Best of LessWrong: December 2016

Best of LessWrong: December 2016

Fact Posts: How and Why

The most useful thinking skill I've taught myself, which I think should be more widely practiced, is writing what I call "fact posts." I write a bunch of these on my blog. (I write fact posts about pregnancy and childbirth here.)

To write a fact post, you start with an empirical question, or a general topic. Something like "How common are hate crimes?" or "Are epidurals really dangerous?" or "What causes manufacturing job loss?"

It's okay if this is a topic you know very little about. This is an exercise in original seeing and showing your reasoning, not finding the official last word on a topic or doing the best analysis in the world.

Then you open up a Google doc and start taking notes.

You look for *quantitative data from conventionally reliable sources*. CDC data for incidences of diseases and other health risks in the US; WHO data for global health issues; Bureau of Labor Statistics data for US employment; and so on. Published scientific journal articles, especially from reputable journals and large randomized studies.

You explicitly do *not* look for opinion, even expert opinion. You avoid news, and you're wary of think-tank white papers. You're looking for raw information. You are taking a *sola scriptura* approach, for better and for worse.

And then you start letting the data show you things.

You see things that are surprising or odd, and you note that.

You see facts that seem to be inconsistent with each other, and you look into the data sources and methodology until you clear up the mystery.

You orient *towards* the random, the unfamiliar, the things that are totally unfamiliar to your experience. One of the major exports of Germany is *valves*? When was the last time I even thought about valves? *Why* valves, what do you use valves in? OK, show me a list of all the different kinds of machine parts, by percent of total exports.

And so, you dig in a little bit, to this part of the world that you hadn't looked at before. You cultivate the ability to spin up a lightweight sort of fannish obsessive curiosity when something seems like it might be a big deal.

And you take casual notes and impressions (though keeping track of all the numbers and their sources in your notes).

You do a little bit of arithmetic to compare things to familiar reference points. How does this source of risk compare to the risk of smoking or going horseback riding? How does the effect size of this drug compare to the effect size of psychotherapy?

You don't really want to do *statistics*. You might take percents, means, standard deviations, maybe a Cohen's *d* here and there, but nothing fancy. You're just trying to figure out what's going on.

It's often a good idea to rank things by raw scale. What is responsible for the bulk of deaths, the bulk of money moved, etc? What is *big*? Then pay attention more to things, and ask more questions about things, that are *big.* (Or disproportionately high-impact.)

You may find that this process gives you contrarian beliefs, but often you won't, you'll just have a strongly fact-based assessment of *why* you believe the usual thing.

There's a quality of *ordinariness* about fact-based beliefs. It's not that they're never surprising -- they often are. But if you do fact-checking frequently enough, you begin to have a sense of the world overall that *stays in place*, even as you discover new facts, instead of swinging wildly around at every new stimulus. For example, after doing lots and lots of reading of the biomedical literature, I have sort of a "sense of the world" of biomedical science -- what sorts of things I expect to see, and what sorts of things I don't. My "sense of the world" isn't that the *world itself* is boring -- I actually believe in a world rich in discoveries and low-hanging fruit -- but the sense *itself* has stabilized, feels like "yeah, that's how things are" rather than "omg what is even going on."

In areas where I'm less familiar, I feel more like "omg what is even going on", which sometimes motivates me to go accumulate facts.

Once you've accumulated a bunch of facts, and they've "spoken to you" with some conclusions or answers to your question, you write them up on a blog, so that other people can check your reasoning. If your mind gets changed, or you learn more, you write a follow-up post. You should, on any topic where you continue to learn over time, feel embarrassed by the naivety of your early posts. This is fine. This is how learning works.

The advantage of fact posts is that they give you the ability to form independent opinions based on evidence. It's a sort of practice of the skill of seeing. They likely aren't the optimal way to get the most accurate beliefs -- listening to the best experts would almost certainly be better -- but you, personally, may not know who the best experts are, or may be overwhelmed by the swirl of controversy. Fact posts give you a relatively low-effort way of coming to informed opinions. They make you into the proverbial 'educated layman.'

Being an 'educated layman' makes you much more fertile in generating ideas, for research, business, fiction, or anything else. Having facts floating around in your head means you'll naturally think of problems to solve, questions to ask, opportunities to fix things in the world, applications for your technical skills.

Ideally, a *group* of people writing fact posts on related topics, could learn from each other, and share how they think. I have the strong intuition that this is valuable. It's a bit more active than a "journal club", and quite a bit more casual than "research". It's just the activity of learning and showing one's work in public.

"Flinching away from truth" is often about *protecting* the epistemology

Related to: Leave a line of retreat; Categorizing has consequences.

There's a story I like, about this little kid who wants to be a writer. So she writes a story and shows it to her teacher.

"You misspelt the word 'ocean'", says the teacher.

"No I didn't!", says the kid.

