

The FDA approved vilazodone to treat depression, so it “works” in that sense. But does its high-tech promise pan out? Is it really faster-acting, anxiety-busting, and less likely to cause sexual side effects?

On a chemical level, things look promising. This study finds that vilazodone elevates serotonin faster and higher than Prozac does in mice. And this is kind of grim, but toxicologists have noticed that vilazodone overdoses are much more likely to produce serotonin toxicity than Prozac overdoses, which fits what you would expect if vilazodone successfully breaks the negative feedback system that keeps serotonin in a normal range.

On a clinical level, maybe not. Proponents of vilazodone got excited about a study where vilazodone showed effects as early as week two. But “SSRIs take four weeks to work” is a rule of thumb, not a natural law. You always get a couple of people who get some effect early on, and if your study population is big enough, that’ll show up as a positive result. So you need to compare vilazodone to an SSRI directly. The only group I know who tried this, Matthews et al, found no difference – in fact, vilazodone was nonsignificantly slower (relevant figure). There’s no sign of vilazodone working any better either.

What about sexual side effects? Vilazodone does better than SSRIs in rats, but whatever. There’s supposedly a human study – from the same Matthews et al team as above – but sexual side effects were so rare in all groups that it’s hard to draw any conclusions. This is bizarre – they had a thousand patients, and only 15 reported decreased libido (and no more in the treatment groups than the placebo group). Maybe God just hates antidepressant studies and makes sure they never find anything, and this is just as true when you’re studying side effects as it is when you’re studying efficacy.

Clayton et al do their own study of vilazodone’s sexual side effects. They find that overall vilazodone improves sexual function over placebo, probably because they used a scale that was very sensitive to the kind of bad sexual function you get when you’re depressed, and not as sensitive to the kind you get on antidepressants. But they did measure how many people who didn’t start out with sexual dysfunction got it during the trial, and this number was 1% of the placebo group and 8% of the vilazodone group. How many people would have gotten dysfunction on an SSRI? We don’t know because they didn’t include an active comparator. Usually I expect about 30 – 50% of people to get sexual side effects on SSRIs, but that’s based on me asking them and not on whatever strict criteria they use for studies. Remember, the Matthews study was able to find only 1.5% of people getting sexual side effects! So we shouldn’t even try to estimate how this compares. All we can say is that vilazodone definitely doesn’t have no sexual side effects.

I can’t find any studies evaluating vilazodone vs. anything else for anxiety, but I also can’t find any patients saying vilazodone treated their anxiety especially well.

There’s really only one clear and undeniable difference between vilazodone and ordinary SSRIs, which is that vilazodone costs $290 a month, whereas other SSRIs cost somewhere in the single digits (Lexapro costs $7.31). If you’re paying for vilazodone, you can take comfort in knowing your money helped fund a pretty cool research program that had some interesting science behind it. But I’m not sure it actually panned out.

V.

Encouraged by Merck’s success…

(not necessarily clinical success, success at getting people to pay $290 a month for an antidepressant)

…Takeda and Lundbeck announced their own antidepressant with 5-HT1A partial agonist action, vortioxetine. They originally gave it the trade name Brintellix®, but upon its US release people kept confusing it with the unrelated medication Brilinta®, so Takeda/Lundbeck agreed to change the name to Trintellix® for the American market.

Vortioxetine claimed to have an advantage over its competitor vilazodone, in that it also antagonized 5-HT3 receptors. 5-HT3 receptors are weird. They’re the only ion channel based serotonin receptors, and they’re not especially involved in mood or anxiety. They do only one thing, and they do it well: they make you really nauseous. If you’ve ever felt nauseous on an SSRI, 5-HT3 agonism is why. And if you’ve ever taken Zofran (ondansetron) for nausea, you’ve benefitted from its 5-HT3 antagonism. Most antidepressants potentially cause nausea; since vortioxetine also treats nausea, presumably you break even and are no more nauseous than you were before taking it. Also, there are complicated theoretical reasons to believe maybe 5-HT3 antagonism is kind of like 5-HT1A antagonism in that it speeds the antidepressant response.

