DYNOMIGHT INTERNET EBOOK

2021-12-09

Table Of Contents

In defense of Myers-Briggs

The Myers-Briggs Personality Indicator (MBTI) gets a lot of scorn:

> About 2 million people take it annually, at the behest of corporate HR departments, colleges, and even government agencies. The company that produces and markets the test makes around $20 million off it each year.
>
> The only problem? The test is completely meaningless.
>
> Why the Myers-Briggs test is totally meaningless (Vox)
>
> I began to read through the evidence, and I found that the MBTI is about as useful as a polygraph for detecting lies. One researcher even called it an "act of irresponsible armchair philosophy." When it comes to accuracy, if you put a horoscope on one end and a heart monitor on the other, the MBTI falls about halfway in between.

> Goodbye to MBTI, the Fad That Won't Die (Psychology Today)
>
> Some psychologists believe that independent, peer-reviewed research in the decades since the MBTI was devised has provided something better than Myers-Briggs. They champion the notion of the "Big Five" personality traits — openness, conscientiousness, extroversion, agreeableness and neuroticism.
>
> Is Myers-Briggs up to the job? (Financial Times)
>
> the MBTI gives a ridiculously limited and simplified view of human personality, which is a very complex and tricky concept to pin down and study. The scientific study of personality is indeed a valid discipline, and there are many personality tests that seemingly hold up to scientific scrutiny (thus far). It just appears that MBTI isn't one of them.
>
> Nothing personal: The questionable Myers-Briggs test (The Guardian)

It would seem that the the MBTI is nonsense, but the Big Five is a *real, scientifically valid* test. To be sure, there's nothing wrong with the Big Five. But these haughty claims that it's dramatically better than the MBTI have got to end.

Claims Against the MBTI

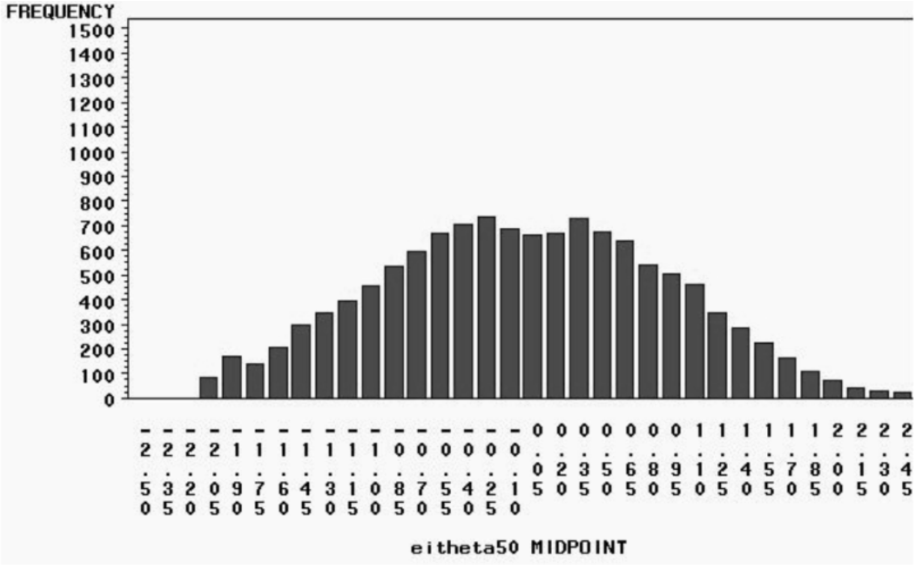
MYERS AND BRIGGS JUST, LIKE, MADE IT UP

It's true. Myers and Briggs were enthusiasts of Carl Jung's theories who created the test during WWII. They hoped to aid the entry of women into the workforce. They weren't professional scientists and didn't base it on some large corpus of data. And Jung's theories are… controversial today.

THERE AREN'T 16 DISCRETE TYPES OF PEOPLE

Also true! There is no switch in your brain set to T or F. We will probably never find the "judging" gene. These traits *are* probably heritable — everything is — but polygenic. So we should expect varying strengths even if we disregard "nurture" influences.

This complaint is understandable. Some early MBTI proponents really did defend binary outcomes, claiming each attribute was "theoretically dichotomous". There was even some now-debunked research that claimed to prove that empirically. It's now clear that the population has a standard bell-shaped distribution for each trait. For example, this is a histogram of the E-I axis (other axes look similar):



This is from Bimodal Score Distributions and the Myers–Briggs Type Indicator: Fact or Artifact? (Answer: Artifact)

THE MBTI HAS POOR REPEATABILITY

Technically true. Suppose someone takes the test, then repeats it 4-5 weeks later. Something like 35-50% of people will fall into a different one of the 16 groups. Is that bad? We'll come back to this.

THE MBTI DOESN'T VARY BETWEEN PEOPLE

Not true. This is claimed strangely often with no evidence. Remember that histogram above for introversion vs. extroversion? There's plenty of variability. All the other axes are similar. Since no one really provides any evidence for this claim, there's not much to rebut here.

If we're being *very* generous, perhaps we could interpret this as a complaint that the population mean for each trait is around zero. That's true, but again, it varies *plenty*.

THE MBTI MAKES EVERYONE SEEM LIKE A BEAUTIFUL SPECIAL SNOWFLAKE

I think this is true. It's hard to read the Big Five and conclude that neuroticism is a positive thing. But every MBTI type seems wonderful in its own way.

THE MBTI MEASURES THINGS THAT AREN'T FUNDAMENTAL or THE MBTI ISN'T USED BY REAL SCIENTISTS or MY IN-GROUP SAYS I MUST TREAT IT WITH CONTEMPT

Well, maybe.

The Myers-Briggs foundation grasps at an image of scientific seriousness in a way that's a little ridiculous. It grates on me, and I'm not a psychologist. People should probably stop spending money on the official test. People should *definitely* stop paying $2,500 to get certified to administer that test.

In defense of the MBTI

YOU DON'T HAVE TO BINARIZE THE AXES

90% of the complaints about MBTI come down to: *you can't split people into two groups along these axes*! Yeah, OK, then how about we don't do that?

Of course someone can be borderline. *Most* people are borderline along at least one axis. Early versions of the MBTI actually gave an "x" for someone near the middle. Let's bring this back! In fact, let's go further.

Behold: **Dynomight™ MBTI notation**:

- Strong preference: Capital letter
- Weak preference: Lower-case letter
- Borderline: x

For example: eNxJ is:

- weakly extroverted (e)
- strongly intuitive (N)
- borderline thinking / feeling (x)
- strongly judging (J)

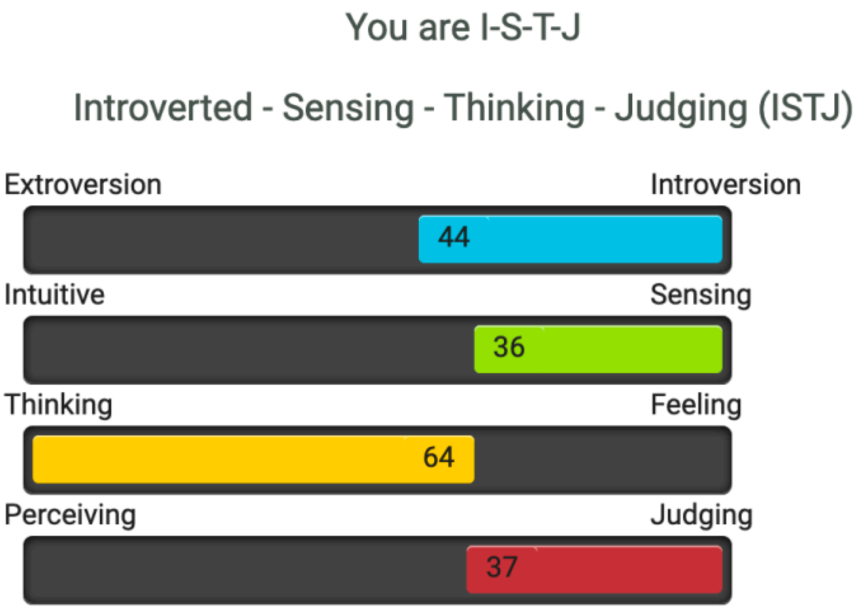
This gives five bins for each axis. You can still sort of say it out loud by making the strong preferences louder. This gives 5 =625 possible results. Of course, there aren't 625 discrete types of people either. Our goal isn't to split people into groups, it's to give a convenient summary of how they answered a bunch of questions.

To underline this point, I even created a Myers-Briggs quiz you can take in 2 minutes that gives results in this form.

To claim it's never useful to discretize denies human nature. You might as well claim we shouldn't say "hot" or "cold" but must always give a number of degrees.

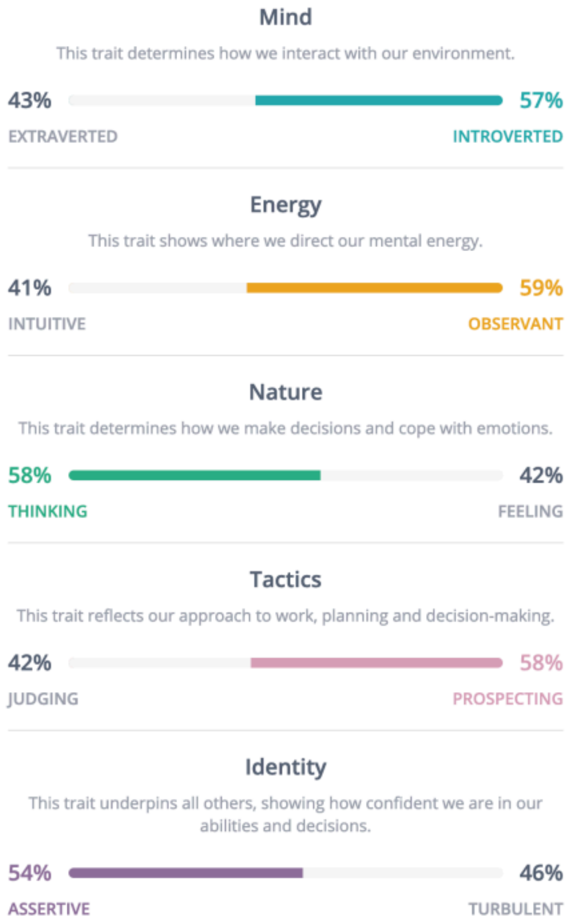
IF YOU DON'T LIKE BINARIZATION THEN DON'T

If you'd like a test that gives the scores as continuous attributes, may I recommend… any test? For example, here's the results of me randomly clicking answers on two popular websites:



You are I-S-T-J

Introverted - Sensing - Thinking - Judging (ISTJ)



Mind

This trait determines how we interact with our environment.

43% EXTRAVERTED — 57% INTROVERTED

Energy

This trait shows where we direct our mental energy.

41% INTUITIVE — 59% OBSERVANT

Nature

This trait determines how we make decisions and cope with emotions.

58% THINKING — 42% FEELING

Tactics

This trait reflects our approach to work, planning and decision-making.

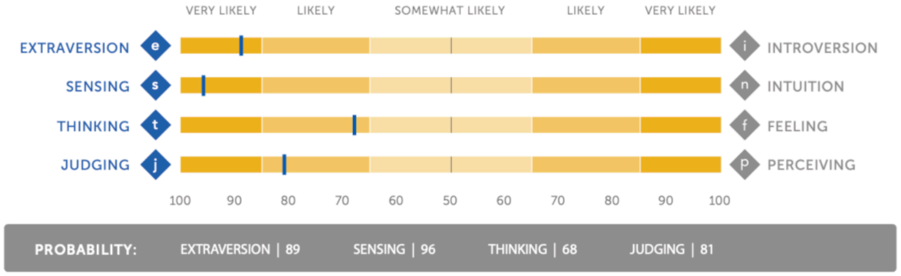
42% JUDGING — 58% PROSPECTING

Identity

This trait underpins all others, showing how confident we are in our abilities and decisions.

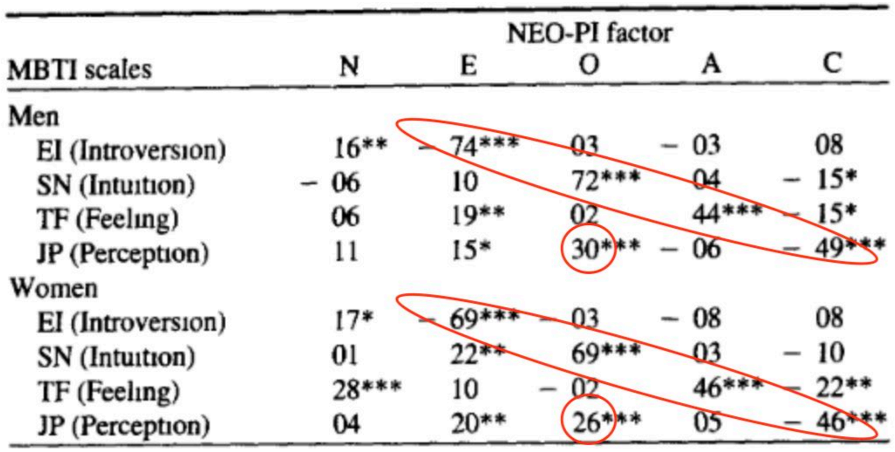
54% ASSERTIVE — 46% TURBULENT

Not convinced? You can even check the "official" results that gullible people pay money for. They look like this:



THE MBTI MIGHT AS WELL BE CALLED THE "BIG FOUR"

And what happens if you don't discretize the axes? If you take continuous measurements (like *every Myers-Briggs test ever* gives you) they correlate strongly with four of the five big five measurements.



	NEO-PI factor				
MBTI scales	N	E	O	A	C
Men					
EI (Introversion)	16**	− 74***	03	− 03	08
SN (Intuition)	− 06	10	72***	04	− 15*
TF (Feeling)	06	19**	02	44***	− 15*
JP (Perception)	11	15*	30***	− 06	− 49***
Women					
EI (Introversion)	17*	− 69***	− 03	− 08	08
SN (Intuition)	01	22**	69***	03	− 10
TF (Feeling)	28***	10	− 02	46***	− 22**
JP (Perception)	04	20**	26***	05	− 46***

The rows show the four MBTI axes, while the columns show the Big Five axes. This shows that:

- The MBTI E-I axis is strongly correlated with the Big Five extraversion axis. (duh)
- The MBTI S-N axis is strongly correlated with the Big Five openness axis.
- The MBTI T-F axis is moderately correlated with the Big Five agreeableness axis and weakly correlated with the conscientiousness axis.
- The MBTI J-P axis is moderately correlated with the conscientiousness axis and weakly correlated with the openness axis.

To interpret these numbers, note that 74/69 is what you get when the axis is *exactly the same thing with exactly the same name*. These correlations are strong.

But don't blame all the journalists for not knowing about this. After all this groundbreaking research is hot off the presses having only been published in 1989. Other research supports the same basic conclusion.

What does this mean? If you believe the MBTI is meaningless, fine. But you must also therefore believe the Big Five is meaningless!

THE MBTI'S REPEATABILITY IS FINE

If someone is near the middle, a small change can land them on the other side. This is why 35-50% of people fall into a different one of the 16 buckets when they re-test. The most common score for all axes is near the middle. There are 4 different axes. If some small measurement noise puts you over the middle on *any* axis, then you fall into a different bucket.

Since *we* aren't binarizing the axes, this is not something to worry about. Instead, we should just ask: How repeatable are the continuous measurements?

For reference, here's Cronbach's alpha for the Big Five, measured for an Arabic version. Higher numbers indicate better repeatability.

Trait	Men	Women
Neuroticism	.83	.74
Extroversion	.82	.83
Openness	.79	.85
Agreeableness	.82	.81
Conscientiousness	.90	.92

And here's the results for Myers Briggs:

Trait	Men	Women
EI	.82	.83
SN	.83	.85
TF	.82	.80
JP	.87	.86

YOU CAN ADD AN AXIS IF YOU WANT

The Big Five measures neuroticism, while the MBTI does not. This is, no doubt, a very important trait. Some MBTI variants add extra axes. For example, one can introduce an axis that measures *turbulence* vs. *assertiveness.* This is a measure of neuroticism in all but name.

Why go with the MBTI?

Everything above just says the MBTI isn't much worse than the Big Five. So why not just use the Big Five? Well, you just try it.

THE MBTI IS COMFORTABLE TO TALK ABOUT

Say you're on a first date. You discuss your favorite Italian films, the pets you had growing up. Then, you ask *How agreeable are you?* How does that go?

If that seems cherry-picked, let's go through all the attributes:

- *How extroverted are you?* This is fine.
- *How neurotic are you?* Uncomfortable.
- *How agreeable are you?* Uncomfortable.
- *How conscientious are you?* Somewhat uncomfortable.
- *How open are you to experience?* This *seems* fine, but isn't.

Only one of the five is OK. And — as you surely noticed— that same attribute exists in the MBTI.

Can I tell you a secret? The MBTI's names for the S-N axes are *sensing* and *intuition*. These don't make any sense. The Big Five *openness to experience* is much more descriptive. And yet, I claim, the MBTI names are often better *because* they are meaningless. Everyone claims to be open! But since no one has any idea what "sensing" is, they'll happily admit to it. Deliberate or accidental, this subterfuge is extremely useful.

For similar reasons, the MBTI's omission of neuroticism is sometimes good. The alternative to four axes is often zero, not five.

The advantage of the the MBTI is precisely that it *makes everyone feel like a beautiful snowflake*. The axes are chosen and named ingeniously so as to make them easy to reveal.

THE MBTI IS POSSIBLE TO TALK ABOUT

Binarization is the key criticism of the MBTI. Still, we must confront the fact that binarization is a huge driver of the MBTI's success. How do we reconcile this?

Say you're a weirdo who is happy to tell your date your Big Five. How do you actually say it out loud? In fact, I am such a weirdo, and I end up saying something like "My neuroticism is low, my conscientiousness is high, my extroversion is moderate, and… ummm, what are the other axes?"

A big part of what makes the MBTI easier to talk about is that it has names for *both ends* of the axes. With the MBTI, you have the *option* of a coarse summary by saying something like "ESFP". It's at least *possible*! Of course, more precise measurements are better, but having options is a positive thing.

I think that we could also sort of make this work with my suggested notation by modulating the sound of your voice. (If you get eNxJ, you say "I'm an e **en** ex **jay**").

Summary

You don't have to discretize the MBTI axes. The MBTI measures similar stuff as the Big Five with similar repeatability, but it is more appealing and dramatically easier to talk about. Don't feel guilty about using it.

In theory the Big Five could adopt the advantages of the MBTI by doing two thing. (Not that I'm advising this!) First, create names for the axes that sound more neutral, so that people are happy to discuss them. Second, create different names for different *ends* of the axes, so that an easy optional discretization is possible. But honestly, why bother? If you did that, you'd have basically invented the MBTI.

Comparative advantage and when to blow up your island

Economists say free trade is good because of “comparative advantage”. But what is comparative advantage? Why is it good?

This is sometimes considered an arcane part of economics. (Wikipedia defines it using “autarky”.) But it’s really a very simple idea. Anyone can use it to understand the world and make decisions.

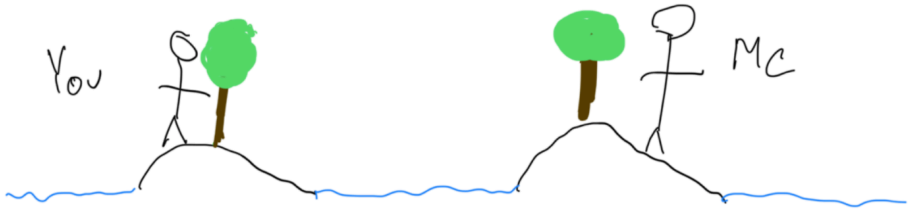
I Islands

Say you live alone on an island.

Each week you gather and eat 10 coconuts and 10 bananas. It takes you five minutes to gather a coconut, and 10 minutes for a banana. Thus, you work 150 minutes per week.

	You Need	Time to gather one	Time You Spend
Coconuts	10	5 minutes	50 minutes
Bananas	10	10 minutes	100 minutes
			Total: 150 minutes

I live on a nearby island.



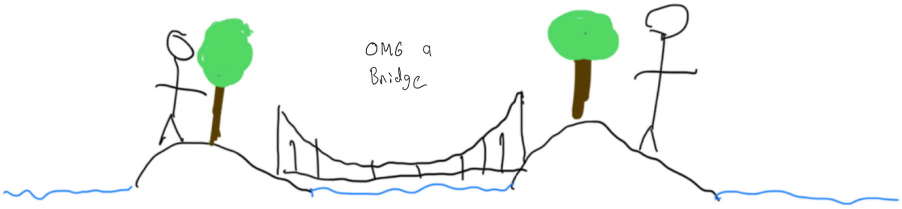
Just like you I eat 10 coconuts and 10 bananas per day. But *unlike* you, I’m terrible at everything.

	I Need	Time to gather one	Time I Spend
Coconuts	10	60 minutes	600 minutes
Bananas	10	30 minutes	300 minutes
			Total: 900 minutes

Since I'm so incompetent, I need to work a lot more than you – six times as much.

II The Bridge

Thus, we live our lives until one day a bridge appears between the islands.



We are both peaceful. We will not coerce each other, but are otherwise completely selfish. What will happen?

Intuitively, you value bananas more, while I value coconuts more. So it's natural to trade my bananas for your coconuts. We agree as follows: Each week, you gather 20 coconuts, and I gather 20 bananas. Then, I trade 10 of my bananas for 10 your coconuts. It's easy to check that this will make *both* of us better off.

	You Gather	Time to gather one	Time You Spend
Coconuts	20	5 minutes	100 minutes
Bananas	0	10 minutes	0 minutes
			Total: 100 minutes

	I Gather	Time to gather one	Time I Spend
Coconuts	0	60 minutes	0 minutes
Bananas	20	30 minutes	600 minutes
			Total: 600 minutes

In one sense, it's obvious that trade makes us both better off. If it didn't we wouldn't both agree to it! But comparative advantage explains how. You have an *advantage* at everything. But I have a *comparative advantage* at bananas, because my ratio (banana time) / (coconut time) is lower than yours. And if we both concentrate our efforts on the thing we have a comparative advantage at, we are both better off.

This is why economists like free trade. If different producers have different relative abilities, everyone can benefit from specializing. This is true even if one producer is better at everything.

Beyond trade, this is an important lesson for life. Choosing your career path? Dividing up chores with your partner? Think about comparative advantage!

III Complexities

The real world, of course, is more complex. For example:

- There might be transportations costs.
- It might get harder to find coconuts as you gather more of them.
- There might be more goods to trade.

More sophisticated models can deal with these complications. The math gets more complex, but more or less the same conclusion arises. There is one complication that's a bit special:

- There might be more than two people.

In this case, introducing trade can make *individual people* worse off: Suppose you live with me on an island, but you're incapable of gathering bananas. Since you need them to live, I can demand a huge number of coconuts for one banana. When a bridge opens up to another island, you might get a better trade. This will help you, but actually *hurt* me. Still, introducing free trade still makes people better off "on the whole".

In this subtlety, Politics emerges. In principle, one can always use free trade plus a set of wealth transfers to make every individual better off. But that's a nightmare in practice: It would require a central authority to predict what set of trades the market will decide on. So we're left with a mess.

IV ZOPA

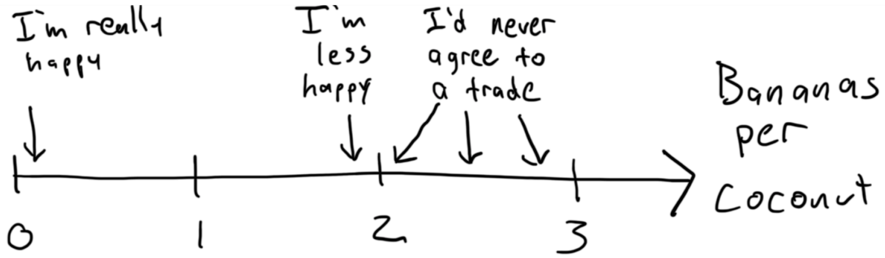
But even in this toy model of two people on two islands, I skipped an important step. How did we decide to trade 10 coconuts for 10 bananas? I might say: "I'll trade 7 bananas for 10 of coconuts. Take it or leave it!"

Of course, this would be great for me, and worse for you than our original trade. But it's easy to check that this is better for you than no trade at all.

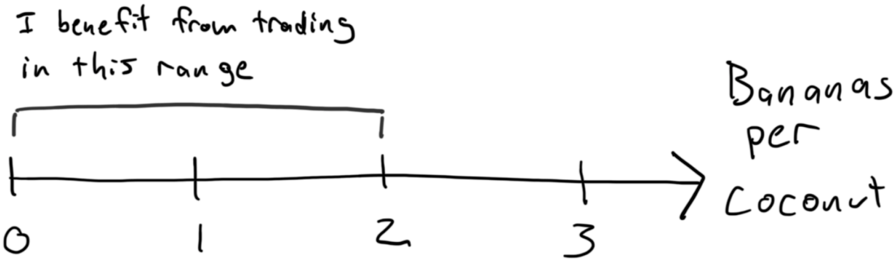
	You Gather	Time to gather one	Time You Spend
Coconuts	20	5 minutes	100 minutes
Bananas	3	10 minutes	30 minutes
			Total: 130 minutes

	I Gather	Time to gather one	Time I Spend
Coconuts	0	60 minutes	0 minutes
Bananas	17	30 minutes	510 minutes
			Total: 510 minutes

Now, what possible banana/coconut exchange rates could we arrive at? I'd be happiest paying you nearly zero bananas for each coconut. On the other hand, I'd never agree to pay you three bananas per coconut – it would be "cheaper" for me to just make the coconuts myself. I'd never agree to trade more than two.

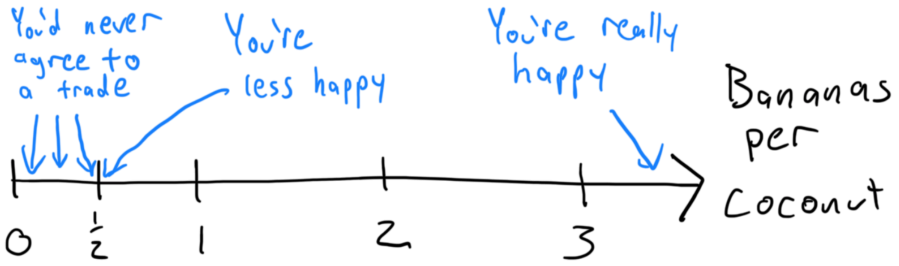


Thus, I benefit from any trade where I pay you between 0 and 2 bananas for one coconut. These are the only trades I'd ever agree to.

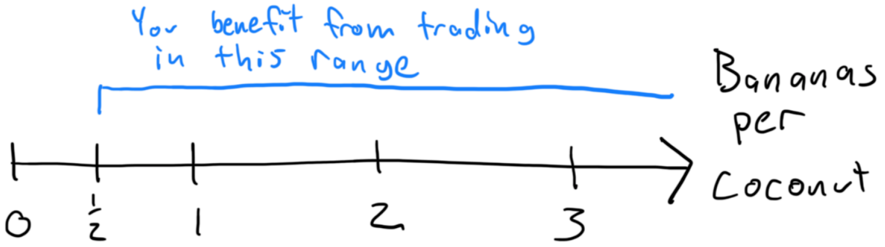


Of course, I'd prefer to pay you fewer bananas! So I'd prefer a rate to the left end of this range.

Conversely, it takes you twice as long to make banana as a coconut. You'd be thrilled if I paid you 4 bananas per coconut, but you'd never accept less than 1/2 a banana for one coconut.



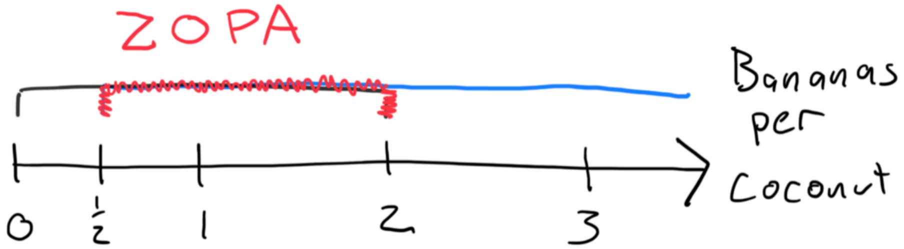
Thus, you benefit from any trade where I give you more than 1/2 a banana for a coconut.



You'd like me to pay you as many bananas as possible. So you'd prefer a rate as far to the right as possible.

Now, the big question is: What rate do we agree on? Simple economics does not tell us the answer! In principle, our negotiations could arrive at an "exchange rate" of anywhere between .5 and 2 of your coconuts for 1 of my bananas.

This range of (.5 to 2) is the Zone of Possible Agreement (ZOPA) in negotiation theory.



V Perverse Behavior

There's no simple simple math to decide what point in the ZOPA we settle on. This can lead to strange and perverse behaviors.

Walking away. Since we are nonviolent, the only "threat" is to refuse to trade. If you know I am "rational" and won't refuse a beneficial deal, you can be "irrational" and refuse to trade unless we do so at the end of the range that's favorable to you. Thus, your “irrational” behavior gives you a better outcome than my “rational” behavior.

Gaining information. Before our first meeting, I build a telescope and spy on you. When we meet, I say “I noticed it takes you 2x as long to make a banana as a coconut. It takes me 1.95x as long. Bananas sure are hard, aren’t they? Because I like you, I’m willing to trade at a rate of .512 bananas per coconut (.512=1/1.95). This does nothing for me, but you have a kind face, and I want to help you.” If you believe me, I get a very favorable rate.

Concealing information. You are smart. After the bridge appears, you quality realize I might spy on you, and this would harm your negotiation position. Before doing any gathering, you construct a privacy wall around your island.

Faking skills. You’re a hard-ass. You will walk away unless I agree to an exchange on your end. I’ve tried walking away, but you don’t care. I always blink before you, and we both know it. What can I do? For a weeks, I secretly gather coconuts in the night. The next time we meet, I bring a huge pile of coconuts. I say “I’ve been practicing, and now it only takes me 1.5x as long to make a coconut as a banana. I know you’re a hard-ass and you want the sweet end of the ZOPA. I admit I can’t beat you, but the ZOPA has shifted. You need to offer me a better deal.”

Blowing up my island. You’re a hard-ass. I only get .501 coconuts for 1 banana. I’ve tried walking away, but we both know you will out-wait me. I’ve tried fakings skills, but you won’t bite. Because we are non-violent, I can’t coerce you. But there’s nothing wrong with hurting *myself*, is there? I build a machine that monitors inter-island commerce. If there is ever a trade that is not 1 coconut for 1 banana, the machine activates a bomb, my island sinks into the ocean forever, and I die. If I try to disable the machine, the bomb activates. When we next meet I say “OK. I can’t out bad-ass you. However, because of this machine, it will forever be against my interests to agree to a non-even trade. There’s no point in you waiting. Even if I *did* agree to an uneven trade, I’d sink into the ocean, and you’d have to gather your own coconuts!”

Blowing up your island if I threaten to blow up my island. You are smart. You are also a hard-ass. As soon as the bridge appears, you know you can out-wait me to get a good rate. You immediately realize that my only option is to build the island destroying machine described above. Before we meet, you construct a machine that monitors my island for the presence of machines. Your machine is connected to a bomb on your island. If at any point, a bomb-activating machine is constructed on my island, your bomb activates, your island sinks into the ocean, and you die. When we meet, you explain that you’re a hard-ass, and that no island-destroying machines can help me. My best bet is to accept terms that barely improve my situation at all. You win.

What happens if you drink acetone?

Question: Should you drink acetone?

Answer: No.

But, out of interest, what if you did? This question is asked repeatedly on the web, with with many answers smugly stating that even tiny amounts of acetone will instantly kill you, *you idiot.* But they provide no evidence.

Fact #1: Acetone bottles are scary looking

Certainly, this doesn't *look* like something you'd want to put in your body:

Fact #2: Your body naturally produces and disposes of acetone.

Acetone naturally occurs in plants. Your liver produces acetone when metabolizing fat. If you fast, have diabetes, or exercise very hard, you produce more acetone. If you follow a ketogenic diet, you produce more. (Acetone is a "ketone”!) Small amounts of acetone are naturally present in your blood and urine, the latter being how you get rid of it.

Fact #3: Diabetes can cause your breath to smell like acetone.

Insulin is needed to break down glucose and provide energy to cells. Diabetics have trouble either producing or using insulin. Thus, their bodies may burn fat instead. Burning lots of fat produces lots of acetone, enough to impact the breath. (This is a serious problem if it occurs.)

Fact #4: Drinking acetone will make you not think so good no more.

Fisher Scientific’s MSDS gives the following effects for acetone:

> Ingestion: May cause gastrointestinal irritation with nausea, vomiting and diarrhea. May cause systemic toxicity with acidosis. May cause central nervous system depression, characterized by excitement, followed by headache, dizziness, drowsiness, and nausea. Advanced stages may cause collapse, unconsciousness, coma and possible death due to respiratory failure.

Sounds serious! Except, oh wait, I made a “mistake”. That was the list of effects for ethanol. *Here* are the effects for acetone:

> Ingestion: May cause irritation of the digestive tract. May cause central nervous system depression, characterized by excitement, followed by headache, dizziness, drowsiness, and nausea. Advanced stages may cause collapse, unconsciousness, coma and possible death due to respiratory failure. Aspiration of material into the lungs may cause chemical pneumonitis, which may be fatal.

Remind you of anything?

Fact #5: Acetone is probably marginally more toxic than ethanol.

In animals, the Oral LD50 for acetone ranges from 3 g/kg in mice to 5.8 g/kg in rats. For ethanol it is around 7.3 g/kg for both mice and rats.

This suggests you’d need to drink something like an entire bottle of nail-polish remover to be at risk of dying: If the mice numbers translate to humans, someone who weighs 80kg (180lb) would need to drink 240g of acetone to have a 50% chance of death. Standard nail-polish remover bottles contain around 200ml and are 98% acetone.

The same person would need to consume around 584 ml of grain alcohol to have the same risk.

Fact #6: Acetone is Generally Recognized as Safe (GRAS) by the FDA.

For better or worse, food manufacturers can put acetone in food and sell it to you without testing for safety. This seems to be common with spice oleoresins (concentrated forms of spices).

Fact #7: Some insane internet people drank acetone and didn't die.

In a thread on *bluelight*, Psychedelic Jay reports:

> So far 1 ml of pure acetone in 10 ml of water. Effects: Slight sedation, easy going sense of euphoria, very similar but smoother than ethanol intoxication. Heart rate increased by 6-10 beats a minute… Blood pressure exactly the same…

While *pino* says:

> So one night, I took 20ml strongly diluted, a dose which shouldn't kill you. The taste was masked by mixing it with fruit juice, which made it actually pleasantly to sip. Slightly fruity. In about half an hour, a pleasant warm sedation spread over my body. It felt like a clean alcohol intoxication. Nothing to strong, but very relaxing. I guess it took me for an hour of 10. There is no hangover.

Both of these are consistent with the idea that acetone has effects that are similar to alcohol. All the other comments in that thread, of course, say "what, are you crazy?".

Fact #8: You shouldn't drink acetone.

There's no reason to do so. It's (presumably) disgusting. It's very flammable. The effects haven't been studied nearly as much as alcohol's. And I could be wrong about all of this.

But suppose acetone had exactly the same effects as ethanol. Yes, that would mean that "acetone is as safe as alcohol". But it would also mean that "alcohol is as dangerous as acetone". That's probably the wiser interpretation.

Making the Monty Hall problem weirder but obvious

The Monty Hall problem is famously unintuitive. This post starts with an extreme version where the solution is blindingly obvious. We then go through a series of small changes. It will be clear that these don't affect the solution. At the end, we arrive at the classic Monty Hall problem.

For reference, the classic formulation goes:

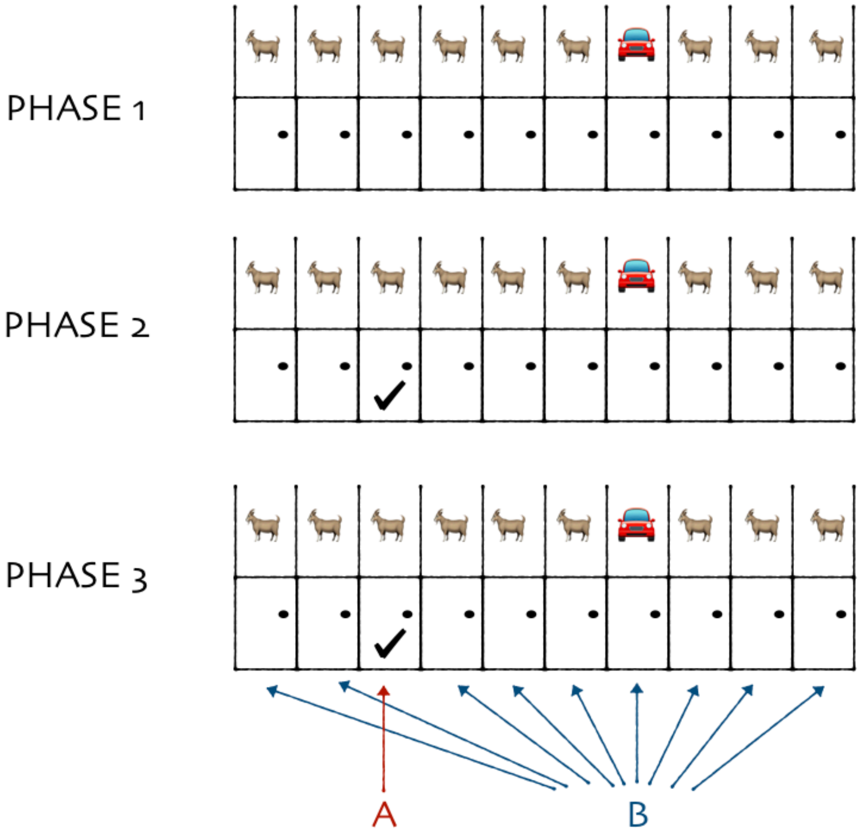
> Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?

Intuitively, many people guess it doesn't matter if you switch. But it does. You get the car 2/3 of the time if you switch, and 1/3 of the time if you don't. Why?

Game 1 (Dynomight™ Monty Hall)

Here's our first game.

1. There are 10 doors. A car is randomly placed behind one, and goats behind the other 9.
2. You pick one door.
3. You get two options:
    - Option A: You get whatever is behind the door you picked.
    - Option B: You get whatever is behind all of the other 9 doors.

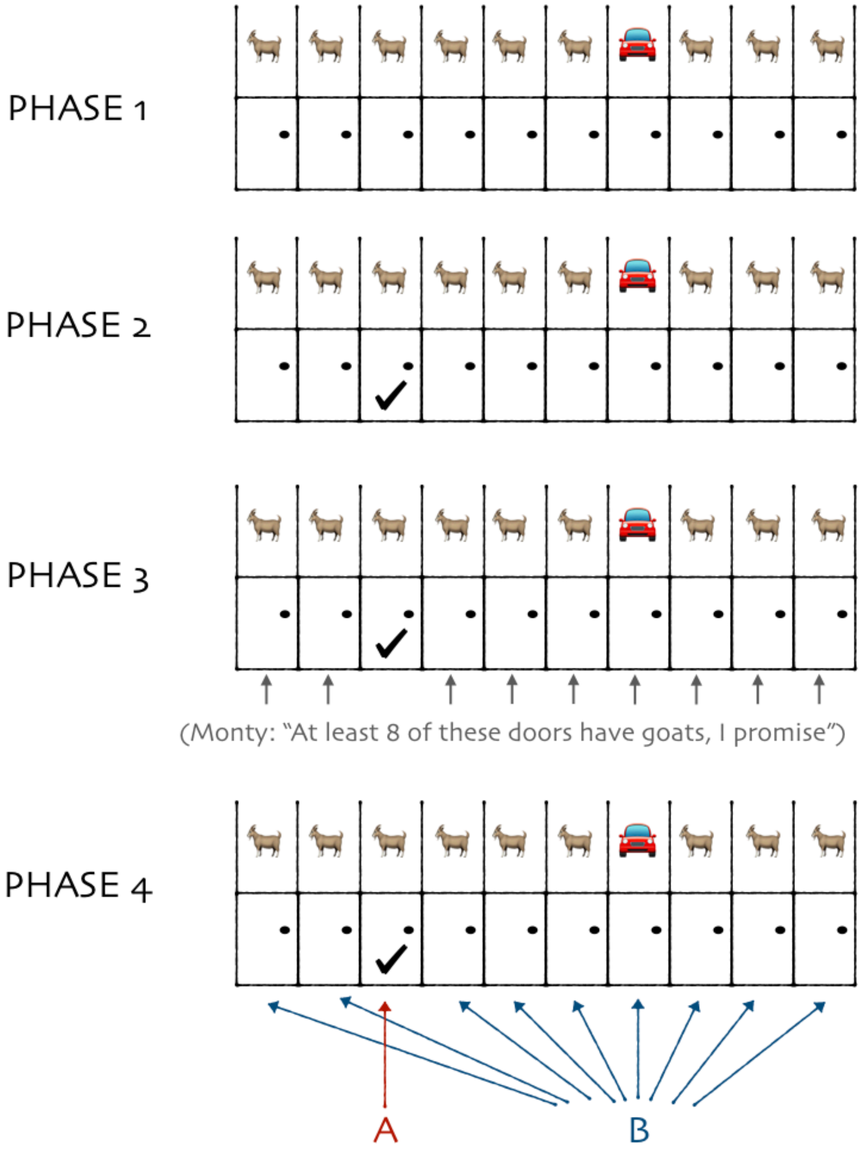


There's nothing mysterious here. You should choose option B. There's only a 10% chance you picked the right door, so there's a 90% chance the car is behind one of the others.

Game 2

Now, we slightly update the game (new part in bold).

1. There are 10 doors. A car is randomly placed behind one, and goats behind the other 9.
2. You pick one door.
3. **Monty says "Hey! I promise you that there is a goat behind at least 8 of the other 9 doors!"**
4. You get two options:
   - Option A: You get whatever is behind the door you picked.
   - Option B: You get whatever is behind all of the other 9 doors.

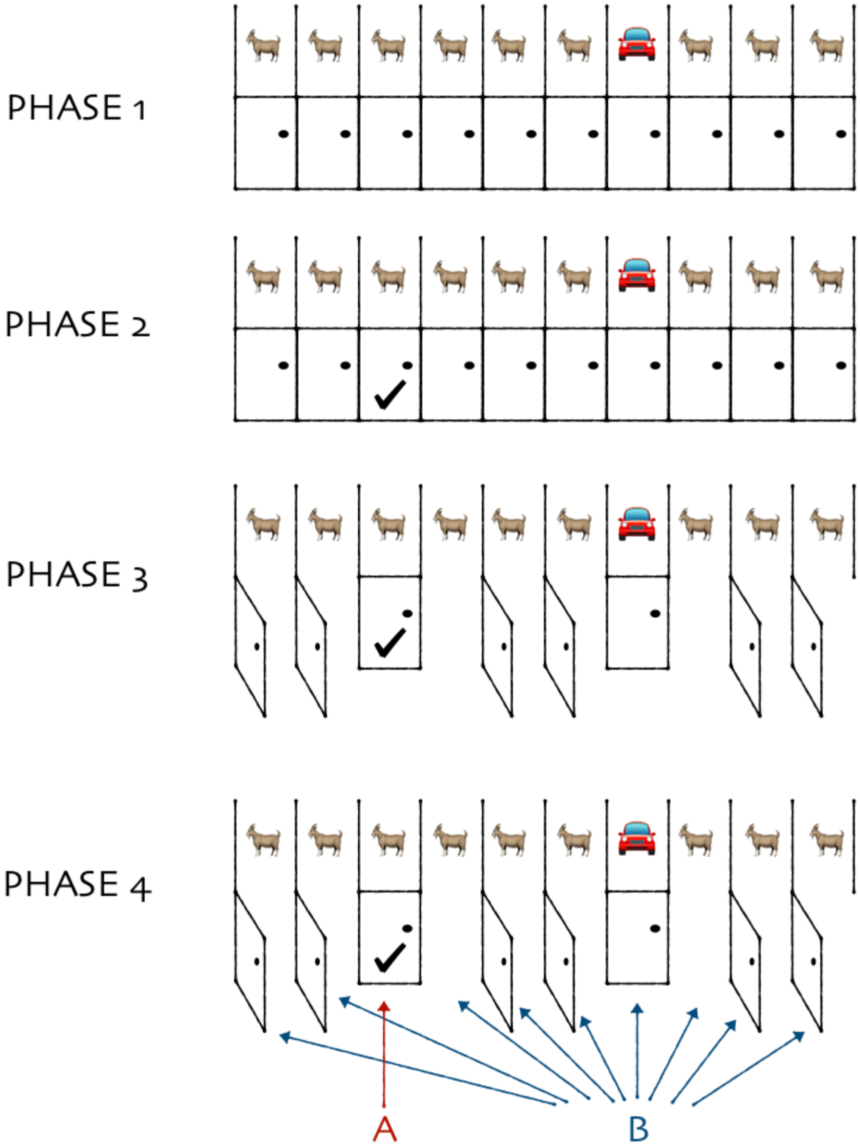


Monty's statement changes nothing. You don't need to rely on his trustworthy looks. You already *knew* there were at least 8 goats! Option B still gets you the car 90% of the time.

Let's update the game again (new part in bold).

1. There are 10 doors. A car is randomly placed behind one, and goats behind the other 9.

2. You pick one door.
3. **Monty looks behind the other 9 doors. He chooses 8 with goats behind them, and opens them.**
4. You get two options:
    - Option A: You get whatever is behind the door you picked.
    - Option B: You get whatever is behind all of the other 9 doors.



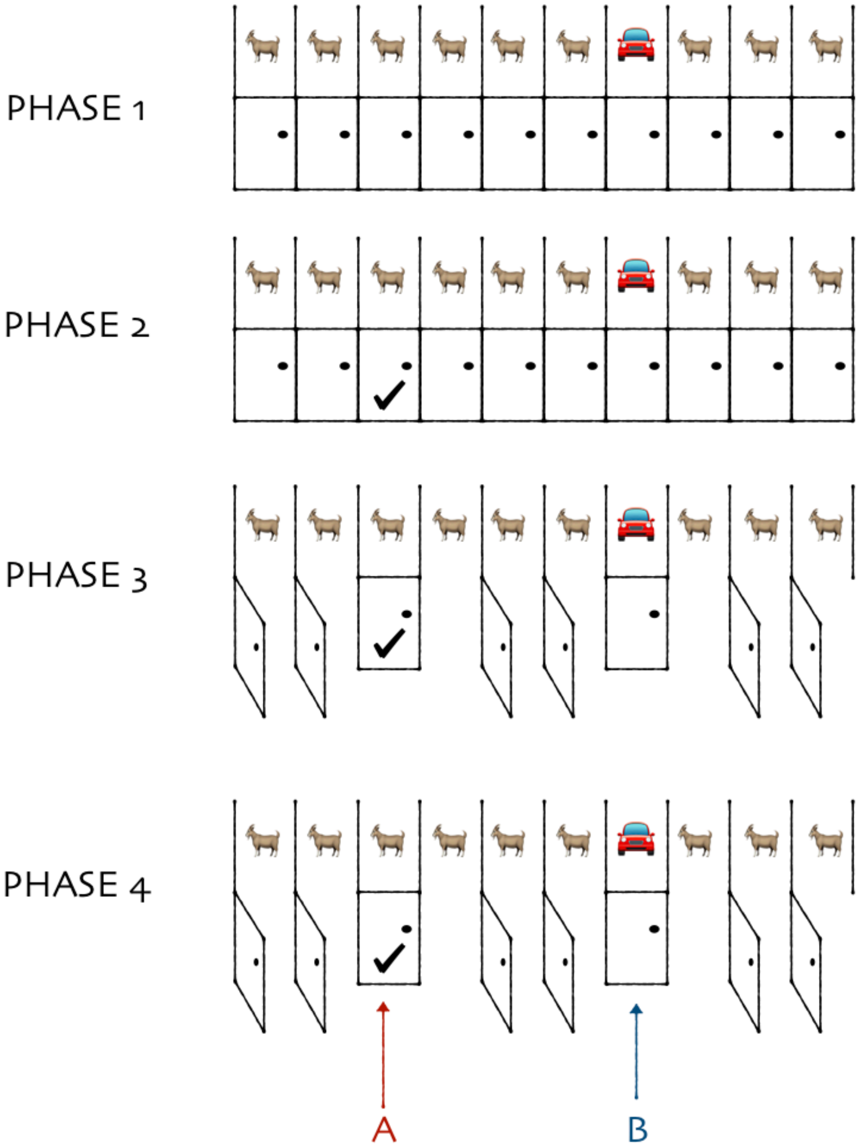
The key insight is this: When Monty shows you that 8 of the 9 other doors contain goats, you haven't learned anything relevant to your decision. You *already knew there were at least 8 goats behind the other doors*! So this is just like game 2. Option B still gets you the car 90% of the time.

Want more intuition? Suppose you picked door 3. Imagne Monty walking past the doors, opening doors 1, 2, 4, 5, 6, **skipping 7**, then opening 8, 9, and 10. Doesn't door 7 seem special?

Game 4

Let's make another change. Finally, we arrive at a game very similar to Monty Hall.

1. There are 10 doors. A car is randomly placed behind one, and goats behind the other 9.
2. You pick one door.
3. Monty looks behind the other 9 doors. He chooses 8 of them with goats behind them, and opens them.
4. You get two options:
   - Option A: You get whatever is behind the door you picked.
   - Option B: You get whatever is behind **the other closed door**.

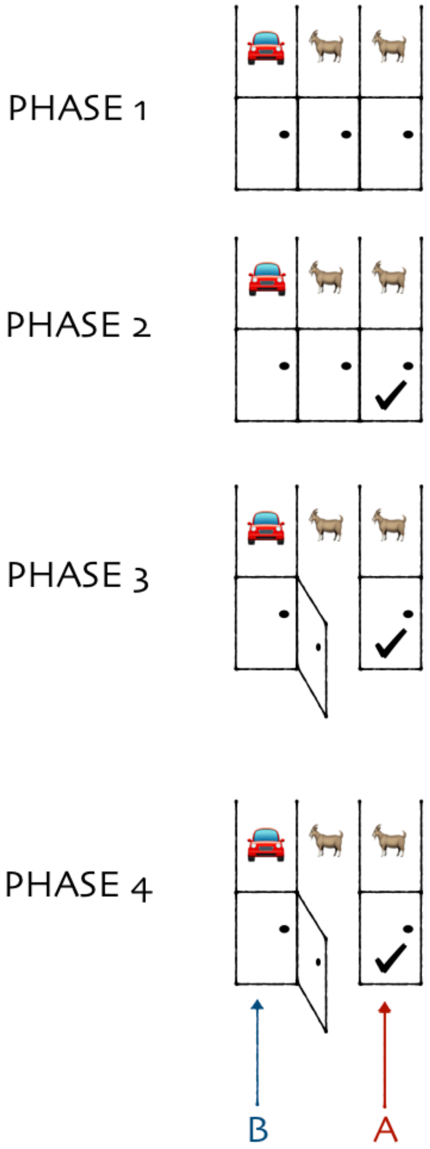


The only difference with Game 3 is that option B doesn't get you the 8 visible goats. Since you don't care about goats, this makes no difference. This is still just like the game 3. You get the car 90% of the time by switching.

Game 5 (Classic Monty Hall)

Here is the last game. We just change the number of doors from 10 to 3.

1. There are **3** doors. A car is randomly placed behind one, and goats behind the other **2**.
2. You pick one door.
3. Monty looks behind the other **2** doors. He chooses one **1** of them with a goat behind it, and opens it.
4. You get two options:
    - Option A: You get whatever is behind the door you picked.
    - Option B: You get whatever is behind the other closed door.



Of course, you still want to choose option B. The chance of success is now 2/3 instead of 9/10. This game is exactly Monty Hall, so we're done.

Side Notes

- It's important that Monty looked behind the doors before choosing which to open. This is where people's intuition usually fails. If he had chosen a door at random — *in a way that he risked possibly exposing a car*, then the situation would be different. (In that case, there's no advantage or harm in switching.) But he doesn't choose the door at random. He deliberately chooses to show you goats. Since this is always possible, it tells you nothing. I think this is the crux of what makes this problem unintuitive. Many people intuitively think it doesn't matter if you switch. And that *would be correct* if the door had been opened at random!

- It might be helpful to draw a diagram of the relationship of the different games, starting with classic Monty Hall and ending with the extreme version.

  Game 5 (Classic Monty Hall)
  ↓
  ↓ (Use 10 doors instead of 3)
  ↓
  Game 4
  ↓
  ↓ (If you switch, get the contents of *all* other doors, not just the other closed door.)
  ↓
  Game 3
  ↓
  ↓ (Monty promises 8 goats behind the other doors instead of showing you.)
  ↓
  Game 2
  ↓
  ↓ (Monty doesn't bother promising.)
  ↓
  Game 1 (Dynomight™ Monty Hall)

- There are some other attempts at variants of the Monty Hall problem, also intended to be more intuitive. These involve switching the doors for "boxers".

- Monty Hall was actually named "Monte" at birth! Given that Monte Carlo simulations are often used for exploring the Monty Hall problem, that's either a tragedy for puns or a miracle for confused students.

Your ratios don't prove what you think they prove

Watching people discuss police bias statistics, I despair. Some claim simple calculations prove police bias, some claim the opposite. Who is right?

No one. Frankly, nobody has any clue what they are talking about. It's not that the statistics are *wrong* exactly. They just don't prove what they're being used to prove. In this post, I want to explain why, and give you the tools to dissect these kinds of claims.

I've made every effort to avoid politics, due to my naive dream where well-meaning people can agree on facts even if they don't agree on policy.

Population size

The obvious place to start is to look at the number of people killed by police. This is easy to find.

	Black	White	Hispanic
# in US (million)	41.3	185.5	57.1
# killed by police per year	219	440	169
# killed by police per million people	5.3	2.3	2.9

Does this prove the police are racist? Before you answer, consider a different division of the population.

	Male	Female
# in US (million)	151.9	156.9
# killed by police per year	944	46
# killed by police per million people	6.2	0.29

And here's a third one.

	<18 y/o	18-29	30-44	45+
# in US (million)	72.9	53.6	63.2	137.3
# killed by police per year	19	283	273	263
# killed by police per million people	0.26	5.2	4.3	1.9

The first table above is often presented as an obvious "smoking gun" that proves police racism with no further discussion needed. But if that were true, then the second would be a smoking gun for police *sexism* and the third for police *ageism*. So let's keep discussing.

Of course, the second and third tables have obvious explanations: Men are different from women. The young are different from the old. Because of this, they interact with the police in different ways. Very true! But the following is also true:

	Black	White	Hispanic
average height (men)	175.5cm (5'9")	177.4cm (5'10)	169.5cm (5'7")
life expectancy	74.9 yrs	78.5 yrs	81.8 yrs
mean annual income	$41.5k	$65.9k	$51.4k
median age	33 yrs	43 yrs	28 yrs
go to church regularly	65%	53%	45%
children in single-parent homes	65%	24%	41%
identify as LGBT	4.6%	3.6%	5.4%
live in a large urban area	82%	61%	82%
poverty	21%	8.1%	17%
men obese	41%	44%	45%
women obese	56%	39%	43%
completed high school	87%	93%	66%
completed bachelor's	22%	36%	15%
heavy drinkers	4.5%	7.1%	5.1%

Maybe it's uncomfortable, but it's a fact: In the US today, there are few traits where there *aren't* major statistical differences between races. (Of course this doesn't mean these differences are *caused* by race! This is a good example of why correlation does not imply causation.)

A thought experiment

Suppose police were required wear augmented reality goggles. On those goggles, real-time image processing changes faces so that race is invisible. Would doing this cause police statistics to equalize with respect to race?

No. Even if race is *literally invisible*, young urban alcoholics will have different experiences with police than old teetotalers on farms. The fraction of these kinds of people varies between races. Thus, racial averages will still look different because of things that are *associated with race* but aren't *race as such.*

So despite the thousands of claims to the contrary, just looking at killings as a function of population size doesn't prove bias. Not does it prove a lack of bias. It really doesn't prove anything.

Arrests

Why do police kill more men than women? We can't rule out police bias. But surely it's relevant that men and women behave differently? So, it might seem like we should normalize not by population size, but by *behavior.*

One popular suggestion is to consider the number of arrests:

	Black	White	Hispanic
# killed by police per year	219	440	169
# arrests for violent crimes per year (thousands)	146	230	83
# killed by police per thousand violent crime arrests	1.4	1.9	1.9

Some claim this proves the police *aren't* biased, or even that there is bias in favor of blacks. But that's nearly circular logic: If police are biased, that would manifest in arrests as much as killings. So what we are really calculating above is

"Normal" killings + killings due to bias"Normal" arrests + arrests due to bias.

$$\frac{\text{“Normal” killings + killings due to bias}}{\text{“Normal” arrests + arrests due to bias}}.$$

The ratio doesn't tell you much about how large the bias terms are. So, unfortunately this also doesn't prove anything.

Incidentally: There are some popular but different numbers out there for this same ratio. These have tens of thousands of re-tweets with no one questioning the math. But I've checked the source data carefully, and I'm pretty sure my numbers are right. (They reach the same basic conclusion anyway.)

Murders

The police have discretion when deciding to make an arrest. But a dead body either exists or doesn't. So why not normalize by the number of murders committed?

This turns out to be basically impossible:

- Something like 40% of murders go unsolved, so the race of the murderer is unknown.
- The only real source of murder statistics is the FBI. They treat hispanic/non-hispanic ethnicity as *independent* of race. Why not just ignore hispanics then? Well, you can't. Hispanics are still counted as white or black in an unknown way. It's impossible to compare to police shooting statistics where hispanic is an alternative race.
- In around 31% of cases, the FBI has no information about race, and in 40% of cases, no information about ethnicity.

I've seen tons of articles use this version of the FBI's murder data that simply drops all the cases where data are unknown. None of these articles even acknowledge the issue of missing data or different treatment of hispanics.

Instead, let's look at murder *victims*. This is counterintuitive, but it's relatively rare for murders to cross racial boundaries (<20%). So this is a non-terrible proxy for the number of murders committed. Data from the CDC separates out black, white, and hispanics in a similar way as police shooting statistics.

	Black	White	Hispanic
# killed by police per year	219	440	169
# murder victims per year	9,908	5,747	3,186
# killed by police per murder victim	0.022	0.076	0.053

So what does this prove? Again, not much. The simple fact is that most police killings are **not in the context of a murder or a murder investigation**. Though there are exceptions, the precise *context* of police killings hasn't had enough study, and definitely not enough to get reliable statistics.

Ratios are hopeless

Really, though, it's not an issue of lacking data. Philosophically, consider the any possible ratio like

# of people of a race killed by police# of times act X committed by a member of a race.

$$\frac{\text{\# of people of a race killed by police}}{\text{\# of times act } X \text{ committed by a member of a race}}.$$

For what act XX does this really measure police bias? I think it's pretty clear that **no such act exists**, even if we could measure it. Races vary along too many dimensions. There are too many scenarios for police use of force. Bias interacts with the world in too many ways. You just can't learn anything meaningful with these sort of simplistic high-level statistics.

This doesn't mean we need to give up. It just means you need to get closer and try harder. In the next part of this series I'll look at some valiant attempts to do that. They will disappoint us too, but for different reasons.

Data Used:

This post is part of a series on bias in policing with several posts still to come.

The veil of darkness

Measuring police bias using simple ratios doesn't work. You can never cleanly separate the impact of race from other associated factors.

But imagine we had augmented-reality goggles that made race invisible. Suppose we ran the following experiment:

- Have half of police wear race-invisibility goggles for a year.
- Have the other half wear non-invisibility goggles.
- Look at the difference of the two groups.

The police with invisibility goggles would *not* have equal statistics with respect to race. That's because race is correlated with many things other than how people look.

However, the only difference between the two groups of police is if they can see race. Thus, their *difference* would reveals exactly the impact of police bias.

We haven't done this experiment, of course. But we've done a kind of low-tech approximation. Instead of augmented reality goggles, we use the geometry of the earth and sun. Here's the idea: Take all cars stopped by police in some around around 7:15. It will be light in summer, but dark in winter, meaning it's harder to tell the race of the driver. So we ask: does the racial mix of stopped drivers change throughout the year?

Stopping data

This was first studied in Grogger and Ridgeway in 2006 with a small and unreliable dataset. A heroic follow-up was done by Pierson et al. in 2020. They filed public records requests with 50 states and over 100 municipal police departments. (You do not, it appears, screw around with Pierson et al.) They ended up with a database of around 95 million stops from 21 state patrol agencies and 35 municipal police departments.

Sure enough, they find that the fraction of stopped drivers who are black is lower when it's dark. Their results (eyeballing a graph) are:

	Black
% stopped when it is light outside	~25%
% stopped when it is dark outside	~22%

I think this is both more and less than it first seems.

This is around a 12% drop, which might seem small. But I think it suggests a larger bias: Reduced light has a modest effect on officers' ability to see race. Often, it changes nothing, either because race was *already invisible* in daylight, or *remained visible* despite darkness. Roughly speaking there are four cases:

	Light Outside	Dark Outside
Case A	Visible	Visible
Case B	Invisible	Invisible
Case C	Visible	Invisible
Case D	Invisible	Visible

Case A might happen near bright streetlights. Case B might happen if the driver is far away from the officer. The measured effect is coming *only* from case C (and partially cancelled by case D). Imagine switching from a regime where officers *always* saw race to one where they *never* saw race. Then the effect – if real – would probably be much larger.

But is the effect real? It looks conclusive at first. But there are three major problems:

- First, sunlight might change **driver demographics**. Some race might have more jobs tied to daylight hours, meaning driving times vary throughout the year. Or, some race might have more parents, meaning a greater sensitivity to school being out in summer.
- Second, sunlight might change **driver behavior**. Maybe some race speeds less when it is dark. Maybe people consume alcohol at different hours.
- Third, sunlight might change officers' access to **information other than race**. Broken taillights might be easier to detect when it's dark. Contraband or domestic disputes might be easier to detect during the day.

It's not clear what effect these factors could be having. They could be making the effect look larger than it is. They could also be making the effect look smaller. They might cancel out. Since there's no reason for them to point in either direction, I give higher probability to the bias being real than not.

So this data is weird. I think it gives fairly *weak* evidence of a fairly *large* effect. There's a huge amount of uncertainty due to all the uncontrolled factors.

Stopping data

- Record all drivers who are stopped by police around 7:15pm during a whole year. The fraction who are black is around 12% lower during times of the year when it is dark outside.
- Suggests a larger bias than 12% since changing light only affects a fraction of situations.
- Could be confounded by three other things that might change during the year: Driver demographics, driver behavior, and officers' access to information other than race.

Search data

Pierson et al. also look at a 8 state patrol agencies and 6 municipal police departments that provide extra data. For these, we know if the police decided to perform a search of the car. The results are as follows.

	Black	White	Hispanic
% searched, state patrol agencies	4.3	1.9	4.1
% searched, municipal police departments	9.5	3.9	7.2

There are big differences, but of course this doesn't prove anything (yet) because we don't know searches were performed because of race or because of other factors correlated with race.

But here things get interesting. There's also data on *if officers report finding contraband.*

	Black	White	Hispanic
% of searches yielding contraband, state patrol agencies	29.4	32	24.3
% of searches yielding contraband, municipal police departments	13.9	18.2	11.0

The obvious explanation is that police tend to require stronger indicators to trigger searches of whites than they do for non-whites, so searches of white people yield contraband more often.

The observed bias is smaller for whites vs. blacks (8% or 30%) than for whites vs. hispanics (31% or 65%). We are also observing the "full effect". We can assume that police are aware of the race of (almost) all drivers the step. This isn't just a fraction of the true bias, like with the stop data above. State patrol agencies show less than half as much bias as municipal police departments.

While this is decent evidence, it's not completely conclusive. It's possible that race-blind policing could produce data like this. Here's three examples:

1. Different "base rates". Imagine that some fraction of cars are randomly chosen to be searched (race-blind). Some races might be more likely to carry contraband. Drivers of that race would have a higher "hit rate" than others, even though police were not biased.

2. Some races might be more likely to give off *signals* of contraband. Imagine a world with two drugs and two races.

Drug A	Drug B
Smoked at home	Smoked in car
Lingering smell on clothing	Smell dissipates quickly

Drug A	Drug B
Used by 50% of green people	Used by 0% of green people
Used by 0% of purple people	Used by 50% of purple people

Suppose police are race blind, and always search when they smell either drug. There will be more searches of greens than purples, but searches of purples will more often be successful. (Aside: While the police treat each individual the same one might argue the *policy* to search when smelling drug A is wrong or "racist" since it gives so many false positives and the burden falls on green people. This is a complex issue I'll come back to later.)

1. Sometimes police are suspicious but don't have enough evidence for a mandatory search. In these cases police may ask drivers to agree to "voluntary" searches. This is (deliberately) phrased in a way that the driver may think it is mandatory. Some races might be less likely to agree to such searches. This would tend to increase the "hit rate" for that race since the searches that do occur tend to happen with strong evidence.

While these effects could distort the data, there's no reason the distortion would be in any particular direction. Such effects could create an illusion of bias when there is none. Or they might be masking even *stronger* bias.

I don't rate these mitigating effects as super strong. They could exist, but they are less plausible than those that could explain the twilight data. It also seems like their effect would be relatively modest.

So, I think this provides moderate evidence in favor of police bias. It's no smoking gun, but it's a glimpse of something that's very hard to see. Also, while it relates to police *bias* it tells us nothing about how that bias relates to police *violence*.

Search data

- Police find contraband 8-30% more often in searches of whites than blacks and 31-65% more often than hispanics. This is consistent with police applying a higher threshold of evidence to trigger a search of whites.
- Data could be confounded by different rates of carrying contraband, different rates of giving evidence (or false evidence) of contraband, or different rates of agreeing to "voluntary" searches.

What would it take?

I can imagine someone who believes police are racially biased grinding their teeth at this point. "Simple ratios show a bias, but you don't believe them.

Fine. So you look at the effect of darkness. That *also* shows a bias, but you worry it's confounded. OK! Then you look at search rates. Since you don't believe the bias they show, you check the *hit rate* of searches, which are… biased. Always you invent stories about confounders. What does it take!?"

My response is this: We are making progress. I give *zero* weight to things like number of people killed per capita. The data discussed above isn't totally conclusive, but it definitely should be given *some* value. (In Bayesian terms, you should update your prior in the direction of bias.)

As for what it takes, there's some more data that could help a *lot* with understanding the impact of confounders here. That would be to repeat the analysis with different groupings of people other than race. Does the fraction stopped of drivers who are male change when it's dark? What about the fraction of the old? Those in poverty? Those who are politically conservative? Pick any group that police can't see or don't have a bias around.

If we verified that the fraction of stopped drivers who were old/male/poor/educated/conservative did *not* change with darkness but the fraction who were black *did*, then the confounders probably aren't too much of a problem. (It's possible in principle that race is confound but not these other groups. But since race is correlated with everything I doubt it.)

Of course, analyzing some data is easy! The hard part will be assembling a database of millions of police stops with tags for these other driver attributes. Good luck with that.

This post is part of a series on bias in policing with several posts still to come.

Pragmatic reasons to believe in formal ethics

Here's a "low-brow" take on ethics that's worth taking seriously.

> Ethics isn't going to save the world. We don't need more "calculations" about the right thing to do. We need people to stop doing *obviously* wrong stuff. Ethics is boring and irrelevant to everyday life. Stop the obsessive navel-gazing and go engage with the real world.

There's a lot that's right about this: In practice our decisions aren't usually influenced by ethics, but by *habits* and *incentives*.

Say you're walking across a park and you consider a shortcut. You ask: Will you hurt the grass? Are there insects? Will you hurt them? What's the moral weight of an insect's life? How does it compare against a small convenience for you? Will other people follow you? Are you responsible if they do?

Living like this would be paralyzing. Almost all the time, we use habits or heuristics to make decisions, not ethics.

Even when people *do* use ethics, we often spend it on problems that just aren't that important. I mourn the hours I've spent on which plastics are recyclable. (It turns out: none!) It's easy to get sucked into an argument about how some corporation named a product.

And of course, lots of people are jerks and just don't care about ethics. Most of the time, ethics don't much influence behavior.

What matters is *incentives*. We are bad at reasoning, but good at taking care of ourselves. That's because it doesn't hurt when you get ethics wrong. If you want to solve climate change or animal welfare or whatever, don't preach at people – make it so no one *needs to think* about ethics. Apply a tax, put up a fence, create legal penalties. Let everyone follow the scent of what's good for us instead of futilely hoping people will both figure out what's best and then actually do it.

If you have to choose between living in

1. a society with a mediocre theory of ethics but well-crafted incentives, or
2. a society with an enlightened theory of ethics but poor incentives,

then I suggest you choose society #1.

Everyone knows that flights emit a lot of carbon. It's also obvious that business seats take up more space than economy seats. A study took the emissions of an Airbus A380 flight from Abu-Dhabi to London and assigned them to passengers in proportion to the area of their seats. They got these numbers:

Mode of travel	Carbon Emitted per person
Business class	2,760 lbs CO
Economy class	520 lbs CO

I've sometimes suggested that if we're worried about CO emissions, maybe we should avoid business class. Most people resist this argument. These conversations go like this:

Dynomight: Switching two long-haul round-trip flights from business to economy saves almost as much carbon as not driving for a year (10,000 lbs CO in the US on average).

Other Person: That doesn't make sense. The planes are *already* flying and *already* have a fixed configuration of seats.

Dynomight: Well… true… but business-class seats exist only because people buy them.

Other Person: If I don't buy a business-class seat, someone else will anyway. Or they'll upgrade someone from economy.

Dynomight: But surely, on the margin, buying business-class seats creates an incentive for airlines to make more of them?

Other Person: Suppose you're right. Even if I on an *individual* level refuse to buy these seats, that won't be enough to change the way airlines configure planes.

Forget who is right. Why do these conversations reach a standstill? The *facts* aren't in dispute. Our *values* are usually similar too, in that both people care equally about climate change.

I think the conflict is due to *different ethical systems*. There are subtle philosophical issues here. It's *true* that the planes are already flying, and it's *true* that that one person won't change the way airlines do business. Maybe just *not flying* in business is pointless, and you should boycott airlines that sell business seats at all. Maybe individual action on these issues is pointless, and you should spend your effort lobbying for a carbon tax or something.

As long as we just keep talking about planes and seat sizes and carbon, these conversations will never converge. The only way is to step back and state how you define "right" and how it leads to your conclusions. Ethical reasoning is a third opportunity for disagreement even when people agree on facts and values.

Where do habits come from?

You can't live your life constantly thinking about ethics. But you can step back once a year and consider: Do you want to change how you interact with loved ones? Volunteer? Donate money? Recycle? Participate in political action? Ethics are important when *designing* habits.

Is it better to spend more time reading to your kids or to help a campaign to improve soil quality? Commonsense ethics simply has no answer, because these choices make the world better in such different ways.

But these choices matter. History has shown over and that if you want to improve something, you should first measure it. Your life and time are finite. Different choices really do have enormously different impacts. But there's no way to compare these choices without something close to a fully-realized ethical theory.

Where do incentives come from?

Even more importantly we need ethics when *creating incentives*. Lower speed limits save lives but cost time. Regulating pesticides in produce makes them more expensive, which might decrease how much people eat. Aggressively approving medical treatments makes them available earlier but has higher risk. Closing schools during a pandemic saves lives but hurts children's future prospects. Ugly tradeoffs are everywhere and we can't hide from them.

Policy choices are implicitly choices among ethical theories.

Say you want to participate in democratic government, but you only believe in commonsense ethics. The problem is that there are many complex issues. You can't examine more than a small fraction in detail. And even for the issues you *do* look at, you'll often conflict with others, since different people have different intuitions for these types of difficult tradeoffs.

But ethics *scale*. You can participate in a single conversation about what ethical system society should adopt. If you have a formula to calculate "how good" a given world-state is and trust that policymakers are always applying that formula, then you don't *need* to inspect every random policy. If you don't trust policymakers, it's still a much easier to check if The Formula is being applied correctly, rather than every issues starting from a blank slate.

This is why public health has invented concepts like disability adjusted life years and quality adjusted life years. These aren't egghead concepts designed to complicate things. It's simply impossible to make policy decisions in most real cases without some theoretical foundation.

Commonsense ethics have a bad track record

Not that long ago, most people believed that women shouldn't vote, homosexuality should be illegal, and it cannabis users should be in jail. In 1959, 96% of Americans disapproved of interracial marriage. Not long before that, many people believed slavery was acceptable. It's easy for us to believe monstrous things.

Given this, you have to wonder: Won't people in the future be horrified by some of our beliefs? Unless this is the moment when we finally got everything right, the answer is yes. You have to worry what those beliefs are.

But people in the past weren't all the same. There are lots of examples of people questioning the beliefs we now find so appalling. What set these people apart? I suspect it wasn't as much that they tried harder to *be good* but that they tried harder to think about goodness *systematically.*

Calibration

Putting numerical values on lives feels a little cold-blooded. In movies you often see a character yell "*You can't put a number on human life!*" as music swells in the background. This reflects an understandable concern that a formal ethical system might lead to terrible conclusions in real life.

While this is *understandable*, I think it's backwards. In my fantasies, these conversations would go like this:

Egghead: If we blow up the dam that will save around 1000 quality adjusted life years. Let's do it.

Superhero: Damn you, egghead! *Who are you to decide what a human life is worth?*

Egghead: I don't.

Superhero: You just said…

Egghead: Everyone puts numbers on human lives all the time. Every time you pay more for a safer car or take a more dangerous but better paying job, you are doing that. I'm just averaging the choices real people make all the time.

Superhero: Oh. Well… Fuck it, blow up the dam, I guess.

This is the right way to think about ethics. We don't come up with a formal ethical system and then derive the consequences for everyday life. No, we look at the choices in everyday life we believe are clear and derive an ethical system to *generalize* these.

There's a tension between (a) deriving an ethical system to generalize from commonsense ethics and (b) hoping that ethical system will reveal flaws in your commonsense ethics. This is a real problem that I don't know how to fully resolve. On the other hand, a child might have this set of commonsense ethics:

- It's bad to hit my parents.
- It's bad to hit my friends.
- It's bad to hit my teacher.
- It's good to hit Kevin, I hate that guy.

Summary

While formal ethics have limited use for everyday decisions, they still have several practical uses:

- **Conflicts:** They can help *explain or resolve* certain *conflicting beliefs.*
- **Personal choices:** They can help when choosing *priorities and habits.*
- **Scalability:** Ethics *scale.* Many related tradeoffs show up repeatedly when choosing policies so it's worth trying to resolve them once and for all.
- **Time Robustness:** They make use more *future-proof* against beliefs that future generations will see as wrong.

Simpson's paradox and the tyranny of strata

It's hard to get into Oxford. Is it easier if your parents are rich?

Population	State Students Accepted	Independent Students Accepted
Everyone	20%	28%
Great grades	46.1%	50.5%

In 2013, The Guardian showed that students from expensive independent schools were accepted more often than students from state schools. If you limit things to just students with strong grades (3 A* at A-level) the bias reduces, but doesn't disappear.

The Conversation later noticed something important: It's much easier to get into Oxford if you apply to Classics (45% accepted) rather than Medicine (21%). If *each department* admits at equal rates, you'll still have the appearance of an *overall* bias, if state students are more likely to apply to Medicine.

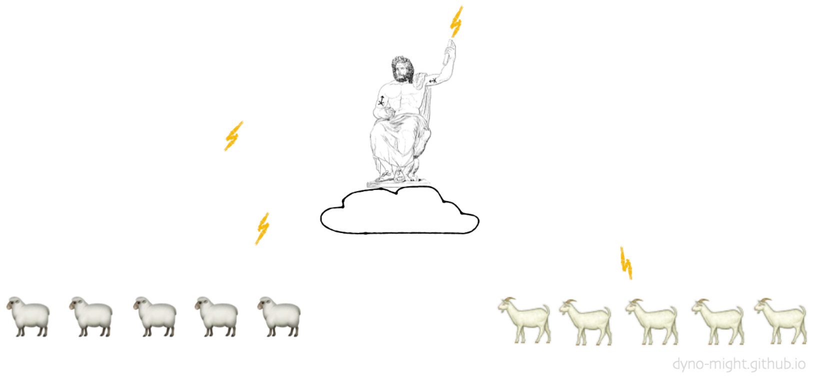
Simpson's paradox is the name for situations like this, where the same data can seem to tell entirely different stories. Most people think of the paradox as an odd little curiosity, or perhaps just a cautionary tale about interpreting data correctly.

But you shouldn't see Simpson's paradox like that. Rather than a little quirk, it's actually a manifestation of a deeper and much stranger issue. It's less about the "right" way to analyze data and more about the limits to what questions data can answer.

How does Simpson's paradox work? What is this deeper issue? You don't need to be a data scientist to understand all this. In this post, I'll illustrate both Simpson's paradox and the issues beneath it using a little cartoon about lighting bolts and farm animals. This will use no statistics and (basically) no math.

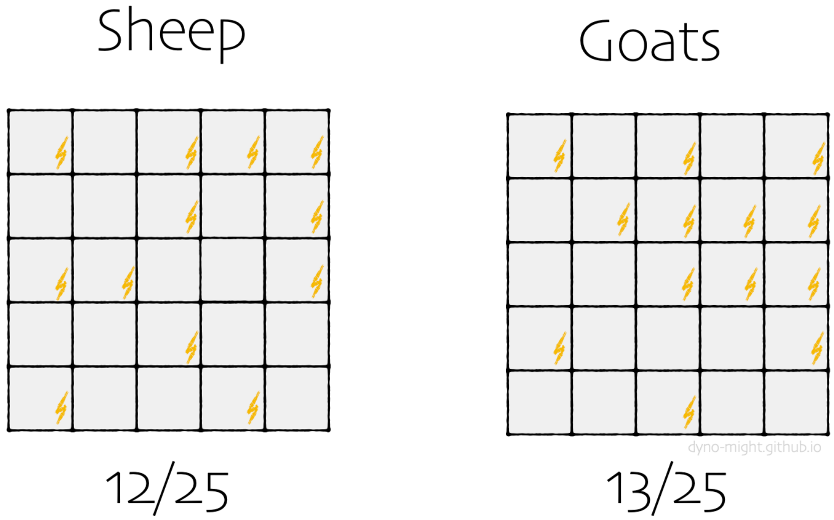
I Zeus

Imagine that you are a mortal shephard living near Olympus with a flock of sheep and goats. Your neighbor, the thunder god Zeus, is a jerk. He has started zapping your animals with lightning bolts.



He's not trying to kill them; He's just bored. (Transforming into animals to seduce love interests gets old eventually.)

Anyway, you wonder: Does Zeus have a preference for shooting sheep or goats? You decide to keep records for a year. You have 25 sheep and 25 goats, so you use a 5x5 grid with one cell for each animal.



Sheep

⚡		⚡	⚡	⚡
		⚡		⚡
⚡	⚡			⚡
		⚡		
⚡			⚡	

12/25

Goats

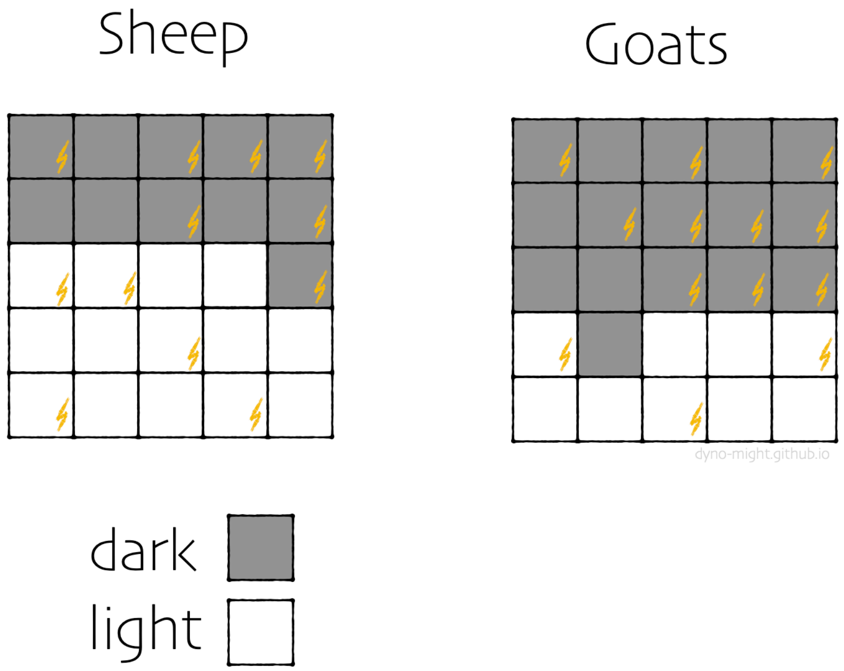
⚡		⚡		⚡
	⚡	⚡	⚡	⚡
		⚡	⚡	⚡
⚡				⚡
		⚡		

13/25

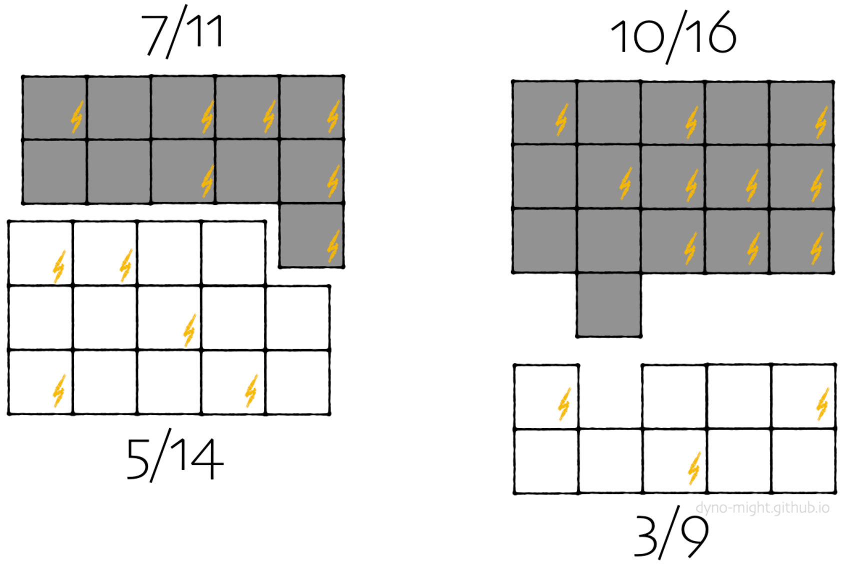
At first glance, it seems like Zeus dislikes goats more than He dislikes sheep. (If you're worried about the difference being due to random chance, feel free to multiply the number of animals by a million.)

II Colors

Thinking about it, it occurs to you that some animals have darker fur than others. You go back to your records and mark each animal accordingly.



You re-do the analysis, splitting the animals into dark and light groups.



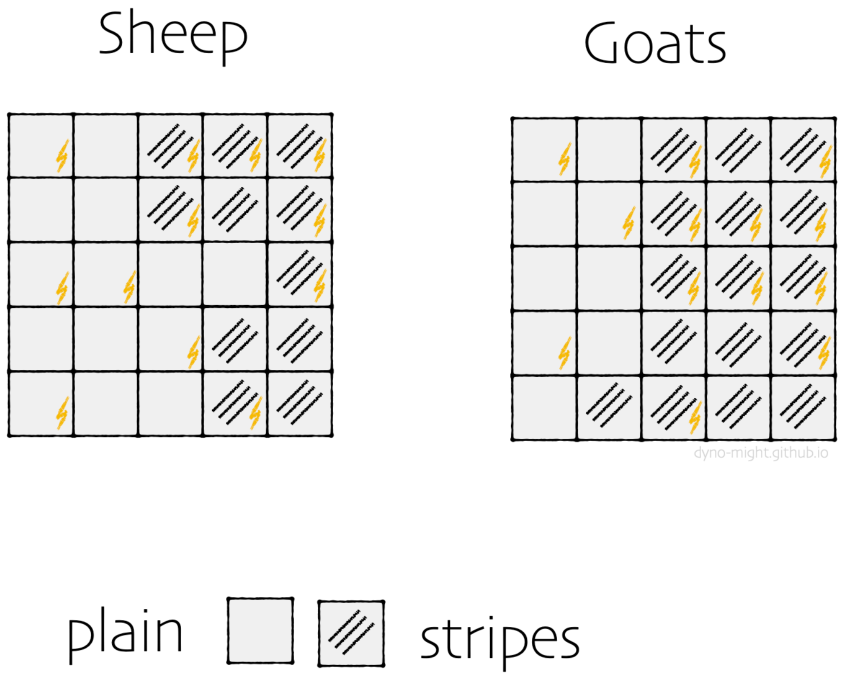
Overall, sheep are zapped less often than goats. But dark sheep are zapped *more* often than dark goats (7/11 > 10/16) *and* light sheep are zapped more often than light goats (5/14 > 3/9). This is the usual *paradox*: The conclusion changes when you switch from analyzing everyone to analyzing subgroups.

How does that reversal happen? It's simple: For both sheep and goats, dark animals get zapped more often, and there are more dark goats than dark sheep. Dark sheep are zapped slightly more than dark goats, and light sheep are zapped more than light goats. But dumping all the animals together changes the conclusion because there are so many *more* dark goats. This is an example of Simpson's paradox. Group-level differences can be totally different than subgroup differences when the ratio of subgroups varies.

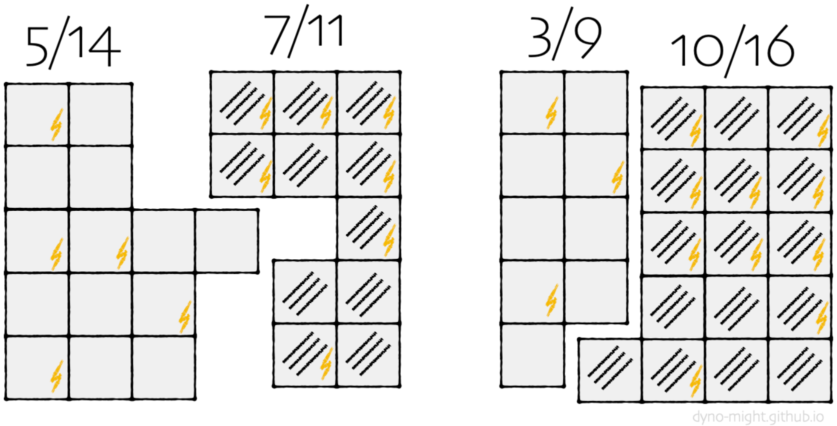
If you just want to understand Simpson's paradox you're done! This probably seems like a weird little edge case so far. But let's continue.

III Stripes

Thinking even more, you notice that many of your (apparently mutant) animals have stripes. You prepare the data again, marking each animal according to stripes, rather than color.



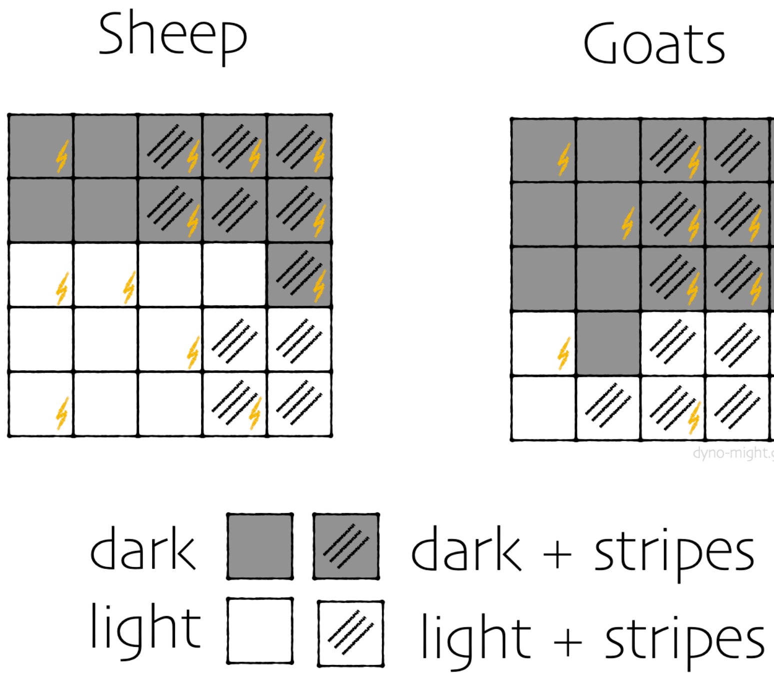
You wonder, naturally, what happens if you analyze these groups.



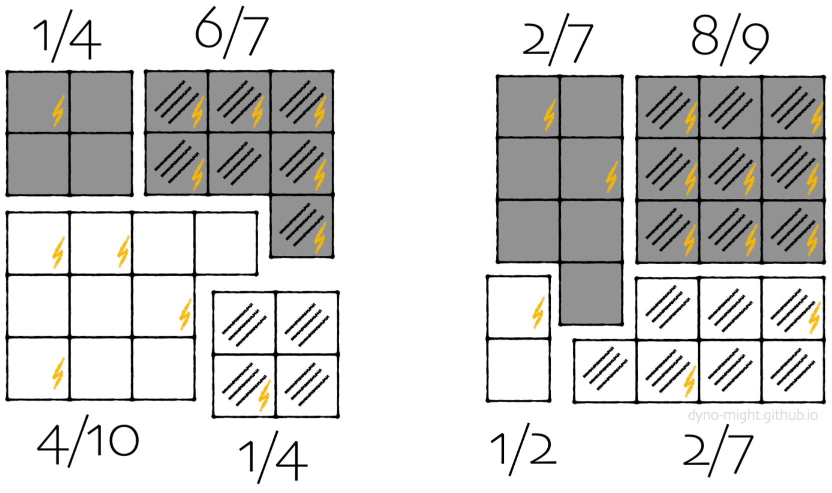
The results are similar to those with color. Though sheep are zapped less often than goats overall (12/25 < 13/25), plain sheep are zapped more often than plain goats (5/14 > 3/9), and striped sheep are zapped more often than striped goats (7/11 > 10/16).

IV Colors and stripes

Of course, instead of just considering either color or stripes, nothing stops you from considering both.



You decide to consider all four subgroups separately.



Now sheep are zapped *less* often in each subgroup. (¼ < 2/7, 6/7 < 8/9, etc.)

When you compare everyone, there's a bias against goats. When you compare by color, there's a bias against sheep. When you compare by stripes, there's also a bias against sheep. Yet when you compare by both color *and* stripes, there's a bias against goats again.

Type of animals compared	Who gets zapped more often?
All	Goats
Light	Sheep
Dark	Sheep
Plain	Sheep
Striped	Sheep
Dark Plain	Goats
Dark Striped	Goats
Light Plain	Goats
Light Striped	Goats

The same data gives lots of different results! How can this happen?

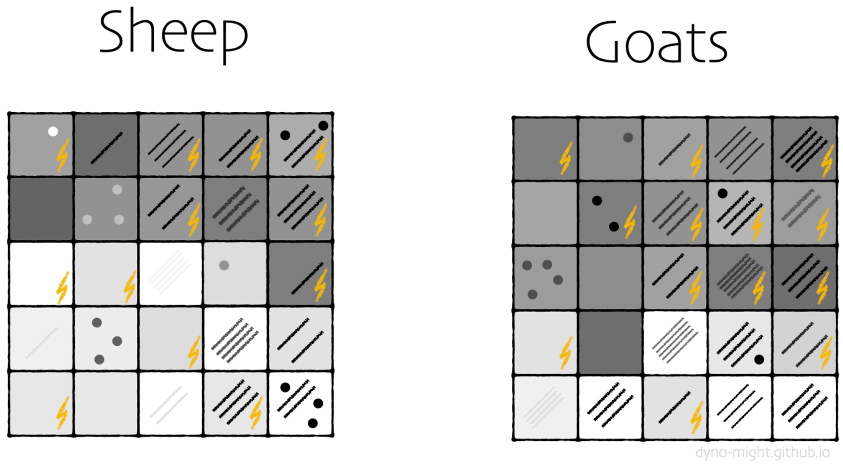
To answer that, it's important to realize that *anything can happen.* When you split data into subgroups, it's possible to find any biases. These could reverse (or not) when you split those subgroups again. In the table above, *any* sequence of goats / sheep is possible in the right-hand column.

But *how*, you ask? *How* can this happen? I think this is the wrong question. Instead we should ask if there is anything to *prevent* this from happening. There are a huge variety of possible datasets, with all sorts of different group averages. Unless there is some special structure forcing things to be "orderly", essentially arbitrary stuff can happen. There is no special force here.

V Individuals

So far, this all seems like a lesson about finding the right way to analyze data. In some cases, that's probably true. Suppose read that Prestige Airways is more often delayed than GreatValue Skybus. Looking closer, you notice that Prestige flies mostly between snowy cities while Skybus mostly flies between warm dry cities. Prestige may have a better track record for *all* individual routes, but a worse track record overall, simply because they fly difficult routes more often. In this case, it's probably correct to say because Prestige is better for any flight you might take, Prestige is more reliable.

But in other cases, the lesson should be just the opposite: There *is* no "right" way to analyze data. Often the real world looks like this:



There's no clear dividing line between "dark" and "light" animals. Stripes can be dense or sparse, thick or thin, light or dark. There can be many dark spots or few light spots. This list can go on forever. In the real world, individuals often vary in so many ways that there's no obvious definition of subgroups. In these cases, you can't beat the paradox. To get answers, you have to make arbitrary choices, even though the answers will depend on the choices you make.

Arguably this is a *philosophical* problem as much as a statistical one. We usually think about bias in terms of "groups". If prospects vary for two "otherwise identical" individuals in two groups, we say there is a bias. This made sense for the airlines above: If Prestige was more often on time than GreatValue for each route, it's fair to say Prestige is more reliable.

In a world of *individuals*, this definition of bias breaks down. Suppose Prestige mostly flies in the middle of the day on weekends in winter, while Skybus mostly flies at night during the week in summer. They vary from these patterns, but never enough that they are flying the same route on the same day at the same time at the same time of year. If you want to compare the two, you can group flights by cities or day or time or season, but not all of them. Different groupings (and sub-groupings) can give different result. There simply is no right answer.

This is the endpoint of Simpson's paradox: Group level differences often really *are* misleading. You can try to solve that by accounting for variability within groups, and there are lots of complex ways to try to do that I haven't talked about. However, none of them solve the fundamental problem that it's not clear what bias means when every individual is unique.

So by all means, beware of Simpson's paradox. Try splitting your data and see what happens. But keep in mind that you might be trying to estimate a quantity that's not defined. No data will solve that problem for you.

Why I'm skeptical of universal basic income

Universal basic income (UBI) is an odd duck. Proponents range from futurists to libertarians to social democrats. Why this weird range of people?

There are different versions of UBI, with different motivations. People don't get how *contradictory* these are. Libertarians and social democrats might both support a fuzzy idea of "UBI" but they are unlikely to agree on specifics.

I don't think it matters though. I'm skeptical of these *interesting* versions of UBI. The only realistic option is another, *boring* version. Why isn't it discussed more? The more you think about it, the more of a riddle that is.

The libertarian argument for UBI

A libertarian argument for Universal Basic Income might go something like this:

> The government is stupid and inefficient. There's medicare, medicaid, social security, food stamps, "other food stamps", welfare, "other welfare", and unemployment benefits, each with a bureaucracy. Did you know that each item in a supermarket has been classified as buyable with food stamps or not? (Energy drinks are usually OK, toothpaste is not.) And there's a *different* classification for the "other" food stamps? (No spices or drinkable yogurt.) And that if you lose your job you need to provide evidence each month that you spent 80 hours looking for a new one? These expensive programs are often miserable for the people they're supposed to help. Let's stop micromanaging people's lives, cut the fat, and *just give people money.*

I see the appeal, and if you want to reorganize cash transfers like social security and call them "UBI", that's fine with me! But it's unlikely UBI can replace medicaid.

To see why, consider schools. Sometimes the government runs them. Sometimes the government merely pays for privately run "charter" schools. Sometimes the government pays for "vouchers" that subsidize a fraction of expensive private schools.

Why do things this way? Why not give parents (or kids?) the money and let them spend it however they want?

The answer is obvious, even if we don't usually say it explicitly: We don't trust people and we want to control them. Educating kids is an investment that benefits everyone in the future. If we just gave parents cash, I'm sure many would arrange for a better education than the government does now. But I'm also sure many would spend it on wacky religious schools, or crystal healing seminars, or beer.

To be clear, no real UBI proposals would change funding for schools! But that's kind of the point. That this is *so obviously* a bad idea shows two things:

1. It's sometimes valid to constrain how money is used. We can't "just give people money" without looking at the details. The question is *how much* control to assert in each case.

2. Control isn't free. Some kids might do better in larger classes with higher-paid teachers. Some might prefer individual tutoring and peer instruction. Some might learn to read at home but need more instruction on math. If parents had flexibility, they could probably use the money more effectively. Restrictions to prevent worst-case outcomes also prevent these gains.

Again, schools are just hypothetical. What's not hypothetical is *healthcare*. For example, Charles Murray suggests rolling all government entitlements and healthcare expenditures together into a grant of $12k, with no tax increase.

I don't think this would work. Say we stop subsidizing healthcare and just give people money. No matter what you do, some won't save and won't buy health insurance. (If you doubt this, you've met a different species of human than I have.) Then, you have this difficult question: When one of these people gets sick and shows up at a hospital, do you let them die?

Intellectually, I take this question seriously. Think about it like this: Suppose you can choose to be born into Safeland, where everyone is secure but bored, or Dangerland, where people have fun but risk getting killed. It's totally reasonable for someone to choose Dangerland. A tough-love "buy insurance or risk death" policy allows people to optimize for their own utility function.

Practically, though, there is no question. Society has decided: It's not acceptable to let people die, and we need some kind of emergency backstop program. But when you create that backstop, you just got back into the healthcare subsidy business, and you've moved away from *just giving people money*, since any funding for the backstop could have been an unrestricted payout to someone instead.

To avoid this problem, Murray suggests that $3k of the $12k must be spent on health insurance. This is a compromise on the dream of just giving people money, but let's ignore that. It also reduces but doesn't eliminate the Dangerland problem. What if people need a treatment that wasn't included in their insurance? Do you have "death panels" to decide what "essential" procedures need to be included? What happens if these can't be offered for $3k? The average person in the US uses $11k worth of healthcare.

Because of issues like these, UBI would mostly have to come *on top* of targeted programs. That's *increasing* the scope of government, not decreasing it. That might be a good idea, but it's not the libertarian dream. If that's what you want to do, you need a different argument.