After this Takeda and Lundbeck kind of just went crazy, claiming effects on more and more serotonin receptors. It’s a 5-HT7 antagonist! (what is 5-HT7? No psychiatrist had ever given a second’s thought to this receptor before vortioxetine came out, but apparently it…exists to make your cognition worse, so that blocking it makes your cognition better again?) It’s a 5-HT1B partial agonist! (what is 5-HT1B? Apparently a useful potential depression target, according to a half-Japanese, half-Scandinavian team, who report no conflict of interest even though vortioxetine is being sold by a consortium of a Japanese pharma company and a Scandinavian pharma company). It’s a 5-HT1D antagonist! (really? There are four different kinds of 5-HT1 receptor? Are you sure you’re not just making things up now?)

If we take all of this seriously, vortioxetine is an SSRI with faster mechanism of action, fewer sexual side effects, additional anti-anxiety effect, additional anti-nausea effect, plus it gives you better cognition (technically “relieves the cognitive symptoms of depression”). Is any of this at all true?

A meta-analysis of 12 studies finds vortioxetine has a statistically significant but pathetic effect size of 0.2 against depression, which is about average for antidepressants. In a few head-to-head comparisons with SNRIs (similar to SSRIs), vortioxetine treats depression about equally well. Patients are more likely to stop the SNRIs because of side effects than to stop the vortioxetine, but SNRIs probably have more side effects than SSRIs, so unclear if vortioxetine is better than those. Wagner et al are able to find a study comparing vortioxetine to the SSRI Paxil; they work about equally well.

What about the other claims? Weirdly, vortioxetine patients have more nausea and vomiting than venlafaxine patients, although it’s not significant. Other studies confirm nausea is a pretty serious vortioxetine side effect. I have no explanation for this. Antagonizing 5-HT3 receptors does one thing – treats nausea and vomiting! – and vortioxetine definitely does this. It must be hitting some other unknown receptor really hard, so hard that the 5-HT3 antagonism doesn’t counterbalance it. Either that, or it’s the thing where God hates antidepressants again.

What about sexual dysfunction? Jacobsen et al find that patients have slightly (but statistically significantly) less sexual dysfunction on vortioxetine than on escitalopram. But the study was done by Takeda, and the difference is so slight (a change of 8.8 points on a 60 point scale, vs. a change of 6.6 points) that it’s hard to take it very seriously.

What about cognition? I was sure this was fake, but it seems to have more evidence behind it than anything else. Carlat Report (paywalled) thinks it might be legit, based on a series of (Takeda-sponsored) studies of performance on the Digit Symbol Substitution Test. People on vortioxetine consistently did better on this test than people on placebo or duloxetine. And it wasn’t just that being not-depressed helps you try harder; they did some complicated statistics and found that vortioxetine’s test-score-improving effect was independent of its antidepressant effect (and seems to work at lower doses). What’s the catch? The improvement was pretty minimal, and only shows up on this one test – various other cognitive tests are unaffected. So it’s probably doing something measurable, but it’s not going to give you a leg up on the SAT. The FDA seriously considered approving it as indicated for helping cognition, but eventually decided against it on the grounds that if they approved it, people would think it was useful in real life, whereas all we know is that it’s useful on this one kind of hokey test. Still, that’s one hokey test more than vilazodone was ever able to show for itself.

In summary, vortioxetine probably treats depression about as well as any other antidepressant, but makes you slightly more nauseous, may (if you really trust pharma company studies) give you slightly fewer sexual side effects, and may improve your performance on the Digit Symbol Substitution Test. It also costs $375 a month (Lexapro still costs $7). If you want to pay $368 extra to be a little more nauseous and substitute digits for symbols a little faster, this is definitely the drug for you.

VI.

In conclusion, big pharma spent about ten years seeing if combining 5-HT1A partial agonism with SSRI antidepressants led to any benefits. In the end, it didn’t, unless you count benefits to big pharma’s bottom line.

I’m a little baffled, because pharma companies generally don’t waste money researching drugs unless they have very good theoretical reasons to think they’ll work. But I can’t make heads or tails of the theoretical case for 5-HT1A partial agonists for depression.

For one thing, there’s a pretty strong argument that buspirone exerts its effects via dopamine rather than serotonin – a case that it seems like nobody, including the pharma companies, is even slightly aware of. If this were true, the whole project would have been doomed from the beginning. What happened here?

For another, I still don’t get the supposed model for how buspirone even could exert its effects through 5-HT1A. Does it increase or decrease serotonergic transmission? Does it desensitize presynaptic autoreceptors the same way SSRIs do, or do something else? I can’t figure out a combination of answers to this question that are consistent with each other and with the known effects of these drugs. Is there one?

For another, the case seems to have been premised on the idea that buspirone (a presynaptic 5-HT1A agonist) would work the same as pindolol (a presynaptic 5-HT1A antagonist), even after studies showed that it didn’t. And then it combined that with an assumption that it was better to spend hundreds of millions of dollars discovering a drug that combined SSRI and buspirone-like effects, rather than just giving someone a pill of SSRI powder mixed with buspirone powder. Why?

I would be grateful if some friendly pharmacologist reading this were to comment with their take on these questions. This is supposed to be my area of expertise, and I have to admit I am stumped.

# Slightly Skew Systems Of Government

[Related To: Legal Systems Very Different From Ours Because I Just Made Them Up, List Of Fictional Drugs Banned By The FDA]

I.

Clamzoria is an acausal democracy.

The problem with democracy is that elections happen before the winning candidate takes office. If somebody’s never been President, how are you supposed to judge how good a President they’d be? Clamzoria realized this was dumb, and moved elections to the last day of an official’s term.

When the outgoing President left office, the country would hold an election. It was run by approval voting: you could either approve or disapprove of the candidate who had just held power. The results were tabulated, announced, and then nobody ever thought about them again.

Clamzoria chose its officials through a prediction market. The Central Bank released bonds for each candidate, which paid out X dollars at term’s end, where X was the percent of voters who voted Approve. Traders could provisionally buy and sell these bonds. On the first day of the term, whichever candidate’s bonds were trading at the highest value was inaugurated as the new President; everyone else’s bonds were retroactively cancelled and their traders refunded. The President would spend a term in office, the election would be held, and the bondholders would be reimbursed the appropriate amount.

The Clamzorians argued this protected against demagoguery. It’s easy for a candidate to promise the sun and moon before an election, but by the end of their term, voters know if the country is doing well or not. Instead of running on a platform of popular (but doomed) ideas, candidates are encouraged to run on a platform of unpopular ideas, as long as those unpopular ideas will genuinely make the country richer, safer, stronger, and all the other things that lead people to approve of a President’s term after the fact. Of course, you’re still limited by bond traders’ ability to predict which policies will work, but bond traders are usually more sober than the general electorate.

This system worked wonderfully for several decades, until Lord Bloodholme’s administration. He ran for President on an unconventional platform: if elected, he would declare himself Dictator-For-Life, replace democracy with sham elections, and kill all who opposed him. Based on his personality, all the bond traders found this completely believable. But that meant that in the end-of-term election, he would get 100% approval. His bond shot up to be worth nearly $100, the highest any bond had ever gone, and he won in a landslide. Alas, Lord Bloodholme was as good as his word, and – after a single sham election to ensure the bondholders got what they were due – that was the end of Clamzoria’s acausal democracy.

II.

Cognito is a constitutional mobocracy.

It used to be a regular mobocracy. It had a weak central government, radicals would protest whenever they didn’t like its decisions, the protests would shut down major cities, and the government would cave. Then people on the other side would protest, and that would also shut down major cities, and the government would backtrack. Eventually they realized they needed a better way, made a virtue out of necessity, and wrote the whole system into their constitution.

The Executive Branch is a president elected by some voting system that basically ensures a bland moderate. They have limited power to make decrees that enforce the will of the legislature. The legislature is the mob. One proposes a bill by having a protest in favor of it. If the protest attracts enough people – the most recent number is 43,617, but it changes every year based on the population and a few other factors – then the bill is considered up for review. Anyone can propose amendments (by having a protest demanding amendments) or vote against it – (by having a protest larger than the original protest demanding that the bill not be passed). After everyone has had a fair chance to protest, the text of the bill supported by the largest protest becomes law (unless the largest protest was against any change, in which case there is no change).