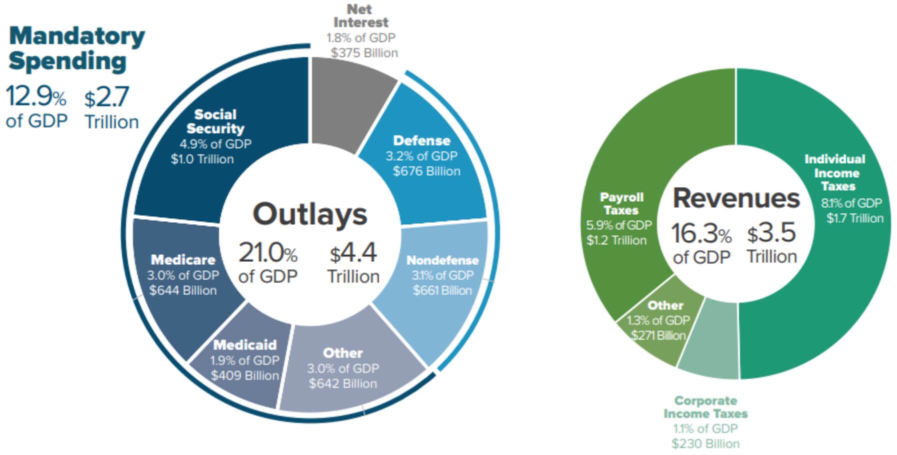
The Silicon Valley Argument for UBI

Here's a different argument for basic income.

> Artificial intelligence is coming, and jobs going extinct. Without action, soon all income will go to the wealthy or to the rare people who work on robots/AIs. But jobs getting automated should be good news! There will be plenty of money. In a sane world, when an AI replaces a truck driver, we'd keep paying them without asking them to come to work. For all of history, people have spent their lives mostly scrambling to survive. It's time for us to be liberated, to spend our time on art, science, politics, whatever. Let the human spirit soar.

I have three responses to this.

First, there isn't plenty of money, not yet. Giving 256 million adults the average truck driver salary of $67k per year would cost $17 trillion. That's 4 times the entire current federal budget:



It's almost as large as the entire *economy* ($21 trillion). Giving everyone a modest $12k per year would cost $3 trillion. That's more than all current benefit programs combined, but not impossible. Suppose we make some difficult cuts, add some new taxes and make it happen. Do we expect anyone who previously had a good job to be happy with $12k? I'm sure it's *helpful*, but it's no *substitute* for a job.

Second, this isn't how economic revolutions worked in the past. Everyone knows the story: We moved from farms to manufacturing to services as each sector got automated. Keynes 90 years ago predicted that we would soon only need to work 15 hours a week. He was right that existing jobs would mostly disappear. What he missed was that we'd invent so much new stuff and then decide we couldn't live without it. Are we *really* sure this time is different? A possible counter-response is that “every trend continues until it stops”. Previously, there were obvious productive things for people to shift their effort to. Now, that’s not clear. My counter-counter-response is: Are you *sure*?

Both of these first two responses boil down to “your argument for UBI is premised on some future that’s not here yet”. So you might ask: Say it’s 30 years from now, the economy is 50× bigger and all the jobs are all gone. What then?

A **third** response is that this argument for UBI assumes jobs are all about money. Joe and Jill Biden are worth around $10 million. So why does Jill Biden teach English at Northern Virginia Community College? People get much of their identity and meaning in life from their jobs. We are sort of like border collies: Without a role to play, we go crazy. Most people don’t want a handout, they want to feel that they’ve *earned* what they have.

To be sure, *some* people, if their current job was replaced with a UBI, would devote themselves to other passions. (Can anyone doubt that a thousand rationalist-adjacent blogs would bloom?) But others would get bored and depressed. Don’t believe me? Check out Financial Independence / Early Retirement (FIRE) forums. There are tons of people who worked hard, saved money, and retired young. They find that relaxation and “fun” get old quickly. Some occupy themselves with volunteer work or creative projects, but others just spiral into depression until they eventually decide to take another job.

Maybe society shouldn’t be like this. I don’t really *like* this connection between work and value. Maybe social norms just need to adjust. Maybe FIRE folks would love retirement if their friends were also free to hang out in the middle of the day. Perhaps, but I think we also need to prepare for the possibility that this doesn’t work.

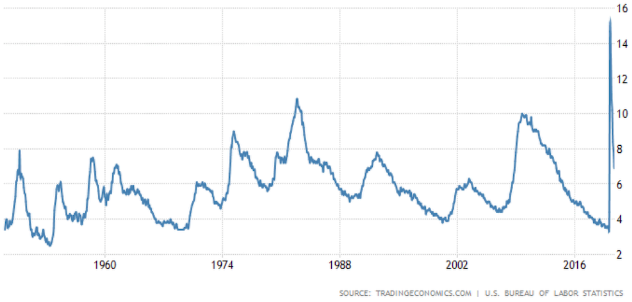
Here, a UBI proponent might object. “Well, look! Suppose this jobless world comes as I predict. What then? At least I’m proposing a solution! All you’re doing is claiming the solution won’t work. What do you suggest instead?”

I agree. We don’t have a button to stop the world from shifting under our feet. If jobs really do go obsolete, we’ll have to figure something out. I’d support *experimenting* with UBI to see if it works or not. But I’d suggest we *also* experiment with how to make jobs stick around. It’s easy to open a building downtown with a “fake jobs for insecure people” sign, but this wouldn’t work. We need to believe our jobs are valuable.

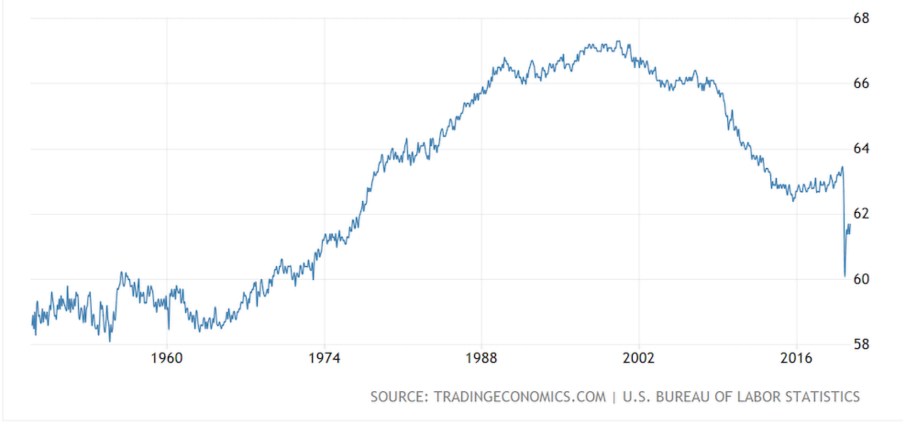
Aside: How to create jobs that look useful? In a future where AI is so powerful that normal human jobs are gone, shouldn’t we be worried about AI risk? Can we create bottlenecks in processes that have to be filled by humans? How to do this in a way that *actually* reduces AI risk is a hard problem, but surely we can at least make it look plausible?

Is the world changing under our feet, though? *Are* jobs going obsolete? Sure, the future is unknowable, we can’t rule it out, blahblahblah. But is it a pressing issue we need to address now, or a long-term concern? If it's the latter, what's the hurry?

Here's a plot of the unemployment rate going back to 1950. I don't see any clear trend.

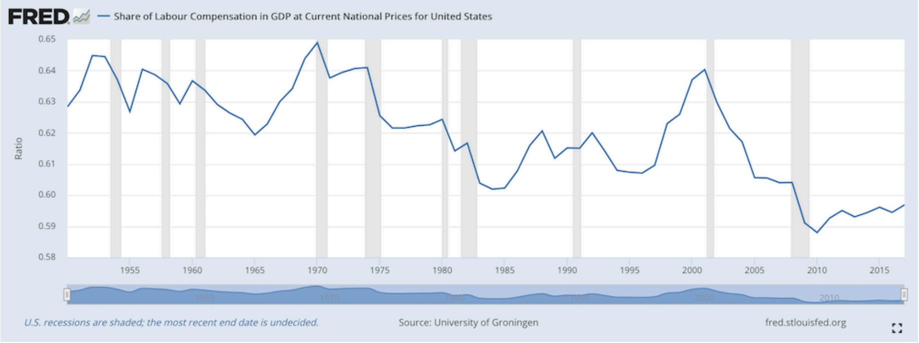


Of course, the unemployment rate only counts people who are *trying* to get a job. Here's a plot of the labor force participation rate (the fraction of 16-64-year-olds who are either working or trying to get a job.)



This strange curve makes a bit more sense when broken down. (The strata refuse to be ignored.) If you dig into the data you'll find participation from men has steadily decreased but participation from women increased between 1950 and 1995 and then plateaued. Is that technological or sociological change? I'm not sure.

As another view, here's a plot of the percentage of national income that goes to labor. If jobs are becoming increasingly irrelevant to the economy, this should be in decline.



Share of Labour Compensation in GDP at Current National Prices for United States

U.S. recessions are shaded; the most recent end date is undecided.

Source: University of Groningen

fred.stlouisfed.org

There's a decline from around 64% to 60%. But the biggest decline was between 1970 and 1985. It's actually rising slowly over the past 10 years and is currently only 0.5% less than in 1984. It's not obvious that AI is the culprit here.

For some problems, we need to think ahead. Climate change is a good example: We need a massive investment to deal with it, but we understand pretty well what we're dealing with, so we can invest effectively. Disappearing jobs are the opposite. We have little idea what it would look like or when it might come. If we wanted to implement UBI, there's no technological hurdle. Maybe there are interesting social questions, but we can't really answer those until the jobless future is here.

Maybe the jobless future is coming, but it's not here yet. I just don't see the argument for massively restructuring the economy *now* to try to address a problem that doesn't yet exist. So I don't—for now at least—buy the Silicon Valley argument for UBI.

The populist argument for UBI

There's one more argument for UBI, one with a charming simplicity.

> We should increase redistribution and make it less conditional.

I don't want to debate if this is *correct* or not. Still, we can see that the argument is *coherent*. If the goal is to increase redistribution, then it's logical to leave existing health care programs in place and just increase taxes. It's also logical to do a small-ish UBI since you don't need to replace jobs.

I think this is the "real" UBI. We can't eliminate targeted programs, and we can't do away with jobs. Realistically, if anything called "UBI" were to become law, it would be this version. So why isn't this the most common argument? Why was Andrew Yang's 2020 proposal all about technological change?

There are a few possibilities:

First, maybe Andrew Yang's proposal is about technological change because that is what Andrew Yang believes. Call me wide-eyed and naive, but I guess this is the answer. More broadly, the Silicon Valley UBI argument is *really* appealing to a minority of people. It's doubtful that milquetoast "more redistribution!" would have grabbed enough attention to propel a presidential campaign.

Second, maybe it's a cynical "trick" to get people to support redistribution without realizing it. This is hard to reconcile with the data. One of the great open secrets of American politics is that redistribution is *very popular*. Recently, 53% of *Republicans* agreed that "The very rich should contribute an extra share of their wealth each year to support public programs," while 40% disagreed. Yang's campaign raised $40 million. The odds they are unaware of these polls is zero.

Third, maybe it's branding. Maybe voters only like redistribution when it's described the right way. Well, here are the results of a 2018 poll on healthcare:

Policy	Support	Opposed
Universal health care	56%	33%
Socialized medicine	52%	34%

Turns out, people aren't all complete idiots! They understand that "socialized medicine" is a valid description of "universal health care", and are totally fine with that.

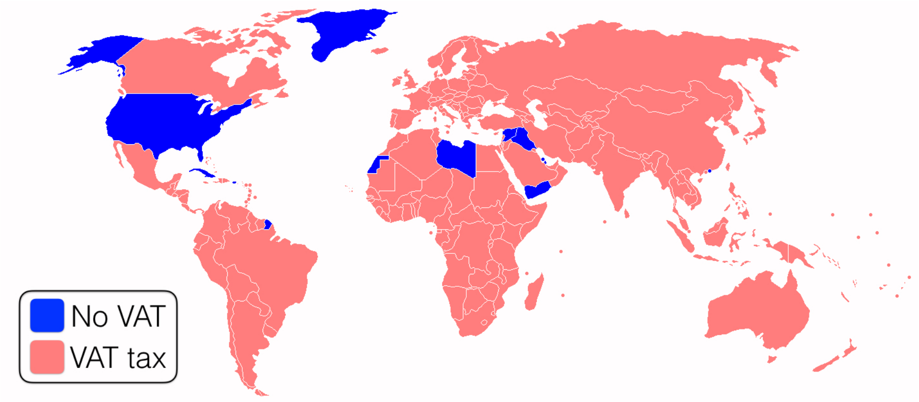
(So why don't politicians who support redistribution win every election, you ask? Because those politicians bundle redistribution with other policies and cultural values that are not popular.)

So what's the deal? Whatever its merits, this argument is dead simple, intellectually honest, and seemingly popular. The biggest apparent disadvantage is that it's kinda… boring? It makes UBI look like just an update to the existing web of taxes and redistributive programs. But so what? Flossing is boring. Cleaning is boring. The only realistic version of UBI is the boring one. Just own it.

Sales tax creates more unnecessary pain than value added tax

It turns out that sales tax has a huge, gigantic, terrible flaw: It punishes specialized businesses. A value added tax (VAT) has no such problems.

The US has sales tax. Most of the planet has VAT.



Maybe it's not the most important issue in the world, but it's just so *clear*. Sales tax is dumb and VAT is better.

Before We Begin

Many people apparently believe that in the US today, sales tax is only paid by final consumers. **THIS IS FALSE**. It varies hugely by state, but the current situation is a hybrid between a "pure final retail consumer only" sales tax and what the toy model below describes. You can debate if it's "sales tax" or "gross receipts tax" or whatever, but it's a fact that *businesses pay tax on business inputs* all the time. You can find proof of this here or here or here or here.

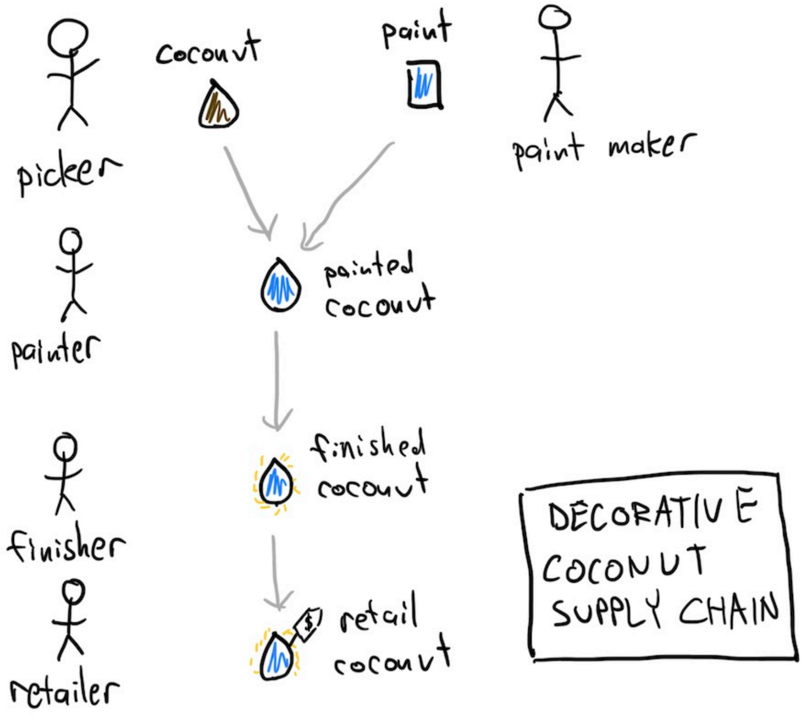
I emphasize that the explanation below is a toy, intended to illustrate in the simplest possible way how specialization gets punished when transfers are taxed in proportion to their values. The current reality not *nearly* this bad due to many complex exemptions, as I discuss at the end. But the flaw described *does* exist and *does* punish specialization.

The Decorative Coconut Model

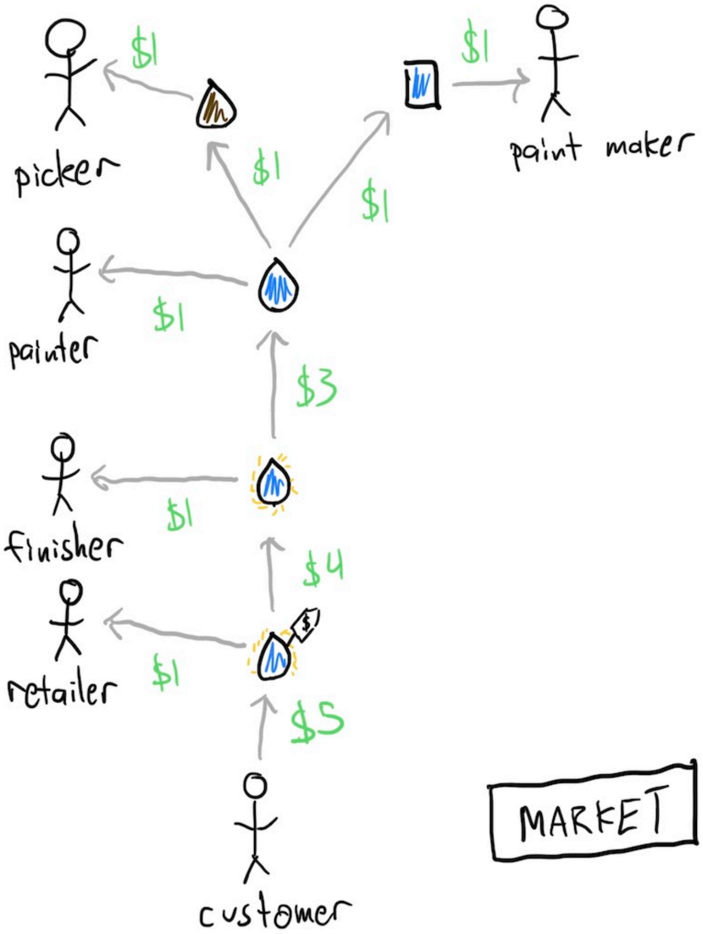
Say you decide to get into the decorative coconut manufacturing business.

You're good at painting coconuts. You find a friend who is good at picking them, and another who's good at making coconut paint. You find a third friend who's a genius with applying finishing lacquer and a fourth who runs a store.

You buy coconuts and paint, apply the paint, then sell to the finisher. He applies lacquer and sells to a retailer.



After negotiating prices, you settle on $1 for a raw coconut, $1 for a coconut's worth of paint, $3 for a painted coconut, $4 for a finished coconut, and $5 retail. This works out to everyone making $1 of profit.



Here's a table showing the accounts:

Item	Cost of inputs	Profit	Price
Raw coconuts	$0	$1	$1
Paint	$0	$1	$1
Painted coconut	$2 (raw coconut+paint)	$1	$3
Finished coconut	$3 (painted coconut)	$1	$4
Retail coconut	$4 (finished coconut)	$1	$5

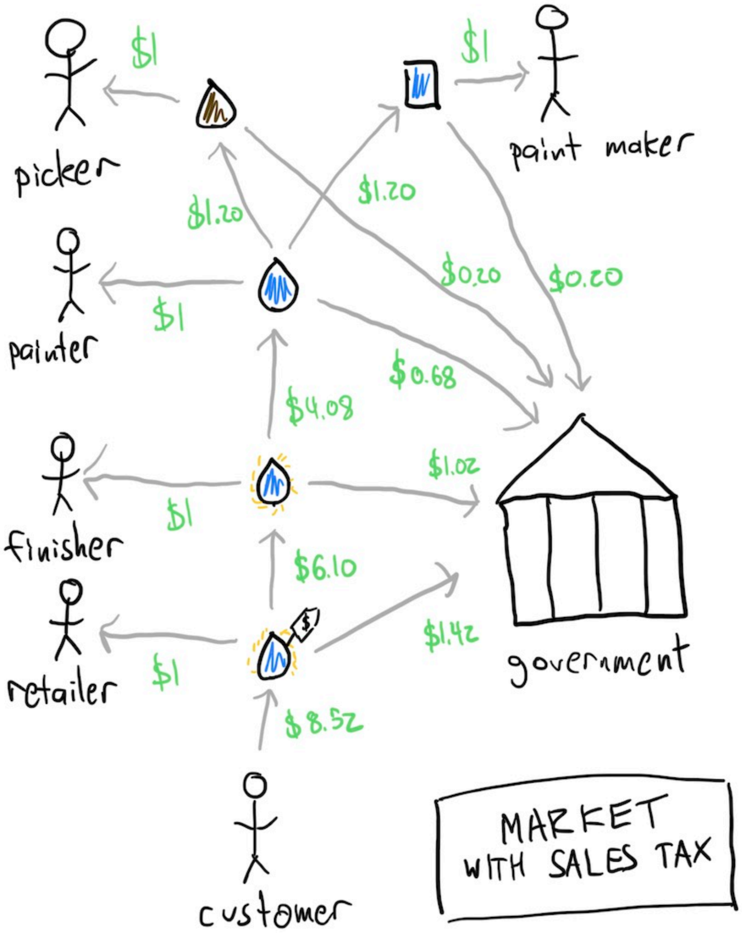
The Sales Tax Regime

For a while, everything runs beautifully. Every day you wake eager to help capture more beauty in coconut form — and then the government announces a

20% sales tax. Whenever you sell something, you need to pay 20% of the sale price to the government.

You talk it over. Everyone feels they still deserve to make the same $1 profit as before. Since you now pay $1.20 for a raw coconut and $1.20 for paint, you need to mark up to $3.40 before tax, and $4.08 after.

After everyone marks up their prices in this way, here are the results:

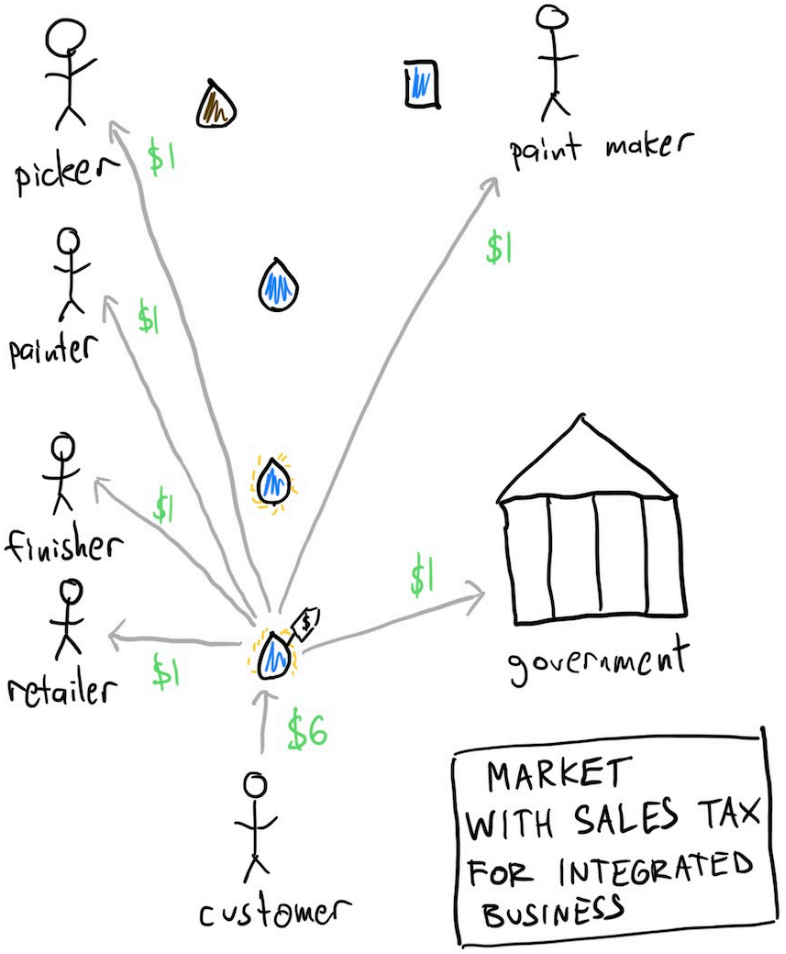


Item	Cost of inputs	Profit	Price	Price after tax
Raw coconut	$0	$1	$1	$1.2
Paint	$0	$1	$1	$1.2
Painted coconut	$2.4	$1	$3.40	$4.08
Finished coconut	$4.08	$1	$5.08	$6.10
Retail coconut	$6.10	$1	$7.10	$8.52

Your customers aren't thrilled about the increase in price, but what are they going to do — live *without* painted coconuts? So they pay the higher price, the government gets its tax, and life continues.

Your Annoying Cousin

A few months later, your unscrupulous cousin hears about your business. He's the jealous type and decides to try stealing your customers. He opens a store and finds four friends to help make coconuts. Unlike you, however, he hires everyone as *employees*. He sells the coconuts for $6 ($5 plus tax) and gives everyone $1 per coconut in wages.



Your cousin and friends don't appreciate the subtle art of coconut decoration.

Everyone agrees yours are better but they start to complain: Why are you charging $8.52 when a slightly worse product is available for only $6? Slowly, your loyal customers drift away and you go out of business.

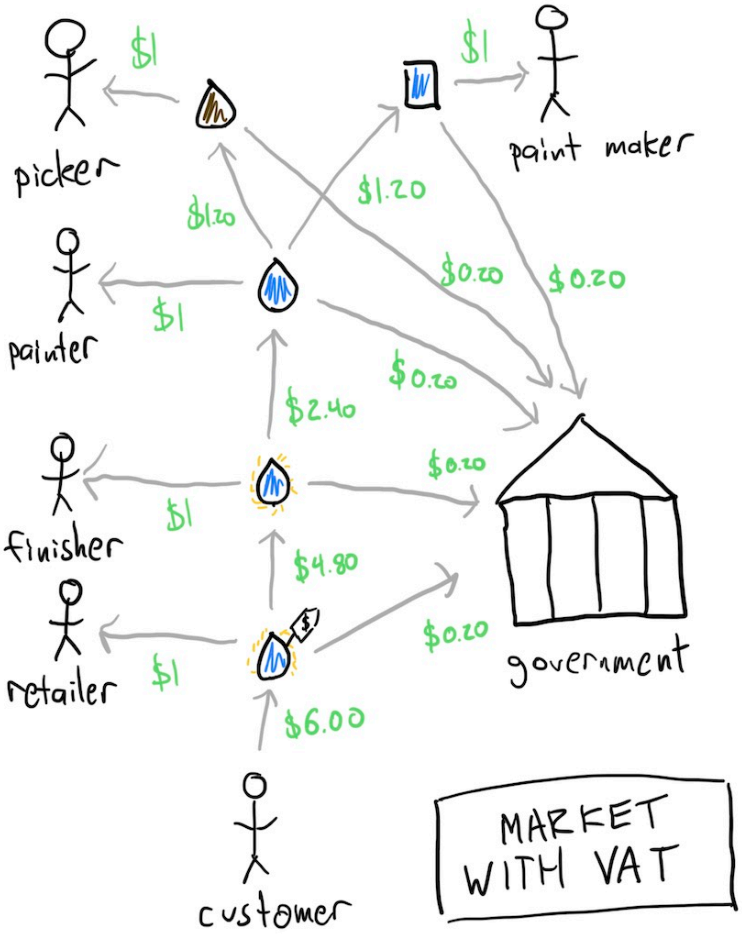
How could this happen? Your team was asking for the same profit while doing a better job! Yet everyone is left with your cousin's knock-off coconuts.

The VAT Regime

Suppose the government had instead announced a 20% VAT. With a VAT, whenever you sell something, you only pay tax on the sale price *minus the price of the stuff you bought to make it.*

As before, you'll need to pay $1.20 for raw coconuts and $1.20 paint. You charge $3.40 for painted coconuts, now you're only taxed on the profit of $3.40-$2.40=$1.00. The price with tax is now $3.60.

Here are the final prices as they go through through the system. Everyone is making a profit of $1, so everyone pays a tax of $0.20.



Item	Cost of inputs	Profit	Price	Price after tax
Raw coconut	$0	$1	$1	$1.2
Paint	$0	$1	$1	$1.2
Painted coconut	$2.4	$1	$3.40	$3.60
Finished coconut	$3.60	$1	$4.60	$4.80
Retail coconut	$4.80	$1	$5.80	$6.00

The final price is $6.00. Since your coconuts are better, your cousin won't be able to drive you out of business with his low-grade stuff.

The Problem with Sales Tax

What happened? Your cousin created a *vertically integrated* business. A sales tax is collected every time someone buys something. If you just do it yourself, no tax is collected.

Are vertically integrated businesses bad? Not necessarily.

However, take your chain of independent independent artisans making and selling coconut products. Imagine someone invents a paint that customers prefer. You almost *have* to switch, or some other painter will drive you out of business. Contrast this with cousin's integrated business making all coconuts. In theory, the inventor could convince your cousin to hire him or license the paint process. If he won't be convinced, the only way for that paint to get to customers is if the inventor develops an entire independent coconut manufacturing chain. Vertical integration means there are price signals at fewer points during production, which tends to make it harder for innovations to thrive.

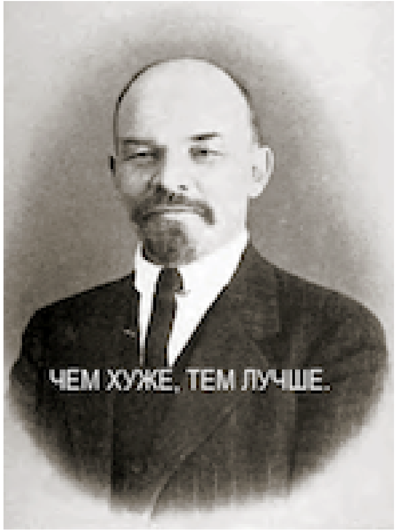
There are times where vertical integration is better. (If everyone is independent, a lot of time is spent on negotiations!) That's perfectly fine. What we *don't* want is to artificially encourage vertical integration even when it's less efficient, which sales tax does.

The VAT Advantage

Another advantage of the VAT is it tends to be easier to enforce. When I sell something, I need to provide certificates proving I paid VAT on my inputs. This gives everyone an incentive to ensure compliance in the previous layer of the chain. With a sales tax, the government needs to watch every single transaction.

Of course, people know sales tax is distortionary. Many exceptions exist to minimize the worst distortions. For example, a retailer usually won't pay sales tax on a manufactured good they intend to a consumer in the same form. Without this exception, we'd probably have a crazy economy where manufacterers sell directly to consumers. The messy patchwork of exceptions reduces the problems with sales tax but doesn't eliminate them.

I think there are two major reasons to oppose replacing sales tax with VAT. The first is a Leninist "worse is better" attitude. If you think *all taxes are bad* then you'd want to keep them painful and visible so people will be maximally annoyed by them. The second is that VAT is complicated to administer, particularly when sales tax can be different in each local area. This might be true, but I find it a bit hard to believe. VAT is more self-enforcing and sales taxes are *already* a nightmare, particularly for anyone selling to different cities/states. If we're keeping the sales tax to keep things simple, where's the payoff?



Notes

- The initial map is based on Wikipedia, but found that many places (Thailand, Saudi Arabia, Iran, Oman, UAE, Kuwait, Angola, Liberia) have recently implemented VAT. I checked that most of the others (Jordan, Greenland, French Guinana, Cuba, Libya, Hong Hong) still do not have a VAT.
- To be sure, if you could implement a sales tax that only applied to final consumers, that would be economically equivalent to VAT. Is that how state taxes work in the US? It's hard to make simple generalizations because (1) it's sometimes hard to say what a "final consumer" is (2) there are different laws in each state and (3) the relevant tax is sometimes called a "gross receipts" tax. The important question is: **Does the US have taxes on intermediate products** that "cascade" like described in the above model? The answer to that question is YES.

Experiments on a $50 DIY air purifier you can make in 30s

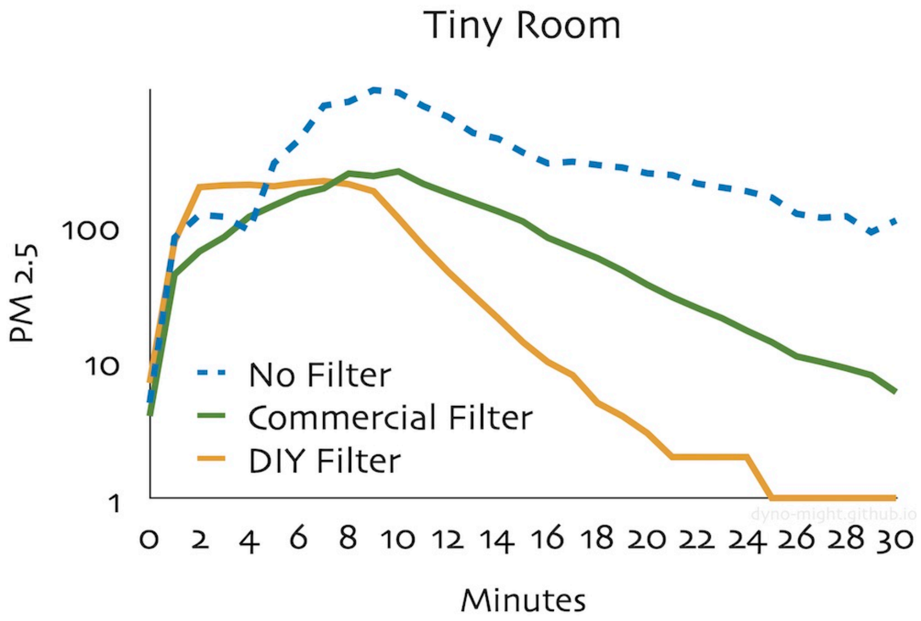
Bad air is bad for you. The air purifier market, though, is a mess. Every purifier uses incompatible proprietary filters, presumably to lock you into buying replacements. How do we know these actually work? Few seem to publish lab tests. And why does it cost $100-$300 for a big plastic box with a fan and a filter inside?

It's common to build DIY air purifiers by basically strapping a filter to a fan. I like the idea of these, but again, it's hard to be confident they really work. There's a few experiments out there, but not enough to make me comfortable.

So I decided to do some experimenting of my own. I made a purifier, generated smoke, and measured how well it removes tiny particles.

TL;DR: YES IT WORKS

If you're in a hurry, this post says that if you strap two HEPA filters to a box fan, it will clear the air of basically all the particles we can measure, and it will do it faster than a commercial filter that costs twice as much.



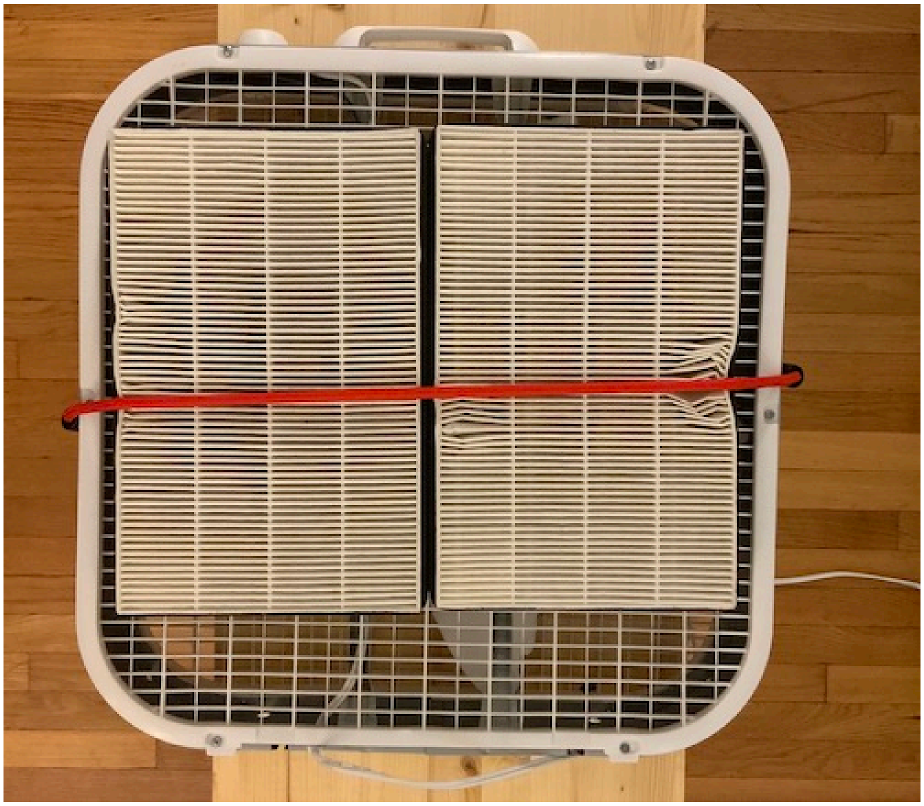
The experiment

DIY purifier

My DIY purifier was *very* simple. (I don't want to promote any particular brands here. Contact me if you want the exact products.)

- A standard box fan. (Cost: $19)
- Two HEPA air filters, each approximately 32 cm x 22 cm and 5cm thick. (Cost: $35 for both)
- A bungie cord. (Cost: Free)

Assembly takes about 30s. You put the filters on the intake side of the fan and strap them on with the bungie cord. Here's a picture:



Timeless elegance and grace, it is not. I get the shakes just looking at that bit of crinkled filter.

Commercial purifier

As a comparison, I got a $100 air purifier from a well-known brand that's intended for small rooms. It uses uses a single HEPA filter that's about 25cm x 12cm and 4cm thick. Replacement filters currently cost around $25.

Smoke

It's surprisingly hard to repeatedly generate a consistent amount of smoke. I tried burning various things (paper, cardboard) and found that the number of particles generated can vary by an order of magnitude, depending on the burn pattern. This is difficult to control and effectively random.

Ideally, I'd have liked to burn some food product like oil, since the kitchen is usually the biggest source of indoor air pollution. I couldn't figure out a good way of doing this, either: You'd need to have the same amount of oil distributed in the same way and heated to the same temperature.

I finally settled on using incense. I cut sticks to the length of a standard credit card and then attached the ends horizontal to the ground. This seemed to be pretty consistent. In retrospect, I bet that burning toast in a toaster would work well. (I didn't have one on hand.)

Measurements

I borrowed a cheap-ish ($100) air quality monitor from a friend. I think it's made by some company in China and then re-sold by various white-label brands. I can't figure out who the original manufacturer is. Based on data I've seen for the reliability of other air quality monitors, I wouldn't trust the absolute numbers, but the I think the *relative* measurements should still be OK.

The typical measurement for particulate pollution is "PM 2.5" which is in units of g/m^3. This is intended to measure what you'd get if you did the following:

- Take a cubic meter of air.
- Filter all the solid particles out of the air.
- Keep only the solid particles that are are 2.5 micrometers (m) or smaller.
- Weigh all the particles you kept in micrograms (g).

Here are some ways to interpret these numbers:

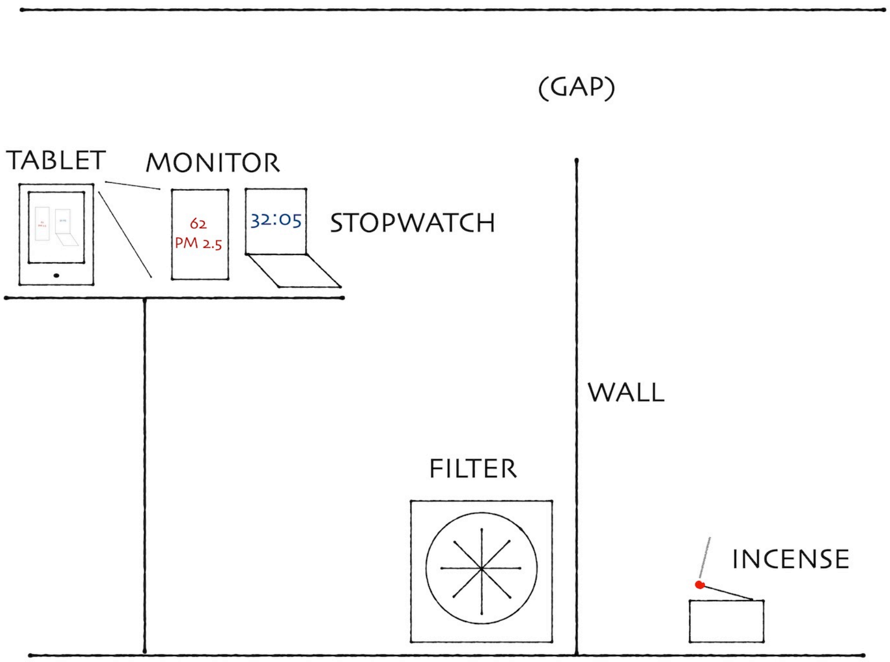
- The EPA says yearly averages should be below 12 and daily averages below 35.
- The average outdoor level ranges from 6 in Finland to almost 100 in Nepal. Rich countries are typically under 15. The highest levels are typically found in Asia and Africa.
- Cooking can easily cause PM 2.5 measurements to spike into the hundreds. I've observed myself that this can happen with only a small amount of visible smoke.

Logging

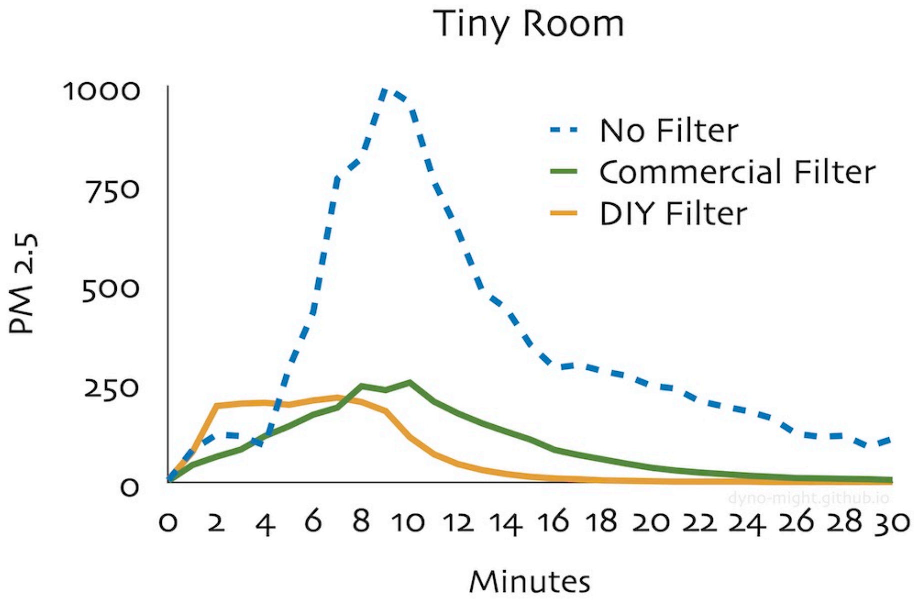
Since the air quality monitor doesn't log data, I used an ultra-hacky alternative: I set the monitor next to a laptop running a stopwatch. I then aimed a tablet at both of those screens and took a timelapse video. Finally, I manually transcribed the data by going to each minute marker in the data. (This was even more tedious than it sounds.)

Results in a Tiny Room

I ran a first experiment in a tiny room of around 8 . Due to worries that wind from the purifiers might change the speed the incense burned, I placed it on the opposite side of a wall, with a gap of around 20 cm near the ceiling.

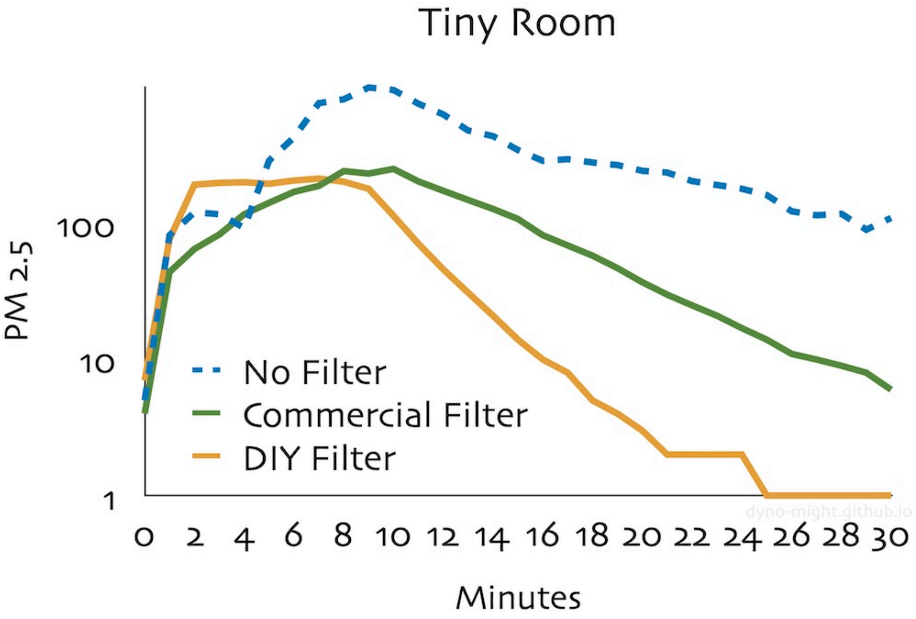


I repeated the experiment once with no filter, once with a commercial filter, and once with the DIY filter. Here are the results:



Things are a bit random around the beginning, probably due to the drifting of the smoke before it's equalized in the room. With no filter at all, this spikes all the way to 1000 g/m^3, the maximum the instrument can show.

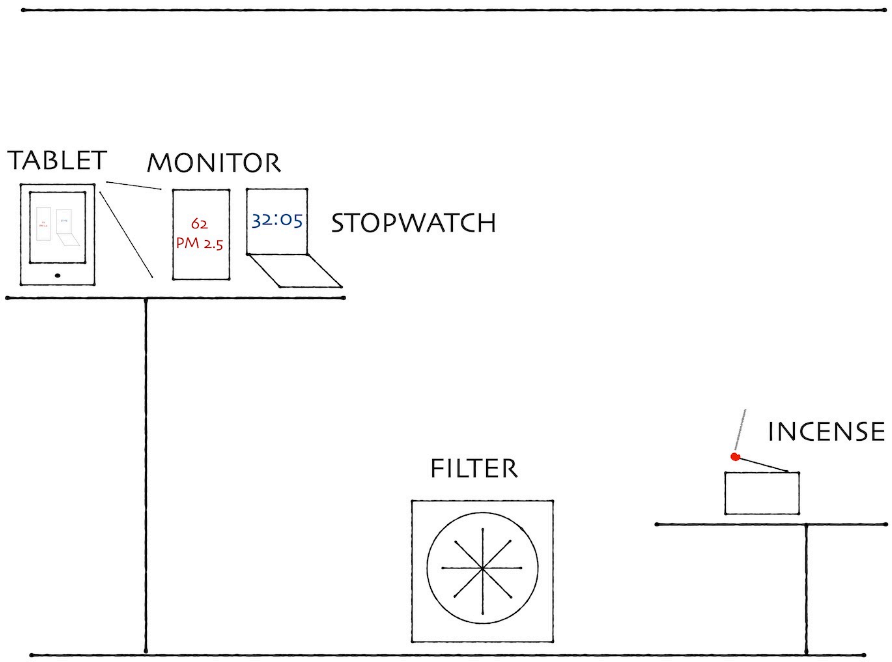
If we make the y-axis logarithmic, it becomes quite clear that the DIY filter is cleaning the air at a better rate. (This is the picture from the top of this page.)



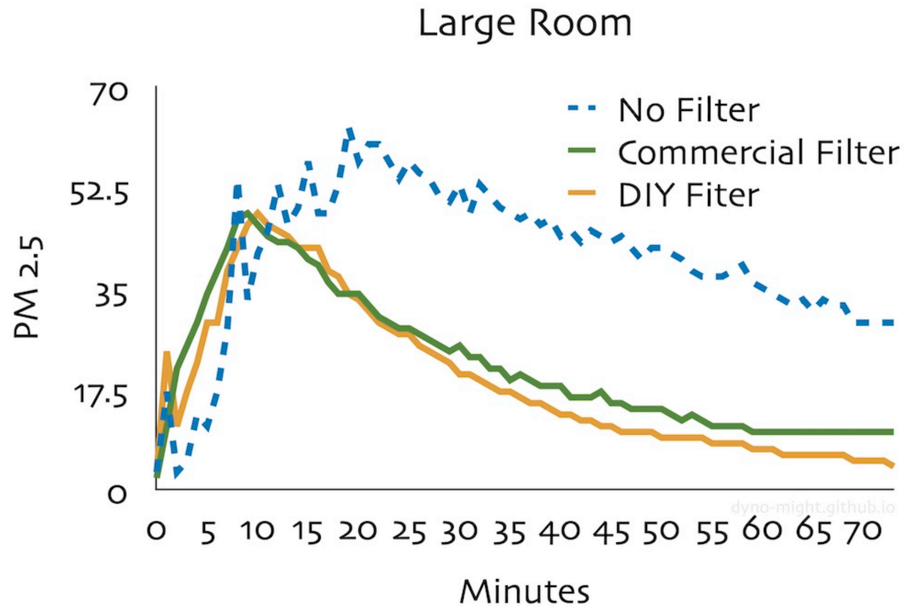
If we take the EPA's threshold of 12 g/m^3, the DIY filter gets there in around 15 minutes, while the commercial filter take around 25 minutes.

Results in a Large Room

Thankfully, I don't spend most of my time in an 8 room. Thus, I repeated the experiment in a large room of around 100 . Here there was no wall between incense and purifier. Instead I left around a meter of distance between the incense and purifier and the purifier and the monitor.

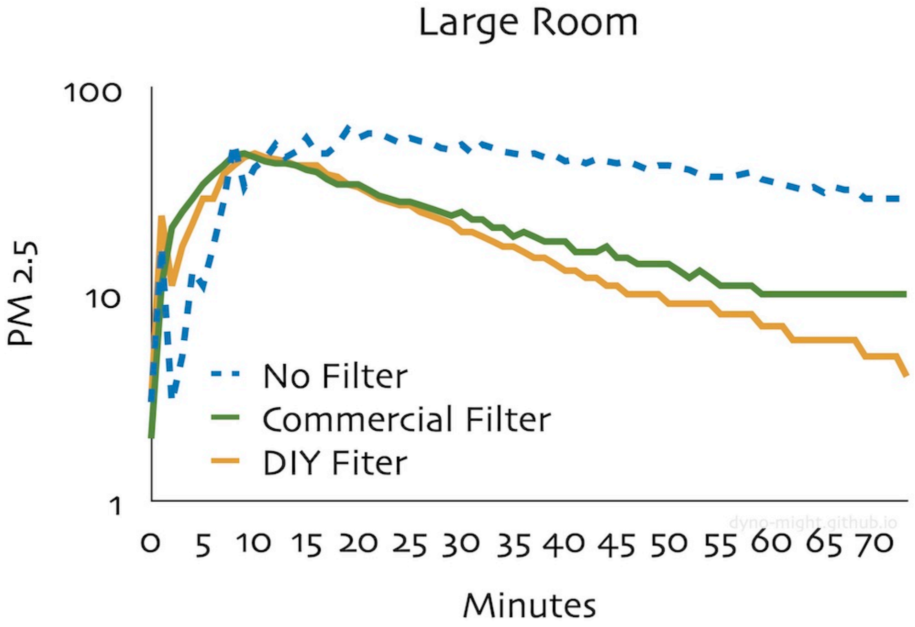


Here are the results:



There's even more randomness around the beginning, probably just due to how the smoke drifts around. Based on the room volume we'd expect a peak concentration with no filter of around 80 g/m^3 = 1000 g/m^3 * (8/100). Reassuringly, this is pretty close to what we see.

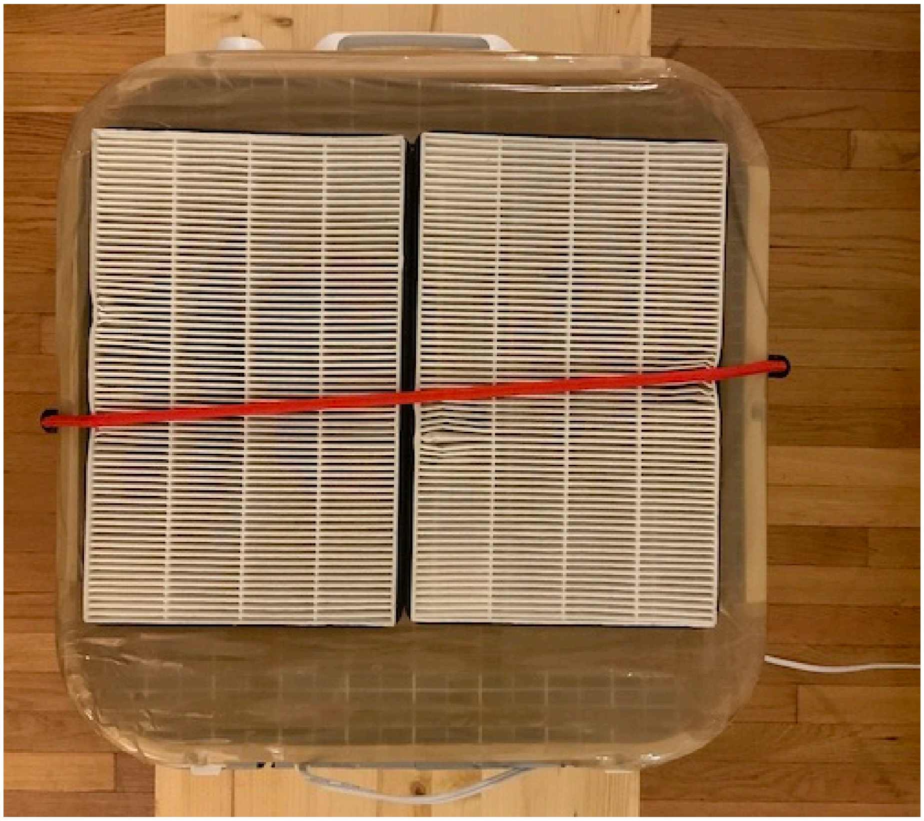
The DIY purifier looks a bit better. If we plot in log space, it's more clear that it is indeed filtering at a better rate:



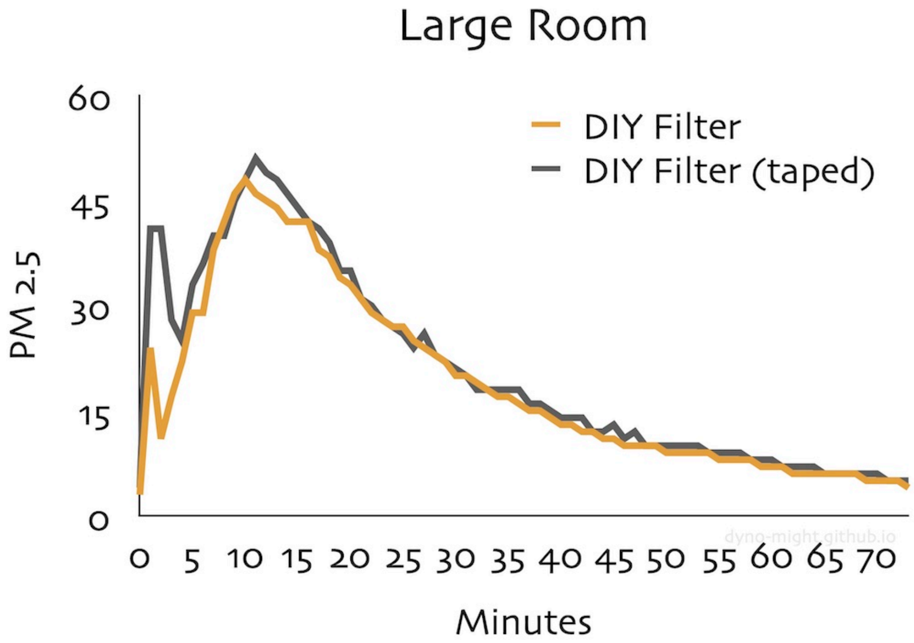
Taping

It's common advice for DIY purifiers like this to seal around the edges of the filter so that all air must pass through it. I share the intuition that this would help, but it's hard to be sure: If you block airflow, you slow down the fan. This could be counterproductive.

In this case at least, experiment is easier than theory. I took packing tape and carefully sealed around the intake side.



And the results are...



…nothing!?

This was unexpected. I thought the tape would help, but I wouldn't have been surprised if it hurt instead. Instead, there's basically no difference at all. I don't know enough about fluid dynamics to even speculate about what's happening here, so I won't try.

There could be some weird quirk in how I ran this experiment. This doesn't necessarily mean that all the advice to tape around the filter is *wrong*. However, I've never seen any experiments that show taping helps either.

Thoughts

Cost. The DIY purifier isn't dramatically cheaper than the commercial one, but I expect the filters would need to be replaced much less often. The commercial purifier uses a single filter with an area of 300 cm^2, whereas the DIY purifier uses two filters with a total area of around 1400 cm^2, and also slightly thicker. It's reasonable to assume the DIY filters could remove ~4 times as many particles before replacement.

Durability. One concern is that box fans aren't meant to be used with filters attached and could wear out. This is reasonable. However, box fans are much cheaper than commercial purifiers, and I've been using this particular fan with various filters attached for several years now without issue. The bigger concern here is probably that heat build-up could in principle cause a fire, which could be dangerous. Also, there are some suggestions that heat buildup on the electrical components could create toxic particles, which might be dangerous for birds. I don't know how serious this danger is.

Electricity. The cost of electricity is another factor. Typical box fans seem to use around 55W, whereas commercial purifiers typically use 30-45W. If electricity costs $0.13 / kWh, the box fan would cost around $62 to operate 24 hours a day for a year, while a 30-watt purifier would cost around $34. Obviously, these numbers decrease if you run the purifier less. Some (more expensive) commercial purifiers have air quality sensors built in and automatically turn on only when needed.

MERV or HEPA? Most people who build box-fan purifiers use MERV-rated filters intended for furnaces. Commercial air purifiers use HEPA-rated filters. Roughly speaking, HEPA filters are "better" in that they are rated to remove a higher percentage of particles in one pass. It's not clear that HEPA filter will actual perform better when attached to a fan, though: A filter that catches fewer particles in one pass might still be better if it allows for faster airflow.

That one video. If you're reading this article after it was linked from some forum, I'd bet you that someone in the comments links to this video from the Michigan Sinus Center. I found this inspirational, but note a couple of things: First, while the description says they use HEPA filter, the video clearly indicates a MERV filter. Again, that's not necessarily bad! They claim that around 90% of particles sized 0.3 microns are larger are eliminated in a single pass. That's good, but not totally reassuring. The question is, does it remove 99% in two passes? If 90% of the particles in the ambient air were large and the filter only catches large particles, then additional passes would never get rid of the most dangerous small particles. This is why I trust HEPA filters a bit more: since they remove almost all particles in one pass, I'm confident they should remove almost all particles eventually. This is also why I strongly prefer experiments that actually measure particles removed from the air in a room, rather than just the air coming out of the purifier.

Further questions. There's a lot of things that further experiments could look at:

1. Does the fan speed make a huge difference?
2. How does the purifier compare to larger commercial purifiers?
3. How do MERV-rated furnace-type filters compare under the same conditions?
4. How can it not matter if there's tape around the filters!?
5. Does fan speed matter? (I always ran the box fan at maximum speed.)
6. Is it better to put the filters on the intake or outtake side of the fan? Intuition suggests the intake side, but as we've seen, surprises happen.

On purifiers and health. If the outside air is clean where you live, your first priority probably shouldn't be to get an air purifier. Instead, I'd recommend you *be careful when cooking.* Having an air quality monitor around really drives home how even small amounts of cooking smoke can cause gigantic spikes in particulate levels. For most people, *not creating* these particles in the first place is the best strategy. Try not to create any smoke at all. If you have a range hood that vents outside, use it! If you do create smoke, open windows immediately. One difficulty is that it's actually quite hard to guess how many particulates you've created. There's a reasonable case that you should spend your money on an air-quality monitor before a purifier.

Incense. It's probably bad for you. I'd avoid it.

How to run without all the pesky agonizing pain

I used to think the people I saw running were insane. They were confused about life. Whatever the benefits of running, nothing could justify that much suffering. Runners were cut from a different cloth. They had a strength of will I lacked. I would never be one of them.

Now I run regularly. Not because I developed stronger willpower, or because I feel an obligation to my health. I run because I like it. Somehow, I became one of them.

Here's the secret to running: The pain you feel when you run? You don't need to endure that.

Untrained runners typically have this experience: You resolve to start running. The first session, you take off at a fast pace. After a minute or two, your heart and lungs are struggling to keep up, and soon your entire body is in pain. This is *terrible.* You don't have the willpower to run through that suffering day after day, so you quit after a few sessions.

For some reason, people don't tell you an important fact: That horrible feeling almost completely disappears within a few weeks of training. Your cardiovascular system develops quickly. Instead, you run until your legs get tired — an infinitely more pleasurable experience. The secret of all those "crazy people" on the street is that they aren't suffering (or at least, not much).

But even this is misleading. It suggests you need to "break through" to the second stage, and only then running becomes easy. It suggests you must summon ultimate willpower for a few weeks in order to level up.

No! You don't need ultimate willpower, even at the beginning. That's the second thing people don't tell you: You don't need to suffer to get through to the second stage. You just need *consistency.*

You should think as follows: You are starting a habit you will keep for decades. It *doesn't matter* how hard you run today. What matters is (1) that you *do* run, and (2) that you enjoy it enough that you'll run again tomorrow. That's it. Are there are people who run three times a week for 5 years and still suck at running, however little their willpower is in each session? No! Just go, take it easy, and let time do the work.

Behold, the official Dynomight™ running program:

- Schedule three 20 minute sessions per week.
- In each session, jog as long as you feel like it. When this gets hard, *stop jogging and walk.*
- Over time, try to spend a higher fraction of your time jogging. Don't measure the fraction. Don't stress. Just keep going.

At the beginning, you might only get in a handful of 30 seconds jogs during your 20 minutes. That's fine.

When I suggest this, people often say "that's too easy" or "I need to get in shape faster". They might start some aggressive training plan and follow it for a week or two. They hate it, of course, and soon they've quit. A year later, they are exactly where they started.

The famous couch to 5k program is of a similar spirit. It starts with walking and sslloowwlllyyy mixes in running over time. It takes two months to go from nothing to running 25 minutes nonstop. But you're in this for the long haul, right? Does it matter if it takes you four months instead? Or six months? A year?

Lots — perhaps most — exercise advice is counterproductive for regular people. It's oriented towards elite athlete. If that's you, fine. But the other 99% of the population with modest ambitions can have essentially all of the health benefits with almost none of the suffering.

Let's consider two possible failure modes of your running plan.

- **Failure mode A**. Your plan is too aggressive. You dread going each time and eventually quit.
- **Failure mode B**. Your plan is too easy. You have a good time but improve your health slightly more slowly than possible.

Now, some questions.

1. Which failure mode is more likely?
2. Which failure mode is more harmful if it happens?
3. Then why, oh why, are you optimizing your exercise plan around the other failure mode?

In reality, *suffering is bad for you.* If you suffer, you'll hate running. If you hate running, you'll stop. So not suffering is your top priority.

Philosophically, should I have known running wasn't interminable pain forever? In retrospect, it seems ridiculous to think I was "different" from runners, physically or mentally. Why didn't I consider the possibility that my random guesses about what they were experiencing was wrong?

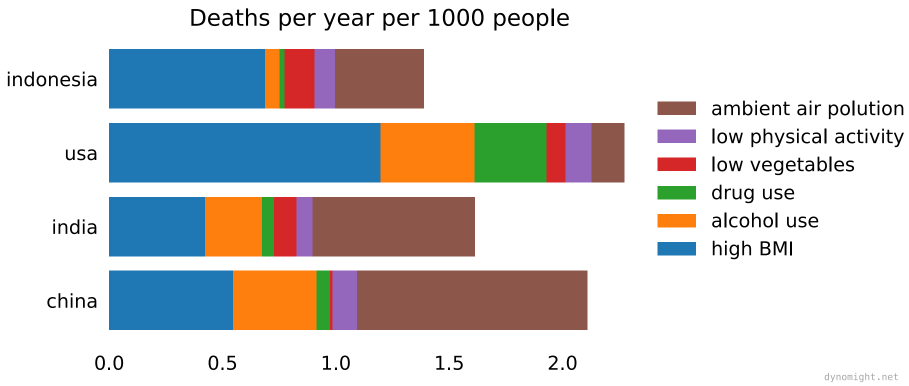
Maybe there's a lesson here about how easy it is to make assumptions, or imagine deep personality differences when mundane explanations suffice. Maybe the fundamental attribution error does exist. Or maybe it's just to hard to resist passively bragging by letting everyone think running must be difficult.

Better air quality is the easiest way not to die

What do you worry about more: Getting exercise, eating vegetables, or the air you breathe?

While most things that clearly improve health are well known, one is insanely underrated: Fixing your air. I suspect this is often **the most effective health intervention, period**. Nothing else is so *important* while also being so *easy to address*.

Let's do a sanity check: Take the four biggest countries in the world and compare how many people died from various causes in 2019.



dynomight.net

It's hard to prioritize health advice. I'm told I should limit salt and eat cruciferous vegetables and do cardio and sleep well and limit alcohol and reduce stress and go for regular checkups. But *how much* do each of these matter? If you're a fallible hairless ape, what should you do first?

To answer this, we need numbers. Below, I've estimated how much various things impact life through air quality.

Lifestyle	Life cost
Be a Viking	4 years
Live in Delhi	3 years

Lifestyle	Life cost
Commute by train from Newark to NYC	½ year
Live in an average part of US	¼ year
Breathe second-hand vape smoke	near zero?

Single Event	Life cost
Live near 2020 US west-coast wildfires	2.4 days
Have a really smoky fire at home	2 days
Burn a cone of incense	2.3 hours
Use an ultrasonic humidifier for one night	50 min?
Broil fish with windows closed	45 min
Burn a stick of incense	27 min
Use hairspray	14 min
Smoke one cigarette	11 min
Blow out a candle before sleep	10 min

This is good news—you can buy extra life with minimal cost in money, time, effort, or willpower!

By all means, control your body-mass, eat well, and start running. Those are important, but they're also kind of hard. You might fail to lose weight, but if you try to fix your air, you'll succeed. You should put the stuff with the highest return on effort first, and that's air.

What I recommend

If you don't want to read this long, *looooooong* article (I'm sorry) just do these things in this order:

1. If you have an ultrasonic humidifier, kill it.
2. Monitor local air quality like the weather.
3. No incense.
4. Extinguish candles with a lid.
5. Be careful about smoke when cooking.
6. Get a particle counter.
7. Use an air purifier at home all the time. (Move this to #1 if the outdoor air has high particulate levels where you live.)
8. Install a HEPA cabin air filter in your car.
9. Avoid aerosols.
10. Use a mask *very carefully* when in dirty air.

A strange list, admittedly, but it's where the evidence seems to lead.

Contents

Particles and the trouble they cause

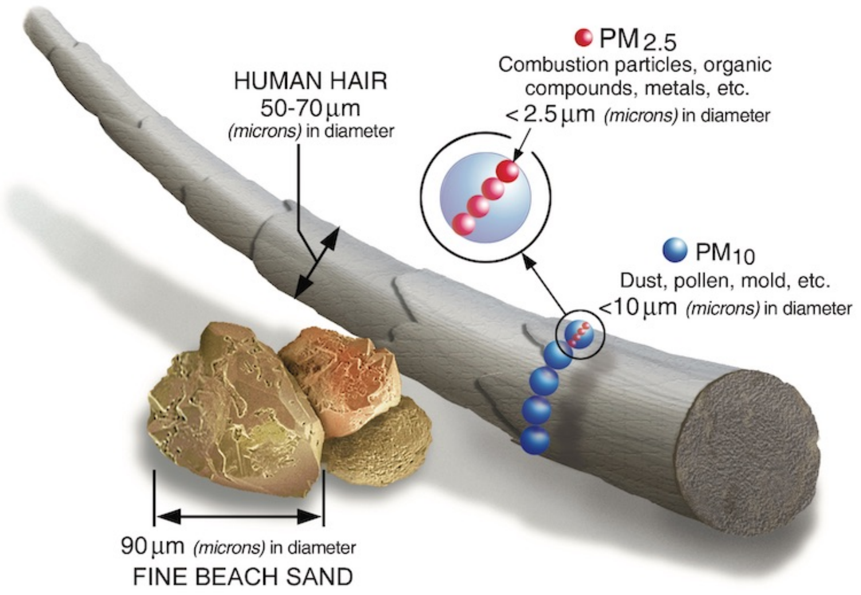
What we're measuring

We measure particles in terms of $PM_{2.5}$. To understand what these numbers mean, here's how you could, in theory, measure them:

- Take a cubic meter of air.
- Filter out all of the solid particles.
- Keep only the particles that are 2.5 microns (m) or smaller.
- Weigh the remaining particles in micrograms (g).

The units are g/m^3, since you're weighing particles (in g) in one m^3 of air.

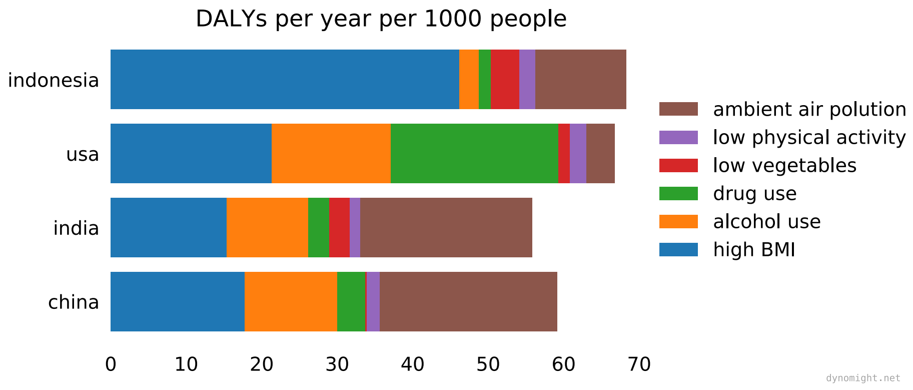
For reference, human hair is around 70 microns wide, bacteria are 1 to 10 and viruses are 0.02 to 0.4. The EPA gives a helpful visualization:


PM2.5
Combustion particles, organic compounds, metals, etc.
< 2.5 µm (microns) in diameter
HUMAN HAIR
50-70 µm
(microns) in diameter
PM10
Dust, pollen, mold, etc.
<10 µm (microns) in diameter
90 µm (microns) in diameter
FINE BEACH SAND

You might ask: Does it matter what the actual particles are? Is 50 g/m^3 created by burning coal or manufacturing cement or natural dust equally harmful? The answer is *no* (heard of asbestos?) but we don't really know how much these differences matter in practice.

Quantifying harms

A better measure than deaths is disability-adjusted life years or DALYs. This is the number of *years* of life lost plus an adjustment for non-lethal conditions that make life worse. For example, schizophrenia is pretty bad, so this measure counts someone who becomes schizophrenic for a year as losing half a DALY. Looking at DALYs lost to different causes gives a similar picture to deaths:

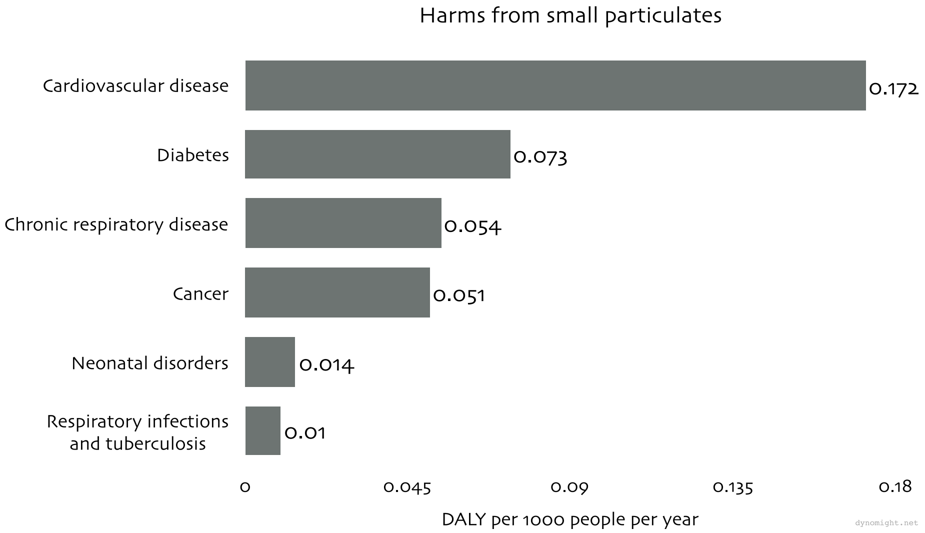


These numbers are only for *ambient* air pollution, e.g., due to cars, power plants, and manufacturing. Indoor air pollution comes on top of this.

How particles hurt you

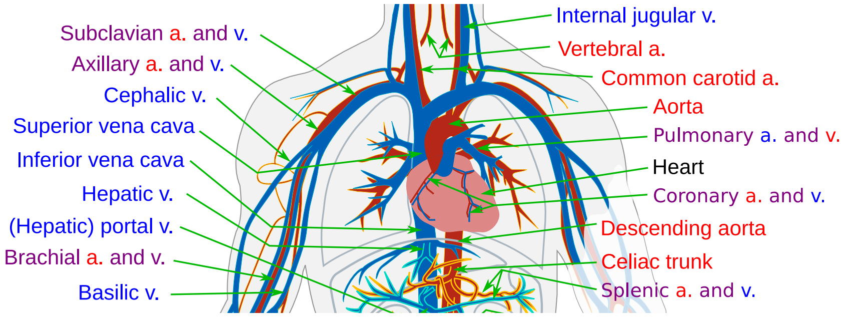
We worry about tiny particles because they seem most harmful, particularly in terms of how chronic exposure leads to long-term health problems.

Tiny particles do cause lung cancer, but that's one of the *smaller* harms. More than half of harms don't come from the lungs at all, but from diabetes and cardiovascular disease. Here are the DALYs lost in the US in 2019:



How particles manage to do all this is still a subject of research. The basic story seems to be that tiny particles easily get through the lungs into the bloodstream. These foreign particles then activate your immune system, which sort of goes on a rampage, causing tons of problems.

Why so much cardiovascular disease? Well, you can probably guess where the blood goes after it leaves the lungs.



A heuristic to quantify harms

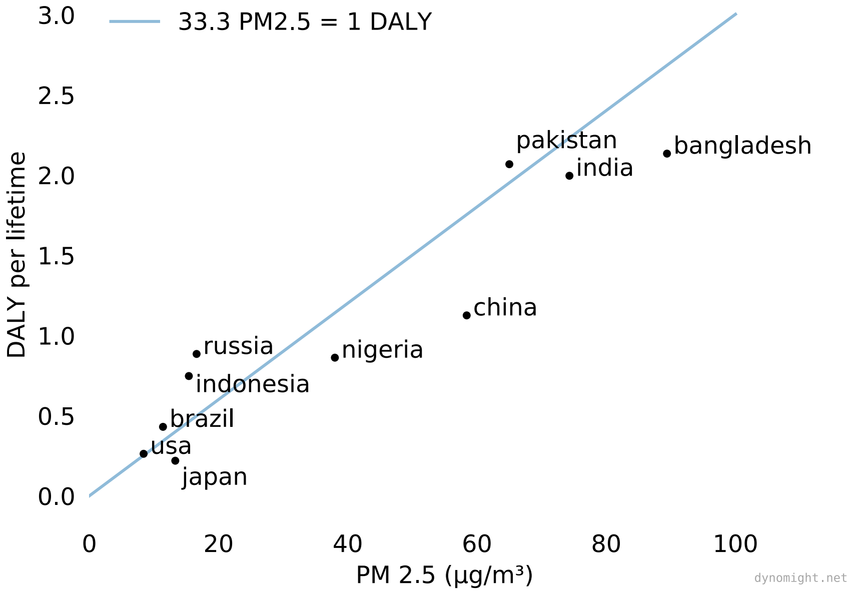
How *much* do particles hurt you? While it's hard to be precise, this section will give two simple heuristics:

- **A life-long exposure of 33.3 PM$_{2.5}$ costs 1 DALY**. This is best for lifestyle changes. For example, moving from somewhere with no particulates to somewhere with a level of 100 costs 3 DALY.
- **At 2500 PM$_{2.5}$, you lose disability-adjusted life in real time.** This is best for one-off events. For example, if you're exposed to a level of 5000 for 3 hours, you lose 6 disability-adjusted life hours.

Of course, the only way to be *sure* would be to do randomized experiments where we lock people inside identical environments for their whole lives, vary the particulates they're exposed to, and see how long they live. Since we can't do that, we're left with observational studies.

A 2013 paper looked at life expectancies and particle levels in 545 US counties, while controlling for confounding variables like wealth, smoking, and demographics. They found that each 28.5 g/m^3 of particulates costs 1 year (not disability-adjusted.)

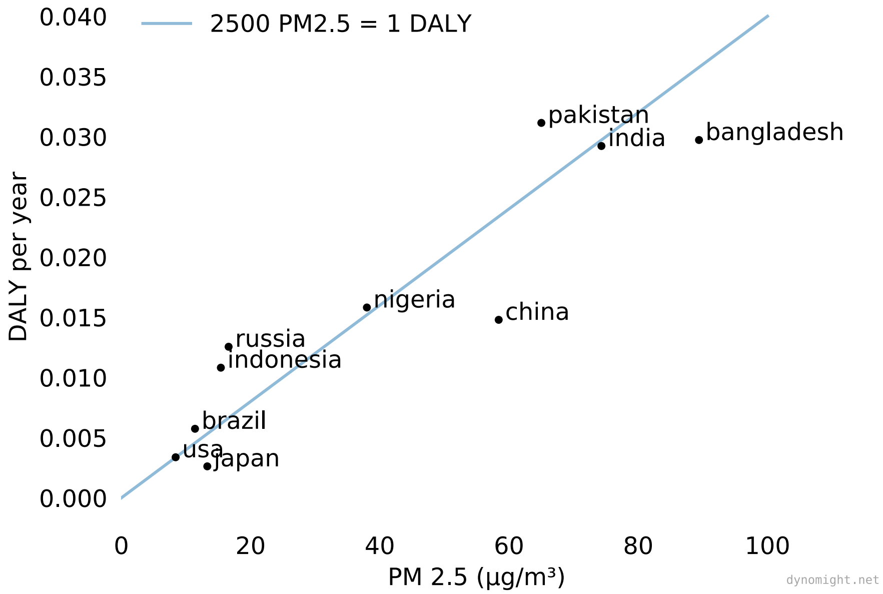
Should we trust this number? Most papers focus on public-health questions, but we can extract estimates. A comprehensive 2017 paper estimates the population mean PM$_{2.5}$ in different countries, as well as the health costs of ambient air particles. I took their estimates and multiplied them by the WHO's life expectancy figures to get an estimate of the total DALYs a person loses over a lifetime. Here we compare these to particle levels.



The straight line shows a fit. It suggests that if you are exposed to 33.3 $PM_{2.5}$ for your whole life, you'll lose around 1 DALY as a result. I wouldn't put too much faith in this exact number. It varies from country to country, and this is all built upon some very tricky observational statistics. Still, the estimate is reassuringly close the the number from the 2013 paper.

You might ask: Shouldn't the same level of particles cause more harm somewhere where people live longer since they have more time to spend breathing?

Maybe. I tried relating the particles to the DALY loss in a single year, without multiplying by life expectancy. It suggests that being exposed to 2500 $PM_{2.5}$ for one year costs 1 DALY.



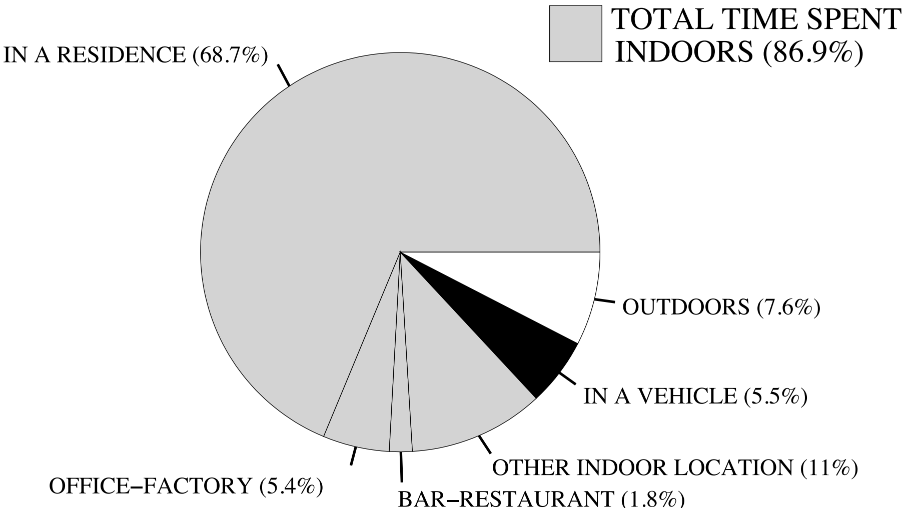
Put another way, if you breathe particles at a concentration of 2500, you double the speed at which you move towards your (disability-adjusted) destiny. If you're exposed to a level of **x** for **h** hours, then you lose **h** × **(x / 2500)** hours.

In practice, this isn't too different from the previous estimate since life expectancies don't vary *that* much between countries.

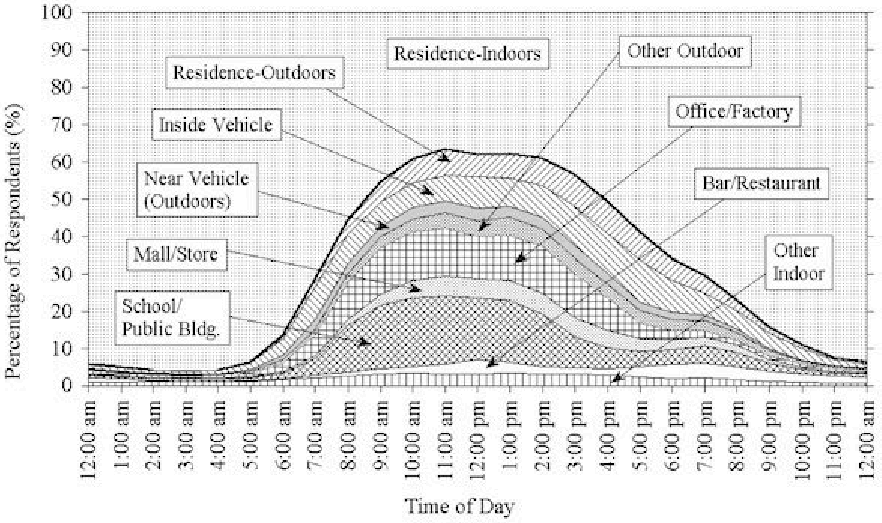
Personal exposure

So, we can estimate how much harm particulates do. Your next question should be, how many particles are you exposed to? *Probably* the answer is ambient levels plus some extra that's created indoors, but it's hard to say how large that extra amount is.

There are great records of outdoor levels, but people aren't outdoors very much. Here's the NHAPS survey on how Americans spent their time from 1992 to 1994.

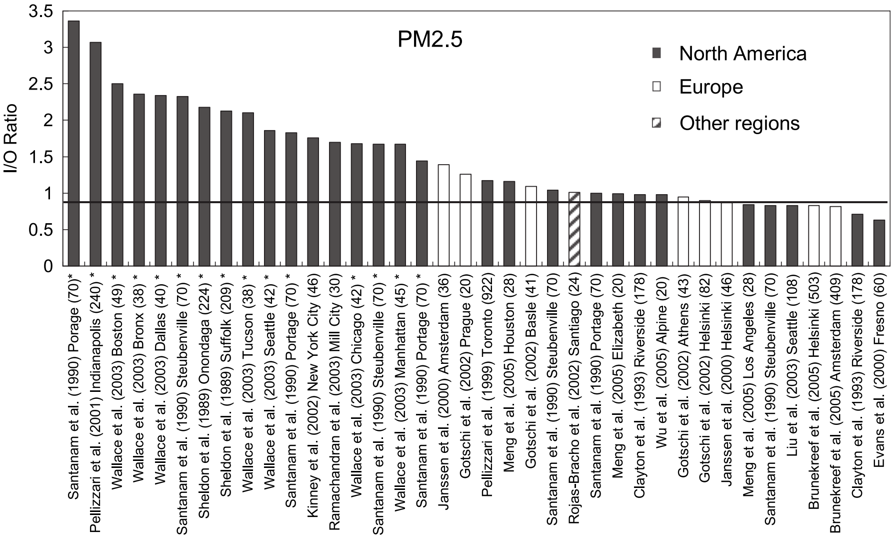


No recent survey seems to equal this. They even have curves of where people are throughout the day.



Personal vs. outdoor exposure

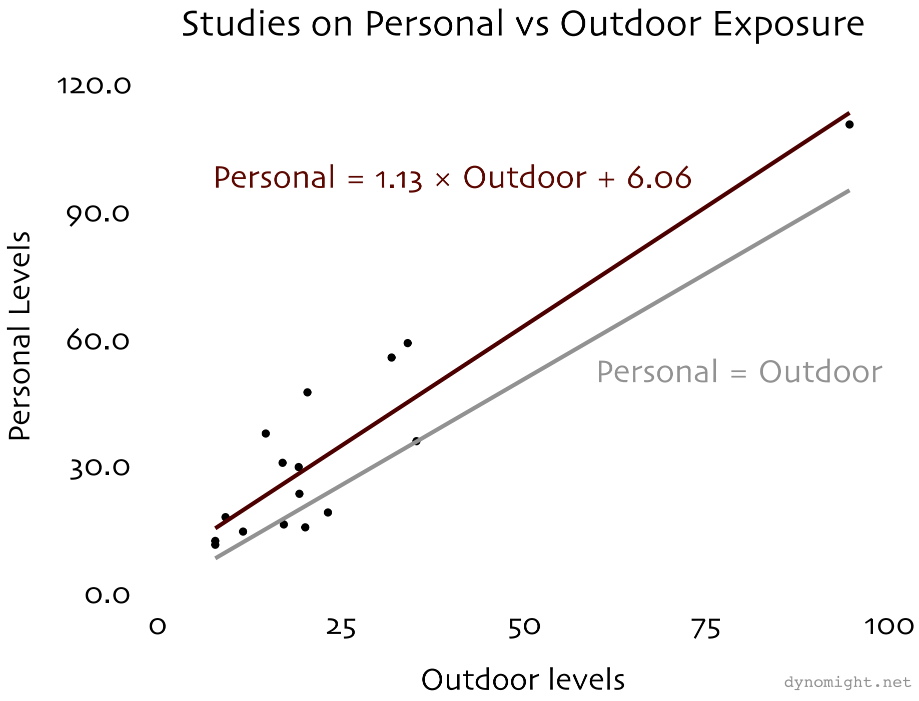
Are levels indoors the same as outdoors? There's a strong correlation—particularly if people keep their windows open—but it varies. Typically, levels indoors are higher than outdoors. Chen and Zhao review a bunch of papers that try to estimate the indoor/outdoor (I/O) ratio:



None of this really matters though. Indoor levels vary in space. What you care about is your *personal* exposure—the air that actually goes into your lungs.

This is hard to study since someone has to carry a measurement device on their person. Still, it's been done a few times. Early attempts put a small tube near the mouth that passed that air through a filter that was later weighed. More recent studies use devices that measure light scatter.

I couldn't find any good reviews, so I did my own. Here's a comparison of the mean personal and outdoor levels in all the studies I found:



You can expand a table with details on all the studies here.

Here's all the studies I found that try to compare mean personal (P) exposure to indoor (I) or outdoor (O) exposure.

Study	Where	I	O	P	
Williams 2003	North Carolina	19.1	19.3	23.0	
Meng 2005	Los Angeles	16.2	19.2	29.3	
	Elizabeth, New Jersey	20.1	20.4	46.9	
	Houston	17.1	14.7	37.2	
Koutrakis 2005	Baltimore		20.1	15.1	seniors, winter
	Baltimore		23.2	18.6	children, summer
	Boston		11.6	14.1	seniors, winter
	Boston		17.0	30.3	children, summer
Sørensen 2005	Copenhagen	13.4	9.2	17.5	< 8C (median exposure)
	Copenhagen	9.5	7.8	11.9	> 8C (median exposure)
Johannesson 2007	Gothenburg, Sweden	9.7	7.8	11.0	
Suh 2010	Atlanta		17.17	15.78	
Lei 2016	Shanghai		94.5	110	
Chen 2018	Hong Kong		35.3	35.4	88% had windows open
Sanchez 2019	Villages near Hyderabad		34.1	58.5	Women (half cooked with biofuels)
	Villages near Hyderabad		31.9	55.1	Men

Personal exposure is strongly correlated with outdoor levels, but typically a bit higher. If you trust the trendline, it predicts that someone with an outdoor level of 50 would have a personal exposure of 62.5.

However, personal exposure is much less predictable than the graph above would suggest. Each point is averaged over many people. Individual studies that broke things down person by person show a huge range.

So, your exposure is probably ambient levels plus some amount that depends on what you do and where you go. To figure that out, we'll have to dig deeper.

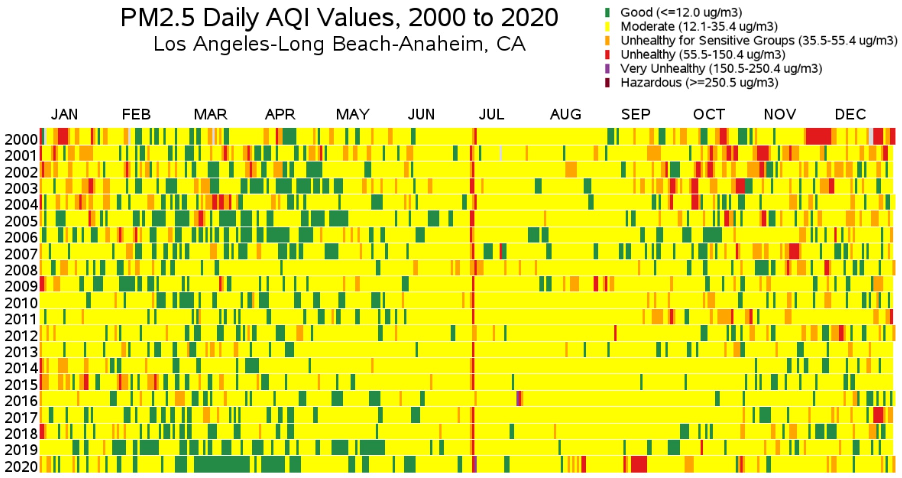
Particles Outdoors

Most particles are introduced by human activity. Common sources are power plants (particularly coal, but also natural gas and oil), factories, human-made fires, cars, and trucks. Natural sources are dust, wildfires, and (oddly) sea spray.

Variance with respect to city/region- Large: Levels vary hugely throughout the world. Estimates vary, but some countries are really low (New Zealand, Canada) and some are an order of magnitude higher (India, Qatar). Different locations inside of countries are correlated partly because of the same air blowing around and partly because of shared emissions controls. Still, if your city has no wind and lots of factories and cars, levels will be higher.

Variance with respect to location within a city/region- Smallish: How much variability is there between different places in the same city? The answer seems to be *some*. A massive study in various locations in Europe addressed found that particle levels near streets were around 20% higher than urban measurements which were, in turn, around 20% higher than regional measurements.

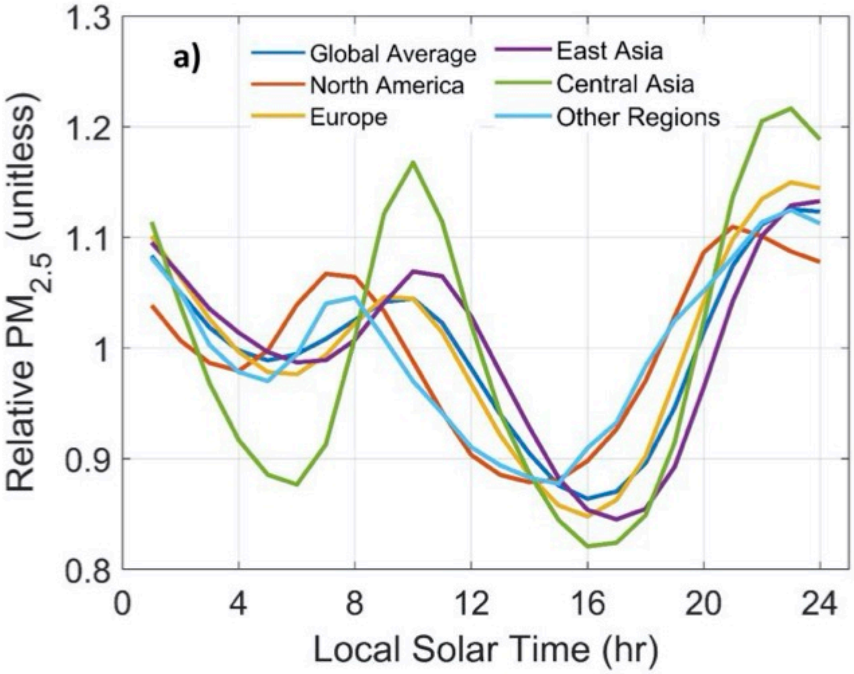
Variance with respect to date- Moderate: At the same location, levels vary throughout the year. Here's the mean concentration in Los Angeles for every day for 20 years, courtesy of the EPA.



You can see the impact of people blowing up fireworks to celebrate their freedom (July 4 every year), everyone staying at home because of a pandemic (March-April 2020), and massive nearby wildfires (September 2020).

Incidentally, how much did those wildfires matter? Many areas saw their levels rise to around 100 for a few weeks and some spiked as high as 200-500. As a pessimistic estimate, suppose your levels rose by 200 for a full month. That raises your yearly average by 16.67, which would cost ½ a DALY if it happened every year. If it just happened once, you'd lose 200/2500=.08 months or 2.4 adjusted life days.

Variance with respect to time of day- Smallish: At the same location and date, levels change by the hour. Manning (2018) combined measurements from 3110 sites around the world to get the following graph.



Here, we have a great riddle: Why are levels lowest in the mid to late afternoon? They suggest that "this remarkable global homogeneity in diurnal $PM_{2.5}$ cycles suggests the influence of common factors including the diurnal cycle of mixed layer depth modulated by other processes such as diurnally varying emission patterns."

What I *think* this means is this: The sun heats up the earth in the morning. This causes the air near the earth to rise, mixing up the different layers of air. This pulls lots of particles from close to the ground up into the sky, decreasing the density. After the earth cools, the air stops mixing around so much. Also, maybe it has something to do with when people commute and so on.

I'm fascinated by this phenomenon of mixed layer depth but haven't been able to figure out much about it. Does it change throughout the seasons? Is that why wintertime air quality is often worse? I don't know.

Particles While Commuting

Driving: Cars generate lots of particles. If you're driving near lots of other cars particle levels are probably higher than the overall ambient air. A couple of studies I saw found levels in the range of 50-100 in cars. Others suggest it's a not so bad and similar to just being outside. It probably depends on your car, traffic, local emissions controls, and weather.

Walking, biking, and running: If you're on a street, your exposure is maybe a bit higher than the background, but not a ton. As mentioned above, street measurements are typically a bit higher than urban measurements. If you're biking or running, you're breathing harder. It seems that a typical adult breathes around 15 times per minute but can speed up by a factor of 4 if exercising hard. This might mean that particles accumulate 4 times as fast, but I can't find any clear evidence.

We could estimate the harms from pollution as a result of running or biking, but I won't since exercise also *improves* cardiovascular health, and I don't know how to calculate the tradeoff.

Riding the subway: Luglio (2020) measured particles for all the major train systems in the northeastern United States. Levels in above aboveground stations were always low (10-25).

City	On Trains	Underground stations
Boston	182	327
New Jersey to New York (PATH trains)	449	779
New York (MTA trains)	343	547
New York to Long Island (LIRR trains)	11.6	91.2
Philadelphia	55.7	112
Washington DC	205	362

For the worst-offending train system, taking take a single trip from Newark, New Jersey to the World Trade Center in New York should cost 7.5 life minutes. Doing that as a commute five days per week would cost 0.56 DALYs.

The trip takes 25 minutes. Suppose you spend a combined 10 minutes in the two stations. Your exposure for the hour you do that trip is $779 \times 10/60 + 449 \times 25/60 = 316.9$, leading to a cost of $316.9/2500 = .126$ hours.

There are 10 total trips, each raising your exposure by 316.9 for one hour. This raises your average weekly exposure by 316.9×10/(7×24)=18.86 with a cost 18.86/33.33 = 0.56 DALY.

Smith (2020) found that in the London underground, the Victoria line had a median level of 361, the Northern line 194, a couple other around 50, and the rest even lower. The Victoria line is highest because it is entirely underground, meaning that particles have nowhere to go.

Martins (2015) reviews many previous studies and found levels over 100 in London, Buenos Aires, Paris, Beijing, New York, Stockholm, Shanghai, Barcelona, and Seoul. There were lower levels in Budapest, Guangzhou, Helsinki, Los Angeles, Mexico City, Taipei, and Sydney.

Rant: Let's face it, these levels are a scandal. Some places are working on it, but progress is slow because it's hard to retrofit trains. Hear me, transit agencies: *Don't retrofit the damn trains.* Just install normal air purifiers in *stations*. Do this because:

- The problem is that particles build up slowly in tunnels with no place to escape. We can solve the problem at the source by slowly removing particles.
- It's *easy* to put static purifiers in stations. Space and power aren't as much an issue. You can use standard components.
- The particles in trains are coming from the stations and tunnels.
- People also breathe the air in stations.

Are air purifiers too expensive? Well, the New York subway (MTA) has 275 underground stations. Assume pessimistically that each station needs 50 purifiers, and it costs $1000 each per year to operate them. (The MTA is fond of vastly overpaying.) The cost would be 13.75 million per year, less than 0.1% of the MTA's budget. Seems worth it to avoid exposing millions of people to air five times more hazardous than the most polluted cities in the world.

Particles Indoors

Suppose you're inside your home. The quality of the air you breathe can be reduced to five factors:

1. Particle levels outdoors.
2. How long particles hang in the air indoors.
3. The exchange rate between indoor and outdoor air.
4. Stuff you do that creates particles indoors (cooking, candles).
5. Stuff you do to remove particles indoors (running a purifier).

We've covered #1 already. Let's do the rest.

How long particles hang in the air

Maybe indoor air is somehow automatically cleaner than outdoor air? Public health guidance to stay indoors during poor air quality suggests this.

Left alone, particles do settle out of the air. With totally still air, this happens deterministically with larger particles falling faster. With real (turbulent) air, particles bounce around until they stick to a surface. This leads to an exponential decay with a rate depending on the particle size and air turbulence. Experiments suggest something like the following:

Particle size	half-life with stirred air
10 m (largest PM$_{10}$)	2 minutes
2.4 m (largest PM$_{2.5}$)	3 hours
1 m	1 day
.1 m	1 month
.01 m	1 year

I have trouble determining how much these times depend on air turbulence or if the stirring rate resembles real-world conditions.

Half-lives of air in homes

Even with all windows closed, air is constantly moving through cracks in the building. Let's quantify this as a "ventilation half-life", the amount of time after which half of indoor air has been replaced from outdoors. In typical homes this seems to range from around 1 to 5 hours. It's longer in more energy-efficient homes and (much) shorter if windows are open.

It's more common (and more confusing) to use something called an "air exchange rate".