The Cognitans appreciate their system because protests are peaceful and nondisruptive. The government has a specific Protesting Square in every city with a nice grid that lets them count how many protesters there are, and all protests involve going into the Protesting Square, standing still for a few minutes to let neutral observers count people up, and then going home. It’s silly to protest beyond this; your protest wouldn’t be legally binding!

There’s been some concern recently that corprorations pay protesters to protest for things they want. Several consumer watchdog organizations are trying to organize mobs in favor of a bill to stop this.

III.

Yyphrostikoth is a meta-republic.

Every form of government has its own advantages and disadvantages, and the goal is to create a system of checks and balances where each can watch over the others. The Yyphrostikoth Governing Council has twelve members:

The Representative For Monarchy is a hereditary position.

The Representative For Democracy is elected.

The Representative For Plutocracy is the richest person in the country.

The Representative For Technocracy is chosen by lot from among the country’s Nobel Prize winners.

The Representative For Meritocracy is whoever gets the highest score on a standardized test of general knowledge and reasoning ability.

The Representative For Military Dictatorship is the top general in the army.

The Representative For Communism is the leader of the largest labor union.

The Representative For Futarchy is whoever has the best record on the local version of Metaculus.

The Representative For Gerontocracy is supposedly the oldest person in the country who is medically fit and willing to serve, but this has been so hard to sort out that in practice they are selected by the national retirees’ special interest group from the pool of willing candidates above age 90.

The Representative For Minarchy is an honorary position usually bestowed upon a respected libertarian philosopher or activist. It doesn’t really matter who holds it, because their only job is to vote “no” on everything, except things that are sneakily phrased so that “no” means more government, in which case they can vote “yes”. If a Representative For Minarchy wants to vote their conscience, they may break this rule once, after which they must resign and be replaced by a new Representative.

The Representative For Republicanism is selected by the other eleven members of the council.

The Representative For Theocracy is the leader of the Governing Council, and gets not only her own vote but a special vote to break any ties. She is chosen at random from a lottery of all adult citizens, on the grounds that God may pick whoever He pleases to represent Himself.

Long ago, the twelfth Councilor was the Representative For Kratocracy (rule by the strongest). The Representative For Kratocracy was whoever was sitting in the Representative For Kratocracy’s chair when a vote took place. This usually involved a lot of firefights and hostage situations, which was fine in principle – that was the whole point – except that the rest of the Governing Council kept getting caught in the crossfire. During the Nehanian Restoration, the Representative For Kratocracy’s chair was moved to a remote uninhabited island, with the Representative permitted to vote by video-link, but environmentalist groups complained that the constant militia battles there were harming migratory birds. Finally, a petition was sent to the Oracle of Yaanek, asking what to do. The God recommended that the position be eliminated, and offered to decide who filled the newly vacated seat Himself; thus the beginning of the Representative For Theocracy.

The Constitution was never fully amended, so technically the position is still the Representative For Kratocracy, and technically anyone who kills the Representative For Theocracy can still take his seat and gain immense power. But for some reason everyone who tries this dies of completely natural causes just before their plan comes to fruition. Must be one of those coincidences.

# NYT Is Threatening My Safety By Revealing My Real Name, So I Am Deleting The Blog

[EDIT 8/4: This post is now six weeks old. The current situation is unclear, but it’s less urgent to take any of the actions described below.]

So, I kind of deleted the blog. Sorry. Here’s my explanation.

Last week I talked to a New York Times technology reporter who was planning to write a story on Slate Star Codex. He told me it would be a mostly positive piece about how we were an interesting gathering place for people in tech, and how we were ahead of the curve on some aspects of the coronavirus situation. It probably would have been a very nice article.

Unfortunately, he told me he had discovered my real name and would reveal it in the article, ie doxx me. “Scott Alexander” is my real first and middle name, but I’ve tried to keep my last name secret. I haven’t always done great at this, but I’ve done better than “have it get printed in the New York Times“.

I have a lot of reasons for staying pseudonymous. First, I’m a psychiatrist, and psychiatrists are kind of obsessive about preventing their patients from knowing anything about who they are outside of work. You can read more about this in this Scientific American article – and remember that the last psychiatrist blogger to get doxxed abandoned his blog too. I am not one of the big sticklers on this, but I’m more of a stickler than “let the New York Times tell my patients where they can find my personal blog”. I think it’s plausible that if I became a national news figure under my real name, my patients – who run the gamut from far-left to far-right – wouldn’t be able to engage with me in a normal therapeutic way. I also worry that my clinic would decide I am more of a liability than an asset and let me go, which would leave hundreds of patients in a dangerous situation as we tried to transition their care.

The second reason is more prosaic: some people want to kill me or ruin my life, and I would prefer not to make it too easy. I’ve received various death threats. I had someone on an anti-psychiatry subreddit put out a bounty for any information that could take me down (the mods deleted the post quickly, which I am grateful for). I’ve had dissatisfied blog readers call my work pretending to be dissatisfied patients in order to get me fired. And I recently learned that someone on SSC got SWATted in a way that they link to using their real name on the blog. I live with ten housemates including a three-year-old and an infant, and I would prefer this not happen to me or to them. Although I realize I accept some risk of this just by writing a blog with imperfect anonymity, getting doxxed on national news would take it to another level.

When I expressed these fears to the reporter, he said that it was New York Times policy to include real names, and he couldn’t change that.

After considering my options, I decided on the one you see now. If there’s no blog, there’s no story. Or at least the story will have to include some discussion of NYT’s strategy of doxxing random bloggers for clicks.

I want to make it clear that I’m not saying I believe I’m above news coverage, or that people shouldn’t be allowed to express their opinion of my blog. If someone wants to write a hit piece about me, whatever, that’s life. If someone thinks I am so egregious that I don’t deserve the mask of anonymity, then I guess they have to name me, the same way they name criminals and terrorists. This wasn’t that. By all indications, this was just going to be a nice piece saying I got some things about coronavirus right early on. Getting punished for my crimes would at least be predictable, but I am not willing to be punished for my virtues.

I’m not sure what happens next. In my ideal world, the New York Times realizes they screwed up, promises not to use my real name in the article, and promises to rethink their strategy of doxxing random bloggers for clicks. Then I put the blog back up (of course I backed it up! I’m not a monster!) and we forget this ever happened.

Otherwise, I’m going to lie low for a while and see what happens. Maybe all my fears are totally overblown and nothing happens and I feel dumb. Maybe I get fired and keeping my job stops mattering. I’m not sure. I’d feel stupid if I caused the amount of ruckus this will probably cause and then caved and reopened immediately. But I would also be surprised if I never came back. We’ll see.

I’ve gotten an amazing amount of support the past few days as this situation played out. You don’t need to send me more – message very much received. I love all of you so much. I realize I am making your lives harder by taking the blog down. At some point I’ll figure out a way to make it up to you.

In the meantime, you can still use the r/slatestarcodex subreddit for sober non-political discussion, the not-officially-affiliated-with-us r/themotte subreddit for crazy heated political debate, and the SSC Discord server for whatever it is people do on Discord. Also, my biggest regret is I won’t get to blog about Gwern’s work with GPT-3, so go over and check it out.

There’s a SUBSCRIBE BY EMAIL button on the right – put your name there if you want to know if the blog restarts or something else interesting happens. I’ll make sure all relevant updates make it onto the subreddit, so watch that space.

There is no comments section for this post. The appropriate comments section is the feedback page of the New York Times. You may also want to email the New York Times technology editor Pui-Wing Tam at pui-wing.tam@nytimes.com, contact her on Twitter at @puiwingtam, or phone the New York Times at 844-NYTNEWS.

(please be polite – I don’t know if Ms. Tam was personally involved in this decision, and whoever is stuck answering feedback forms definitely wasn’t. Remember that you are representing me and the SSC community, and I will be very sad if you are a jerk to anybody. Please just explain the situation and ask them to stop doxxing random bloggers for clicks. If you are some sort of important tech person who the New York Times technology section might want to maintain good relations with, mention that.)