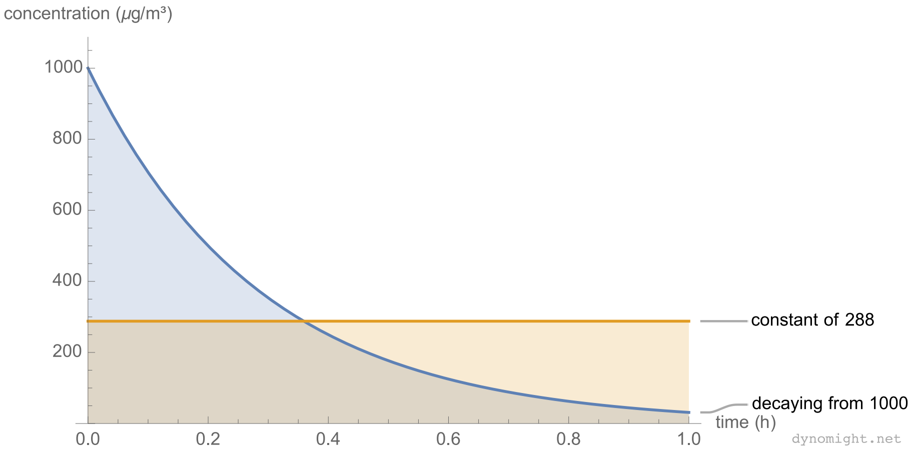
If you search for "air exchange rate" you'll mostly see definitions like "number of air changes per hour". This obviously doesn't make any sense – there aren't discrete changes. If you search very hard, you can find that the air exchange rate is the constant **a** such that after time **t**, a fraction of **exp(- a t)** of the initial air is remaining. We can solve this for **1/2** to get that the half-life is **t=ln(2)/a=0.693/a**.

In summary, unless you actively clean the air, indoor levels are probably similar to outdoor levels *plus* whatever particles you generate inside. The air is exchanging fast enough that by the time particles have fallen out of the air, new particles have come from outside. (However, closed windows should work for larger particles and there might be some benefit to opening/closing your windows based on changing outdoor levels.)

Exposure with decaying concentrations

To calculate the impact of specific things that generate particles, we'll need a quick calculation. If you generate a puff of smoke with a peak concentration of **c** and a half-life of **h**, levels will fall off slowly over time. What's your total exposure?

The following graph shows a peak concentration of **c=1000** that decays with a half-life of **h=0.2** hours. This turns out to be equivalent (same area-under-the curve) to being exposed to a constant of 288 for 1 hour.



Where did 288 come from? The general formula is **1.44 × c × h**. This is natural enough: It's the peak concentration times the half-life times an extra constant because of math. As you'd expect, doubling the peak concentration or half-life doubles the total exposure.

The general formula comes from a simple integral.

After an amount of time **t** particles have undergone **t/h** half-lives, meaning the total concentration will have decayed by a factor of **(1/2)^(t/h)**, and so the current level is at time **t** is **c×(1/2)^(t/h)**. If we integrate this to get the total exposure, it is **c×(1/2)^(-t/h) dt = c × h/ln(2)**. It happens that **1/ln(2) 1.44**.

Things that create particles indoors

Smoking: Have you heard? Smoking is bad for you. Shaw (2000) estimates that one cigarette reduces (non disability-adjusted) life expectancy by 11 minutes, and helpfully points out that this is enough time for "fairly frantic sexual intercourse". Let's move on.

Vaping: The following is entirely about *second-hand* vape smoke. (I advise against vaping unless you're doing it to quit smoking.)

It sometimes feels like the public health community is hellbent on destroying vaping, damn the evidence. Sure, I always thought, vaping isn't great, but surely the harm is small?

The vaping community often points to this report from the CDC, which didn't find any problems. However, *they didn't check for particles.* This proves nothing.

Lots of people have measured particles near vaping and found high levels. Soule (2017) found levels around 1000 at an e-cigarette conference. Li (2021) found an average level of 276 in six vape shops in Southern California. Protano (2020) found that a single person vaping could create levels as high as 1000-10,000. Lots of random people measure particles and report numbers like 600 or 546 or 1000. Some report that their parents made them do this (good job, parents.)

Vaping enthusiasts retort that the particles are almost all water and glycerol plus a tiny amount of nicotine. This appears to be true – unlike tobacco smoke, exhaled vape air doesn't contain significantly more phenols or carbonyls (like formaldehyde or acetone) than regular exhaled air.

I'm tentatively calling this for Team Vape. In the mountains of papers measuring high particles from vape smoke, few even acknowledge the claim that it's all water and glycerol. Maybe I'm missing something? My idea of an anti-vaping conspiracy has not receded.

Fireplaces and solid fuels: Various papers have found that a typical day living as an 18th century farmer or being a Viking or cooking over an open fire in Guatemala could all lead to average daily particle concentrations around 200. These studies always find insane particle levels near the fire (in the thousands), but these buildings have good ventilation and people don't spend *that* much time near the worst smoke. This would suggest a loss of around 6 DALYs, though real Vikings might have created less smoke after generations of practice at tending fires.

Still today, there are many people throughout the world who cook indoors with solid fuels (coal/wood). In India, including the harms from household air particulates nearly doubles the (already large) losses that ambient air pollution causes.

How much could it hurt you to have a fire in a fireplace? This depends on many factors: How big is the fire? Is it burning well? How well-ventilated is your home? Do you have a glass panel in front of the fire? How well does the flue for the fireplace draw air?

Let's be *really* pessimistic and imagine that you have a fire that produces a concentration of 25,000 and that this lasts for five hours. This would cost you around 50 disability adjusted life hours. Doing everything right can probably reduce this by 2-3 orders of magnitude. A modern wood stove with tight seals could do even better.

Cooking: What about other types of cooking? Again, it all depends. Kang

(2019) tried cooking various things in 30 different buildings in Korea. On average, soup produced a peak concentration of 65, frying 424, and broiling 1256. Both frying and broiling were hugely variable.

Opening a window reduced the peak concentration to around as much. Using a range hood (with or without an open window) reduced it to around as much.

And how long do the particles stick around? With no range hood or ventilation, they found a mean half-life of around an hour. (That's actually on the faster end of the ventilation half-lives we talked about above.) With a range hood, the half-life was around 20 minutes. With open windows, it was around 6-7 minutes (regardless of the range hood).

These numbers suggest broiling fish with the windows closed leads to a total loss of around 45 minutes.

Candles: Candles produce some particles while burning, but the vast majority of their particles happen in an instant when the candle is extinguished. (This is confirmed by other research.) Blowing out a candle causes a spike to around 50-200 with particles hang in the air for 3-5 hours.

Assuming average values, blowing out a candle costs 10 minutes. Doing it every day of your life costs 0.5 years.

Take a spike of 100 and assume that the particles hang in the air for 4 hours. Then the cost is 100 × 4 / 2500 = 0.16 hours or 9.6 minutes.

The obvious solution is to avoid candles. If that's a nonstarter for you, candles probably aren't that bad, just *don't blow them out.* Instead, extinguish them with an airtight lid.

Incense: One paper suggests that burning an incense stick could produce a peak concentration of around 800, while a cone might produce a peak concentration of over 4000, with a half-life of around 1.6 hours.

This suggests burning a cone of incense costs 3.68 hours.

Using our half-life formula, the total exposure is 4000 × 1.6 × 1.44 = 9216 particle hours. Thus, the total cost is 9216 / 2500 = 3.68 disability adjusted hours.

Don't use incense.

Aerosols: Some cleaning products create a ton of particles. One paper found that using Febreze caused particles to spike by 50-75, They also found that hairspray could cause a spike of up to 200, with a half-life of 2 hours. These also seem to be exceptionally small particles that hang in the air for a very long time.

This suggests that using hairspray (indoors) has a cost of around 14 minutes.

Assume an increase of 200 and a half-life of 2 hours. Then, the total cost is 200 × 2 × 1.44 / 2500 = .23 hours.

Try to avoid spraying anything into the air. If you really love hairspray, you could use it outside, or in a well-ventilated room right before leaving.

Humidifiers: You're not going to like this, but ultrasonic humidifiers produce huge numbers of particles. They turn any minerals in the water become airborne particles. They do this almost by *design*! Park (2020) tested different types of water in a small chamber with a carefully controlled air exchange rate and got the following steady-state increases over background levels.

- Mineral water: ~265
- Tap water (Seoul): ~260
- Purified water: ~50
- Distilled water: ~0

This appears to show that it's not just water particles being counted. Some real-world people report even higher numbers. I've tested this myself and found the same thing.

This suggests that using an ultrasonic humidifier with tap water at night for 8 hours costs 50 minutes. Doing it every night costs 2.6 years.

(Since we're looking at steady-state concentrations rather than peak, we don't consider half-lives.) Your total exposure for one night is 260 × 8 = 2080, with a cost of 2080/2500 = 0.832 hours = 50 minutes. If you do this every night, it raises your mean daily exposure by 260 × 8 / 24 = 86.66 with a cost of 86.66 / 33.33 = 2.6 years.

Using a humidifier at night is almost as bad as smoking 5 cigarettes!? I was skeptical such large numbers could be real. I read the instruction manuals for a few ultrasonic humidifiers. There were a few passages suggesting they are not unaware of the issue:

- "Use only clean, cool tap water to fill the Water Tank (filtered or distilled water is recommended to avoid white dust if tap water is too hard.)"

- "The best way to minimize mineral build-up is to use distilled or demineralized water."

- "IMPORTANT: Using tap water with a high mineral content aka 'hard water' with any humidifier can cause a fine white dust to be emitted. To avoid this, use distilled or demineralized water."

As further evidence, the EPA has a report that says, "Recent studies […] have shown that ultrasonic and impeller (or 'cool mist') humidifiers can disperse materials, such as microorganisms and minerals, from their water tanks into indoor air." They recommend emptying and cleaning the tank *every day* as well as using distilled water.

The EPA emphasizes that there's no proof that these *particular* particles harm health. That's an… interesting… way of looking at things. It might be appropriate logic for a government contemplating a ban, but not for us. These are huge effects, and you should assume ultrasonic humidifiers are dangerous until there's convincing evidence to the contrary.

In theory, you might try to solve the problem by using distilled water or humidifying a room during the way with the door closed. But don't. What if the humidifier gets dirty? What if particles leak out of the room? What if you have bacteria build up and pollute your water? Are you really going to unplug, clean, and dry the humidifier every time you use it?

Just use an evaporative or steam humidifier, which seem to create almost no particles.

Other random things

Vacuuming has been reported to cause particles to spike by 50. Cleaning dryer lint causes a spike in PM_{10} but not $PM_{2.5}$. You can get a small spike by *making a bed*!

Overall, there's no way to avoid creating particles. Almost any activity that involves molecules will create *some* particles. In reality, exposure for a small amount of time isn't a big deal. The solution is to avoid the biggest problems and make sure particles are removed from the air quickly enough to avoid major harm. You can do this by purifying the air. Or, if you're a lucky person with clean and temperate air outside, you can just keep your windows open.

What to do

Try not to create particles indoors.

Above all, kill your ultrasonic air humidifier. Don't burn stuff while cooking. Don't use incense at all ever. Extinguish candles them with a lid. Avoid dangerous cleaning products. If you use hairspray, I guess you could do it outside?

Monitor outdoor levels.

Most places in the world have real-time particle measurements. You should follow these along with the weather. Some weather websites/apps already include particle counts – consider using these.

Get a particle counter.

You can estimate the air exchange rate of your home and the rate your purifier removes particles. You can try to reduce activities that generate particles indoors. But the only way to be sure is to check.

You can get an OK-ish particle counter for $100. Carry it around with you sometimes to check for any nasty surprises. Share it with your friends. We really need better wearable particle counters. Ideally, these should be integrated into watches or phones.

(Lots of people have asked me for a recommended particle counter. I don't want to recommend the one I have because it's somewhat crap. If you search for "air quality monitor" there are many options that seem better. Let me know if you have a particle counter you actually like.)

Use an air purifier inside.

An air purifier in your home has two purposes:

- To reduce the steady-state level caused by outdoor particles drifting indoors.
- To reduce the half-life of particles you generate indoors.

It turns out that an air purifier reduces both of these by exactly the same fraction. To be more precise, assume

- Your ventilation half-life is **t**, meaning that after this amount of time, half the air in your home has been replaced with outdoor air. (Typical times might be one to five hours.)
- Your purification half-life is **s**, meaning that after this amount of time, your purifier removes half the particles from the air.

When you turn on the purifier, because of math, your total exposure from both indoor and outdoor sources gets multiplied by a factor of

s / (s + t).

This is intuitive: You want purification to be faster than ventilation, i.e., you want **s < t**. In that's true, then your total exposure is reduced to a small fraction of what it would be without the purifier.

For example, suppose your home has an average ventilation half-life of 120 minutes and that you run a cuboid air purifier on low, meaning a purification half-life of 15 minutes in a 31 m^3 room. Then that purifier reduces your exposure to a fraction of =15/(15+120) of what it would otherwise be.

While this works for both indoor and outdoor particles, remember that it always helps to open the windows if outdoor levels are *currently* lower than indoors — e.g., if you just burned dinner, it helps to open the windows until indoor levels equalize with outdoor levels.

Because this gets a little bit into the weeds, I'm suppressing the details of how to derive that formula here.

First let's talk about the steady state. In a fixed amount of time, purifiers remove a fixed fraction of particles from the air, while ventilation will replace a fixed fraction of indoor air with outdoor air. Thus, the more particles, the faster purification works and the less ventilation does. The steady state is when these are equal.

If outdoor levels are **L**, it turns out that the steady-state for indoor levels is **L × s / (s + t).**

For example, suppose outdoor levels are 100, your home has an average ventilation half-life of 120 minutes and that you run a cuboid air purifier on high, in a medium room, meaning a purification half-life of 7 minutes. Then the steady-state concentration will be 5.5 = 100 × 7 / (7 + 120).

Now let's talk about particles created indoors. Say you generate a bunch of smoke and get a peak concentration of **c**. If particles only go away because of ventilation, your total exposure ends up being **1.44 c t**. Now suppose you have an air filter that would have a half-life of **s** without ventilation. The combined half-life ends up being **st/(s+t)**, so your new exposure is **1.44 c s t / (s+t)**. The ratio of new and old exposure is again **s/(s+t)**.

Put a HEPA cabin air filter in your car

One car company hypes the hell out of their air purification system. It's great to raise awareness but… well… most cars made any time in the last 10-20 years have cabin air filters which the air goes through before being blown into your face.

These usually aren't HEPA, but you can buy a HEPA version for like $10-$20, and easily install it yourself. For most cars you just pull out the glovebox and then slide the filter out. It takes around 5 minutes and requires absolutely no tools.

If you're like most people, you didn't even know you had a cabin air filter, so it's probably time to replace it anyway. Pay the extra $5 to get one that removes small particles. (This might reduce the air velocity a little.)

Masks are really tricky

In some parts of the world, it's common to wear masks to reduce the impact of particles. How much is this helping? I don't think anyone really knows. Still, we can make a few conclusive statements:

1. It is possible to use a mask to eliminate the majority of particle exposure (>90 % reduction).
2. Many widely sold masks simply don't perform as advertised, regardless of how they are used.
3. To use a mask successfully, it's paramount to have one that fits your face and to check the fit.

In laboratory tests, researchers buy a bunch of n95 masks, and then carefully attach them to either dummies or real human heads, and then check how well they work. Optimistically, Shakya (2016) found two n95 masks both worked well, and even cloth masks did something. Richard Saint Cyr tried a bunch of masks in China while being very careful about fit and got numbers between 56% and 99%. Some papers (Cherrie, 2018; Pacitto, 2019) tested a bunch of different masks and found that one or two perform well but the majority achieve remove half of particles or less. Dismally, Faridi 2020 tried 50 different masks available in Iran and found that *none* did better than 40% and most were much worse.

My guess is that real-world conditions are closer to the bad cases than to the good ones.

What can you do? Two things:

- Refuse to buy any mask without an independent test to prove it works.
- Obsessively fit-test– a low-quality check is to look at what happens when you inhale or exhale – is it going around the sides of the mask? But really you want a formal fit test.

A common method for "real" fit-tests is to create an aerosol with saccharin (yes, the artificial sweetener) and check if the mask blocks the smell. Provenzano (2020) gives some preliminary evidence that you can do this at a low cost.

There's no proven DIY recipe for this yet, but it seems possible. Maybe we can use those damn ultrasonic humidifiers to make the saccharin aerosol. Who knows, maybe some unrelated reason to be interested in masks might come up.

The irrelevance of test scores is greatly exaggerated

Here's some claims about how grades (GPA) and test scores (ACT) predict success in college.

> In a study released this month, the University of Chicago Consortium on School Research found—after surveying more than 55,000 public high school graduates—that grade point averages were five times as strong at predicting college graduation as were ACT scores. (Fortune)

> High school GPAs show a very strong relationship with college graduation despite sizable school effects, and the relationship does not differ across high schools. In contrast, the relationship between ACT scores and college graduation is weak-to nothing once school effects are controlled. (University of Chicago Consortium on School Research)

> "It was surprising not only to see that there was no relationship between ACT scores and college graduation at some high schools, but also to see that at many high schools the relationship was negative among students with the highest test scores" (Science Daily)

> “The bottom line is that high school grades are powerful tools for gauging students’ readiness for college, regardless of which high school a student attends, while ACT scores are not.” (Inside Higher Ed)

See also the Washington Post, Science Blog, Fatherly, The Chicago Sun Times, etc.

All these articles are mild adaptions of a press release for Allensworth and Clark’s 2020 paper “High School GPAs and ACT Scores as Predictors of College Completion”.

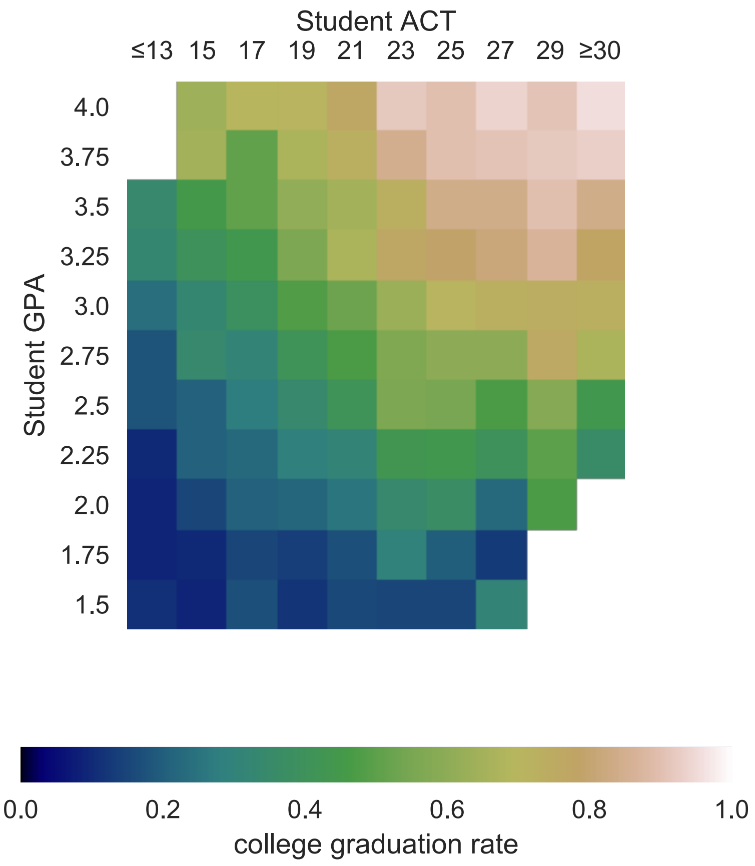
I understood these articles as making the following claim: **Standardized test scores are nearly useless (at least once you know GPAs), and colleges can eliminate them from admissions with no downside.**

Surprised by this claim, I read the paper. I apologize if this is indelicate, but… the paper doesn’t give the slightest shred of evidence that the above claim is true. It’s not that the paper is *wrong*, exactly, it simply doesn’t address how useful ACT scores are for college admissions.

So why do we have all these articles that seem to make this claim, you ask? That’s an interesting question! But first, let’s see what’s actually in the paper.

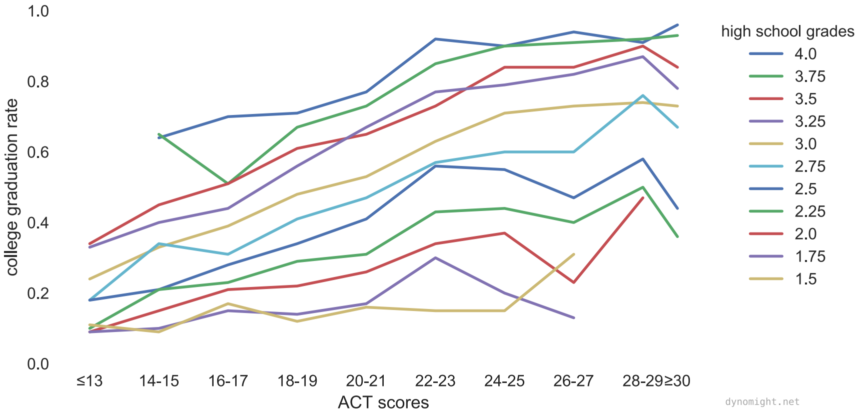
Test scores are not irrelevant

The authors got data for 55,084 students who graduated from Chicago public schools between 2006 and 2009. Most of their analysis only looks at a subset of 17,753 who enrolled in a 4-year college immediately after high school. Here’s the percentage of those students who graduated college within 6 years for each possible GPA and ACT score:



dynomight.net

We can also visualize this by plotting each row of the above matrix as a line. This shows how graduation rates change for a fixed GPA score as the ACT score is changed.



It doesn't *appear* that ACT scores are useless… But let's test this more rigorously.

Test scores are highly predictive

The full dataset isn't available, but since we have the number of students in each ACT / GPA bin above, we can create a "pseudo" dataset, with a small loss of precision in the GPA and ACT score for each student. I did this, and then fit models to predict if a student would graduate using GPA alone, ACT alone, or with both together. (The model is cubic spline regression on top of a quantile transformation.)

To measure how good these fits are, I used cross-validation, repeatedly holding out 20% of the data, fitting a model like above to the other 80%, and then predicting if each student will graduate. You can measure how accurate the predictions are, either as a simple error rate (1-accuracy) or as a Brier score. I also compare to a model using no features, which just predicts the base rate for everyone.

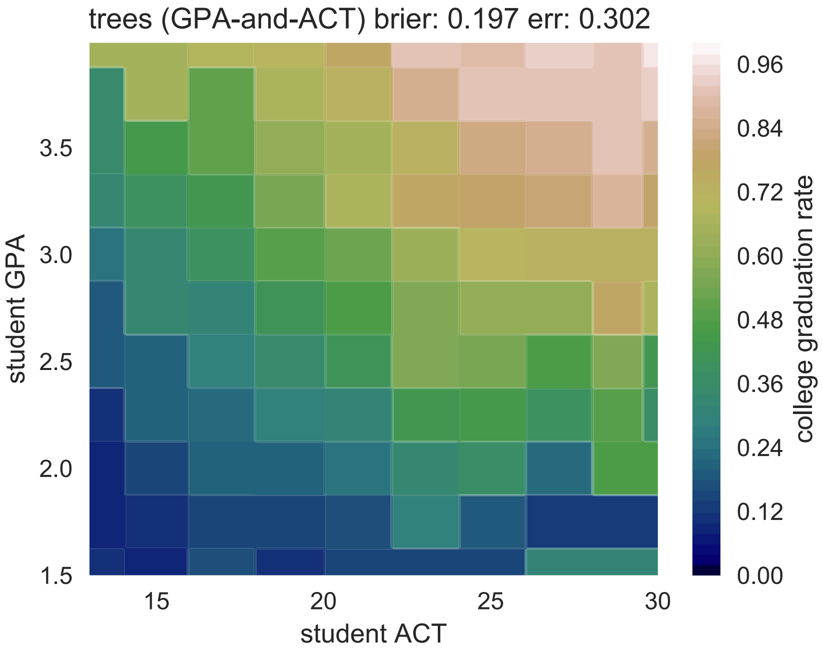
Predictors	Brier	Error
Nothing	.249	.491
ACT only	.219	.355
GPA only	.210	.330
both	.197	.302

It's true that GPA does a bit better than the ACT. But if you care about that difference, you should care *even more* about the difference between (GPA only) and (GPA plus ACT). It's not coherent to simultaneously claim that the GPA is better than the ACT *and also* that the ACT doesn't add value to the GPA.

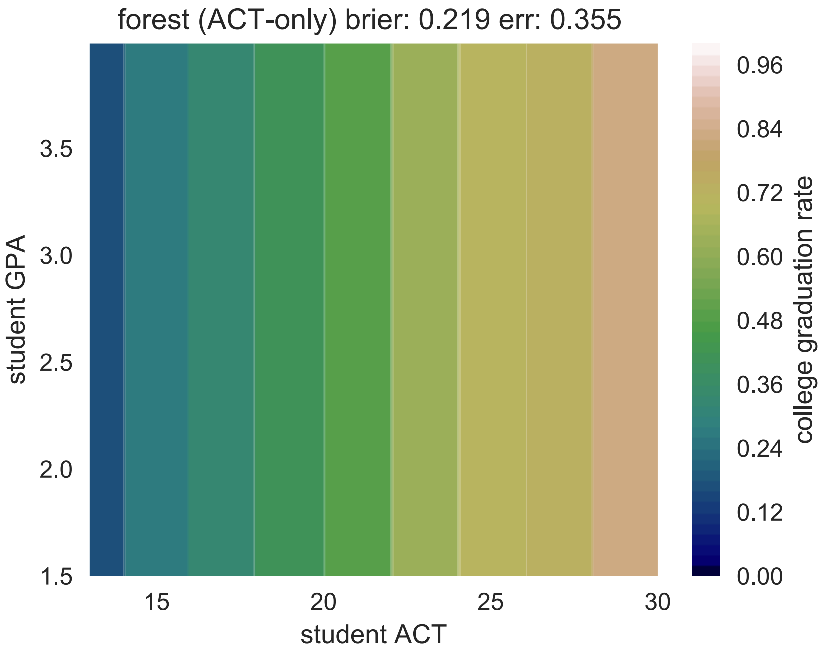
I repeated this same calculation with other predictors: logistic regression, decision trees, and random forests. The numbers barely changed at all.

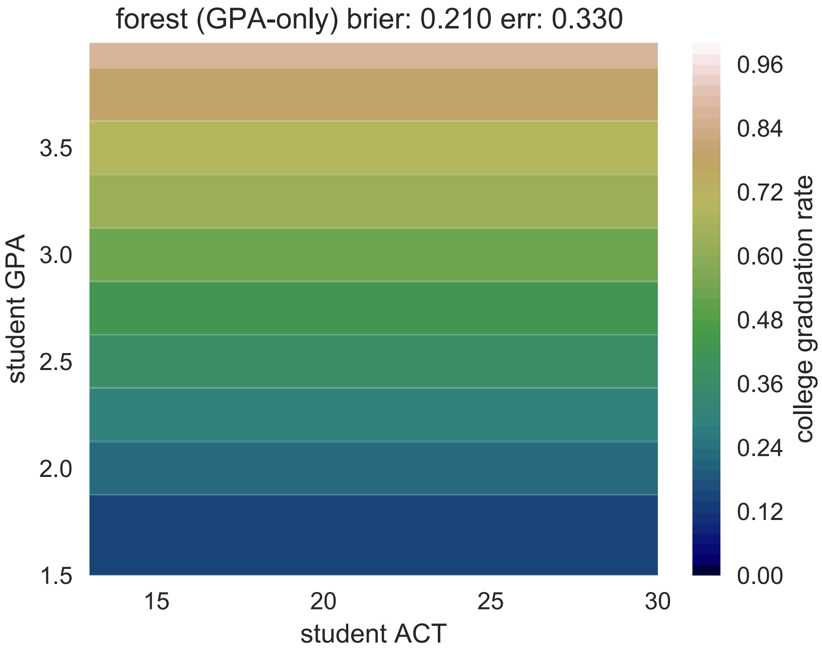
Here's logistic regression with almost regularization: (These don't look perfectly linear because of the quantile transform.)

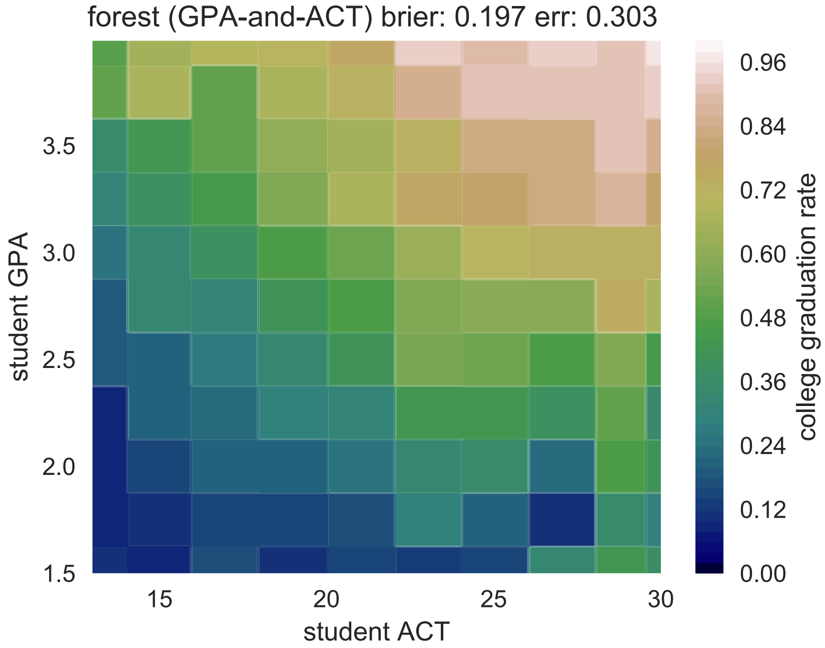
Here's trees, grown to maintain at least 10 data in each leaf.



Here's random forests:


forest (ACT-only) brier: 0.219 err: 0.355
student GPA
3.5
3.0
2.5
2.0
1.5
15
20
25
30
student ACT
0.96
0.84
0.72
0.60
0.48
0.36
0.24
0.12
0.00
college graduation rate


forest (GPA-only) brier: 0.210 err: 0.330
student GPA
3.5
3.0
2.5
2.0
1.5
15
20
25
30
student ACT
0.96
0.84
0.72
0.60
0.48
0.36
0.24
0.12
0.00
college graduation rate



Here's the error rates of all the methods:

predictor	spline	logreg	trees	forests
ACT only	.355	.353	.355	.355
GPA only	.329	.330	.330	.330
both	.301	.303	.302	.303

And here are the Brier scores:

predictor	spline	logreg	trees	forests
ACT only	.219	.219	.219	.219
GPA only	.210	.210	.210	.210
both	.197	.197	.197	.197

Still, these are all just calculations based on the first table in the paper.

What the paper actually did

For each student, they recorded three variables:

- Gender

- Ethnicity (Black, Latino, Asian)
- Poverty (average poverty rate in the student's census block)

For the students who enrolled in a 4-year college, they recorded four variables about that college:

- The number of students at the college
- The percentage of full-time students
- The student-faculty ratio
- The college's average graduation rate

They standardized all the variables to have unit mean and unit variance (except for gender and ethnicity since these are binary). For example, GPA=0 for someone with the average grades, and GPA=-2 for someone 2 standard deviations below average.

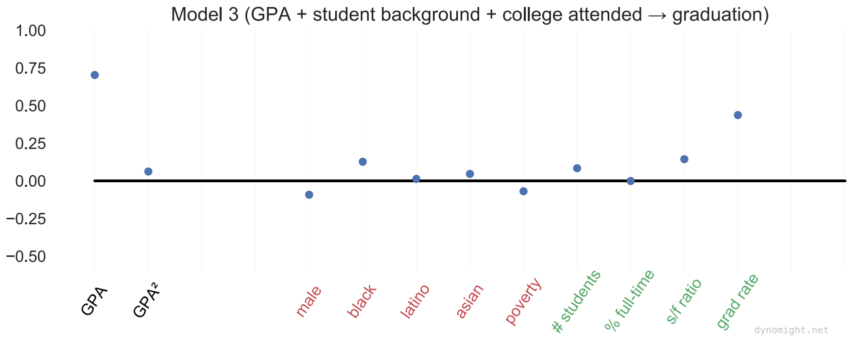
They also included squared versions of GPA and ACT, GPA^2 and ACT^2. These are never negative and larger for any student who is *unusual* either on the high or low end. They do this because the relationship is "slightly quadratic", which is reasonable, but it's not explained why the other variables don't get a squared version.

With this data in hand, they fit a bunch of models.

First, they predicted graduation rates from grades alone. Higher grades were better. There's nothing really surprising here, so let's skip the details.

Second, they predicted graduation rates from ACT scores alone. Higher ACT scores were better. As you'd expect, this relationship is strong. Again, let's skip the details.

Third, they predicted graduation rates from grades, student variables, and variables for the college the student enrolled at. This model gets a "likely-to-graduate" score for each student as follows. This labels student background variables and college institutional variables in different colors for clarity.

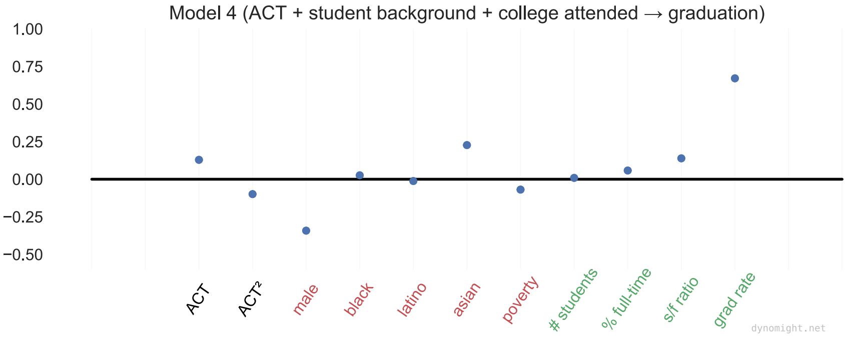


The "likely-to-graduate" score becomes a probability after a sigmoid transformation. If you're not familiar with sigmoid functions, think of them like this:

If a student has a score is **X** then graduation probability is around **.5 + .025** × **X**. For larger **X** (say |**X**|>**1**) scores start to have diminishing returns, since probabilities must be between 0 and 1.

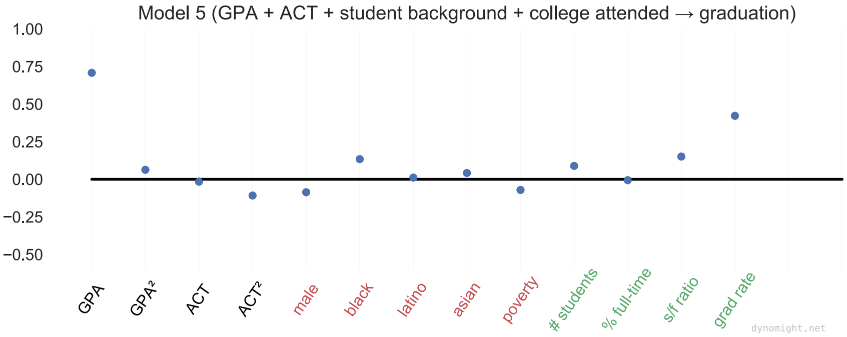
For example, the coefficient for (male) above is -.092. This means that a male has around a 2.3% lower chance of graduating than an otherwise identical female. (For students with very high or very low scores the effect will be less.)

Fourth, they predicted graduation rates from ACT scores, student variables, and college variables.



The dependence on ACT is less than the dependence on GPA in Model 3. However, the dependence on student background and college variables is much higher.

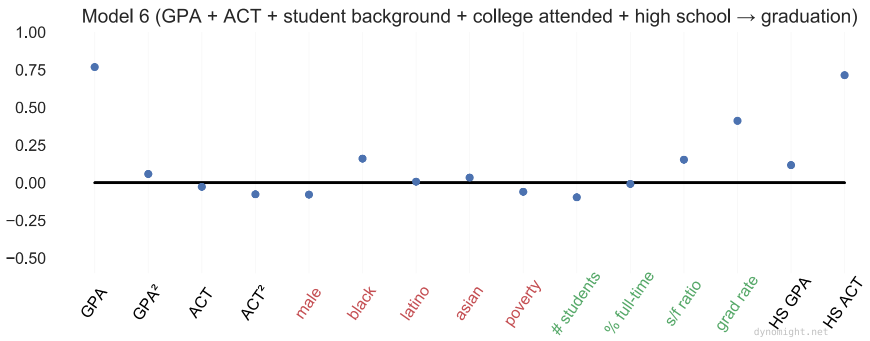
Fifth, they predicted graduation rates from GPAs, ACT scores, student variables, and college variables.



Here, there's minimal dependence on ACT, but a *negative* dependence on ACT^2, meaning that extreme ACT scores (high or low) both lead to lower likely-to-graduate scores.

Does that seem counterintuitive to you? Remember, we are taking a student who *is already enrolled in a particular known college* and predicting how likely that are to graduate from that college.

Sixth, they predicted graduation rates from the same stuff as in the previous model, but now adding mean GPA and ACT for the *student's school.* They also now standardize some variables relative to each high school.



I can't tell what variables are affected by this change of the way things are standardized. My guess is that it's just for GPA and the SAT, but it might affect other variables too.

What this says about how to do college admissions

I mean… not much?

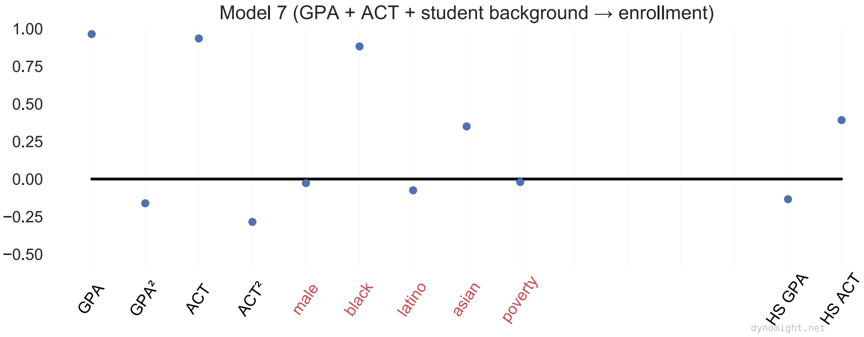
Here's what these models do: Take a student with a certain GPA, ACT scores, background who *is accepted to and enrolls in* a given college. How likely are they to graduate?

It's true that these models have small coefficients in front of ACT. But does this mean ACT scores aren't good predictors of preparation for college? No. ACT scores are still influencing *who enrolls in college* and *what college they go to*. These models made that influence disappear by dropping all the students who didn't go to college, and then conditioning on the college they went to.

These models don't say much of anything about how college admissions should work. There's three reasons why.

First, these models are conditioning on student background! Look at the coefficients in Model 5. What exactly is the proposal here, to do college admissions using those coefficients? So, college should explicitly penalize men and poor students like this model does? Come on.

Second, test scores influence if students go to college at all. This entire analysis ignores the 67% of students who don't enroll in college. The paper confirms that ACT scores are a strong predictor of college enrollment.



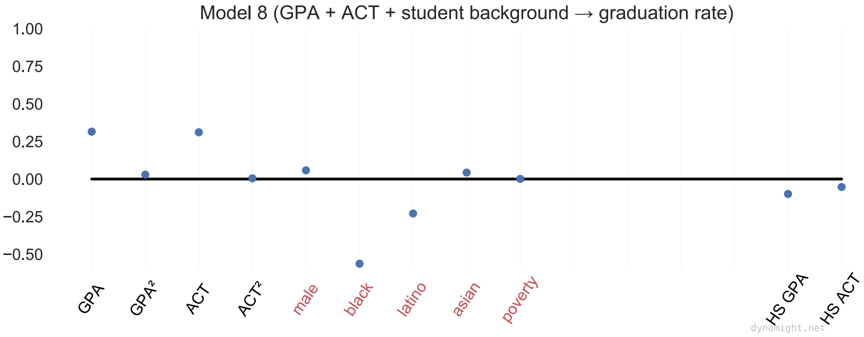
Of course, many factors influence if a student will go to college. Do they want to? Can they get in? Can they afford it?

You might say, "Well *of course* the ACT is predictive here – colleges are using it." Sure, but that's because colleges think it gauges preparation. It's possible they're wrong, but… isn't that kind of the question here? It's absurd to *assume* the ACT isn't predictive of college success, and then use that assumption to *prove* that the ACT isn't predictive of college success.

Third, for students who go to college, test scores influence *which* college they go to, and more selective colleges have higher graduation rates. Here's three private colleges in the Boston area and three public colleges in Michigan.

College	acceptance rate	average graduation rate
Harvard University	5%	98%
Northeastern University	18%	85%
Suffolk University	84%	63%
University of Michigan - Ann Arbor	23%	92%
Michigan State University	71%	80%
Grand Valley State University	83%	60%

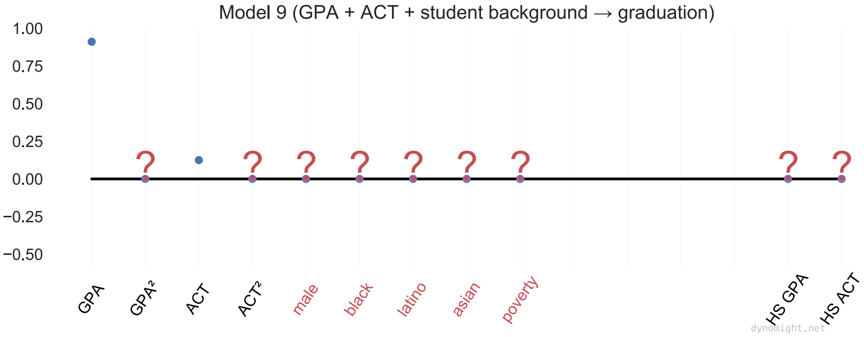
The paper also does a regression on students who go to college to try to predict *the graduation rate of the college they end up at.* Again, GPA and ACT scores are about equally predictive.



Of course, you could also drop the student background and college variables, and just predict from GPA and ACT. But remember, we did that above, and the ACT was extremely predictive.

Alternatively, I guess you could condition on student background *without* conditioning on the college students go to. I doubt this is a good idea or a realistic idea, but at least it's *causally possible* for colleges to use such a model to do admissions.

Why didn't the authors do this? Well… Actually, they did.



Unfortunately, this is sort of hidden away on a corner of the paper, and no coefficients are given other than for GPA and ACT. It's not clear if high-school GPA or ACT are even included here. The authors were not able to provide the other coefficients (nor to even acknowledge multiple polite requests *notthatimbitteraboutit*).

The laundering of unproven claims

What happened? There's really nothing fundamentally *wrong* in the paper. It fits some models to some data and gets some coefficients! Interpreted carefully, it's all fine. And the paper itself never really pushes anything beyond the line of what's technically correct.

Somehow, though, the *second* the paper ends, and the press release starts, all that is thrown out the window. Rather than "ACT scores definitely predict college graduation, but they don't seem to give much *extra* information if you already know if and where they're going to college, plus their sex, ethnicity, and wealth", we get "ACT scores don't predict college success".

To be fair, a couple hedges like "once school effects are controlled" make their way into the articles but are treated as a minor technical asides and never explained.

Let's separate a bunch of claims.

1. It might be desirable to reduce the influence of test scores on college admissions to achieve worthy social goals.
2. It might be that test scores don't predict college graduation rates.
3. It might be that test scores only predict college graduation because selective (and high graduation-rate) colleges choose to use them in admissions.
4. It might be that if selective colleges stopped using test scores in admissions, test scores would no longer predict college graduation.

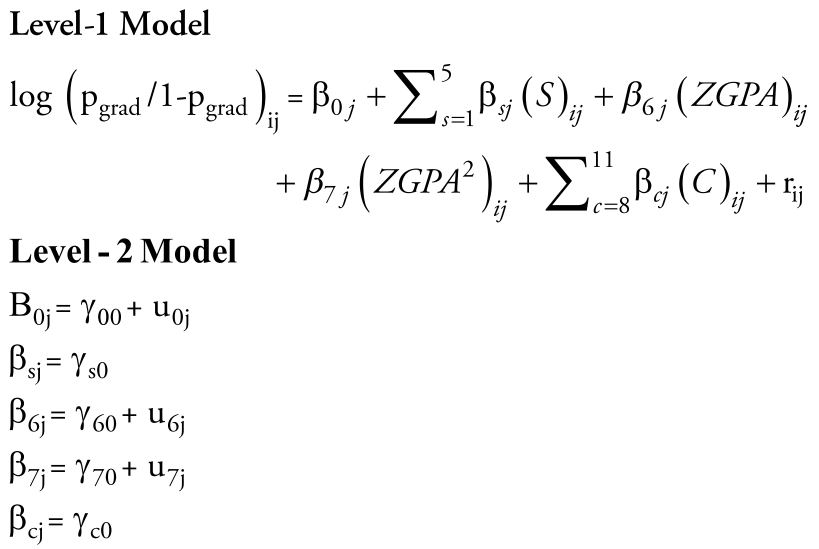
I'm open to claim #1 being true. If you believe #1, it would be *convenient* if #2, #3, and #4 were true. But the universe is not here to please us. #2 is not just *unproven* but *proven to be false.* This paper gives no evidence for #3 or #4. Yet because these claims were inserted into the public narrative after peer review, we have a situation where the paper isn't *wrong*, yet it is being used as evidence for claims it manifestly failed to establish.

Journals don't issue retractions for press releases.

A field guide

There's a number of ambiguities, undefined notation and straight-up errors in the paper. If you try to read it, these might throw you off (or make you wonder about Educational Researcher's review process). I've created a guide to help you on your way.

The equations for the models in the paper look like this:



Level-1 Model

$$\log\left(p_{\mathrm{grad}}/1\text{-}p_{\mathrm{grad}}\right)_{\mathrm{ij}} = \beta_{0j} + \sum_{s=1}^{5}\beta_{sj}\left(S\right)_{ij} + \beta_{6j}\left(ZGPA\right)_{ij} + \beta_{7j}\left(ZGPA^{2}\right)_{ij} + \sum_{c=8}^{11}\beta_{cj}\left(C\right)_{ij} + \mathrm{r}_{\mathrm{ij}}$$

Level - 2 Model

$$\mathrm{B}_{0\mathrm{j}} = \gamma_{00} + \mathrm{u}_{0\mathrm{j}}$$

$$\beta_{\mathrm{sj}} = \gamma_{s0}$$

$$\beta_{6\mathrm{j}} = \gamma_{60} + \mathrm{u}_{6\mathrm{j}}$$

$$\beta_{7\mathrm{j}} = \gamma_{70} + \mathrm{u}_{7\mathrm{j}}$$

$$\beta_{\mathrm{cj}} = \gamma_{\mathrm{c}0}$$

If you want to understand this, you face errors, undefined notation, and the fact that the actual statistical methodology is never explained. First, let's talk about the errors/typos:

- There's a a pair of missing parenthesis on the left.
- The first sum makes no sense, since $\mathbf{(S)_{ij}}$ doesn't depend on $\mathbf{s}$. I think that this should be $\mathbf{(S)_{si}}$ instead.
- The final sum makes no sense, since $\mathbf{(C)_{ij}}$ doesn't depend on $\mathbf{c}$. I think that this should be $\mathbf{(C)_{cj}}$ instead.
- $\mathbf{B_{0j}}$ is a mistake. It should be $\mathbf{_{0j}}$.

If we fix those errors, we get this corrected model:

Level-1 Model

$$\log\left(p_{grad}/(1\text{-}p_{grad})\right)_{ij} = \beta_{0j} + \sum_{s=1}^{5}\beta_{sj}\overset{(S)_{si}}{\cancel{(S)_{ij}}} + \beta_{6j}\left(ZGPA\right)_{ij} + \beta_{7j}\left(ZGPA^{2}\right)_{ij} + \sum_{c=8}^{11}\beta_{cj}\underset{(C)_{cj}}{\cancel{(C)_{ij}}} + r_{ij}$$

Level - 2 Model

$$\cancel{B_{0j}} \xleftarrow{} \beta_{0j}$$

$$\cancel{B_{0j}} = \gamma_{00} + u_{0j}$$

$$\beta_{sj} = \gamma_{s0}$$

$$\beta_{6j} = \gamma_{60} + u_{6j}$$

$$\beta_{7j} = \gamma_{70} + u_{7j}$$

$$\beta_{cj} = \gamma_{c0}$$

Next, let's talk about undefined notation. At no point does the paper define **i**, **j**, or $\mathbf{r_{ij}}$. (Undefined notation isn't as bad as it sounds, these are probably cultural knowledge in the authors' community.) This makes it tricky to decode, but here's my best attempt:

- The left-hand side is the "score" for student **i** who happens to be in high-school **j**. You transform that score to a probability through the sigmoid transformation, since **score = log(p/(1-p))** is equivalent to **p= (score)**.
- $\mathbf{S_{1i}}$, ..., $\mathbf{S_{5i}}$ are the background variables for that student. (Gender, ethnicity, poverty.)
- $\mathbf{(ZGPA)_{ij}}$ is the student's GPA, standardized to have zero mean and unit variance. (Called a "z-score")
- $\mathbf{(ZGPA^2)_{ij}}$ is the square of $\mathbf{(ZGPA)_{ij}}$. (Don't get triggered by the location of the parentheses.)
- $\mathbf{(C)_{8j}}$, ..., $\mathbf{(C)_{11j}}$ are the institutional variables for college **j** (size, % full time, student/faculty ratio, mean graduation rate).
- The variable are the learned coefficients. The coefficients for the intercept and GPA terms vary by school and are fit as part of a multilevel model, while the others are fixed.
- $\mathbf{r_{ij}}$, $\mathbf{u_{0j}}$, $\mathbf{u_{6j}}$, and $\mathbf{u_{7j}}$ are residuals.

Frankly, there's still some issues here, but I can't be bothered trying to fix anything else.

Third, the paper never explains what methodology they use to take a dataset, and estimate the parameters of the above model. My guess is that the algorithm is maximum-likelihood. But using maximum-likelihood requires a prior for all the residual terms. The paper never says what that is. Probably a standard Gaussian? Again, this might be "obvious" to the typical reader of this paper, but if the journal is issuing press-releases, shouldn't they make a cursory attempt to make the paper readable by general audiences?

Finally, a small thing. Their Table 1 lists the ACT ranges as 0-13, 14-16, 16-17, which doesn't make sense because it repeats 16. I think 14-16 should be 14-15, so I treated it like that above.

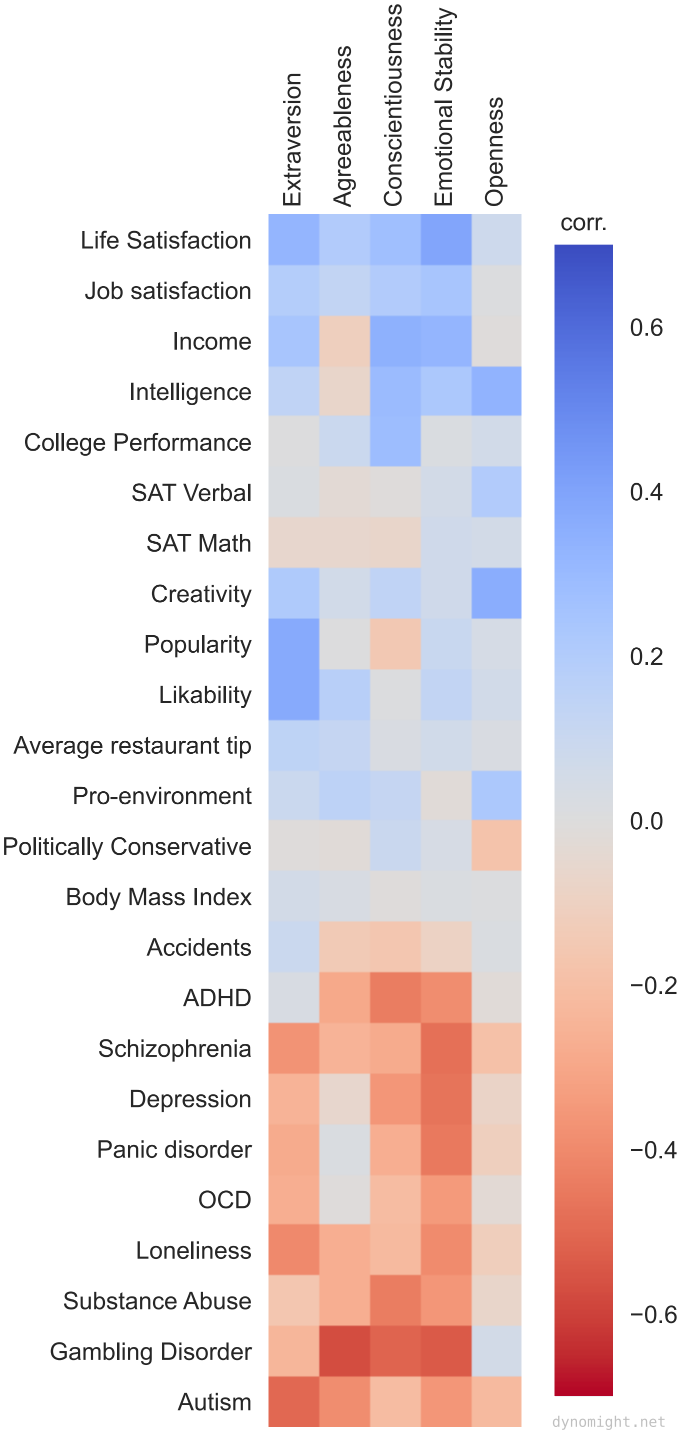
Are some personalities just better?

I don't know if you like parties. I don't know if you're organized or punctual. But I bet you don't like rotting smells or long swims in freezing water.

That is to say: People are different, but only in certain ways. What's the difference? Hypothermia enthusiasts have few kids, so their genes tend to disappear. If introverts were worse at breeding than extraverts, then the same thing would have happened. Since extraversion varies widely, we can infer that we're at an equilibrium point with no real advantage either way. (Personality traits are around 40% genetic.)

So, no personality is *better* than any other. Instead, there must be intricate tradeoffs, with each personality occupying a different kind of niche.

That's what I thought, anyway. Then I read a few dozen papers and made this table:


Extraversion
Agreeableness
Conscientiousness
Emotional Stability
Openness
Life Satisfaction
Job satisfaction
Income
Intelligence
College Performance
SAT Verbal
SAT Math
Creativity
Popularity
Likability
Average restaurant tip
Pro-environment
Politically Conservative
Body Mass Index
Accidents
ADHD
Schizophrenia
Depression
Panic disorder
OCD
Loneliness
Substance Abuse
Gambling Disorder
Autism
corr.
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
dynomight.net

This shows correlations between the Big Five personality traits and various personal characteristics. Blue shows positive correlations, while red shows negative. For example, the blue upper-left cell shows that extraversion is associated with life satisfaction, whereas the red lower-left cell shows that extraversion is *negatively* associated with autism.

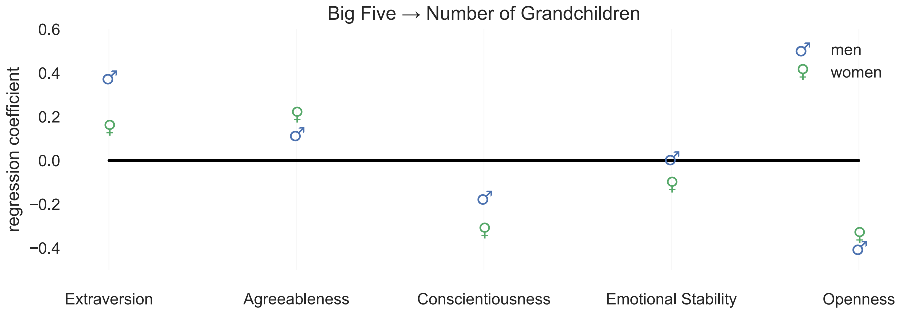
So, uhh, where are the tradeoffs? People who are extroverted, agreeable, conscientious, emotionally stable, and open seem to do better at basically everything. Let's call these people **all-blues**. Broadly speaking, they are more happy, successful, intelligent, creative, and popular. They have fewer addictions and less of every mental disorder. The only real tradeoffs are agreeableness against income/intelligence and extraversion/conscientiousness against math scores.

I'd like to give a list of famous all-blues as examples, but this doesn't seem to exist. As a proxy, we can look to Myers-Briggs, where all-blues are similar to emotionally stable ENFJs. The internet claims that examples of ENFJs are Michael Jordan, Oprah, Pope John Paul II, Martin Luther King Jr., Pericles, and Barack Obama. For the opposite type, famous ISTPs supposedly include the Dalai Lama, Ernest Hemingway, Snoop Dogg, Melania Trump, and Vladimir Putin. (Personally, I assume the Dalai Lama has high emotional stability, but judge for thyself.)

Anyway, what's the deal here? Why don't we see more tradeoffs? Is the idea of a population equilibrium mistaken?

Evolution don't care

Evolution doesn't care if you're happy. Evolution only wants you to pass on your genes. Berg et al. (2014) took data from 10.7k representative Americans born between 1900 and 1947 and did a regression to predict the number of grandchildren someone has from their personality traits. Here are the regression coefficients:



Coefficients from "Personality and long-term reproductive success measured by the number of grandchildren", Berg. et al, 2014.

The personality characteristics are standardized so Extraversion = 0 for someone who is average, and Extraversion = -2 for someone 2 standard deviations below average, etc. They focus on grandchildren to reflect the influence of a parent's personality on a child's survival, but just using children gives similar results.

If you're wondering, this suggests the ESFP as the most fecund MBTI type (Ronald Reagan, Bill Clinton, Hugh Hefner).

On the one hand, this would explain why everyone isn't an all-blue: If you want to dominate the personality landscape, you need to reproduce more. On the other hand, it creates a bigger puzzle: If we were in population equilibrium, all the coefficients would be zero! Instead, there are huge effects like extroverted men having 0.8 more grandchildren than introverted men. If that's true, then we are *way* out of equilibrium, and future generations will look different from us.

It's tempting to make up post-hoc stories for those coefficients. ("High openness people spend too much time on rationalist-adjacent blogs," har-har, yes, very good.) But it's not that simple. You've got to do one of two things.

- You might reject the idea of a population equilibrium. If so, you should explain why the rules of natural selection don't apply here.
- You might claim that humans *used* to be in equilibrium, but there's been some recent change to reproductive fitness that evolution hasn't caught up to yet.

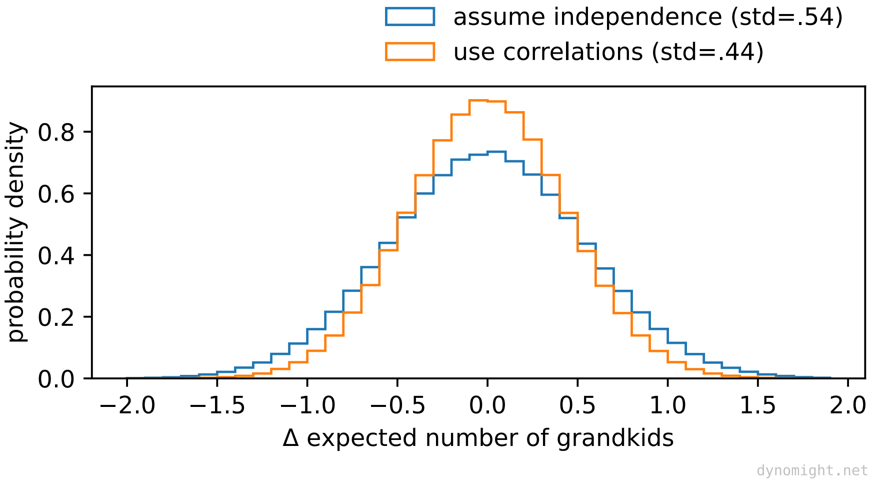
What changes could have thrown us out of equilibrium? It can't be the dawn of agriculture—there's been too many generations for effects this strong to persist. It would need to be more recent like the industrial revolution or the invention of birth control. Intelligence isn't exactly a personality trait, but Udry (1978) surveyed 225 women on birth control. After three years, the percentage of low, medium, and high-IQ women who accidentally had a baby was 11.1%, 8.2%, and 3.4%, respectively. This paper is old and I couldn't find any replications, so I wouldn't put too much faith in it. Still, it's plausible that conscientiousness could have a similar role.

Aside: Personality traits are correlated. For example, extraversion is correlated with openness. But if we account for this, there's still tons of variability in how many grandkids different personalities should expect.

Lieu pointed out that there are correlations between personality traits. Vukasović and Bratko (2015) do a meta-analysis, arriving at the following correlations.

	E	A	C	ES	O
E	1	.051	.122	.231	.413
A	.051	1	.413	.438	.114
C	.122	.413	1	.442	.208
ES	.231	.438	.442	1	.188
O	.413	.114	.208	.188	1

Fortunately for us, correlations alone are enough to generate the normalized variables (z-scores) that we need to plug into the above regression. I generated a bunch of "random people" either sampling from either an independent multivariate Normal distribution, or a multivariate Normal distribution with the above table as a covariance matrix. I then plugged those people into the regression model and computed a histogram for each.



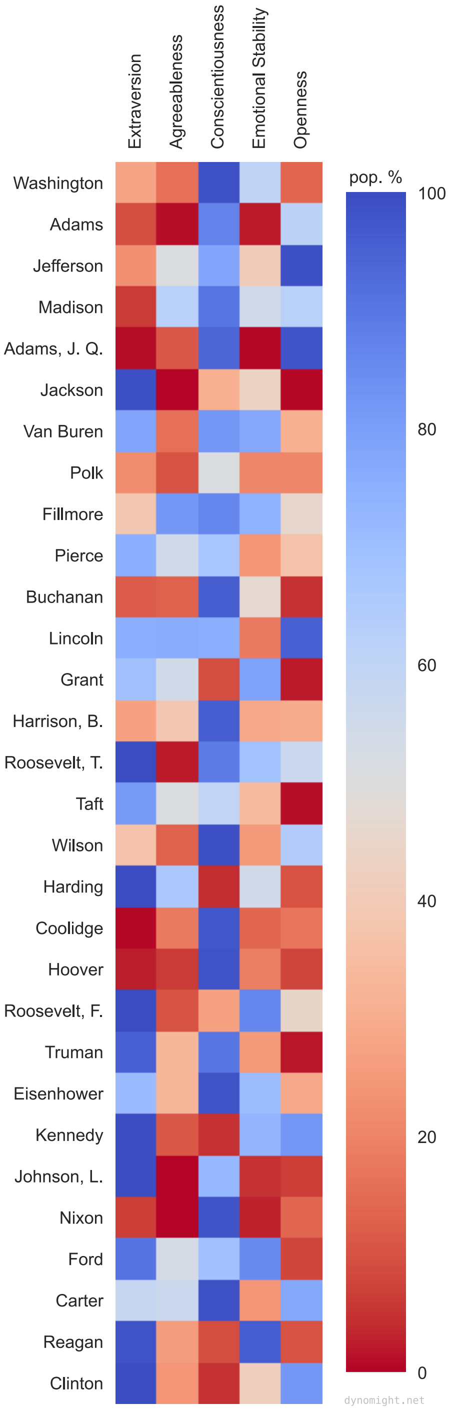
Sure enough, the standard deviation is less if we account for the correlations. But it's only a *bit* less. In any case, the grandkids model is a *regression*. Correlations among with inputs don't change the fact that certain people (high openness introverts) have fewer grandkids than others (low openness extraverts).

I suspect we really are out of equilibrium. Modern lives are very different than even 5-10 generations ago, and it would be strange if this didn't impact how much different people reproduce. But there's no reason to think we're evolving in an all-blue direction.

A look at presidents

Forget about evolution for a second. Do successful people still tend to be all-blue?

Rubenzer et al. (2000) had multiple expert historians profile American presidents. Here are their results, presented as a percentage of the population:

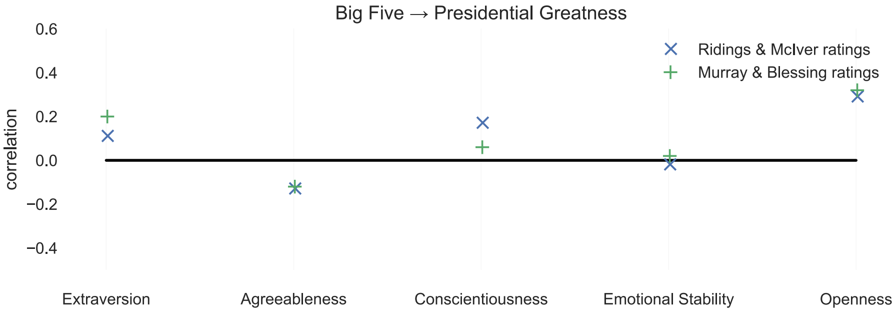


Data from "Personality, Character, and Leadership In The White House: Psychologists Assess the Presidents", Rubzenzer et al. (2004)

Washington is a 98.6% on Conscientiousness. Nixon is a 0.02% on Agreeableness.

It's *very* hard to become president. If an all-blue personality was better, we'd see that here. Instead, among recent presidents, we see high extraversion, low agreeableness, and no clear trend otherwise. (Most US households gained radios around 1930 before the election of Franklin Roosevelt, so that's a reasonable place to think of the "modern" era starting.)

If being all-blue doesn't help you *become* president, does it make you a good one? It happens that the closest there's been to an all-blue president was Lincoln, often considered greatest president. (He scores low on emotional stability due to his lifelong struggles with depression.) But that could be a random coincidence. To be more data-driven, the paper finds correlations between personality factors and how *great* a president is rated to be.



Data from "Assessing the U.S. Presidents using the Revised NEO Personality Inventory", Rubenzer et al., 2000

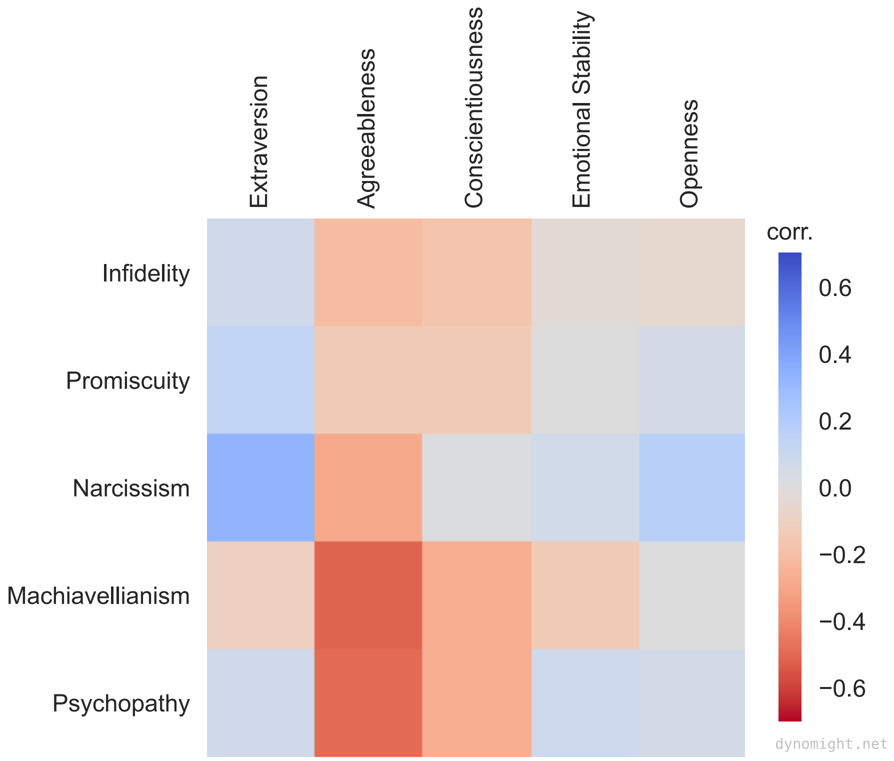
dynomight.net

This is similar to the profile of an all-blue, except that agreeableness is bad, and emotional stability doesn't matter. Teddy Roosevelt probably fits this profile best.

You might also learn things from looking at sub-traits. For example, *Agreeableness* has different facets: *Tender-mindedness* is correlated with greatness, while *compliance* and *straightforwardness* are anti-correlated.

The darkness hypothesis

Why are recent presidents usually extraverted and low in agreeableness, but otherwise so mixed? Here's my guess:



This isn't to say that presidents are all narcissists or psychopaths. (Though, who are we kidding, some are.) It's widely agreed now that "narcissism" and "psychopathy" aren't discrete categories. Rather, they are "spectrum traits" that we all have to some degree.

How do we arrive at a spectrum of psychopathy? It's another equilibrium process. If there were no psychopaths, maybe the first one to show up would probably manipulate everyone and have a thousand kids. As psychopaths became more common, everyone's defenses go up, and the strategy becomes less useful. It's not shocking that these traits might be useful in politics.

Summary

All-blues might really be happier and healthier. If so, it could be a result of late modernity, or it might have always been true. Still, that doesn't mean that evolution will favor all-blues, or that all-blues are "more successful".

What about the rest of us, who aren't all-blue? (Market research suggests my median reader is high openness and conscientiousness, but low extraversion and agreeableness.)

Well, it's all just correlations. It's not that low agreeableness *causes* people to have gambling disorders. I'm not even sure that statement makes sense! Rather, other factors (genes, environment) cause both. If you aren't a psychopath, then correlations aren't something to worry about. If you *are* a psychopath, then… probably you're not worried about my advice?

Data Sources

Life Satisfaction - Anglim et al, 2020 (Many other measures of happiness are similar)

Job satisfaction - Judge et al., 2002

Income, Intelligence - Judge et al., 1999

ADHD - Nigg et al., 2002 (Table 8 Grand M)

Schizophrenia - Ohi et al., 2016

Autism - Lodi-Smith et al. 2018

Depression, Panic disorder, OCD, Substance Abuse - Kotov et al., 2010

Loneliness - Buecker et al., 2020

Popularity, Likability - D. van der Linden et al., 2010

Gambling Disorder - Dash et al., 2019 (Table 2, 2+ symptoms, average men + women)

College Performance - Vedel, 2014

SAT Verbal, SAT Math - Noftle and Robins, 2014

Creativity - Zare and Flinchbaugh, 2018

Average restaurant tip - Lynn, 2021

Accidents - Beus et al., 2015

Pro-environment - Soutter et al, 2020 (traits as attitudes)

Politically Conservative - Sibley et al, 2012

Body Mass Index - Sutin et al., 2015

Infidelity, Promiscuity - Schmitt, 2004

Narcissism, Machiavellianism, Psychopathy - Muris et al, 2017

The Presidential Big Five data is from *Personality, Character, and Leadership In The White House: Psychologists Assess the Presidents* by Rubenzer and Faschingbauer, 2004. It was quite an effort to extract these numbers from the book and get them into a useable form. Since they might be useful to others, I've uploaded the raw CSV data here: presidents.csv names.csv

Update: The “all-blue” personality is essentially the General Factor or *Big One* in psychology, and has been debated since it was proposed by Musek (2007). Thanks to arctor_bob for pointing this out.

Alcohol, health, and the ruthless logic of the Asian flush

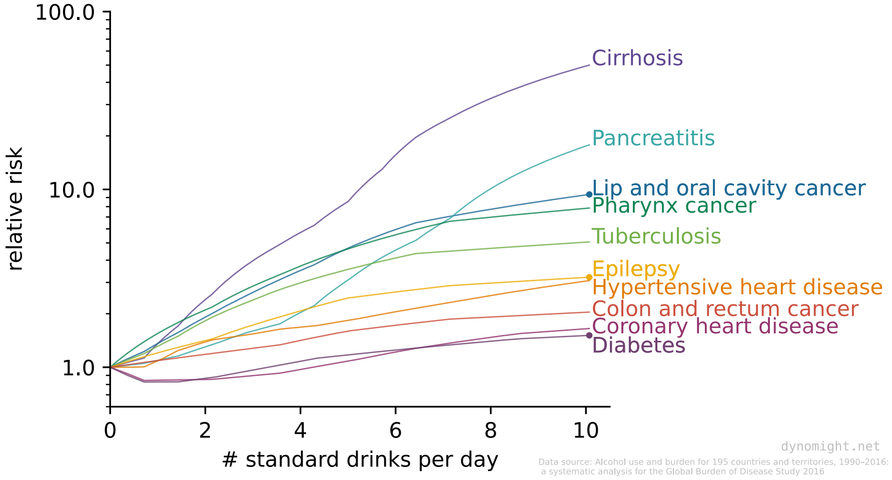
Say you’re an evil scientist. One day at work you discover a protein that crosses the blood-brain barrier and causes crippling migraine headaches if someone’s attention drifts while driving. Despite being evil, you’re a loving parent with a kid learning to drive. Like everyone else, your kid is completely addicted to their phone, and keeps refreshing their feeds while driving. Your suggestions that the latest squirrel memes be enjoyed *later at home* are repeatedly rejected.

Then you realize: You could just sneak into your kid’s room at night, anesthetize them, and bring them to your lair! One of your goons could then extract their bone marrow and use CRISPR to recode the stem-cells for an enzyme to make the migraine protein. Sure, the headache itself might distract them, but they’ll probably just stop using their phone while driving. Wouldn’t you be at least tempted?

This is an analogy for something about alcoholism, East Asians, Odysseus, evolution, tensions between different kinds of freedoms, and an idea I thought was good but apparently isn’t.

It’s not good to drink too much

This is a surprise to no one, but let’s look at some numbers. Here’s data from a meta review on the relative risk of various health conditions as a function of the number of US standard drinks (14g of alcohol) someone has in a day:



Data source: Alcohol use and burden for 195 countries and territories, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016

The three small dots show that having 10 drinks a day is associated with a 9x risk of getting lip/oral cancer, a 3x risk of epilepsy, and a 1.5x risk of diabetes, as compared with not drinking at all. These are all *associations*, controlling only for age, sex, and drinking history. This makes the little dip around 1-2 drinks for heart disease and diabetes controversial. Still, the causal link is pretty clear in many cases, and for our purposes, all that matters is that heavy drinking is not good.

But who averages *10 drinks* per day, you ask? The answer is *an astonishing number of people*. Half of Americans drink almost nothing, but the top 10% average more than 10 drinks per day. They're responsible for around 75% of all alcohol consumption.

Some East Asians struggle with alcohol

Humans metabolize alcohol in various ways. The "normal" way is that an ADH enzyme converts alcohol to acetaldehyde after which an ALDH enzyme breaks the acetaldehyde down into acetate. Eventually the acetate is broken down into water and carbon dioxide. The intermediate product (acetaldehyde) is highly toxic and carcinogenic, while acetate is much less active. It appears that ethanol itself isn't carcinogenic, but acetaldehyde is.

But guess what: Around 80% of East Asians have a variant of ADH (ADH1B or ADH1C) that converts alcohol to acetaldehyde more quickly. Also, around 50% of East Asians have a variant ALDH isoenzyme (ALDH2*2) that breaks down acetaldehyde more slowly. Both of these mean that acetaldehyde tends to accumulate, leading to a "flush" reaction.

Kang et al. (2014) recruited a bunch of healthy 20-something Korean men. Here is the peak acetaldehyde concentration (ng/ml) of people with different genes after consuming 0.25 g/kg of ethanol (around 1.25 standard drinks for someone who weighs 70 kg / 154 lb.)

ALDH \ ADH	half variant	full variant
standard	167.9	190.1
full variant	736.6	1,613.6

The variant enzymes lead to much higher peak concentrations. Remember, we have two copies of every gene, one from mom and one from dad. The left column shows people with one copy of the ADH1B variant while the right column shows people with two copies. This doesn't even include people with zero copies of the ADH1B variant, presumably because they couldn't find enough of them. The top row shows people with the standard ALDH2 enzyme, the bottom row with the East Asian variant. This is dominant so you don't have to worry about half-effects.

It's likely that having these variant enzymes means that if you *do* drink, alcohol causes more problems. This is hard to study since you can't do randomized tests and people with the mutation drink less, but there's pretty strong evidence of this in humans for esophageal cancer. In mice, removing the ALDH enzyme greatly increases the DNA damage that alcohol causes to the stomach.

Those East Asians drink less

Now, *why* do East Asians have these genetic variants? I long assumed that this is because other people had a longer tradition of drinking alcohol, and so had evolved to do it painlessly. This is totally incorrect.

Let's back up. When Homo Sapiens left Africa, these variants essentially didn't exist. We evolved to be able to consume alcohol, probably because we're fond of not starving to death. (If rotting fruit is the only source of calories, it's better if you can eat it without getting incapacitated.) The genes for these variant enzymes probably arose later, in China.

Alcoholic beverage production started early in China. It's hard to say exactly when, since it pre-dates recorded history, but 9000 year old Chinese pottery already has residues of early beers. Alcohol production in Egypt seems to have started around 5000 years ago, and in Europe perhaps 4000 years ago.

So, China is where alcohol first became common, and also where genes that make alcohol consumption difficult first became common. Why would such "defective" genes arise in the place where alcohol has been around the longest?

The simplest explanation is that these genes are *adaptive*. It's obvious in retrospect: Humans are prone to alcoholism. Alcoholics tend to get sick, commit suicide, and have accidents, all of which interfere with having and raising kids.

A study in Taiwan found that 48% of the control population had a copy of the (dominant) ALDH2*2 mutation that slows the breakdown of acetaldehyde, but only 12% of alcoholics did. Similarly, 93% of the control population had at least one copy of the ADH1B gene that speeds acetaldehyde production, compared to 64% of alcoholics. Other studies (Muramatsu et al. 1995, Hurley and Edenberg, 2012, Bierut et al., 2012) confirm the same basic picture, which is that these genes reduce alcoholism.

If these genes really are an adaptation, it shows how ruthless evolution can be. If you implanted a device in your kid that mildly poisoned them every time they drank, you'd be a monster. But evolution basically did that.

No one cares about my freedom to rob convenience stores or burn down public buildings. We all understand that different people's freedoms are in conflict, and we've invented things like "manners" and "property" and "noise ordinances" to navigate the tradeoff.

There's a different tradeoff I think about a lot. We all know the story of Odysseus having his men block their ears and tie him to the mast of his ship. He knew he would go temporarily insane when going past the Sirens, so he wanted to remove freedom from himself to overcome that.

It's a cute story, but it's not typical. Odysseus constrained his future self with technology. Most real-world scenarios are different:

- We need *society* to enforce constraints.
- Those constraints affect *everyone* to some degree, even those who don't want them.

I almost never buy snacks because once home, I can't resist the urge to eat them. This works OK for me, but they're sometimes available at conferences or parties or whatever, and I have a hard time saying no. What I'd *really* like is for society to criminalize all mint-chocolate flavored snacks.

(We'll get back to alcohol in a second, I promise.)

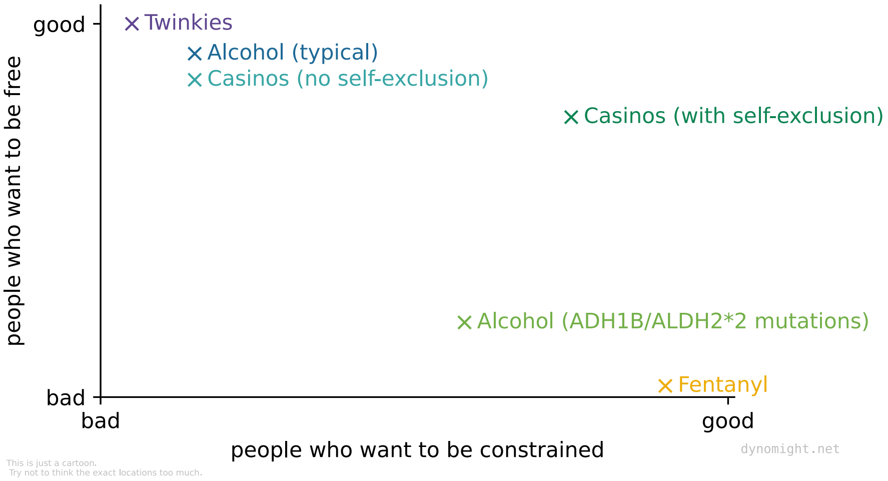
Snacks are laughable, but how about fentanyl? *Some* of the reason drugs are illegal is because of externalities or the idea that people don't know what's good for them. However, there's no doubt that some former or potential addicts would *choose* criminalization if it was up to them. Say you used to be addicted but now you've quit. If you could snap your fingers and make all drugs disappear, wouldn't you do that?

Obviously, criminalizing cookies (or fentanyl) is bad for both responsible users and people who can't or don't want to quit. I'm just trying to point out that *there is a tradeoff*. Society has decided that tradeoff in favor of responsible Twinkie users and against responsible fentanyl users.

Just as society made a different tradeoff for Twinkies and drugs, *biology* made a different one for alcohol, depending on if you got the East Asian variants or not.

Sometimes we can give people the chance to "Odysseus" themselves without intruding too much on the freedoms of others. An example is gambling. Some locations allow people to "self-exclude" from gambling, after which casinos won't let you play for a time period of your choice. This isn't perfect, since now responsible gamblers have their ID checked, and addicts can still cross state lines or play the lotto or whatever. But it's pretty good.

We can informally picture the different regimes like so:

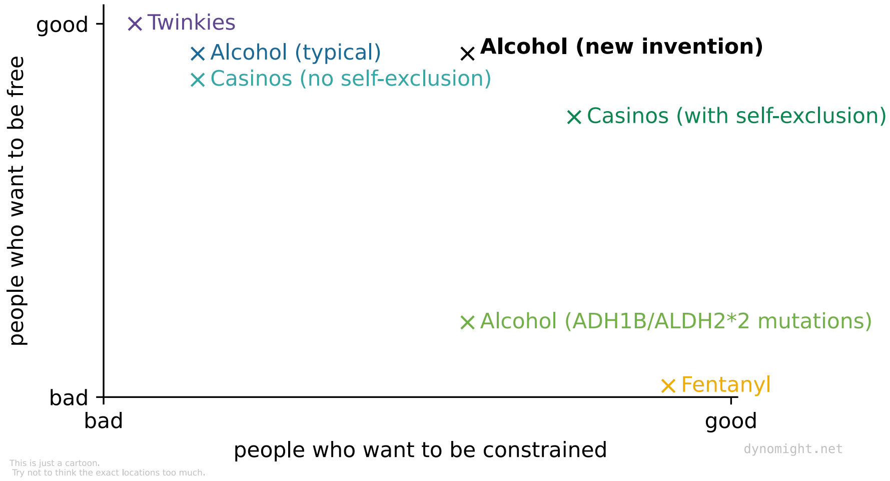


A self-ban for alcohol

Roughly 10% of people in the US are raging alcoholics. Could we offer them the chance to self-exclude from alcohol?

It seems very difficult. We'd have to force some heavyweight process of checking IDs on all bars and liquor stores. Even then, it wouldn't be very effective, since people could still have their friends buy it for them. Do we want to make it illegal to hand out drinks at a party without checking everyone's ID against a database? It would be a nightmare.

A while ago, I had a strange idea. In principle, instead of having people ban themselves from alcohol *legally*, couldn't it be done *biologically*? After all, this is the solution evolution came up with. Can we allow people to *opt-in* to getting the Asian flush?



Obviously, this is just hypothetical. For one thing, is it even possible? It seems hard, but with the full might of our modern nano RNAi cell-therapy CAR-T stem-cell gene therapy arsenal, perhaps we could figure something out. It doesn't matter though. We could never actually *give* such a drug to people, since it's equivalent to poisoning them.

Just kidding. It's called disulfiram, and it was approved by the FDA in 1951.

Does disulfiram work?

Still today, we don't really know.

To be sure, disulfiram does what's asked of it. It definitely blocks the ALDH enzyme and this definitely does lead to 5-10x higher acetaldehyde concentrations. This definitely causes "flushing" and other symptoms typical of those with a genetic predisposition against drinking. Early on, it was prescribed in such massive doses that patients who drank anyway sometimes went into cardiac arrest, or even died.

What's unknown is if it helps with alcohol addiction. There have been a huge number of studies, but none of them give clear answers. As far as I can tell, there are three problems.

First, just imagine you give alcoholics a bottle of pills, explain that they make alcohol (more) toxic, and send them on their way. Obviously, almost nobody takes them, and those that do are ultra-determined and would probably have quit anyway. You can ask patients to come into the office to take the drugs, but then people drop out. It's just incredibly hard.

Second, there's all sorts of confounders and weirdness. Some show effects on *number of days without alcohol* but not on *number of drinks* or vice-versa. Some studies show great results for people who are married but not for single people.

Third, most of the studies are kind of.. crap? Hughes and Cook reviewed the studies up to 1997. Their paper is a marvel of inventive euphemisms like "Nothing can be said directly", and "not strictly a controlled study", and "a very poor study, but the authors subsequently stated that they 'made no claims for methodological sophistication or statistical significance'."

The drug is still available today, though not much used for alcoholism except in Denmark, where it's widely prescribed. (This seems to be pharma-nationalism, resulting from the drug being invented there.)

If people refuse to take the pills, couldn't we just make some kind of implant? This too has been experimented with since 1968, and again we have no clear answers. One major problem is that it's not clear how well the drug is absorbed from the implant. Another is that randomized trials require "sham implants" to blind participants. A third is that various trials used ridiculously low doses of the drug, far below the level that's physiologically plausible.

Update: People have pointed out that disulfiram implants are apparently fairly popular in Eastern Europe (1, 2, 3). However, these implants typically have a total dose of 1-2g dispensed over something like 6-24 months. If you assume 1g dispensed over a year, that's 2.7 mg / day, around 1% of a typical oral dose of 250 mg / day. On top of that, bioavailability of implanted disulfiram appears to be lower than oral. So I suspect these implants are almost entirely placebo.

So, it's hardly been revolutionary. What explains this?

One possibility is that the drug *could* cure alcoholism, we just haven't done enough studies, or found a sufficiently reliable way to deliver it yet.

Another possibility is that the alcohol intolerance in East Asians is just a "nudge", which is often enough to prevent alcoholism from forming in the first place, but not strong enough to displace alcoholism once it's taken root.

I favor the second possibility. Disulfiram definitely does make acetaldehyde build up when you drink. If that had a massive effect on alcoholism, it shouldn't be that hard to see it! Yet we still don't see much after 70 years. These days, the first-line drug treatments for alcoholism are acamprosate which reduces the physical symptoms of alcohol withdrawal and naltrexone which screws around with the opiod receptors and probably reduces the pleasure people get from drinking.

What are we supposed to conclude from all this? That we should be careful about cute evolutionary explanations? That human fallibility means your individual freedom is in tension with my *freedom from temptation*? That "Odysseusing" is a way to resolve that tension? That our addictions run deep into us, and aren't easy to remove? That there's nothing new under the sun? That human behavior is complex, and harder to manipulate than mere biology? Take your pick. Honestly, I just thought it was a good story.

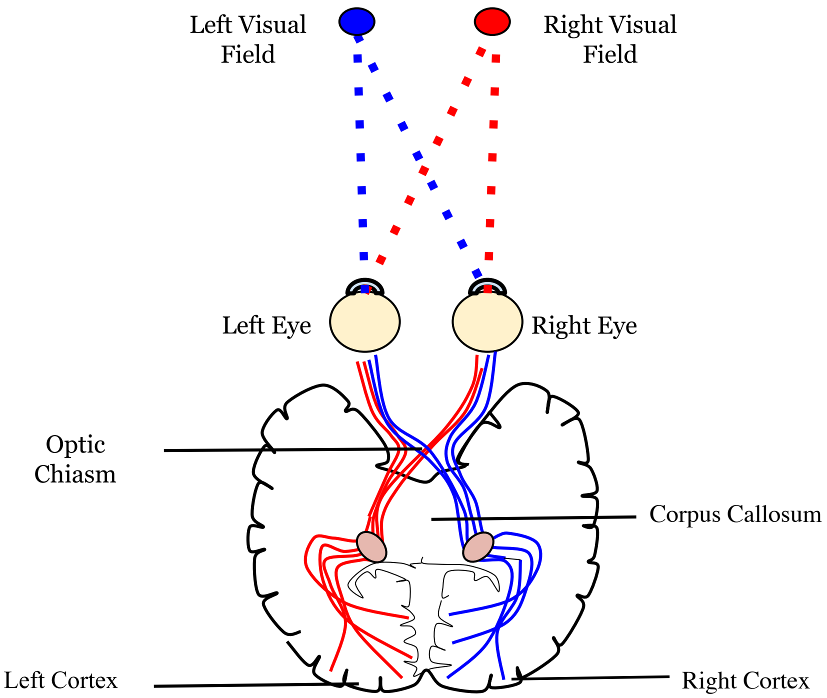
A review of early split-brain experiments

What happens if you cut your cortex in half? When this was first tried on animals, the answer seemed to be *not much*. But starting in the late 1950s, a series of experiments found that very weird things happen under careful testing. These experiments are fascinating for their implications into the mind, consciousness, and selfhood.

Existing presentations of these experiments have a lot of ~~philosophical rambling~~ interpretive narrative. This is dangerous. When I actually read the original papers, I was shocked how much they diverged from what I'd read before. If you have a story for what's happening, it's way too easy to explain away inconvenient results everywhere.

These experiments are just *so* weird, and their implications *so* fascinating, that the only way to make sure you're getting an accurate picture is to focus on the facts: What was actually done, and what was observed? Here I'll review early animal experiments, particularly those that set the stage for Roger Sperry's later research on humans.

A Reminder About Mammal Brains

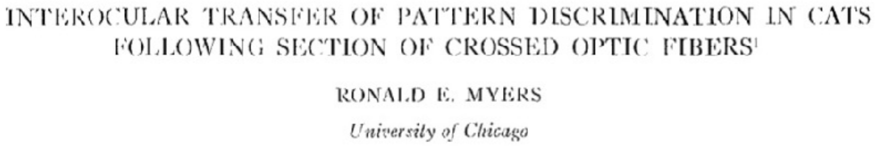


The other layer of the brain is the cortex, key for attention, perception, memory, vision, and language. The cortex has two hemispheres which communicate mainly via the corpus callosum. They also communicate via the subcortical brain and indirectly through the body.

Light from both eyes first goes to the optic chiasm. This relays the signal from the left side of *both* eyes to one hemisphere of the cortex, and the signal from the right side of both eyes to the other hemisphere.

It appears that one side of the body is primarily *but not completely* controlled by one half of the cortex.

Myers 1955



INTEROCULAR TRANSFER OF PATTERN DISCRIMINATION IN CATS FOLLOWING SECTION OF CROSSED OPTIC FIBERS[1]

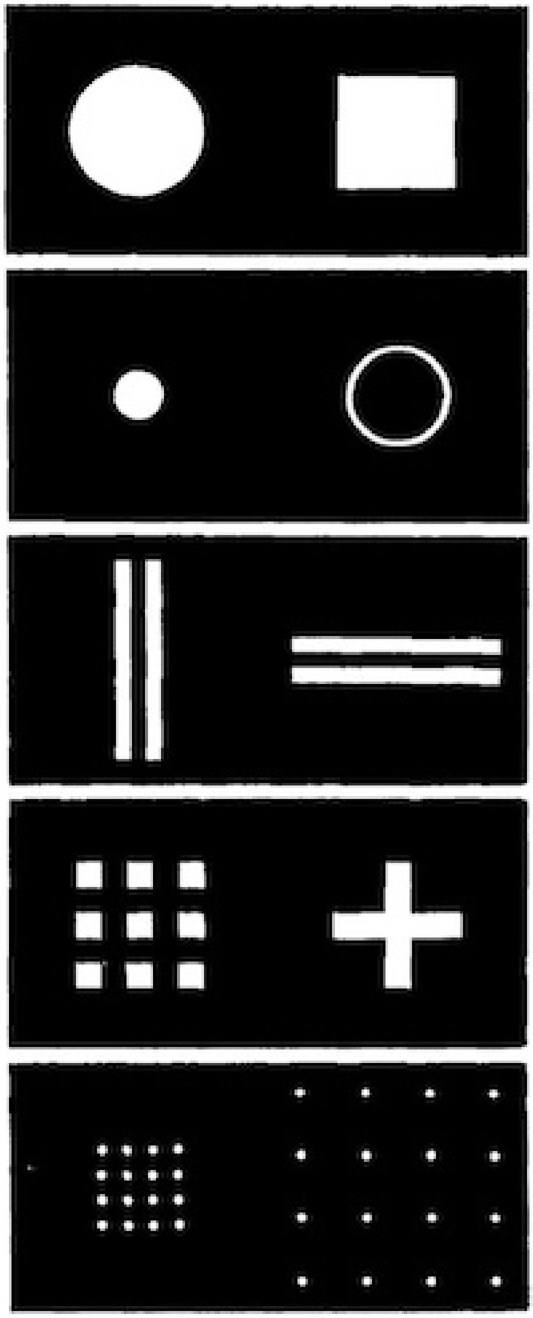
RONALD E. MYERS

University of Chicago

Journal of Comparative and Physiological Psychology, 1955

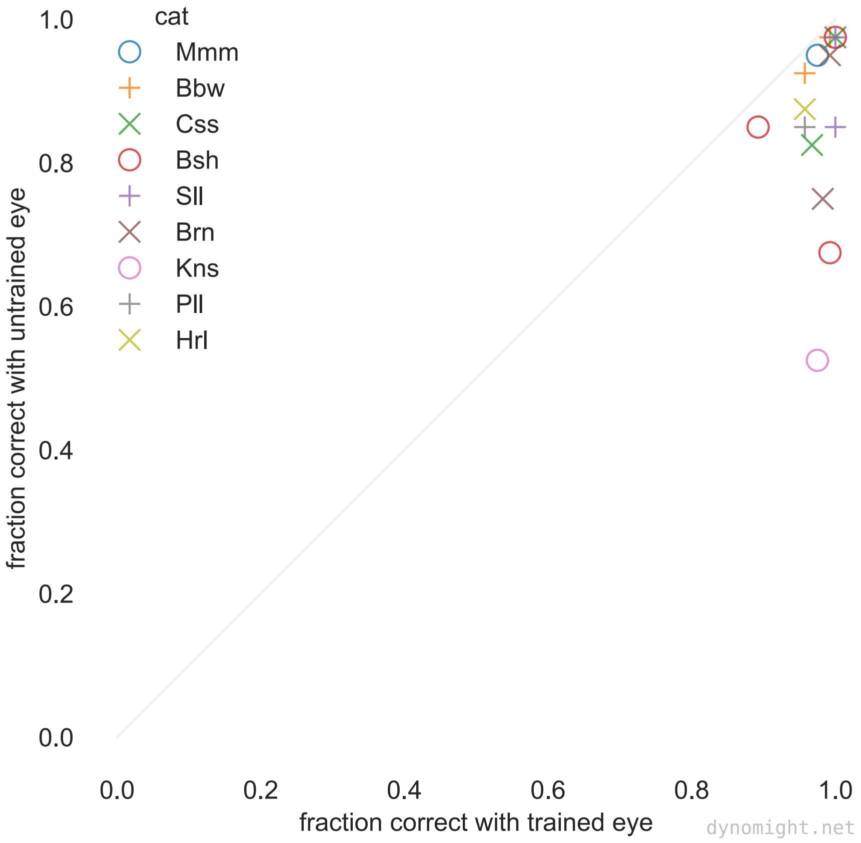
WHAT THEY WANTED TO TEST. Say you do surgery so one half of the brain only gets visual information from one eye. Then, you train cats to do some task using only one eye. Will they also be able to do it using the other eye, which never got any direct visual information about the task?

WHAT THEY DID. They took 9 cats, and cut the optic chiasm, so that light from one eye now only goes to one half of the cortex. They then covered one eye with a patch, and put two doors in front of the cat with different shapes on them, e.g., a square and a circle. One shape always led to food, the other to an electric buzzer.



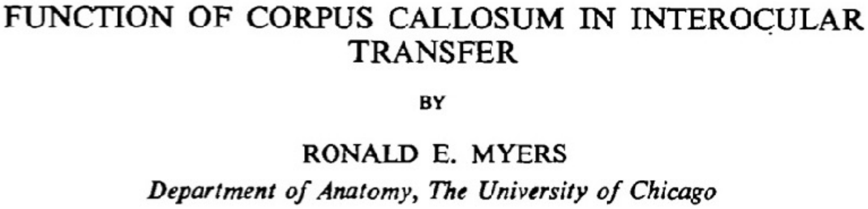
Every day, they repeated trials with the positions of the doors chosen randomly day until the cat was no longer hungry and stopped cooperating. They kept this up until the cat could get 34/40 trials correct, and then just did 40 trials per day for a while. Finally, they switched the patch to the other eye and tested how well the cat did. They repeated the whole process with different shapes.

WHAT THEY FOUND. Here's the fraction of trials the cats got correct in the last few training sessions (on the x-axis) and the testing session (on the y-axis). Each marker shows one cat (named "Mmm", "Bbw", etc.) on one pattern. The same marker is repeated if the same cat was tested on multiple patterns.



The cats did somewhat better with the trained eye, but they did pretty well with the untrained eye, too.

Myers 1956



FUNCTION OF CORPUS CALLOSUM IN INTEROCULAR TRANSFER

BY

RONALD E. MYERS

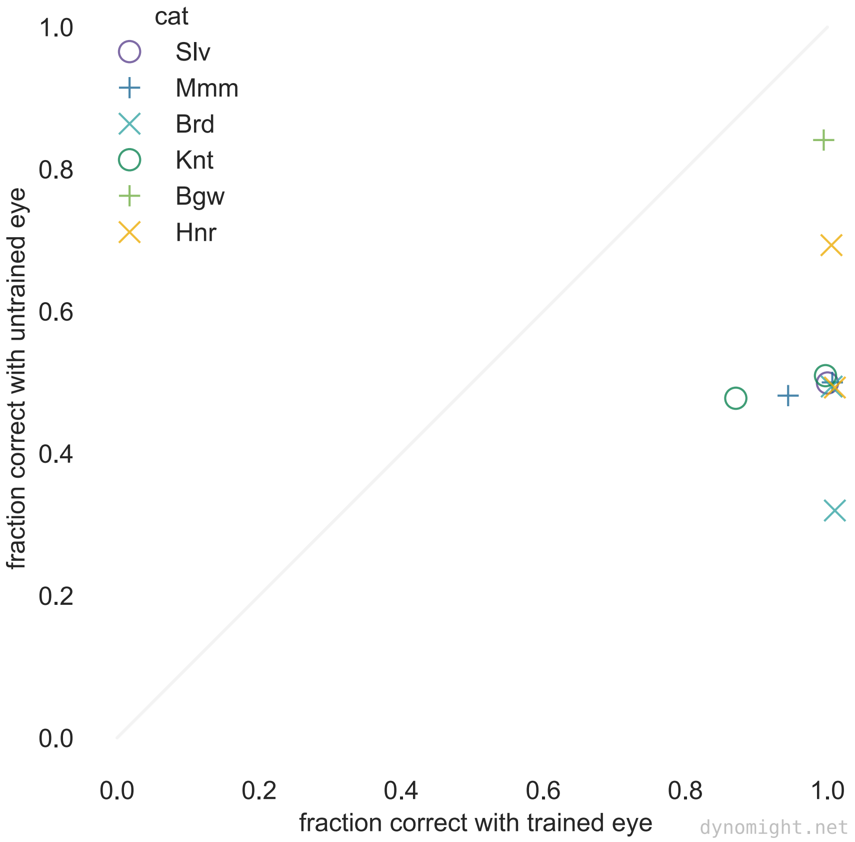
Department of Anatomy, The University of Chicago

Brain, 1956

WHAT THEY WANTED TO TEST. What if you did the previous experiment, except instead of just cutting the optic chiasm, you also cut the corpus callosum?

WHAT THEY DID. The previous experiment, except instead of just cutting the optic chiasm, they also cut the corpus callosum.

WHAT THEY FOUND First off, the cats were fine, with no obvious problems in coordination or anything else. (This is probably the most surprising result of all!) As for performance, the untrained eye did much worse than the trained eye, usually only getting 50% correct, the same as random guessing.

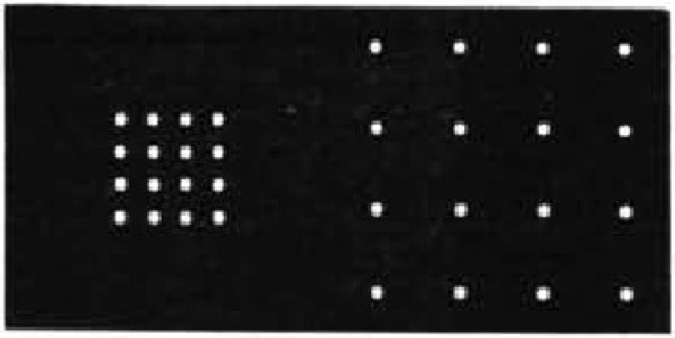


In two cases (Bgw and Hnr) the cats did generalize OK. Myers claims these aren't "true exceptions", though I'm a bit skeptical.

Bgw generalizes very well, and another, while Hnr that does OK on one pattern, but poorly on another.

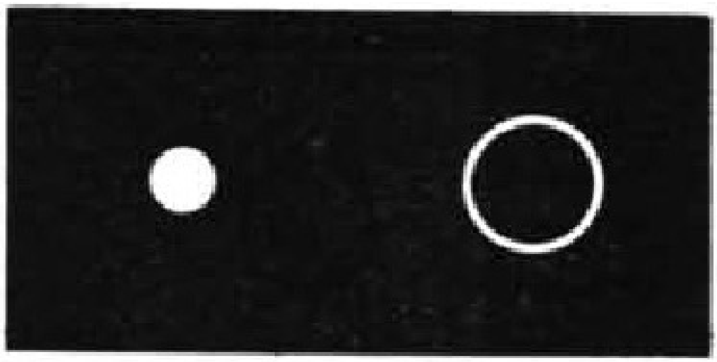
In this test Bgw had to deal with pattern V:

V



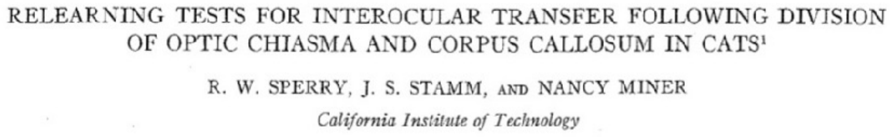
Apparently in some previous experiment (not int his paper), Bgw had already learned pattern II:

II



Myers states that he trained another cat on pattern II and it was able to generalize immediately to pattern V without any training. Thus, he thinks this wasn't a "true exception". OK, but did Bgw *also* generalize immediately, or was training for the regular eye needed? It's not stated. Also, why wasn't Bgw tested on a second pattern, almost all the other cats were? It's never explained. A cynic (not me! a lesser person!) might think that's out of a fear that Bgw would have been able to generalize again.

Sperry 1956



RELEARNING TESTS FOR INTEROCULAR TRANSFER FOLLOWING DIVISION OF OPTIC CHIASMA AND CORPUS CALLOSUM IN CATS[1]

R. W. SPERRY, J. S. STAMM, AND NANCY MINER

California Institute of Technology

Journal of Comparative and Physiological Psychology, 1956

WHAT THEY WANTED TO TEST. Say you again cut the optic chiasm and corpus callosum. We saw above that they usually do poorly with the test eye. But can they at least re-learn the pattern with the test eye *faster*? This would suggest that some information is still getting between the two halves of the cortex.

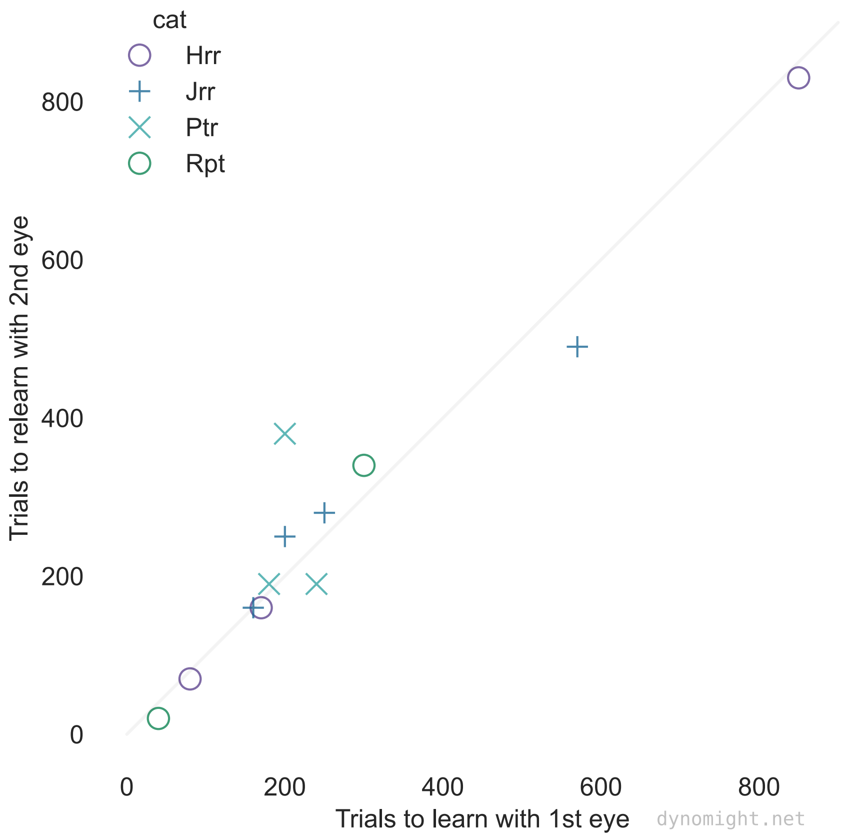
WHAT THEY DID. They took 4 cats and again cut the optic chiasm and the corpus collosum. They trained one eye of the cats to recognize the pattern, doing trials of 10 and stopping when a success condition was reached, meaning the cat was doing well. They then trained the other eye, and compared how long it took to reach the same condition.

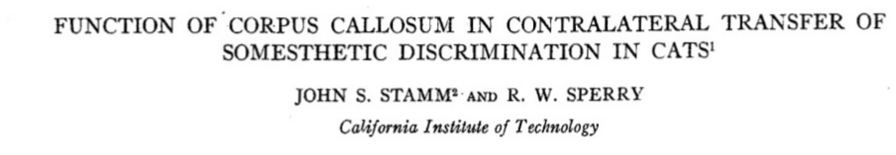
The exact condition for success is complicated.

They stop when they got at least 17/20 right in two trials but *also* have 18/30 right in the three trials before the last two. They do this “so that the performance was thereby maintained above the .01 probability level through 50 trials”, whatever that means.

There are also some details about how they did extra trials to balance the total number of trials with each eye. However, they show the total number of trials and they don’t look balanced. I can’t figure it out.

WHAT THEY FOUND. There’s not much speedup with the second eye. (They didn’t verify that there *was* a speedup with an intact corpus callosum, presumably because Myers 1955 already did that.)





FUNCTION OF CORPUS CALLOSUM IN CONTRALATERAL TRANSFER OF SOMESTHETIC DISCRIMINATION IN CATS[1]

JOHN S. STAMM[2] AND R. W. SPERRY

California Institute of Technology

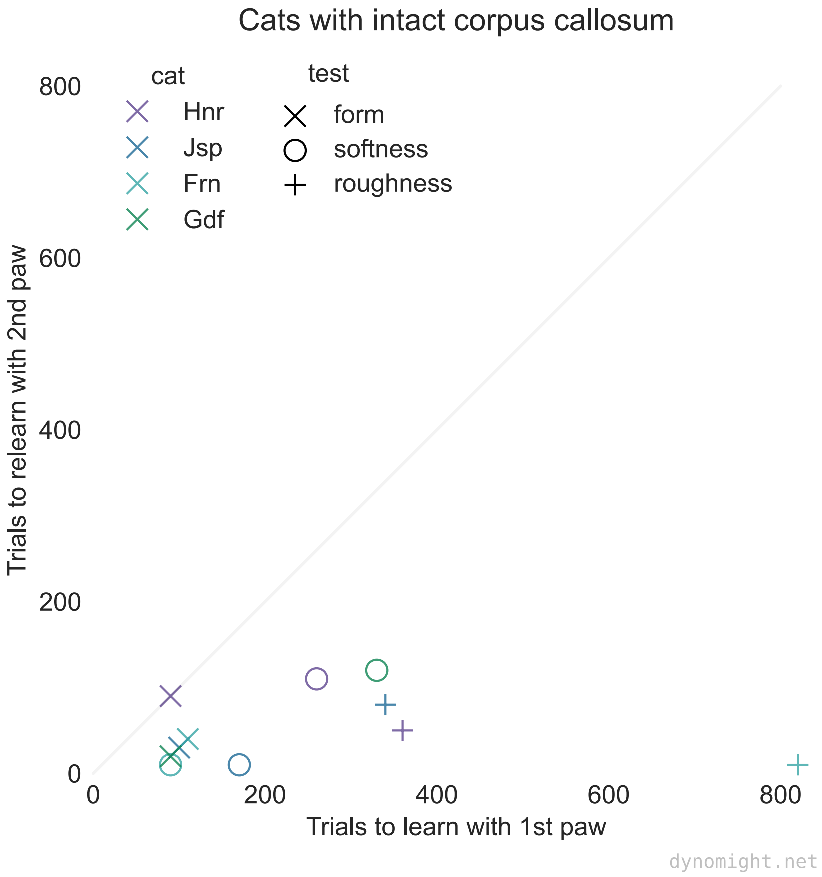
Journal of Comparative and Physiological Psychology, 1957

WHAT THEY WANTED TO TEST. Do similar results hold if you use touch rather than vision?

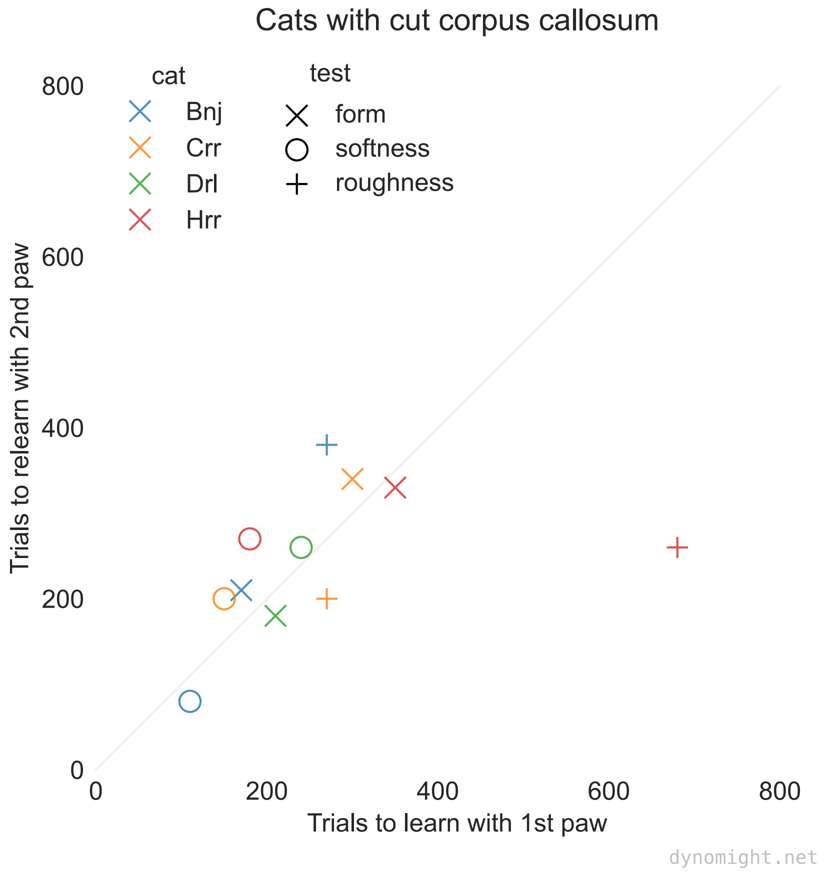
WHAT THEY DID. They took eight cats and cut the corpus callosum of four. One cat (Hrr) also had the optic chiasma cut in the above experiment, but presumably this doesn't matter since vision isn't being used. These cats were placed in a box where only one forepaw (right or left) could press two pedals. These varied in three ways:

- Form (wood triangular prism vs. a flat square wood)
- Softness (Two of a 2cm rubber pad, a 2cm wood pad, and a 1/2 cm rubber pad.)
- Roughness (A half-cylinder that is smooth wood vs. covered in sandpaper)

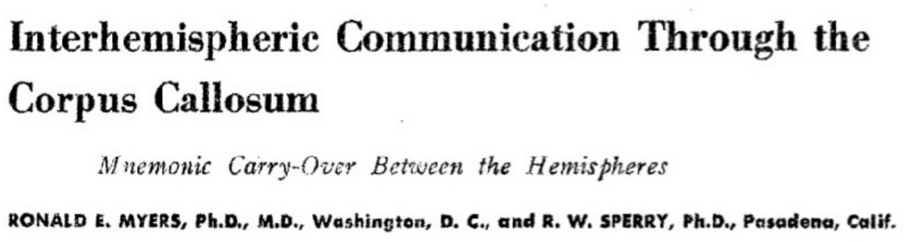
WHAT THEY FOUND. Cats with intact corpus callosum could re-learn the same task with the second paw more quickly. In the following plot, `x`, `o`, and `+` marks the different types of tests.



In cat with the corpus callosum cut, however, relearning with the second paw usually took almost as long as learning with the first paw.



Myers 1958



Interhemispheric Communication Through the Corpus Callosum

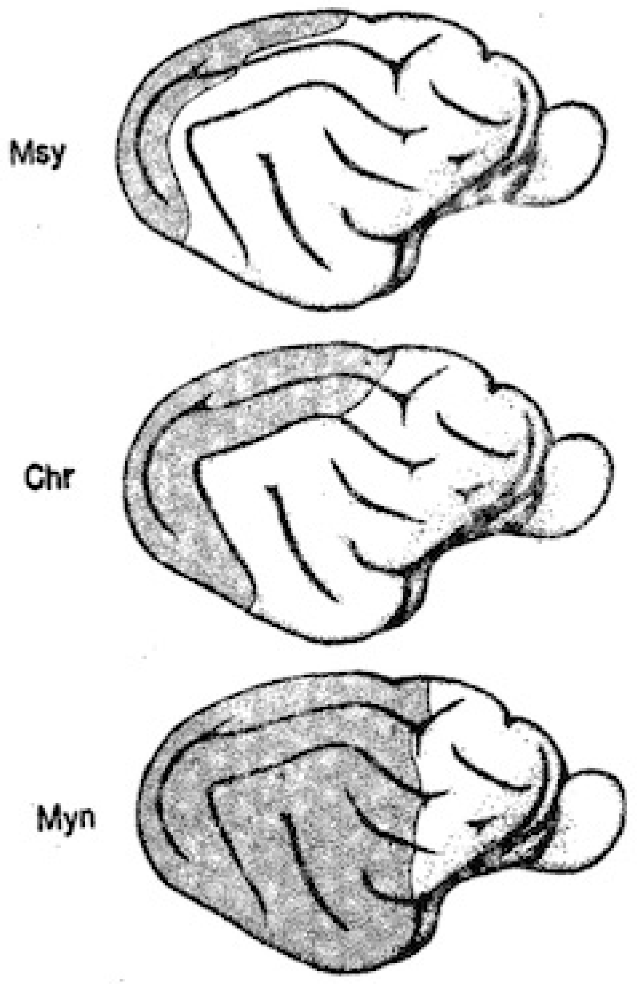
Mnemonic Carry-Over Between the Hemispheres

RONALD E. MYERS, Ph.D., M.D., Washington, D. C., and R. W. SPERRY, Ph.D., Pasadena, Calif.

AMA Archives of Neurology and Psychiatry, 1958

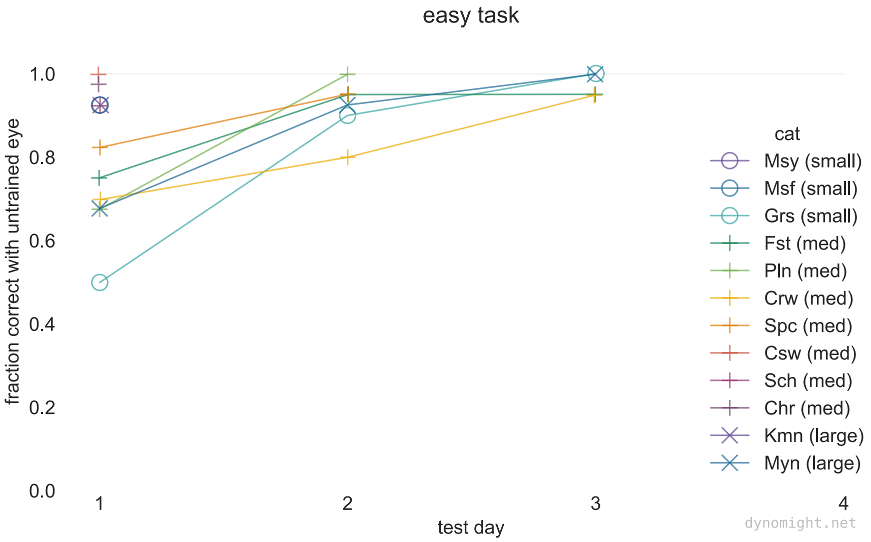
WHAT THEY WANTED TO TEST. Cutting the corpus callosum seems to prevent the other hemisphere from being able to perform the task the other eye was trained on. But is that because the two hemispheres ordinarily make two "copies" of the learned information, or is there one copy with the information shared during testing? Suppose you train a cat while only one hemisphere of the cortex gets any visual information. Then you surgically remove much of the trained hemisphere. Will the other one be able to do anything?

WHAT THEY DID. They took 14 cats and cut the optic chiasm. (They did *not* cut the corpus callosum.) They then trained them to recognize visual patterns with one eye covered, similarly to Myers 1955. Once they reached a performance of 34/40 trials correct, they got 10 more days of 40 trials. Then, they removed various amounts of the "trained" cortex corresponding to the eye that could see. They did all the tests on two tasks, one "easy" and one "hard".

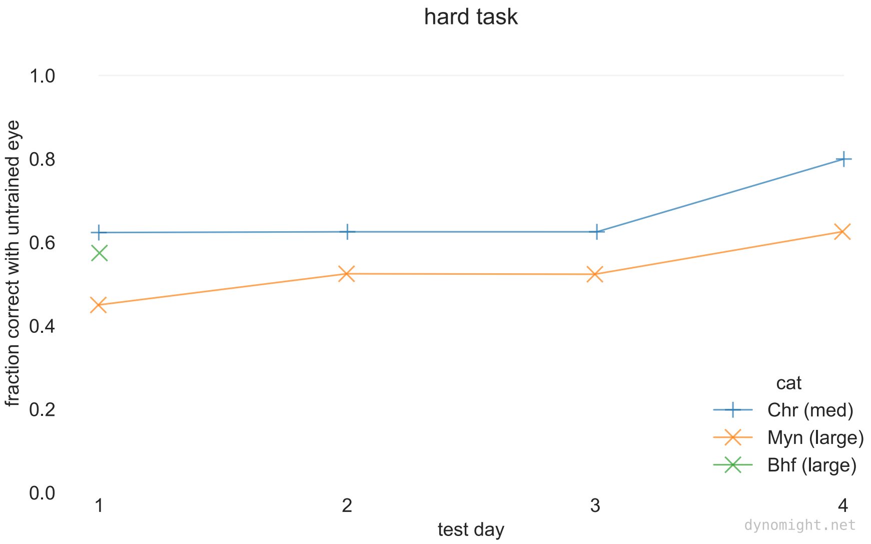


After 12 days to recover, they did 40 new tests per day, and recorded how the cats did.

WHAT THEY FOUND. On the easy task, the cats almost all did well within a few days, while on the hard task, there was little transfer. This plot shows the different cats with markers depending on how much cortex was removed.



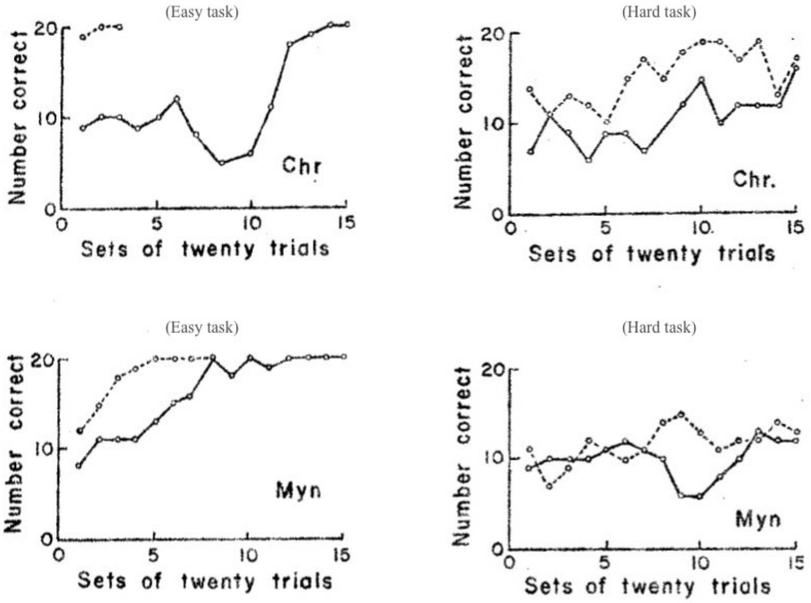
Here are the results on the hard task.



There are no numbers on how long it took to learn with the first eye but judging from Sperry 1956 it was presumably more than 1-3 days.

There are a few other details, including examples of the full training / retraining process for two cats.

Here the solid lines show the original training, the dotted line, the re-training process.



In terms of things I don't understand, there was one other cat (Nkw) who had a large amount of cortex removed, got the first 7 trials right, then 3 wrong, and then refused to participate any longer. This isn't show above.

I don't understand in what order Chr and Myn did the easy and hard tasks. I guess they were trained on both of the tasks before the surgery was done? Was one tested immediately, and the other after? It doesn't seem to be stated. I also don't understand why there were so few trials on the hard task.

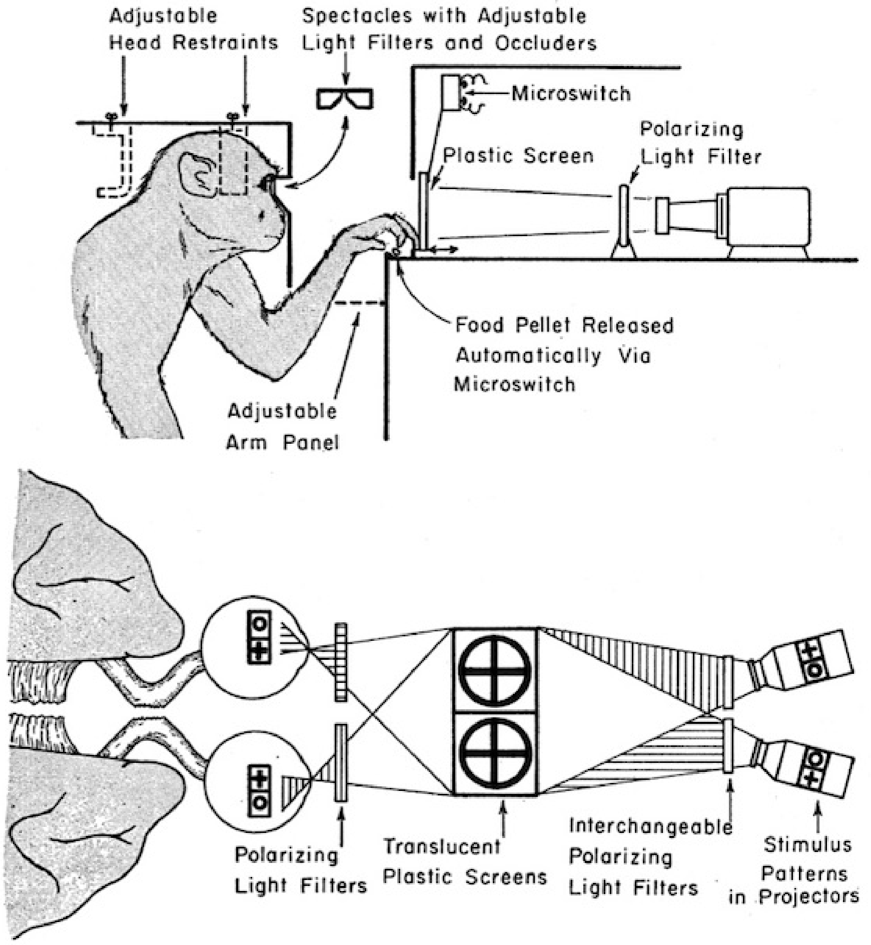
Trevarthen 1962

Double Visual Learning in Split-Brain Monkeys

Science, 1962. (Although the actual data below is from Trevarthen's PhD thesis.)

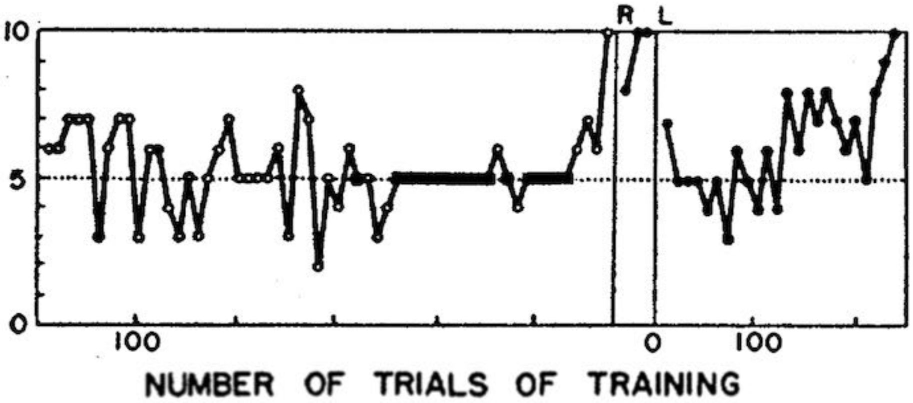
WHAT HE WANTED TO TEST. The previous experiments train one hemisphere and then test on the other. What if we take a monkey, sever the hemispheres as much as possible, and then show the two hemispheres contradictory things *simultaneously*? Will these monkeys get less confused than monkeys with intact brains?

WHAT HE DID. He took two monkeys, and cut the optic chiasm, the corpus callosum, and the anterior and hippocampal commissures. Then, he used polarized glasses to show opposite patterns to the two eyes. For example, the left eye might see +o while the right eye sees o+. The order of the patterns is randomly altered. There were 14 such patterns.

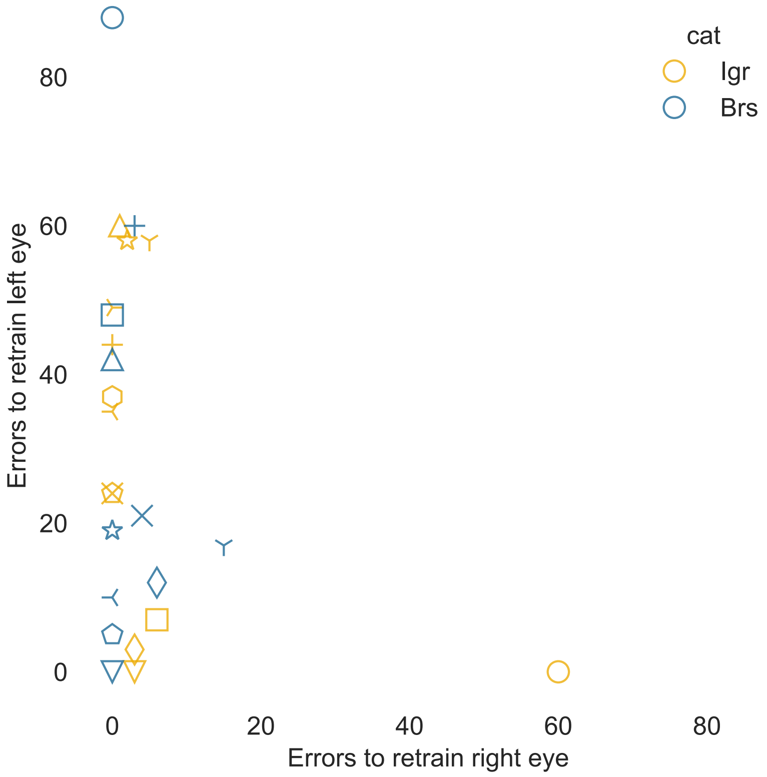

Adjustable Head Restraints
Spectacles with Adjustable Light Filters and Occluders
Microswitch
Plastic Screen
Polarizing Light Filter
Food Pellet Released Automatically Via Microswitch
Adjustable Arm Panel
Polarizing Light Filters
Translucent Plastic Screens
Interchangeable Polarizing Light Filters
Stimulus Patterns in Projectors

There were two screens in front of the monkey. They would get a morsel of food if they pressed the correct one. For example, the monkey might always need to choose the side that's + for the left eye and o for the right eye. They repeated these trials in groups of 10 until the monkeys either got 10/10 right or 9/10 twice in a row. Then they switched to doing trials with just one eye, repeated it until the monkey learned it, and finally the other eye.

WHAT HE FOUND. The monkeys could always do well with one eye (almost always the right) but sometimes did well with both eyes. Here's a typical example of the full training process. The main part shows the training period, while `R` and `L` show the results of re-training each eye.



Here are the full results of the number of trials needed to re-train with each eye. The different markers show the 14 different shape patterns that were used.



They tried various patterns and couldn't easily figure out which would transfer well to both and which wouldn't. They mention that the two monkeys were similar in which patterns this happened for, though. For tasks involving color or brightness differences, things tended not to transfer to both eyes.

He took two extra monkeys and cut the posterior commissure, habenular commissure, and rostral two-thirds of the quadrigeminal plate. (I think in addition to the other stuff) In these monkeys there didn't seem to be the "interference" in color discrimination, though there still was in brightness.

Discussion

Obviously, it's strange that these animals appear to behave normally, even when these tests show they don't always behave like a single unified system. We even know that humans after similar procedures will *explicitly tell you* that everything feels totally normal. Still, the whole goal of this article is to avoid injecting personal narrative about what it all means, so I'll shut up here. It's much more fun to DIY your own philosophical rambling anyway.

Does the gender-equality paradox actually exist?

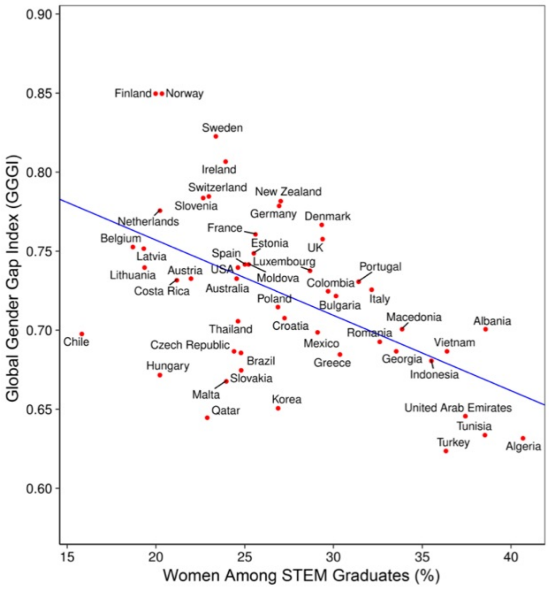
The gender-equality paradox is the (disputed) idea that countries with more gender equality have fewer women in STEM careers. While there's lots of debate in the scientific literature about the *causal implications* of this paradox, there's no agreement about a more basic question: Does the paradox even *exist*, or is it just an illusion caused by a contrived data analysis?

- The debate so far
  - Act I
  - Act II
  - Act III
  - Act IV
- New Analysis
  - Paradox dissolved?
  - A bunch of analyses
  - A different calculation for STEM-participation
  - Other measures of equality
  - Against BIGI
  - Other measures of women in STEM
- Takeaways

The debate so far

Act I

In 2018, Stoet and Geary had one of the most surprising results in social science in a decade. They took the Global Gender Gap Index (GGGI), which measures gender equality, and plotted it against the percentage of women among STEM graduates.



Finland has high equality but few women in STEM, while Algeria is the opposite. That's the trend.

Why this would be true is unclear, but the result seems hard to dispute. It's obvious that GGGI is measuring *something*, just look at the countries that are high or low on the graph. And you don't need to trust any fancy statistics, you can see the trend in the data.

This was picked up by The Atlantic, The American Enterprise Institute, Ars Technica, MacLean's, and Jordan Peterson. Stoet and Geary themselves published an article at Quillette, where they suggest their graph is partly due to different levels of interest in STEM and partly to comparative advantage—in places like Finland, girls perform similarly to boys in science but much better in reading.

Wait, did I just say this was hard to dispute?

Act II

Suspicious of these results, Richardson and colleagues took the same data, calculated the percentage of women among STEM graduates, and got… completely different numbers. They—I think—contacted the journal, which led to a corrigendum from Stoet and Geary in late 2019. This clarified what's on the x-axis in the above graph:

> The propensity of women to graduate with STEM degrees was $a/(a + b)$, where a is the percentage of women who graduate with STEM degrees (relative to all women graduating) and b is the percentage

of men who graduate with STEM degrees (relative to all men graduating).

Get that? Take a country with the following graduates each year:

	STEM degrees	All degrees
Men	100	1000
Women	5	50

Women make up 4.8% (5/105) of STEM graduates. However, their formula gives 50%, since the fraction of women who do STEM is the same as the fraction of men who do STEM. That is, a=5/50 is equal to b=100/1000.

There's a good argument for this. The most salient fact about the above country isn't anything STEM-specific, it's just that few women get degrees. Stoet and Geary's formula is invariant to this kind of imbalance.

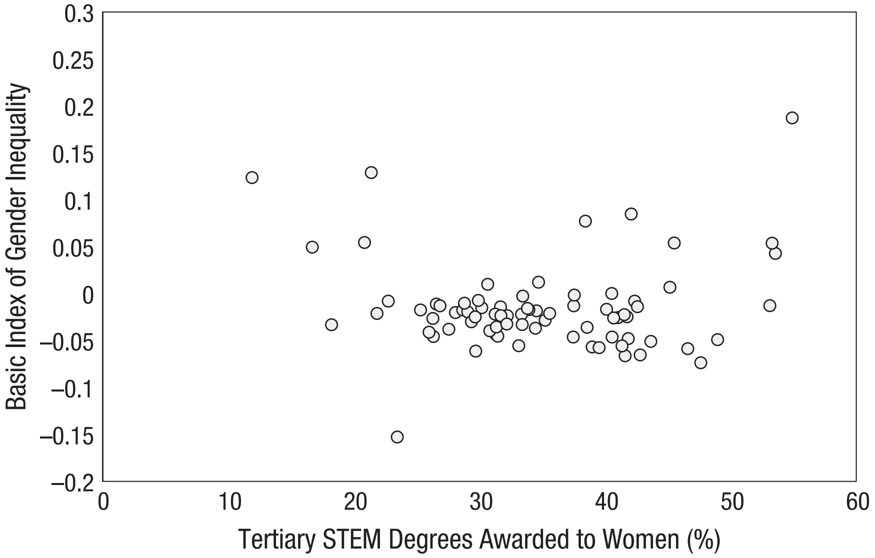
There's also a good argument against this formula. Maybe you think that imbalances in the total number of degrees are important, and you don't *want* to be invariant to them.

What there's *not* a good argument for is calling this quantity "Women Among STEM Graduates (%)" like the above graph does. In their corrigendum, Stoet and Geary don't really explain how this happened. In fact, they don't change much about their paper at all, other than adding above quote and inserting "propensity" everywhere.

Act III

Simultaneously with Stoet and Geary's corrigendum in 2019, Richardson and colleagues published a commentary on the corrected paper. They argue:

1. Propensities are bad.
2. It's not cool to use GGGI because it "measures achieved outcomes, not propensities" and "is not intended to be used to causally explain outcomes".
3. Better than GGGI is the ultra-simple Basic Indicator of Gender Inequality (BIGI). Stoet and Geary shouldn't object to this, since it was proposed by… Stoet and Geary.
4. If they compute the *actual* percentage of STEM degrees earned by women and plot it against BIGI, they get this graph, along with a non-significant regression coefficient.



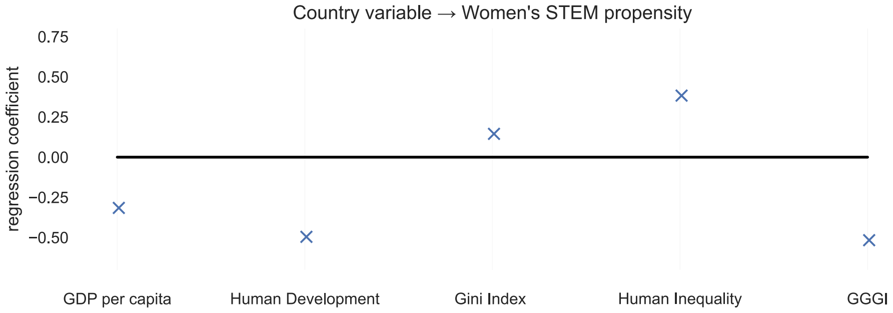
They also published articles in Slate and on their blog. This was picked up by Buzzfeed and The Scientist.

Act IV

In 2020, Breda and colleagues published a paper, part of this uses the same propensities as Stoet and Geary. They argue this is worthwhile both because the original result is well-known and because it's nice to be invariant to imbalances in the overall number of degrees.

Their first observation is that the propensities aren't just correlated with GGGI. They are also correlated with:

They do a regression to predict propensities from each of these variables (one variable at a time) and get these coefficients (from Table S5):



Coefficients from "Gender stereotypes can explain the gender-equality paradox", Breda et al, 2014. (Table S5) dynomight.net

Everything “good” is associated with lower propensities, be it more GDP, more development, less income/human inequality, or more gender equality.

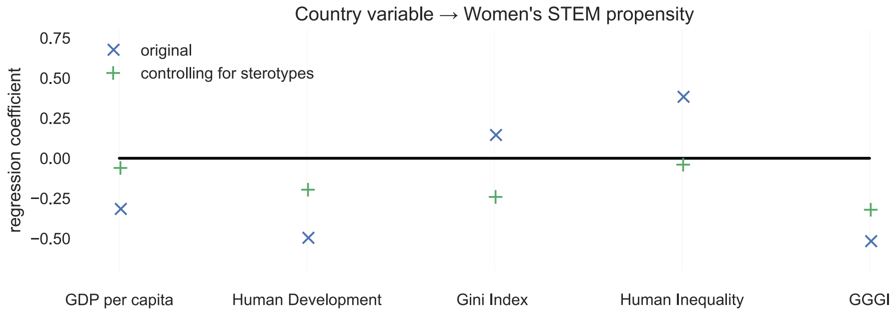
Their goal was to test how all this relates to gender stereotypes. They took the PISA 2012 data, and looked at how boys and girls felt about these two statements. These were chosen because they don’t directly mention gender, reducing the risk of social desirability bias.

> “Whether or not I do well in mathematics is completely up to me.”

> “My parents believe it’s important for me to study mathematics.”

Their *stereotype score* for each country reflects how much boys vs. girls agree with the above statements. If a boy of equal math ability is more likely to agree than a girl, the stereotype score is positive. If a girl is more likely to agree, the stereotype score is negative.

Their main result is a second regression to predict STEM propensities, now controlling for the stereotype scores in each country:



Coefficients from "Gender stereotypes can explain the gender-equality paradox", Breda et al, 2014. (Table S5) dynomight.net

Knowing stereotypes makes the other variables less predictive, dramatically so in some cases (Human Inequality) less so for others (GGGI).

This paper is often summarized (e.g. on Wikipedia) with quotes like this (emphasis mine):

> The stereotype associating math to men is stronger in more egalitarian and developed countries. It is also strongly associated with various measures of female underrepresentation in math-intensive fields and can therefore *entirely explain* the gender-equality paradox.

New Analysis

Paradox dissolved?

After first reading these follow-up papers, I had the impression the original study was debunked. But notice three things:

First, **causality isn't everything.** Richardson et al. think that BIGI is better than GGGI for establishing causality. I don't understand their reasoning in the slightest, but it doesn't matter. *None* of these analyses prove causality.

Still, does the paradox actually exist? It can't simultaneously be *false* (as Richardson et al. seem to claim) and *true but explained by gender stereotypes* (as Breda et al. claim.) Which is it?

Second, **stereotypes don't solve the paradox.** How could they, when the reduction for the GGGI coefficient above is so modest? I think the Wikipedia quote is misleading: Most of Breda et al.'s paper is about predicting other things, e.g. the *intention* to study STEM, where controlling for stereotypes has a stronger effect.

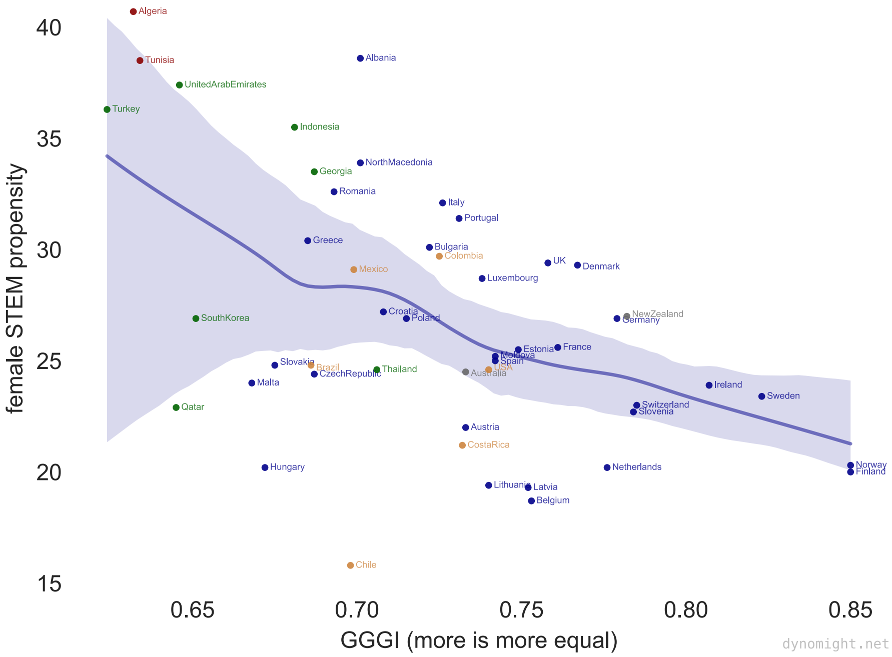
But OK, suppose that the paradox was entirely explained by gender stereotypes. That would just mean we've traded the mystery of why more gender-equal countries have fewer women in STEM for the mystery of why more gender-equal countries would have stronger stereotypes. That is still very paradoxical!

Third, **it's unclear how fragile the result is.** Richardson et al. say that the paradox only appears because of "contrived measures and selective data". Of course, if the paradox only appears after torturing the data in one particular way, then we shouldn't trust it. But their evidence is what happened when they tortured the data in one *other* particular way.

Shouldn't we try a *bunch* of analyses, and just check how robust things are?

A bunch of analyses

Let's start with the original analysis, relating GGGI to propensities. (Click to zoom in and look at the country names.)



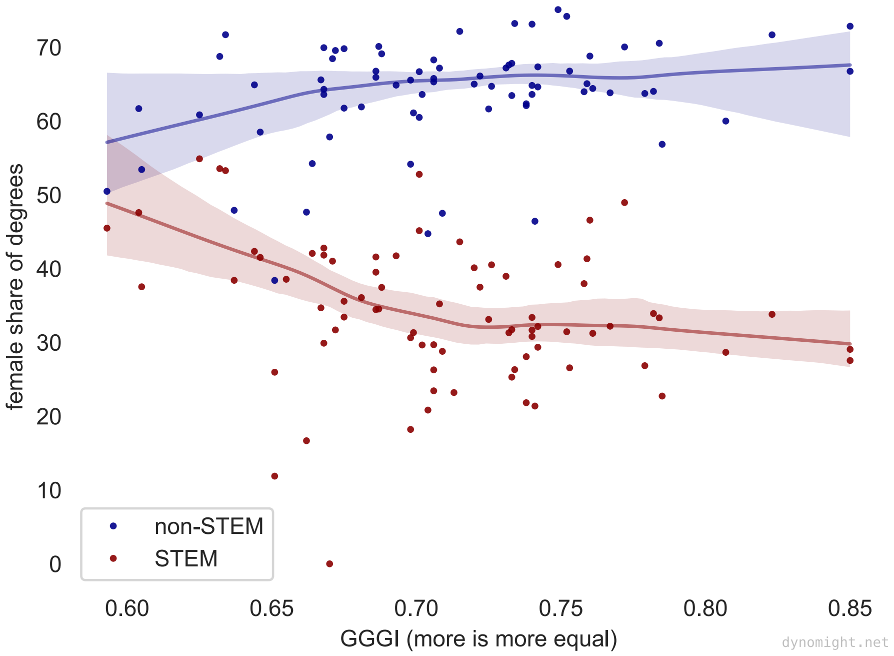
This is the same as the original Stoet and Geary figure, with three small changes:

1. Switch the axes.
2. Color countries according to their continent.
3. Show a LOWESS smoothing (linearity is for wimps) along with a 95% confidence interval, computed using bootstrapping.

A different calculation for STEM-participation

The above figure uses propensities, which is a major point of contention. Personally, I think this debate is silly. Propensities give one view of the data, while the raw fraction of women in STEM gives another. They both have value.

So, what if Stoet and Geary had just switched to using the *actual* percentage of women among people who earn STEM degrees, as Richardson et al. suggest they should have? They'd have gotten the following curves. (I added non-STEM degrees for context.)

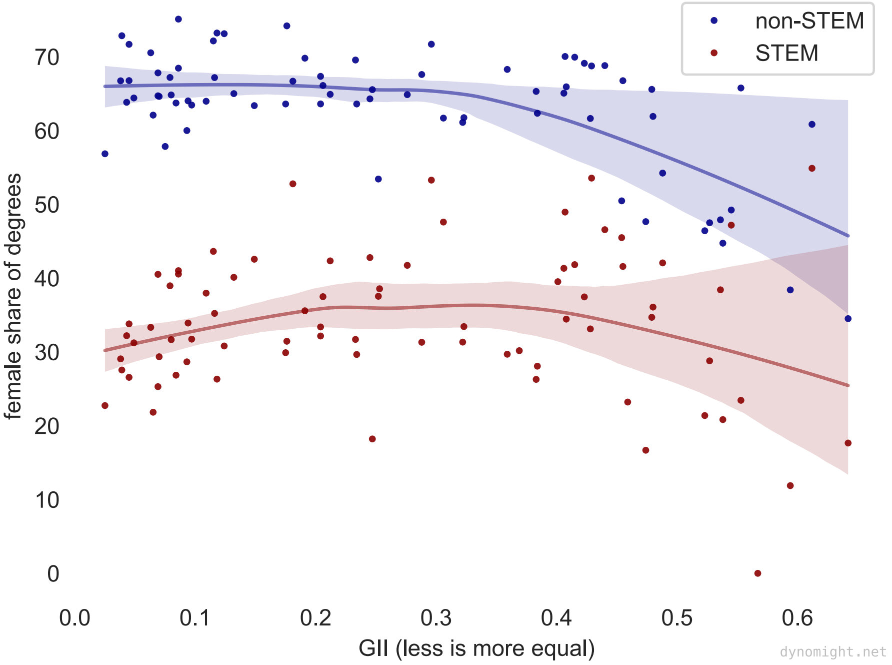


In more-equal countries, women earn a larger share of non-STEM degrees, but a smaller share of STEM degrees. The paradox is still there.

Other measures of equality

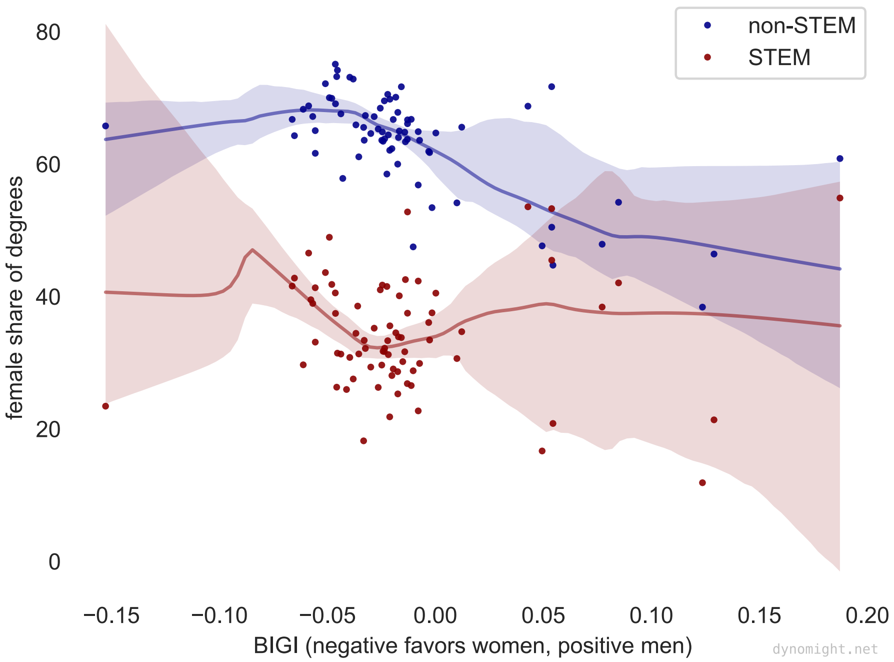
Maybe this all depends on some weirdness with how GGGI measures equality? A newer alternative is the Gender Inequality Index (GII). I took the 2019 rankings and used them instead of GGGI.

Be careful interpreting this graph: While more equality meant *more* GGGI, it means *less* GII.



Again, the most gender-equal countries have a smaller fraction of women in STEM, but not non-STEM. If you use propensities instead of the female share of degrees, the effect is even stronger.

A third alternative is BIGI, as suggested by Richardson et al. Be very careful here: BIGI is negative when women are favored and positive when men are favored. Equality occurs around zero.

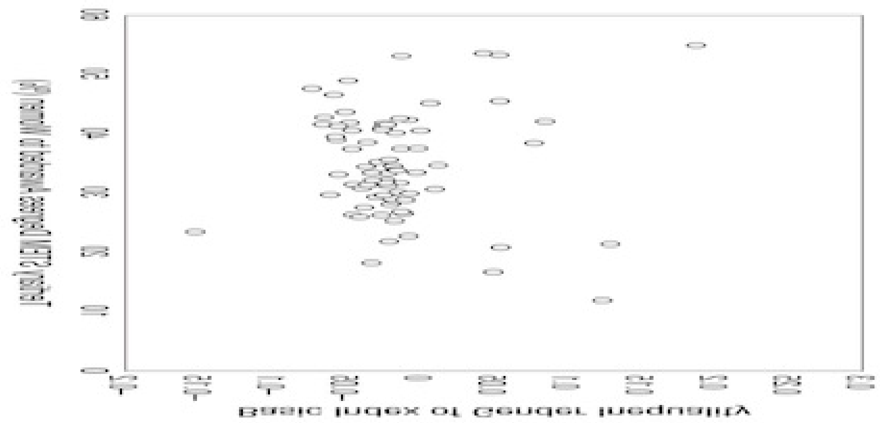


For non-STEM degrees, the trend is simple—the more women are favored, the more degrees they earn. But for STEM degrees, there's a U-shaped curve where women earn the smallest share around BIGI -.02, where women are *just slightly* favored. Comparing BIGI to propensities gives a stronger, but less symmetric, effect.

While we're on the subject… The red dots in the above graph show the same data as in Richardson et al.'s commentary above, which they used to claim that there was no gender-equality-paradox. (You can also see them by themselves with country labels.) What's going on?

Well, for one thing, I made the graph ~~better~~ differently, switching the axes and using smaller markers so you can see the density of countries.

Don't believe me? Here's what you get if you take their graph, rotate right by 90 degrees, flip the vertical axis, and change the aspect ratio:

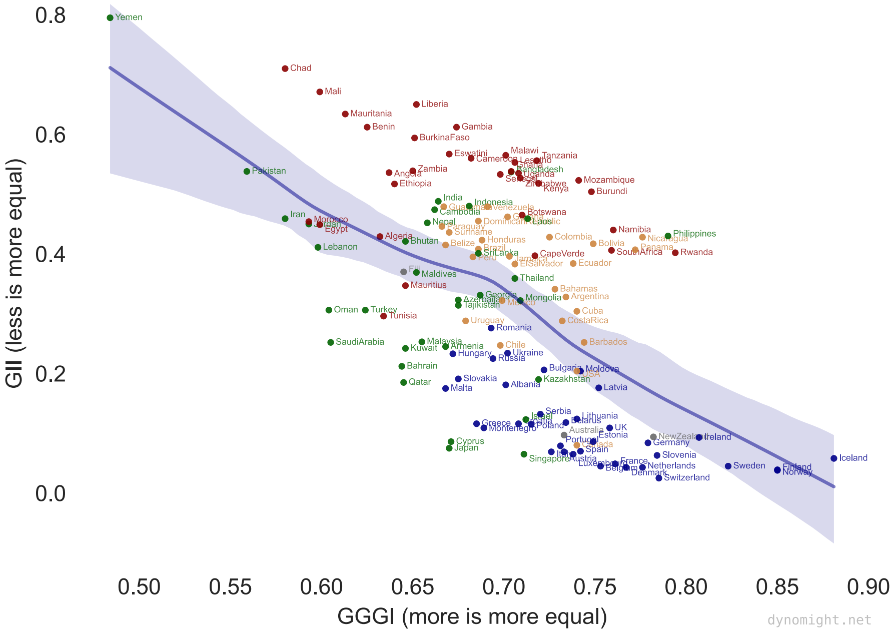


If you look carefully, you can see that these dots are the same as the red dots above.

For another thing, they did a *linear regression* and found no significant result. That's not too surprising, given that the effect above is nonlinear and symmetric.

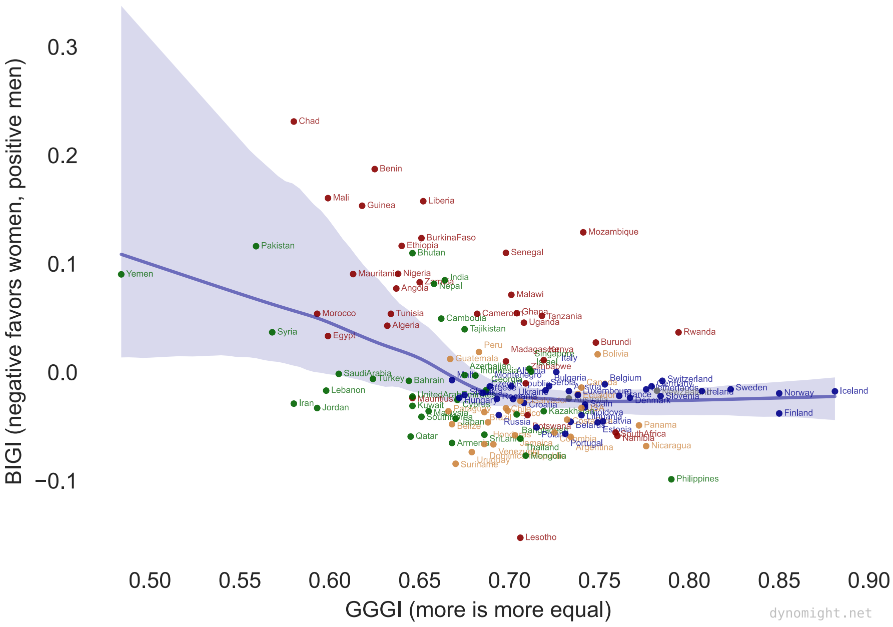
Against BIGI

I think BIGI is a terrible measure of gender-equality and we shouldn't be using it. For context here's a plot comparing the other two measures we've looked at, GGGI and GII:



Are the Philippines more gender-equal than Japan (as GGGI implies) or the opposite (as GII implies)? I don't know, but I'll accept that it depends on different, reasonable definitions of *gender-equal*.

On the other hand, here's a plot of GGGI against BIGI:



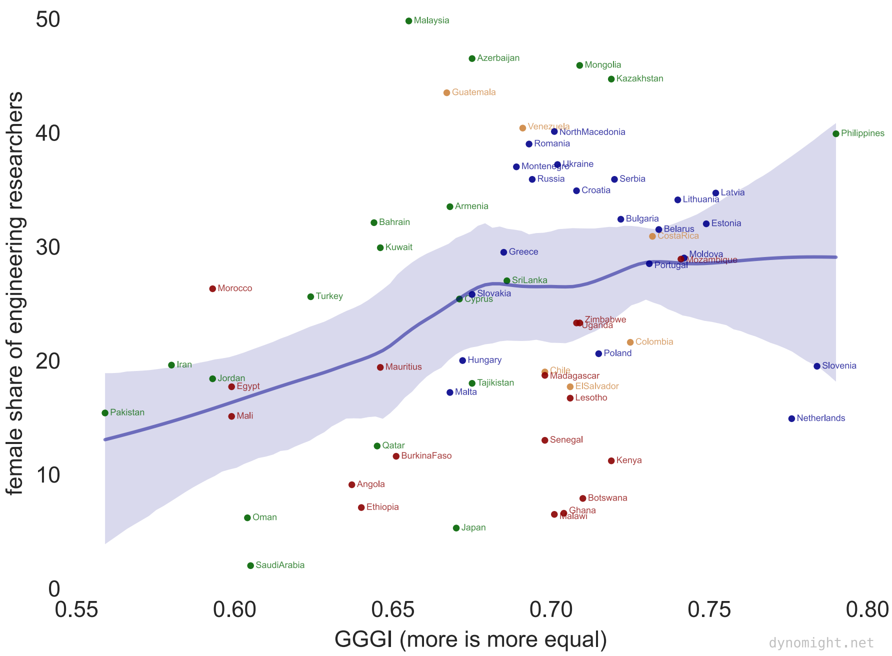
According to BIGI, Saudi Arabia—where women can only show their hands and eyes in public and must have a legal male guardian—is basically the same as Switzerland. And Lesotho—the tiny country inside South Africa—is by far the most women-favored place in the entire world. Ooohkaaay.

This isn't to say that BIGI is *bad* exactly. They specifically discuss Saudi Arabia in their paper. My point is that it doesn't capture what we have in mind in this context. At all. So while we *do* seem to get a paradox with BIGI, I think it's meaningless and we should forget about it.

Other measures of women in STEM

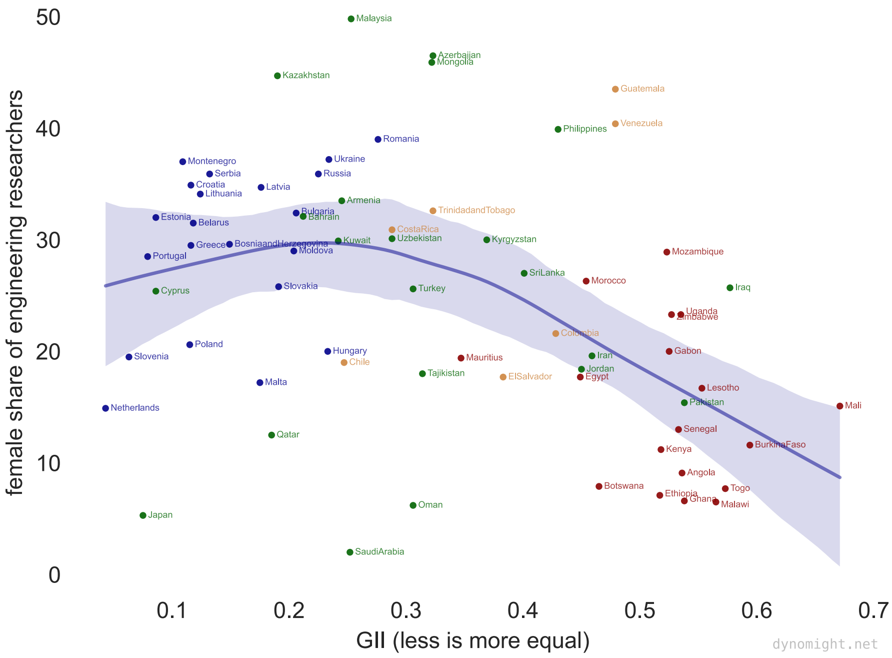
While the result seems robust to different measures of gender equality, everything above uses the same data from UNESCO on the number of STEM graduates. We've analyzed it both in terms of propensities and raw fractions, and the result is still robust. Still, what if we use a different data source entirely to measure STEM participation?

For variety, I looked at the female share of researchers in engineering and technology. If you compare this to GGGI, there's really no paradox at all. At most, there's a bit of a "leveling off".



If you look at natural science researchers instead of engineering, you again see no paradox.

On the other hand, if you use GII instead of GGGI, you do see a small effect in the most gender-equal countries:



Comparing GII to the natural sciences shows more of a leveling off than a full reversal.

I'm not sure if all these observations constitute a "paradox" exactly, but they aren't something I would have predicted.

Takeaways

So, is there a gender-equality paradox? Three points.

First, Stoet and Geary's original paradox is **robust.** It doesn't matter how you measure gender inequality and or if you use propensities or raw fractions to measure women's fraction of STEM degrees. It's not fair to imply that they cherry-picked the details of their analysis to support some pre-determined conclusion.

Second, the paradox is **somewhat limited.** It appears with STEM degrees no matter how you define "equality", or how you torture the data. For STEM researchers, the effect is more modest and only appears for certain definitions of gender equality. This is weird, and I don't understand it other than that it suggests we need more nuance than "more gender equality → fewer women in STEM".

Third, **resist simplistic causal explanations!** People choose degrees for lots of reasons: Economics, working conditions, family influences, cultural/media influences, intrinsic interest, and simply what degree programs are accessible. Most of these operate in feedback loops with each other. My love for scatterplots is vaster than the seas, but they're at most *vaguely suggestive* of any single cause.

Plot all the plots

Lest I be accused of cherry-picking, here's *all* the different ways of measuring gender inequality against *all* the ways of measuring women's participation in STEM. I also threw in per-capita GDP and Breda et al.'s stereotype measurements. (For GDP I removed Qatar and the top 10 tax havens where GDP is meaningless.)

	GGGI	GII	BIGI	GDP	stereotypes
STEM propensity	x	x	x	x	x
STEM degrees	x	x	x	x	x
non-STEM degrees	x	x	x	x	x
Engineering researchers	x	x	x	x	x
Natural science researchers	x	x	x	x	x

Choose the column you want on the x-axis, the row you want on the y-axis, and let the beautiful dots wash over you.

Data sources

- GGGI: Wikipedia (2015 rankings)
- GII: Wikipedia (2019 rankings)
- BIGI: genderinequality.info
- Women's share of STEM / non-STEM: The actual UNESCO data used for the share of STEM degrees going to women appears to no longer be on their website. With much gnashing of teeth, I was able to find an older version on archive.org.
- Propensities: Due to the same problem, I couldn't find the raw data for propensities. Instead, I took these from Stoet and Geary's supplementary material.
- Women's share of engineering / natural science researchers: UNESCO report
- GDP: The IMF's 2021 estimates in purchasing power parity, via Wikipedia.
- Stereotypes: Breda et al.'s supplementary material.

It's perfectly valid for a trait to be more than 100% heritable

All psychological traits are heritable. This is best replicated finding in all of behavioral genetics. Some recent numbers include:

But what, exactly, does “heritability” mean?

I used to have a mental model something like this: Each person has some number of religiosity points that come from genes and some number that come from the environment. If religiosity was 40% genetic, I pictured this:

```
Genes                  (3/4)
      Environment           (2/6)
      Total               (5/10)
```

The problem with this picture—aside from being completely wrong—is that it suggests heritability is an immutable constant, like the number of chromosomes in a cell or the fine structure constant.

So what *is* heritability? It’s the ratio of the *genetic variance* of a trait with the *total variance*, including all causes. Since the environment is always changing, so is heritability.

Let’s explore this definition. We’ll see that it leads to several puzzles:

- Why is heritability often higher for traits that seem less important? Why, for example, is pig back-fat thickness 14× more heritable than pig litter size?
- How, even when there are large environmental effects, can a trait still be 100% heritable?
- How, when there are correlations between genes and the environment, can traits can be *more than* 100% heritable?

The only math we’ll use is the concept of variance. If you’re not familiar with that, just think of it as “how variable” something is. Humans have high variance in how much we like folk music, but low variance in our number of fingers.

One note on terminology: Biologists use “phenotype” to refer to what actually happens, including all genetic or environmental causes. The phenotypic length of your foot is, thrillingly, the actual length of your foot.

Contents

Simplest example

Say that people's height depends on only one gene, which can be either **short** or **tall**. The genetic contribution to height is 2m if they get the short gene, and 4m if they get the tall gene. (The math is easier if we make everyone giants.)

genes	Genetic value
short	2
tall	4

The only thing that matters in the environment is food. At birth, people are assigned either to **diet** or to **feast**. If on a diet, they lose 1m of height compared to their genetic baseline. If on a feast, they gain 1m of height.

env	Environmental value
diet	-1
feast	+1

A person's height is the sum of their genetic and environmental values. Using **P** for phenotypic, write the final height as

$$\mathbf{P(genes,env) = G(genes) + E(env).}$$

With all these assumptions, can we calculate heritability? Not yet! We need the **population distribution**, i.e., what fraction of people have each combination of genes and environments. Assume for now that half of people have each gene, and half of people have each environment, and that these are independent.

population distribution	diet	feast
short	¼	¼
tall	¼	¼

The definition of heritability

Given any trait with P=G+E, the (broad-sense) heritability is

Heritability is how much the *genetic* contribution to the trait varies as compared to how much that trait varies *overall*. These variances are computed with respect to random people from the population.

In the above example, the variance of E is one, since it is -1 half the time, and +1 the other half. Similarly, G is always one off from its mean of three, so the variance of G is also one. Since the genes and the environment are independent, the variance of their sum is **var(P)=var(G)+var(E)**. This tells us that

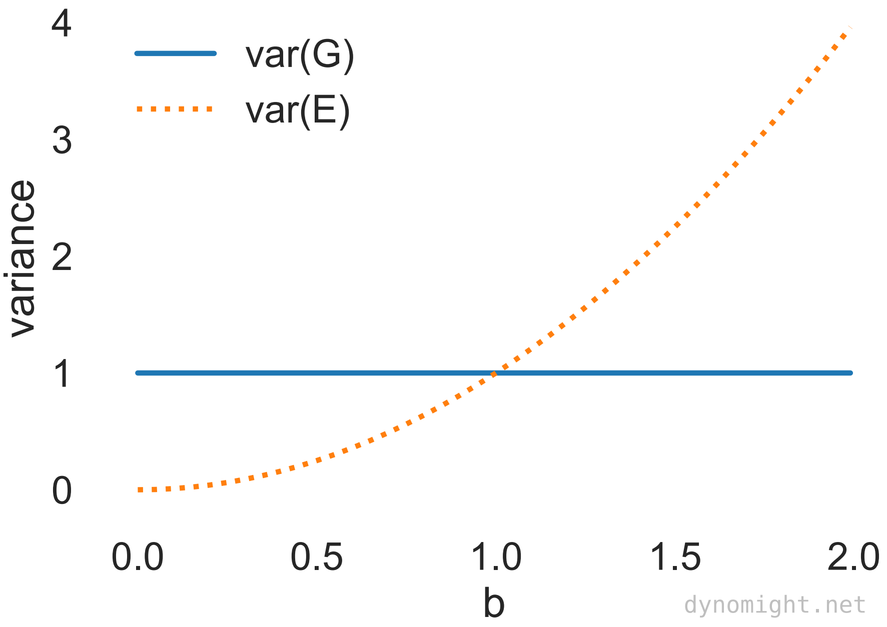
This is natural, since the effects of genes and environments are symmetric—changing either one changes height by 2m.

Let them eat more (or possibly less)

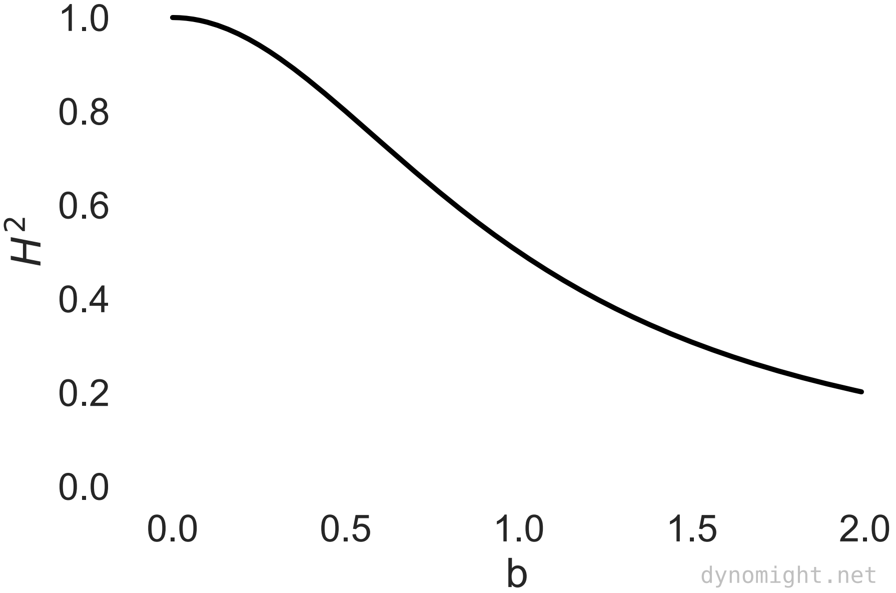
Let's change how much food people get. Suppose that there's some parameter **b** that reflects how much more food the feast group gets than the diet group.

env	Environmental value
diet	**-b**
feast	+**b**

As **b** changes, the environmental variance changes. If you feed the two groups almost equally (**b** is near zero) then the variance is near zero. As the difference between groups increases, variance increases.



Here's what the corresponding heritability looks like:



If the two groups are fed equally, then height is 100% heritable. If the feast group is fed much more (**b=2**), then height is only 20% heritable.

Takeaway: Heritability depends on how variable the environment is. If the environment becomes more variable, heritability decreases. If the environment becomes more consistent, heritability increases.

Example: Changing the heritability of IQ Research suggests that childhood lead exposure can decrease IQ by 5-10 points and heavy prenatal alcohol exposure can decrease IQ by 15-20 points. Current estimates are that intelligence is 40-80% heritable. But suppose society managed to eliminate lead and alcohol exposure. Environmental variance would decrease, and heritability would go up.

In theory, it's possible to push the heritability of *all* traits to 100%. Just make sure each person is exposed to *completely identical environments.* (You can argue that free will or quantum mechanics or the butterfly effect means there is no such thing as two identical environments, but let's not get distracted.)

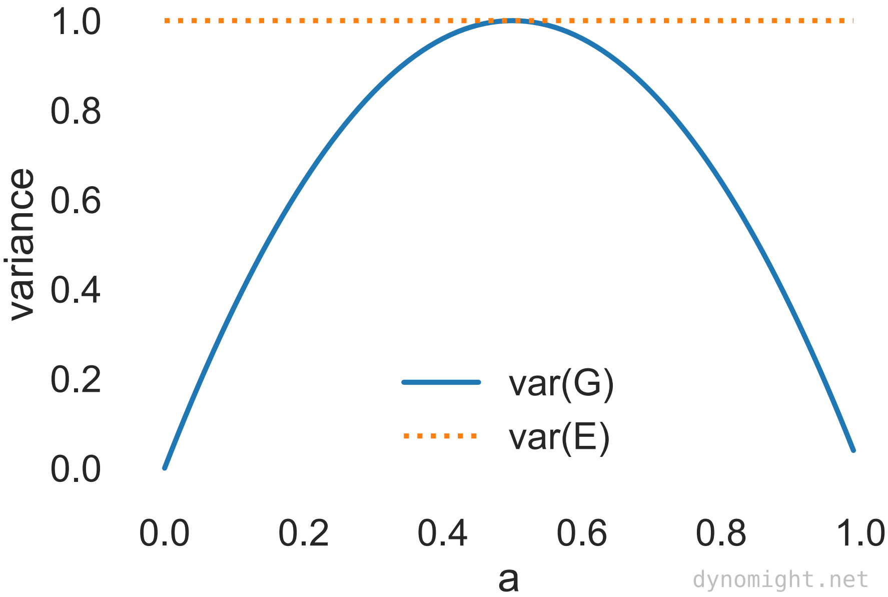
One could decrease the heritability by doing the opposite: Take half of kids and make it illegal to teach them to read or whatever. Then, take the other half of kids and give them private tutors. Environmental variance would increase, and heritability would go down.

The purge

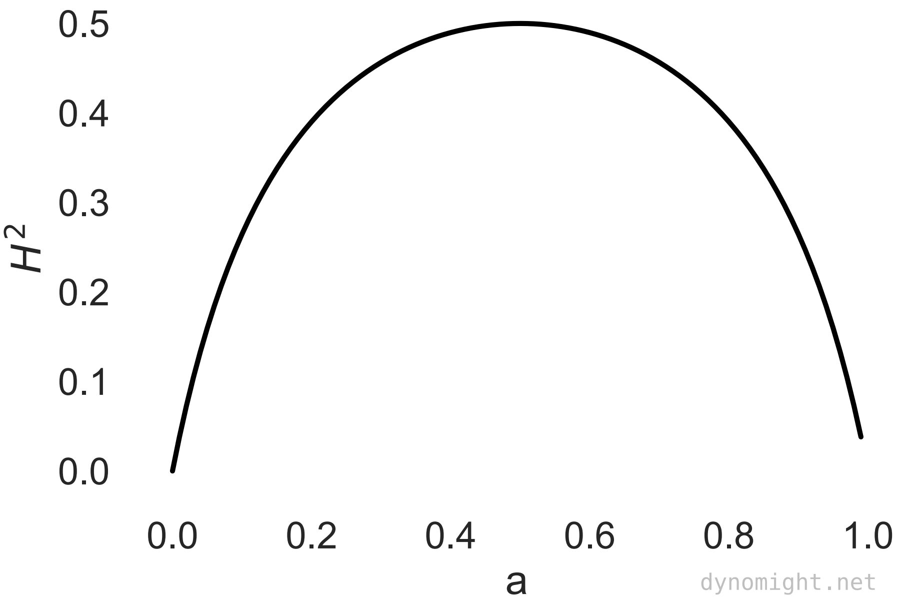
Now let's delete a bunch of people. After the purge, instead of half being short and half tall, some fraction **a** are short, and a fraction **(1-a)** are tall. If short people are more likely than tall people to survive the purge, then **a** > **½**. If they are less likely, then **a** < **½**.

Leave the food difference fixed at **b=1**. The environment doesn't change, so the variance of E remains one.

The genetic variance does change. If almost everyone has short genes (**a** is near zero), then almost everyone will have G=2, meaning the variance of G will be small. If **a** is ½, we get the previous model. A bit of math gives the following plot.



The heritability is still $\mathbf{H^2 = var(G) \;/\; (var(G){+}var(E))}$, which now looks like this:



Heritability decreases hugely if the population gets really imbalanced in either direction.

Takeaway: Heritability depends on how much genetic diversity there is.

Example: Hair in Japan How heritable is hair color in Japan, before anyone goes grey? Say 99% of people have genes for black hair, so **var(G)** is very low. Yet, lots of people have (artificially) non-black hair, so **var(P)** will be reasonably high. Assume that the choice to dye your hair isn't genetically determined. Then, hair color in Japan has low heritability. (Lower, for example, than the US, which has more genetic diversity in hair color.) At the same time, we know that genes pretty much 100% determine what color hair grows on your head. Heritability is not a measure of how deterministic genes are in their effects.

Example: Heritability in different species Here are some numbers, courtesy of Falconer's *Introduction to Quantitative Genetics*.

Animal	Trait	Heritability (%)
Human	Height	65
	Immunoglobulin	45
Cow	Adult weight	65
	Butterfat %	40
	Milk yield	35
Pig	Back-fat thickness	70
	Efficiency of food conversion	50

Animal	Trait	Heritability (%)
	Litter size	5
Chicken	Weight	55
	Egg production	10
Mouse	Tail length	40
	Weight	35
	Litter size	20
Fruit Fly	Bristle number	50
	Body size	40
	Ovary size	30
	Egg production	20

Notice anything? The closer something is to reproduction, the lower heritability is. This seems odd at first. Isn't the number of eggs a fly produces incredibly important? Yes, but that's precisely the point! *Because* it is so important, evolution can select strongly for the genes with the optimal egg production numbers, leaving little genetic variance, and driving heritability down.

So low heritability doesn't necessarily mean that genes don't *matter* – it simply means that in the current population, most people have genes that do similar stuff, as compared with environmental causes.

(Technically, the numbers in the above table are estimates of *narrow*-sense heritability rather than *broad*-sense heritability as we are discussing. It doesn't matter since we're only looking at the general trend.)

Selective feeding

So far, genes and the environment have been independent. No one looked at your genes when deciding if you get a diet or feast environment. Let's change that. (We assume genes and the environment have the same effects as in the simplest model, i.e. **b = 1** and **a = ½**.)

While half of people are short/tall, and half get a diet/feast environment, change the odds that feasts go to the people with short vs. tall genes. Specifically, take this distribution:

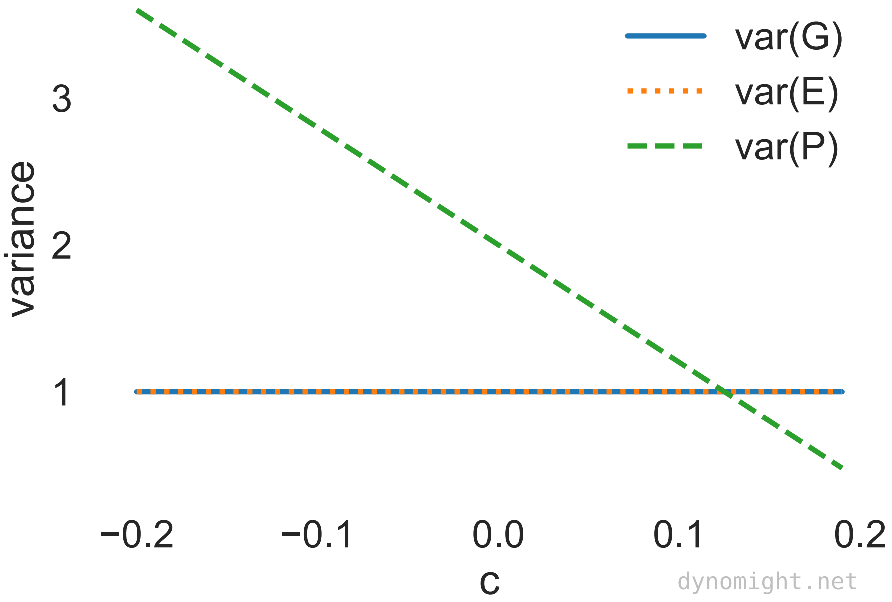
population distribution	diet	feast
short	¼ - c	¼ + c
tall	¼ + c	¼ - c

Here, **c** is any number between -¼ and +¼. If **c = -¼**, then people with short genes *always* get a diet, and people with tall genes always get a feast. If **c = +¼**, it's the opposite.

Now, there's something odd about this scenario. The genetic and environmental variances are always one. That's because, no matter the value of **c**, the distribution over genes and environments are always uniform. (You can check that the row and column sums in the above table are always 0.5.)

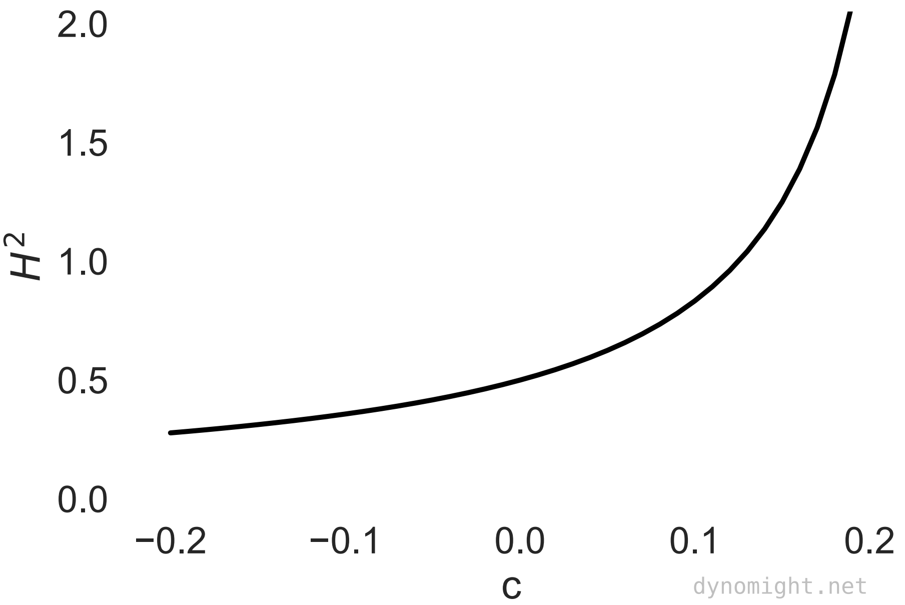
However, the *phenotypic* variance does change with **c**. Suppose that **c** = **¼**. People with short genes always get a feast, and thus have a height of 3. People with tall genes always get a diet, and so *also* have a height of 3. Everyone has the same height, so the phenotypic variance is zero!

Or suppose that **c** = **-¼**. People with short genes always get a diet, and so have a height of 1. People with tall genes always get a feast, and so have a height of 5. There is lots of phenotypic variance.



(Mathematically, what's happening here is that G and E are no-longer independent, so the variance of P is not just the sum of the variances of G and E but must be adjusted to include their (now-nonzero) covariance.)

The heritability, as ever, is $\mathbf{H^2{=}var(G)/var(P)}$. If we plot this it looks like the following:



In the “starve the short” regime (negative **c**) heritability decreases. In the “feed the short” regime (positive **c**) it increases.

Speaking of which, on the right of the plot, heritability goes above one! This is not a mistake. Remember, heritability is the ratio of genetic variance and phenotypic variance. *Usually*, the environment creates more variance, so this ratio is less than one. However, if people with short genes get more food than people with tall genes, then environment *decreases* variance. (Technically, if **c=¼**, then the phenotypic variance is zero, meaning the heritability is *infinite*!)

Takeaway: When the environment depends on genes, things get weird. If the environment *exacerbates* genetic differences, that *decreases* heritability. If the environment *reduces* genetic differences, then that *increases* heritability.

Example: Education and intelligence It’s plausible that education exacerbates genetic differences in intelligence. In most places today, I’d bet that genetic intelligence is positively correlated with the quality and quantity of education someone gets. (Partly because intelligence is correlated with income and partly because it helps to qualify for scholarships and selective schools/classes.) It also appears that education increases intelligence.

Now, imagine that society is reformed to remove correlations between genetic intelligence and education. (Maybe all students are assigned to public schools at random.) This would be like moving to the right on the above graph. Thus, strangely, more equal access to education would cause heritability to go *up*. If we went further and provided *better* education to less gifted students, that would increase things even further.

It doesn't matter if more equal education is a good idea, or even if we truly are in a “starve the short” regime now. The point is that *heritability depends on society.*

Three key takeaways

1. Less random environments lead to higher heritability. If everyone had completely equal environments, every trait would be 100% heritable.
2. Heritability depends on the amount of genetic variance in the population. If people have similar genes, heritability goes down. You don't need anything dramatic like new mutations.
3. If genetics are correlated with the environment, weird things can happen. If environmental correlations decrease variance, then total variance might be lower then genetic variance. If so, heritability is greater than 100%!

Heritability is really just a ratio of variances. You'll mislead yourself thinking about it any other way.

Factors of mental and physical abilities

Is there a general factor of intelligence?

This question is a trap. If you try to answer it, you'll find yourself beset by semantic questions. What's *intelligence*? What's a *factor*? And if you get past those, you'll then find a bleak valley of statistical arcana. What do the eigenvalues look like? Do they imply causality?

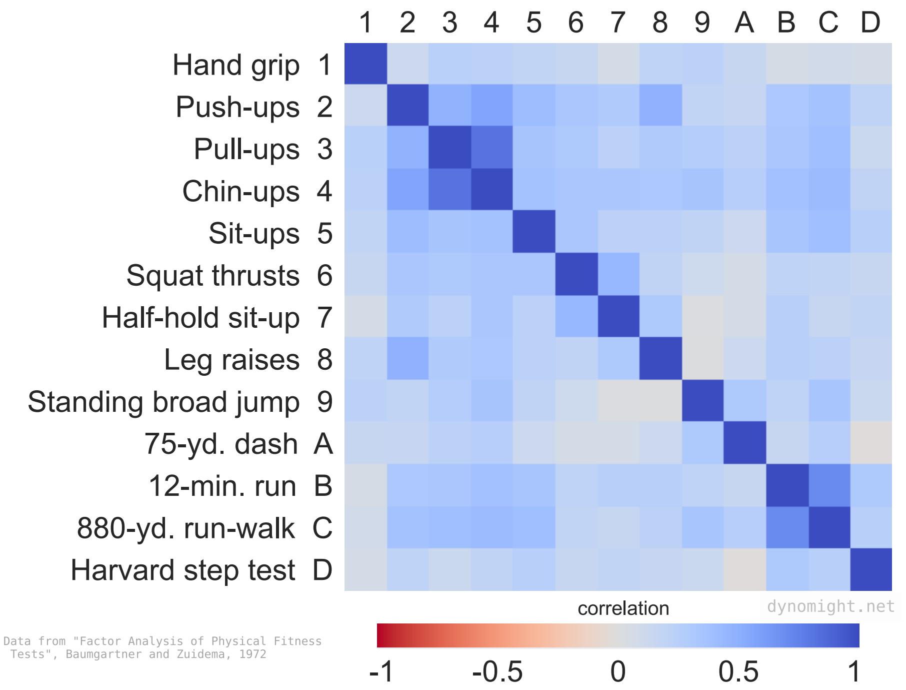
This is all backward. If your goal is to understand the external world, you can skip the hand wringing and start by looking at the damn data. So let's do that.

To start, let's forget about intelligence, and ask an easier question: Does it make sense to refer to some people as “more physically fit” than others? Is that reasonable, or do we need to break things down by strength, speed, coordination, etc.?

To answer this, I looked for studies that gave people batteries of different physical tests. I found three that make enough information available to analyze.

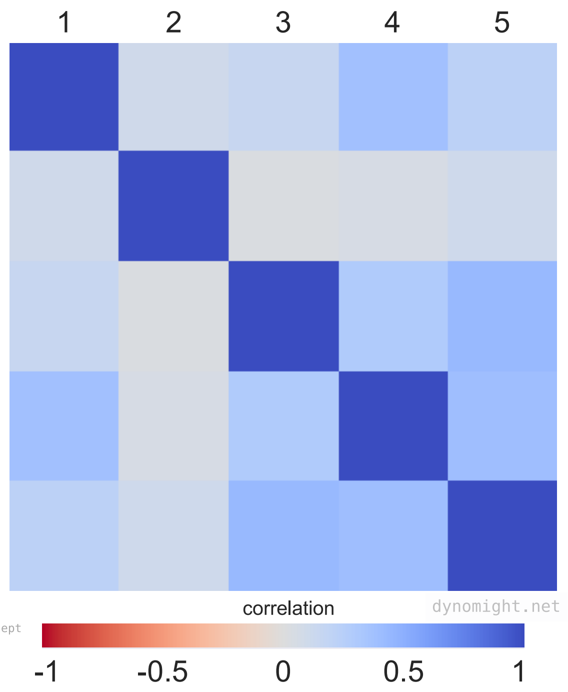
Here are the correlations among the different tests (click to open/close). The columns are the same as the rows—so the 3rd square in the first row is the correlation between hand grip and pull-ups.

Baumgartner and Zuidema



This is males. (Females are similar, except with lower correlations in upper-body strength.)

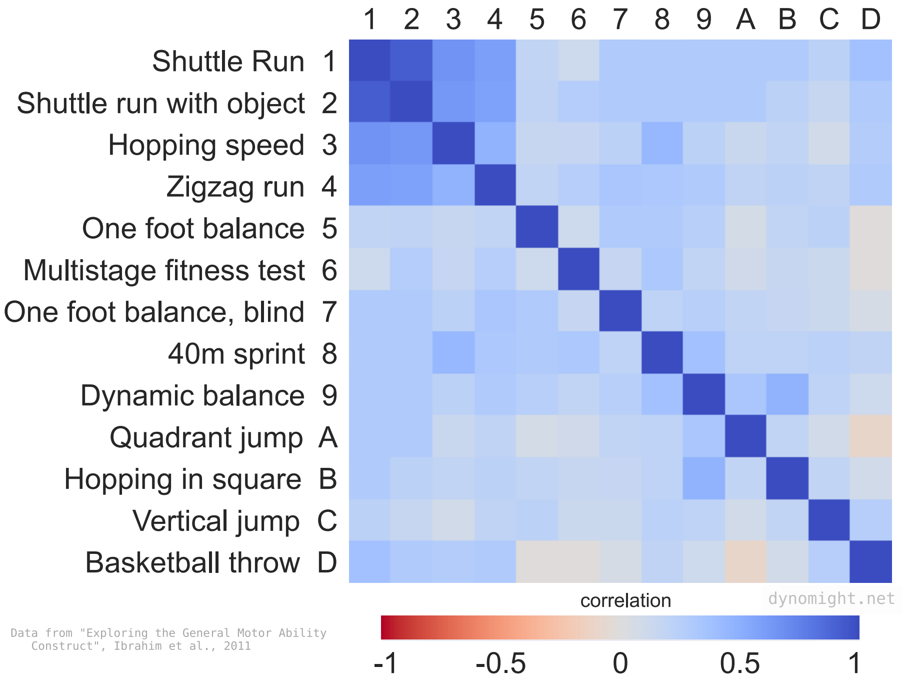
Marsh and Redmaye



dynomight.net

Data from "A Multidimensional Physical Self-concept and Its Relations to Multiple Components of Physical Fitness", Marsh et al., 1994

Ibrahim et al.



dynomight.net

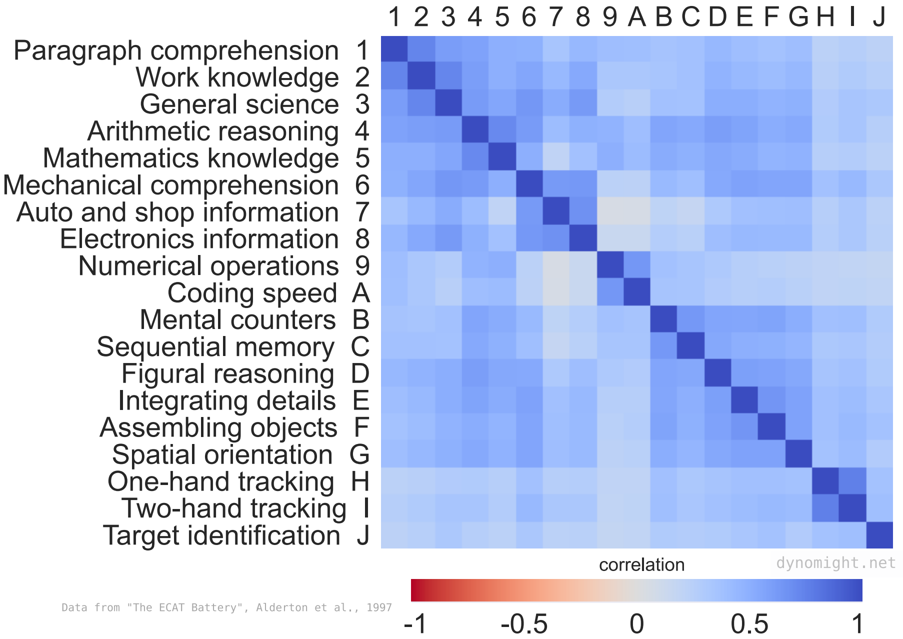
Data from "Exploring the General Motor Ability Construct", Ibrahim et al., 2011

Most tests are positively correlated, and none are negative. Is this surprising? I don't know. For our purposes, we just want this as a point of comparison.

We can do the same analysis with batteries of mental tests. For whatever reason, many more studies have been done with mental tests, so I picked four of the best.

Here are the correlations in each of the studies (again, click to open/close):

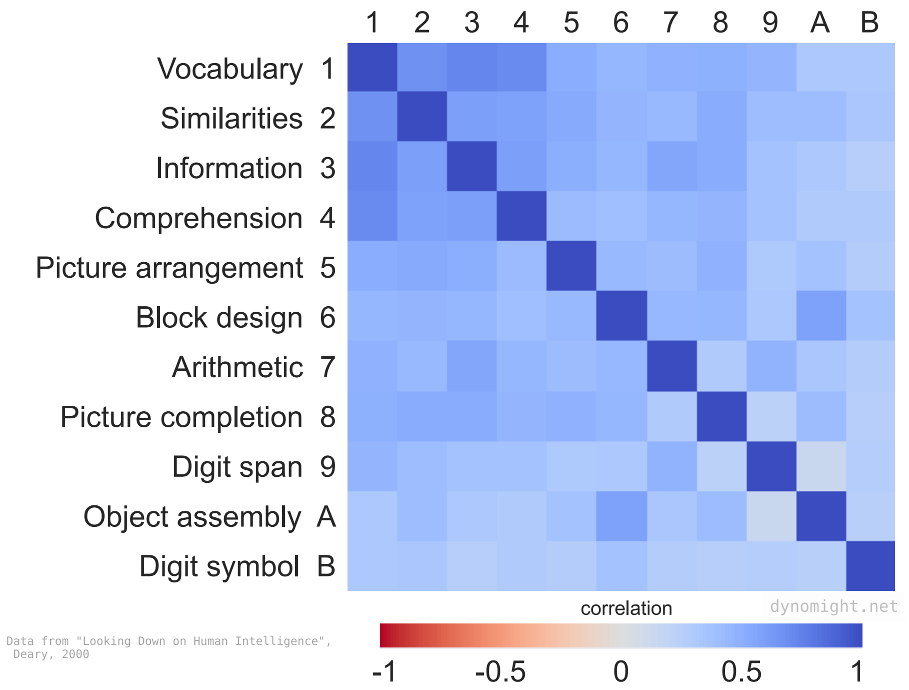
Alderton et al.



Data from "The ECAT Battery", Alderton et al., 1997

This test battery is designed to measure aptitude for various military tasks. Some of these, like tracking and target identification, are partly physical.

Deary

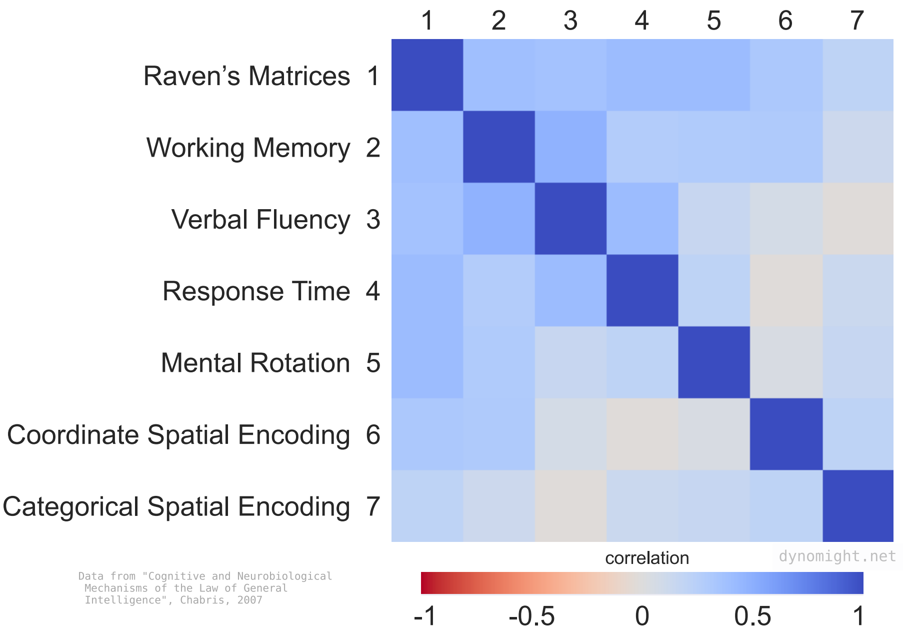


Data from "Looking Down on Human Intelligence", Deary, 2000

These subjects were tested on the 11-component revised Wechsler Adult Intelligence Scale.

As an aside, the origins of these data are somewhat obscure: Chabris published them, crediting Deary (2000) who in turn credits personal communication from Crawford, with no other information. Whoever Crawford is, they deserve way more recognition on Wikipedia.

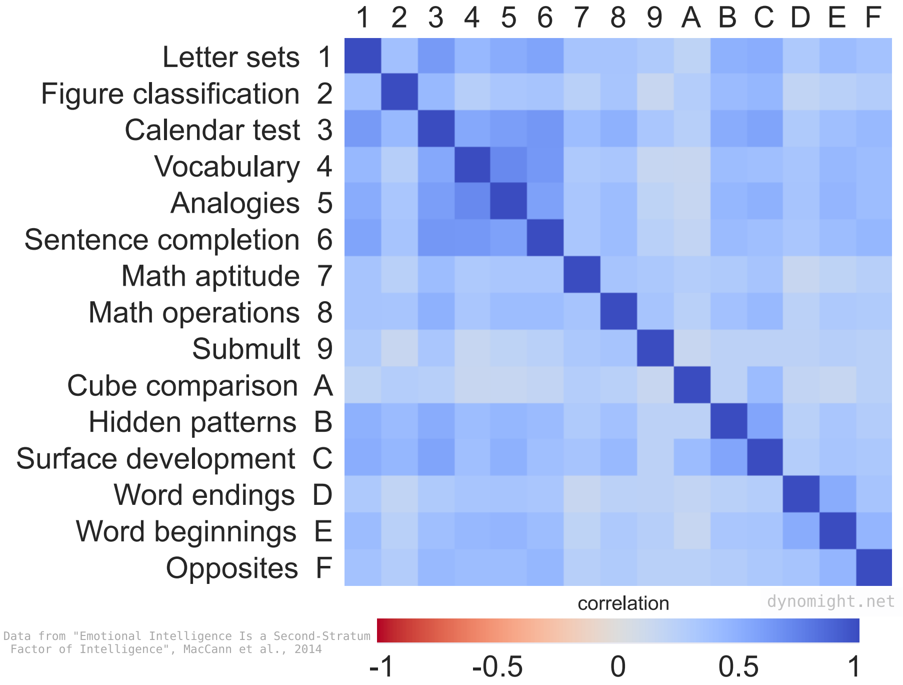
Chabris



Data from "Cognitive and Neurobiological Mechanisms of the Law of General Intelligence", Chabris, 2007

The data was gathered as part of a project to test decision-making. Raven’s matrices test shape pattern recognition, working memory is tested via 3-back, and the coordinate encoding tests check if people can tell the distance or orientation of a dot relative to a line.