If you are a journalist who is willing to respect my desire for pseudonymity, I’m interested in talking to you about this situation (though I prefer communicating through text, not phone). My email is scott@slatestarcodex.com.

[EDIT: I am now over capacity for interview requests. Feel free to send more if you want, but I may not answer, sorry.]

[EDIT2: There’s now a petition you can sign.]

# Update On My Situation

It’s been two and a half months since I deleted the blog, so I owe all of you an update on recent events.

I haven’t heard anything from the New York Times one way or the other. Since nothing has been published, I’d assume they dropped the article, except that they approached an acquaintance for another interview last month. Overall I’m confused.

But they definitely haven’t given me any explicit reassurance that they won’t reveal my private information. And now that I’ve publicly admitted privacy is important to me – something I tried to avoid coming on too strong about before, for exactly this reason – some people have taken it upon themselves to post my real name all over Twitter in order to harass me. I probably inadvertently Streisand-Effect-ed myself with all this; I still think it was the right thing to do.

At this point I think maintaining anonymity is a losing battle. So I am gradually reworking my life to be compatible with the sort of publicity that circumstances seem to be forcing on me. I had a talk with my employer and we came to a mutual agreement that I would gradually transition away from working there. At some point, I may start my own private practice, where I’m my own boss and where I can focus on medication management – and not the kinds of psychotherapy that I’m most worried are ethically incompatible with being a public figure. I’m trying to do all of this maximally slowly and carefully and in a way that won’t cause undue burden to any of my patients, and it’s taking a long time to figure out.

I’m also talking to Substack about moving to their blogging platform. While part of me wants to jump right back into blogging here and pretend nothing ever happened, the Substack option has grown on me. I think I’d feel safer as part of a big group that specifically promises to defend their bloggers when needed. And also, I’d feel safer with a lot of diverse income streams, and Substack has made me an extremely generous offer. Many people gave me good advice about how I could monetize my blog without Substack – I took these suggestions very seriously, and without violating a confidentiality agreement all I can answer is that Substack’s offer was extremely generous.

When I originally asked readers about this possibility, they raised a lot of valid concerns: some of them were confused by Substack’s commenting system, others annoyed by its pop-up reminders to subscribe, others were concerned about being stuck outside a paywall. I’ve talked to Substack about this, and they’ve made some really impressive promises to address these things – they’re going to code a maximally-SSC-like commenting experience, they’re going to let me opt out of the subscription reminders, I won’t have to “paywall” anything besides some Hidden Open Threads. This isn’t the time for me to go over the dozens of examples of concerns I had that Substack went above and beyond to address, but assume I had most of the same ones you did and put a lot of work into addressing them.

(and if you’re worried about the Hidden Open Threads, check out Data Secrets Lox, a forum that has done a great job keeping the SSC Open Thread tradition going over the past few months.)

So that’s where I am right now – trying to wind things down at my day job, very preliminarily planning a private practice, and negotiating writing details with Substack. I’m also looking into some other things to protect my physical safety. When all of that is done, I’ll start blogging again. Right now I’m expecting that to be some time between October and January – and obviously when it happens I’ll let you know. I would appreciate if people continued to respect my preferences about anonymity until then. After that I’ll stop caring as much – though I’ll still go by “Scott Alexander” to keep the brand the same, and I’ll still do what I can to avoid publicity.

I might have hinted at this already, but I should say it explicitly – I’m really grateful for all the support I got throughout this whole incident. You people are all great. I’ll say so at more length later, and talk more about some specific examples, but for now just accept on faith that you’re all great.

I still plan to do the book review contest! I’ll do it sometime after I start the new blog! Those of you who sent me reviews didn’t waste your time! It’s going to happen! Pestilence may afflict every corner of the world, the skies may turn red as blood and the sun go dark at noon, the earth may shake and plagues of locusts cover the land, but never doubt that there will be a book review contest someday, in the golden future, when all of this is over.

As all the kids are saying these days, “thank you for your continuing support during these difficult times”.