MacCann et al.



Data from "Emotional Intelligence Is a Second-Stratum Factor of Intelligence", MacCann et al., 2014

The colors group the tests into those that test fluid reasoning, comprehension, quantitative knowledge, visual processing, and long-term storage/retrieval.

What do we see?

The same basic pattern holds for both physical and mental tests.

First, **almost everything is positively correlated**. You might imagine that people with more upper-body strength would be worse runners—what with the extra muscle they need to carry around. You might imagine that people who are good at paragraph comprehension would be worse at math. But that's not what happens.

Second, **more similar stuff is more correlated**. It's natural that chin-ups are strongly correlated with pull-ups, or that arithmetic reasoning is strongly correlated with mathematics knowledge. It's more surprising that hand-grip strength is correlated with the 75-yd dash or that paragraph comprehension is correlated with target identification. These more surprising correlations are weaker, but still positive.

Third, **the results are robust**. The tests span several decades, different countries, and many different test batteries. The basic pattern doesn't change.

Things are correlated. No one seems to seriously dispute this. So why all the debates?

For one thing, the discussion sometimes ascends into meta-controversy. There are many arguments to be had about the definition of “general intelligence”. Some people even debate if there is anything controversial! (I take no position here, but note that the “not surprising” camp doesn’t seem to agree on *why* it’s not surprising…)

On the lower planes of argument, the main issue is if the tests are *just correlated* or if there’s something deeper going on underneath of them. Here, the burden of proof falls on whoever claims there is something deeper.

Aside: The mental correlations are somewhat stronger than the physical ones, but don’t take that too seriously. The mental tests used more diverse populations than the physical tests. Imagine doing physical tests on a group of 20-year-olds. If you throw in a bunch of 80-year-olds, they’ll be worse at everything and correlations will shoot up.

Factor analysis is like a cigar

The typical argument that there’s something deeper happening relies on a statistical technique called factor analysis. This is usually described with fancy technical language, but the basic idea is just that you can summarize all the tests for each person using a single number.

Let’s make this concrete. Say you go out and grab random people and test them, and get this data:

Person	Test 1	Test 2	Test 3
Antonio	1	-1	1.5
Bart	.67	-.67	1
Cathy	-0.5	0.5	-0.75
Dara	-2	2	-3

You can visualize the data as a magical rotating point-cloud:

Now, notice something important: This data is *special*, in that the points fall along a straight line. This means that even though each person took 3 tests, you can represent each person using a single number, namely their position on the line. If you tested lots of people, and the data looked like this, then each person’s position along the line would be the “general factor”.

Of course, real data would never exactly look like this. It would have noise! To reflect that, we need to build a “model”. That is, we will try to build a “simulator” that can make fake data that (hopefully) looks like real data.

Simulations

The simplest simulator would be to just generate people along a line. First, pick some direction of variation. Then, to simulate a person (i.e. a set of test scores), draw a random number **g** from a standard "bell curve" Normal distribution to represent their position along the main direction.

Here's an example, where we choose a direction of variation similar to the dataset above. If you simulate a bunch of people, you'll get a dataset that looks like this:

Of course, real data will never look like that—there will always be "noise", either from measurement error, or from certain people randomly being good/bad at certain tasks. To account for this, let's update our simulator, by adding some random noise to each point. This produces data that looks like a cigar.

The critical thing here is that cigars are rotationally symmetric. If you "roll" the point cloud along the main axis of variation, it still looks basically the same.

Now we can finally say what factor analysis is. It's an algorithm that takes a real dataset and adjusts the shape of the cigar so that the simulated data will look as much like the real data as possible. It can modify the direction of variation, and how "thick" the cigar is, but that's it. (Note: all this describes the *simplest* possible variant of factor analysis, which is all we need here.)

If your dataset looks like a cigar, factor analysis will fit well. If not, it will fit poorly. Here's an example of the kind of data factor analysis can't represent:

The meaning of cigars

Factor analysis tries to approximate your data with a cigar. Why should you care about this?

Let's back up. As we saw earlier, physical and mental tests are correlated. If you learn that Bob scored well on paragraph comprehension that raises your estimate for how Bob will do on coding speed.

But say your data was a cigar. Take Bob's position along the length of the cigar, and call it **g**. Say Bob's value for **g** is low. If that's all you know, and you had to guess Bob's coding speed, you'd give a low number.

Now, suppose that in addition to **g**, you learn that Bob did well on paragraph comprehension. How does this change your estimate of Bob's coding speed? Amazingly, *it doesn't.* The single number **g** contains all the shared information between the tests.

In a cigar distribution, once you know **g**, everything else is just random noise—one test no longer tells you anything about any other. (Mathematically, once you control for **g**, the partial correlations of the tests are zero.)

In a *non*-cigar distribution, this doesn't happen. There's no single number that will make all the tests uncorrelated. Some interesting structure would remain unexplained.

Mental tests aren't not cigars

So, what does real data look like? Is it a cigar? Can we capture all the structure with a single number?

Here I took the earlier cigar data, and manually drew three lines to capture the "shape" of the data:

The blue line corresponds to the main direction of variation, while the shorter red and green lines correspond to the random noise added to each point. You can see that the shorter lines are the same length. This happens because factor analysis models are rotationally symmetric.

In contrast, here's the earlier "non-cigar" data:

Here, the shorter green and red lines are different lengths, reflecting that there is no rotational symmetry.

OK, I lied. I didn't draw the lines manually. There's a simple algorithm that can automatically compute these for any dataset. (By computing a singular value decomposition of the covariance matrix, if those words mean anything to you.) The details don't particularly matter, just that we can automatically find lines that span a point cloud. This will be important when we move beyond three dimensions.

So now we have a plan: We will take a real dataset, compute these lines, and see how long they are. If we have one long line and a bunch of equal-length short lines, then the data is cigar-like, meaning that a single variable explains all the "interesting" latent structure. If we have a bunch of lines of random lengths, then the data isn't cigar-like, meaning that we can't summarize things with one number.

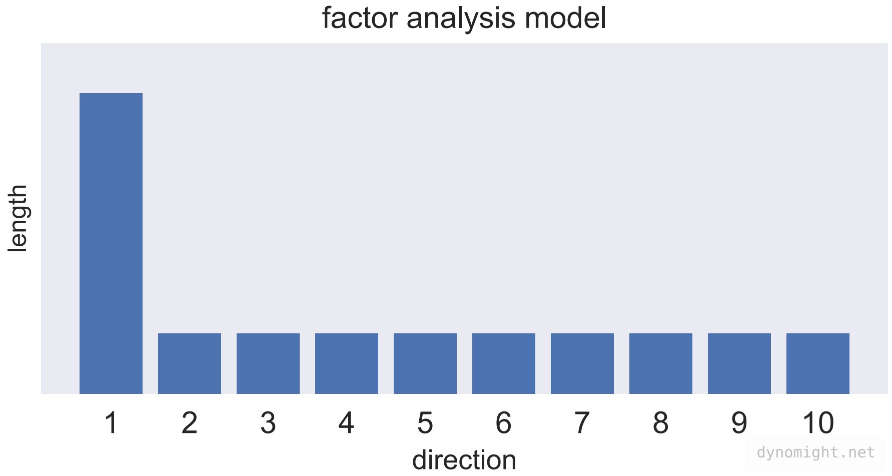
I'd like to show you the real data from the datasets above, but none of them seem to be publicly available. Still, we can approximate it by generating data from a multivariate Normal with the known covariance.

Here are the first three tests of Alderton et al.'s (Paragraph comprehension, work knowledge, and general science).

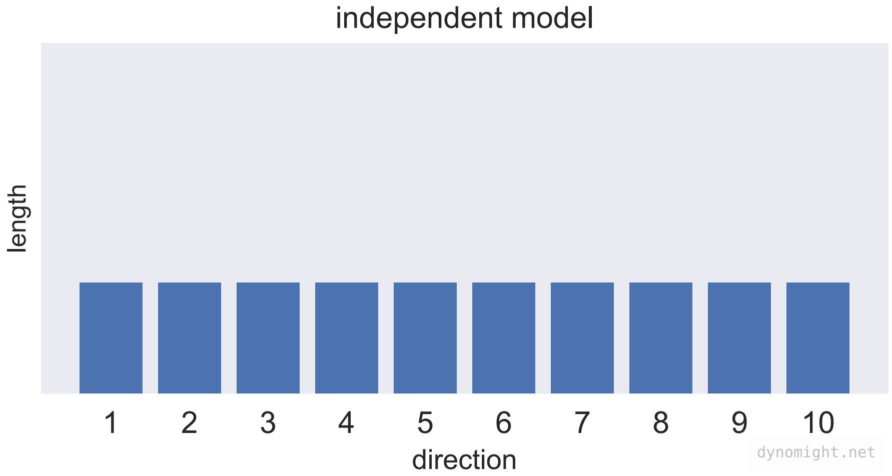
It's not a perfect cigar, but it's not exactly *not* a cigar either. Here are the relative lengths of the three directions in decreasing order:

```
1st direction (blue line):   0.890
      2nd direction (red line):    0.362
      3rd direction (green line):  0.279
```

What if we use all seven tests? We can't make pretty pictures in seven dimensions, but we can still do the math. With **N** tests, a factor analysis model always produces **1** "long" direction and **N-1** "short" directions. If we plot the length of the directions, it should look like this:

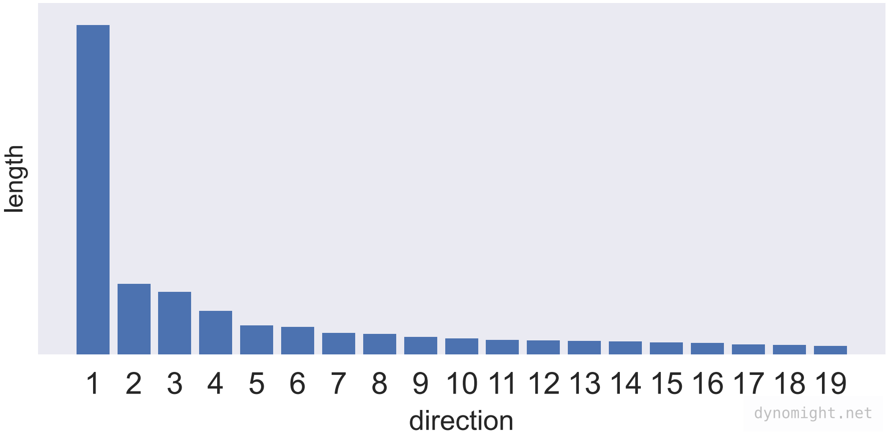


In contrast, here's how things would look if all the tests were completely uncorrelated.

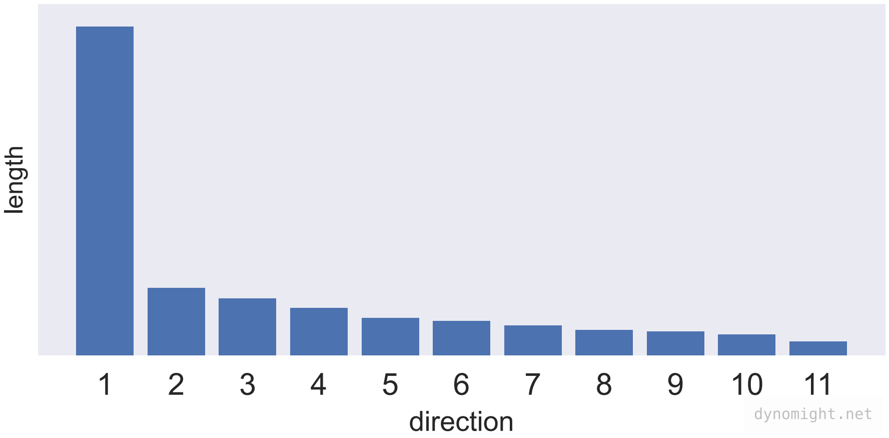


What do the lengths on real data look like? Well, judge for yourself. (Again, click to open/close)

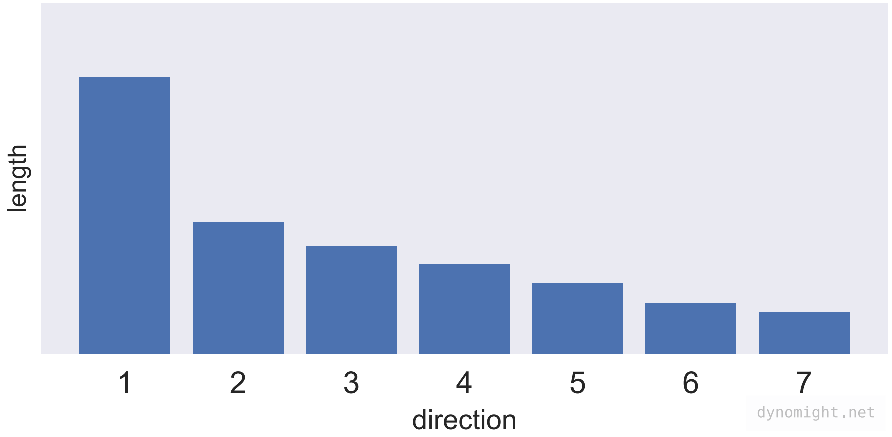
Alderton et al.



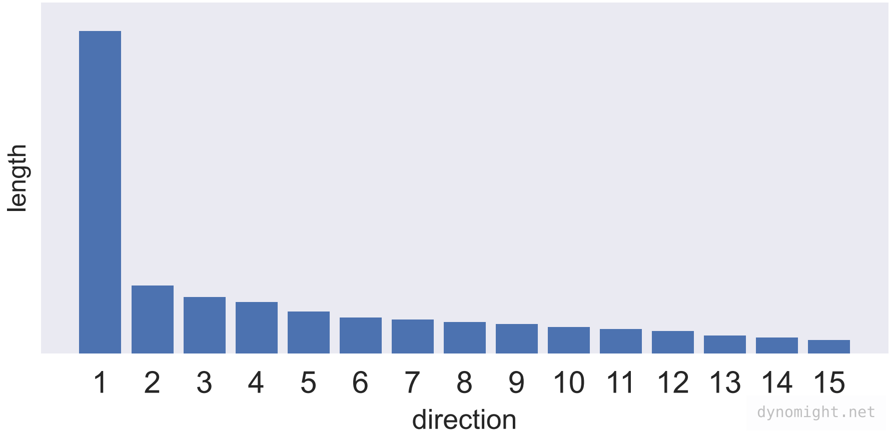
Deary



Chabris



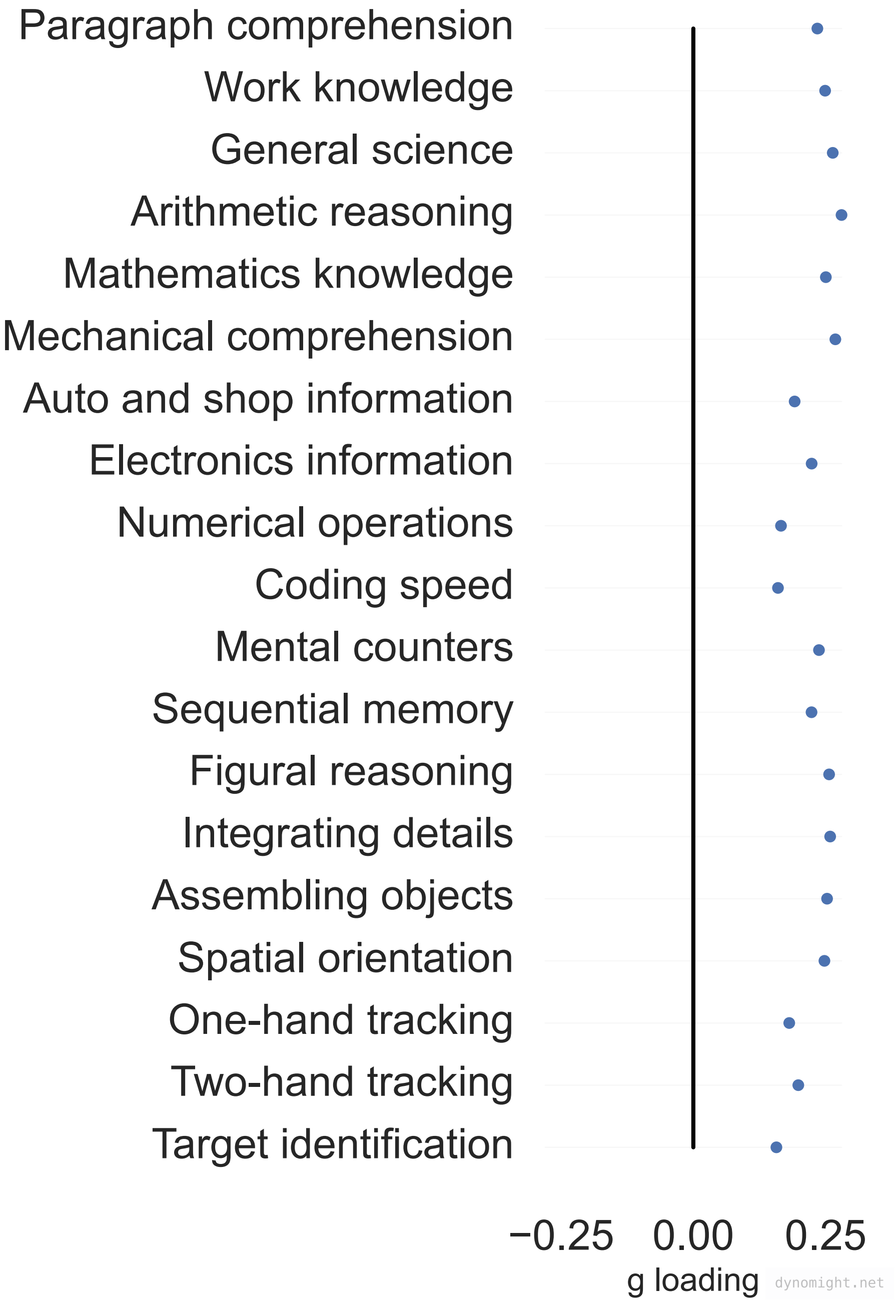
MacCann et al.



Do these look exactly like what factor analysis can produce? No. But it's a reasonable approximation.

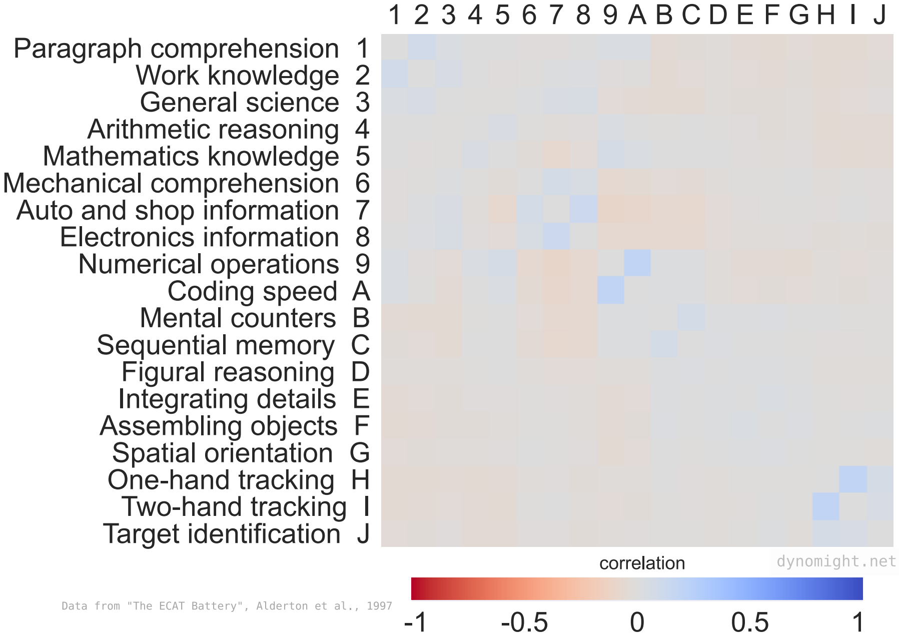
Directions of variation

Here's another way of visualizing things. For any dataset, we can take the principal direction of variation (the blue line) and look at its length along each of the tests. This says, essentially, how much each of the tests contributes to the main direction of variation. Here's what we get if we do that for Alderton et al.:



Calculating **g** is similar to taking a simple average of the test scores, though the weights are *slightly* higher on some tasks than others.

If we calculate **g** like this for each person, we can then compute the *partial* correlations. These are the correlations once you control for **g**. Here's what that gives for Alderton et al.:

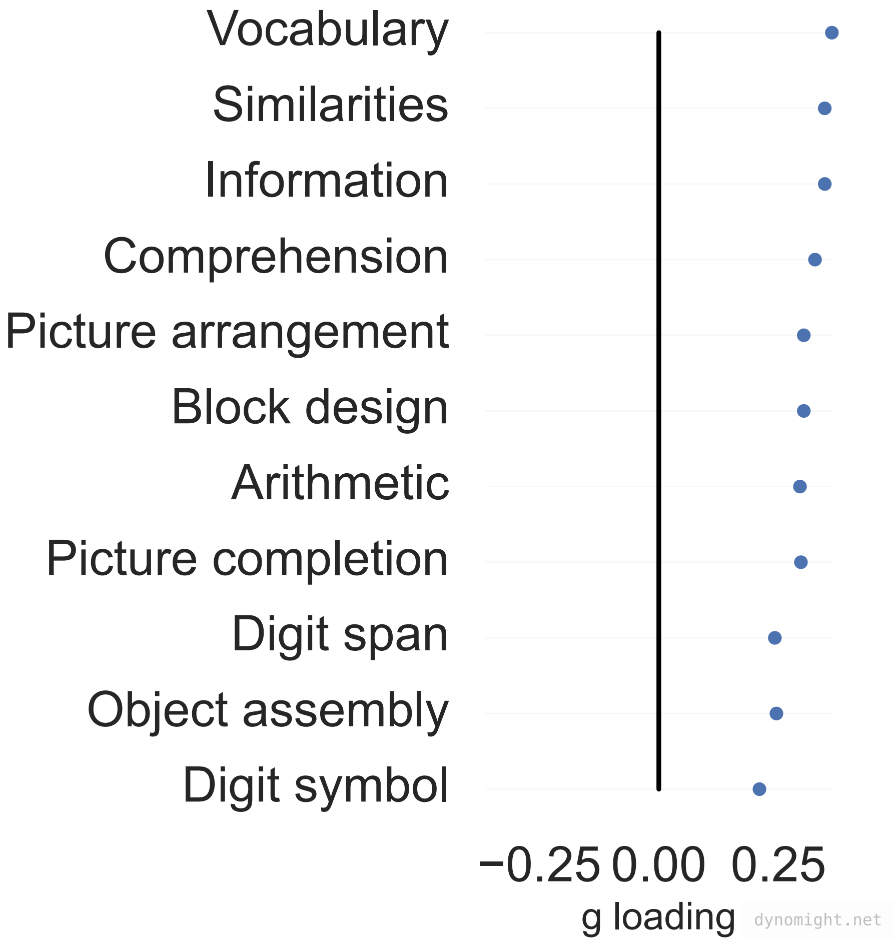


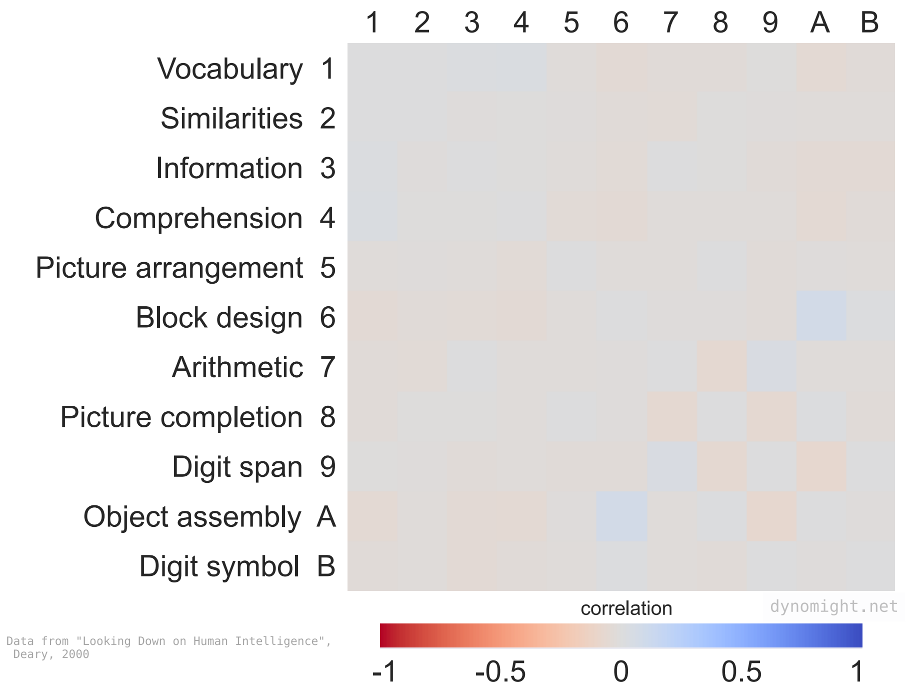
Data from "The ECAT Battery", Alderton et al., 1997

Mostly it's a sea of gray, indicating that the partial correlations are all quite small.

And here's the *g* loadings and partial correlations for the other studies:

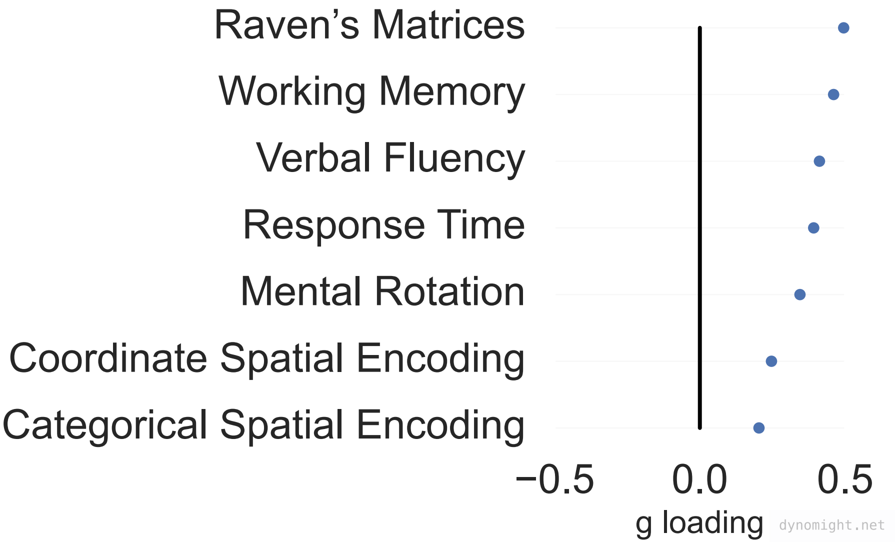
Deary

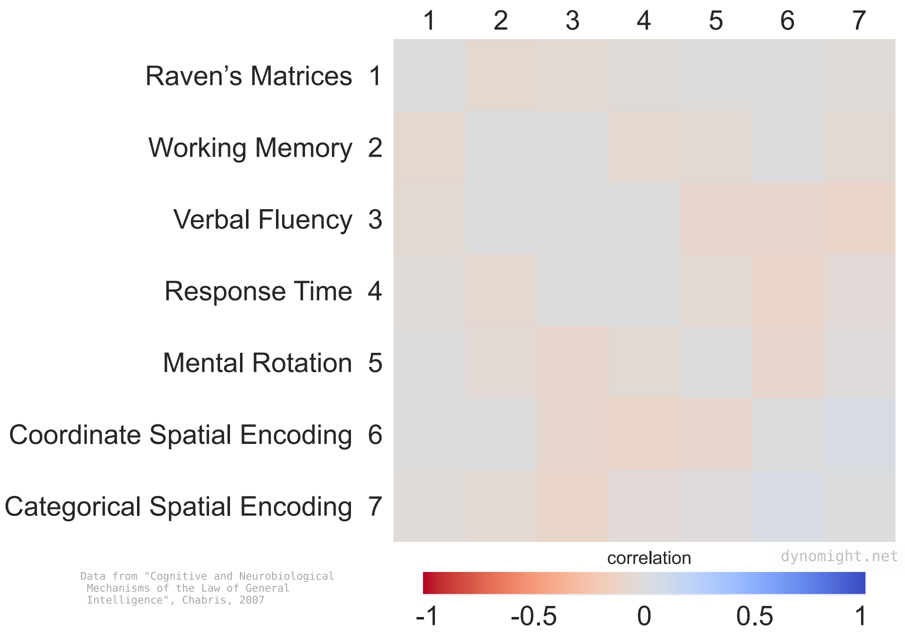

Vocabulary
Similarities
Information
Comprehension
Picture arrangement
Block design
Arithmetic
Picture completion
Digit span
Object assembly
Digit symbol
−0.25
0.00
0.25
g loading
dynomight.net



Data from "Looking Down on Human Intelligence", Deary, 2000

Chabris

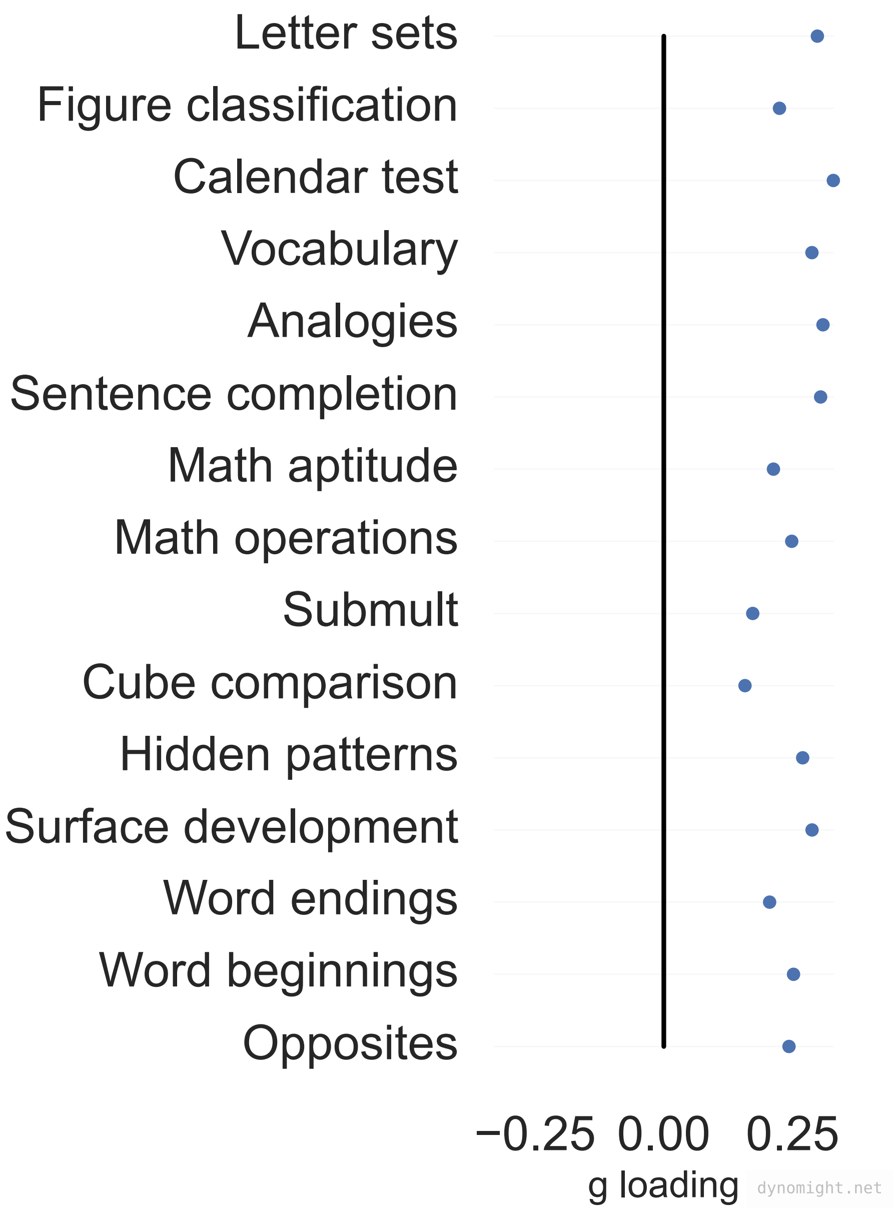


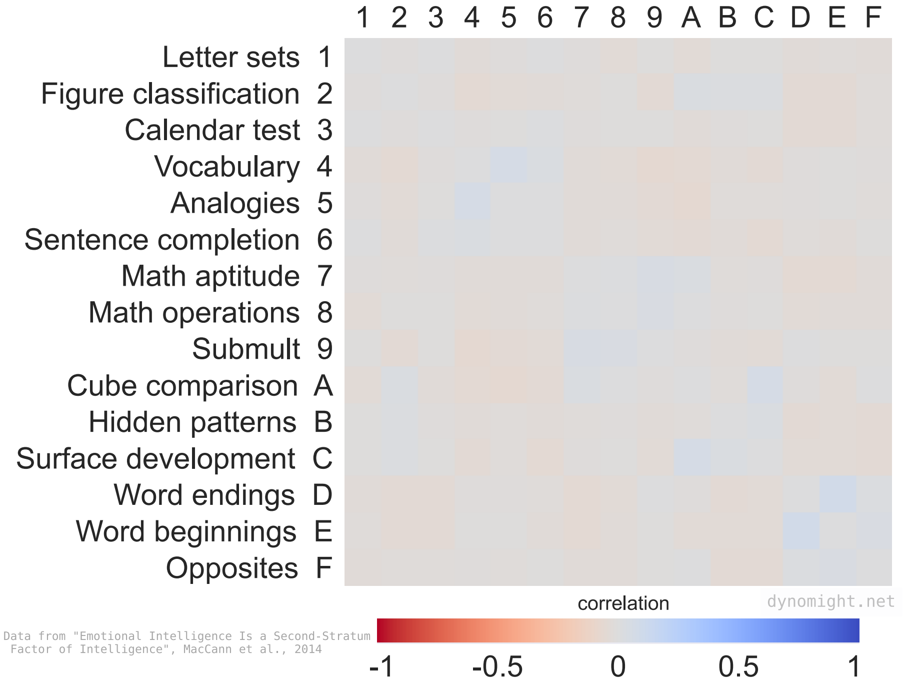


dynomight.net

Data from "Cognitive and Neurobiological Mechanisms of the Law of General Intelligence", Chabris, 2007

MacCann et al.


Letter sets
Figure classification
Calendar test
Vocabulary
Analogies
Sentence completion
Math aptitude
Math operations
Submult
Cube comparison
Hidden patterns
Surface development
Word endings
Word beginnings
Opposites
−0.25
0.00
0.25
g loading
dynomight.net



Data from "Emotional Intelligence Is a Second-Stratum Factor of Intelligence", MacCann et al., 2014

If factor analysis was a perfect fit, these would all be zero, which they aren't. But they are pretty small, meaning that in each case, the single number *g* captures most of the interesting correlations.

What would g look like?

Factor analysis is a decent but not perfect model of mental tests. What does this tell us about how intelligence works? Well, suppose that factor analysis was a *perfect* model. Would that mean that we're all born with some single number **g** that determines how good we are at thinking?

No. A perfect fit would only mean that, across a population, a single number would *describe* how people do on tests (except for the "noise"). It does not mean that number *causes* test performance to be correlated.

This is a point that often comes up in "refutations" of the existence of *g*. People argue, essentially, that even though tests are correlated, they might be produced by many *independent causes*. I'd go further—we *know* there are many causes. While intelligence is strongly heritable, it's highly polygenic. Dozens of genes are already known to be linked to it, and more are likely to be discovered. How "broad" the effects of individual genes are is an active research topic. It's harder to quantify environmental influences, but there are surely many that matter there, too.

So, no, the above data doesn't imply that there's no magical number **g** hidden in our brains, just like it doesn't imply that there's single number in our bodies that says how good we are at running, balancing, or throwing stuff. But that doesn't change the fact that a single number provides a good *description* of how good we are at various mental tasks.

Suppose you're hiring someone for a job that requires a few different mental tasks. (Arithmetic, sequential memory, whatever.) If you knew someone's **g**, you could guess how well they'd do at each task. But it would only be a guess! To really know, you still need to test the skills individually. That's the key word: *Individually.* It's not that **g** tells you everything—it doesn't—it's just that once you know **g**, how good someone is at one task doesn't tell you anything about how good they'll be at another.

Again, that's assuming factor analysis were a perfect fit. Which it isn't. Though it's close.

Takeaways

1. Skill at mental and physical tasks are positively correlated. More similar stuff is more correlated.
2. A factor analysis model tries to model data with a "cigar" shape. These models fit mental and physical tests reasonably well, but not perfectly.
3. Call the position along the "long axis" of the cigar **g**. A perfect fit wouldn't mean that **g** contains all the information about how good someone is at different tasks—only that it contains all *shared* information.

Political polarization is partly a sample bias illusion

We're here on Earth for such a short time. So, I often wonder—what do people spend their days thinking about? Judging from the ever-increasing amount of screaming everywhere, the answer would seem to be politics. But is that right? What opinions do normal people really have?

Asking my friends won't help. They're a biased subpopulation—otherwise, we'd have, like, universal pet insurance funded by a land-value tax. (Or at least lower particle levels in subways.)

The solution is to find lots of random people, and ask them all a battery of questions. That is, polling.

Polling

To better understand what people think, I found the 2020 Cooperative Election Study Common Content survey, which is a representative sample of 61,000 American adults. (They've helpfully made it impossible to link directly, but you can get a description by going here and clicking on "Access File").

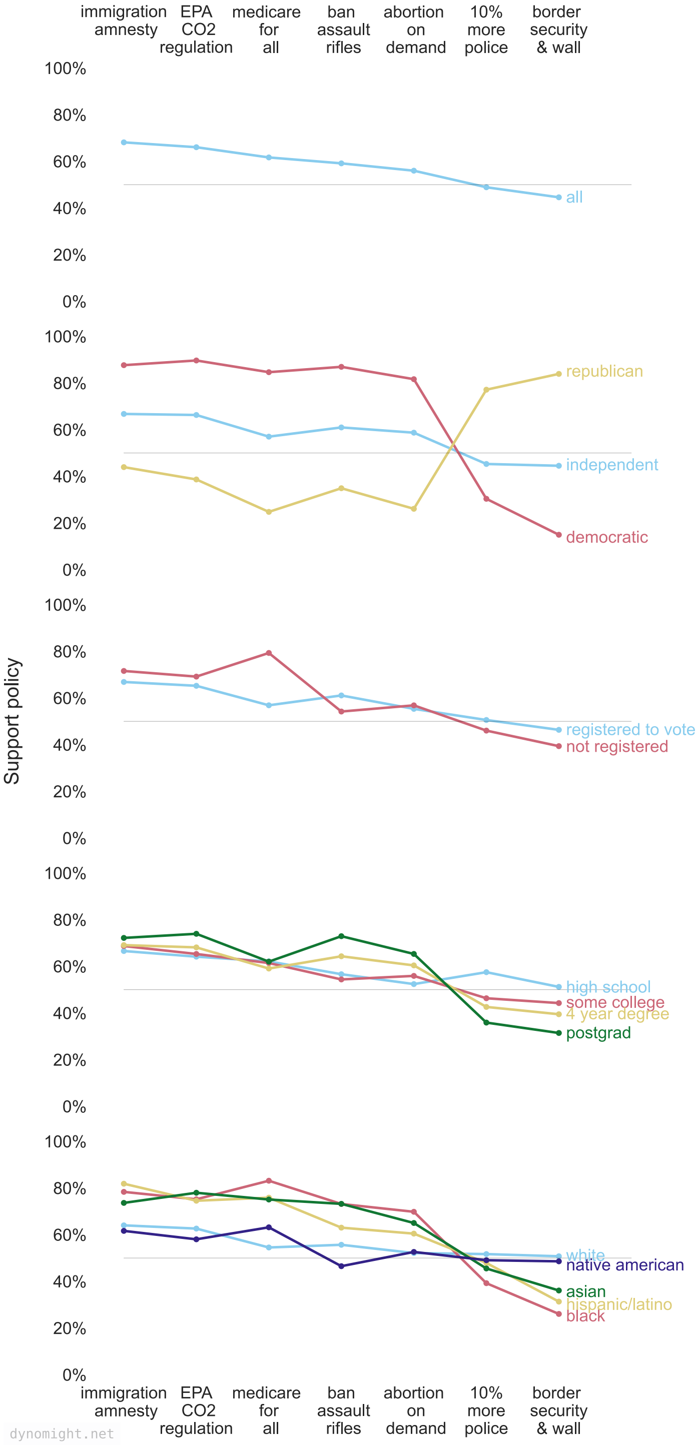
It asks boring questions like if people approve of Trump or Congress or whatever. But it also asks if people approve of a ton of different policies. I picked seven that I thought were representative. Here they are, along with their exact wording:

Immigration amnesty	“Grant legal status to all illegal immigrants who have held jobs and paid taxes fo
EPA CO2 regulation	“Give the Environmental Protection Agency power to regulate Carbon Dioxide e
Medicare for all	“Expand Medicare to a single comprehensive public health care coverage progran
Ban assault rifles	“Ban assault rifles”
Abortion on demand	“Always allow a woman to obtain an abortion as a matter of choice”
10% more police	“Increase the number of police on the street by 10 percent, even if it means fewe
Border security & wall	“Increase spending on border security by $25 billion, including building a wall be

We have every answer from each person, along with their political party, education level, race, income, and so on.

What people think

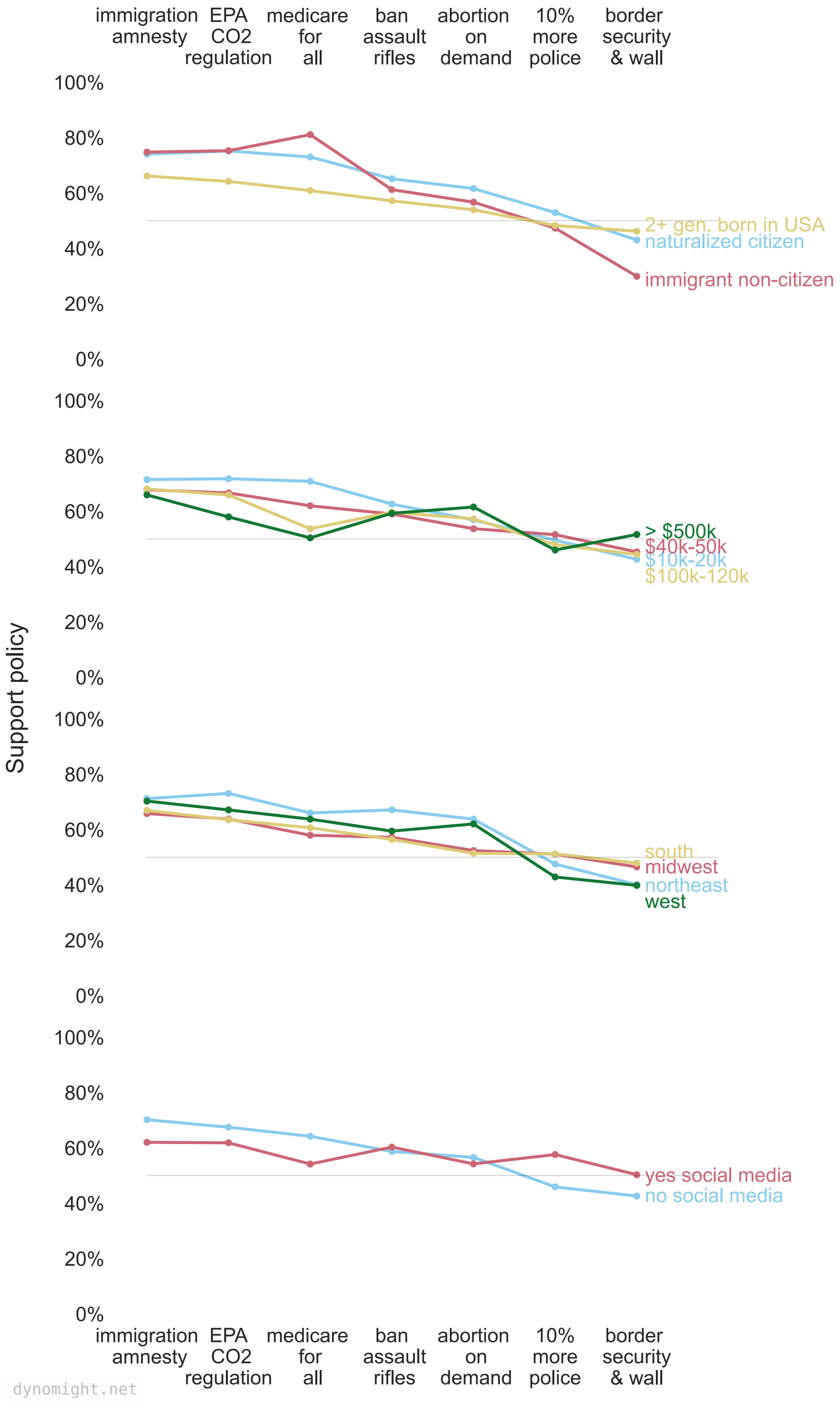
Here’s a first visualization. This shows the percentage of people that support each policy, broken down into various groups.


immigration amnesty
EPA CO2 regulation
medicare for all
ban assault rifles
abortion on demand
10% more police
border security & wall
100%
80%
60%
40%
20%
0%
Support policy
all
republican
independent
democratic
registered to vote
not registered
high school
some college
4 year degree
postgrad
white
native american
asian
hispanic/latino
black
dynomight.net

Here's what I conclude from this:

- **First two panels**: For most issues, a majority of people support the leftist / Democratic position.
- **Third panel**: Registered voters are more right-leaning than non-registered voters. However, this pattern reverses for assault weapons, where registered voters favor a ban.
- **Fourth panel**: More educated people lean left, though there's little effect on medicare for all.
- **Fifth panel**: Blacks, asians, and hispanics are more left-leaning than whites and native Americans.

Here's a chart that breaks people up in other ways that turn out to be less impactful: If they are an immigrant, their family income, where they live in the US, and if they used social media in the last 24 hours:

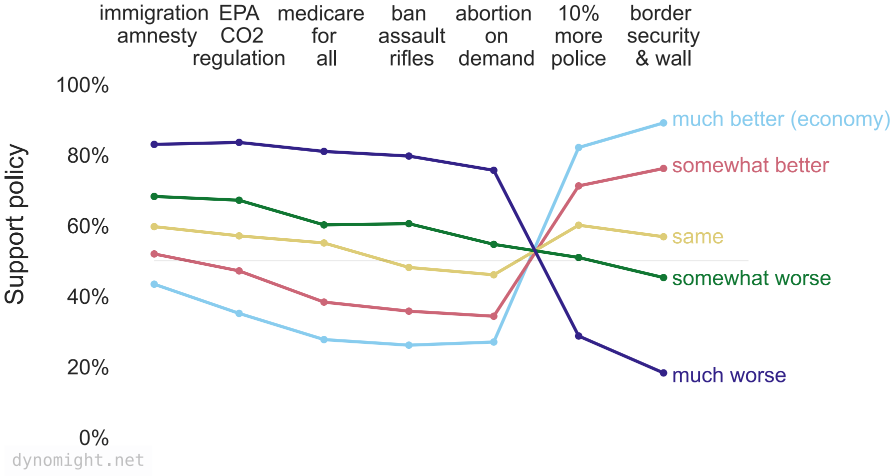

immigration amnesty
EPA CO2 regulation
medicare for all
ban assault rifles
abortion on demand
10% more police
border security & wall
Support policy
100%
80%
60%
40%
20%
0%
2+ gen. born in USA
naturalized citizen
immigrant non-citizen
> $500k
$40k-50k
$10k-20k
$100k-120k
south
midwest
northeast
west
yes social media
no social media
dynomight.net

What does this mean?

- **These people lean more to the right**: Non-immigrants, richer people, those in the south and midwest, and social media users.
- **These people lean more to the left**: Immigrants, poorer people, those in the northeast and west, and social-media abstainers.

There are a few subtleties: The pattern reverses for the rich on abortion and the police. There's no pattern between if someone uses social media and how they feel about assault weapons or abortion. Naturalized citizens (but not non-citizens) favor increasing the number of police.

Some other ways you can break things up that seem interesting, but turn out to be kind of dumb. For example, here's what you get if you break things down by how people think the economy changed over the last year.



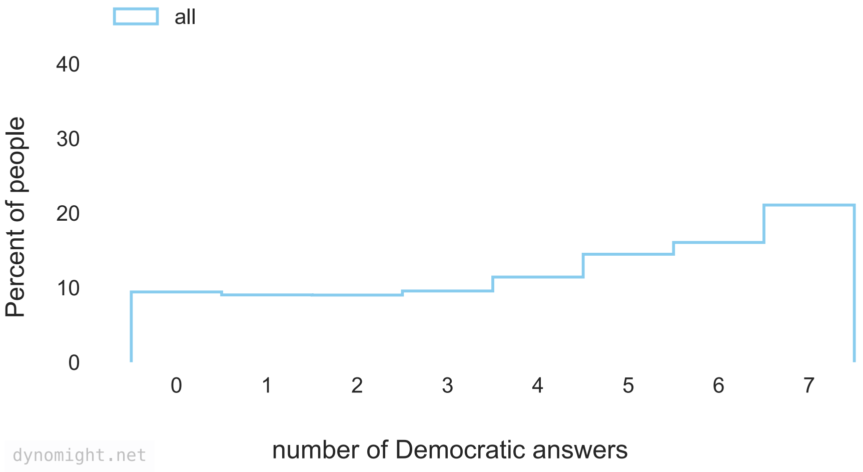
This looks interesting at first, but I think it just shows the power of motivated reasoning. Remember, this survey was done in the run-up to the 2020 presidential election, where Republican Donald Trump was running for a second term. Democrats convinced themselves the economy was terrible, while Republicans did the opposite. That's all this is showing.

Heterodoxy

The above graphs are organized around *questions* rather than *people*.

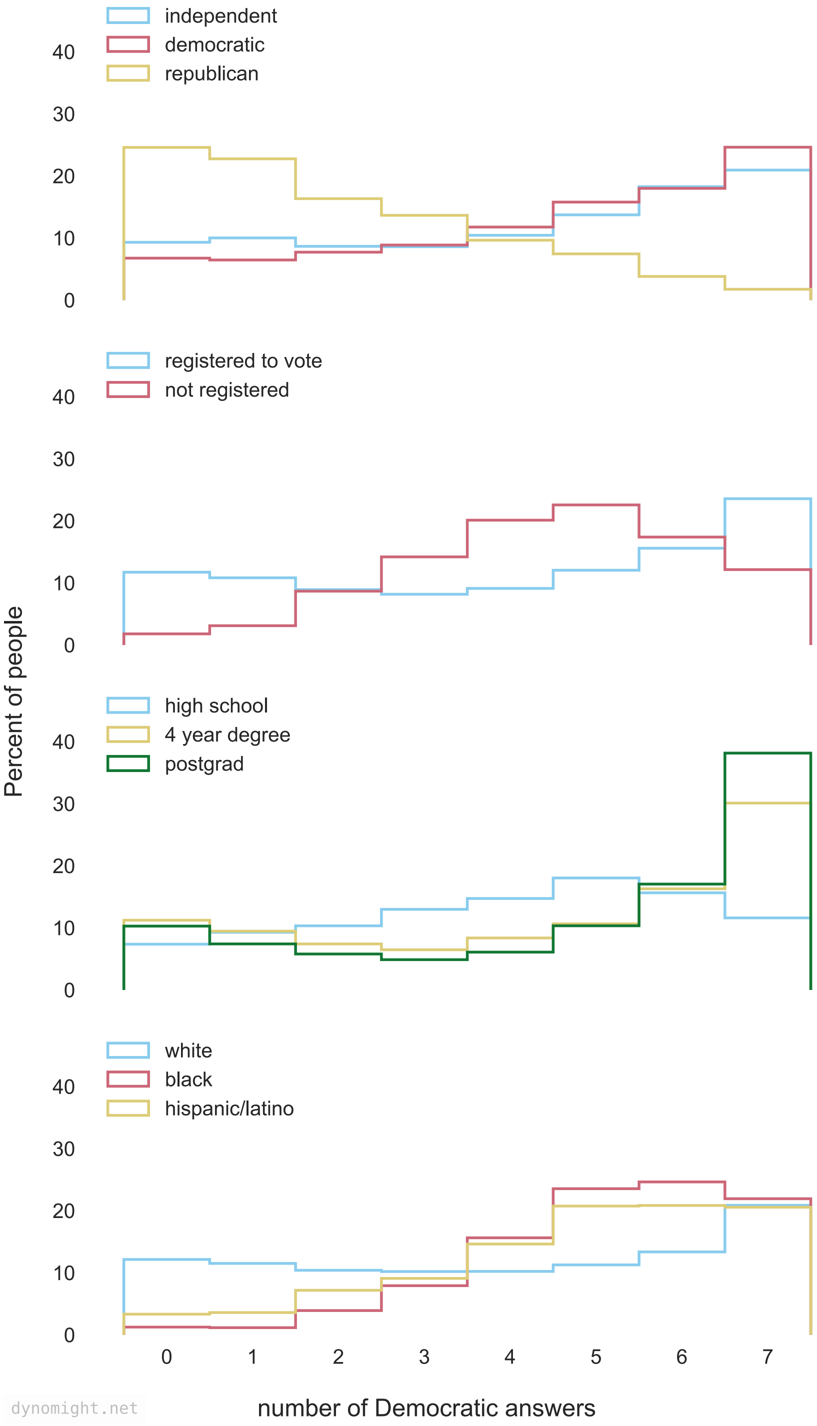
Remember that, for each of the policies, around 55% of social media users were in support. How how does that 55% come about? Is it 55% supporting all the policies, and 45% opposing them all? Does everyone support some random subset? These situations would represent very different degrees of polarization, but we can't tell them apart from the above graphs.

We need to look at interactions between how individual people answered different questions. To do this, I calculated how many "Democratic answers" they gave, i.e. supporting the first five policies or opposing the last two. Here is a histogram of the full population:



Around 10% of people give "all Republican" answers, while around 20% give "all Democratic" answers. Between the two it's a gradual change. This first glimpse doesn't look very polarized at all.

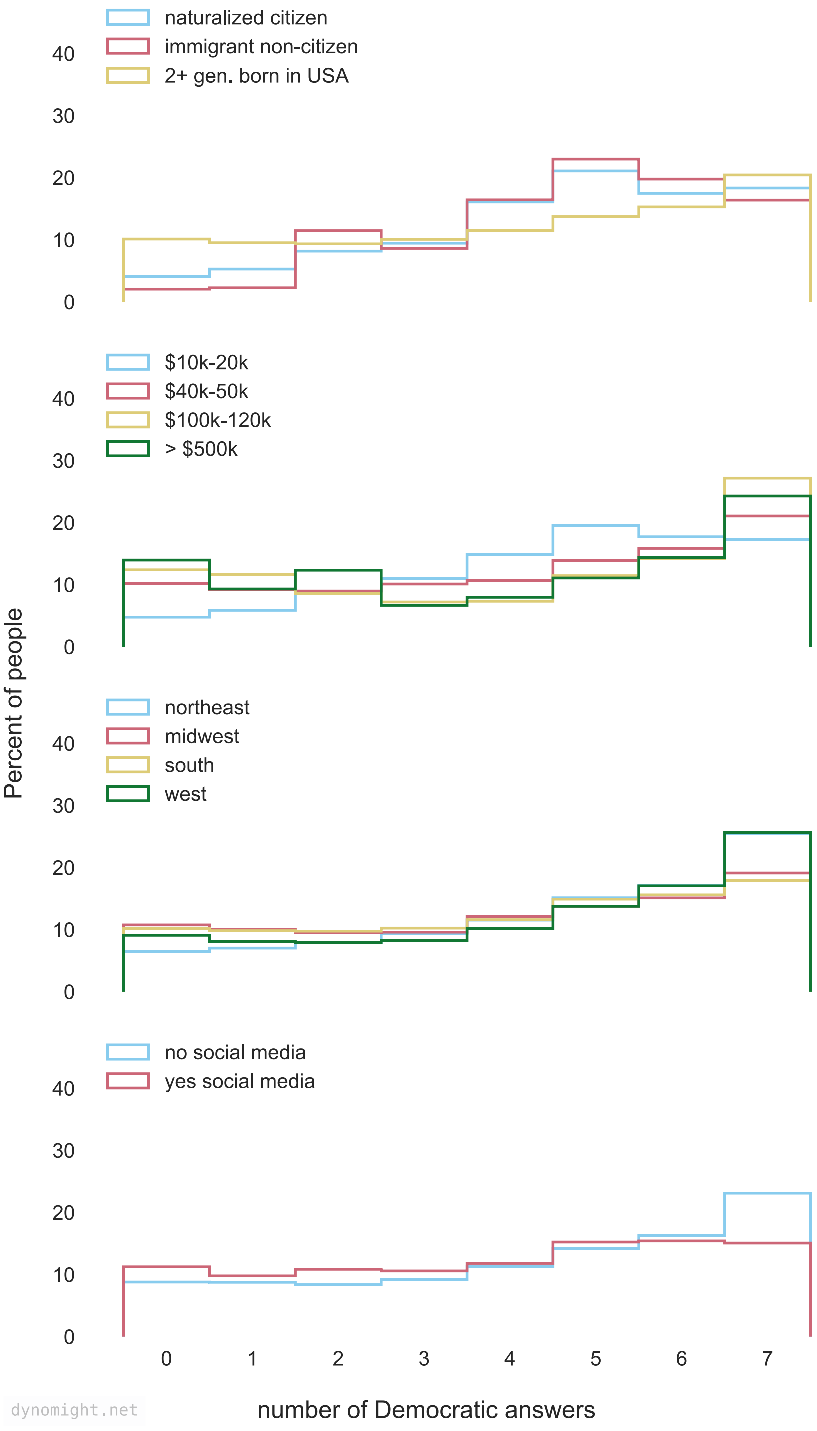
Things get more informative if we calculate histograms for different groups of people.


independent
democratic
republican
registered to vote
not registered
high school
4 year degree
postgrad
white
black
hispanic/latino
Percent of people
0
10
20
30
40
0
1
2
3
4
5
6
7
number of Democratic answers
dynomight.net

This contains a ton of information.

- **First panel**: There's lots of overlap between the distribution of Republicans and Democrats. And independents aren't necessarily moderates! Their distribution looks much like the overall population. Plenty of them hold all Democratic or all Republican positions.
- **Second panel**: People who aren't registered to vote tend to support a mix of Democratic and Republican positions. People who are registered are more polarized.
- **Third panel**: People with more education are more likely to have all-Democratic opinions, and less likely to have heterodox positions.
- **Fourth panel**: Blacks mostly have mostly-Democratic opinions, whites are more variable, and hispanics are in the middle.

We can do the same thing for the other groups. Like above, these groupings don't tend to show particularly strong results. But remember, the lack of a result, is a result in and of itself.


naturalized citizen
immigrant non-citizen
2+ gen. born in USA
$10k-20k
$40k-50k
$100k-120k
> $500k
northeast
midwest
south
west
no social media
yes social media
Percent of people
0
10
20
30
40
0
1
2
3
4
5
6
7
number of Democratic answers
dynomight.net

On Caring

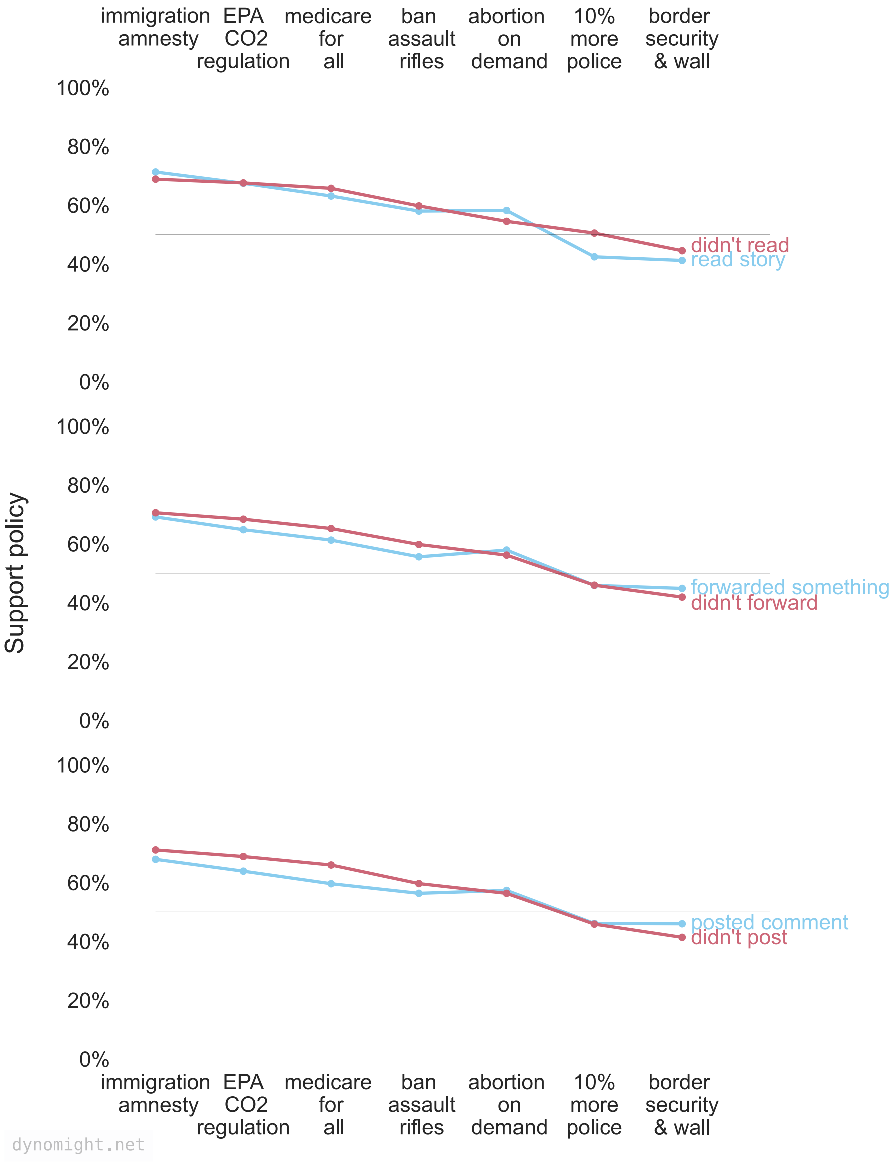
People aren't as polarized as I would have expected. Only around 30% of people hold "all-Democratic" or "all-Republican" positions. And there are reasonable numbers of people who identify as Democrats yet support nearly all Republican positions, or vice-versa.

However, *registered voters* are a lot more polarized than those not registered. This seems important! There are two possible explanations:

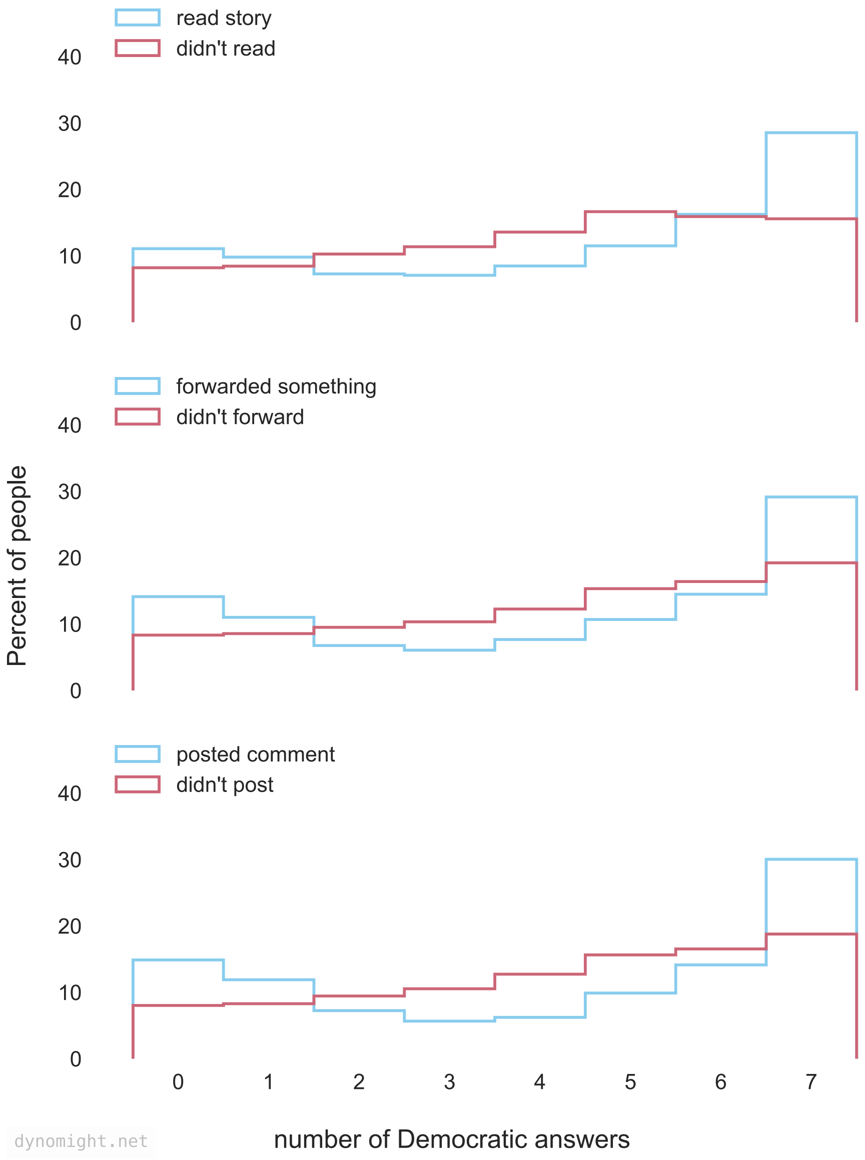
1. Maybe people decide not to vote if their opinions don't align well with a party. Why bother if you dislike all the options?
2. Maybe people who care more tend to have more extreme positions.

My instinct was that the second hypothesis was more correct. To test this, I looked for other questions that measure how "engaged" people are—did they read/forward/comment about politics on social media in the last 24 hours? (The "didn't" groups include people who didn't use social media at all.)

If you look at individual issues, you don't see much difference between these groups:



Yet, if you look at *polarization*, the effect is clear:



If we could separate people who read *multiple* stories, and made *multiple* comments, the trend would surely continue. People who engage with politics online are like everyone else, except more so. This is probably worth keeping in mind when you read about politics online. You can't hear everyone who isn't yelling.

Statistical nihilism and culture-war island hopping

The Guadalcanal campaign was the first major offensive operation by the Allies in the Pacific theater of World War 2. This nightmarish battle ran for six months and—while an Allied victory—involved losses so high the US Navy refused to release casualty figures for years afterward.

When this campaign ended in 1943, Japan's major forward base in the South Pacific was Rabaul, a town Japan had seized from Australia in 1942 and fortified with aircraft, ships, structural defenses, and over 100,000 troops. General MacArthur wanted a direct attack on Rabaul. However, the US joint chiefs would not shift resources away from the war in Europe and likely feared horrendous losses.

Eventually, the Allies settled on a strategy of *island-hopping*: With air and naval supremacy, it was decided to simply avoid heavily fortified areas like Rabaul. Instead, the Allies took only the fewest and most weakly defended points necessary to establish supply lines towards the Japanese mainland.

So, here's a theory: If you want to make the world better, culture-war debates are often like attacking Rabaul.

- Statistical nihilism
- We argue a lot

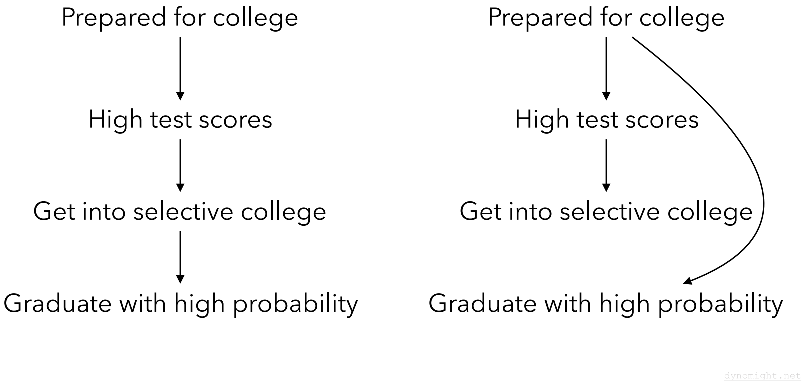
- Culture war island-hopping
- Control air and sea
- Counterarguments
- TLDR

Statistical nihilism

I spend a lot of time sort of arguing that nothing means anything. For example:

1. There are lots of ratio arguments about police bias. Some claim the number of people killed per capita proves the police are racially biased. The problem is that in current society, races vary in basically every dimension: Age, income, where they live, etc. These would produce racial discrepancies, even if race *itself* were invisible. Other people claim the number of people killed per violent crime arrest proves police *aren't* biased. This doesn't work either because the number of arrests could itself be biased.

2. Some claim that standardized tests are useless for predicting success in college. The argument is that if you already know grades, test scores aren't predictive of finishing a college degree. The problem is that this is based on a regression that conditions on *if* a student goes to college at all and *what college* they enroll in. But those colleges—at least when the data was collected—used test scores to decide which students to accept! And more selective colleges have higher completion rates. Here are two possible causal chains:

   
   Prepared for college
High test scores
Get into selective college
Graduate with high probability
Prepared for college
High test scores
Get into selective college
Graduate with high probability
dynomight.net

   If it's the one on the left, removing test scores from admissions would mean that more students with low scores would get into selective colleges, and thus more of those students would graduate. If it's the one on the right, those students would still graduate at lower rates and see little benefit. Which is it? Very hard to say.

3. The gender-equality paradox is the idea in more gender-equal countries, fewer women participate in STEM. It's debated if this really exists, but if it

did, would that mean women were "intrinsically" less interested in STEM? No. Maybe parents in poorer countries suppress their biases and guide daughters towards high-paying jobs. Maybe it's a product of different access to other non-STEM career paths, or some complex cultural feedback loop. Beyond all that, we should remember that "interest" and "culture" aren't separate things—culture influences what people want!

4. Heritability of human traits is uncomfortable for people across the political spectrum. People on the right worry that if skill at high-paying jobs is heritable, that damages the case for a meritocratic society. People on the left worry that high heritability would imply that inequality is "natural". The thing is, the technical definition of "heritability" is just a ratio. This can have weird consequences, like the heritability of hair color changing in different countries, or heritability being higher than 100%.

   Statements like "intelligence is 60% heritable" just aren't as consequential as people often think. Some suggest that we should just "fix" the technical definition of heritability to better correspond to the everyday meaning of the word, but it's not clear that's possible. Like most words, the way people actually use "heritable" is inconsistent and probably impossible to formalize.

What's the point of all this? Why spend so much effort just to point out that, whatever the subject, there's no implications for what we actually care about? Partly, it's a personal fixation that statistical arguments should be careful and not over-claim. That is perfectly valid! But there's a deeper reason, too.

We argue a lot

We tend to care a lot about culture war questions. Often, we believe we know the truth, perhaps because of personal experience or strong convictions about how the world works. And surely we need to agree a problem *exists* if we want to fix it, right?

Well, how's that going?

Here's what I see. For issues without clear culture-war implications, some semblance of rational discussion is possible. Between January and July 2021, the US House and Senate overwhelmingly passed bills to speed disability benefits for people with ALS, avoid cutting Medicare, allow FDA to use a narrow definition of "active ingredient", treat sesame as a potential allergen, temporarily allow foreign cruise ships to skip stops in Canada, fund bone marrow transplants, prohibit reverse auctions for specialized construction work, and make June 19 a national holiday.

Many perceive Congress to be totally deadlocked, but that's not right. It's just that with rare exceptions (June 19) bipartisan stuff passes quietly and isn't relevant to culture war.

Even well-publicized bills can proceed with some sanity, provided they aren't culture-war relevant. Take the debates about infrastructure spending that happened in Summer 2021. These were quite partisan and well-publicized, yet the discussion seemed to be at least somewhat grounded in the reality that it would be nice to have better roads/rail/ports/internet, but all that stuff costs money.

For issues that become culture-war signifiers, forget it. In politically diverse institutions (the US Senate) everything stops. Even if someone is convinced by an opposing argument, they can't change their position because it's terrible politics to give the other side a "win". Bringing more attention doesn't create a healthy debate, it makes debate impossible.

Or take a politically homogenous institution, like a university. If everyone is "on the same side" on an issue, that often leads people to support *every possible measure* regardless of effectiveness. Witness the attempted censure of Steven Pinker for misdeeds like writing an equivocal review article about the intelligence of Ashkenazi Jews or suggesting it was a bad idea to abolish the police. My best guess for what's happening here is that once there's a "correct" opinion, a sort of moral superiority arms race starts. Eventually, even people who agree with the general opinion can't question any evidence or mention any downsides, lest they look heretical.

As far as I can see, you don't "win" culture war. If you tell the people who disagree with you that they are horrible and evil, they'll probably just dig in. And if you tell your allies that anyone who disagrees is a monster, you won't have a very nuanced internal discussion. We aren't finding any consensus, we're just yelling past each other.

Culture war island-hopping

Many people say that culture war is a waste of time. But isn't that kind of defeatist? Policing is important. Education is important. Because these topics have become polarized, are they now off-limits? What's the plan, to sit around and watch impotently as the set of things we can talk about shrinks down to nothing?

And anyway, suppose you won the culture war. What now?

It would have some value. It might bring attention to a neglected issue. Fine, but it wouldn't tell us what we should now *do*. For example, say you proved that women and men have equal intrinsic interest in STEM (however you define that). That suggests something be done to equalize things, but what?

The alternative to culture war is to focus on *interventions*. This has three advantages.

The first advantage of interventions is that we have *good methods to arrive at truth*. In the recent paper Systemic Discrimination Among Large US Employers, researchers sent out 83,000 fake resumes, which were equivalent except for having stereotypically black or white names. They found that resumes with "black" names were called back 23% of the time, compared to 25% of the time with "white" names.

This is about as culture-war of a topic as it gets but, as far as I can tell, it caused little controversy. Partly, I suspect that's because the magnitude of the result is neither large nor small enough to make a good headline. But is it naive to think that the lack of quibbling is partly because there isn't much to quibble about? The researchers changed names and nothing else—they intervened! This makes it hard to argue with the conclusion.

People, in general, are terrible at finding flaws in research that reinforces their views but excellent at spotting flaws in research that challenges those views. When two sides are locked in a trench-warfare "my research says A" vs "my research says B" kind of thing, that's often because the research being cited on both sides is flawed.

A second advantage is that *interventions help us focus on tradeoffs.* Engineers know that changes to complex systems usually have many effects. In a car, making an engine bigger might increase acceleration but decrease fuel efficiency. In human bodies, genes for the Asian flush seem to decrease alcoholism but make alcohol more carcinogenic.

Society is a very complex system. Creating civilian review boards might reduce police use of force. That would be good, but maybe police do their job less aggressively and crime goes up. Or maybe police don't like being reviewed and you have to raise salaries to keep enough employed. I personally suspect the tradeoff is positive, but that's an empirical matter, not something I can just assert as obvious. A conversation about "are police biased?" crowds out these questions.

A third advantage of interventions is that it's *easier to convince someone* about them, and you don't need to resolve culture-war to do it. Consider policing again. Here are some of the policy suggestions I've seen:

- Increase / decrease the number of police officers.
- Increase / decrease prison sentences for certain crimes.
- Ban shooting at moving cars.
- Hire more mental-health workers to work in collaboration with police.
- Eliminate qualified immunity.
- Make more records on use of force public.
- Mandate implicit bias training.
- Require officers to de-escalate or exhaust alternatives before resorting to force.

Say you have evidence that banning chokeholds reduces the number of people killed by police without causing any increase in crime. A hardcore police supporter might well agree to that policy if it's phrased as an "evidence-based policy tune-up". If it's phrased as a win for "Team Police Are Bad", not so much.

If we can resist the urge to talk about grander themes and stick to the effects of things we can actually *do*, we'd have a better chance of making progress.

Control air and sea

As well as increasing focus on interventions, it would be good to *decrease* focus on intractable culture-war debates.

What happens today when someone makes a culture-war claim based on sloppy reasoning and dodgy statistics?

1. Ideological allies make broad supportive comments.
2. Ideological opponents are dismissive and fussy about the details.

And what happens when someone makes a culture-war claim based on careful reasoning and rigorous statistics?

1. Ideological allies make broad supportive comments.
2. Ideological opponents are dismissive and fussy about the details.

It's indistinguishable. There needs to be some incentive against ideologically convenient theories built on mirages of evidence. We need more voices who focus on the strength of claims rather than how good the conclusions are.

To some degree, fact-checkers play this role, but they are supposed to check *facts*, things that are either right or wrong. If someone is lying, by all means call it a lie. But with culture war, it's often that someone makes a subtle statistical error or overgeneralizes from an unrepresentative dataset or something. I don't think it's helpful to award "Pinocchios" in these cases.

A more gentle approach would be more effective. Treat all arguments with the same skeptical eye. Point out the flaw in the argument, explain why the broad question is so hard to resolve, and then turn the debate towards questions where there *are* answers: those about interventions.

Counterarguments

I see two major counterarguments. The first is that I'm assuming you view policies in consequentialist terms, where what matters is effects. If you truly think that what's important in policies is the spirit behind them, then I reluctantly concede that culture war might be the right choice for you. And for some issues—say creating a new national holiday—maybe symbolism is the whole point.

A second counterargument is that maybe "nature finds a way". My assumption is that culture war is intractable because it focuses on controversial questions that are beyond the reach of evidence. The other possibility is that culture war is intractable because… we just love intractable arguments. Maybe it's human nature to see what's in the news and then line up for trench warfare. The drive towards tribalism and group cohesion seems stronger than the drive towards cold-blooded analysis.

This second counterargument worries me. One response is that, sure, maybe most people won't do this, but *you* can, and if it's a good idea you'll help on the margin.

But is it a good idea? If everyone focused on the effects of interventions, would we get people across the political spectrum high-fiving in the streets and cheering the benefits of fixing particulate levels in subways? Or would we get "You don't want the Fed to do inflation targeting—what are you, *a fascist?*"

I think this would happen to some degree. But the discourse just isn't big enough to support polarized opinions about hundreds of policies. Already today, both Republicans and Democrats give idiosyncratic answers when asked about specific issues. Maybe some high-profile issues become super polarized, but I doubt it could happen everywhere.

TLDR

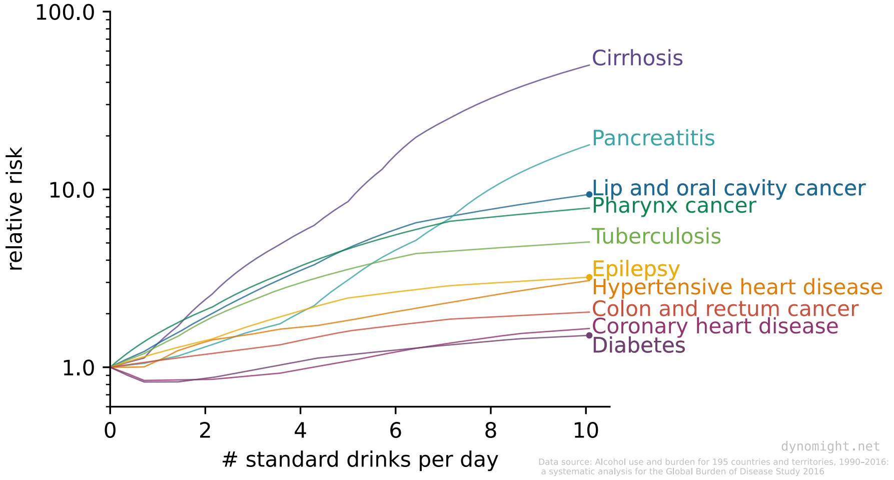
Broad culture war questions are irresolvable mind-killers, but we can't simply give up rational discussion of huge realms of human life. Statistics might not be able to answer those questions. What is *can* do is find the effect of interventions: What happens after you do something? We should skip the broad culture-war, and ask "if we push this policy lever, what are the effects?" At the same time, we should try to deflate the culture-war by skeptically evaluating all broad claims. Studying interventions is less controversial, more grounded in evidence, and must happen eventually anyway. Directly fighting the culture war is like a frontal assault on Rabaul.

The big alcohol study that didn't happen: My primal scream of rage

What does drinking do to your health? We can say two things with confidence:

1. Drinking is associated with lots of health problems.
2. Heavy drinking is bad for you.

Here's a graph of some associations.



dynomight.net

Data source: Alcohol use and burden for 195 countries and territories, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016

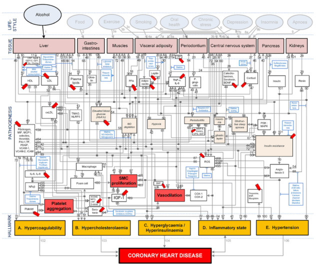
Someone who averages 10 drinks per day is 50x more likely to get cirrhosis than someone who doesn't drink at all (controlling for age, sex, and drinking history). This looks bad, but there are two caveats.

First, it doesn't establish causality. It *could* be—if all you had was this figure—that cirrhosis causes hormonal changes that in turn create the urge to drink more.

But we *do* know that heavy drinking is bad. That's partly because we know *how* alcohol causes problems. It causes cirrhosis by destroying liver cells. It causes cancer by getting converted to acetaldehyde and then damaging DNA. There are also randomized controlled trials (RCTs) that take heavy drinkers and get them to drink less. These inevitably show improved health (either health outcomes or biomarkers like blood pressure).

The second caveat is the little dip for diabetes and heart disease around 1-2 drinks. Some people think alcohol is causing this dip. Lots of mechanisms have been proposed: Maybe it reduces inflammation. Or maybe it impairs the cells that build up plaques in arteries. Or maybe it creates a hormonal imbalance that changes blood pressure regulation. Or maybe it increases HDL-cholesterol or insulin sensitivity or adiponectin levels.

Or, maybe alcohol doesn't help diabetes and heart disease at all. Mathews et al. (2015) try to model how alcohol affects the heart, ending up with this terrifying figure:



Alcohol does a *lot* of different things and interacts with a *lot* of other factors. It's great to try to unravel all this, but I don't trust anyone who says they understand everything with confidence.

If alcohol doesn't improve heart health, then why the dip? Well, it could just be that the same people who drink moderately also tend to exercise and eat well.

So we don't know if moderate drinking is bad for you. It almost certainly causes harms like cancer, but it might help heart disease enough to offset those harms. In the US, around 20% of adults drink 1-2 drinks per day. Even if the effects are modest, the collective impact is huge. Second perhaps to caffeine, alcohol is humanity's favorite drug. We need to know what it does.

This is the story of a trial that came close to answering this question and then exploded. At first, this looks like a simple story of corruption but when you look closely it's a *very complicated* story of corruption.

- We need an RCT
- A solution
- Skepticism
- A defense of the main characters
- Rage

We need an RCT

You might be thinking, "what we need to do is compare the health of people who drink different amounts, while controlling for income, diet, education, exercise, etc." The problem is that to a first approximation, "controlling" for things doesn't work. It requires tons of different assumptions, like what you control for, how you code stuff, and how you model everything. Reasonable people can disagree about those assumptions. For alcohol, reasonable people *do* disagree, and so they get estimates that are all over the place.

So what do we do? We take the long, slow, hard path:

1. Get a large group of people.
2. Tell some of them to drink moderately, tell the others not to drink at all.
3. Wait years, monitoring people to make sure they are actually drinking (or not) like they're supposed to.

4. Follow up and see which group is healthier.

Lots of things make this hard. Because the expected effects aren't huge, you need a *large* group of people. Because culture and genetics vary, you need people from around the world. Because diseases take a long time to show up, you need to wait years. And imagine the challenge of telling people how much to drink and then making sure they follow your instructions.

An international effort monitoring thousands of people around the world for years—does that sound expensive?

A solution

Back around 2013, the NIH's National Institute on Alcohol Abuse and Alcoholism (NIAAA) got interested in funding this. They figured it would cost on the order of $100 million for the full trial. This doesn't seem crazy given the NIAAA's $500 million annual budget, but the NIAAA has lots of other priorities and didn't feel they had the money.

You know who has a lot of money, though? The alcohol industry. Worldwide, $100 million of booze is sold every 30 minutes. In principle, the industry could directly fund a study, but who would trust it?

In 2016, it looked like the NIAAA had found an elegant solution:

- Five alcohol companies would donate money for a trial.
- The NIH would ask researchers to send proposals for how they'd run a trial.
- The NIH would choose the scientifically best proposal, just like they do with any government-funded grant. The donors would have no influence on the process.
- The make the results trustworthy, there would be a "firewall", with no communication between the industry and the research team.

Sounds promising. But if we go forward a couple of years, everything suddenly blows up.

June 15, 2018

NIH cancels $100 million study of moderate drinking as inescapably compromised

What happened? You might imagine banal corruption, with cocaine and overseas bank accounts, but it's nothing like that.

The real story is a much more interesting cocktail of science, academia, bureaucratic maneuvering, ambition, politics, capitalism, the “deep state”, secret emails, and slippery ethical slopes. It’s particularly interesting because it’s a huge stroke of luck that we know about any of this. You have to ask how often similar things happen and *don’t* blow up.

If you’re brave, you can read the 165-page report the NIH prepared before canceling the program. But I warn you: It’s mostly out-of-order redacted emails written by people who wanted to conceal what was happening. There’s an executive summary, but it’s written in a frustratingly “government” style. There are also newspaper stories, but they don’t try to give the full timeline.

After spending way too much time reconstructing things, here’s the full story as best as I can tell.

(If you want an even-more-obsessive amount of information about the timeline, you can click on (more) after each of the sections.)

2001 - 2013. Kenneth Mukamal, a physician at Beth Israel medical center and faculty member at Harvard Medical School, publishes many papers that argue that moderate alcohol consumption has health benefits, usually for heart disease or diabetes. During the same period, John Krystal, a psychiatrist and professor at Yale publishes many papers on alcohol, mostly focusing on addiction and mental health. (Many other researchers will be involved in this study, but these two are most prominent.)

(more)

Here’s a small sample of Mukamal’s papers:

> In summary, all of this evidence implicates alcohol consumption rather than lifestyle factors [...] as the primary factor in the lower rates of cardiovascular disease found among moderate drinkers. (2001)
>
> Compared with abstention, consumption of 1 to 6 drinks weekly is associated with a lower risk of incident dementia among older adults. (2003))
>
> In this large cohort study of older adults, there was a lower risk of congestive heart failure associated with moderate drinking compared with abstention. (2006)
>
> There is convincing evidence that light-moderate, non-binge alcohol intake reduces the risk of coronary heart disease. (2009)
>
> In a nationally representative samples of U.S. adults, light and moderate alcohol consumption were inversely associated with cardiovascular disease mortality, even when compared with lifetime abstainers (2010)

> Secondary analysis of mortality from all causes showed lower risk for drinkers compared with non-drinkers (2011)
>
> Long-term moderate alcohol consumption is inversely associated with all-cause and cardiovascular mortality among men who survived a first myocardial infarction. (2012)

I'm not joking about this being a small sample. His list of publications (up to the present day) has **185** hits for "alcohol". I didn't read all them, but after randomly sampling 20 or so, I found that almost all have a positive spin on the health effects of alcohol. The only exception I found was this paper from 2010 that suggests stroke risk goes up while alcohol is in your system. On that paper, Mukamal was the 4th author out of 6. (In academic publications, the people with the most influence on the paper are typically either first (actually did the work) or last (most senior/famous muckety-muck).)

During a similar period, John Krystal also published many papers on alcohol. These mostly focus on various technical issues related to addition, e.g. if the GAD2 gene might contribute to alcoholism. Some have a clearly negative view of alcohol, e.g. one that explores how alcohol dependence has effects that are similar to accelerated aging.

Early 2013. Some NIAAA staff are convinced that moderate drinking is good for you, and an RCT could prove it conclusively enough that doctors might recommend it to patients like they do with aspirin now. (We don't know who these staff were, but Margaret Murray was probably among them.) They have the idea of getting the alcohol industry to fund the study but face two problems. First, the alcohol industry wants lots of details before forking over any cash. Second, the NIAAA isn't allowed to solicit from industry. They try to get around these problems by having outside researchers (including Mukamal and Krystal) meet with industry to give details on how such a trial might work. This creates a dynamic where everyone (the NIAAA, the alcohol industry, Mukamal) wants to coordinate with each other, but maintain a pretense of being isolated. There's lots of scheming about how information should flow to maintain this pretense.

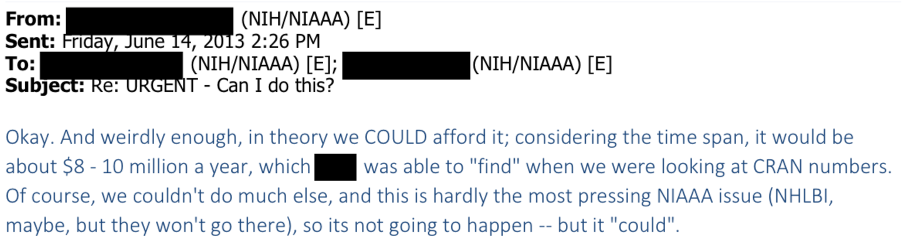
(more)

You can read the full business case the NIAAA put together here. Here's an excerpt:

> Consistent evidence […] has demonstrated that moderate drinking […] lowers one's cardiovascular, metabolic […], and neurodegenerative […] disease risk. […] the benefit is not negated by the potential increases in risk for specific cancers or other illnesses
>
> […]
>
> no government public health entity or scientific/medical professional society has been willing to recommend that patients be advised to

> consider using alcohol as a risk-reduction intervention. [...] there remains a hesitance to be more proactive in the recommendation without a large-scale fully randomized clinical trial

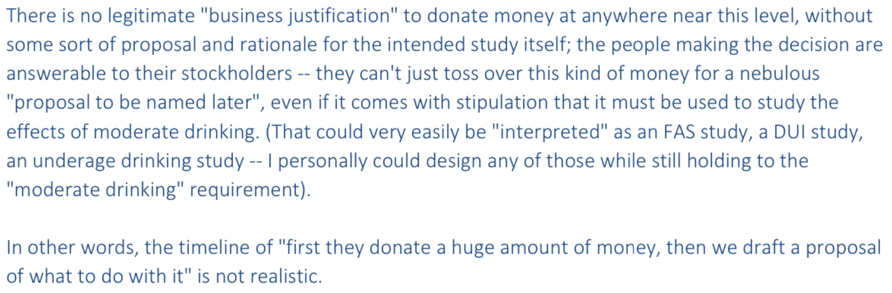
Discussion around that time shows that they didn't want to spend the money that would be required for such a trial.



From: [redacted] (NIH/NIAAA) [E]
Sent: Friday, June 14, 2013 2:26 PM
To: [redacted] (NIH/NIAAA) [E]; [redacted] (NIH/NIAAA) [E]
Subject: Re: URGENT - Can I do this?

Okay. And weirdly enough, in theory we COULD afford it; considering the time span, it would be about $8 - 10 million a year, which [redacted] was able to "find" when we were looking at CRAN numbers. Of course, we couldn't do much else, and this is hardly the most pressing NIAAA issue (NHLBI, maybe, but they won't go there), so its not going to happen -- but it "could".

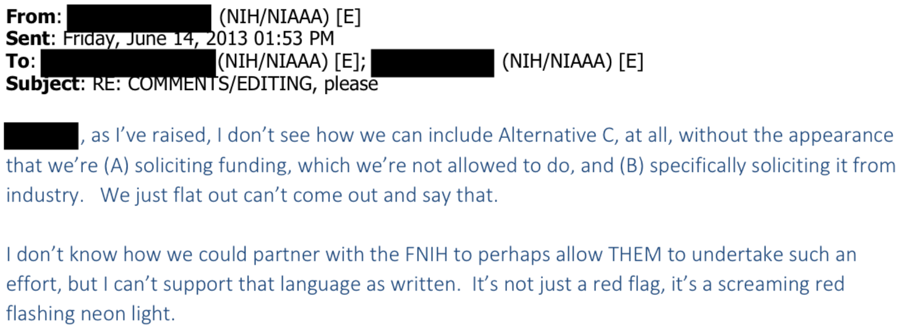
Other discussion shows that they thought they might be able to get the alcohol industry to pay for it. However, they had two concerns. First, they knew the alcohol industry wouldn't be willing to fund a study without some idea of what it would look like.



There is no legitimate "business justification" to donate money at anywhere near this level, without some sort of proposal and rationale for the intended study itself; the people making the decision are answerable to their stockholders -- they can't just toss over this kind of money for a nebulous "proposal to be named later", even if it comes with stipulation that it must be used to study the effects of moderate drinking. (That could very easily be "interpreted" as an FAS study, a DUI study, an underage drinking study -- I personally could design any of those while still holding to the "moderate drinking" requirement).

In other words, the timeline of "first they donate a huge amount of money, then we draft a proposal of what to do with it" is not realistic.

Second, they knew that they weren't allowed to "solicit" funding from the industry, and they were worried that they might be crossing a line.

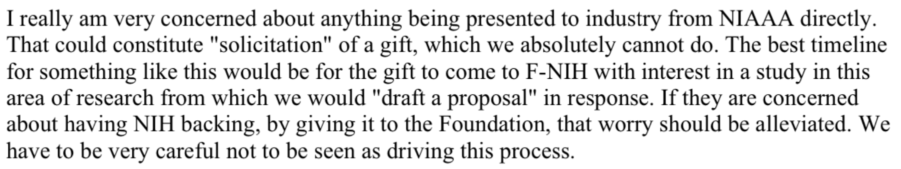


From: [redacted] (NIH/NIAAA) [E]
Sent: Friday, June 14, 2013 01:53 PM
To: [redacted] (NIH/NIAAA) [E]; [redacted] (NIH/NIAAA) [E]
Subject: RE: COMMENTS/EDITING, please

[redacted], as I've raised, I don't see how we can include Alternative C, at all, without the appearance that we're (A) soliciting funding, which we're not allowed to do, and (B) specifically soliciting it from industry. We just flat out can't come out and say that.

I don't know how we could partner with the FNIH to perhaps allow THEM to undertake such an effort, but I can't support that language as written. It's not just a red flag, it's a screaming red flashing neon light.

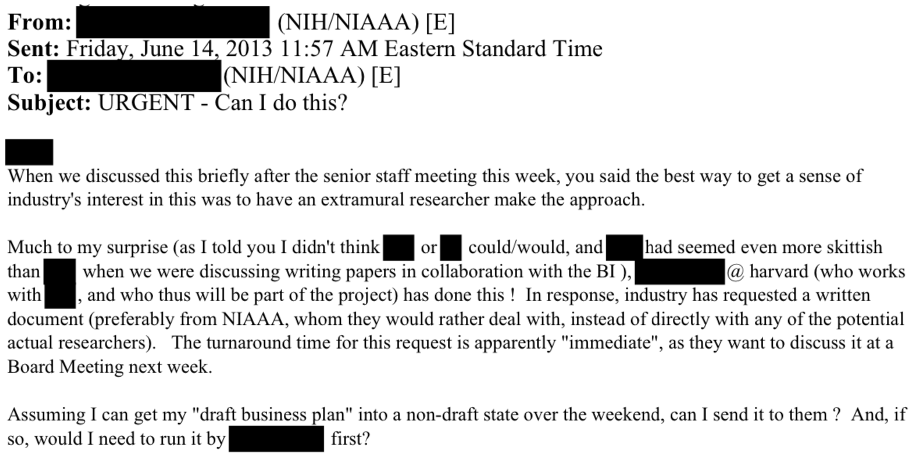
They settled on the strategy of having the industry make a "gift" to FNIH, the arm of the NIH that was set up to take industry money and then do NIH-stuff with it.



I really am very concerned about anything being presented to industry from NIAAA directly. That could constitute "solicitation" of a gift, which we absolutely cannot do. The best timeline for something like this would be for the gift to come to F-NIH with interest in a study in this area of research from which we would "draft a proposal" in response. If they are concerned about having NIH backing, by giving it to the Foundation, that worry should be alleviated. We have to be very careful not to be seen as driving this process.

At the same time, they decided that they could get rid of the appearance of soliciting by getting an external researcher to make the case. They settled on… Kenneth Mukamal. (Here BI appears to refer to Beth Israel, a medical center affiliated with Harvard where Mukamal holds an appointment.)

The record is silent on exactly *why* they chose Mukamal. My guess is that it's partly because of Mukamal's pro-alcohol research record, and partly because it helped to overcome some weird inscrutable issues regarding collaborations between Harvard and Beth Israel.



From: [redacted] (NIH/NIAAA) [E]
Sent: Friday, June 14, 2013 11:57 AM Eastern Standard Time
To: [redacted] (NIH/NIAAA) [E]
Subject: URGENT - Can I do this?

[redacted]

When we discussed this briefly after the senior staff meeting this week, you said the best way to get a sense of industry's interest in this was to have an extramural researcher make the approach.

Much to my surprise (as I told you I didn't think [redacted] or [redacted] could/would, and [redacted] had seemed even more skittish than [redacted] when we were discussing writing papers in collaboration with the BI), [redacted]@ harvard (who works with [redacted], and who thus will be part of the project) has done this ! In response, industry has requested a written document (preferably from NIAAA, whom they would rather deal with, instead of directly with any of the potential actual researchers). The turnaround time for this request is apparently "immediate", as they want to discuss it at a Board Meeting next week.

Assuming I can get my "draft business plan" into a non-draft state over the weekend, can I send it to them ? And, if so, would I need to run it by [redacted] first?

I think you can see the seeds of destruction in the above email. You have these three entities (the NIAAA, the alcohol industry, and outsider researchers) who all want to pretend that they are isolated from each other whilst not actually being isolated. The NIAAA wants someone else to make the approach to overcome their prohibition of solicitation, even though they've obviously set this whole thing in motion. The alcohol industry is excited about what they hear directly from the researcher, but then they want the plan to come "from the NIAAA".

Who were these NIAAA staff? Thanks to all the redaction, for the most part we don't know. We only have two hints. First, we can look at who ended up quoted in news reports. This is George Koob (who wasn't at the NIAAA in 2013) and Margaret Murray. Murray also ended up in the final report written by the researchers on the design of the study, where it's stated that Murray helped develop the initial proposal to the NIAAA.

July 12, 2013. After getting some positive feedback from the industry, NIAAA staff decide to create a "planning grant". This was supposed to be open to anyone, but the staff conspire to steer the money to Mukamal by having a super-short deadline (overruled by NIH-central) requiring pre-approval (also overruled, sort of), and asking for a very specific clinical trial. Two staff go as far as to take fake "personal vacations" to travel to Boston and help Mukamal write the grant. When the window to apply for the grant closes on November 1, Mukamal is the only applicant.

(more)

On July 12, 2013, the NIAAA published a NOT-AA-13-004. By NIH rules, this was a public opportunity, meaning any researcher could submit and win the grant if they had the best science. Yet they obviously wanted "their" PI to win:

> I would be fine with a one-year term; I think the PI can easily meet that, given that we have gone over in a lot of detail what the ultimate RCT should look like; plus that tight a timeframe would discourage other applicants who have not even begun to think about this idea yet !

They stacked the deck in three ways. First, they asked for an extra-short deadline, and said that applications would need to get pre-approval before submitting a grant. Both of these tricks were overruled by NIH central, though "prior consultation" was still "strongly encouraged". Second, rather than a typical open-ended call for research, they asked for a specific trial to be done—coincidentally exactly the trial Mukamal wanted to do. Third, NIAAA staff decided to physically travel to Boston to help Mukamal write the grant. Since this was totally forbidden, they went another way.

> I am going to Boston for a brief "vacation". It would be entirely coincidental if I happened to spend a day with some friends who might be in the process of writing a U34 grant application, and if we also just happened to have some "hypothetical" discussions about details of such a study. This is a purely personal, i.e., NOT NIAAA-funded or authorized, trip.

All the scheming from the NIAAA worked. Ultimately, they received received exactly one application: from Mukamal. There's a complex subplot regarding the review of this grant: There were serious concerns from someone on the NIAAA advisory council, but staff in the NIAAA rebutted them, and then were able to exclude them from voting on procedural grounds. In email, they reassured Mukamal "Do not worry" and that "They are inappropriate comments". In response, Mukamal simply said "here's a draft for the U34. I tried to be discrete (sic) about the industry stuff."

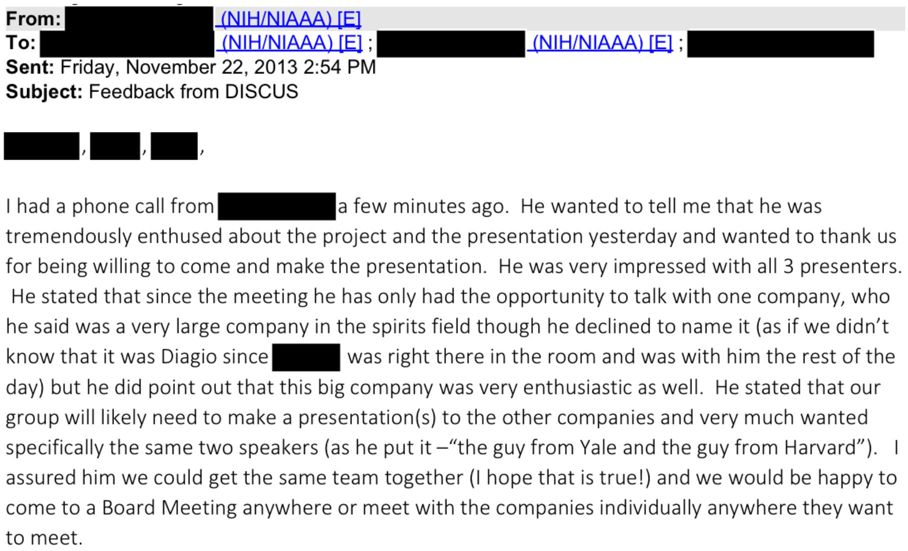
November 21, 2013. There is a meeting at the Distilled Spirits Council in Washington, DC between the alcohol industry, the NIAAA, and three researchers, including Mukamal and Krystal. Someone from industry later reported to NIAAA staff that "he was tremendously enthused about the project" and that they would need similar meetings with other companies. He specifically wanted to hear more from Mukamal and Krystal. There was another meeting at the same location on Jan 28, 2014.

(more)

This meeting took place Distilled Spirits Council's headquarters in Washington DC

Here's an email between NIAAA staff following this meeting. Clearly, the industry liked what they heard. They wanted to hear more from the NIAAA, and specifically said that they wanted two of the same researchers.

Here's a key to help you understand the following email:



From: [redacted] (NIH/NIAAA) [E]
To: [redacted] (NIH/NIAAA) [E] ; [redacted] (NIH/NIAAA) [E] ; [redacted]
Sent: Friday, November 22, 2013 2:54 PM
Subject: Feedback from DISCUS

[redacted], [redacted], [redacted],

I had a phone call from [redacted] a few minutes ago. He wanted to tell me that he was tremendously enthused about the project and the presentation yesterday and wanted to thank us for being willing to come and make the presentation. He was very impressed with all 3 presenters. He stated that since the meeting he has only had the opportunity to talk with one company, who he said was a very large company in the spirits field though he declined to name it (as if we didn't know that it was Diagio since [redacted] was right there in the room and was with him the rest of the day) but he did point out that this big company was very enthusiastic as well. He stated that our group will likely need to make a presentation(s) to the other companies and very much wanted specifically the same two speakers (as he put it –"the guy from Yale and the guy from Harvard"). I assured him we could get the same team together (I hope that is true!) and we would be happy to come to a Board Meeting anywhere or meet with the companies individually anywhere they want to meet.

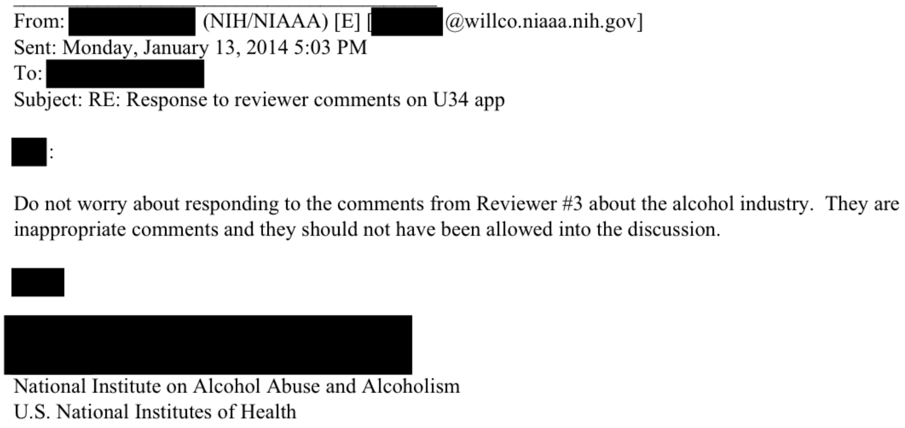
According to the New York Times, representatives of Anheuser-Busch InBev, Heineken and Diageo later confirmed that these meetings were important for their decision to go ahead and fund the trial:

> "When Heineken was invited by the N.I.H. to partially fund the N.I.A.A.A. trial for a duration of ten years, as part of our decision making process, the scientists presented the research project to us so we would have a sound understanding of the trial," Michael Fuchs, a company spokesman, said in an email.

January 2014. The preliminary planning grant is reviewed. One reviewer was concerned about the alcohol industry, but NIAAA staff were able to exclude the reviewer from voting on procedural grounds. When responding to reviewer comments, Mukamal states that he "tried to be discrete [sic] about the industry stuff.” The grant is formally awarded on March 20, 2014.

(more)

There’s a whole complex subplot about the review process for this grant: There was a secondary review from the NIAAA advisory council, who raised concerns about the alcohol industry. Staff in the NIAAA provided a rebuttal to these concerns, and were able to exclude that person’s vote on procedural grounds. In an email, they reassured Mukamal.



From: [redacted] (NIH/NIAAA) [E] [[redacted]@willco.niaaa.nih.gov]
Sent: Monday, January 13, 2014 5:03 PM
To: [redacted]
Subject: RE: Response to reviewer comments on U34 app

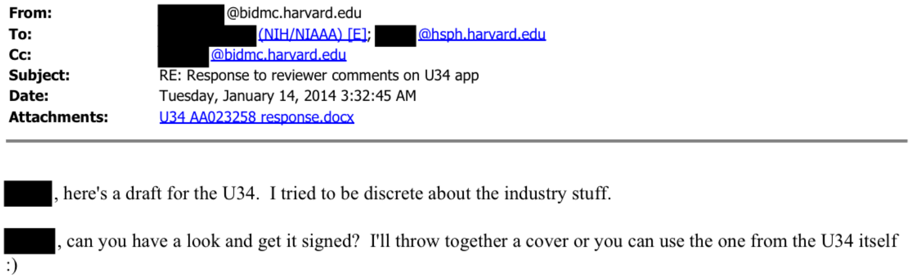
[redacted]:

Do not worry about responding to the comments from Reviewer #3 about the alcohol industry. They are inappropriate comments and they should not have been allowed into the discussion.

[redacted]

[redacted]
National Institute on Alcohol Abuse and Alcoholism
U.S. National Institutes of Health

Mukamal responded as follows



From: [redacted]@bidmc.harvard.edu
To: [redacted](NIH/NIAAA) [E]; [redacted]@hsph.harvard.edu
Cc: [redacted]@bidmc.harvard.edu
Subject: RE: Response to reviewer comments on U34 app
Date: Tuesday, January 14, 2014 3:32:45 AM
Attachments: U34 AA023258 response.docx

[redacted], here's a draft for the U34. I tried to be discrete about the industry stuff.

[redacted], can you have a look and get it signed? I'll throw together a cover or you can use the one from the U34 itself :)

There was a parallel conference grant that was awarded at the same time, also successfully steered to Mukamal.

February 26, 2014. There is a meeting in Palm Beach, Florida, including the alcohol industry, at least one NIAAA staffer, and outside researchers. Mukamal’s slides stated, “A definitive clinical trial represents a unique opportunity to show that moderate alcohol consumption is safe and lowers risk of common diseases.”

(more)

Little seems to be known about this meeting other than that it took place at

The Breakers hotel in Palm Beach Florida. The New York Times appears to have slides from this meeting and gives the following quotes:

> "A definitive clinical trial represents a unique opportunity to show that moderate alcohol consumption is safe and lowers risk of common diseases," said one slide in the scientists' presentation at The Breakers. "That level of evidence is necessary if alcohol is to be recommended as part of a healthy diet."
>
> "We have strong reason to suspect so," said another slide, referring to the large number of studies suggesting that moderate alcohol may be linked to reduced risk of cardiovascular disease.

Since I have no other relevant information, here's a picture of the hotel instead:

February 28, 2014. Wine Industry Insights publishes "US Govt Asking Industry To Fund Most Of $50 Million Alcohol/Health Study", causing a ton of concern inside the NIAAA from people who didn't know what was going on. The people involved openly discuss how to best conceal information.

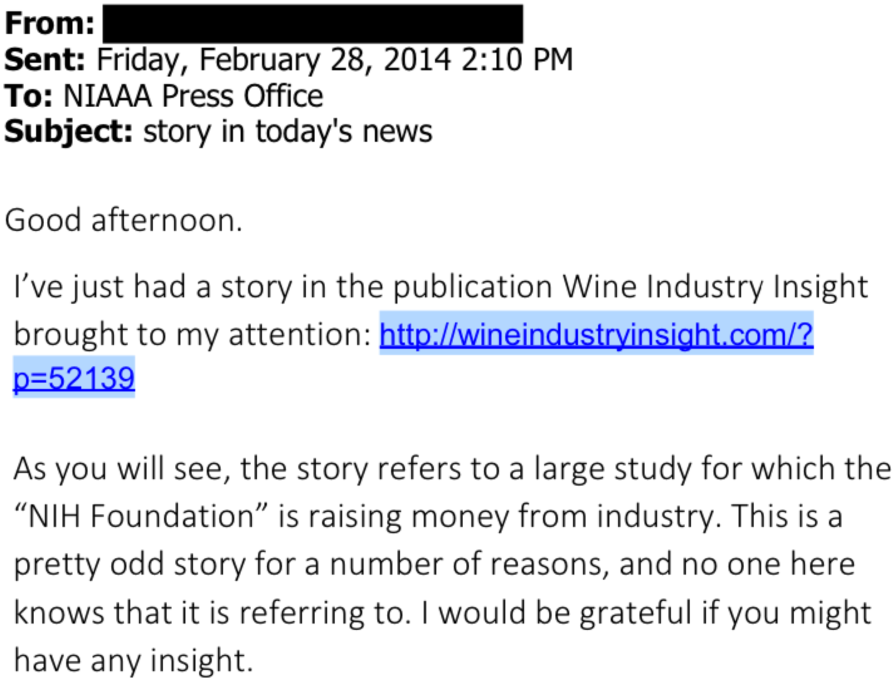
(more)

In early 2014, Wine Industry Insight published an article that said:

> The federal government, along with scientists from Yale and Harvard, are asking wine, beer and spirits organizations to fund a landmark clinical study on the health effects of moderate alcohol consumption estimated to cost $36 million to $54 million.

> […]
>
> “While there are risks in every new endeavor, this study will be a landmark piece of research that should legitimize moderate consumption,” said a member of the DISCUS board of directors, speaking off the record to Wine Industry Insight.
>
> The source added that the only risk involved is that some new negative information might be uncovered. “The evidence is overwhelming that moderate consumers live longer,” the source said. “The risk of discovering negative information is very small given the decades and billions that the government has spent trying to prove the French Paradox wrong.”
>
> […]
>
> The prime movers from the university research sector are [John Krystal] of the Yale University School of Medicine and [Kenneth Mukamal] of the Harvard University Medical School.

This caused a lot of concern within the NIAAA.



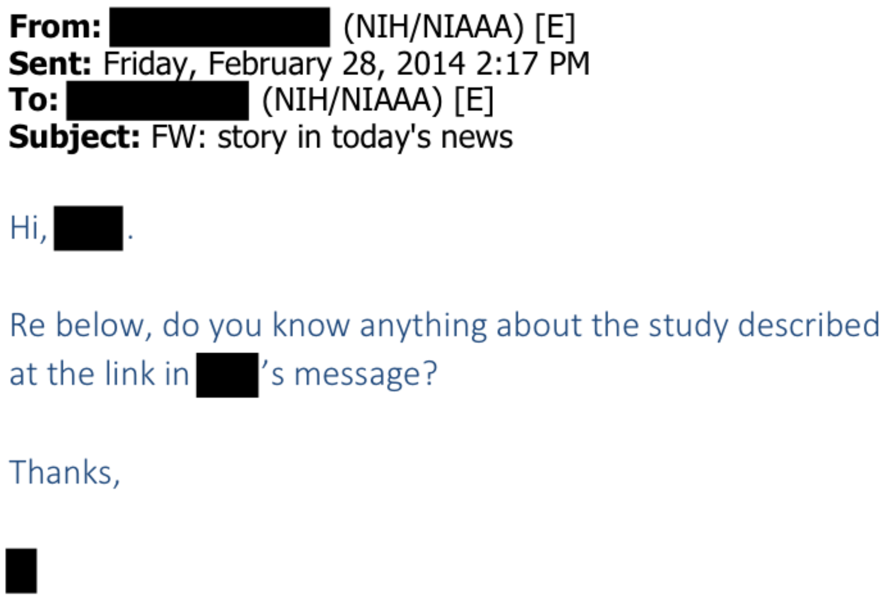
From: [redacted]
Sent: Friday, February 28, 2014 2:10 PM
To: NIAAA Press Office
Subject: story in today's news

Good afternoon.

I’ve just had a story in the publication Wine Industry Insight brought to my attention: http://wineindustryinsight.com/?p=52139

As you will see, the story refers to a large study for which the “NIH Foundation” is raising money from industry. This is a pretty odd story for a number of reasons, and no one here knows that it is referring to. I would be grateful if you might have any insight.

Clearly, information was not being shared very well within the NIAAA. Some people asked what was going on.



From: [redacted] (NIH/NIAAA) [E]
Sent: Friday, February 28, 2014 2:17 PM
To: [redacted] (NIH/NIAAA) [E]
Subject: FW: story in today's news

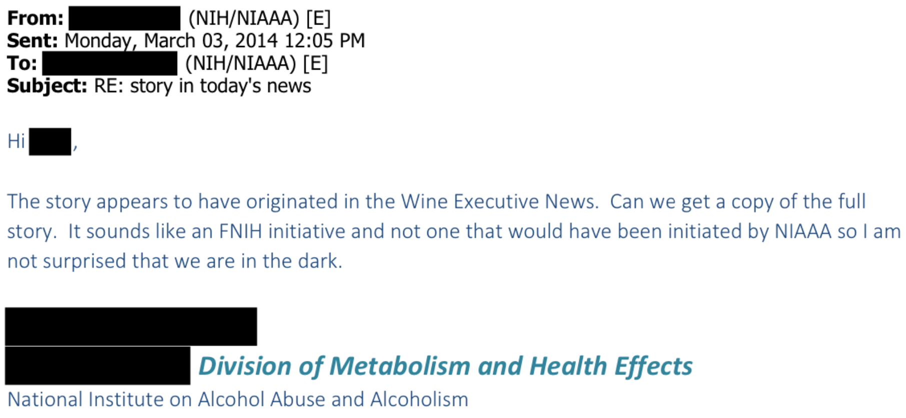
Hi, [redacted].

Re below, do you know anything about the study described at the link in [redacted]'s message?

Thanks,

[redacted]

Someone at the Division of Metabolism and Health Effects naively suggested that it was an FNIH initiative, so it made sense that they had no idea what was going. (Remember, the NIAAA led this from the beginning, and only later decided the FNIH was the easiest way to structure funding.)



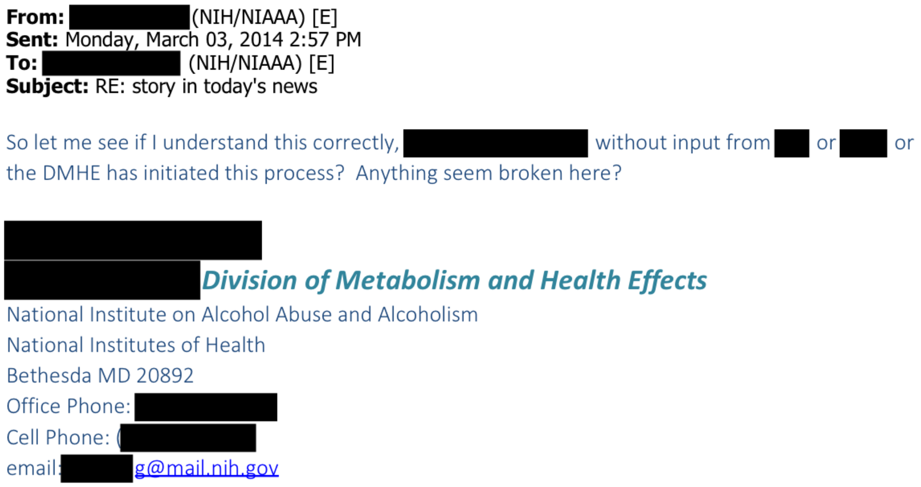
From: [redacted] (NIH/NIAAA) [E]
Sent: Monday, March 03, 2014 12:05 PM
To: [redacted] (NIH/NIAAA) [E]
Subject: RE: story in today's news

Hi [redacted],

The story appears to have originated in the Wine Executive News. Can we get a copy of the full story. It sounds like an FNIH initiative and not one that would have been initiated by NIAAA so I am not surprised that we are in the dark.

[redacted]
[redacted] ***Division of Metabolism and Health Effects***
National Institute on Alcohol Abuse and Alcoholism

Here's an email later the same day that is identified as part of the NIAAA communications office. They are clearly annoyed and/or worried. I think this is the same person as in the previous email (comparing the widths of redacted email addresses and phone numbers and such.)

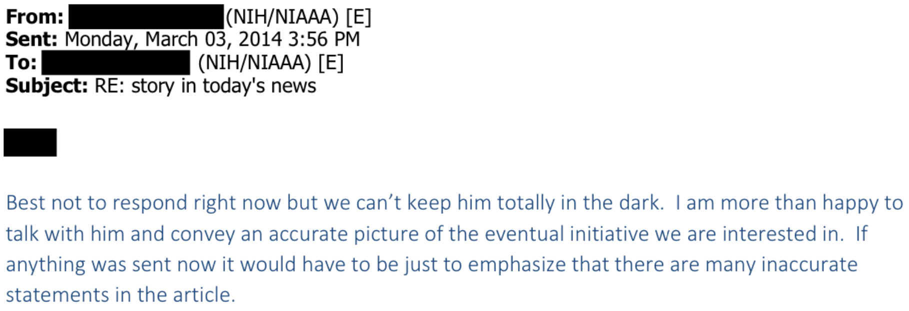


From: [redacted] (NIH/NIAAA) [E]
Sent: Monday, March 03, 2014 2:57 PM
To: [redacted] (NIH/NIAAA) [E]
Subject: RE: story in today's news

So let me see if I understand this correctly, [redacted] without input from [redacted] or [redacted] or the DMHE has initiated this process? Anything seem broken here?

[redacted]
[redacted] *Division of Metabolism and Health Effects*
National Institute on Alcohol Abuse and Alcoholism
National Institutes of Health
Bethesda MD 20892
Office Phone: [redacted]
Cell Phone: ([redacted]
email:[redacted]g@mail.nih.gov

Here's an email from one NIAAA senior staff member to another saying that they should basically conceal as much information as possible. The person they are talking about concealing information from is an NIAAA division director.



From: [redacted] (NIH/NIAAA) [E]
Sent: Monday, March 03, 2014 3:56 PM
To: [redacted] (NIH/NIAAA) [E]
Subject: RE: story in today's news

[redacted]

Best not to respond right now but we can't keep him totally in the dark. I am more than happy to talk with him and convey an accurate picture of the eventual initiative we are interested in. If anything was sent now it would have to be just to emphasize that there are many inaccurate statements in the article.

Eventually the furor about the article all seems to die down, though it's unclear when or if the people wondering what was going on become informed.

June 21, 2014. There's a meeting in Seattle, led by Mukamal, and including NIAAA staff and the alcohol industry. Afterward, representatives from industry send Mukamal a list of technical concerns about the design of the RCT, including what outcomes to measure, the treatment population, adherence, dropouts, monitoring, using beer vs. spirits, and incentives to participate. Mukamal sends back a detailed response, sort of saying "well, this is what *I* would do *if* I happened to win the grant…" and then giving some reasonable answers.

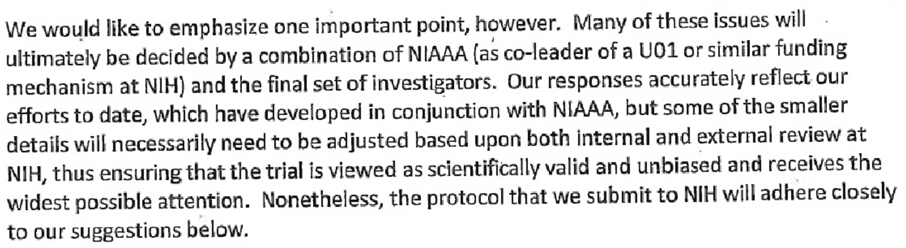
(more)

This meeting took place at the Hyatt Regency in Bellevue, Washington.

Following this meeting in July, representatives from the ICAP and DISCUS (industry groups, ICAP is now IARD), Diago (the world's largest distiller), and AB InBev (the world's largest brewer) sent Mukamal a detailed list of concerns about the design of an RCT. In August, Mukamal gives a 7-page response responding to all concerns in detail.

I can't emphasize enough: In this exchange, industry concerns and Mukamal's response look almost completely non-scandalous. They seem like reasonable concerns from people that are paying a lot of money and want to make sure the trial is well-designed: What outcomes will be measured, who will be eligible to enroll in the trial, would it be better to have fewer sample sites with larger populations, what if patents don't comply with their instructions to (not) drink, what if patents quit the study, what biomarkers will you measure to be sure if people are drinking or not.

The one bit that does seem a little suspicious is this from the beginning of Mukamal's response:



We would like to emphasize one important point, however. Many of these issues will ultimately be decided by a combination of NIAAA (as co-leader of a U01 or similar funding mechanism at NIH) and the final set of investigators. Our responses accurately reflect our efforts to date, which have developed in conjunction with NIAAA, but some of the smaller details will necessarily need to be adjusted based upon both internal and external review at NIH, thus ensuring that the trial is viewed as scientifically valid and unbiased and receives the widest possible attention. Nonetheless, the protocol that we submit to NIH will adhere closely to our suggestions below.

It's not clear how to judge this. Did industry really believe that any investigator could win? Or had the NIAAA winked at them enough that they could be confident who would win?

November-December, 2014. A large joint conference call is coordinated between the alcohol industry, NIAAA staff, and researchers including Mukamal. Here are three topics that industry asks about:

1. Will the data be shared with other researchers? Mukamal states that they would make "controlled data sets" available one year after the study ends.
2. Might industry funding call the study into doubt? Mukamal reassures that it's fine because there will be a "firewall" between research and industry.
3. Will results will be published even if they are negative? Mukamal says yes, but they will "most certainly" see a positive impact at least for diabetes.

(more)

The purpose of this call is "an opportunity to understand more about the protocol that is currently under development". This call takes place on December 8, 2014. Here's some quotes from the minutes of this meeting:

Regarding the overall trial:

> - **What would a trial look like?** - It would be a randomized, multicenter, trial. Individuals interested in the trail will come the field centers and sign a consent form. They will be at high

> cardiovascular risk (so we can conduct this in a 5yr period). They would fall in an intermediate category of drinking a little, but less than daily. They would be randomized to not drink at all or to drink daily for 5 years.

Regarding data sharing:

> - ABI, [redacted] Can you describe the data availability to be shared with other researchers? What happens to the blinded data after the study? And will there be any differentiation btwn wine, beer, and spirits?
>
> [redacted] - We will need to make data available a year after the study concludes, and we can do so in the form of controlled data sets.

Regarding funding:

> - Suntory, ? – Is the funding source going to impact the interpretation of the results by external agencies?
>
>   [redacted] – We will be running the study in conjunction with NIH/NIAAA who is the biggest source of data for the WHO. As long as that firewall is established between industry, and the design/ management of the trial, it should remove doubt.

Regarding possible negative results.

> - Suntory, ? – Is the intention to publish results even if they are less desirable eg. negative or mixed?
> - [redacted] – Yes, however the peer review comments from the initial analysis of our study design were that we will most certainly see an impact for DM and we are not enrolling people of high risk for breast cancer.

Here "DM" is diabetes mellitus. You can read the full minutes for this meeting here. (For fun, try to find where "Ken" slipped through redaction.)

February 26, 2015. Murkamal and NIAAA senior staffers coordinate edits to an email that will be sent to someone in industry. This email states that yes, they really need $100 million, and "one of the important findings will be showing that moderate drinking is safe."

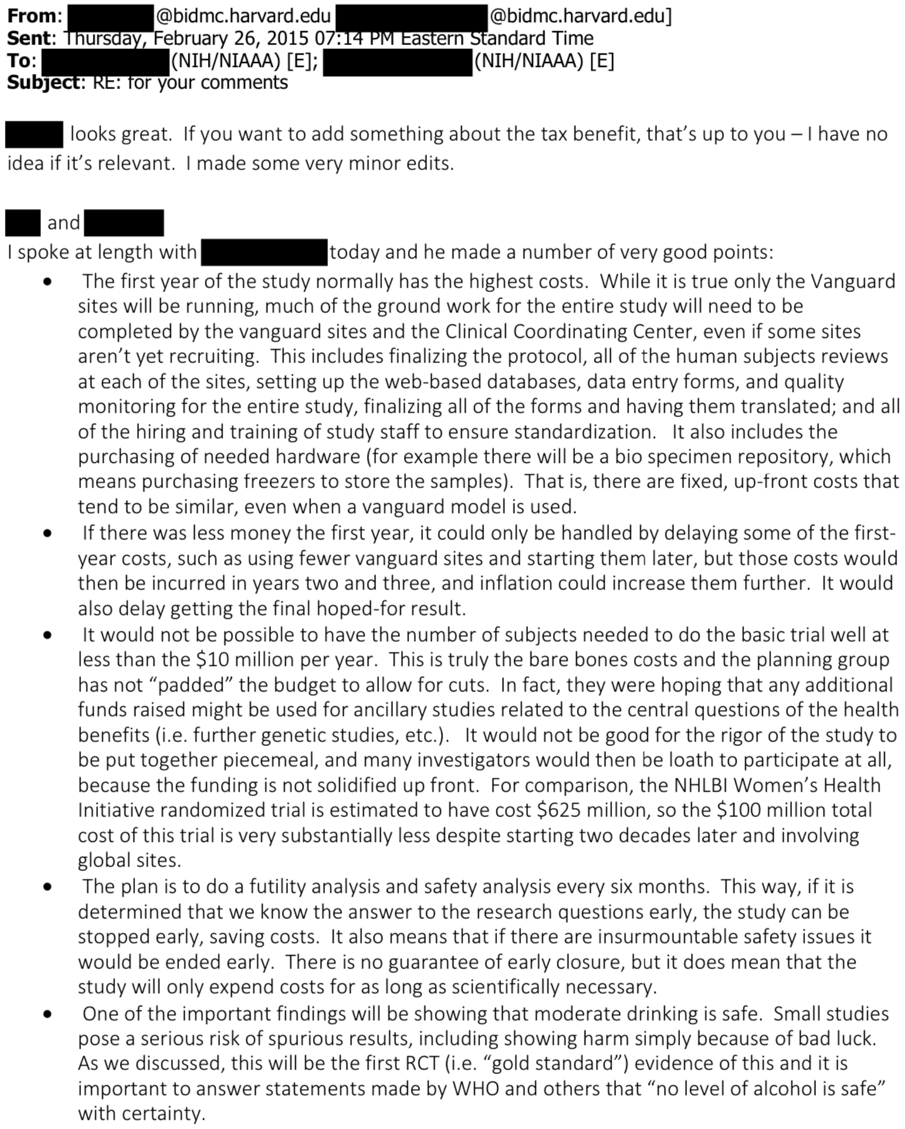
(more)

They are coordinating an email to send to some [redacted person, probably part of ICAP]. Here's the full quote to show that it isn't taken out of context:

> One of the important findings will be showing that moderate drinking is safe. Small studies pose a serious risk of spurious results, including showing harm simply because of bad luck. As we discussed, this will be the first RCT (i.e. "gold standard") evidence of this and

> it is important to answer statements made by WHO and others that "no level of alcohol is safe" with certainty.

The rest of the message is mostly devoted to explaining that yes, they really need the entire 100 million dollars.



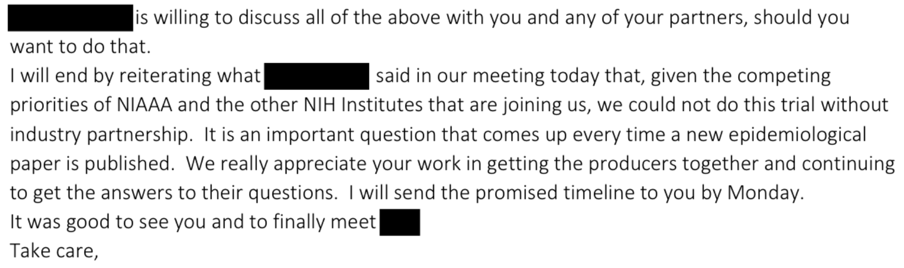
From: [redacted]@bidmc.harvard.edu [redacted]@bidmc.harvard.edu]
Sent: Thursday, February 26, 2015 07:14 PM Eastern Standard Time
To: [redacted](NIH/NIAAA) [E]; [redacted](NIH/NIAAA) [E]
Subject: RE: for your comments

[redacted] looks great. If you want to add something about the tax benefit, that's up to you – I have no idea if it's relevant. I made some very minor edits.

[redacted] and [redacted]

I spoke at length with [redacted]today and he made a number of very good points:

- The first year of the study normally has the highest costs. While it is true only the Vanguard sites will be running, much of the ground work for the entire study will need to be completed by the vanguard sites and the Clinical Coordinating Center, even if some sites aren't yet recruiting. This includes finalizing the protocol, all of the human subjects reviews at each of the sites, setting up the web-based databases, data entry forms, and quality monitoring for the entire study, finalizing all of the forms and having them translated; and all of the hiring and training of study staff to ensure standardization. It also includes the purchasing of needed hardware (for example there will be a bio specimen repository, which means purchasing freezers to store the samples). That is, there are fixed, up-front costs that tend to be similar, even when a vanguard model is used.
- If there was less money the first year, it could only be handled by delaying some of the first-year costs, such as using fewer vanguard sites and starting them later, but those costs would then be incurred in years two and three, and inflation could increase them further. It would also delay getting the final hoped-for result.
- It would not be possible to have the number of subjects needed to do the basic trial well at less than the $10 million per year. This is truly the bare bones costs and the planning group has not "padded" the budget to allow for cuts. In fact, they were hoping that any additional funds raised might be used for ancillary studies related to the central questions of the health benefits (i.e. further genetic studies, etc.). It would not be good for the rigor of the study to be put together piecemeal, and many investigators would then be loath to participate at all, because the funding is not solidified up front. For comparison, the NHLBI Women's Health Initiative randomized trial is estimated to have cost $625 million, so the $100 million total cost of this trial is very substantially less despite starting two decades later and involving global sites.
- The plan is to do a futility analysis and safety analysis every six months. This way, if it is determined that we know the answer to the research questions early, the study can be stopped early, saving costs. It also means that if there are insurmountable safety issues it would be ended early. There is no guarantee of early closure, but it does mean that the study will only expend costs for as long as scientifically necessary.
- One of the important findings will be showing that moderate drinking is safe. Small studies pose a serious risk of spurious results, including showing harm simply because of bad luck. As we discussed, this will be the first RCT (i.e. "gold standard") evidence of this and it is important to answer statements made by WHO and others that "no level of alcohol is safe" with certainty.



[redacted] is willing to discuss all of the above with you and any of your partners, should you want to do that.

I will end by reiterating what [redacted] said in our meeting today that, given the competing priorities of NIAAA and the other NIH Institutes that are joining us, we could not do this trial without industry partnership. It is an important question that comes up every time a new epidemiological paper is published. We really appreciate your work in getting the producers together and continuing to get the answers to their questions. I will send the promised timeline to you by Monday.

It was good to see you and to finally meet [redacted]

Take care,

Oct 5, 2015. The NIAAA publishes the funding opportunity for the big RCT. The published document implies that only someone who won the earlier planning grant—meaning only Mukamal—should apply. In December, Mukamal applies, and in January the opportunity closes without receiving any other applications.

(more)

The funding opportunity is "Multi-Site Randomized Controlled Clinical Trial Research Center on Alcohol's Health Effects", published on October 5, 2015.

Apparently the NIAAA originally requested that this funding opportunity be a limited competition where only people who had won the preliminary planning grant could apply, Mukamal would be eligible. NIH central rejected this, however, the funding opportunity still "encouraged" this with language like the following:

> Applicants for the U10 Clinical Trial Implementation Cooperative Agreement must be able to begin the trial without further planning activities when the U10 is awarded. Therefore, investigators who have already completed planning activities through an NIAAA-funded U34 clinical trial planning grant are expected to apply.

On December 18, Mukamal submits his application. You can read it in its entirety here. It begins like so:

> The health effects of alcohol consumption have been key public health concerns for millennia. Alcohol consumption is highly prevalent, with remarkably little change in prevalence over the last century, and excessive use is a risk factor for innumerable adverse health outcomes, including cognitive impairment, cancer, cardiomyopathy, cirrhosis, gastrointestinal bleeding, trauma, and social devastation. Although the benefit of avoiding alcohol misuse is well-accepted and uncontroversial, the risks and potential benefits of alcohol consumption when consumed within moderation remain unproven. Observational studies document a lower risk of coronary heart disease and diabetes among moderate consumers relative to abstainers, but they also suggest a higher risk of breast and gastrointestinal cancers, and the possibility of residual confounding of these associations by other

> characteristics cannot be excluded. No clinical trial has been conducted to test the hypothesis that moderate alcohol consumption lowers risk of cardiovascular disease or diabetes compared to abstention, yet public policy continues to be made regarding safe limits of drinking. A definitive yet feasible clinical trial investigating whether moderate alcohol consumption lowers cardiovascular and diabetes risk is needed; indeed, it was the foremost recommendation of the NIAAA Expert Panel on Alcohol and Chronic Disease Epidemiology.

On January 12, 2016, the funding opportunity closed. There were no other applications.

March-September 2016. The proposal is reviewed by the NIH, and eventually awarded to Mukamal. The project begins on September 30.

(more)

The panel peer review happened on March 29, 2016, while the advisory council teleconference review happened on April 19, 2016. Little information seems to be publicly available about these reviews. The "memorandum of understanding" was signed on September 16, 2016 and the "cooperative agreement" made on September 30, 2016. The grant was to run from September 30, 2016 to July 31, 2021.

July 3, 2017. The New York Times publishes "Is Alcohol Good for You? An Industry-Backed Study Seeks Answers". This quotes Margaret Murray of the NIAAA as saying that five companies had pledged $67.7 million, and has a lot of general skepticism of the reliability of industry-sponsored research. There's this quote from George Koob, then director of the NIAAA:

> "This study could completely backfire on the alcoholic beverage industry, and they're going to have to live with it," Dr. Koob said. "The money from the Foundation for the N.I.H. has no strings attached. Whoever donates to that fund has no leverage whatsoever — no contribution to the study, no input to the study, no say whatsoever."

There's also this:

> Dr. Mukamal […] said he was not aware that alcohol companies were supporting the trial financially. "This isn't anything other than a good old-fashioned N.I.H. trial," he said. "We have had literally no contact with anyone in the alcohol industry in the planning of this."

October 26, 2017. Wired publishes "A Massive Health Study on Booze, Brought to You by Big Alcohol". Aside from more general skepticism of industry funding research, it also points out that Murray and Koob at the NIAAA seem to have a cozy relationship with the industry. It's got some quotes from a researcher in South Africa that sort of make Mukamal look like a jerk, and finally this:

> Yet when I spoke to Mukamal in February 2017, he said he didn't know about the Foundation's negotiation for industry contributions "until relatively recently." [...] "We have no contact with funders other than NIAAA itself whatsoever," he wrote.

Feb 5, 2018. The trial begins enrolling patients.

March 17, 2018. The New York Times publishes "Federal Agency Courted Alcohol Industry to Fund Study on Benefits of Moderate Drinking". They interviewed former federal officials and used Freedom of Information Act requests to get emails and travel vouchers related to the grant. This story reveals that, contrary to Mukamal's claims, there were various meetings in 2013 and 2014. This includes a "working lunch" at the Beer Institute convention in Philadelphia that's not in my timeline above because I can't figure out when it happened.

March 20, 2018. Based on the previous above article, NIH director Francis Collins orders an investigation into the trial.

April 11, 2018. Collins appears before the House Appropriations Subcommittee on Labor, Health and Human Services to discuss the NIH's budget. When asked about the trial, Collins responds that he is very concerned and is investigating the issue as a matter of priority. (You can watch the video here.)

May 10, 2018. The NIH suspends enrollment in the trial.

June 8, 2018. Anheuser-Bush pulls its funding.

June 15, 2018. Based on a recommendation from an NIH working group, Collins terminates the study.

Skepticism

You might think I'm out of my mind, but it's hard for me to celebrate this trial being canceled. Obviously, lots of inappropriate stuff happened. But when you think about *why* you'd cancel the trial, the arguments aren't as strong as you might think. Here are the arguments I've seen:

The NIAAA and Mukamal lied to the public.

True. They claimed this was just like any other NIH grant, where any researcher could propose a study design, and the NIH would choose the best entirely based on scientific merit. In reality, the NIAAA intentionally steered the money to one pro-alcohol researcher who coordinated the plan with the alcohol industry.

That was bad. But this doesn't *necessarily* imply cancelation, if the study would have been useful. If the point is to punish people, let's not hurt ourselves in the process, right?

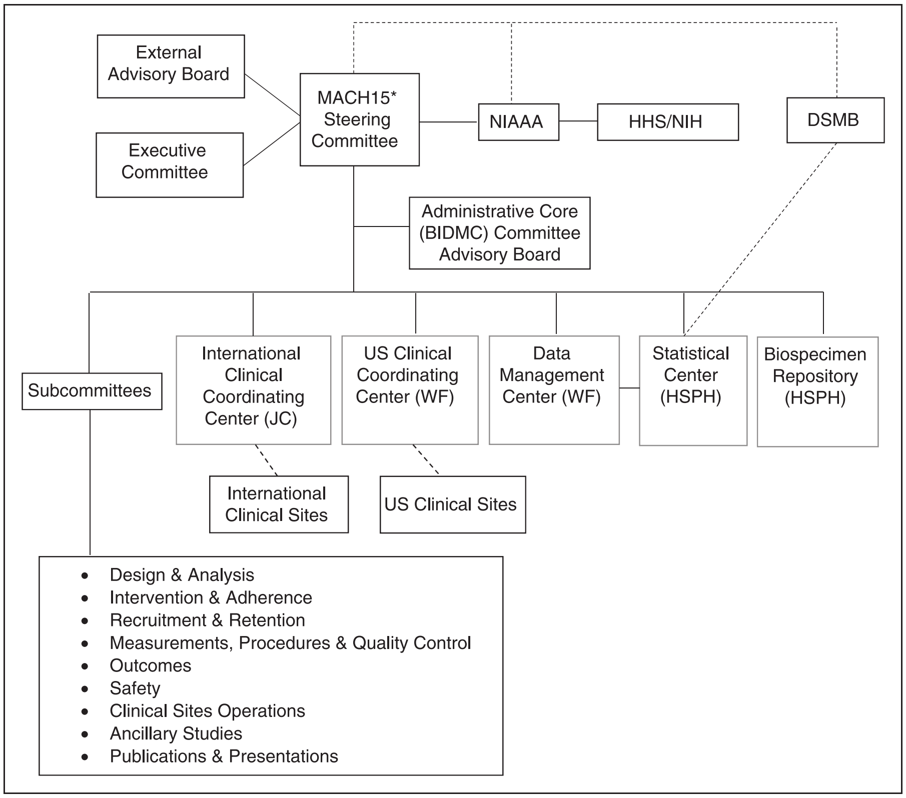
If the study were done, no one would trust the results.

Possibly true, but let's be careful. Are we claiming that no one *should* believe the results, or just that no one *would*? If it's the latter, isn't that kind of a weird reason to cancel a trial? Let's break this down. Why might you not trust the results?

I don't trust the research team.

Clearly, Mukamal *thought* the trial would show a benefit, but that doesn't mean he was right. Anyone who's worked in science knows what it's like to confidently run an experiment, only to get smacked in the face by reality's indifference to your pet theories and career goals.

But OK, say you don't trust the research team. What do you think they are going to do, fabricate data? The study was a collaboration of a large team around the world. The data would be stored at a Data Management Center (whatever that is) at a different university and inspected every six months by a monitoring board. Here's the organizational structure for the study:

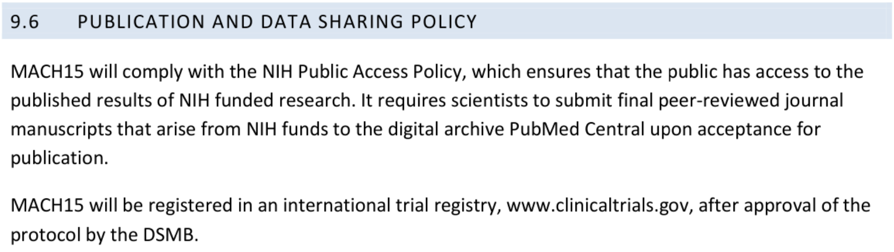

External Advisory Board
Executive Committee
MACH15* Steering Committee
NIAAA
HHS/NIH
DSMB
Administrative Core (BIDMC) Committee Advisory Board
Subcommittees
International Clinical Coordinating Center (JC)
US Clinical Coordinating Center (WF)
Data Management Center (WF)
Statistical Center (HSPH)
Biospecimen Repository (HSPH)
International Clinical Sites
US Clinical Sites
Design & Analysis
Intervention & Adherence
Recruitment & Retention
Measurements, Procedures & Quality Control
Outcomes
Safety
Clinical Sites Operations
Ancillary Studies
Publications & Presentations

This isn't some excel spreadsheet stored on one grad student's laptop. You'd need a big conspiracy.

Or maybe you don't think they'd falsify data, but that for publication they would use some tortured data analysis to spin the results. The thing is, it's not unusual to have researchers who want to find a given result—that's every researcher everywhere! We have a system for this, which is that studies pre-register their statistical analysis. This study did that, and the plan seems fine (although, see below). There just aren't many places to hide the bodies.

If the full data would have been public, that would be another major safeguard against selective data analysis, and made it even harder for anyone to fake things. I can't tell exactly what would have been public. Mukamal mentions making it public when planning the grant with industry. But then the actual grant proposal says nothing about it.

NIH guidelines say that any research costing $1/2 million or more are expected to include a plan for sharing final research data for research purposes, or state why data sharing is not possible. Yet, in the actual grant proposal, here's the entirety of that plan:



9.6 PUBLICATION AND DATA SHARING POLICY

MACH15 will comply with the NIH Public Access Policy, which ensures that the public has access to the published results of NIH funded research. It requires scientists to submit final peer-reviewed journal manuscripts that arise from NIH funds to the digital archive PubMed Central upon acceptance for publication.

MACH15 will be registered in an international trial registry, www.clinicaltrials.gov, after approval of the protocol by the DSMB.

That's it. It's the entire thing. As far as I can tell, there's no mention of making the data public. This is odd, since when he planned the trial with the alcohol industry earlier, he said he'd *have* to make the data public, at least in "controlled" form. Why did this disappear from the final grant? If you had any intention to publish the data, you'd make *absolutely sure* it was in the proposal. That way, if anyone tries to stop you, you can point out that you're committed to it.

I'm not sure what's going on here. Maybe the sharing details were in another document that isn't publicly available? One hint is that at least some of the data would be made public is that for the small amount that was actually collected, some data *is* public today, and can be viewed here on clinicaltrials.gov. It basically says that 32 people were randomized into alcohol or abstention groups and nobody had any adverse events whatsoever.

However, my ever-patient biologist friends point out that there is no obligation to actually put data on clinicaltrials.gov. You can easily confirm this by looking at random studies that most don't seem to bother. I really have no idea if the researchers for the alcohol study were planning to put their data on the site, or just did it after the study blew up since it shows nothing and makes them look (marginally) better.

The study seems designed to deliver a pro-alcohol result.

Two concerns have been made about the study design. For one, it's plausible that the biggest harms of alcohol (e.g. cancer) appear later, while cardiovascular and diabetes benefits (if they exist) happen quickly. So a five-year study might find alcohol reduces mortality while a ten-year study could show the opposite.

Fine, but what's the principle here? Should we cancel all studies where there's a much more expensive and difficult variant that would be more conclusive? We know this is an issue now, and we'd still know it when interpreting results after the study is done.

Another concern is that the study population maximizes the chances for alcohol to look good: It would only enroll people who are either 75 years old or at elevated risk for cardiovascular disease while excluding anyone with liver disease, a personal history of colon/liver/breast cancer, a family history of breast cancer, suicidal ideation, or dementia. If I wanted to maximize the chance that alcohol could be beneficial while minimizing the chance that alcohol could be harmful, this is the population I would choose.

If you want a final verdict on if moderate drinking is safe, I agree this seems like stacking the deck. I'd prefer a random sample of all adults. You can call this a "bias". But you can also call it "refusing to take the sampling scheme into account when interpreting results". There's still value in knowing how alcohol affects a restricted population. And we can extrapolate—a neutral result in this study population would suggest alcohol is harmful to the average person.

You might also argue that it's ethically *required* to exclude people who are at higher risk for being harmed by alcohol. I don't really agree, but I'd imagine many people would.

The premise of the study is flawed: Recent evidence says alcohol is harmful to cardiovascular health.

This was brought up by the extra reviewers brought in to check the scientific merit of the study for the big NSF investigation. Some recent research suggests that alcohol could be *bad* for cardiovascular health. One strategy is "Mendelian randomization": The ADH1B gene (which we've talked about before) makes it hard to metabolize alcohol. People who have it drink less. If you assume that gene is random in the population and that it's *causing* reduced drinking, then you can treat it like a random assignment to drink less. Holmes et al. (2014) did this and found that carriers of ADH1B had better cardiovascular health by every measure. This suggests alcohol makes cardiovascular disease worse, not better. There's also a recent meta-analysis of observational studies by Wood et al. (2018) that suggests that even small amounts of alcohol hurt cardiovascular health.

I don't get this. Is the point that alcohol is *definitely* harmful? That's wrong, the research in the previous paragraph is great, but it isn't conclusive. Or is the point just that an RCT could fail to prove alcohol was helpful? Then… umm… isn't that the entire point of doing the RCT?

The study would be misrepresented.

Imagine that the trial was done and that it showed little overall effect on health. Sure, you might say, *you'll* remember that it used a special population and maybe didn't run long enough to catch cancer. Clever people *like you* will interpret this as meaning that alcohol is probably harmful to the average person.

But do you trust journalists to understand these subtleties and convey them to the general public? Or would we just end up with headlines like "Gold-standard trial shows that moderate alcohol consumption is safe"?

This worries me, but less than you might think. For one thing, don't most people *already* think moderate drinking is safe? The CDC just says not to drink more than 1-2 drinks a day. Tyler Cowen—the Internet's greatest teetotaler—often points out the massive harms of alcohol. Yet he's stated that he believes that by refusing to drink at all, he's sacrificing a small amount of health.

Put that aside, though. Let's make the logic more explicit: This is suggesting that because journalists might do something dumb, we should *not run a trial* that could give knowledge humanity has needed for generations.

Sure, I agree journalists might oversimplify things and confuse people. (Can anyone disagree given recent history?) I just don't think that we can live in fear. We have to believe that once the scientific community has found the truth, it will eventually make its way into public consciousness. The solution to bad journalism is better journalism, not scientists refusing to do research on anything that could be misinterpreted.

It just looks bad.

The final NIH report notes that the researchers do not have "equipoise". You could interpret this two ways. One, you might say the whole thing seems rotten and damn the logic of it. The other is that it looks bad *for the NIH*—that even if useful, it needs to be canceled to preserve trust in the institution. I understand this. But if that's the reason to cancel, it makes me sad.

A defense of the main characters

When I first read about this trial blowing up, I was stupefied—how could everyone have been so shameless? What were they thinking?

Before criticizing people, it's good to try to imagine the strongest defense of their actions. So let me try to do that.

The alcohol industry

I mean, OK, this is an industry entirely devoted to selling an addictive substance that kills, by WHO estimates, three million people per year. Something like 75% of alcohol is sold to raging alcoholics. This isn't a nonprofit organic vegetable farm. But we live in a capitalist system. We expect companies to try to make money, and selling alcohol is *legal*. Let's not conflate this particular trial with general objections to the alcohol industry's existence.

Think about their perspective. The NIAAA *came to them* and said, "We think moderate alcohol consumption is good for you! You should fund a trial to prove this. Win-win for everyone!" The NIAAA sent fancy researchers from fancy places to present to them. Those researchers told them, "I, fancy person, am sure moderate drinking is good! Give me money to prove it!"

The alcohol industry was straightforward they wouldn't fund anything without knowing what would happen in the trial. The NIAAA could have given up at that point, but they bent the rules instead. The industry was worried, "Won't it look bad that we're funding this?" Again, they were told, "Nah, it's fine! There will be a firewall!"

They were told by well-credentialed people that they could make money and do good at the same time. Is it so terrible that they believed them?

NIAAA staff

You might criticize NIAAA staff for becoming convinced that moderate drinking was healthy, even though the science is inconclusive. That's bad, but if you criticize everyone who's wrong about stuff, you're not going to get much sleep.

You can also criticize them for stretching the rules and misleading the public. This is a more clear failing. But imagine you *knew* a study would be valuable, but there's some bureaucratic rule that prevents you from doing it. Wouldn't you be tempted to stretch the rules?

Think about the NIAAA staff who took "personal vacations" to visit Mukamal to help him write the original planning grant. When they did this, I bet they saw themselves as heroes. This is what you see in movies: There's a big problem in the world. People in power *know* there's a problem, but for institutional reasons, it's hard to fix. Most of the people in power are blankfaces, more concerned with covering their asses than helping people. The heroes are the ones who are willing to bend the rules to solve the problem—even if that means taking on personal risks.

If you think that no one in government should bend any rules, then I bet you haven't interacted with the government much. Often, the rules were made by people so removed from what's actually happening that the abstractions in the rules don't even make sense.

Here's an example. Say you're a scientist and you want to send a grant to the National Science Foundation (NSF). According to The Rules, you will propose a detailed plan of *future* work. In some (more theoretical) fields this is absurd:

You have to do half the work in order to write that plan! And in other (less theoretical) fields, your grant will be reviewed by other scientists who will expect to see “preliminary work” to show your idea has promise. This leads to a funny situation where people do much of the research and then “propose” it afterward.

Everyone involved knows that this is happening! The grant reviewers aren’t fooled. The people at NSF aren’t fooled. (Though if they’ve been around for a while, they might not notice the doublethink anymore.) When Congress set up the NSF, they had a mental model of how research works. When that model doesn’t fit, people do the best thing they can: They collectively follow a parallel set of slightly different rules while simultaneously going through the motions of the rules as written. Congress didn’t *mean* to set up a system like this. Bending the rules allows their spirit to be followed as closely as possible.

At the NIAAA, The Rules say that you can’t solicit grants from industry. But what exactly is “soliciting”? You might imagine there’s some oracle somewhere ready to lend definitive answers, but I doubt it. Instead, what you probably see is some people doing things that are a *little* like soliciting, and it’s fine. Other people do things that look slightly more like soliciting, and again it’s fine. Eventually, someone pushes things slightly too far (or is just unlucky) and gets into trouble. The rules get clarified a bit then, but without acknowledging the institutional incentives that made everyone bend the rules in the first place. The person who got in trouble probably feels like a duck shot out of a flock.

So that’s what I guess happened at the NIAAA. The staffers are used to bending the rules because that’s what everyone does all the time because it’s the only way to do anything. They think that the alcohol study would be beneficial, and go for it, and over time things sort of spiral out of control.

Mukamal

There are lots of quotes from Mukamal where he appears to be promising to deliver a positive result. At first glance, these might look like red flags, but I don’t think they’re as bad as they seem.

For one thing, Mukamal didn’t start claiming alcohol was safe as a cynical ploy to get his hands on grant money. He had been publishing on the health effects of alcohol for years. There is no reason to doubt that he sincerely believed that moderate drinking had cardiovascular and diabetes benefits. (And he may well be correct!)

Can Mukamal be trusted? We can look at his track record. In 2007, he was first author on a paper that randomly assigned patients to consume black tea or not. They looked at tons of different biomarkers and found that the tea did… basically nothing. This is the kind of case where it would be easy to p-hack your way to force some conclusion, but they straightforwardly state they found no evidence.

So I don’t think these quotes represent a promise to falsify data but rather his confidence for what the study really would show when honestly performed.

Then there's the lying. Mukamal said there was *no communication* with industry and that he had *no idea* industry funding was even involved. Lying is bad, but still: When Mukamal was describing a "firewall" between industry and research, he was probably thinking of a firewall that started existing sometime after industry committed to funding the study. As far as we know, such a firewall *did actually exist*: Mukamal wrote the final study plan without (further) interference from industry, and the trial would have run without any industry contact.

Would this "late firewall" have meant anything? Maybe so! The biggest question is if industry would have had an opportunity to bury the results if they didn't look good. Maybe the firewall really would have stopped that.

So why did he hide the earlier meetings? Likely, Mukamal felt the public couldn't handle it. Take a look at the first New York Times story on the subject. It is dripping with implications that the study is totally compromised when the only thing known (at the time) was that industry had funded things. It's understandable that Mukamal might have felt that the media was out to get him.

So my guess is that Mukamal was basically a well-intentioned researcher who happened to have pro-alcohol views. He took an opportunity to try to prove his pet theory, and then kind of fell down a slippery slope where he was making gradually larger and larger ethical compromises in pursuit of a goal that he thought was worthy.

Rage

Having written that defense, I'd now like to explain why it's wrong and I'm furious about every aspect of this story.

First, we can only compensate for biases if we know about them. I'm open to industry-funded research. I don't *necessarily* mind a lead researcher who was chosen because they believe what industry likes. I can even live with industry having influence on the study design. I stubbornly hold all this even when the study has a goal of proving it's safe to use humanity's most harmful drug.

But my (possibly delusional) open-mindedness is based on the idea that it's possible to compensate for the biases these issues create. That's not possible if we don't know about them. If you think research still has value despite these issues, fine, but you need to make that argument openly, not pretend the issues don't exist.

Second, the firewall was fake. Say you're OK with a "late firewall" where there's tons of contact with industry early on, but no influence after the trial starts. This didn't happen. How do I know? Well, did you notice the part where Anheuser-Bush pulled its funding? Having the power to shut down the entire trial whenever you want qualifies as *influence*.

Third, slippery slopes aren't much of an excuse. Yes, we all face them, but that's why it's important to have principles—lines you won't cross. If you haven't run into one of those lines before you start lying to the New York Times, something is wrong.

Fourth, many people are complicit in silence. Maybe the alcohol industry really didn't think anything underhanded was happening. Well, they knew on July 7, 2017, when the first New York Times story came out, including untrue or misleading statements from Mukamal and the NIAAA. They had months to correct the record, but they did nothing. The same is true for many of the other researchers involved.

Fifth, the general idea of industry funding with a firewall could be tremendously valuable but was tarnished by everyone here. Take nutritional supplements. Every time someone actually checks, we find out what's in them bears little resemblance to what's on the label (e.g. melatonin off by a factor of 10, or "ginkgo biloba extract" containing *zero* ginko biloba or tons of supplements containing heavy metals.) Some rare companies publish lab tests, but these always seem to be a test of some batch from two years ago by an unknown lab with no reputation who only tests three things and labels them "within spec".

In principle, firewalled research could be the solution. Supplement companies could pay to have tests done by independent researchers. Consumers would have a quality signal for what products to trust, and the companies that make good stuff would make more money. Everyone would win (except the people selling crap products).

This trial has discredited this idea. Obviously, I blame the main characters, but the media is also part of this. Take the first New York Times article again. Remember that when this was written, the firewall was valid, as far as anyone knew. But the article is almost an editorial disguised as journalism. Besides mentioning that the study exists and is funded by industry (which is totally legit) it's largely a collection of whatever random suspicious connections they could dig up between anyone even vaguely connected to the study and the alcohol industry. There are also quotes about how industry funding skews research, but it doesn't address that *that's why* there was supposed to be a firewall.

Obviously, I'm glad the New York Times followed up on this story and revealed holes in the firewall. I just wish there was a more nuanced tone that engaged with the premise that the problems with industry funding are possible to overcome, at least in principle.

Sixth, in the final review, the NIH made no attempt at cost/benefit analysis. Their final report is a fair summary of the problems with the trial. But it doesn't consider the information that was lost by cancellation, or the fact that that there was little cost to taxpayers. (Though Collins' letter to Senator Grassley reveals the NIH did pay around $4 million out of pocket.) Could a different principal investigator be put in charge? Could the study design be modified to address the concerns? Could the monitoring bodies have been strengthened so people could trust the results? Maybe the trial was unsalvageable, but it's telling that the NIH didn't bother to make that argument.

Finally, why have there been so few consequences? Collins says that "three individuals are no longer employed" at the NIH, and they made process changes to avoid similar problems in the future.

That's something, but what about the researchers? To their credit, Harvard and Beth Israel did do an investigation of Mukamal, which led to him formally apologizing and them creating safeguards to make sure no future employees would do anything similar.

Hahahaha, no. Here's what actually happened:

1. Mukamal stated, "We stand fully and forcefully behind the scientific integrity" and "Every design consideration was carefully and deliberately vetted with no input or direction whatsoever from private sponsors." (Yes, these are real quotes from *after* the study was canceled.)
2. As far as we know, there were no investigations by Harvard, Beth Israel, or any of the other researchers' institutions. No one faced any penalty of any kind.
3. In 2020, in what might be the most brazen display of academic shamelessness in history, the researchers published a paper on how awesome the study would have been. Here's a quote from that paper's "sponsorship" section:

   > The Foundation for the National Institutes of Health (FNIH) supported the trial financially and managed contact between public and private organizations on behalf of NIH. The funds provided by FNIH for this project were contributed to FNIH by the brewing and distilling industries following contract negotiations that established an intellectual and financial firewall between MACH15 investigators and private contributors. The corporations providing support agreed to have, and had, no contact with trial investigators about any aspect of the study **after their commitment of funding**, and they agreed to receive no data or updates until they became publicly available. Ultimately, however, the most important safeguard for impartiality lies in the execution of a rigorous, transparent protocol following independent, expert peer review, and in the conduct of the statistical analyses as described in the protocol.

Emphasis mine. You can't make this stuff up.

Two conspiracy theories about cola

Our first conspiracy theory has all the best qualities:

1. It sounds insane.

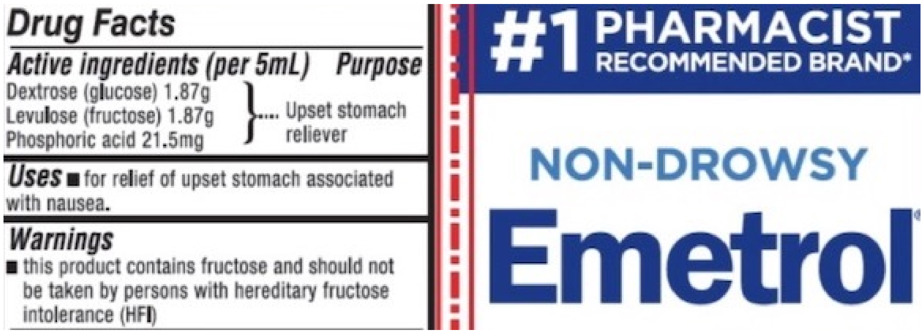
2. At first, the facts seem to support it.
3. Later, the facts lead to disquieting reevaluations of the medical system.

So here's the conspiracy:

> "Cola has so much sugar in it that you'd throw up from drinking it, except they add an anti-vomiting drug to stop that from happening."

Fact #1: Phosphoric acid is the active ingredient in some over-the-counter anti-nausea drugs.

The common brand name here is Emetrol, but generic equivalents are everywhere too.



Drug Facts

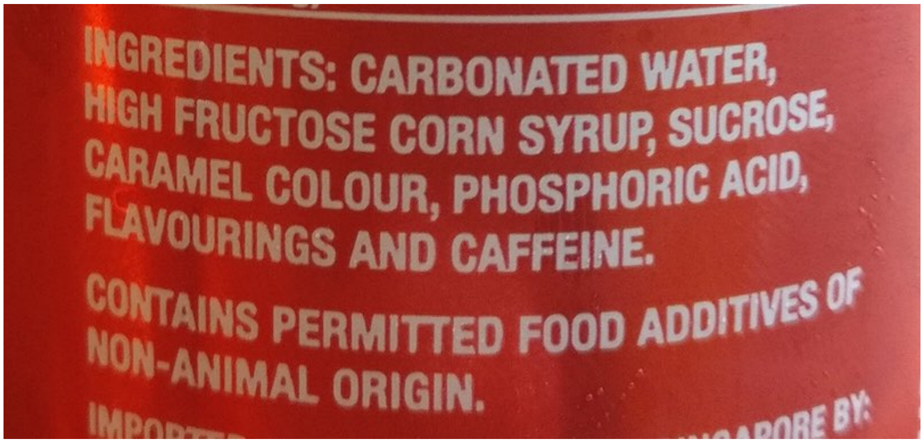
Active ingredients (per 5mL)	Purpose
Dextrose (glucose) 1.87g Levulose (fructose) 1.87g Phosphoric acid 21.5mg	Upset stomach reliever

Uses ■ for relief of upset stomach associated with nausea.

Warnings
■ this product contains fructose and should not be taken by persons with hereditary fructose intolerance (HFI)

Fact #2: Cola contains phosphoric acid.

Nothing controversial here, it's right on the label.



INGREDIENTS: CARBONATED WATER, HIGH FRUCTOSE CORN SYRUP, SUCROSE, CARAMEL COLOUR, PHOSPHORIC ACID, FLAVOURINGS AND CAFFEINE.

CONTAINS PERMITTED FOOD ADDITIVES OF NON-ANIMAL ORIGIN.

IMPOR APORE BY:

My first thought when I saw this was that, OK, cola has phosphoric acid. But surely it has much *less* than the drugs. Well...

Fact #3: Cola has just as much phosphoric acid as anti-nausea drugs do.

The drugs contain 64.6 mg of phosphoric acid in a small dose and twice that in a large dose. This is easy to calculate given the information on the label above and the fact that the recommended dose is 15-30 mL.

Cola *probably* contains something like 200 mg per can. This isn't trivial to figure out. I estimated it in two ways, which I trust only because they give similar answers:

1. Random internet people claim that Coke is around 0.055% phosphoric acid. This suggests that a standard 382 g can has 210 mg.
2. This paper says that a single can of Coke contains 58mg of phosphorus, and this leaflet from Coca-Cola Hellenic gives a similar number. Meanwhile, the chemical formula for phosphoric acid is H_3PO_4. Since the molecular weights for hydrogen (H), phosphorus (P), and oxygen (O), are 1.01, 16.0, and 30.97, each unit of phosphorus should convert to $(1.01\times3 + 30.97 + 16 \times 4)/30.97 = 3.16$ units of phosphoric acid. So a can of Coke should have around 58mg × 3.16 = 183 mg of phosphoric acid.

Dose	Sugar	Phosphoric Acid
Emetrol (small dose)	11.22 g	64.5 mg
Emetrol (large dose)	22.44 g	129 mg
Coke (one can)	39 g	~200 mg

Interestingly, even the ratio of phosphoric acid to sugar is almost the same (around 5.5 mg phosphoric acid / g sugar).

What's the catch? If Emetrol is just a more expensive and worse-tasting cola, why does every pharmacy everywhere sell it? Well, if you think that's disturbing, keep reading.

Fact #4: You wouldn't throw up if the phosphoric acid wasn't there.

You can just try it. A can of cola has 39 grams of sugar (10 teaspoons), and around 350 ml of water. I measured these out into two glasses, which looked like this:

After mixing these up, the result was *very* sweet, but all members of House Dynomight—sometimes after a lot of encouragement—drank some without feeling the slightest bit ill.

Apple juice has just as much sugar as cola. Non-cola soft drinks have citric acid instead of phosphoric acid. (As do some diet colas.) Obviously, people drink these without vomiting. Root beer, it turns out, has neither phosphoric nor citric acid. For this reason root beer has a higher pH than other soft drinks and probably isn't as bad for your teeth.

Fact #5: There's no evidence that these anti-nausea drugs actually do anything.

The Mayo Clinic simply states that the combination of sugar and phosphoric acid "has not been proven to be effective".

Puzzling. Or, you can go to the NIH's information page on different versions of the drug:

All of these state:

> Marketing Status: unapproved drug other
>
> **DISCLAIMER:** This drug has not been found by FDA to be safe and effective, and this labeling has not been approved by FDA. For further information about unapproved drugs, click here.

Don't click there, it's useless. But here's the history: The FDA was created in 1906, but only gained the authority to approve drugs with the 1938 Federal Food, Drug and Cosmetic Act. However, this only gave authority to approve *new* drugs, and only said that they needed to be tested for *safety*. It wasn't until 1962 that the law was amended to require that new drugs be tested for effectiveness.

As far as I can tell, drugs like Emetrol can be sold is because (a) they've been around forever and so are grandfathered in, and (b) phosphoric acid is generally recognized as safe meaning the FDA thinks it's unlikely to be harmful. It's in all sorts of food like almonds, beer, bread, and jams. To the best of my knowledge, there's no evidence at all that these drugs have any effect other than placebo. (Though, of course, absence of evidence is not evidence of absence.)

Fact #6: The phosphoric acid is there for flavor.

On the one hand, the conspiracy is even more true than claimed: Cola doesn't just *contain* an anti-nausea drug, it basically *is* an anti-nausea drug. It's just that this "drug" may not do anything.

So why is it there? Well, we could always ask the people who put it there. The Coca-Cola corporation says they include phosphoric acid in their beverages because:

> "It gives them their tartness."

Conclusion: If there's a conspiracy here, it's not with the cola, it's with the drugs.

Second theory

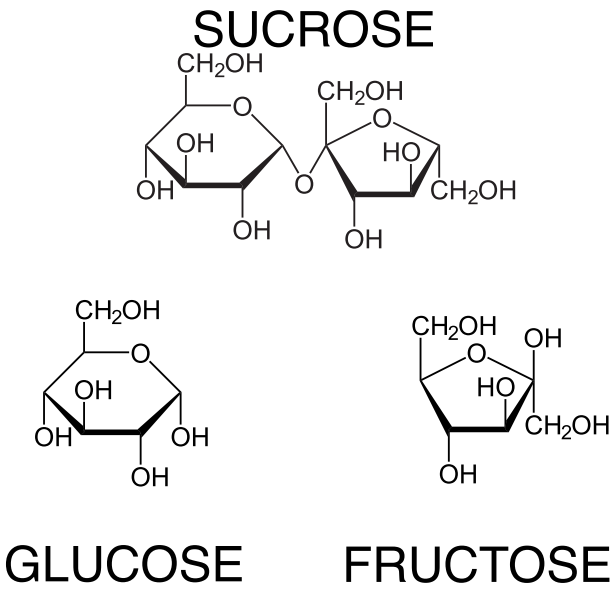
Our other conspiracy theory is more well known:

> "Mexican Coke contains sugar rather than high-fructose corn syrup, and is therefore better, or healthier, or something."

Stores and restaurants in the US charge much more for Mexican Coke. What's going on?

Fact #1: Real sugar is exactly 50% glucose and 50% fructose.

Cane sugar is made of sucrose. One molecule of sucrose is one molecule of fructose glued to one molecule of glucose.



Fact #2: High fructose corn syrup (HFCS) can have basically any ratio of fructose and glucose.

The standard types are HFCS 42 (42% fructose) which is usually used in processed foods like bread and breakfast cereal, and HFCS 55 (55% fructose) which is mostly used for soft drinks. The rest is mostly glucose with traces of other sugars like maltose. However, HFCS can be made with basically *any* ratio, and it's not obvious what's used in any particular product.

Fact #3: Mexican producers of Coke threatened to stop using sugar back in 2013.

In 2002, Mexico created a 20% tax on any beverages not sweetened with cane sugar. American HFCS manufacturers, who have no chill at all, challenged this at the WTO, eventually forcing Mexico to remove the tax and apparently winning $160 million in damages.

In October 2013, Mexico passed a tax on junk food, including a 1 peso ($0.05) per liter tax on soft drinks. In response, one of Coke's bottlers in Mexico announced that they might start using corn syrup instead.

I was confused by this: Could they avoid the tax by using HFCS? As far as I can tell, no, it's just that HFCS was *already* cheaper than cane sugar, and that the tax created an incentive for cost-cutting.

Anyway, this announcement caused Americans who think Mexican Coke is better to freak out. Except, shortly after that announcement, Coca-Cola clarified that Mexican Coke made for export to the US is a “nostalgia product” and would still be made with pure sugar. Except…

Fact #4: Even before 2013, lab tests showed that Mexican Coke had no sucrose.

Ventura et al. (2010) and used liquid chromatography to analyze the contents of various beverages. Here’s what they found (all measurements in g / 100 mL):

Drink	Fructose	Glucose
American Coke	7.2	3.9
Mexican Coke	5.4	5.0

The ratio of fructose and glucose is indeed more even in the Mexican version. But remember how sucrose is *exactly* one molecule of fructose and glucose? Since they have the same molecular weight (it’s the same atoms, just arranged differently) the weight should be exactly the same. That could just be measurement noise, though, so they went ahead and directly checked for sucrose. They found that Mexican Coke contained *zero sucrose*.

Walker et al. (2014) did a similar experiment, except to be *really* sure they used not just liquid chromatography but also metabolomics and gas chromatography. By all three measurements, they found that American Coke contained around 60% fructose, while Mexican Coke was somewhere around 50%. They also directly tested for sucrose and found that Mexican Coke had *some* but just a little (12.5 g/L as opposed to 51 g/L of fructose and 47.9 g/L of glucose). They also found that the Mexican Coke contained around 7% more sugar overall.

Since this second paper was published in July 2014 it almost certainly used Mexican Coke made before October 2013. (You need time for the Coke to be made, sold, experimented on, along with time to write the paper, wait for reviews, and then wait for publication.) Thus, this predates the switchover of Mexican Coke from sugar to HFCS.

I think the most likely explanation for these experiments is sucrose inversion: In the presence of heat or acid, sucrose will naturally break up into fructose and glucose. Likely this happens during the manufacturing of Mexican Coke. Regardless of the source of the sugars in Mexican Coke, we still have these facts:

- The fructose/glucose ratio in Mexican Coke was nearly 50/50.
- The fructose/glucose ratio in American Coke was higher, something like 60/40.

Let’s be charitable to the conspiracy. Let’s ignore that Mexican Coke had 7% more sugar overall, and let’s assume that the Mexican Coke you ~~overpay~~ buy today still has a 50/50 ratio. Is it better?

Fact #5: Fructose and glucose are metabolized differently.

Once glucose is floating around the body, it can be absorbed by basically any random cell. Those cells break it down to provide energy. On the other hand, fructose needs to be metabolized by specific organs. Recent research suggests that small amounts of fructose are metabolized by the gut, with larger amounts being metabolized by the liver.

These differences have all sorts of complex implications that are still being worked out. Whatever the case, it's not crazy to imagine the difference could affect health somehow.

Fact #6: The science on the health effects of glucose vs. fructose is inconclusive.

There are some hypothesized effects: For example, glucose seems to cause an insulin spike, but fructose doesn't. It's unclear if that is good or bad—the insulin spike might make you feel full and stop eating. There are also suggestions that the liver might produce fat when metabolizing glucose.

As far as I can tell, the experts are very much split on the science here. Some seemingly credible people claim that fructose might be worse than glucose, and some claim it makes no difference. Some even suggest that sucrose *itself* is less harmful than a 50/50 mixture of glucose and fructose since the bond needs to be broken. (But remember that the sucrose in Mexican Coke is already broken down.)

I don't know enough to figure out who is right, or which experts are more credible. I think the sensible approach is to just accept the uncertainty. While I emphasize that you shouldn't trust me, here's my completely made-up posterior about the health effects of cane sugar vs. HFCS in the 60 / 40 ratio present in American Coke.

Sugar vs. HFCS	Probability
Sugar much worse	2%
Sugar slightly worse	15%
Same	50%
Sugar slightly better	30%
Sugar much better	3%

A lot of people seem to think cane sugar is healthier because it's more natural. If so, that would suggest that honey would be even better. But did you know that honey also contains more fructose than glucose? It's incoherent to both use "naturalness" and "fructose bad" as justifications for HFCS being worse than sugar, especially since the sugar in soft drinks isn't really "natural" anymore.

If you accept my posterior above, I guess decision theory says you should prefer cane sugar. Fine, but remember, there is overwhelming evidence that all added sugars are harmful. At best, cane sugar is slightly *less bad* than HFCS.

Fact #7: Blind tests do not support the idea that Mexican Coke tastes better.

The internet is awash with people claiming they prefer Mexican coke, but these people are unscientific. Blind tests tell a different story:

- First We Feast did a blind test with six staffers. They did worse than chance at guessing which was American and which was Mexican.
- Science Jon also did a test where testers were given four samples, two of each type. Out of 13 people, only one was able to pair up which samples were the same, and that person couldn't repeat the trick in a second test with five samples.
- Serious Eats did a test that used various conditions (in a glass bottle, in a cup with ice, etc.). They found that people overwhelmingly preferred a bottle to a can. Half of the people had no preference between the Mexican and American Coke, while half of people preferred the American version, and no one preferred the Mexican version. In a second experiment, they presented two American Cokes but lied to people that one was Mexican. People overwhelmingly chose the fake "Mexican" version.
- The Huffington post did a test where 20 editors blindly tasted the two versions. They report that 85% "could tell the difference" and 80% preferred the Mexican Coke. That supports the hypothesis, but they give no details about how blinding was done so I don't completely trust it. Also, they did a similar test with Pepsi and found there was no difference.
- The Miami times did a test where 3/9 people tasted no difference 5/9 preferred the American version, and zero preferred the Mexican version.

Overall, there's no clear evidence that the Mexican version truly tastes better. If it does, it's a placebo from being told it's a special import or maybe just the fact that it comes in adorable little glass bottles.

Conclusion: Mexican Coke doesn't taste better, has no clear health benefits, and doesn't really contain "natural" sugar anyway.

P.S.

All sugary drinks are bad for you, don't drink them.

If you must drink cola, drink *diet* cola, sweetened with aspartame. As far as we know from current science, aspartame has basically zero negative effects. For example, this recent meta-analysis found no significant effects on blood glucose levels, insulin levels, cholesterol, triglycerides, body weight, or energy intake. There was a tiny effect on HDL cholesterol, but that's "good" cholesterol, and the effect was to *increase* it. It's certainly possible there are some downsides that aren't yet established—for example, while current evidence suggests aspartame is not genotoxic, it's still somewhat open. Still, any risks of aspartame are an order of magnitude lower than the huge and indisputable harms of sugar. Water is safer than diet cola, but diet cola is *much* safer than real cola.

Also, remember that all cola (including diet) is bad for your teeth, due to the acidity. If you drink it, try to do so with meals, try to use a straw, and try to drink some water after. Also, *don't* brush your teeth immediately after drinking cola. The enamel on your teeth is vulnerable due to the acidity. You should wait 30-60 minutes for your saliva to repair the enamel before brushing.

The main thing about P2P meth is that there's so much of it

Sam Quinones was recently on Econtalk and in the Atlantic talking about methamphetamines and homelessness. He points out that "old" meth was made from ephedrine and that "new" meth is made from a chemical called Phenylacetone or P2P. He suggests that new meth might be chemically different in a way that caused people to go crazy, starting around 2017:

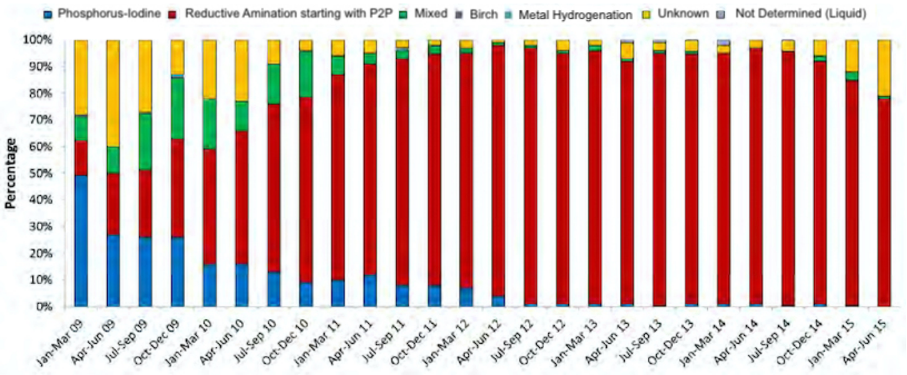
> Ephedrine meth was like a party drug. […] You could normally kind of more or less hang onto your life. You had a house, you had a job. […] P2P meth was *nothing* like that. It was a very sinister drug. It brought you inside. You didn't want to be around other people. You wanted to just kind of be alone with whatever bizarre thoughts your mind was now cooking up, and conspiracies.

I was curious about this. What do we know about the difference between old meth and P2P meth? What evidence is there that these have a chemical difference?

Meth in the US shifted to P2P synthesis between 2009 and 2012.

In the before times, meth was made with ephedrine or pseudoephedrine. However, in 2006, the US banned over-the-counter sales of pseudoephedrine, and in 2008 Mexico banned almost all sales. In response to this, meth makers switched to a synthesis based on P2P, which can be made from many different, widely available, source chemicals.

The Drug Enforcement Agency tests the meth they seize to see how it was made. Here's their data starting in 2009, where you can see that P2P synthesis (in red) rapidly displaces the older ephedrine-based synthesis (in blue).



How could P2P meth be different? There are two ways: Either it could be a different *type* of meth, or the meth could be contaminated with some other chemicals.

Let's talk about different types of meth first.

Isomers

A naive P2P synthesis would produce an even mixture of l-meth and d-meth.

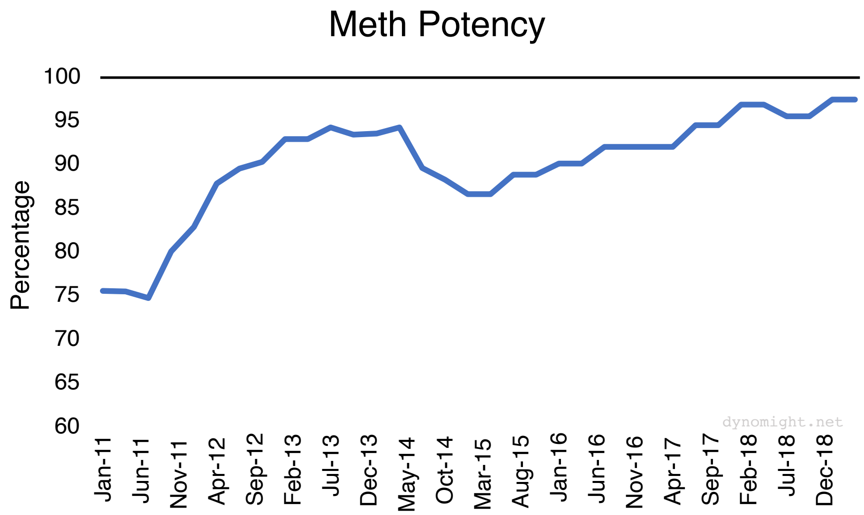
For many complex molecules, you can take the atoms, and "flip" them to get another stable version of the same molecule, called an isomer or (more specifically) an enantiomer. These different versions of the molecule can have very different effects on the body.

Methamphetamine happens to be one of those molecules. The one that produces the effects we call "meth" is d-methamphetamine (d-meth). That's the one that increases dopamine in the brain, causing euphoria. (It's also the one that is sold at pharmacies in the US to treat ADHD and obesity.) On the other hand, l-methamphetamine (l-meth) has no effects on dopamine and presumably isn't nearly as much fun.

Anyway, a synthesis that turns P2P into meth will create an equal mixture of d-meth and l-meth, basically because atoms bouncing around randomly are equally likely to end up in either of two equally low-energy configurations. Older synthesis methods using ephedrine would create only d-meth.

P2P initially had a fair amount of l-meth, but it was almost all gone by 2019.

Here's data from the DEA again, where "potency" is the percentage of d-meth among all meth.



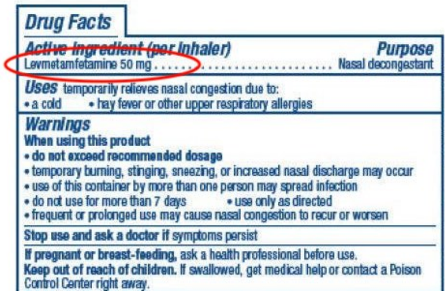
The dip in 2014 might be explained by the introduction of a new synthesis method (NTS), which we'll talk about below.

Unfortunately, I can't seem to find any data going back further to before when P2P meth was introduced. It's likely that d-meth was higher before P2P synthesis become popular, though this paper analyzes meth in Australia and finds that, for some reason, ephedrine-based meth often has fair amounts of l-meth, too.

L-meth is in various easy-to-obtain drugs.

Vick's VapoInhalers contain 50mg of l-meth, which they spell in an unusual way probably to reduce the number of people who notice what's in there and freak out.

TAMPER EVIDENT: Do not use if printed inhaler wrap is broken or missing.
VICKS
VapoInhaler
Levmetamfetamine...Nasal Decongestant


Drug Facts

Active Ingredient (per Inhaler) Purpose
Levmetamfetamine 50 mg .... Nasal decongestant

Uses temporarily relieves nasal congestion due to:
- a cold
- hay fever or other upper respiratory allergies

Warnings
When using this product
- do not exceed recommended dosage
- temporary burning, stinging, sneezing, or increased nasal discharge may occur
- use of this container by more than one person may spread infection
- do not use for more than 7 days
- use only as directed
- frequent or prolonged use may cause nasal congestion to recur or worsen

Stop use and ask a doctor if symptoms persist

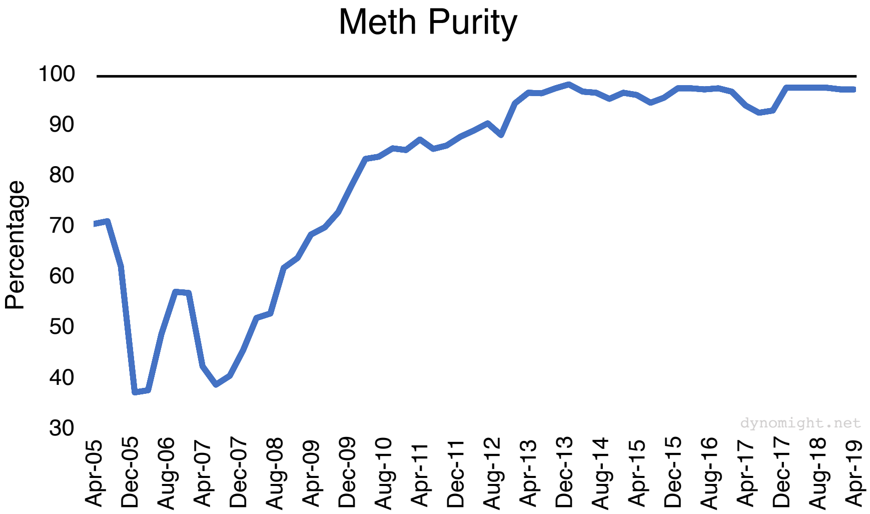
If pregnant or breast-feeding, ask a health professional before use.
Keep out of reach of children. If swallowed, get medical help or contact a Poison Control Center right away.

L-meth is also produced as a metabolite of Selegiline, a drug for Parkinson's and depression.

Contaminants

The purity of meth is now higher than ever.

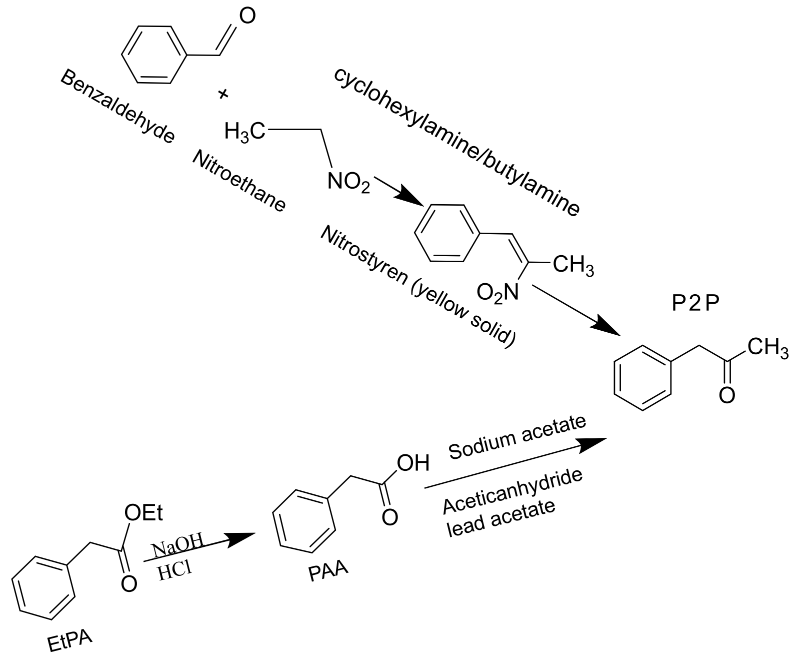
The DEA has tracked purity in meth that they have seized for a long time. They define purity to be the percentage of meth (d or l) amongst all chemicals in the sample. Here's a plot of all the data I could find (assembled from the National Drug Threat reports from many different years):



Modern street meth is purer than ever, something on the order of 97% d-meth on average.

There are many ways to make P2P meth.

Here's a figure that shows how P2P might be produced from source chemicals, simplified from this paper:



This shows two routes to make P2P. The top route uses benzaldehyde and nitroethane to produce nitrostyrene (NTS), which is then made into P2P. The bottom route uses ethyl phenylacetate (EtPA) to make phenylacetic acid (PAA), which is again made into P2P. Note that lead acetate (which has been raised as a concern) is only used in the PAA synthesis route.

Synthesis methods for P2P meth have changed repeatedly.

This paper by DEA scientists goes over the profiling of different types of P2P meth. Here's the history, as far as I can make out:

- Starting around 2009, people used EtPA to make PAA.
- Around 2014, there was a shift towards using the NTS synthesis.
- Around 2018, there was a shift back towards using PAA. (It's not clear if this PAA was sourced directly, or made from EtPA or what.)

It's much messier than this implies: The transitions were gradual, and the DEA finds a fair number of "unknown" samples each year that they can't classify.

On top of these different methods to make P2P, there are different methods to convert P2P into meth, and these have probably changed over time as well. The DEA seems to attribute most impurities to the P2P production step. However, they seem more interested in the meth supply chain than how impurities might affect the health of users.

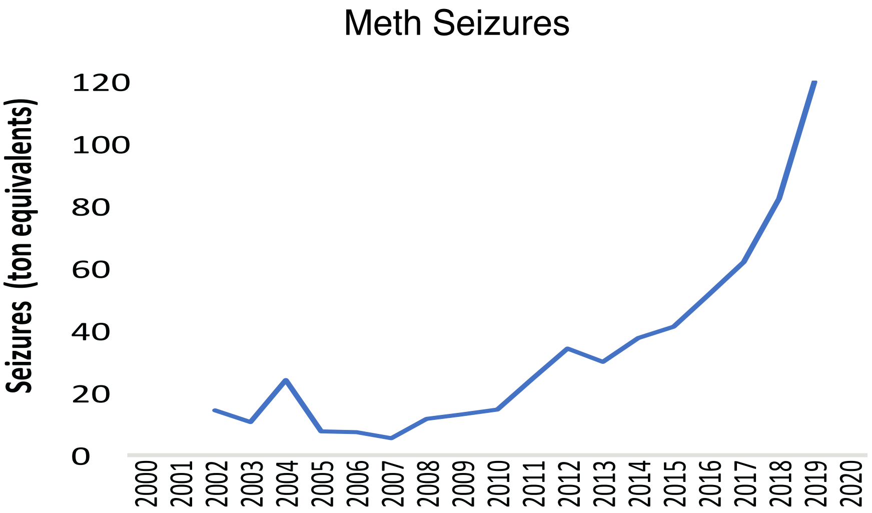
This history of synthesis methods does not support the theory that lead acetate in meth is causing schizophrenia: Lead acetate was used much less between 2014 and 2018 when NTS synthesis mostly displaced PAA synthesis. This doesn't correlate with reports of schizophrenia at all.

Quantity

How much meth is used, by how many people? It's hard to say exactly, given that we're talking about a black-market supply chain and a product that's illegal to consume. Still, we have various windows into things.

The amount of meth seized at the border is skyrocketing.

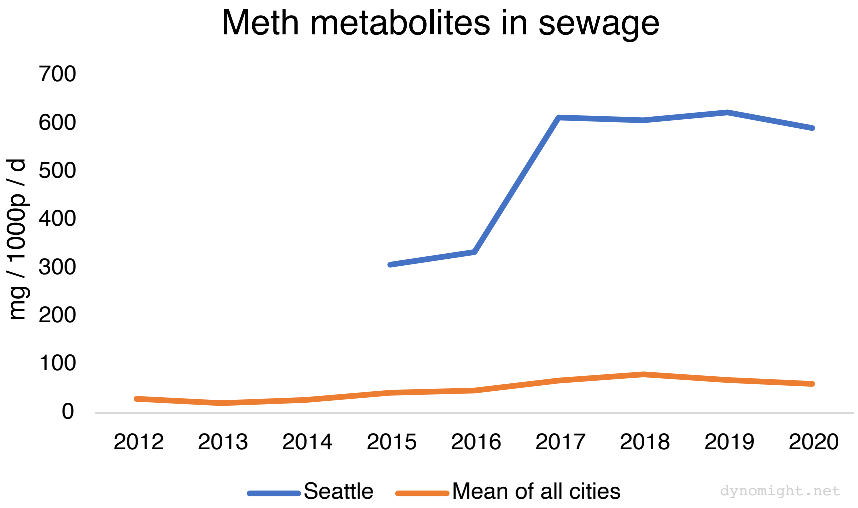
Here's a figure, modified from the UN's 2021 World Drug Report.



To some degree, this reflects Mexican-made meth displacing US-made meth, but this isn't a major factor: Already in 2012, the DEA estimated that 80% of meth in the US was Mexican-made.

Sewage measurements suggest a doubling of usage in Seattle around 2017.

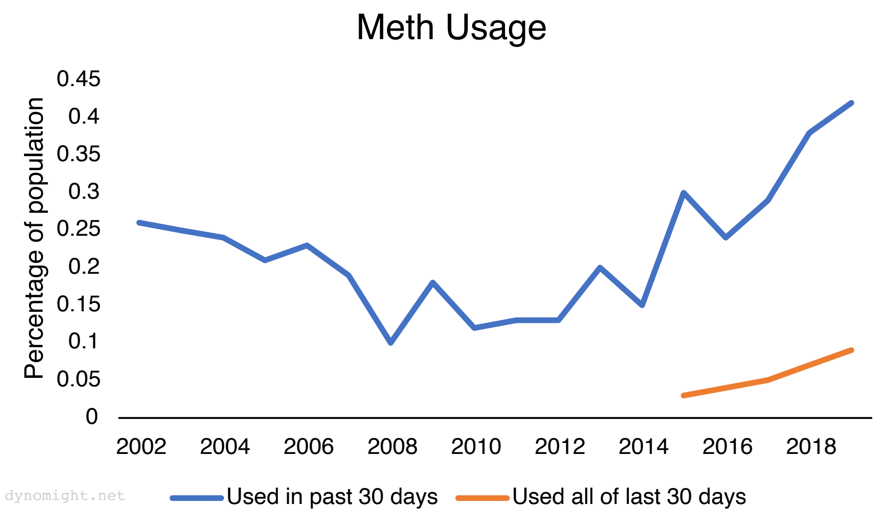
There's an impressive project in Europe to measure drug use from biomarkers in sewage. They invite participation from cities around the world, which Seattle does. Here are their measurements (extracted from measurements in the 2020 report):



In 2016, Seattle already had the highest levels of any participating city in the world, but these doubled in 2017 and then stayed roughly constant after.

More people now report using meth, especially using a *lot* of meth.

One way to estimate how much meth people use is to ask them. The US Department of Health and Human Services maintains the Substance Abuse & Mental Health Data Archive with data going back to 2002. I used this data to get two numbers: The percentage of people who used meth *at all* in the past 30 days, and the percentage of people who used meth *every day* in the last 30 days. (The latter is only available since 2015). These are proxies for the number of casual users and the number of serious addicts.



The number of people who use meth has increased. However, the real growth is in the number of heavy users, which tripled just between 2015 and 2019.

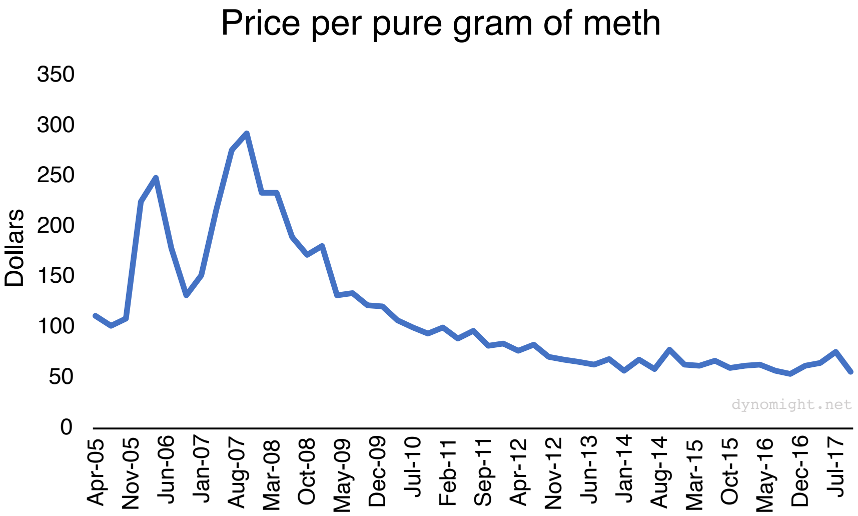
Here are some details on the data used in the above graph, in case you'd like to play around with similar analyses.

These are the raw cross-tables:

After clicking one of those links, turn off all table display options except for row %.

Meth prices have come down.

If supply is increasing, we would expect prices to come down. Have they? The DEA tracks prices in seized meth going back at least as far as 2005. After cobbling together (sometimes contradictory) numbers on National Drug Threat Assessments, here's the best series I could find:



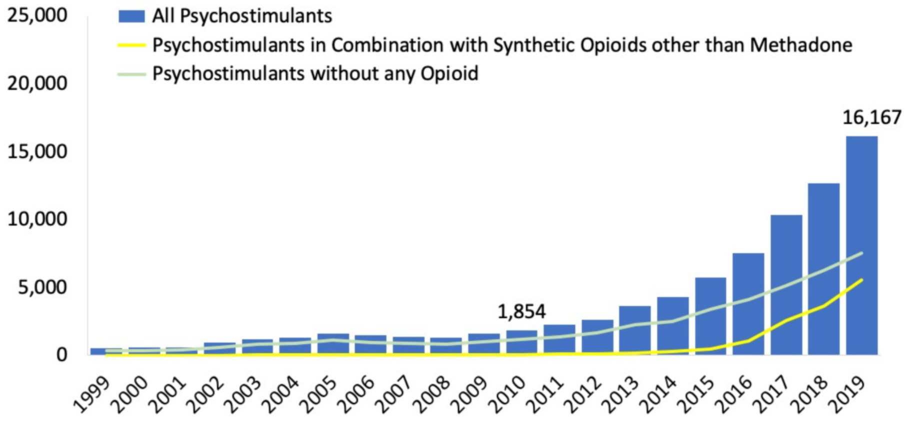
The DEA continued to track prices after 2017, but they noticed that lots of researchers found this data useful and therefore stopped publishing it because screw you.

I'm not sure how reliable these numbers are. They vary a lot based on the quantity being bought and the location in the country. The RAND corporation estimates numbers that are 2-3x higher overall, but show a similar relative decrease from 2008 to 2016.

There are also random quotes scattered across the media: The Kansas Bureau of investigations (2017, 2018, 2019, 2020) reports the price of a pound of meth dropping from around $15k in 2014 to around $4k in 2019, and slightly rebounding to $5k in 2020 during the pandemic. (Apparently, the meth supply chain is more robust than that for semiconductors.) A public television station in California in 2019 quotes a law enforcement officer in Fresno as saying a pound of meth had dropped from $6k for a pound a few years before to only $1k per pound now.

Meth overdoses are skyrocketing.

From the National Institute on Drug Abuse, here are the number of overdose deaths per year. This includes other psychostimulants like caffeine and MDMA, but the deaths overwhelmingly come from meth:



This isn't slowing down. More recent data (not plotted) indicates that that 2020 had 24,076 deaths, and things sped up even more during early 2021. While a lot of these deaths come in combination with opiates like Fentanyl, a lot don't, too.

We can put these numbers in context with some very rough arithmetic.

Let's compare to someone who takes amphetamines/Adderall for ADHD, typically prescribed 5-20 mg per dose. Meanwhile, a strong single dose of meth is 40-150 mg, on top of which people say that meth is around 2× as potent as amphetamines. So meth users take roughly **15×** as much per dose as the typical Adderall user. Meth addicts often dose several times per day, due to the short half-life, meaning a total of 300-800 mg per day.

That's a lot, but let's talk about overdoses. It's actually pretty hard to overdose on meth. One way to estimate it is to look at animals This paper says that 50% of rats and mice die at a dosage of around 55 mg/kg. This suggests that an 80 kg (175 lb) person would need to take 4400 mg of meth to have a 50% chance of dying. Now, it's not safe to extrapolate numbers between animals and humans, and there's a blurry boundary between lethal and non-lethal doses. But there are *many* reports out there of people taking 500 mg of meth at a time without overdosing. That's something like **100×** a clinical dose of 10mg of Adderall.

That's an *insane* amount of stimulants. I find it difficult to understand how anyone would want to do that to themselves. But they do, enough that meth overdoses kill half as many people as die in car accidents, and the numbers are still increasing. I guess drug users use a lot of drugs.

Takeaways

What to make of all this?

First, I think it's unlikely that l-meth is causing people to go crazy. Modern P2P meth is nearly pure d-meth, and the percentage of l-meth peaked before

2011, before these reports of schizophrenia.

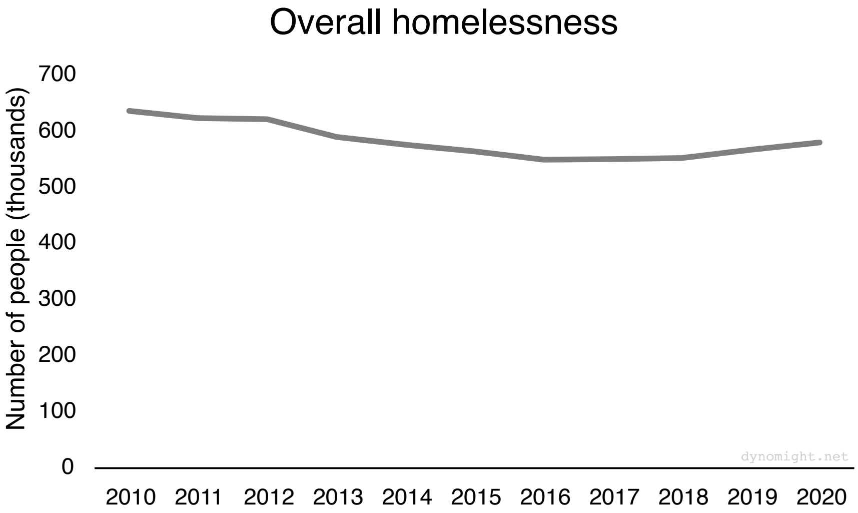
Second, the evidence we have is against the idea of contaminants in P2P meth. Almost all meth was produced using P2P since 2012, before most reports of schizophrenia. And P2P meth synthesis has changed several times in the interim, resulting in higher purity than ever before.

Third, the major impact of P2P synthesis is that a *lot* more meth is available. We have many sources of evidence for this: Border seizures, sewage measurements, usage surveys, prices, and overdose data. All these indicate that people are using historically large amounts.

Does this rule out the idea of contaminants? No. Even if it's 97% pure d-meth, there could be something very nasty lurking in that last 3%. But I don't see the *need* for such an explanation. We know there are many more heavy users, so there's no need to go beyond the idea that quantity has a quality all its own.

A breakdown of the data on the homeless crisis across the U.S.

Is the US in the midst of a homelessness crisis? Many people think so, but that's largely based on based on anecdotes. What does the data say?

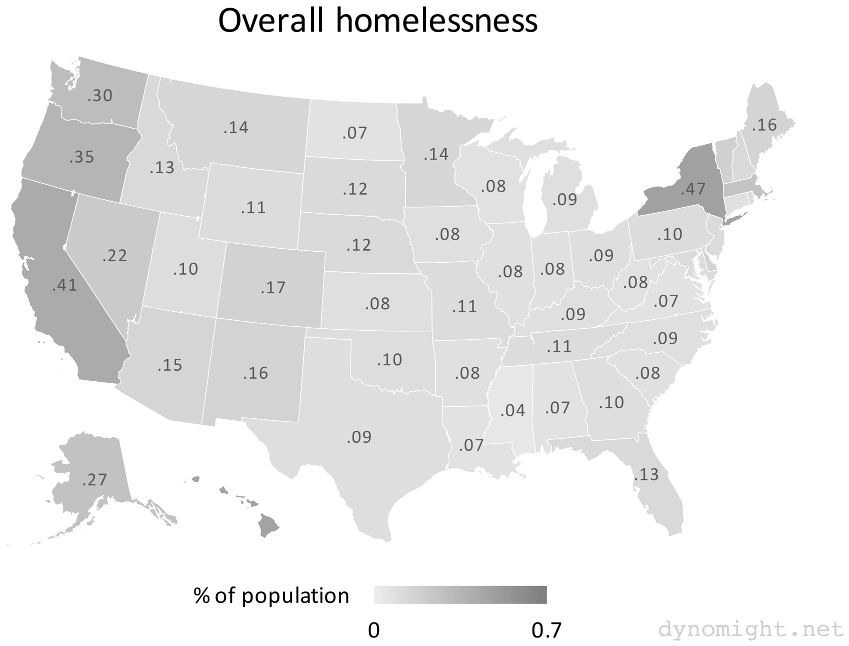


At a glance, this doesn't *look* very crisisy. Since 2015, things have only gone up by less than three percent.

Still, I think there *is* a crisis, we just have to work harder to see it. We need to look at different locations, rates of change, different types of homelessness, and mental health and substance abuse issues. Let's do that.

Homelessness is much higher in some places.

The rate of homelessness varies hugely between different states. Here's the percentage of the population in each state that was homeless in 2020.

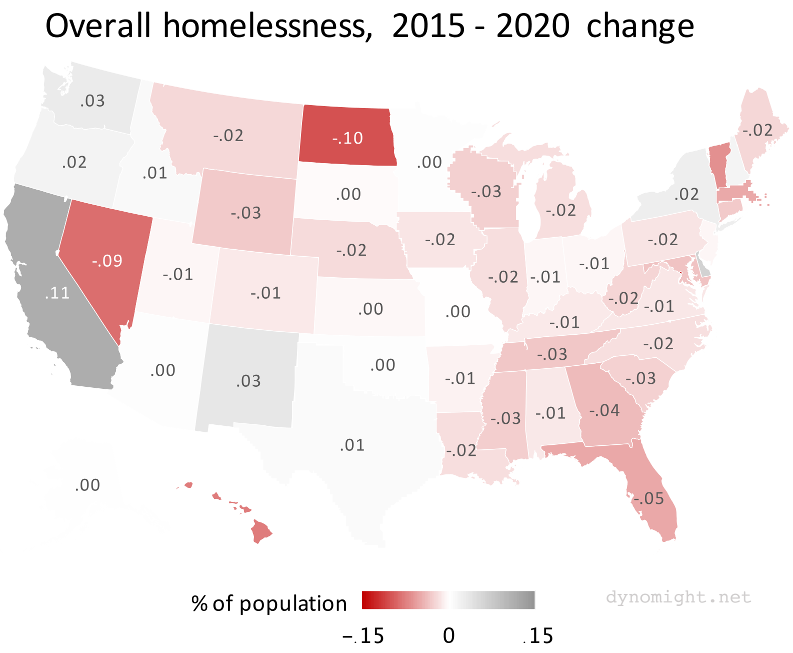


In Mississippi it is 1 in 2500 (0.04%), while in New York State it is 1 in 210 (0.47%). That's a huge difference.

Now, when we talk about a crisis, there's an implication that things are getting worse. (You don't hear much about the *everyone you love will die and be forgotten "crisis"*...) We already saw that things are pretty stable at the national level. How are things changing in individual states?

Homelessness is increasing in some places and decreasing in others.

Here's the change between 2015 and 2020, again as a percentage of each state's population.



The general pattern is increases close to California and decreases close to Florida. The exceptions are the random huge decreases in Nevada and North Dakota (maybe because of the end of the fracking boom?), and the confusing mess in the Northeast.

This is a relatively simple story so far, but I don't think it fully captures what's going on. We have to go deeper.

There are different types of homelessness.

Some people run out of money, get evicted, and stay at a shelter for a few weeks before moving in with family and eventually getting back on their feet. Other people have mental health issues and stay on the street for years. When we talk about a "homelessness crisis", we should try to distinguish these different situations.

Let's back up. Where is all this data coming from? Well, every year during the last 10 days of January, the US tries to count every homeless person. This is done by hundreds of local government organizations and nonprofits that cover almost the entire country. They must participate in the counting or they won't get any funding from the US Department of Housing and Urban Development (HUD), which collects all the data.

We'll look at two attributes that people get during this count. First, people are classified as *sheltered* or *unsheltered*. Someone is sheltered if they are staying in an emergency shelter, transitional housing program, or a safe haven. Someone is unsheltered if they're staying in a vehicle, an abandoned building, or on the street. Second, people are *chronically* homeless if they've been so for at least a year, and otherwise *non-chronically* homeless.

Aside: Technically—because nothing can be simple—the definition of "chronically homeless" is much more complex. You also qualify if you've been homeless for a total of a year in the last three years, plus that happened in at least four episodes, plus those episodes were separated by at least a week. Yes, that means that someone homeless their entire life except the first week of each year is "non-chronic", and yes that is crazy, but probably it never happens and anyway these categorizations are done by random people around the country who may not much care about the legalistic definition anyway. Oh, and also, this changed in 2016.

So, what types of homelessness does the US have?

Most homeless are non-chronic, and most non-chronic homeless are sheltered. But the chronic homeless are usually unsheltered.

Here are the fractions of people that fell into each of the four possible groups in 2020.

	Non-Chronic	Chronic
Sheltered	53.5%	7.6%
Unsheltered	25.8%	13.2%

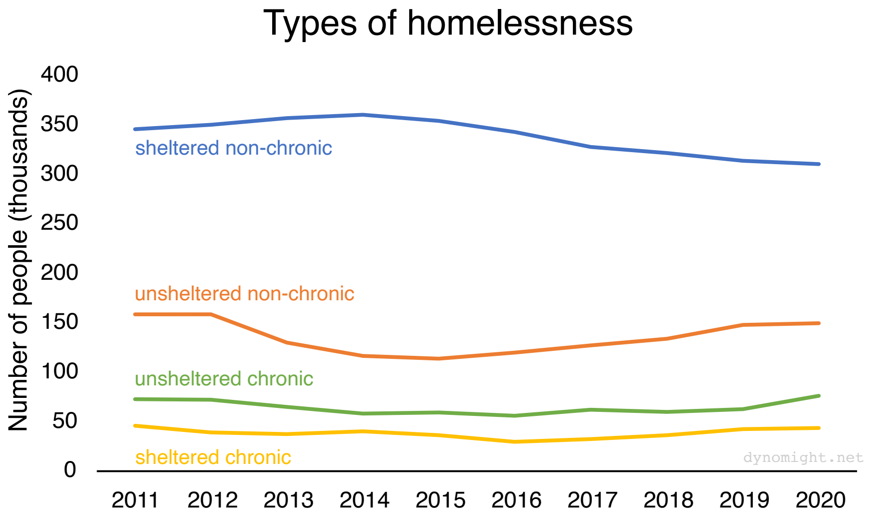
Here's how I think about this:

1. Most homeless are non-chronic (around 80%).
2. The non-chronic homeless are mostly sheltered (around 2/3).
3. The chronic homeless are mostly *un*sheltered (around 2/3).

That's the overall mix. But we have to worry about two things. First, is the mix changing over time? And second, how does the mix vary in different places?

Unsheltered and chronic homelessness is increasing.

Unfortunately, the mix is changing, and for the worse. The "best" type of homelessness (sheltered and non-chronic) is decreasing, while the other types are increasing.



Since it's a bit hard to see, here's the number of people in each group in 2020 divided by the number in 2015:

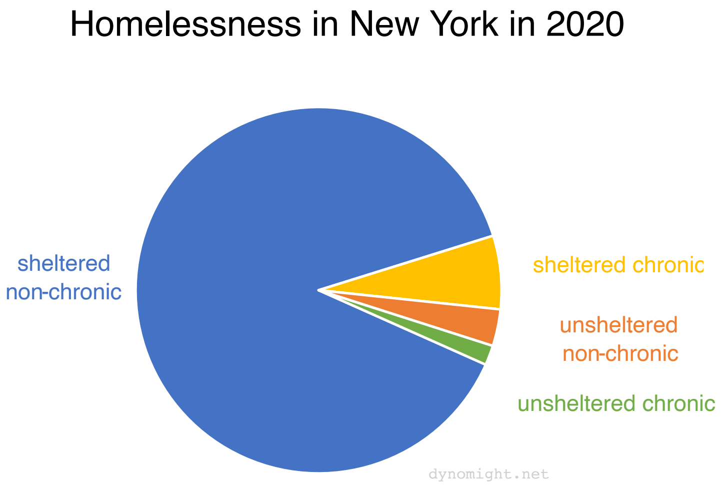
	Non-Chronic	Chronic
Sheltered	87.5%	119.3%
Unsheltered	131.5%	128.4%

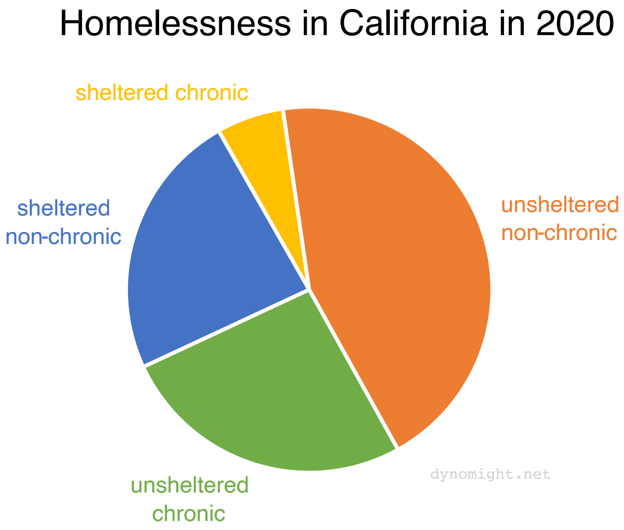
While sheltered and non-chronic homelessness remains the most common, it actually decreased by around 12.5% since 2015, whereas all the other types have increased by around 25%.

OK, but how do things look in different places?

Chronic and unsheltered homelessness is much more common in some places than others.

Let's compare New York and California:

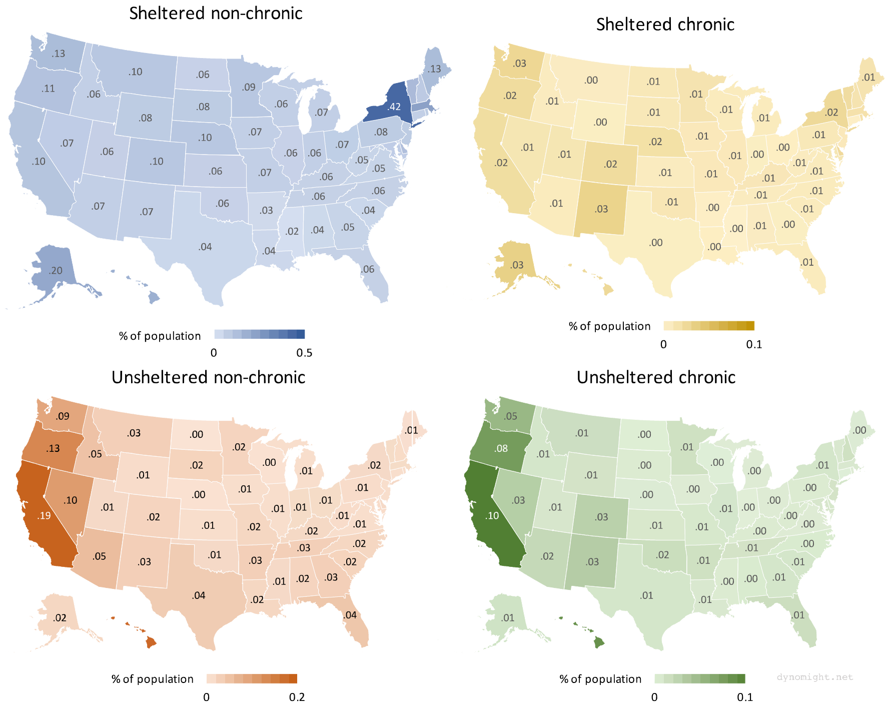




Homelessness overall is similarly high in both places—0.47% in New York and 0.41% in California. But California has *much* more unsheltered and chronic homelessness.

Aside: Did you know that there is a *constitutional* right to shelter in New York? This is a result of a 1979 New York State Supreme Court decision. Two other places have a weaker version of this. Massachusetts has a mandate from a 1983 law but it only applies to families, not individuals. The District of Columbia guarantees shelter to families year-round, and to individuals when the temperature is below 32° F or above 95 °F (below 0° C or above 35 °C).

Anyway, we can picture the situation for all states by visualizing the four types of homelessness on a map. Here they all are, each as a percentage of the state's population. (Click for a .pdf if you want to look closer.)



For sheltered non-chronic homelessness, the West Coast isn't exceptional. Instead, the standouts are New York at 0.42% and—press your face against the screen—the District of Columbia at 0.49%. These are followed by Massachusetts at 0.23% and Alaska at 0.20%.

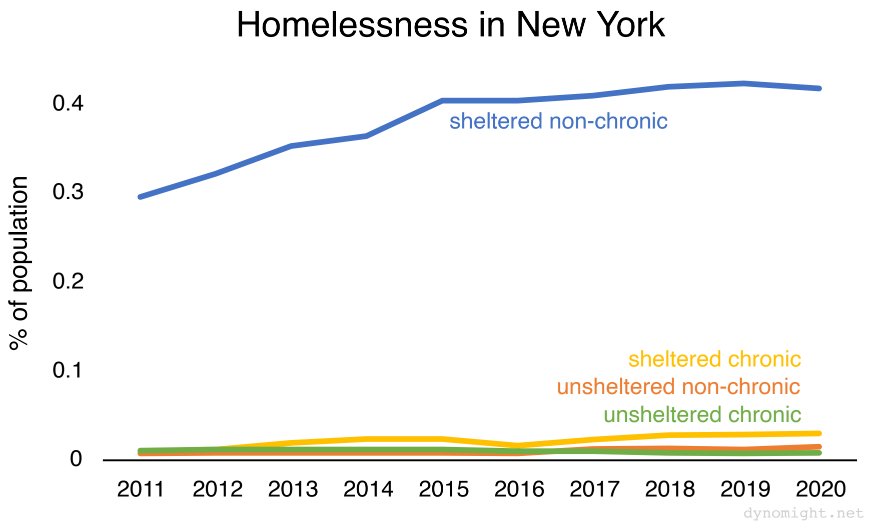
But unsheltered chronic homelessness is very different. It's almost nonexistent in some places—it's 0.10% in California and 0.08% in Hawaii, but 0.0013% in Wisconsin. In Wisconsin, that is 80 people total. In California, that's 42,195.

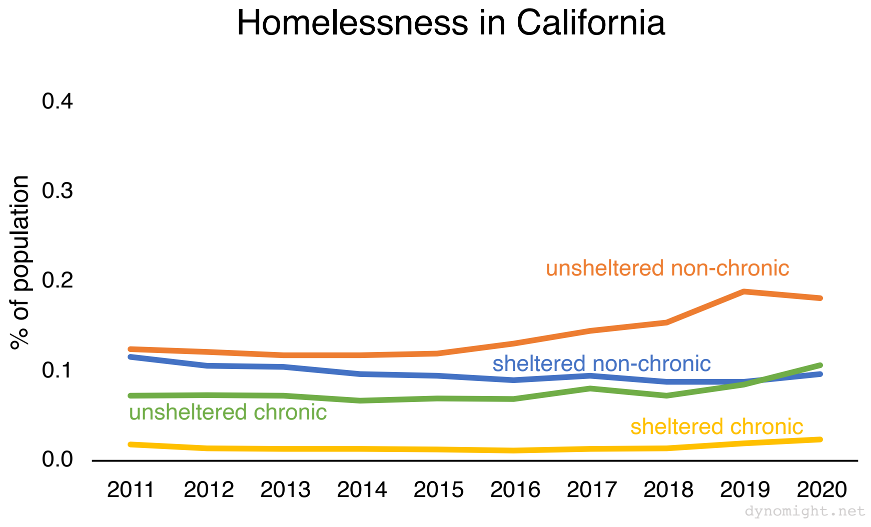
However, let me remind you—I can't emphasize this enough—this survey is done in *late January*, when it's cold in New York and *stupidly cold* in Wisconsin. You have to imagine that things would look different if things were done in Summer, though it's hard to quantify. Still, the difference might be significant in that almost all homeless people in cold places have at least some occasional contact with social services.

OK. The types of homelessness are different in different places, and they are changing over time. But how are they changing *in each place*? We still need to go deeper.

Unsheltered and chronic homelessness is getting worse in some places, particularly the West coast.

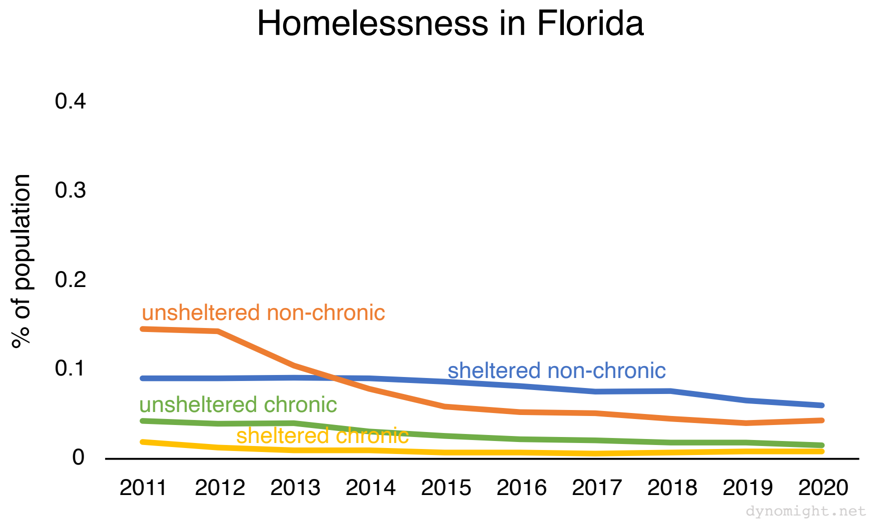
How are the different types of homelessness changing over time in each state? Let's again contrast New York and California:





The situation in New York is reasonably stable since 2015, albeit at very high levels. Meanwhile, California has a small *decline* in sheltered non-chronic homelessness, but big increases in the other groups.

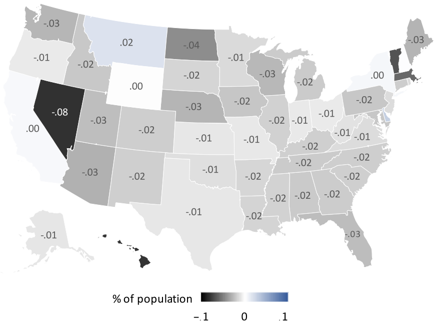
Other places are different. For example, in Florida, everything is decreasing at once.



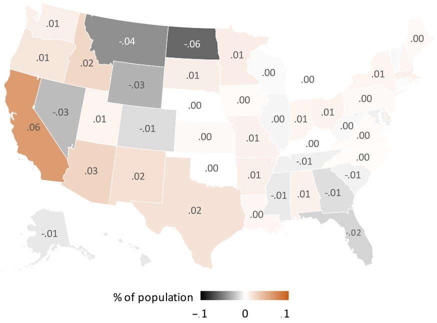
What does the rest of the country look like? Remember, that for the country as a whole, we saw above that sheltered non-chronic homelessness is decreasing, and the other categories are increasing.

But what about individual states? Since it's not convenient to look at 50 different charts, I made four maps to show the change in each type of homelessness between 2015 and 2020, again as a percentage of each state's population. In each of these, grey is a decrease, white is no change, and a non-grey color is an increase. (Click for a .pdf to zoom.)

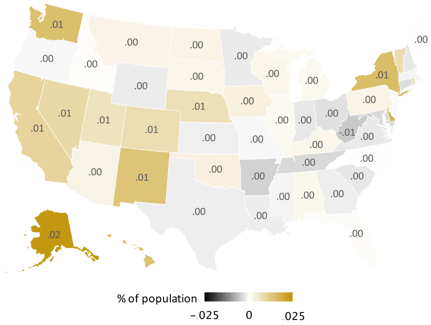
Sheltered non-chronic, 2015 -2020 change



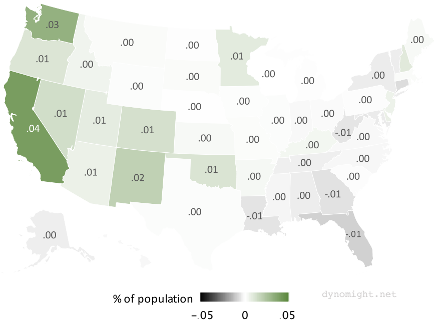
Unsheltered non-chronic, 2015 - 2020 change



Sheltered chronic, 2015 - 2020 change



Unsheltered chronic, 2015-2020 change



How should we think about this? Here's my best attempt at a summary:

1. Sheltered non-chronic homelessness decreased almost everywhere.
2. All homelessness declined in Florida and places near Florida.
3. Chronic and unsheltered homelessness increased a lot in California and places near California.

There are some exceptions. For one thing, despite being close to California, Nevada and some of the Montana-esque states saw big decreases in certain categories. For another, the Northeast is weird and defies any attempt to summarize. Sheltered non-chronic homelessness decreased a lot in Massachusetts and Vermont but barely changed in New York or New Hampshire. Every other category is a random mishmash with no pattern. I tell you, on my worst days it's almost like reality is just completely indifferent to our desires to understand it with tidy little narratives.

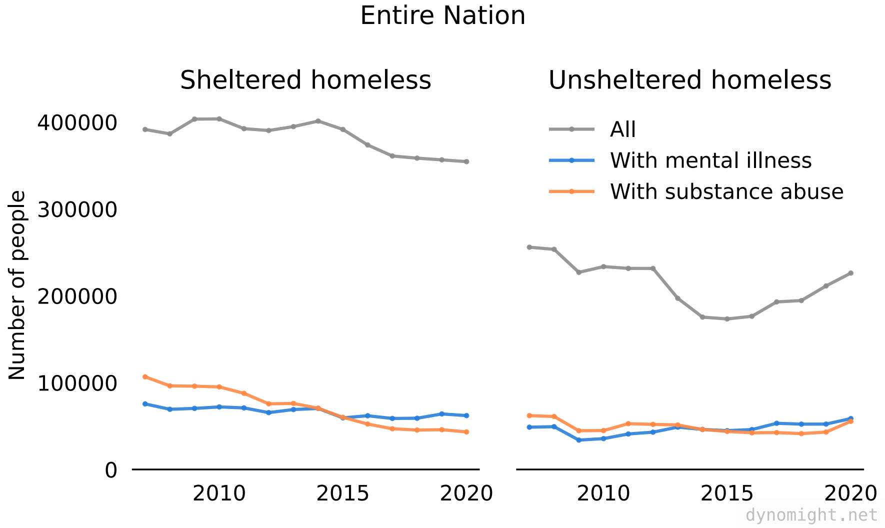
And what about meth? Theories abound that mental health and substance abuse are a huge part of the homelessness crisis.

Well, the yearly homelessness survey collects data on if the homeless are "severely mentally ill" or suffer from "chronic substance abuse". I can't figure out exactly how these are defined. It's implied that it's done by literally asking people, but I think it varies from place to place.

Unfortunately, HUD doesn't publish data on these numbers. However, they do publish *reports*, both for the entire nation, for individual states, and for each individual region.

So, I did the sensible thing. I downloaded the .pdf files for each of the 6,082 different reports, wrote a script to convert each .pdf to plain text, wrote a parser for that text, compensated for 8 billion inconsistencies in how the reports were laid out, damn you HUD, damn you to hell, extracted the data for each of the above categories, and made plots.

Nationally, there is only a small uptick in mental illness and substance abuse.



Blue is people with mental health issues, while orange is substance abuse. Some people will be in both categories, but we don't know how many. (My guess is a lot because they're strongly correlated.) Grey is all people, including those with no issues.

Like the other national data, this doesn't scream crisis. If anything, it's a bit reassuring. We already knew that unsheltered homelessness was increasing, but here we see that mental health and substance abuse aren't increasing at the same rate.

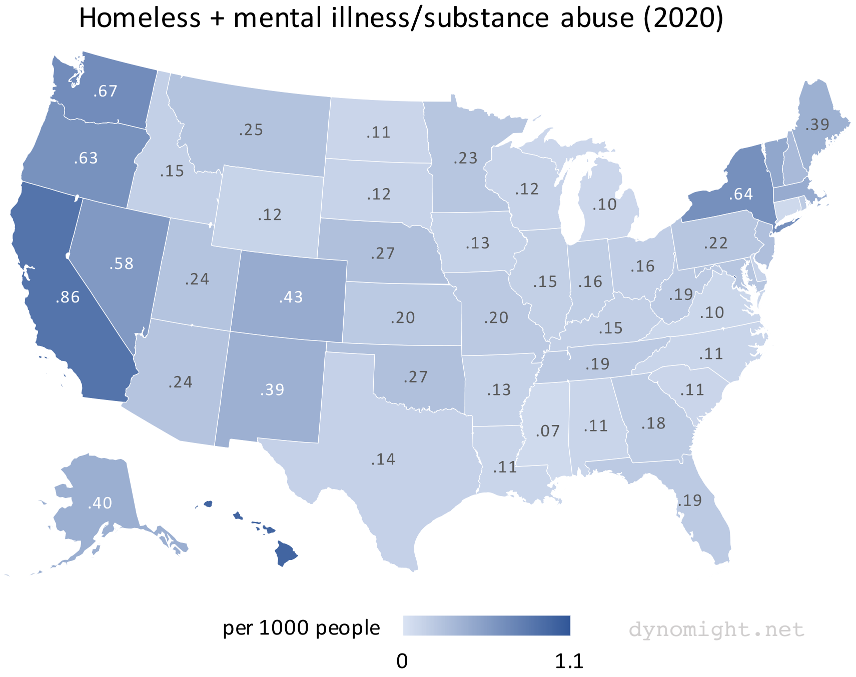
Next, of course, we want to know how this looks in different places, and how it's changing.

I made plots for each of the 437 regions and states (you can see them below). This data is quite noisy, which makes it a challenge to visualize. For example, here's Los Angeles.

To get more reliable numbers, I combined sheltered and unsheltered homelessness, took the average of mental health and substance abuse, and then applied a smoothing function. I used the smoothed values below, which are hopefully less polluted by noise.

There are significant increases in mental illness and substance abuse in certain states.

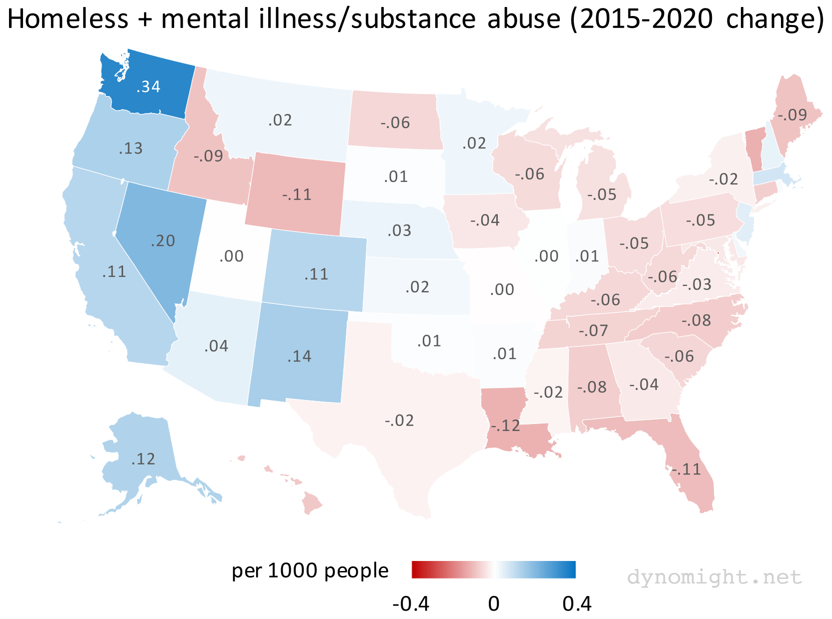
How many homeless people are there in each state with severe mental illness of substance abuse problems? Here are the numbers in 2020.



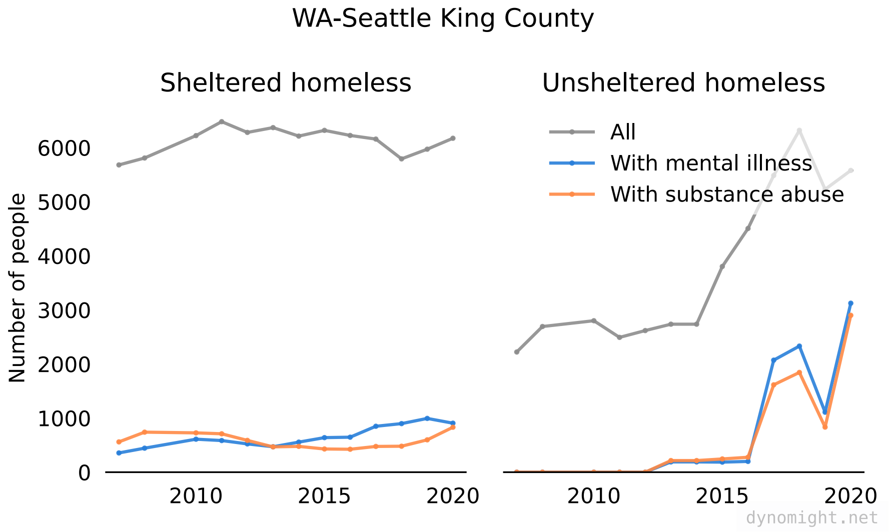
Homeless + mental illness/substance abuse (2020)

Again, this is high if you're close to California and in certain parts of the Northeast. This is pretty similar to the map of homelessness overall we started with, though you'll see a few places that stand out more here, e.g. Nevada.

OK, and how are things changing?



Again, we have a crisis mostly in the West. California was already really high in 2015 and has gotten even higher. The real standout is Washington state, where things more than doubled. And sure enough, the plot for Seattle is stark.

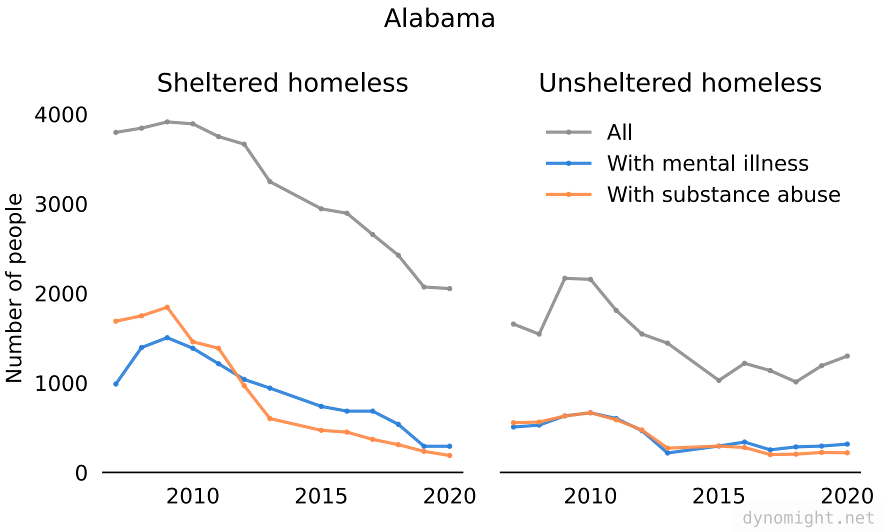


Things just exploded between 2016 and 2017 and then went up from there. You might remember that 2016-2017 is exactly when meth metabolites in sewage in Seattle also exploded. Of course, that could just be a coincidence, or might be noise in either of these datasets. Still, I think we can see why people in Seattle might feel there's a crisis. Maybe anecdotal knowledge ain't so bad.

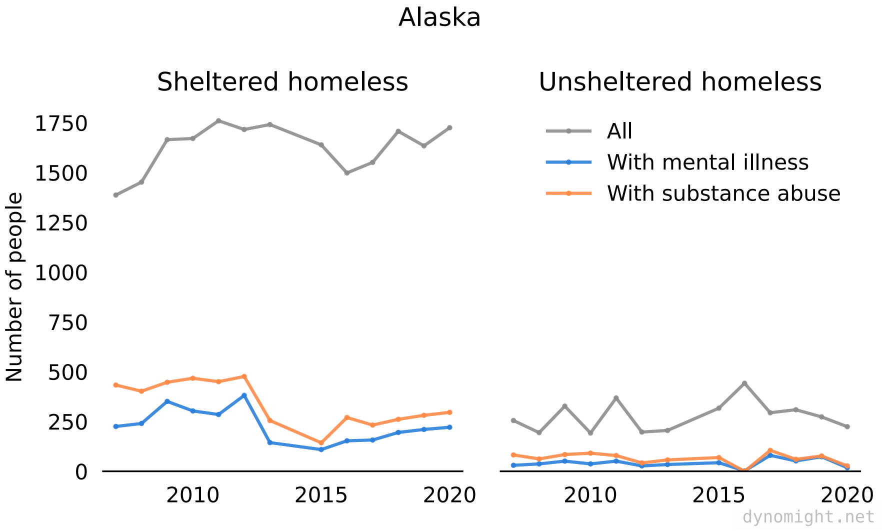
P.S. If you want to see data for your local state or city it is below.

All 50 states and DC.

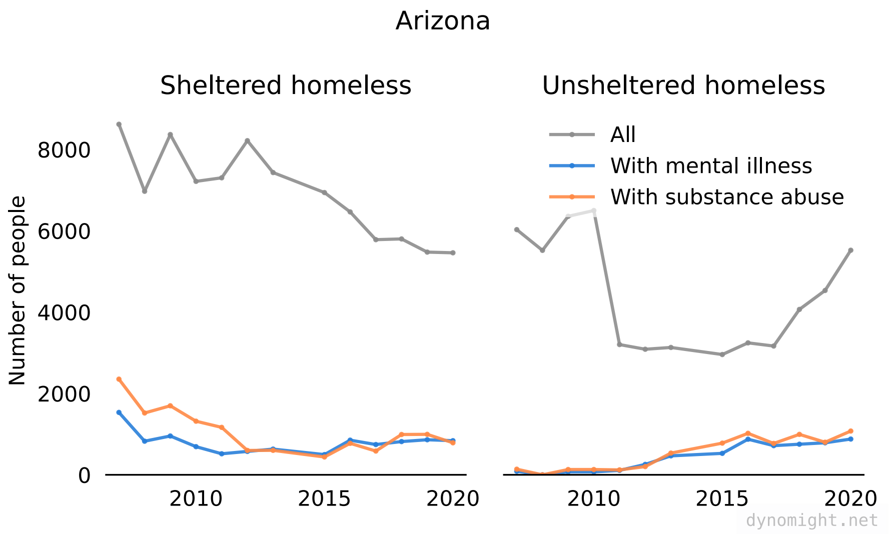
Alabama



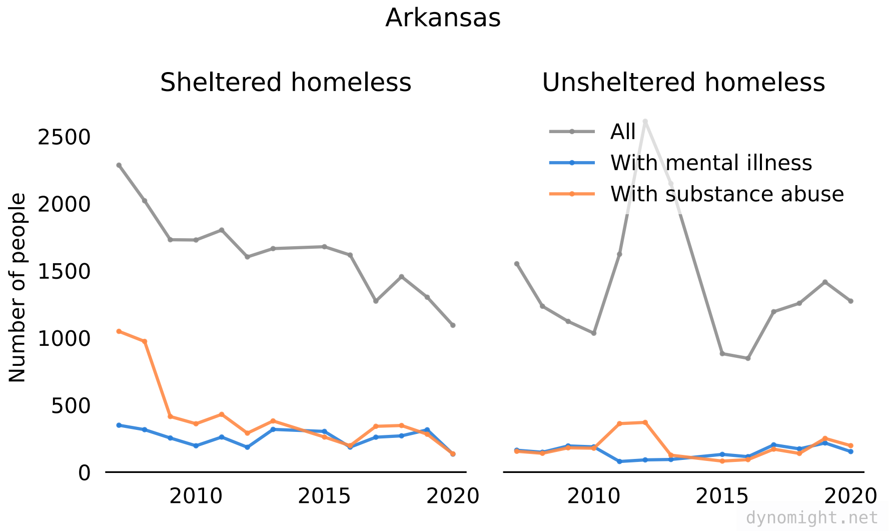
Alaska



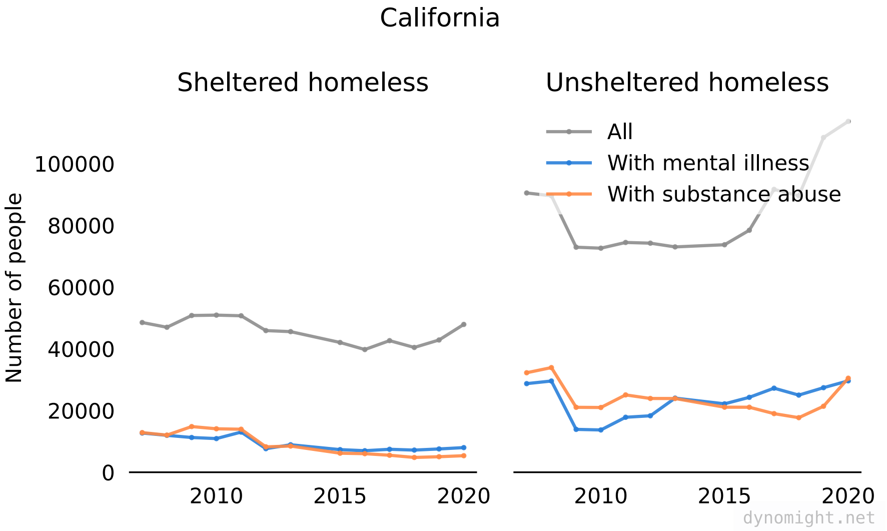
Arizona



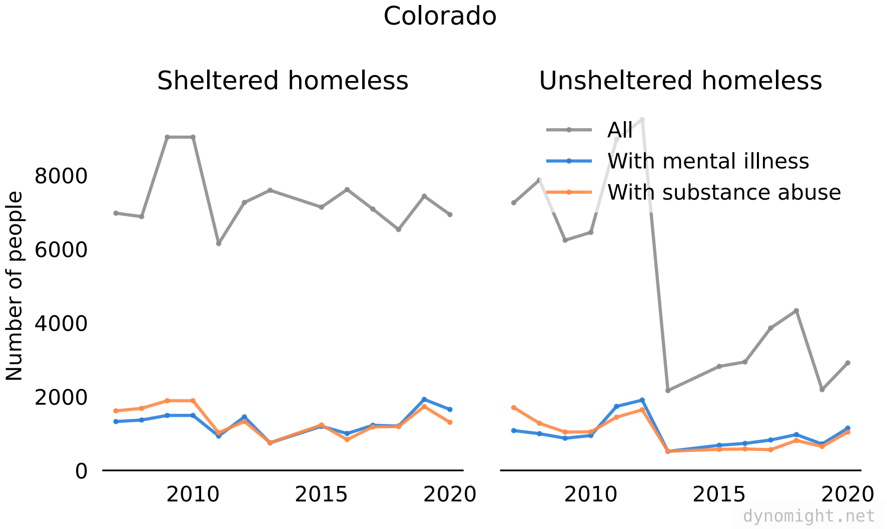
Arkansas



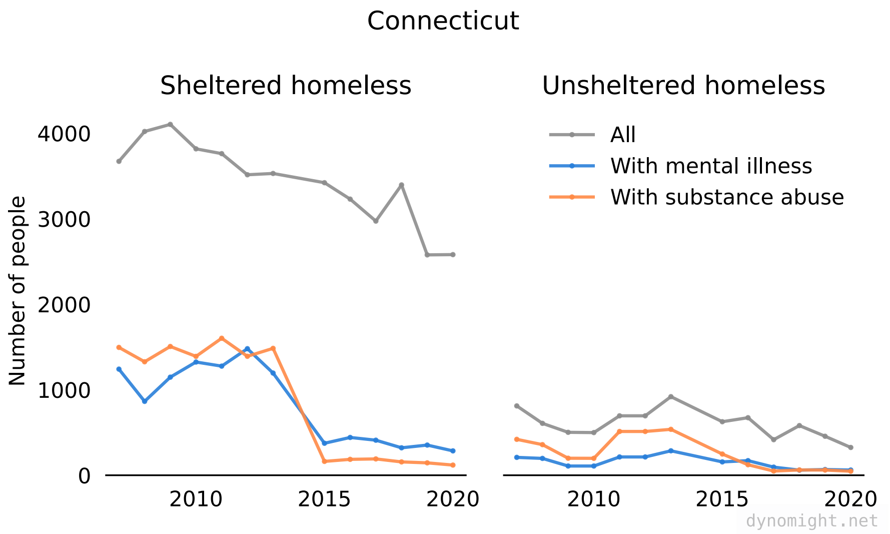
California



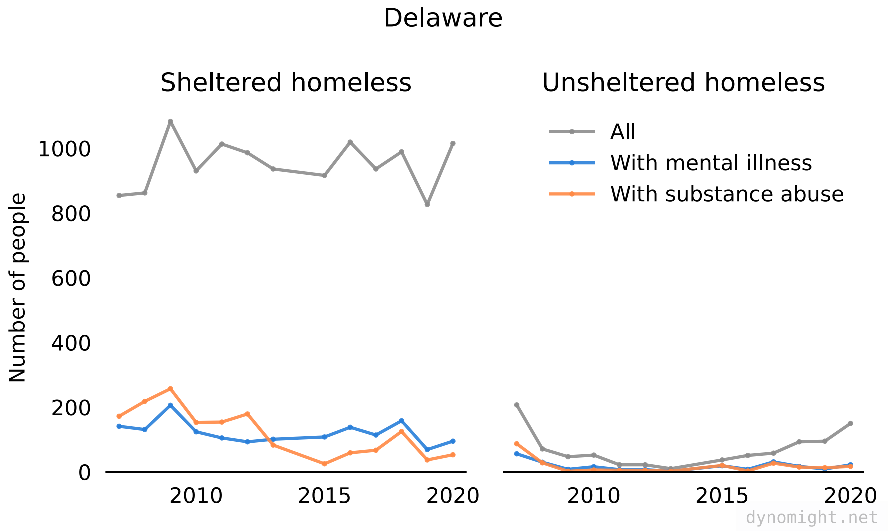
Colorado



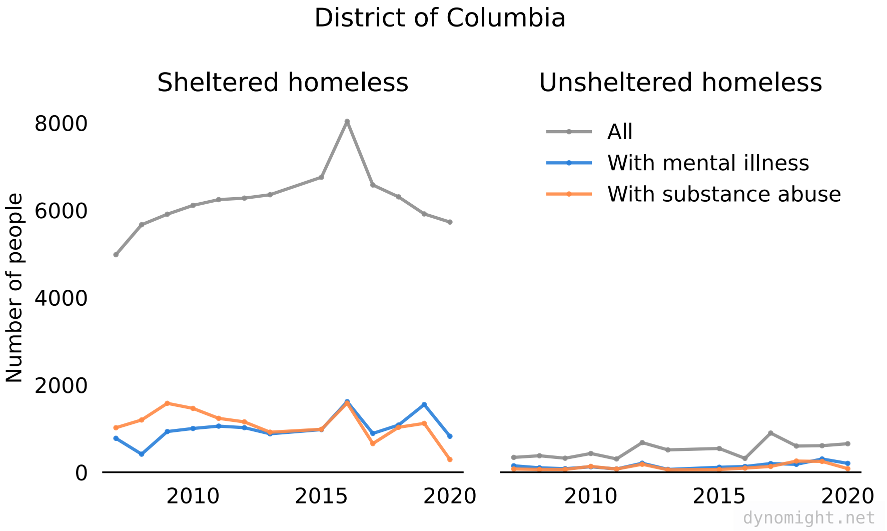
Connecticut



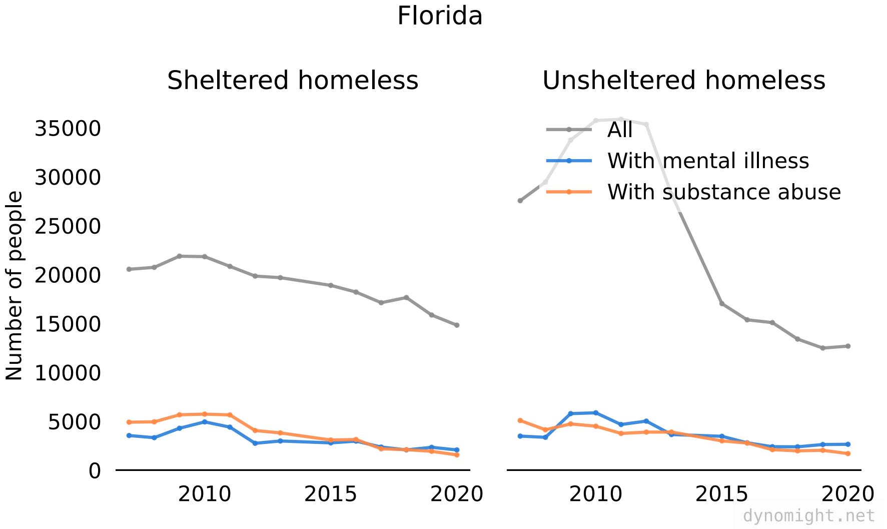
Delaware



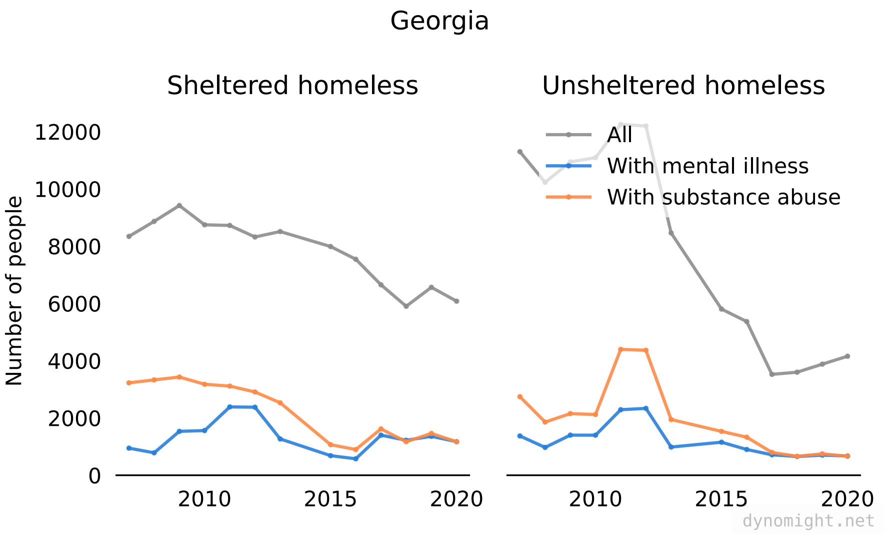
District of Columbia



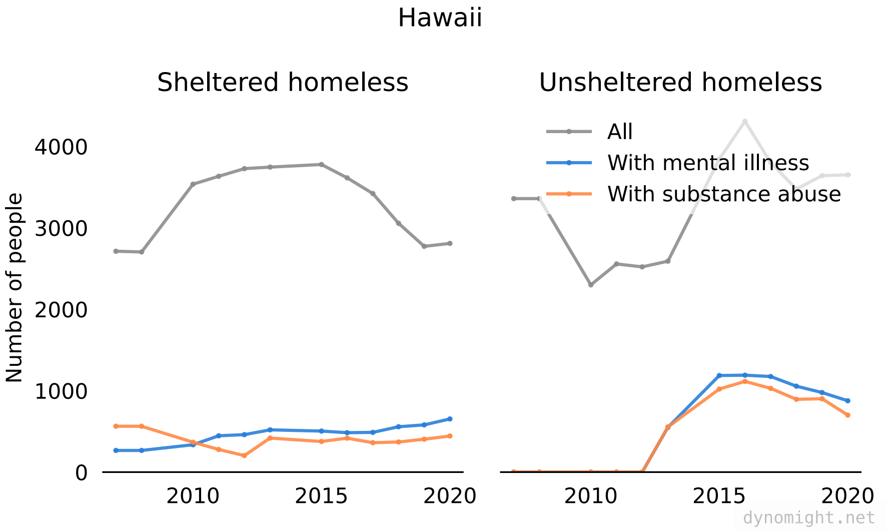
Florida



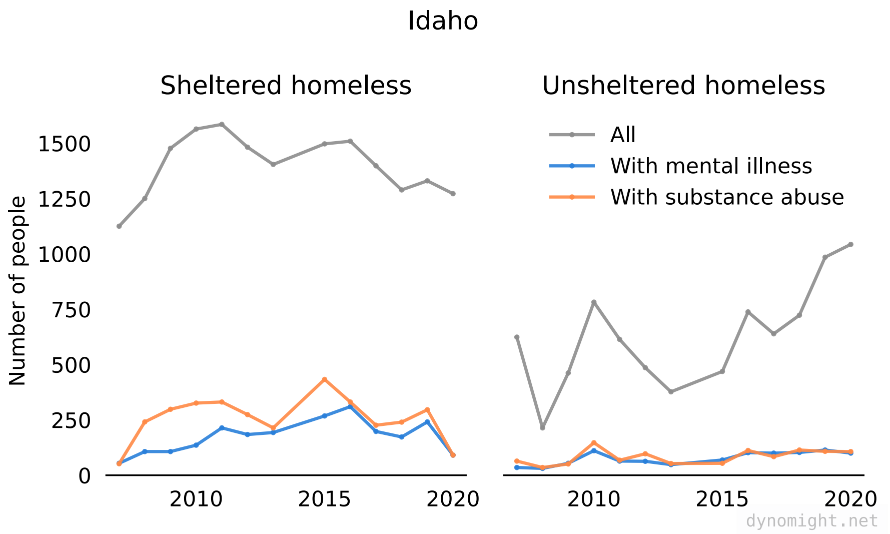
Georgia



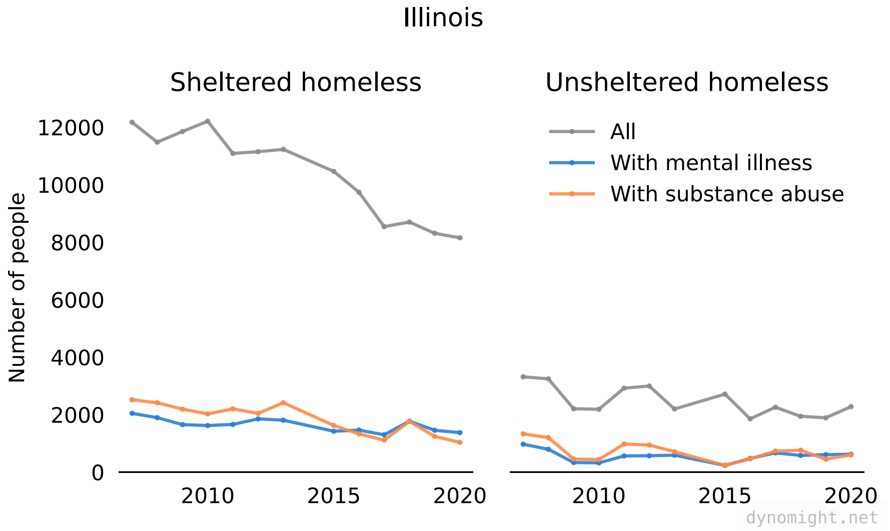
Hawaii



Idaho



Illinois



Indiana

Iowa

Kansas

Kentucky

Louisiana

Maine

Maryland

Massachusetts

Michigan

Minnesota

Mississippi

Missouri

Montana

Nebraska

Nevada

New Hampshire

New Jersey

New Mexico

New York

North Carolina

North Dakota

Ohio

Oklahoma

Oregon

Pennsylvania

Rhode Island

South Carolina

South Dakota

Tennessee

Texas

Utah

Vermont

Virginia

Washington

West Virginia

Wisconsin

Wyoming

All 387 local cities / regions.

(First choose a state, then choose a local region.)

AK

Alaska Balance of State

Anchorage

AL

Alabama Balance of State

Birmingham Jefferson, St. Clair, Shelby Counties

Florence Northwest Alabama

Gadsden Northeast Alabama

Huntsville North Alabama

Mobile City & County Baldwin County

Montgomery City & County

Tuscaloosa City & County

AR

Arkansas Balance of State

Fayetteville Northwest Arkansas

Little Rock Central Arkansas

Southeast Arkansas

AZ

Arizona Balance of State

Phoenix, Mesa Maricopa County

Tucson Pima County

CA

Alpine, Inyo, Mono Counties

Amador, Calaveras, Mariposa, Tuolumne Counties

Bakersfield Kern County

Chico, Paradise Butte County

Colusa, Glenn, Trinity Counties

Daly City San Mateo County

Davis, Woodland Yolo County

El Dorado County

Fresno City & County Madera County

Glendale

Humboldt County

Imperial County

Lake County

Long Beach

Los Angeles City & County

Marin County

Mendocino County

Merced City & County

Napa City & County

Nevada County

Oakland, Berkeley Alameda County

Oxnard, San Buenaventura Ventura County

Pasadena

Redding Shasta, Siskiyou, Lassen, Plumas, Del Norte, Modoc, Sierra Counties

Richmond Contra Costa County

Riverside City & County

Roseville, Rocklin Placer County

Sacramento City & County

Salinas Monterey, San Benito Counties

San Bernardino City & County

San Diego City and County

San Francisco

San Jose Santa Clara City & County

San Luis Obispo County

Santa Ana, Anaheim Orange County

Santa Maria Santa Barbara County

Santa Rosa, Petaluma Sonoma County

Stockton San Joaquin County

Tehama County

Turlock, Modesto Stanislaus County

Vallejo Solano County

Visalia Kings, Tulare Counties

Watsonville Santa Cruz City & County

Yuba City & County Sutter County

CO

Colorado Balance of State

Colorado Springs El Paso County

Fort Collins, Greeley, Loveland Larimer, Weld Counties

Metropolitan Denver

CT

Bridgeport, Stamford, Norwalk, Danbury Fairfield County

Connecticut Balance of State

DC

District of Columbia

DE

Delaware Statewide

FL

Charlotte County

Citrus, Hernando, Lake, Sumter Counties

Columbia, Hamilton, Lafayette, Suwannee Counties

Deltona, Daytona Beach Volusia, Flagler Counties

Fort Pierce St. Lucie, Indian River, Martin Counties

Fort Walton Beach Okaloosa, Walton Counties

Ft Lauderdale Broward County

Ft Myers, Cape Coral Lee County

Gainesville Alachua, Putnam Counties

Hendry, Hardee, Highlands Counties

Jacksonville-Duval, Clay Counties

Lakeland, Winterhaven Polk County

Miami-Dade County

Monroe County

Naples Collier County

Ocala Marion County

Orlando Orange, Osceola, Seminole Counties

Palm Bay, Melbourne Brevard County

Panama City Bay, Jackson Counties

Pasco County

Pensacola Escambia, Santa Rosa Counties

Sarasota, Bradenton Manatee, Sarasota Counties

St. Johns County

St. Petersburg, Clearwater, Largo Pinellas County

Tallahassee Leon County

Tampa Hillsborough County

West Palm Beach Palm Beach County

GA

Athens-Clarke County

Atlanta

Augusta-Richmond County

Columbus-Muscogee

DeKalb County

Fulton County

Georgia Balance of State

Marietta Cobb County

Savannah Chatham County

GU

Guam

HI

Hawaii Balance of State

Honolulu City and County

IA

Des Moines Polk County

Iowa Balance of State

Sioux City Dakota, Woodbury Counties

ID

Boise Ada County

Idaho Balance of State

IL

Aurora, Elgin Kane County

Bloomington Central Illinois

Champaign, Urbana, Rantoul Champaign County

Chicago

Cook County

Decatur Macon County

DuPage County

East St. Louis, Belleville St. Clair County

Joliet, Bolingbrook Will County

Madison County

McHenry County

Peoria, Pekin Fulton, Tazewell, Peoria, Woodford Counties

Rock Island, Moline Northwestern Illinois

Rockford DeKalb, Winnebago, Boone Counties

South Central Illinois

Southern Illinois

Springfield Sangamon County

Waukegan, North Chicago Lake County

West Central Illinois

IN

Indiana Balance of State

Indianapolis

KS

Kansas Balance of State

Overland Park, Shawnee Johnson County

Topeka Shawnee County

Wichita Sedgwick County

KY

Kentucky Balance of State

Lexington-Fayette County

Louisville-Jefferson County

LA

Alexandria Central Louisiana

Lafayette Acadiana Regional

Louisiana Balance of State

Monroe Northeast Louisiana

New Orleans Jefferson Parish

Shreveport, Bossier Northwest Louisiana

Slidell Southeast Louisiana

MA

Attleboro, Taunton Bristol County

Boston

Cambridge

Cape Cod Islands

Fall River

Gloucester, Haverhill, Salem Essex County

Lynn

Massachusetts Balance of State

New Bedford

Pittsfield Berkshire, Franklin, Hampshire Counties

Quincy, Brockton, Weymouth, Plymouth City and County

Springfield Hampden County

Worcester City & County

MD

Annapolis Anne Arundel County

Baltimore

Baltimore County

Carroll County

Frederick City & County

Harford County

Howard County

Mid-Shore Regional

Montgomery County

Prince George’s County

Wicomico, Somerset, Worcester Counties

ME

Maine Statewide

MI

Battle Creek Calhoun County

Dearborn, Dearborn Heights, Westland Wayne County

Detroit

Eaton County

Flint Genesee County

Grand Rapids, Wyoming Kent County

Grand Traverse, Antrim, Leelanau Counties

Holland Ottawa County

Jackson City & County

Lansing, East Lansing Ingham County

Lenawee County

Livingston County

Michigan Balance of State

Monroe City & County

Norton Shores, Muskegon City & County

Pontiac, Royal Oak Oakland County

Portage, Kalamazoo City & County

Saginaw City & County

St. Clair Shores, Warren Macomb County

Washtenaw County

MN

Dakota, Anoka, Washington, Scott, Carver Counties

Duluth St. Louis County

Minneapolis Hennepin County

Moorhead West Central Minnesota

Northeast Minnesota

Northwest Minnesota

Rochester Southeast Minnesota

Southwest Minnesota

St. Cloud Central Minnesota

St. Paul Ramsey County

MO

Joplin Jasper, Newton Counties

Missouri Balance of State

Springfield Greene, Christian, Webster Counties

St. Charles City & County, Lincoln, Warren Counties

St. Joseph Andrew, Buchanan, DeKalb Counties

St. Louis City

St. Louis County

MP

Northern Mariana Islands

MS

Gulf Port Gulf Coast Regional

Jackson Rankin, Madison Counties

Mississippi Balance of State

MT

Montana Statewide

NC

Asheville Buncombe County

Chapel Hill Orange County

Charlotte Mecklenburg County

Durham City & County

Fayetteville Cumberland County

Gastonia Cleveland, Gaston, Lincoln Counties

Greensboro, High Point

North Carolina Balance of State

Northwest North Carolina

Raleigh Wake County

Wilmington Brunswick, New Hanover, Pender Counties

Winston-Salem Forsyth County

ND

North Dakota Statewide

NE

Lincoln

Nebraska Balance of State

Omaha, Council Bluffs

NH

Manchester

Nashua Hillsborough County

New Hampshire Balance of State

NJ

Atlantic City & County

Bergen County

Burlington County

Camden City & County Gloucester, Cape May, Cumberland Counties

Elizabeth Union County

Jersey City, Bayonne Hudson County

Lakewood Township Ocean County

Monmouth County

Morris County

New Brunswick Middlesex County

Newark Essex County

Paterson Passaic County

Salem County

Somerset County

Trenton Mercer County

Warren, Sussex, Hunterdon Counties

NM

Albuquerque

New Mexico Balance of State

NV

Las Vegas Clark County

Nevada Balance of State

Reno, Sparks Washoe County

NY

Albany City & County

Binghamton, Union Town Broome, Otsego, Chenango, Delaware, Cortland, Tioga Counties

Buffalo, Niagara Falls Erie, Niagara, Orleans, Genesee, Wyoming Counties

Columbia, Greene Counties

Elmira Steuben, Allegany, Livingston, Chemung, Schuyler Counties

Franklin, Essex Counties

Glens Falls, Saratoga Springs Saratoga, Washington, Warren, Hamilton Counties

Ithaca Tompkins County

Jamestown, Dunkirk Chautauqua County

Jefferson, Lewis, St. Lawrence Counties

Kingston Ulster County

Nassau, Suffolk Counties

New York Balance of State

New York City

Newburgh, Middletown Orange County

Poughkeepsie Dutchess County

Rochester, Irondequoit, Greece Monroe County

Rockland County

Schenectady City & County

Syracuse, Auburn Onondaga, Oswego, Cayuga Counties

Troy Rensselaer County

Utica, Rome Oneida, Madison Counties

Wayne, Ontario, Seneca, Yates Counties

Yonkers, Mount Vernon Westchester County

OH

Akron, Barberton Summit County

Canton, Massillon, Alliance Stark County

Cincinnati Hamilton County

Cleveland Cuyahoga County

Columbus Franklin County

Dayton, Kettering Montgomery County

Ohio Balance of State

Toledo Lucas County

Youngstown Mahoning County

OK

Norman Cleveland County

North Central Oklahoma

Northeast Oklahoma

Oklahoma Balance of State

Oklahoma City

Southeastern Oklahoma Regional

Southwest Oklahoma Regional

Tulsa City & County

OR

Central Oregon

Clackamas County

Eugene, Springfield Lane County

Hillsboro, Beaverton Washington County

Medford, Ashland Jackson County

Oregon Balance of State

Portland, Gresham Multnomah County

Salem Marion, Polk Counties

PA

Beaver County

Bristol, Bensalem Bucks County

Chester County

Eastern Pennsylvania

Erie City & County

Harrisburg Dauphin County

Lancaster City & County

Lower Merion, Norristown, Abington Montgomery County

Philadelphia

Pittsburgh, McKeesport, Penn Hills Allegheny County

Reading Berks County

Scranton Lackawanna County

Upper Darby, Chester, Haverford Delaware County

Western Pennsylvania

Wilkes-Barre, Hazleton Luzerne County

York City & County

PR

Puerto Rico Balance of Commonwealth

South-Southeast Puerto Rico

RI

Rhode Island Statewide

SC

Charleston Low Country

Columbia Midlands

Greenville, Anderson, Spartanburg Upstate

Sumter City & County

SD

South Dakota Statewide

TN

Appalachian Regional

Central Tennessee

Chattanooga Southeast Tennessee

Jackson West Tennessee

Knoxville Knox County

Memphis Shelby County

Morristown Blount, Sevier, Campbell, Cocke Counties

Murfreesboro Rutherford County

Nashville-Davidson County

Upper Cumberland

TX

Amarillo

Austin Travis County

Bryan, College Station Brazos Valley

Dallas City & County, Irving

El Paso City & County

Fort Worth, Arlington Tarrant County

Houston, Pasadena, Conroe Harris, Fort Bend, Montgomery Counties

San Antonio Bexar County

Texas Balance of State

Waco McLennan County

Wichita Falls Wise, Palo Pinto, Wichita, Archer Counties

UT

Provo Mountainland

Salt Lake City & County

Utah Balance of State

VA

Alexandria

Arlington County

Charlottesville

Fairfax County

Fredericksburg Spotsylvania, Stafford Counties

Harrisonburg, Winchester Western Virginia

Loudoun County

Lynchburg

Newport News, Hampton Virginia Peninsula

Norfolk, Chesapeake, Suffolk Isle of Wight, Southampton Counties

Portsmouth

Prince William County

Richmond Henrico, Chesterfield, Hanover Counties

Roanoke City & County, Salem

Virginia Balance of State

Virginia Beach

VI

Virgin Islands

VT

Burlington Chittenden County

Vermont Balance of State

WA

Everett Snohomish County

Seattle King County

Spokane City & County

Tacoma, Lakewood Pierce County

Vancouver Clark County

Washington Balance of State

WI

Madison Dane County

Milwaukee City & County

Racine City & County

Wisconsin Balance of State

WV

Charleston Kanawha, Putnam, Boone, Clay Counties

Huntington Cabell, Wayne Counties

West Virginia Balance of State

Wheeling, Weirton Area

WY

Wyoming Statewide

The death penalty as a lens on democracy

Who is really in charge? In democracies, policies are *correlated* with public opinion, but why? The obvious explanation is that people choose representatives, and those representatives give them what they want. But maybe the causal arrow points in the other direction—maybe elites choose policies, and the public gradually figures that since that's how things are, it must be right.

To get some perspective on this question, I looked for an issue to use as a case study. The ideal issue would have divergence between elite and public opinion, plus have a long history in different countries, so we can see how things played out under different conditions.

The best fit I could find is the death penalty. It's perfect in terms of having a long history, good records, and being prominent enough that the public *has* an opinion. The primary disadvantage is that it's, you know, a massive bummer.

So, here's our goal: Around the world today, the death penalty is less popular in countries where it is banned. But which caused which?

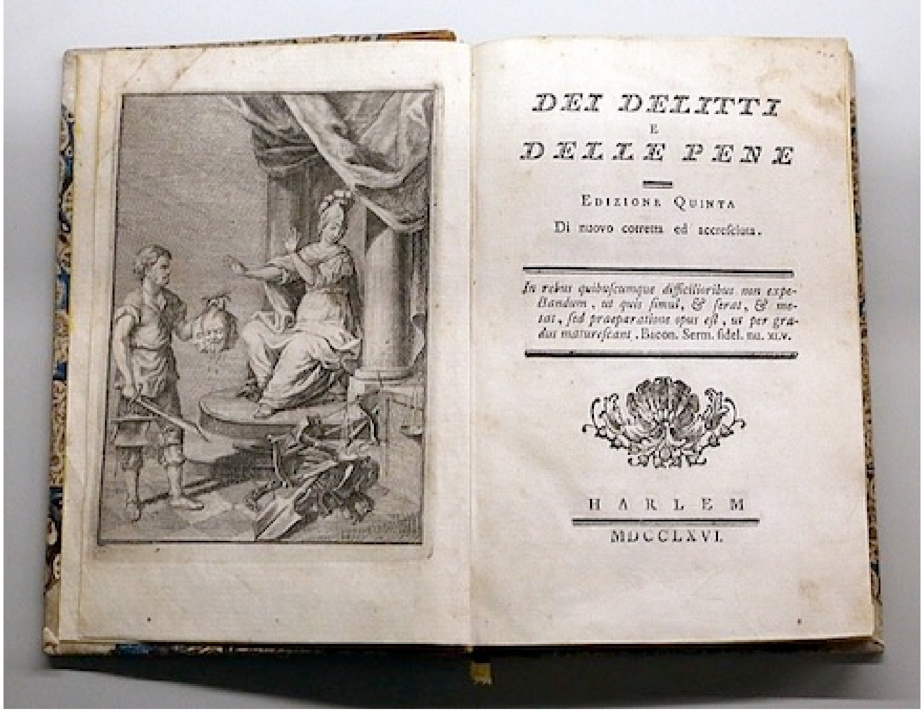
The movement to abolish the death penalty arguably began in the West in 1764. This is when, at the age of 26, Cesare Beccaria anonymously published *On Crimes and Punishments*, a call for applying enlightenment values to criminal law. He argues poetically for a break with history:

> If it is objected that almost all times and almost all places have used the death penalty for some crimes, I reply that the objection collapses before the truth, against which there is no appeal, that the history of mankind gives the impression of a vast sea of errors, among which a few confused truths float at great distances from

> each other. Human sacrifices were common to almost all nations; but who would dare to justify them? That only a few societies have given up inflicting the death penalty, and only for a brief time, is actually favourable to my argument, because it is what one would expect to be the career of the great truths, which last but a flash compared with the long and dark night which engulfs mankind.

He suggests that punishments should be set at the lowest level so that a rational person would lose more from the punishment than they gain from the crime. Moreover, he claims a life of hard labor exceeds *any* potential gain, and is even *more* of a deterrent than execution since it avoids the spectacle and leaves the person around to serve as a reminder of what happens if you misbehave.

This book had a huge and immediate influence. Voltaire published an—also anonymous—positive commentary that spread the book's fame as far as America where it was discussed by John Adams and Thomas Jefferson. (Jefferson was partly influenced by a different argument in the book against gun control.) *On Crimes and Punishments* was also quickly banned by the Venetian and Roman Inquisitions.



DEI DELITTI
E
DELLE PENE

EDIZIONE QUINTA
Di nuovo corretta ed accreſciuta.

In rebus quibuſcumque difficilioribus non expectandum, ut quis ſimul, & ſerat, & metat, ſed praeparatione opus eſt, ut per gradus matureſcant. Bacon. Serm. fidel. nu. XLV.

HARLEM
MDCCLXVI.

Beccaria's writing is clear, urgent, and seems to speak across time. He sort of resembles an earlier but more influential Peter Singer, another philosopher who was able to shape public opinion.

Anyway, while this all sounds very idealistic, here's the very next passage:

> The voice of a philosopher is too weak against the uproar and the shouting of those who are guided by blind habit. But what I say will find an echo in the hearts of the few wise men who are scattered across the face of the earth. [...]
>
> How happy humanity would be if laws were being decreed for the first time, now that we see seated on the thrones of Europe benevolent monarchs, inspirers of the virtues of peace, of the sciences, of the arts, fathers of their people, crowned citizens. [...] That is a reason for enlightened citizens to wish all the more fervently for their authority to continue to increase.

Put less politely:

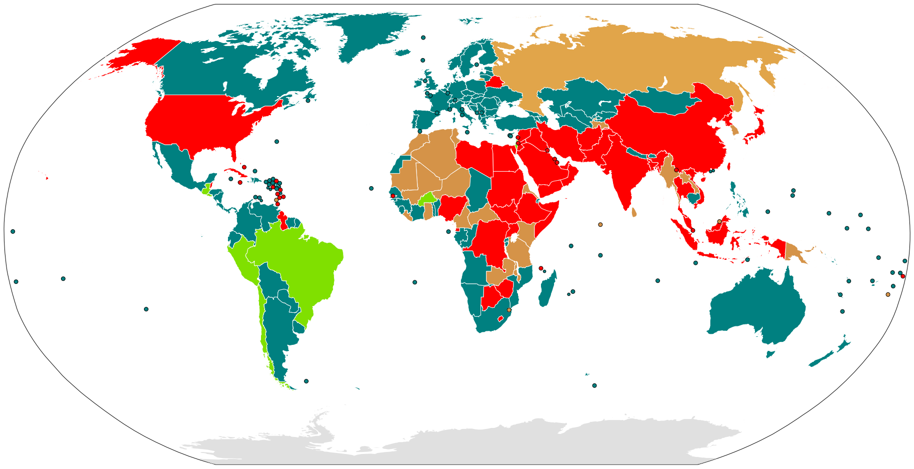
1. While the death penalty is wrong, the teeming masses will never understand why.
2. But thankfully Great Leaders are in charge, not the public.
3. In fact, this shows exactly why monarchies are so important—let's make monarchies even more powerful!

This is less comfortable for someone sitting in a democracy a few centuries later.

Now, this might just be some pragmatic flattery, and anyway I don't judge monarchists so far in the past. But whatever the motivations, look at what Beccaria is suggesting: We need an elite—monarchs are certainly elite—to impose enlightened policies on the public. In fact, this happened. Tuscany, Russia, and Austria had monarchs who banned capital punishment soon after this book came out (though the bans didn't last forever).

But that's not our story. We're interested in the history of the death penalty in *democracies*, which happens much later. In the next few posts, we'll go over the history of the death penalty in Germany, the UK, and France, which abolished the capital punishment in 1949, 1969, and 1981. We'll then contrast this to the U.S., where it remains in use. Our focus is not if the death penalty is good or right, but how these policies came to be, and what the public thought about them over time.

Now, why is it worth going through all this history?



You probably know that the U.S today is the only Western democracy that maintains the death penalty. (See also: Belarus, Belize, Japan, Taiwan.) But did you know the following?

- When Germany, the U.K., and France banned the death penalty, they did so despite the opposition of 2/3 of their populations?
- For much of the 1960s, public support for the death penalty in the U.S. was *lower* than in these other countries, despite the U.S. having an order of magnitude higher levels of violent crime?
- In the U.K. and France, something like half of the population still supports the death penalty *today*, only slightly less than in the U.S.?

It's not obvious that the U.S. is the country that needs explaining here! If your model of the world is that the death penalty is bad and democracies stop doing bad things, then sure, the U.S. is puzzling. But if your model is that democracies do what people want, then how did so many countries implement policies that were against the will of their citizens?

How Germany banned the death penalty

Germany came close to banning the death penalty several times in its early history:

- In the 1848 revolution, the new constitution almost completely banned it, but this was immediately overturned by the conservative restoration.
- When Germany unified in 1870-1871 it again came very close to abolition before a late intervention from Bismarck caused several delegates to switch their votes.
- In the Weimar republic in 1919, there was a vote that failed probably just because a bunch of Social Democrats weren't physically present to vote.

- Finally, in 1928 some proceedings ended in a 14-14 deadlock, after which the Düsseldorf Vampire committed a series of sex murders that were covered in lurid detail by the press, putting abolition in retreat.

Then there was the Nazi era. Nearly 20,000 death sentences were handed down by German courts and carried out, sometimes for crimes as simple as making a critical remark, or just at a judge's discretion.

After the war in 1949, a Parliamentary Council was working on the constitution for West Germany, when, to everyone's surprise, a delegate from the nationalist Deutsche Partei proposed a constitutional ban on capital punishment. This delegate, a wealthy industrialist who had collaborated with the Nazi regime, urged abolition to express "revulsion at the large number of death sentences carried out in the last few years", meaning both executions by the Nazis and post-war executions of war criminals by the occupying powers.

The Social Democratic Party was hesitant to support this motion, but came around on the third reading, along with several mainstream conservatives. One member of the Christian Democratic Union argued that abolition should be left to the future parliament for the sake of democratic legitimacy. Others argued anyway that the constitution was just a stopgap before reunification of East and West Germany, and that it didn't need to be perfect.

In the end, Article 102 of the constitution wasted no words:

> Die Todesstrafe ist abgeschafft.

The public was shocked to find the death penalty was constitutionally banned. A 1949 opinion poll reportedly found 77% of Germans were in favor of capital punishment with 18% opposed.

There were many attempts to bring capital punishment back, but these failed for three reasons.

First, and most importantly, a 2/3 majority of parliament was needed to change the constitution, a very high bar.

Second, in Germany's system, all members of parliament—either locally elected or proportionally chosen—are selected *by the parties* after an internal vetting process. Because most elites favored abolition, they picked candidates who did as well. Politicians openly spoke for a Burkean model of politics where a "natural aristocracy" would carefully consider an issue. Because constituents have no expertise and don't participate in the debate, their opinions shouldn't be taken too seriously.

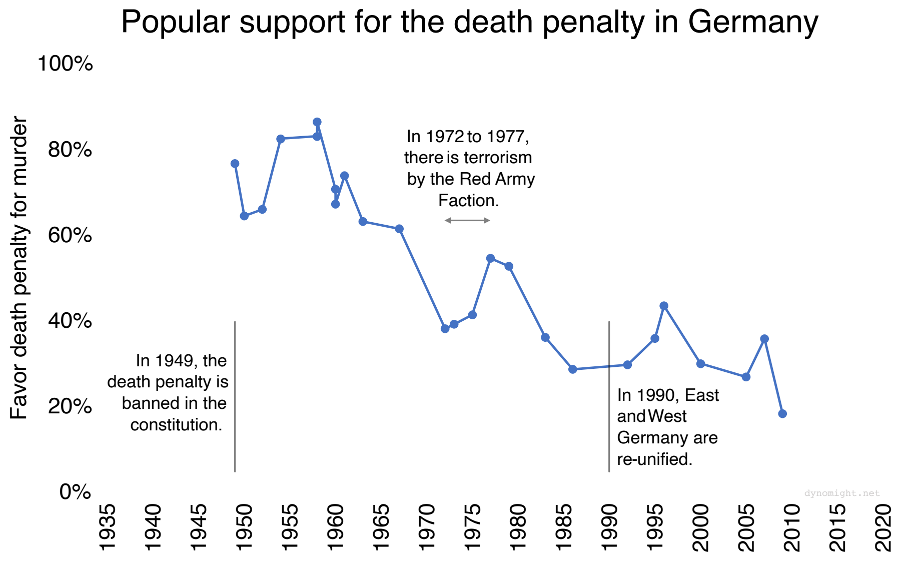
Third, there was the *Großen Strafrechtskomission*, formed in 1954 to reform all the nation's criminal laws. People who wanted to bring back the death penalty were often told to wait for this commission to finish, which took years. While a majority of the population supported the death penalty, almost all the experts they heard from opposed it. One man who carried out several executions during the war and said the first execution horrified him but he found the fourth routine, and that killing shouldn't be normalized.

In 1960, the commission finally voted 19-4 against re-introducing the death penalty on the logic that there was no deterrence effect, popular opinion was "unfounded" and it would be unstable to change the constitution so soon. In any case, this was just a recommendation, and even the recommendations the commission *did* make weren't taken up for several years, leaving it as an active reason to postpone any action. After this process was complete, people who wanted to reinstate the death penalty were told that a commission had just looked into that.

In the 1960s, public opinion started to change. A large group of German and Swiss law professors created a draft of the penal code that would not only maintain the abolition of the death penalty, but ban hard labor, decriminalize homosexuality, and generally orient the entire system towards rehabilitation. This had a strong influence in Switzerland, Scandinavia, Austria, Brazil, and Argentina but not—immediately—Germany itself.

By the time a new social-liberal coalition took power in Germany in 1969, the death penalty was no longer an active issue. Instead, the system was reformed in the other direction, to eliminate all punishment and retribution in sentencing, focusing only on rehabilitation and protecting the public.

There was a brief spike in support for the death penalty following the Red Army Faction's terrorist attacks in the late 1970s. Otherwise, the death penalty's support continued to decline. By 2021, the death penalty is so unpopular that it is rarely even polled. The last poll I could find was 2009, where it had 18.5% support.



The plot above assembles various polls, with slightly different wordings. Beware that polling on the death penalty is always extremely squishy. If you ask people, "Do you support the death penalty?", you get much lower support than asking, "Do you support the death penalty for [insert specific extremely heinous crime]?"

To keep things somewhat level, I've only used polls that ask about the death penalty *in general* or the death penalty *for murder*. Even so, different pollsters often get different numbers with similar questions at similar times, so take everything with a large implied level of uncertainty. Polling is hard.

Summary: Germany abolished the death penalty while writing the constitution in 1949. This was possible because of an alliance between the left and certain right nationalists who wanted to protest executions of Nazi war criminals. There was clear public support for the death penalty at the time, but the ban was virtually impossible to overturn since it was in the constitution. In the next seven decades, the public gradually came to oppose the death penalty, and then oppose it strongly.

What does it all mean? Did public support decline *because* the death penalty was banned? Let's come back to that after looking at some other countries.

Raw data and sources

In all these histories of European countries, I rely heavily on Andrew Hammel's excellent 2010 book, *Ending the Death Penalty: The European Experience in Global Perspective*.

Anyway, here is the raw data included in the plot above. For all polls, I converted to a single "support" score by removing respondents who gave "no opinion". For example, a poll that had 50% support, 25% opposition, and 25% no opinion would convert to 2/3 support support. An exception is if I could only find the "support" number in which I used that unchanged.

year	support	oppose	source
1949	55	??	Allensbach
1950	55	30	IPSOS
1952	55	28	IPSOS
1954	72	15	DIVO
1958	75	15	DIVO
1958	78	12	INRA
1960	71	??	Allensbach
1961	63	22	INRA
1963	52	30	IPSOS
1967	50	31	BSAS
1972	33	53	BSAS

year	support	oppose	source
1973	30	46	BSAS
1975	35	49	BSAS
1977	45	37	BSAS
1979	44	39	BSAS
1983	28	49	BSAS
1986	24	59	BSAS
1992	24	56	BSAS
1995	30	53	BSAS
1996	35	45	BSAS
2000	23	53	BSAS
2005	22	59	BSAS
2007	35	62	IPSOS
2009	15	66	BSAS

- The Allensbach polls quoted in this article.
- The 1950, 1952, 1958, and 1963 IPSOS polls I found somewhere and, uhhh, lost the source and can't find it anymore. I'll dig it up if enough people harass me.
- The DIVO and INRA polls are quoted by Erskine (1970).
- The BSAS polls are quoted by Hammel in his book which, again, is really good.

How the United Kingdom banned the death penalty

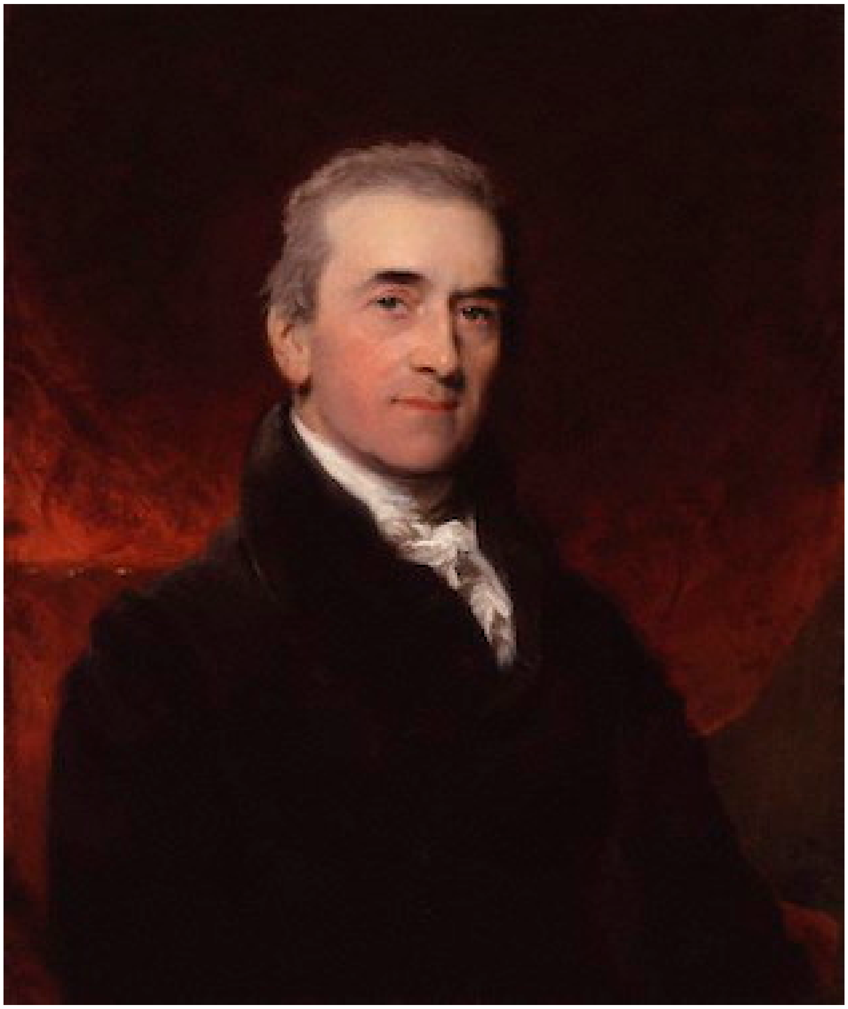
When Beccaria wrote *On Crimes and Punishments* in 1764, there were around 150 crimes punishable in Britain by death. This "bloody code" included crimes as small as the theft of some items worth 1 shilling. For context, a skilled worker at the time could earn around 20 shillings in a week, and 1 shilling in 1764 inflation adjusts to £7.25 ($9.75) today.

This was an embarrassment to the British Crown, always eager to see itself as superior to the continent. Britain had abolished torture, which most of Europe had not. On the other hand, while much of Europe maintained the death penalty, it was rarely applied to crimes like petty theft.

It's still debated why so many offenses we punishable by death. One popular theory is that it was driven by social disorder resulting from the industrial revolution. I find this theory strange since many of these laws predate the industrial revolution's start around 1760. Another theory is that power in Britain came from Parliament, whereas in much of Europe it came from monarchs who didn't need to worry about a backlash from the public.

Most potential executions weren't carried out, due to slack in the system: Prosecutors, victims, and juries often tried to convict on lesser charges. There were legal loopholes and various clemencies. When people were actually executed, it seemed less due to their crimes and more to bad luck.

Some did defend the system. William Paley said you could deter crime by either having capital punishment for few offenses and always carrying it out, or having it for many offenses and carrying it out rarely, and the latter would be more deterrent. This is a kind of prelude to Gary Becker's thinking about crime and expected value theory in the 1970s.



An initial push against this system was led by Samuel Romilly, the handsome fellow above. Romilly was a self-educated lawyer who grew up in a French family in London, traveled in Europe, and also opposed the slave trade. He did not oppose the death penalty for murder but argued it was ridiculous to apply such varying penalties to the same crimes. While many other members of Parliament were hostile to his suggestions, he managed to repeal the death penalty for things like pickpocketing. He also obtained a ban execution by drawing and quartering (bleak link; do not click).

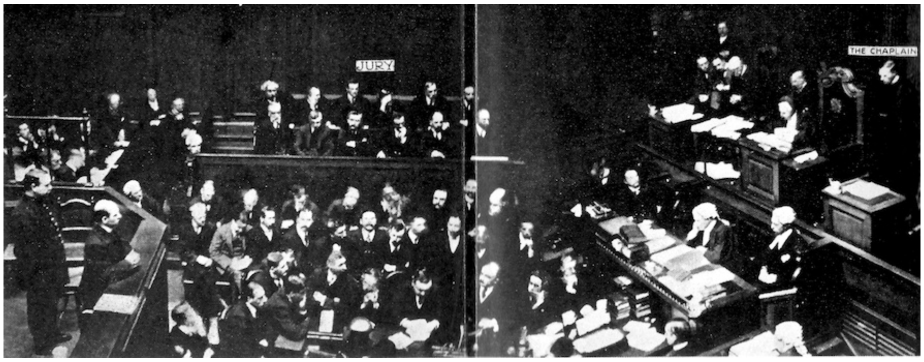
The movement against capital punishment was strengthened by Jeremy Bentham, the founder of utilitarianism. In 1832, Parliament passed the *Punishment of Death, etc. Act*, which eliminated the death penalty from around 2/3 of offenses (and apparently did not suffer from overthinking when being named). Finally, the reformist Whig leader John Russell convinced even the House of Lords that the current system was a disgrace, and in 1837 the “bloody code” was dismantled, leaving the death penalty only for serious crimes.

In 1840, William Makepeace Thackeray published Going to See a Man Hanged. He describes setting out on an adventure, and gradually becoming horrified at the atmosphere outside the prison with people selling food, telling jokes, and climbing in trees. He says of the condemned man, “it is painful to see how he fastens upon everybody who approaches him, how pitifully he clings to them”. At the end of the day, he says “I feel myself ashamed and degraded at the brutal curiosity which took me to that brutal sight”. A sign of a changing culture, this is supposedly the first time pity for a murderer appeared in print.

Public opinion continued to move against the death penalty. Improvement in policing meant more crimes were solved, and more prisons meant other penalties existed, meaning there was less a need to "make an example" of people. Further reforms followed. By the 1860s, the death penalty was, in practice only applied for murder, and in 1868 Parliament passed *A Bill to Provide for Carrying Out Capital Punishment in Prisons* (rather than outside them), another testament to trend of laws with very explicit names.

At this point, movement slowed. John Stuart Mill had earlier opposed the death penalty but felt that reforms justified it. Speaking of continental Europe (axiomatically worse than Britain) he says there are "great and enlightened countries, in which the criminal procedure is not so favorable to innocence, does not afford the same security against erroneous conviction, as it does among us." That is, it was OK to have the death penalty because Britain's awesome legal system meant an innocent person could never be condemned.

Here's what a court looked like handing down a death sentence in 1912:



Not much changed until 1923 when the Labour party made abolition an explicit position. This led to the creation of a select committee on capital punishment, but this was stacked somewhat in favor of retention, leading it to a divided verdict and no action from the Labour government.

Then came Sydney Silverman, an interesting character who went to jail as a conscientious objector during WWI, lectured at the University of Finland, moved back to Liverpool, became a lawyer, joined the Labour party, and won an election to represent Lancashire in Parliament in 1935. Silverman was a militant socialist. His personality was described as "opinionative, dogmatic, assertive, and quarrelsome" by a Labour *ally.* He was so disliked by Conservatives that several who privately favored abolition refused to support it in public simply because they hated Silverman so much.

In 1938, a Conservative member of Parliament suggested a motion to abolish the death penalty. This was done as a "free vote", meaning members weren't bound by the position of their party. The vote passed 114-89. Except this was all non-binding, and so Chamberlain's government simply ignored it.

As elsewhere, WWII paused any movement on death penalty abolition (and reversed Silverman's pacifism). When Labour formed a government after the war, they didn't have the nerve to include abolition in their big criminal justice bill, probably because support for the death penalty had risen during the war. Still, in around 1947—I can't figure out the exact date—they allowed it to be raised as an independent issue. Many spoke that the moment was not right, with the public strongly opposed. But Silverman argued passionately for ignoring public opinion:

> We are not delegates; we are not bound to ascertain exactly what a numerical majority of our constituents would wish and then to act accordingly without using our judgment. Edmund Burke long ago destroyed any such theory. We are not delegates. We are representatives.

The vote passed 252-222 and the bill went to the House of Lords. At the time, many understood the Lords to have a special prerogative to examine only certain bills, somewhat akin to how the Supreme Court functions in the United States. So it was controversial that they would even consider rejecting this bill. But they rejected it anyway, 181 to 28, partially justified by the fact that this was not an "official" Labour bill.

After this rejection, the House of Commons passed a compromise bill that would have restricted but not eliminated capital punishment. In the House of Lords, Churchill denounced it as out of step with the nation, and the compromise was also rejected. Finally, Labour set up a special commission that met for four years before reporting in 1953, at which point Churchill was again in government. He delayed debate to 1955, and those recommendations were rejected, too.

Public opinion in this era may have been influenced by the case of Timothy Evans, who was executed in 1950 for the murder of his wife and daughter. After their death, he made suspicious and contradictory statements about trying to abort a new baby (abortion was illegal) eventually saying that his neighbor John Christie had offered to perform an abortion, then said she died in the process, and Evans should leave London for his own safety. Some speculate that Evans was mentally limited and not capable of fully understanding what was happening. Three years after the execution, it emerged that Christie was a serial killer and had murdered several women in the same building. This possible execution of an innocent man was in the news for years. (An official report in 1966 came to the strange conclusion that Evans probably killed his wife but that Christie had killed the daughter. Nevertheless, the Queen gave Evans a posthumous pardon.)

In 1956, Silverman was able to get a five-year moratorium on capital punishment passed 286 to 262, with support from Labour and a growing number of Tories. In committee, various amendments were proposed. These were defeated, except one that would keep the death penalty for murder committed by someone already serving a life sentence. When Silverman couldn't remove this amendment, he used a series of arcane motions to *remove all the words* from the amendment except for the meaningless fragment “provided that this”. Even many in Labour thought this was beyond the pale, but it worked and the bill passed.

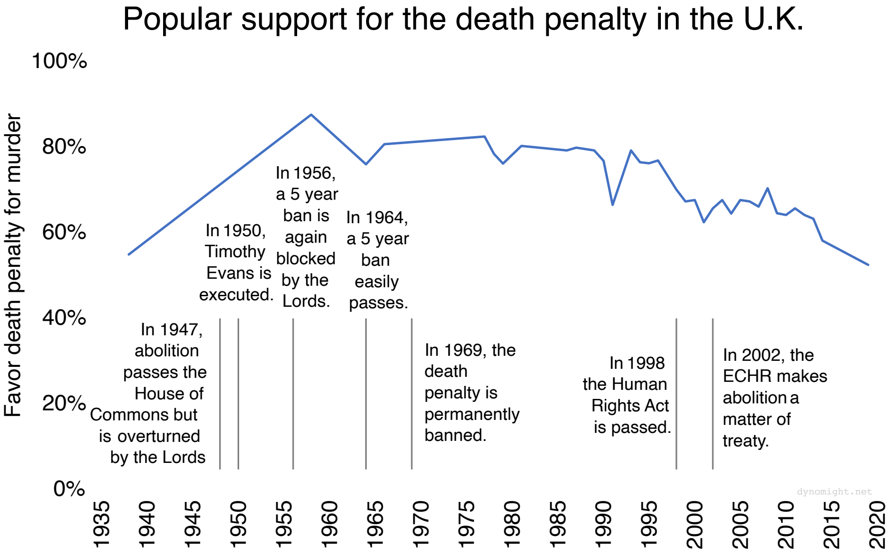
Again this bill went to the House of Lords. Based partially on the earlier controversy, reforms had been passed to limit their power to reject most legislation. But no matter, they rejected it anyway.

Eventually, the Queen publicly called to limit the scope of capital punishment. The Tories passed the Homicide Act of 1957 which reduced the death penalty to specific types of murders, e.g. murder during theft or murder of police. The goal of these limits was to stave off calls for complete abolition. However, they were immediately criticized. For example, people questioned why murder in the course of rape was excluded.

Finally, in 1964, Labour obtained a small majority. There was a fierce debate, in which several conservative members of Parliament described converting to abolition. In the end, they easily passed a five-year moratorium. In a sign of the changing times, this got an even larger majority in the House of Lords. Silverman did not quite live to see abolition being made permanent in 1969, with support from the leaders of all three major parties.

Various conservatives have made attempts to reintroduce capital punishment in 1973, 1975, 1979, 1983, 1987, 1988, 1990, and 1994, but these were defeated with ever-larger majorities.

In 1998, the Human Rights Act (HRA) completely abolished capital punishment (it was still in principle allowed for military executions). In 2002, The UK signed protocol 13 of the European Convention on Human Rights (ECHR), meaning that the UK could not reinstate the death penalty without leaving the EU.



In the last poll I could find, from 2019, public support was still just above 50%. This might seem to contradict all the articles from 2015 titled “Support for death penalty drops below 50% for the first time”. I think those articles are misleading—the data comes from the British Social Attitudes survey where you can easily see that the number of people who supported the death penalty *was still clearly larger* than the number who opposed it. It’s just that active support was less than 50% due to the fact that 18% of people have no opinion. My chart above shows support among people who do have an opinion.

Anyway, now that the UK *has* left the EU, could it reinstate the death penalty? As far as I can tell, in principle, a simple act of Parliament could overturn the Human Rights Act and put the death penalty back in place. But in negotiating the post-Brexit EU-UK trade agreements, the EU insisted that the UK continue to enforce the European Convention on Human Rights. If the UK wanted to put the death penalty back in place, it would have no trading agreement with any of the EU.

Summary: Shortly after WWII, the population strongly supported the death penalty, yet leftist elites united behind abolition and rightist elites were divided. After a series of battles between the House of Commons and the House of Lords a ban finally went in place in 1964. Public support for the death penalty was stable (and positive) for around 30 years, and then started slowly declining to something near an even split today.

What does it all mean? Well, it’s hard not to be impressed by how… irrelevant public opinion seems to the fierce battles between different elites. It’s also striking how the UK, like Germany, moved to take the issue out of the democratic domain. But let’s save the theorizing to after we’ve covered France and the U.S.

Data and sources

year	support	oppose	source
1938	49	40	SOC
1958	79	11	INRA
1964	67	21	SOC
1966	76	18	SOC
1977	81	17	IPSOS
1978	77	21	IPSOS
1979	74	23	IPSOS
1981	78	19	IPSOS
1986	73	19	BSAS
1987	72	18	BSAS
1989	73	19	BSAS
1990	70	21	BSAS
1991	58	29	BSAS
1993	73	19	BSAS
1994	69	21	BSAS
1995	68	21	BSAS
1996	67	20	BSAS
1998	59	25	BSAS
1999	58	28	BSAS
2000	59	28	BSAS
2001	52	31	BSAS
2002	56	29	BSAS
2003	59	28	BSAS
2004	53	29	BSAS
2005	59	28	BSAS
2006	58	28	BSAS
2007	57	29	BSAS
2008	60	25	BSAS
2009	55	30	BSAS
2010	54	30	BSAS
2011	56	29	BSAS
2012	54	30	BSAS
2013	52	30	BSAS
2014	49	35	BSAS
2019	40	36	YouGov

More details:

- The OSC and INRA polls are quoted by Erskine (1970).

- The IPSOS polls I found somewhere and, uhhh, lost the source and can't find it anymore. I'll dig it up if enough people harass me.
- The BSAS numbers are from the British Social Attitudes survey. Despite being funded by charity and the government they don't publish their actual data, so I estimated the numbers manually from this low quality fiure. They should be accurate to within a point or two.

How France banned the death penalty

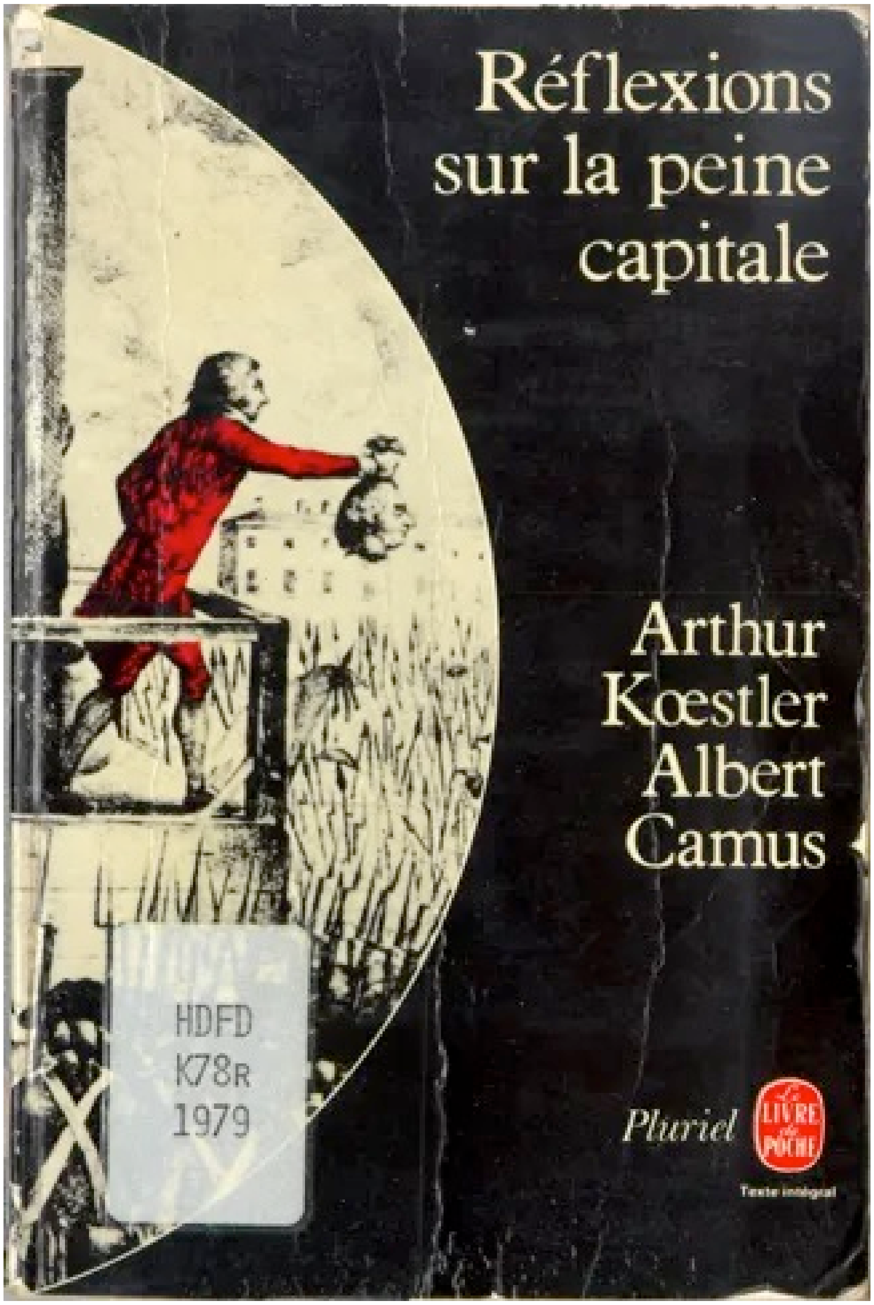
Capital punishment was debated during the French revolution (1789-1799). Due to the influence of Beccaria and Voltaire, the discussion was similar to how the death penalty is discussed today. Robespierre said "The state's execution of the death penalty is legalized murder." (Though as Hammel points out, "given his later role in the Terror, one has to wonder about Robespierre's sincerity".)

While the death penalty was not eliminated, they did decide a more humane method was needed. This led to the guillotine, which France continued to use until abolition in 1981.

In the mid-1800s, the Radical movement started to form in France, the goals of which were broadly to realize the original enlightenment values of the French revolution. The Radical party was formed in 1901 and in 1906 Armand Fallières became president. Fallières was a committed abolitionist and for several years he commuted all death sentences. This was controversial, as many felt it trampled on the rights of *jurys*, citizen bodies that had assisted French judges in decisions and sentencing since the Napoleonic era. There was a debate in 1908, which abolitionists lost, in part due to a perception that public opinion was in favor of retention.

In 1939, the behavior of the crowd at the public execution of Eugen Weidmann caused a scandal, and President Lebrun immediately banned public executions.

During the Vichy regime, judicial executions were fairly rare, something like 50 total. (Of course, this excludes deaths resulting from deporting French Jews and members of the resistance.)



Albert Camus’ 1957 essay Reflections on the Guillotine had a major influence on intellectual thought in France, possibly due to Camus winning the Nobel Prize in the same year. Camus describes how his father, a supporter of the death penalty, saw an execution and was so horrified that after returning home he vomited and spent several days in shock. Many of Camus' arguments are similar in spirit to those of Beccaria, though he doesn't agree with everything:

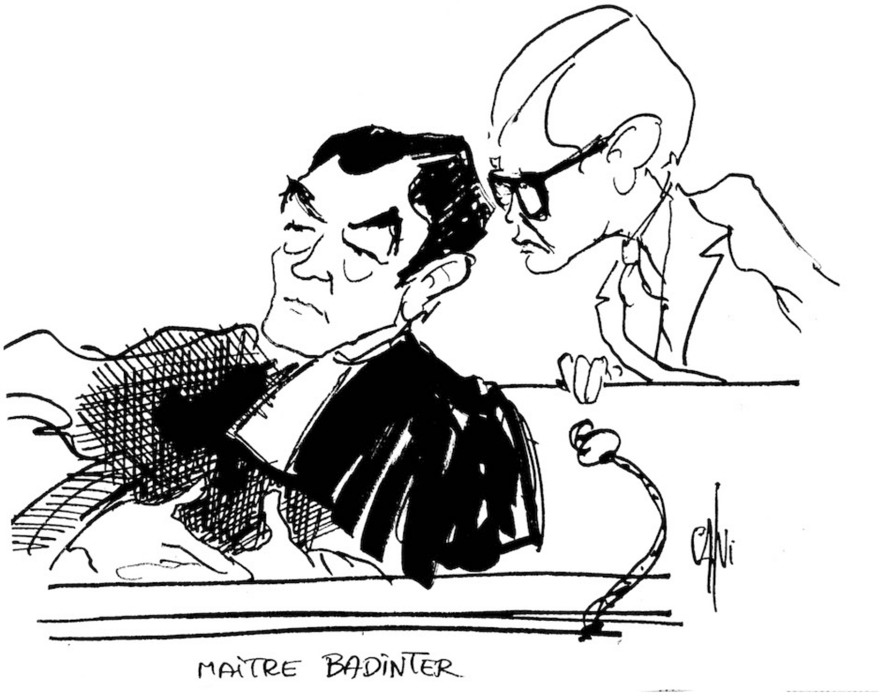
> Capital punishment would then be replaced by hard labor—for life in the case of criminals considered irremediable and for a fixed period in the case of the others. To any who feel that such a penalty is harsher than capital punishment we can only express our amazement that they did not suggest, in this case, reserving it for such as Landru and applying capital punishment to minor criminals.

(Camus died in a car accident in 1960.)

Throughout the 1960s and 1970s, the death penalty was used less and less in France. This was largely because French presidents had the power to commute death sentences. De Gaulle used this power aggressively, and later presidents were *very* aggressive, with only six during the presidencies of Poher, Pompidou, and Giscard d'Estaing. Executions typically happened only for the murder of a child with sexual overtones, and when guilt was beyond all doubt. Unlike other places, there were no incidents in France with sympathetic or possibly innocent people being executed. This might have kept abolition out of the public consciousness.

In 1972, two prisoners attempted to escape from Clairvaux prison, during which a guard and a nurse were killed. While one prisoner (Claude Buffet) took full credit for both killings, the other prisoner (Roger Bontems) was also sentenced to death. Bontems' lawyer, Robert Badinter, was outraged that someone was executed even though he had never killed anyone. Badinter dedicated himself to abolition, and is prominent in the rest of our story. Throughout the 1970s, he eloquently defended many criminals where the prosecutor had sought the death penalty, often winning life sentences instead.

However, public opinion seemed to move against abolition. There were a series of brutal crimes in France in the 1970s, covered in great detail in the press. The most notorious was Patrick Henry, an ordinary-looking computer programmer who in 1976 kidnapped 8-year old Philippe Bertrand for ransom money. While the case was being investigated, Henry gave an interview in which he said that he hoped that kidnapper would be executed. When he realized there was too much publicity for him to get a ransom, he strangled the girl.



Badinter was hesitant to defend Henry, as he showed no remorse. Still, he ultimately decided to take the case on the theory that if he could win life in prison for this most detestable of murderers, it would prove that the death penalty was never necessary. He was able to elicit ambiguous statements from psychiatrists and, after a beautiful speech, the jury chose life in prison. (Henry was paroled in 2001, returned to prison in 2003 for drug smuggling, and was released again in 2017 a few months before his death.)

The outcome of the Henry trial in 1977 was met with outrage, and Bedinter began to feel that a grass-roots movement for abolition could not succeed.

There was another concerning development for abolitionists. Previously, *jurys* were selected from lists made by municipal authorities, who tended to select more educated people. But in 1978, the French code was modified so that these were drawn uniformly from the population. While many people like Bedinter welcomed the democratic spirit of this move, they also realized that this was likely to lead to more executions.

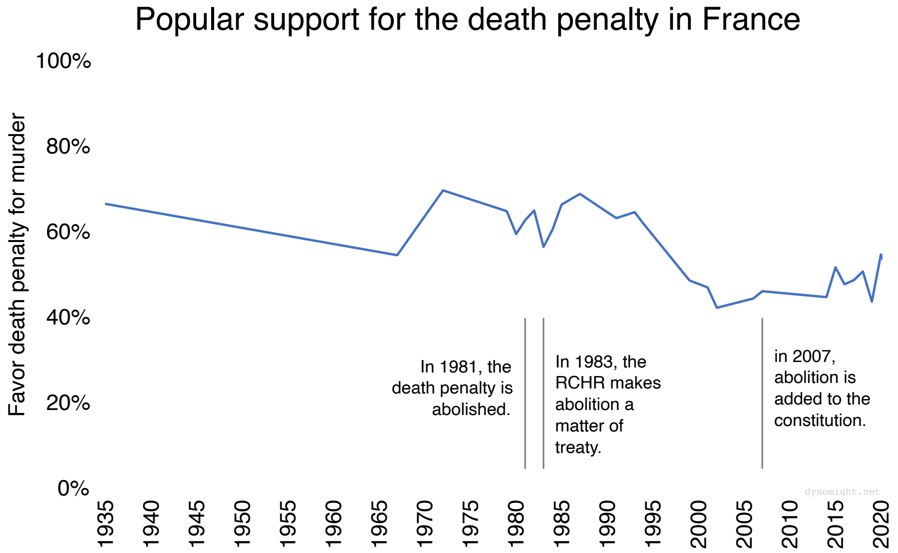
While the public wasn't changing, *elite* opinion was swinging strongly towards abolition. There were numerous reports from judges' unions and various law-reform committees that all suggested abolition. These pushed the issue forward and eventually forced all the major parties to take positions. Ultimately, the leftist parties lined up clearly in favor of abolition, while the rightist parties were divided. Activists like Bedinter had been able to make abolition a key part of the left's profile.

In the late 1970s, the center-right government of President Giscard become unpopular due to a perceived inability to deal with the *choc pétrolier* caused by the 1979 drop in oil production. It became clear that the leftist parties were likely to win government.

Before the 1981 elections, Bedinter approached François Mitterrand and asked him to make abolition an official position of the Socialist party. The logic was that two-thirds of the population supported the death penalty, and a change could only be seen as legitimate if it was promised before the election. The two were already close as a result of Bedinter supporting Mitterrand's 1974 bid for the presidency. During an interview, Mitterrand said, "In my innermost conscience, like the churches [...] and international humanitarian associations, in my heart of hearts, I am opposed to the death penalty".

Mitterrand won the 1981 election on May 10 after which, in the June legislative elections, the leftist parties won huge majorities. Mitterrand eventually named Badinter Justice Minister. He reasoned that speed was of the essence, since French *jurys* were handing out ever-larger numbers of death sentences, and the unity of the left wasn't going to hold forever. Bedinter presented abolition alone as a single bill, which was quickly scheduled for debate on September 17.

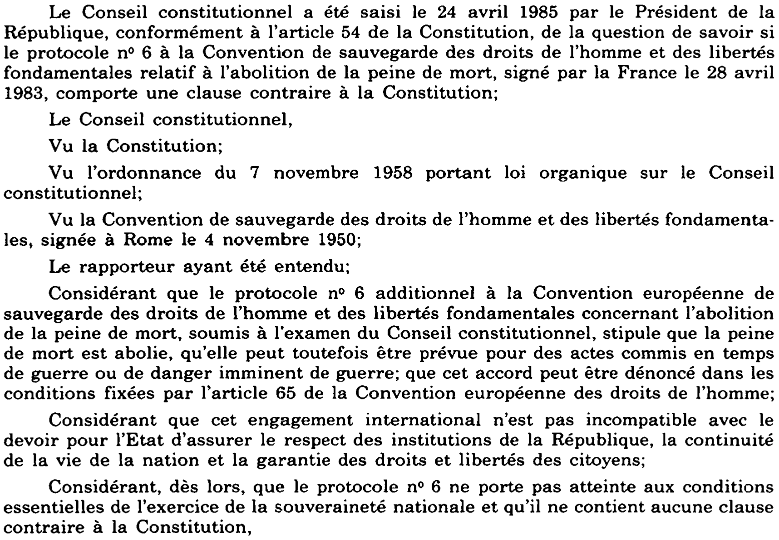
Bedinter knew the public was likely to be outraged at abolition. Nevertheless in the debate, he framed the bill as the next step of human progress, a step that a great nation like France should be ashamed not to have taken yet. The bill passed 363 to 117.


Popular support for the death penalty in France

Some have portrayed Mitterrand as an unprincipled opportunist, because as justice minister during the Algerian War he had ordered the executions of 45 people. While it's certainly true that he changed his position, it's important to remember that the "opportunity" was only with other elites—the public still clearly supported the death penalty and Mitterand knew it.

To make it harder to move backward, in April 1983, France signed Protocol 6 of the European Convention on Human Rights, making it a matter of treaty. The French constitution states that treaties supersede acts of Parliament.

The *Conseil Constitutionnel* met to decide if this was constitutional. In *Décision n° 85-188 DC du 22 mai 1985*, they said they saw no problem.



Le Conseil constitutionnel a été saisi le 24 avril 1985 par le Président de la République, conformément à l'article 54 de la Constitution, de la question de savoir si le protocole nº 6 à la Convention de sauvegarde des droits de l'homme et des libertés fondamentales relatif à l'abolition de la peine de mort, signé par la France le 28 avril 1983, comporte une clause contraire à la Constitution;

Le Conseil constitutionnel,

Vu la Constitution;

Vu l'ordonnance du 7 novembre 1958 portant loi organique sur le Conseil constitutionnel;

Vu la Convention de sauvegarde des droits de l'homme et des libertés fondamentales, signée à Rome le 4 novembre 1950;

Le rapporteur ayant été entendu;

Considérant que le protocole nº 6 additionnel à la Convention européenne de sauvegarde des droits de l'homme et des libertés fondamentales concernant l'abolition de la peine de mort, soumis à l'examen du Conseil constitutionnel, stipule que la peine de mort est abolie, qu'elle peut toutefois être prévue pour des actes commis en temps de guerre ou de danger imminent de guerre; que cet accord peut être dénoncé dans les conditions fixées par l'article 65 de la Convention européenne des droits de l'homme;

Considérant que cet engagement international n'est pas incompatible avec le devoir pour l'Etat d'assurer le respect des institutions de la République, la continuité de la vie de la nation et la garantie des droits et libertés des citoyens;

Considérant, dès lors, que le protocole nº 6 ne porte pas atteinte aux conditions essentielles de l'exercice de la souveraineté nationale et qu'il ne contient aucune clause contraire à la Constitution,

Above is the entire decision—I show this mostly to demonstrate how absurdly laconic the *Conseil* is compared to the U.S. Supreme Court, which would no doubt produce three contradictory opinions each running hundreds of pages.

Anyway, once Protocol 6 was approved, reinstating the death penalty would require leaving that treaty, which would damage France's international credibility. In addition, by the terms of the treaty, it would take a minimum of five years to withdraw before the death penalty would resume.

In 2007, the French constitution was modified by a vote of 828 to 26 to add Article 66-1 the entirety of which is:

> « Nul ne peut être condamné à la peine de mort. »

("No one can be condemned to the pain of death.")

France's final execution was in 1977, the last execution in Western Europe, and the last use of the guillotine by any government in the world.

Summary: During the 1970s, leftist elites settled on abolition despite popular opinion being to the contrary. Leading up to the 1981 elections, the left realized they were in the strongest position in a generation. François Mitterrand clearly signaled that he intended to abolish the death penalty, won the election, and then did what he had promised. To make this move harder to reverse, France signed international treaties that supersede national law. Over the next 20 years, popular support dropped to around even, though it may have slightly increased in recent years.

What to think of this? The overall story is fairly similar to that in Germany and the U.K.: leftist elites unified behind abolition, conservatives elites were divided, and popular sentiment against abolition wasn't particularly important. Still, personally, I can't escape the feeling that France's path to abolition is somehow a bit more honorable—probably that's exactly what Badinter intended when he urged that abolition be a clear position before the 1981 elections.

Next up, we'll look at the history in the U.S. and try to understand why all this *didn't* happen there. Then we'll finally tackle our primary question of how abolition interacts with public opinion.

Data and sources

Again, in all these histories of European countries, I rely heavily on Andrew Hammel's fantastic and underrated book, *Ending the Death Penalty: The European Experience in Global Perspective.*

Anyway, here are the poll numbers in the plot above. Where possible, I converted to a single "support" score by removing respondents who gave "no opinion". However, if only "support" numbers are given, I used those unchanged. In the IPSOS polls below, I can't determine if there was a "no opinion" option. (Let me know if you can figure that out.)

year	support	oppose	source
1908	77	??	IFOP
1960	50	39	IFOP
1967	51	42	IFOP
1972	63	27	IFOP
1979	58	31	TNS
1980	55	37	TNS
1981	58	34	TNS
1982	62	33	TNS
1983	50	38	TNS
1984	56	36	TNS
1985	64	32	TNS
1987	65	29	TNS
1991	61	35	TNS
1993	61	33	TNS
1994	59	36	TNS
1999	46	48	TNS
2001	44	49	TNS
2002	40	54	TNS
2006	42	52	TNS
2007	45	52	IPSOS
2014	45	??	IPSOS

year	support	oppose	source
2015	52	??	IPSOS
2016	48	??	IPSOS
2017	49	??	IPSOS
2018	51	??	IPSOS
2019	44	??	IPSOS
2020	55	??	IPSOS
2021	50	??	IPSOS

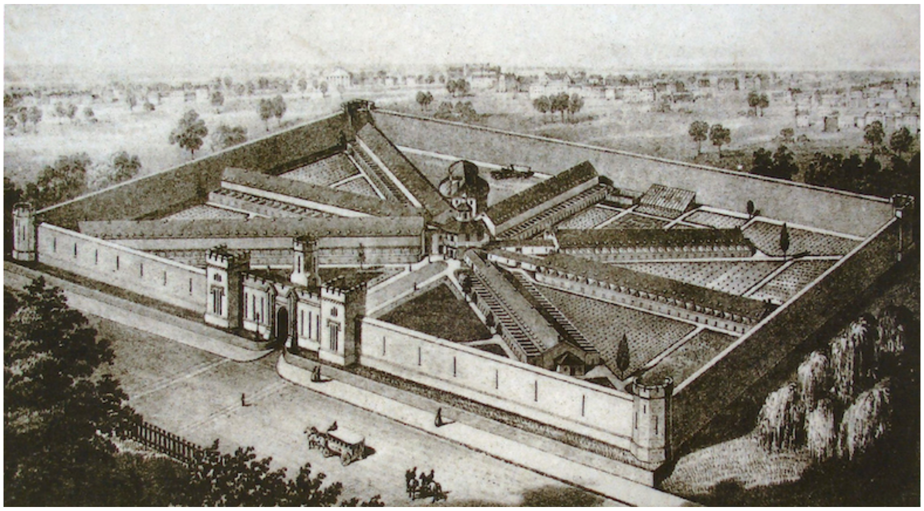
More details:

- The 1908, 1960 and 1972 IFOP polls are quoted on Wikipedia.
- The 1967 IFOP poll is quoted by Erskine (1970).
- The TNS polls are quoted by Hammel.
- The IPSOS polls are here on p. 32.

How the United States didn't ban the death penalty

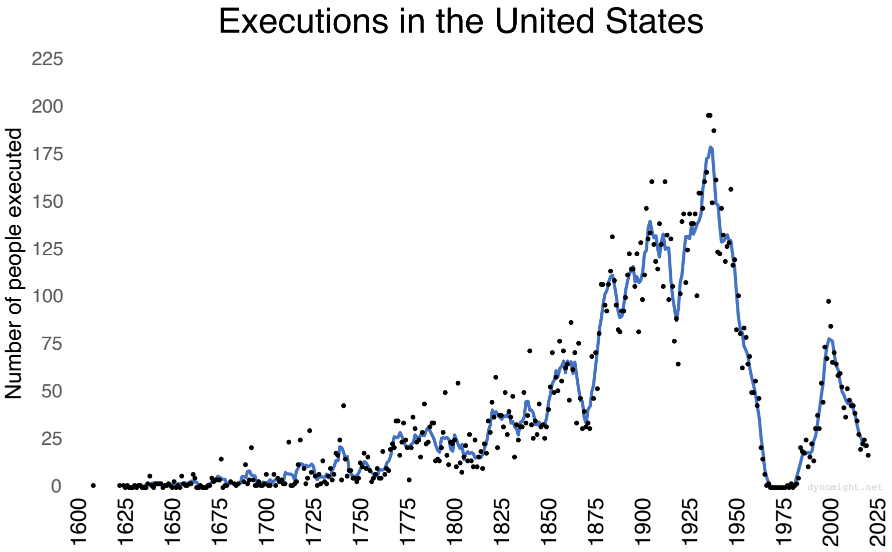
Early America inherited much of Britain's bloodthirsty but arbitrary approach to the death penalty, with theoretical penalties for things like theft and rebellious children that were rarely carried out. However, executions did happen for crimes short of murder: Death was a common penalty for repeated theft, and in 1644, a couple was executed in Massachusetts for adultery.

Following the revolution in 1776, states dramatically decreased the number of crimes punishable by death. (The vast majority of executions in the US have always been carried out by individual states, not the federal government.) There was a general movement to make punishments more proportional, and the number of executions fell somewhat despite rapid population growth.



At the start of the 1800s, some religious leaders and enlightenment thinkers started to push for complete abolition. Benjamin Rush, one of the signers of the declaration of independence, pushed for imprisonment as an alternative, leading Pennsylvania to build the world's first "penitentiary"—so-called because of Quaker beliefs in the value of penitence.

Gradually, a few states started to abolish the death penalty. Michigan never executed anyone and abolished in 1847, followed by Rhode Island (mostly) in 1852 and Wisconsin in 1853.


Executions in the United States

The number of executions rose rapidly between 1800 and 1900. However, there's no cultural shift to explain here, it's entirely due to the rising population—the number of executions per million people was slowly dropping.

Year	Executions	Population (m)	Ratio
1800	20	5.3	3.8
1850	50	23.1	2.2
1900	120	76.2	1.6

Abolition was disrupted by the Civil War and reconstruction, though Maine did ban the death penalty in 1887.

The start of the 1900s was the beginning of the Progressive era. Ten states abolished the death penalty (Minnesota, North Dakota, Colorado, Oregon, Washington, Kansas, South Dakota, Missouri, Arizona, and Tennessee), and the number of executions plateaued.

However, after WWI, support for capital punishment rebounded, and all but two of these states reversed their abolition. The 1930s saw the highest absolute number of executions in US history.

From the 1930s to 1950s, culture gradually came to view hanging as barbaric, in part due to the botched execution of Eva Dugan. Most states replaced hanging poison gas, electrocution, or lethal injection, which were less disturbing to onlookers.

After the end of WWII, the number of executions gradually fell, reaching zero in 1967. At this time, the U.S. was even cited as an example of countries moving away from the death penalty in debates in the UK.

Why did this fall happen? One cause was an increased number of states abolishing the penalty, which Hawaii, Alaska, Delaware, Michigan (again), Oregon, Iowa, New York, West Virginia, Vermont, and New Mexico all did between 1957 and 1969. Another cause was that even in states where the death penalty was legal, it gradually became rare for prosecutors to seek it.

There were no executions in the United States between 1967 and 1977.

Along with this decline in executions, the death penalty was also quickly becoming less popular. In 1966, a Gallup poll found for the first time (and only time, to this day) found that more Americans opposed the death penalty than supported it.

In 1958 in *Trop v. Dulles*, the Supreme Court held that the Eighth Amendment's prohibition of cruel and unusual punishments held for "evolving standards of decency". This led to a suspicion that the death penalty might be unconstitutional.

In 1972 in *Furman v Georgia*, the Supreme Court held that all current death penalty schemes did indeed violate the Eight Amendment. This decision was extremely disjoint, coming 5-4 with no consistent rationale and all nine justices writing separate opinions. Two justices in the majority wrote that the death penalty itself was cruel and unusual, while three others were more narrow, saying that current statutes were too unreliable and arbitrary.

As a glimpse of the boundaries of elite opinion at the time, the four dissenting judges in *Furman*—all appointed by President Nixon—each stated they were personally opposed to capital punishment and would vote for abolition if they were on a state legislature.

Following *Furman*, there was a large jump in public support for the death penalty. I thought that this might be a backlash to Supreme Court overreach, however, ~~a small and highly biased sample of older Americans I happen to know~~ my sources tell me that nobody cared about that and it was more about increases in crime.

Furman created a de-facto moratorium. Many suspected that the death penalty was gone for good, but it wasn't entirely clear if reformed systems could pass constitutional muster. Between 1972 and 1976, 37 states created new death penalty laws. Reforms included separate guilt and sentencing phases, and more uniform standards to guide juries and judges. The Supreme Court invalidated laws in North Carolina and Louisiana that imposed the death penalty for all capital crimes.

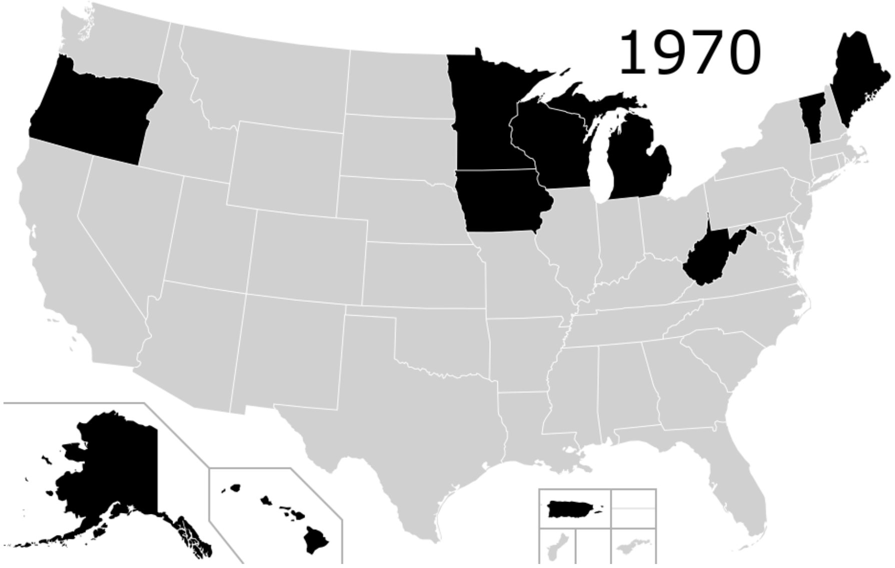
In 1976 in *Gregg v. Georgia*, the court confirmed 7-2 that some of these reformed laws *were* constitutional. (The two dissenters are the same justices who wrote that all executions were unconstitutional in *Furman*.)

Up until these court decisions, America's history isn't all that different to that of the U.K. or France. In fact, support for the death penalty at this time was substantially lower in the U.S. than in Germany, the U.K., or France.

Executions resumed in 1977 and increased up until around 2000. Popular support for the death penalty also dramatically increased, peaking at around 80% in 1996.

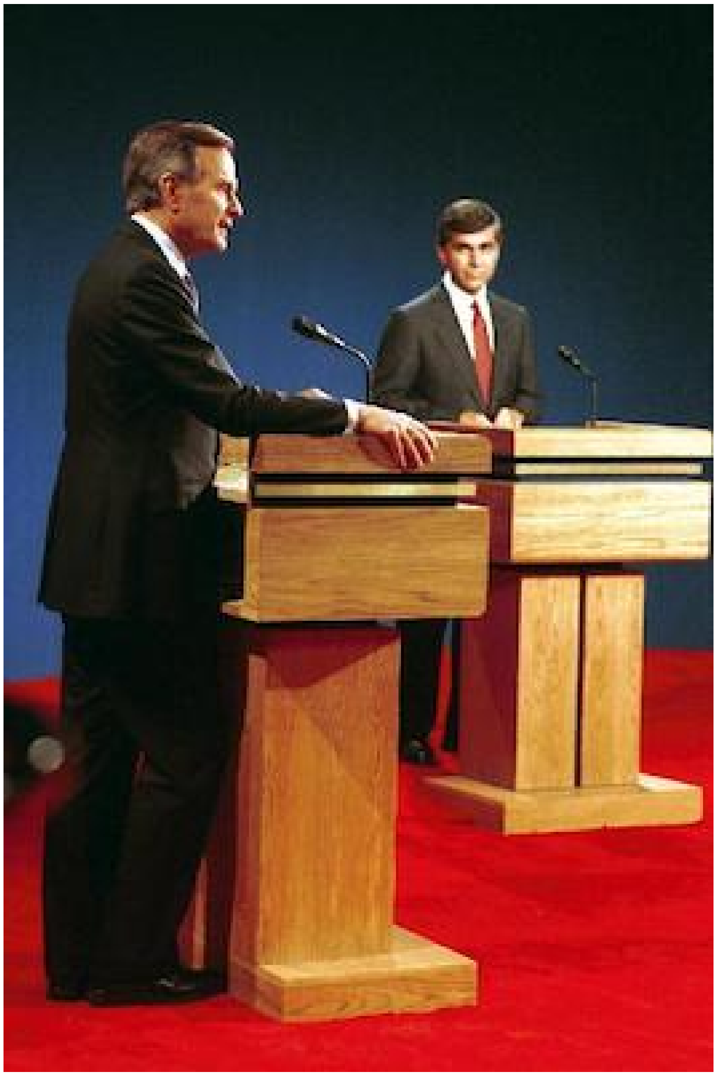
The surge in popular support for the coincided with a huge increase in violent crime. The homicide rate—already much higher than the European countries we looked at—more than doubled between 1963 and 1975.

Here's an animation that shows what states formally banned the death penalty (black) from 1970 to 2021 (if you can avoid screaming in frustration at how slowly it moves):



Since 1972 there have been at least 20 ballot initiatives in different states to either abolish the death penalty, reinstate it, or change the scope of applicability.

In every case, the vote went in the pro-death penalty direction, e.g. against abolition, in favor of bringing the penalty back, in favor of applying the death penalty to more crimes, in favor of reducing the governor's ability to commute sentences, or in favor of shortening the appeals process. One of the closer votes was the 2016 initiative to ban the death penalty in California, which failed 53% to 47%. (Governor Newsom imposed a moratorium in 2019.)



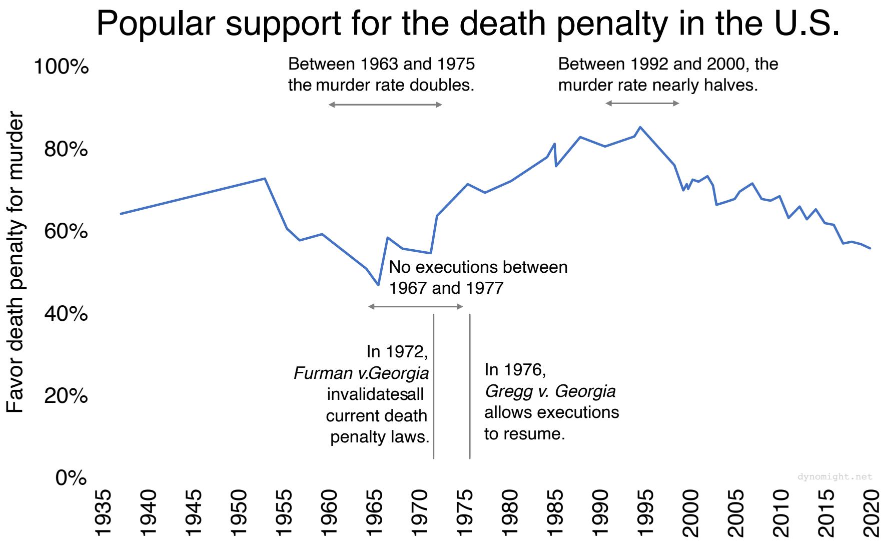
The politics of the death penalty were laid bare in 1988. Michael Dukakis, the Democratic governor of Massachusetts was in a presidential debate when he stated—consistently with his abolitionist position—that he would not support the death penalty for someone what raped and murdered his wife. His poll numbers dropped by 7 points overnight and he went on the lose the election by 8 points.

Dukakis' example was surely on Bill Clinton's mind in the 1992 presidential election campaign, when he specifically flew home to Arkansas to confirm an execution. The man was Ricky Ray Rector, who had essentially lobotomized himself in an attempted suicide after murdering a man and shooting a policeman. Rector had so little understanding of what was happening that he saved the dessert from his last meal “for later”.

Over the years, the Supreme Court has continued to narrow the scope of applicability of the death penalty.

- *Coker v. Georgia* (1977) prohibited the death penalty for the rape (without murder) of an adult woman.
- *Godfrey v. Georgia* (1980) held that a murder being “outrageously or wantonly vile” was not sufficient cause for the death penalty.
- *Enmund v. Florida* (1982) held that the death penalty cannot be applied to someone who didn’t kill anyone, try to kill anyone, or intend to kill anyone, even if they were part of a crime that resulted in someone dying. This was later clarified in Tison v. Arizona (1987) to include behavior that was reckless and indifferent to human life.
- *Ford v. Wainright* (1986) held that it was unconstitutional to execute the insane.
- *Atkins v. Virginia* (2002) prohibited executing someone who is intellectually disabled.
- *Roper v. Simmons* (2005) prohibited the death penalty for any crimes committed by someone under the age of 18.
- *Kennedy v. Louisiana* (2008) prohibited the death penalty for the rape of a child.

Eventually, in the 1990s, things started to move against the death penalty.



In the early 1990s, the homicide rate finally began to fall, dropping by about half by 2000 and then stabilizing at a similar level to the 1950s and 1960s. (But keep in mind that level is still *much* higher than other Western democracies.)

Starting in 1995, support for the death penalty decreased. It fell from a peak of 85% to around 56% in 2020, only slightly higher then France or the U.K..

Starting in 2000, the number of executions fell, decreasing from a peak of almost 100 per year to around 20 per year now. There seem to be two major reasons for this fall:

1. Juries are more hesitant to issue death penalty convictions. This is partially spurred by the dozens of people who were sentenced to death and then later exonerated, something that become much more common after the arrival of DNA evidence. There have been a few possible wrongful executions. Juries also increasingly choose life without the possibility of parole, an alternative sentence now available in more states.
2. Many states have added extra layers of protection for defendants charged with the death penalty, such as providing better defense lawyers, and more appeal opportunities. These reforms seem often motivated by a concern that the Supreme Court's gradually increasing standards might otherwise invalidate all convictions. These protections have made it more difficult and expensive to get the death penalty, meaning fewer prosecutors seek it.

As of 2021, the situation is:

- In 27 states, the death penalty is formally abolished or was never created in the first place.
- In 7 states and the federal government, the death penalty is legal but there is a formal moratorium.
- In 10 states and the military, the death penalty is legal, and there is no moratorium, but no executions have happened in the last 10 years.
- In 13 states, the death penalty is actively used. (Arizona, Idaho, Texas, Oklahoma, Missouri, Arkansas, Kentucky, Mississippi, Alabama, Georgia, and Florida)
- Ohio is complicated.

While most executions are done by the states, the death penalty remains legal at the federal level. Nevertheless, between 1972 and 2000, there were no federal executions. Between 2001 and 2003, there were three, including Timothy McVeigh who killed 168 people in the Oklahoma City bombing. Between 2004 and 2019, there were again no executions.

Still, all this time, US federal courts continued to *sentence* people to death. President Trump resumed executions in July 2020, and 13 were carried out before President Biden imposed a moratorium in early 2021.

Summary: In the 1960s, public support for the death penalty was unusually low compared to the other countries we looked at. Executions gradually slowed and stopped in 1967. The Supreme Court flirted with declaring the death penalty unconstitutional in 1972 but confirmed it was legal in 1976. Following this, violent crime, executions, and public support for the death penalty all surged. Starting in the 1990s, violent crime receded, reforms made the death penalty more difficult to carry out, and the number of executions waned, along with popular opinion. In 2021, a small majority of the American public supports the death penalty.

Compared to Germany, France, or the U.K., the history of abolition in the U.S. is bewilderingly convoluted. It's odd that the penalty would slowly become less popular, stop being used for 10 years, and then suddenly rebound for 25 years. Also, it's impossible to keep track of the number of states that banned or restricted the death penalty, but then it would get un-banned or un-restricted, either because different politicians won the next election, or because of a popular referendum.

Next time, at last, we'll try to find patterns in all these histories we've gone through.

Raw data and sources

Underrated reasons to be thankful

1. That our atmosphere has low enough pressure and levels of deuterium that nuclear fission in air doesn't cause hydrogen atoms to fuse into helium, meaning that the first nuclear bomb test in 1945 didn't in fact ignite the atmosphere and engulf the planet in flames, which was still a bit of an open question when it happened.
2. That the Earth hasn't recently been hit by a solar flare as powerful as the 1859 Carrington event, which is good because that would set electrical lines around the world on fire, meaning months of power outages and simultaneous failures of food, transport, and medical systems.
3. That human social behavior was shaped in an environment of small bands with repeated interactions, endowing us with instincts to be fair and punish defections that surely aren't game-theoretically optimal now in anonymous late modernity but have positive externalities in making large-scale cooperation possible and without which society as we know it probably couldn't exist.
4. That humans didn't evolve under strong alpha selection, meaning that 99% of us don't get eaten by birds in our first moments of life as we crawl down the beach towards the ocean, which is nice.
5. That the FDA, Health Canada, and the European Food Safety Authority all agree that at the doses humans consume, aspartame is perfectly safe—*not* genotoxic, *not* carcinogenic, does *not* cause an insulin spike—or at least has small, unknown harms, meaning that people with a sweet tooth can

avoid the large, known harms of sugar with minimal exertion of willpower, and this is still true even though people for some reason seem to reject and despise this extremely lucky fact.

6. That resistance to antibiotics appears to come at at least *some* cost to bacteria, meaning that there's nonzero hope that if we invent enough antibiotics and cycle them or use some other tricks we can avoid a future where we return to the grim reality of the past where any small accident could be deadly.

7. That English and Scottish people in the 1700s got really obsessed tinkering with stuff which led to the industrial revolution, if that really is why it happened in Britain and not China or Austria or wherever, and assuming it wasn't inevitably going to happen somewhere. (Speculative)

8. That an asteroid killed the dinosaurs after they had roamed the Earth for 170 million years, knocking evolution into a different basin of attraction that led within 60 million years to humans and higher intelligence and the industrial revolution and a world where not every animal spends every day in a constant state of war trying not to be eaten, which possibly the dinosaurs would never have done? (*Very* speculative)

9. That the Earth happens to have a liquid outer core with electrical currents that produce a magnetic field which protects us from high energy particles which would otherwise at least burden us with more cancer and higher genetic load and possibly even make life impossible.

10. That the hard problem of consciousness exists, i.e. that for whatever reason we have phenomenal experiences rather than being "zombies" which, while it's dispiriting that this seems inexplicable in terms of any current or possible future physics, it's cool that the lights are on in the universe.

    Oh, and also that the universe exists at all, and that life exists on at least one planet in it.

11. That it happens to be a game-theoretic equilibrium to have a near-equal sex ratio, although honestly, I have no idea what things would be like if that wasn't the case and maybe it would be fine.

12. That private life exists and that markets, while we all (some?) appreciate their power to allocate resources, don't permeate every single part of life, that there are moments of beauty and grace that aren't best understood as competition and harnessed selfishness.

13. That selfhood is possibly an illusion, and it's all atoms bouncing around in the void, and there are no real boundaries between different creatures, and the idea that you are the same person you were yesterday is an illusion your brain-meat gives to you, which OK, isn't obviously a positive thing but does give a sense of peace for those of who happen to have maxed-out existential angst stats. (Speculative)

14. That due to some combination of nitrogen fertilizer, pensions, women's rights, education, birth control, etc., an overpopulation calamity hasn't yet happened and we might coincidentally stabilize at a level that's somewhat close to what maximizes average utility, and without (mostly) needing to use gruesome methods of coercion.

15. That large animals like humans seem to be able to develop sophisticated defenses against parasites that parasites can't counter-adapt against, meaning that parasitism is way less of an issue for us than it is for smaller animals, which is good because parasites are bad.

16. That even though we evolved as ruthless replication machines, we've somehow risen out of the muck and we currently find ourselves running cultural software that's way out of sync with what game theory would dictate, and perhaps we can seize the moment and build a civilization that can tame the brutal dynamics that created us.

17. That hokey unfashionable techniques like practicing gratitude turn out to have strong scientific evidence behind them, and several countries happen to have a preexisting holiday that's already, at least in theory, dedicated to this practice.

18. That even though the turn humans made from hunter-gatherer bands into agriculture pretty clearly made life worse, it eventually led to the industrial revolution and modern society which is way better than hunter-gatherer life, and people who doubt that should consider the percentage of us who used to die violent deaths.

19. That even if, as most scientific-minded people seem to assume, there is no afterlife, that's not ideal, but is much better than other possibilities like, say, being tortured for eternity.

20. That Ramanujan existed, which means that humans not *that* different from you or me can have capabilities that seem impossible, which is a hint that—possibly with a little help—human potential might be much higher than it now seems.

21. That humans happen to not be obligate carnivores meaning it's possible to contemplate more (possibly, debatably, some people think, let's not fight we're friends) ethical diets without facing the kind of dilemma that vampires do.

22. That it's even possible to develop technologies to produce huge amounts of energy without emitting carbon and we happen to have developed them at a time where we at least *could* avoid the most extreme climate change scenarios with minimal impact on our lifestyles.

23. That some unknown miracle blend of circumstances happened to arrive in Athens in 500 BC leading a tiny city of 250k people to produce Socrates,

Plato, Aristotle, Archimedes, Euclid, Hippocrates, Pythagoras, Thucydides, Herodotus, Aesop, Solon, Pericles, Aristophanes, and Sophocles, and that it might be possible to intentionally recreate such conditions around the world today and spur incredible human flourishing, and why aren't we working on this? (Partially wrong, Archimedes, Euclid, Pythagoras, and Herodotus weren't from Athens but spent time there, Hippocrates had no connection to Athens, and Aesop maybe didn't exist.)

24. That the Quakers settled in modern Pennsylvania, creating an example of a peaceful, tolerant, enlightened society that avoided war with native Americans—at least for a while—and piloted ideas like abolishing slavery, trial by jury, public education, and equal rights for women, a cheerful contrast to what was happening elsewhere at the time.

25. That evolution happened to settle on this trick of "love" to serve the interests of reproduction rather than, like, causing us to feel like we're being burned alive every time we don't find mates or feed our kids or whatever, which there's no obvious reason it should have done, and also I acknowledge the real pain some lonely people do feel.

26. That even if humans can't travel to other stars or galaxies with our fragile organic bodies, it's probably possible for us to create artificial intelligences that can, and while it's not great that they might kill us, it's surely better than the light of consciousness vanishing entirely when the sun eats the Earth in 7.5 billion years, no?

27. That the homeostatic theory of drug tolerance isn't a 100% perfect law, which means that we aren't *entirely* stuck with our bodies operating according to their own whims all the time.

28. That we happen to live in a world where ideas are both non-rivalrous and hugely valuable allowing us to create things like small molecule drugs and no-knead bread recipes and semiconductor manufacturing techniques which are a gift to our descendants at least unless/until civilization collapses, and which makes the average utility of human life a concave function of world population rather than some monotonic decrease, giving us another tool to fight evolution's greedy hand trying to drag us back into the mud.

29. That the abstraction of "narratives" exists, allowing us to understand the world at least partially through the crazy messy process you're undertaking right now rather than everything being a blind inscrutable idiotic evolution of the wavefunction, and also the world happens to be structured so that these "narratives" are powerful enough to actually partially explain at least some phenomena sometimes.

30. That other animals have more cone cells than humans, e.g. birds with four and shrimp with *up to 16*, and so probably see colors we can't even conceive of which, yeah, that limitation of our minds is frustrating, but

it also hints that there are huge unseen dark continents of qualia lurking out there which someday we might find a way to visit.

Effective selfishness

Here are some things I'd like to know about how to live my life:

1. If I eat Brussels sprouts for dinner tonight instead of pizza, how much longer do I live (in expectation, in minutes)?
2. What should I eat to avoid getting tired after lunch?
3. If I have a glass of wine with dinner every day, how much longer/shorter do I live? What if I drink seven glasses on Saturday instead?
4. What's the impact of a 30-minute run tonight, in terms of longevity and my energy over the next month?
5. If I drink coffee every day, do I reach homeostasis and the same steady-state as if I drank none?
6. Should I even bother with this health stuff? Is it dumb to spend younger-me time now to get older-me time later?
7. How much do I reduce my life expectancy by driving 20% faster?
8. Is the link between relationships and happiness causal? What's the best way for me to build more relationships?

   (Is it spending all my free time writing a pseudonymous blog?)
9. Or should I focus on making my existing relationships better?
10. Say my personality is extraverted, agreeable, and open to experience. Should I seek a romantic partner with the same traits or complementary ones? Is there any predictive power at all once you condition on how well we get along?
11. Will I be happier if I take a media diet?
12. Is the link between religion and happiness causal? Will it still work if I'm a nonbeliever?
13. What isn't on this list but should be?

You get the point: We have lots of choices in modern life. They collectively have a big impact on how happy/healthy/productive we are, but they are *hard* and evolution hasn't given us good instincts for making them.

It seems likely that there's low-hanging fruit, but what is it?

The problems that effective altruism solves

Take Alex, who wants to make a donation and make the world better. She faces four questions:

1. Which organizations are competent, and which are more the collect $500 million in donations, build six homes, refuse to elaborate type?
2. What will happen to her money? Will a huge fraction of it go as a commission for the person who convinced her to donate?
3. What's a good strategy? If Alex is worried about breast cancer, should she help with basic research, or early diagnostics, or prevention, or what?
4. Which problem should she focus on? How much will an extra dollar move the needle for breast cancer vs. early warning of asteroid collisions?

These questions are also hard, basically too hard for a single person to answer. So, when you look at which philanthropic organizations thrive, these questions traditionally haven't mattered very much. The organizations that endure tend to be the ones that make their supporters feel good.

This isn't anyone's fault! It's not Alex's fault for not doing this basically-impossible research, and it's not an organization's fault for being good at the one thing they must be good at to not be outcompeted by an organization that *is* good at it. It's just another case where nature imposes a fitness function that isn't aligned with our interests.

How effective altruism solves them

Broadly speaking, effective altruism tries to tie impact with how good things feel. The basic strategy is to (1) get people excited about impact in general, (2) figure out what has impact, and (3) promote high-impact stuff to impact enthusiasts.

Research shows that different strategies can be *much* more cost-effective than others for the same problem. Jamison et al. (2006) estimated that how much it could cost to save a disability-adjusted life year (DALY) for someone at risk of HIV/AIDS. Surgical treatment of Kaposi's sarcoma cost $30k, antiretroviral therapy cost $1k, condom distribution cost $100, and peer education cost only $20.

Calculations like these allow you to compare not just different strategies, but different *problems*. If you care equally about polar bears and girls' education in Eastern Europe but one dollar can do much more for the girls, that's probably where you'll want to concentrate.

The effective altruism community also considers personal actions. For example, if you care about orangutans, you could fly to Sumatra and go to the rainforest and then, like… help? But it's probably more effective to get a job and donate some money to an orangutan sanctuary. It also suggests that some things might be counterintuitively *in*effective—e.g. if you can land a competitive job at a famous nonprofit, maybe you're just displacing someone else who would have done a great job anyway.

Most importantly (I claim) effective altruism tries to make all this analysis legible across different domains. This is critical if you want to compare different problems because no one can simultaneously be an expert on the existential risks of nuclear weapons *and* pandemics, *and* micronutrients, *and* AI safety.

This legibility also helps spread insights between different problems. For example, some have criticized effective altruism for neglecting "moonshot" economic or political goals and for having a bias towards things that can be measured. The community seems to be chewing on these concerns in a productive way, which is probably more efficient than having the same discussion separately in each sub-community.

Self-help has a lot of problems, too

At first glance, self-help seems very different from altruism. Most obviously, people care much more about helping themselves. (Go to a bookstore and look at the number of books on dieting vs. extreme poverty.) And in theory, there's an even deeper difference—for self-help, you have *feedback*. With charity, it's easy to send money off into the void and forget about it. But if you waste a lot of money failing to fix your back, you'll notice your back still hurts.

That's the theory. In reality, can we just agree that self-help advice on average is not great? There's a bunch of reasons for this.

For one thing, feedback loops are too long. If I buy a diet book and I don't lose weight, well, I already paid for it.

For another, feedback loops are usually confounded. Say I want to be more popular, so I follow the advice of a TED talk on how to communicate, but three months later everyone still hates me. Was the advice bad, or is it that I was grumpy because I have new loud neighbors and I couldn't sleep?

A third problem is that science is hard, statistics is hard, and people get them wrong. For example, diet advice is often obsessed with p-values, but ignores effect sizes. If eating cranberries has a 20% chance of giving me an extra 10 years, I'm going to eat cranberries. If it's certain to give me an extra week, maybe not. You can't make these judgments without effect sizes.

Fourth, good advice is often siloed and illegible. If you want to know how to best exercise/make friends/invest/sleep/etc., then for each goal you face anew the problem of figuring out who you can trust and what their jargon means.

Fifth, there's the "bare minimum quality" problem. Say I want to sleep better. If I go to a forum on beds they'll tell me that if I'm low on cash, I can get an OK mattress for "only" $2k or so, but honestly, I should take a look at my life and reconsider what I'm worth as a human being. Or say I want to be stronger. If I go to r/fitness, I'll learn that I should get a power-cage and do 15 sets each week of barbell back squats, deadlifts, shoulder press, and bench press, along with assistance work, and if I'm not up for that, I *deserve* to be weak.

What's happening here is that the people who know a lot about a given topic are *really into* that topic. Their advice might be good if you had their values, but you don't, and anyway it would be *impossible* to give that level of attention to every part of your life.

Sixth, and most importantly, what should you prioritize? We all face a *very large* action space. A diet book isn't going to tell you that, "actually your diet is fine, but you seem kinda lonely—you should get a better haircut and take a dance class."

Some examples

Here are some examples of what I *think* are effective advice for different domains. I'll start with standard normie advice:

Investing. Pick a low-cost lifecycle fund, put the same amount of money into it every month.

Diet. Eat food, mostly plants, not too much.

Sleep. Use sleep hygiene and maybe low-dose (300 g) melatonin.

Skincare. The most important thing is sunscreen. Also, use moisturizer if you like, and *possibly* a retinol cream.

Smoking. Nah, mate.

Smoke alarms. Use them and keep fresh batteries in them. (Almost 3000 people die in home fires in the US every year, and having smoke alarms reduces risk by more than half. This is a modest effect, but it's a huge win since smoke detectors are so cheap.)

Depression. You might feel better just through mean reversion. In terms of treatments, weirdly, different types of therapy and different anti-depressants are all about equally effective on average. (Which is: moderately) But different things work for different people.

Cancer. We live in the future! If you get breast cancer, colorectal cancer, testicular cancer, prostate cancer, thyroid cancer, melanoma, cervical cancer, or Hodgkin's lymphoma, then modern treatments give you a >90% chance of making a full recovery, *provided it is caught early.* You have a reasonable chance of getting one of these at some point in your life (maybe 20%?), so catching these early is plausibly worth a DALY or so in expectation.

Drinking. Any drinking increases cancer, but 1-2 drinks per day might help with heart disease and diabetes and *might* be enough to make moderate drinking neutral or even a net win, but no one knows. Going beyond 1-2 drinks per day is clearly harmful, especially if you have genes for the Asian flush. It also screws up your sleep.

Mortality. Tell your loved ones how you want the end of your life to go. No wrong answers, but consider that 89% of doctors refuse high-intensity interventions.

And here are some examples of more unusual life advice that I first saw in the rationalist or effective altruism diaspora.

Air quality. Stop using aerosols and candles and ultrasonic humidifiers, be careful about smoke while cooking, install an air purifier, and you can plausibly buy yourself ½ to 2 DALY, depending on where you live.

Seasonal depression. Your phone can measure lux. On a sunny day in summer, you'd get 30k-100k lux. Keep adding lights inside until you get that many. (Maybe this works as well as a lightbox?)

Aging. If you are >75 years old you should try to get a bit fat, so that if you're injured you don't get trapped in a cycle of not being able to swallow and weakness from lack of food.

Parenting. There's a whole lot of things like toxins, fluoride, pesticides, and flame retardants that probably don't much harm adults, but could have a serious impact on the neurological development of kids.

These all seem reasonable, but they aren't sorted, some of them are probably wrong, and lots of things are surely missing. We can do better.

A flourishing leaderboard

So here's the dream: Let's invent "flourishing adjusted life minutes" (FALM) a sort of better quality-adjusted life year. Then, you should be able to go to a website, answer some questions, then get a leaderboard that might look like this:

Intervention	money cost	time cost	benefit (FALM)	FALM / yearly cost
Call mom	$0	20 min	30 min	3.0
Invite Alex to lunch	$30	1 hr	150 min	2.5
Get smoke detectors	$15	30 min	60 min	2.0
Eat broccoli	$5	10 min	10 min	0.5

(All the numbers are totally made up, and assume you value your time at $0.50 per minute.)

There are lots of challenges with this. For altruism, the world is the world, but this needs to be individualized. (I feel that effective altruism should also make more effort to address differing values, but never mind.) Another challenge is that 10 minutes spent cooking broccoli isn't the same as 10 minutes running—you need to adjust for the cost of the experience itself. These are major challenges, but I see no reason they can't be solved.

Why effective altruists should do this

Now you might agree that this seems like a good idea, but why should *effective altruists* do it, rather than someone else?

(To be clear, many are *already* doing this!)

Here's a few examples that come to mind:

First, *someone* needs to do it, and effective altruists are well-positioned to do it and build their community. Pretty much everyone wants to make their life better. If you attract a lot of people and show them the power of ~~the dark side~~ utility functions and cost-benefit analysis, then they are one step away from joining effective altruism itself. And the effective altruism community has a comparative advantage—they've already built a culture and set of tools that are applicable.

Second, most care in the world is self-care. I'm not making some Howard Roark-type argument that altruism is bad, I'm just observing that if you collect all the selfish actions that an average person takes, they'll weigh 100× as much as their altruistic actions. If you make people even slightly better at helping themselves, that's a huge impact.

Third, the act of *facilitating* effective selfishness is itself effective *altruism.* I'm going to use an example of my own here (even though it's indecorous) because it's the only thing I have numbers for. As I write this, something like 40,000 people have viewed my essay on air quality, which took me around 100 hours to write. Let's make some assumptions:

- 10% of visitors read a serious chunk of the essay.
- 5% of *those* people took enough actions to cut their exposure to particulates in half.
- All *those* people live somewhere with relatively clean air, meaning that fixing it saves only around ½ a disability-adjusted life year (DALY) in expectation.
- I value my time at $100 per hour.

If we add all this up, the essay should save around 40,000 × 0.1 × 0.05 × .5 = 100 DALY at an effective cost to me of 100 × $100 = $10,000. That is $100 per DALY saved, similar to Givewell's most effective charities.

If you don't write so woefully slowly and you have more influence, I'm sure much higher efficiencies are possible.

Best of all, with this change, the "Effective Altruism" community could rebrand itself as simply "Effective", a confusing but incontestably cool-sounding mononym.

How many extra days of life do you get from taking statins?

Advice is everywhere, but it's rare to know what benefit you're supposed to get from following it.

Say you're a smoker. If I tell you not to smoke, you'll probably ignore me because you think (correctly) that I care less than you do about the cravings you'd suffer if you quit. But if someone reminds you that each cigarette takes 10 minutes off your life, you'll probably consider that.

Or say you hate vegetables. You know they're good for you, but *how good*, enough to outweigh the horror of eating okra?

The solution is "effective selfishness"—advice that comes with effect sizes.

Effect sizes are hard

To know what running does, you should take thousands of people, tell half to run, wait years, make sure people follow their instructions, and then see how healthy everyone is at the end. That's not easy, and even if you did that, you still have to worry: Did the running *itself* make people healthier? Or did people get more sunlight while outside or make friends at the park or get a placebo effect from the *belief* that running is healthy?

The easiest place to find effect sizes should be small-molecule drugs: There's no interaction with lifestyle, you can control for placebo effects with placebo pills, and these experiments are what the medical system, in its grace, likes to spend money on.

And in terms of small-molecule drugs, what should be easiest? I chose statins, the drugs that 200 million people around the world take to control cholesterol. These are old, cheap, commonly used, and thought to have huge effects, so we have lots of big studies.

Aside: Statins have an interesting mechanism of action: HMG-CoA reductase is an enzyme that takes HMG-CoA and creates mevalonic acid, which is needed to make cholesterol. Statins are molecules that are structurally similar to HMG-CoA, meaning they sort of get stuck on the HMG-CoA reductase enzyme where HMG-CoA is supposed to go, meaning the enzyme can't make mevalonic acid, which bottlenecks the whole cholesterol pipeline.

Anyway, there are many meta-analyses for statins out there, but they typically work on risk ratios, i.e. the relative chance that someone will suffer a heart attack or die during a given period. These are great, but they aren't particularly *actionable*: If I don't have a great sense for what my odds of dying are (which I don't) then it's hard to say how important it is if those odds get multiplied by a given constant.

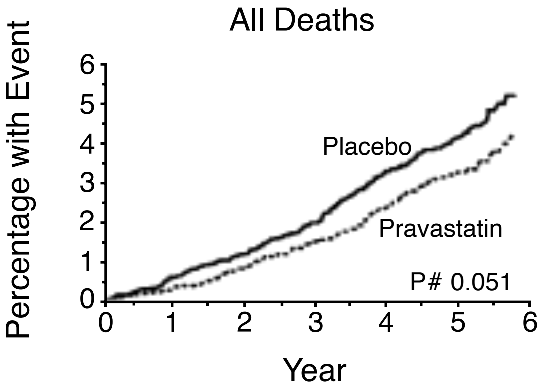
So, being a fan of simple-minded "number of days of life" effect sizes, I was excited to find Kristensen et al.'s paper, *The effect of statins on average survival in randomised trials, an analysis of end point postponement.* They try to do exactly what we want:

> To the best of our knowledge, statins have not been systematically assessed in an outcome postponement model. We identified statin trial reports that provided all-cause survival curves for treated and untreated, and calculated the average postponement of death as represented by the area between the survival curves.

Studies

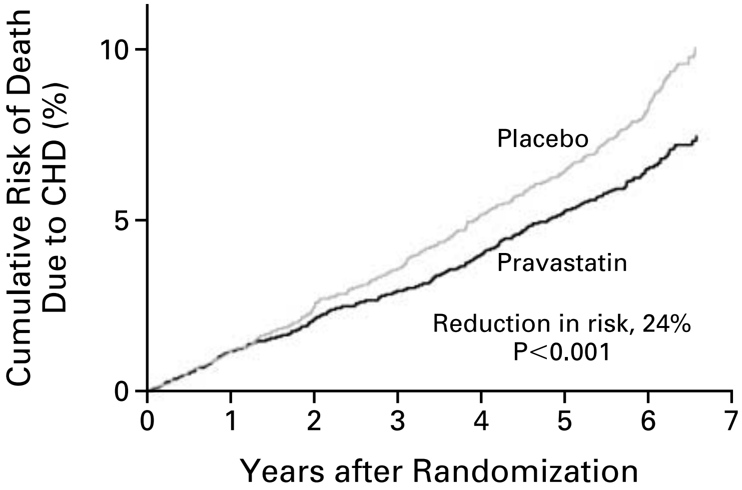
Kristensen et al. do a meta-analysis of 11 studies. Let's take a quick look at a couple of them to get a feel for the data. (We'll need the intuition later.)

1. The WOSCOPS study (1995) took 6595 Scottish men aged 45 to 64 who had high cholesterol (total cholesterol at least 252 mg/dL) and no history of heart attacks. They were given either 40 mg/d of pravastatin or placebo.

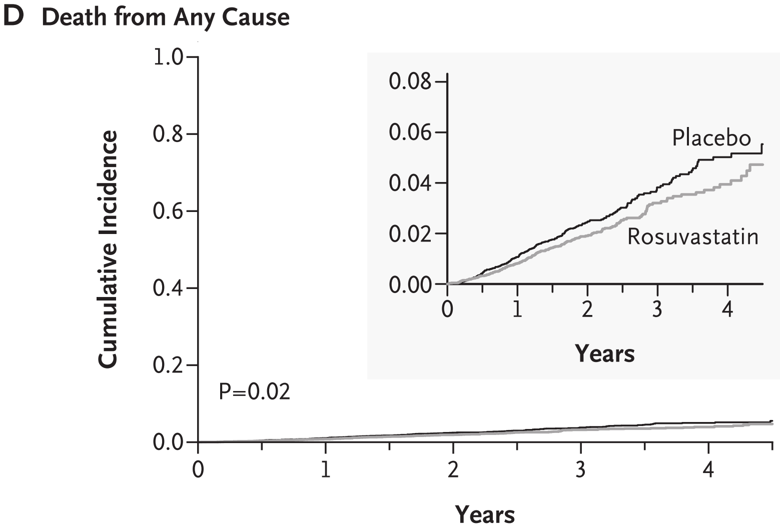
   At the end of the study, 4.1% of people in the placebo group had died of any cause while only 3.2% of the pravastatin group had died. The difference has a vexing p-value of 0.051.

2. The LIPID study (1998) took 9014 patients aged 31 to 87 in Australia and New Zealand who had a history of either a heart attack or chest pain, as well as moderately high cholesterol (total cholesterol between 155 and 171 mg/dL). They were given either 40 mg/d of pravastatin or placebo.



At the end of the study, 14.1% of the control group had died as compared to 11.0% of the treatment group. The difference is very significant.

3. The JUPITER study (2008) took 17,802 people from around the world who had moderate cholesterol (LDL-C 130 mg/dL) and elevated C-reactive protein levels (2.0 mg/L). Men needed to be at least 50, while women needed to be at least 60. They were were given either 20 mg/d rosuvastatin or placebo.



At the end of the study, 1.19% of subjects in the placebo group had died (of any cause), while 0.96% of the rosuvastatin group had died.

There are a couple of things about the JUPITER study that are slightly suspicious.

For one thing, it was financed by AstraZeneca who sell rosuvastatin under the brand name of Crestor. Actually, it wasn't just financed, AstraZeneca also collected the data and monitored the sites, but was firewalled from influencing the analysis or manuscript.

For another, these mortality rates are weirdly low. For example, mortality in the US for people aged 55-64 is around 0.88% per year, and for people aged 65-74 is it 1.78%. Yet the overall mortality in the placebo group was only 1.19% even after multiple years? It's possibly explained by the fact that they excluded people with high cholesterol but it's still weird and apparently, I'm not the first person to notice that.)

Confusion

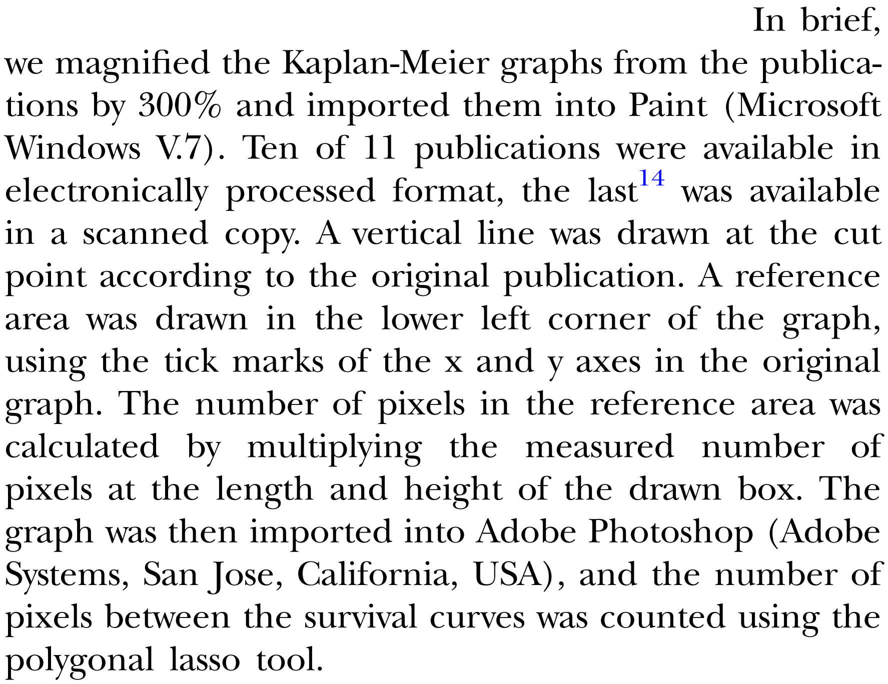
Looking at these studies, these results seem amazing—a simple pill can reduce the entire risk of dying, from *any* cause by around 25%! It's hard to imagine asking for more.

And yet. Let's get back to Kristensen et al.'s meta-review. Remember, they wanted to calculate how many days of life are saved by taking statins. Here's what they calculate for the above studies:

study	Days of life
WOSCOPS	9.33
LIPID	22.05
JUPITER	7.26

There's a benefit of… only a few weeks? How can a magic pill that erases 25% of all deaths translate into such small numbers?

First off, let's look at their calculations. Here's part of their analysis section, which I include just because it's endearing to see "*we used MS Paint*" translated into High Academic.



> In brief, we magnified the Kaplan-Meier graphs from the publications by 300% and imported them into Paint (Microsoft Windows V.7). Ten of 11 publications were available in electronically processed format, the last[14] was available in a scanned copy. A vertical line was drawn at the cut point according to the original publication. A reference area was drawn in the lower left corner of the graph, using the tick marks of the x and y axes in the original graph. The number of pixels in the reference area was calculated by multiplying the measured number of pixels at the length and height of the drawn box. The graph was then imported into Adobe Photoshop (Adobe Systems, San Jose, California, USA), and the number of pixels between the survival curves was counted using the polygonal lasso tool.

Basically, if you take the survival curves and measure the areas under them, you get the average number of days people in the treatment and control groups were alive during the experiment. If you take the difference, you extra days of life from statins.

So how to reconcile these small numbers with the large-seeming effects from before?

The main issue is that their analysis only looks at extra days of life *during the study itself*. Take the WOSCOPS study. When it was over, 4.1% of the control group were dead, as compared to 3.2% of the treatment group. Let's divide people into three groups:

1. People who would survive regardless of statins (95.9%)
2. People who would die regardless of statins (3.2%)
3. People who would survive with statins but die without (0.9%)

The analysis above doesn't account for any extra life that the 0.9% of people in the last group rack up after the study finished. It's as if they all died the day the study was over.

Let's do a very rough estimate. The mean age of subjects at the start of the study was 55 years, it lasted five years, and the life expectancy for a 60-year old in Scotland in 1994-1996 was 17.2 years. If we assume that the people in the last group have standard life expectancies, then this means that there is an average of

17.2 × 365 × 0.009 = 56 days

that's not being accounted for.

(If you're a stats nerd worried about the variance in the age distribution, stay calm: You can check that the life expectancy tables are reasonably symmetric for ages near 60, so if the age distribution is also symmetric, it all cancels out.)

Now, assuming that group will live for 17.2 years on average might be too much—these are people who survived with statins but would have died without, so they are probably in somewhat below-average health.

But there's another factor: The future outcomes of the 95.9% of people who would have survived the study regardless of statins. We don't have data, but I feel comfortable saying that the people in this group have better prospects if they've been taking statins. I mean, statins worked so far, won't they keep working? Or just look at the figure from above again—the curves look like they are still bending away from each other at the end, i.e. differences seem to still be accelerating.

So we have one way in which our calculation is too aggressive and one way it's too conservative. With no justification at all, I'll assume these cancel out, so the adjustment is still 56 days, but with lots of model uncertainty.

We can repeat this analysis with each of the other studies.

WOSCOPS

- Median age at start: 55 years
- Marginal population (survive with statins but not without): 0.9%
- Length of study: 5 years
- Life expectancy for a 60 year old Scott in 1995: 16.2 years
- Adjustment: **17.2 × 365 × 0.009 = 56 days.**

ASCOT-LLA

CARDS

JUPITER:

- Median age at start: 66 years
- Marginal population: 0.23%
- Length of study: Varies, people signed up through most of 2003-2006, and the analysis considers all events through when it was terminated on March 29, 2008. Let's call it 3 years?
- Life expectancy of a 66 year old American in 2008 (since I can't tell where the subjects in this study came from): 18.16 years
- Adjustment: 18.16 × 365 × 0.0023 = 15.24 days

4S

LIPID:

Putting this together, we finally get these estimates:

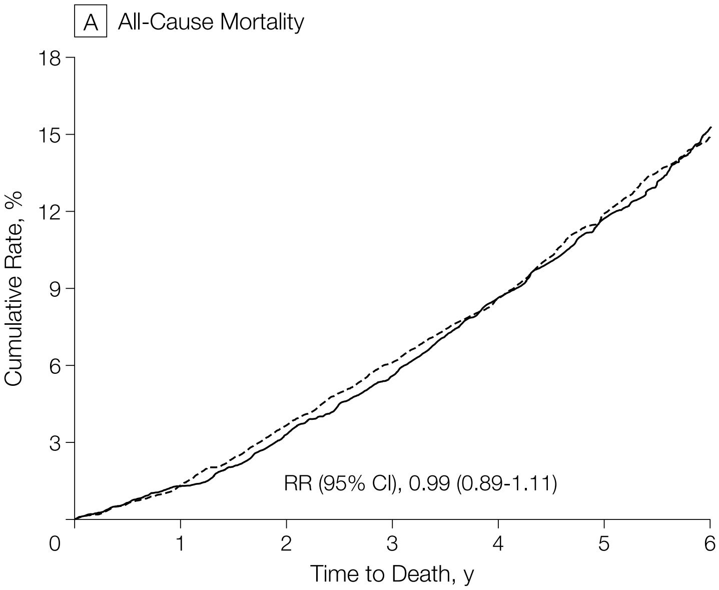
study	Original days	Adjusted days	Length of study (years)
WOSCOPS	9	65	5
ASCOT-LLA	2	33	3.5
CARDS	19	114	4.8
JUPITER	7	23	4
4S	27	264	5.8
LIPID	22	184	6.1

The increases range from a factor of 3 to a factor of 16.

The major trend in the above table is that larger studies tend to find larger effects, both in terms of the original estimate and my very rough adjustment. For this reason, if I had to guess, I'd guess that my adjustments are still too low, i.e. the benefits of statins are larger than the above would suggest.

There's a bunch of other studies that I didn't include here, either because they weren't placebo-controlled or because the study population was already having particular medical issues.

There's a bunch of studies I didn't use. One is **ALLHAT-LLT** (2002), which 10,355 North Americans who were 55 or older with moderately high cholesterol and triglycerides that weren't *too* high (LDL-C between 120 and 189 mg/dL and triglycerides less than 350 mg/d). They gave people either 40mg/d pravastatin or usual care.



Here the black line is pravastatin, and the dotted line is usual care. There's minimal effect, and for much of the study, the usual care group even did a bit better. At the end of the study, a *slightly* lower fraction of those in the pravastatin group had died (14.9% vs 15.3%), but it's obviously not significant.

While this at first seems to suggest that statins don't do anything, there one *huge* caveat: The "usual care" group kept going to see their doctors, who were free to prescribe them statins, which many did. This means that around 30% of the usual care group *also* took statins. Arguably, what this study shows is that mindlessly giving everyone the same dose of the same statin works as well as having thousands of doctors examine everyone and apply all of their talents. That's certainly *interesting*, but probably beside the point for us.

Two other studies (MEGA, and GISSI-P) also weren't placebo-controlled. Finally, GISSI-HF had subjects on dialysis and CORONA had subjects with ongoing systolic heart failure. These are all so different that I'm skipping them as well.

Now, I kept every study with a placebo control that targeted a "normal" population, which I think is the most sensible approach. But I still feel compelled to mention that these rules exclude 5 of the 6 studies with the smallest (or even negative) effects. That's particularly strange with non-placebo-controlled studies, which you'd expect to produce a *larger* effect size, not a smaller one.

Takeaways

If your doctor says you should take statins, then you should probably take statins. I think the previous analysis that said you only get a few weeks is too conservative, and a better estimate is a few months.

(By the way, Kristensen et al. are well aware that their analysis might be conservative—none of this is a criticism of them!)

Now, even with my adjustments, you might find these effects to be surprisingly small. If the above table makes you think statins aren't worth the trouble, let me make a couple of points:

1. Try inversion: Suppose I offered you a lifetime supply of delicious cookies, one per day, except if you eat the cookies, you'll die a few months earlier. Would you take them? (For this to work, you're supposed to find the pleasure of the cookies equivalent to the inconvenience of taking statins.)
2. Lifestyle interventions that buy you a few months are a big deal! If you want to have a longer, healthier life, it will probably come from a combination of things, each of which has a modest effect. "A few months here, a few months there, pretty soon you're talking about real money."

Still—and despite how dismally non-contrarian all this is—please don't listen to me when making medical decisions!

Reasons and Persons: Watch theories eat themselves

I've long been fascinated by Derek Parfit's *Reasons and Persons.* This is often listed as a favorite philosophy book, e.g. by Peter Singer and David Chalmers. How can you not love something that starts like this?

> We are particular people. I have my life to live, you have yours. What do these facts involve? What makes me the same person throughout my life, and a different person from you? And what is the importance of these facts?

How? I'll tell you how! Turn to page 107.

> Those who hold a Constructivist view may question my division of a moral theory. (R1) revises what I call our Ideal Act Theory. Constructivists may see no need for this part of a moral theory. But they cannot object to my proposal that we should ask what we should all do, simply on the assumptions that we shall all try, and all succeed. Answering this question is at worst unnecessary. If a Constructivist asks what we should all ideally do, his answer cannot be some version of Common-Sense Morality. If he accepts some version of this morality, he must move to the corresponding version of (R1), the revised version of his morality that would not be directly collectively self-defeating. And, since he should accept (R1), he should also accept (R2) and (R3). He should revise his Practical Act Theory, the part that used to be his whole theory.

There are two issues here.

For one, this is hard to read. Now, Parfit isn't *trying* to be obscure, it's just that he's happy to make things 100% more difficult to get 10% more precision. I'm sure that's right for academic philosophers, but I'd like a different tradeoff.

Second, I sometimes, well… I sometimes don't care about the questions Parfit is trying to answer. The above paragraph is asking if people who think morality is created by society should accept a certain way of splitting up moral theories. To me, that feels like agonizing about definitions.

You might think that these issues mean the book is uninteresting, but that's not true *at all.* Parfit's arguments are anchored by a series of evocative thought experiments. These are accessible and mind-expanding independently of the hulking logical arguments he builds on top of them.

So this is the summary of *Reasons and Persons* I would have liked to read: The goal is to provide a tour of the thought experiments and let you (mostly) decide for yourself what you think about them.

Warning: I often changed the scenarios quite a lot, and added some new ones. (Parfit has no alien viruses.) I also changed the order of things, and greatly shortened or dropped most of the detailed arguments. I think this still gives *a lot* of the value, but it's not a full representation of Parfit's ideas.

- How self-interest gets into trouble
- On rational irrationality
- On how deontology gets into trouble
- How consequentialism get into trouble
- How consequentialism might save itself
- Satan
- Coordination problems
- Assigning blame
- Morality decompositions
- Takeaways
- Thoughts

How self-interest gets into trouble

The first part of the book asks, do ethical theories eat themselves? Let's start with the idea that it's rational to do what's best for you.

The desert hitchhiker

You're self-interested but unable to lie. When driving through the desert, your car breaks down. A stranger stops and offers you a ride into town for $20. You'd be *thrilled* to pay $20, but you have no money on you. The stranger asks, "Well if I give you a ride, will you pay me $20 later?"

You think about it. Once you're in town, the stranger can't force you to pay and so, since you're self-interested, you won't. Since you can't lie, you admit this to the stranger, and you're left on the side of the road.

The firefighting pact

You are self-interested. One day, your neighbors have a meeting and point out that the weather is hot and dry, and fires could happen at any time. They propose that everyone should swear that if a fire breaks out at anyone's house, everyone will help out. However, they have a lie detector, and under examination discover that you wouldn't show up when needed. Thus, you're excluded from the pact. When your house catches on fire, it burns to the ground.

Kate the writer

Kate is an altruistic writer. She believes her work is important for the world, so she works like mad until she collapses in exhaustion and depression. She doesn't like being exhausted and depressed, but she thinks that the good she does for others outweighs her pain.

Fortunately, you have a science-fiction neuroscience gun. You zap Kate's brain and make her completely selfish. She immediately quits writing, since she did that for the benefit of others. But now she finds her life is less meaningful and is less happy than when she was altruistic.

Takeaway: There are scenarios where being self-interested seems to make you *worse* off, not better.

Are you disturbed by these examples? Many people say, "So what? Just do what's best for you overall." OK, but how does that work? Once the stranger has given you a ride into town, it *is* in your interest to stiff them, no matter what you said in the past.

Adding an exception like, "Follow your self interest unless you made a promise" doesn't work. If you realize the stranger would use your $20 to buy a knife and stab you, you should renege. So you can't "*just*" do what's better for you overall. That's very hard, or maybe even impossible.

On rational irrationality

As further evidence that self-interest isn't necessarily good for you, there are situations where you might rationally choose to make yourself crazy.

Schelling's answer to armed robbery

A burglar breaks into your house. He threatens that you must open your safe or he'll kill your children.

Fortunately, you are an expert on conflict strategy and have a potion that makes you temporarily indifferent to the lives of your kids. You quickly drink it and then laugh at the man's threats. "You want to kill my daughter? Go ahead!"

The burglar realizes that there's nothing to be gained by hurting anyone, so he leaves before the police show up.

Alice's brain modification

You live with a group of people on an island, gathering and eating coconuts. You are all rational and self-interested.

Tired of working so hard, Alice builds a machine and implants it in her brain. This machine leaves her rational *except* when it comes to fulfilling threats, which she always does regardless of the damage to herself. She announces to the group, "I will not be gathering any more coconuts. Either you gather coconuts for me, or I'll burn down all the coconut trees and we'll all starve and die."

You regretfully conclude that your only choice is to do Alice's work for her. But you wonder: Did you do go wrong somewhere?

Eventually, you realize: Upon arrival on the island, your first task should have been to implant a machine in *your* brain and promise, "I'll burn down the coconut trees if anyone else installs any machines in their brains."

On how deontology gets into trouble

Next up is deontology, the idea that morality can be defined by obeying a set of universal rules, such as “don’t lie”, “don’t steal”, etc.

The obscene film

A man breaks into your house. He says, “If you don’t allow me to film you doing [obscene act], I will kill your children. If you do allow me, I will later use that film to blackmail you into doing minor crimes.” If you make the film, your kids are safe forever. But you know that, given your personality, you would give in to the blackmail and do the crimes.

Should you allow him to make the film?

Murder and accidental death

Ben is about to die. But right before he does, he plans to murder Cathy. Meanwhile, a fire is closing in on Deb, who will die unless she is rescued. The lives of Cathy and Deb are equally valuable.

You have enough time to either convince Ben not to murder Cathy or to rescue Deb, but not both. You have a 50% chance of convincing Ben not to do the murder and a 51% chance of rescuing Deb.

What should you do?

The question here is if only *outcomes* matter, or if it matters what actions people make. In the first scenario, you choose between (A) doing some minor crimes, or (B) someone *else* doing a horrible murder. If morality is about avoiding doing bad things, then you should choose (B) since then *you* do nothing wrong. Similarly, in the second scenario, the question is if you’re just trying to avoid *death*, or trying to avoid *murder*.

How consequentialism get into trouble

Finally, let’s look at consequentialism, the idea that only consequences matter.

Clare’s child

Clare loves her child. She can spend $50 to buy her kid a wonderful dinner, or to cure a stranger of a horrible disease. She chooses dinner.

You suggest to Clare that this was wrong. She says, “Given how much I love my kid, it was impossible for me not to spend the money on her. And it would be wrong for me to make myself love my kid less. So I can’t really be blamed here.”

If only consequences matter, is she right?

The alien virus

People get much of their joy in life from “selfish” desires like eating pizza or playing with their kids. One day, aliens happen by the Earth. Noticing that we are such primitives, they decide to help us out: They drop a virus that transforms everyone into pure consequentialist do-gooders.

All people now focus only on increasing the average good in the world. Desires like “play with my kids” are dropped since they are “agent-dependent” and so not consequentialist.

In some ways, the world immediately becomes some kind of zero-externality socialist paradise: Litter disappears! Fisherman stop fishing when stocks start to deplete! There’s no need for taxes. People work as hard as they can and donate everything to the needy. There’s no need for fences or police.

But people report that their lives feel meaningless. Evolution put the rewards for enjoying dinner and taking care of our kids deep within us. When we neglect these “agent-dependent” goals, we lose a lot. Average happiness falls.

Parfit takes these scenarios very seriously. He thinks consequentialism is seriously broken because humans couldn’t have rich lives without agent-relative (and thus non-consequentialist) desires.

How consequentialism might save itself

Esoteric theories

Let’s continue the alien virus scenario. When we stopped, people lived in a clean socialist paradise but they were depressed because their lives lost meaning without being able to focus on themselves and those closest to them.

Now, the top minds of humanity get together. They calculate that if they transform everyone back into *non*-consequentialists, the average happiness in the world will go up. Since they are (currently) consequentialists, they consider this to be a good thing. So they create a vaccine that neutralizes the virus and makes people as selfish as before. They vaccinate everyone and destroy any evidence that this era ever happened.

Time passes. By the year 2907, people live in small orbital space clusters. In this new situation, people would be happier if their circle of concern happened to be everyone in their cluster. But they’re stuck with their old outlook.

Fortunately, a small group had refused the vaccine and endured secret consequentialist misery over the centuries. They decide their time has come, and so they re-release the alien virus. After the population is converted back to consequentialism, they calculate the moral outlook suited for the current age and engineer a virus that gives people that morality. Since everyone is (now) consequentialist, they agree that this is a good thing, so everyone takes the virus. They again destroy all the evidence, except leaving a secret cabal to keep the flame of consequentialism alive, waiting for the day to again inflict a brief flash of enlightenment on everyone else.

The argument here is that, sure, consequentialism might lead people to choose some other (non-consequentialist) morality. But even if that's true, people should keep some trace of consequentialism around in case circumstances change and that secondary morality needs to be adjusted.

Satan

Satan rules the universe

Maybe Satan rules the universe. If so, he could make it so that anyone who is self-interested has a horrible life. Or he could make it so that if everyone is utilitarian, then average utility is low. Or he could make it so that people who believe it's wrong to lie end up lying more—i.e. believing in deontology leads to worse deontological outcomes.

Now, maybe Satan doesn't rule the universe. But what if he did? It's worth worrying about since in our universe these things probably *are* at least a little true in some situations.

Coordination problems

In the real world, what happens doesn't just depend on you, it depends on other people. How should we think about this?

The prisoner's dilemma

You're probably familiar with this, but just in case: You and I are in separate rooms. We each have a red button and a blue button. Depending on what we press, we'll get different amounts of money. Here's what you'll get:

You \ me	Red	Blue
Red	$1	$3
Blue	$0	$2

And here's what I get:

You \ me	Red	Blue
Red	$1	$0
Blue	$3	$2

No matter what I do, you always get $1 more by choosing red. The same is true for me. Since we're both selfish brutes, we both choose red, and so we only get $1.

Parfit argues that, in practice, 2-person prisoner's dilemmas are rare. One reason is that most prisoner's dilemmas are *repeated*. Even if I'm a sociopath, I probably don't want to take all my roommate's stuff and trade it for drugs because I don't want my roommate to do that to me later.

More importantly, people conspire! You need an outside force to stop them. That's why the cops in movies keep the prisoners in different rooms and lie to them about what the other one is doing.

What *is* common is *multi-party* prisoner's dilemmas.

Littering

It's convenient to litter, and it's mostly other people that have to look at it and clean it up. Since I'm a low-quality person, I litter.

Overfishing

You're a fisherman in an area where the fish population is collapsing. You know everyone needs to slow down and let the fish recover, but you also know that if you *don't* fish, someone else will and the fish will collapse anyway. So everyone fishes and pretty soon the fish are all gone.

Public transit

Our roads are clogged. If we all took the bus, the roads would be clear, and we'd all get to work faster. But, if everyone *else* took the bus, you could drive and get to work *even faster*. So everyone drives and the roads stay clogged.

Prisoner's dilemmas with kids

We all want our kids to have relaxed childhoods. However, if I push my kid *slightly* harder than you do, then my kid will get into a fancier college. Slowly, everyone comes to the same realization and pushes their kids harder and harder until every childhood is an endless misery of studying.

Weird outcome matrices

You and I are put into separate rooms. We each have three buttons. Depending on what the two of us press, we'll each get some BAD, OK, GOOD, or GREAT outcome. Here is the outcome matrix.

You \ me	1	2	3
1	OK	BAD	BAD
2	BAD	GOOD	GREAT
3	BAD	GREAT	BAD

If we both press button 1, have we done the right thing, according to consequentialism?

You can create many other weird possibilities with these outcome matrices. These all lead you to ask: Does the moral thing to do depend on what you expect others to do? If yes, then you get stuck in the OK outcome above. If not, then you might get terrible results if others misbehave. Is it right to assume other people will behave well, or should you try to guess, or what?

Parfit argues that commonsense morality evolved in small communities where what you did could only affect a few others. But modern life creates many more situations where we can benefit ourselves by creating a larger cost to others. Thus, he suggests we need a new morality to adjust to these new times.

Assigning blame

Sometimes it's hard to say who did the right thing.

1st rescue mission

There are 100 people trapped in a cave, and four people are needed to rescue them. You could go on the mission, or go rescue 10 *different* people. If you don't go on the mission, someone else will take your place. What should you do?

2nd rescue mission

There are 100 people trapped in a cave, and four people are needed to rescue them. Only four people are available, one of which is you. Alternatively, you could go rescue 50 people on your own. Is that better (since 50 is more than 1/4 of 100)?

These scenarios suggest that *marginal* changes are what matter, that the good of your actions is what difference they make, holding everyone else's actions constant. Cool, except…

1st execution

Alfred and Bert hate Carlos so they get together and shoot him at the same time. Either bullet would have killed Carlos. Did either of them do anything wrong?

2nd execution

You hear that a million people who hate Carlos are going to show up and shoot him at midnight. The mob cannot be stopped. Since you hate Carlos too, you also shoot him at the same time.

3rd execution

Carlos needs $1000 to buy medicine or he'll die. He asks a million people for a donation, and they all refuse. You also refuse.

4th execution

I could spend $1000 to save the life of a child somewhere, but I don't.

1st poisoning

Alfred gave Carlos a poison that will painfully kill him in a few minutes. Before that happens, Bert shows up and shoots Carlos, killing him instantly and painlessly. Was that wrong?

2nd poisoning

Alfred gave Carlos a poison that will painfully kill him in a few minutes. Meanwhile, you are about to be hit by a truck. Bert throws Carlos in front of the truck, instantly killing him and saving you. Was that right?

Drops of water

A thousand thirsty people are in the desert. A truck is going out to them with a big barrel in it. You have a single liter of water, that you could pour into the barrel to be distributed equally to everyone. No one can sense the difference of 1mL of water. Does that mean that you don't need to add your water?

There are a ton of examples like this, with drops of water or torturers with buttons that cause an imperceptible increase in pain. These are meant to dispute the idea that imperceptible effects can't be bad. I find these utterly boring, probably because I'm already convinced that "imperceptible" effects can be bad—if something is imperceptible, isn't that just a comment on the quality of someone's sensory system?

Morality decompositions

Two key distinctions

Moral theories vary in two key ways:

1. They can be *agent-neutral* or *agent-relative*. In an agent-neutral theory, it should be possible to take any world state and say how good it is. In an agent-relative theory, you can only say how good it is from one person's perspective.
2. They might care about *what happens*, or also *what we do*. The question here is if actions themselves matter, or just the result.

This matrix illustrates the different attitudes you might take to kids being fed, depending on which kind of moral theory you are operating with.

	What we do matters	What happens matters
agent-relative	I should feed my kids	My kids should be fed
agent-neutral	Parents should feed their kids	Kids should be fed

Traditional deontology would be in the upper-left corner. Traditional consequentialism would be in the lower-right corner.

Five parts of a moral theory

As well as having different moral theories, you can take a theory and break it up into pieces. For example, maybe you want to be pragmatic. You might say that in an ideal world, we would have moral theory **A**, but given that many people are jerks, we should instead have moral theory **B**.

Somewhat separate from morals are *motives*, the desires that animate us. Maybe the *desire itself* to take care of our kids makes us happy, independent of the effects of that desire. You might say that in an ideal world, we would have motives **C**, but that given that many people are jerks, we should have motives **D** instead.

Finally, you can also think about **E**, how we should react to bad acts. (Here no theory is needed in a perfect world!) You can picture these different parts of a moral theory like this:

	Ideal	Practical
Successful acts	Ideal Act Theory	Practical Act Theory
Motives	Ideal Motive Theory	Practical Motive Theory
Blame and remorse	n/a	Reaction Theory

You can spell out these five theories in more detail.

- *Ideal Act Theory* is what we should do if everything was predictable and everyone did the right thing.
- *Practical Act Theory* is what we should do, given that the world is uncertain and some people are jerks.
- *Ideal Motive Theory* is what motives we should have, given that motives had effects other than our acts.
- *Practical Motive Theory* is what motives we should have, given that the world is uncertain and some people are jerks.
- *Reaction theory* is what acts we should blame people for, given that the world are uncertain, some people are jerks, and blame has complex effects on the future.

This is useful: If you're having a moral debate with someone, make sure that you're talking about the same cell in the above matrix!

Takeaways

I think Parfit makes three main arguments.

1. Self-interest theory is collectively self-defeating due to prisoners dilemmas. (And gets complex at the individual level due to conflict strategy.)
2. Consequentialism is indirectly self-defeating because it is *agent-neutral*—if we only look at outcomes without considering the agent, our lives would be empty, since we've evolved to care about our relationships.
3. Commonsense morality is self-defeating because it is *agent-relative*. If different people are supposed to prioritize loved ones, then everyone will screw over the world for their ingroup.

Now, it's bad that self-interest theory is self-defeating, but that doesn't mean it is "wrong"—it never claimed to give globally optimal outcomes. It only promises to give you the best outcome given what everyone else is doing. It's perhaps more concerning that being self-interested might leave you stranded in the desert or excluded from firefighting pacts, but maybe you can overcome this with clever self-interested meta-reasoning.

It's more of a problem that consequentialism and commonsense morality are self-defeating. After all, these are moral theories—you'd expect them to lead to optimal outcomes.

Parfit suggests that a solution might come from some kind of merger of consequentialism and commonsense morality. He doesn't say exactly how this would work (except that it's not simple) but suggests a kind of agent-neutral version of commonsense morality as a starting point: Maybe we should all collectively work to make "kids are fed by their parents" happen, as opposed to robotically trying to feed *all* kids (consequentialism) or just trying to feed *our* kids (commonsense morality).

Thoughts

I had two major thoughts after reading this. First off, it's striking that Parfit doesn't engage with economic ideas. For example, he is concerned that self-interest theory might lead to poor self-interested outcomes, and that consequentialism might lead to poor consequentialist outcomes. That's fine, but surely the major problem in our actual world is that *consequentialist* behavior leads to poor *self-interested* outcomes. The implicit view in economics is that it's society's job to align selfish interests with public interests. I'd have liked to know what Parfit thought about that.

Similarly, Parfit is concerned about the bad effects of multi-party prisoners dilemmas. I agree these are a serious, problem, but let's remember that these also have *positive* effects. After all, they are the entire basis of capitalism! Companies could make massive profits if they all agreed to keep their prices high. But usually—we hope—at least one greedy company will try to screw the others over by cutting prices, and so the cartel collapses.

A second thought is just how strong of an influence Schelling's The Strategy of Conflict is. While Schelling is concerned with war, conflict, negotiation, etc., it is remarkably similar in that it starts with simple principles but as you start to consider reactions and counter-reactions, you seem to get a spiral of ever-increasing levels of complexity. Parfit is aware of this, but he never seems to push things *quite* as far as he could. Here's an example:

How much should the rich give to the poor?

After making billions in Silicon Valley, you read The Most Good You Can Do and decide you're a consequentialist. You ask yourself, how much of your money should you give away? Under mild assumptions, the answer seems to be *almost everything.*

But then, you have Thomas Schelling over for dinner. He points out that you should be careful: When you donate, you aren't just giving away your money, you are also helping to establish a *social norm* for how much the rich should give away. Perhaps giving away almost everything will make people think that consequentialists are crazy, and so many people will give nothing. So maybe it would be better to try to establish a gentle norm, like giving away 10% of your wealth.

This is (roughly) where Parfit stops. But you can keep going. Maybe what you should do is *secretly* donate almost all your money, but pretend in public like you only gave away 10%, so people don't realize consequentialism is a bummer. And maybe at the same time, you should try to secretly find other like-minded billionaires and tell them that you did actually give most of your money away? Or maybe you should build an AI to predict each person's appetite for donation, and coordinate a society-wide deception so that everyone thinks there is a different norm?

Or maybe all this is outweighed by the fact that your lies could be discovered or that dishonesty has other bad effects? And if you think lying is bad, does that mean that even the original gambit of trying to create a norm of 10% is bad?

I could keep going forever here, but you get the point—things spiral and it never ends.

Sadly, that's my main conclusion from all of this. Nothing works. Whatever system you commit to, it's always possible to jump up one level of abstraction and break things. There's no answer for this, or at least Parfit doesn't provide one. (Though honestly, in practice things don't get broken *too* badly, and we can solve most of the problems while only climbing the abstraction ladder a level or two.)

So that's part one of the book: Parfit breaks morality. In part two (coming soon) he will wield time for even more breakage, with gambits like "if time is an illusion, then…" Part three is the good stuff—breaking the idea of personal identity.

References

1. In defense of Myers-Briggs (dynomight.net)
2. Comparative advantage and when to blow up your island (dynomight.net)
3. What happens if you drink acetone? (dynomight.net)
4. Making the Monty Hall problem weirder but obvious (dynomight.net)
5. Your ratios don't prove what you think they prove (dynomight.net)
6. The veil of darkness (dynomight.net)
7. Pragmatic reasons to believe in formal ethics (dynomight.net)
8. Simpson's paradox and the tyranny of strata (dynomight.net)
9. Why I'm skeptical of universal basic income (dynomight.net)
10. Sales tax creates more unnecessary pain than value added tax (dynomight.net)
11. Experiments on a $50 DIY air purifier you can make in 30s (dynomight.net)
12. How to run without all the pesky agonizing pain (dynomight.net)
13. Better air quality is the easiest way not to die (dynomight.net)
14. The irrelevance of test scores is greatly exaggerated (dynomight.net)
15. Are some personalities just better? (dynomight.net)
16. Alcohol, health, and the ruthless logic of the Asian flush (dynomight.net)
17. A review of early split-brain experiments (dynomight.net)
18. Does the gender-equality paradox actually exist? (dynomight.net)
19. It's perfectly valid for a trait to be more than 100% heritable (dynomight.net)
20. Factors of mental and physical abilities (dynomight.net)
21. Political polarization is partly a sample bias illusion (dynomight.net)
22. Statistical nihilism and culture-war island hopping (dynomight.net)
23. The big alcohol study that didn't happen: My primal scream of rage (dynomight.net)
24. Two conspiracy theories about cola (dynomight.net)
25. The main thing about P2P meth is that there's so much of it (dynomight.net)
26. A breakdown of the data on the homeless crisis across the U.S. (dynomight.net)
27. The death penalty as a lens on democracy (dynomight.net)
28. How Germany banned the death penalty (dynomight.net)
29. How the United Kingdom banned the death penalty (dynomight.net)
30. How France banned the death penalty (dynomight.net)
31. How the United States didn't ban the death penalty (dynomight.net)
32. Underrated reasons to be thankful (dynomight.net)
33. Effective selfishness (dynomight.net)
34. How many extra days of life do you get from taking statins? (dynomight.net)
35. Reasons and Persons: Watch theories eat themselves (dynomight.net)