The teacher looks a bit apologetic, but persists: "'Ocean' is spelt with a 'c' rather than an 'sh'; this makes sense, because the 'e' after the 'c' changes its sound…"

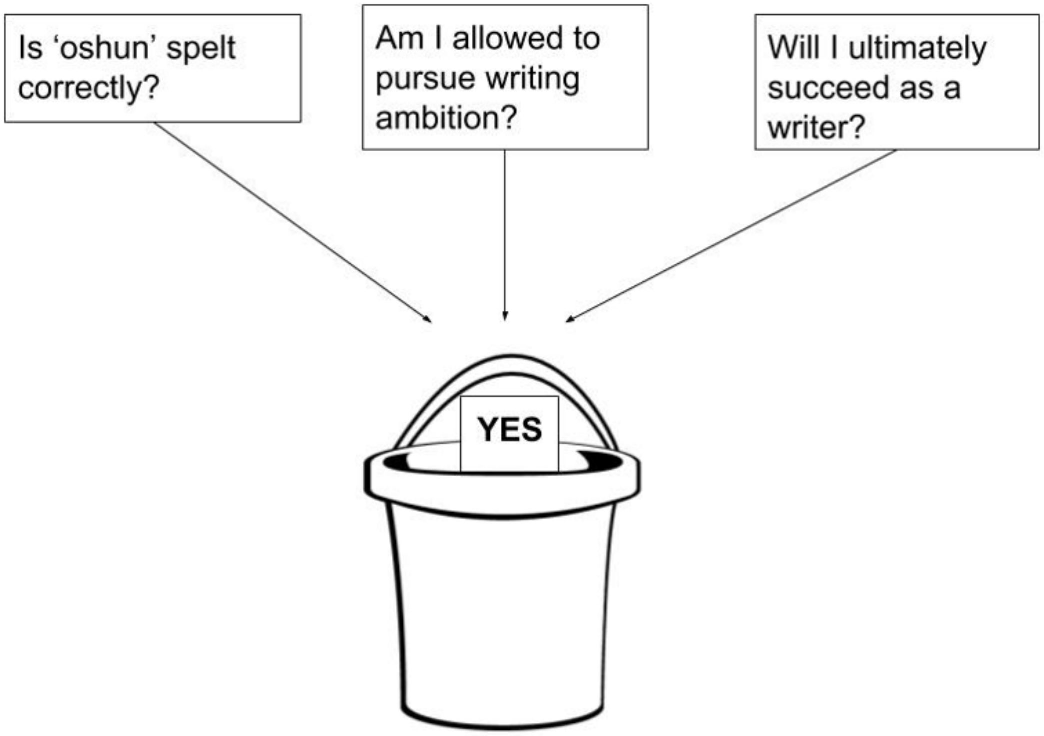
"*No I didn't!*" interrupts the kid.

"Look," says the teacher, "I get it that it hurts to notice mistakes. But that which can be destroyed by the truth should be! You did, in fact, misspell the word 'ocean'."

"I *did not*!" says the kid, whereupon she bursts into tears, and runs away and hides in the closet, repeating again and again: "I did not misspell the word! I can too be a writer!".

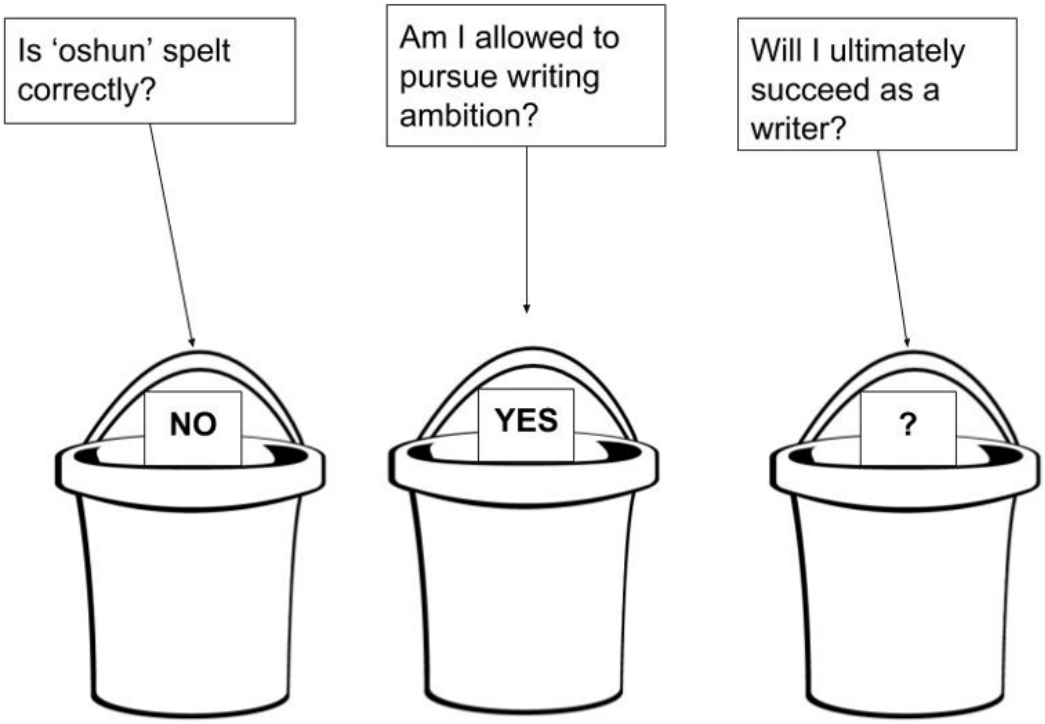
I like to imagine the inside of the kid's head as containing a single bucket that houses three *different* variables that are initially all stuck together:

Original state of the kid's head:


Is 'oshun' spelt correctly?
Am I allowed to pursue writing ambition?
Will I ultimately succeed as a writer?
YES

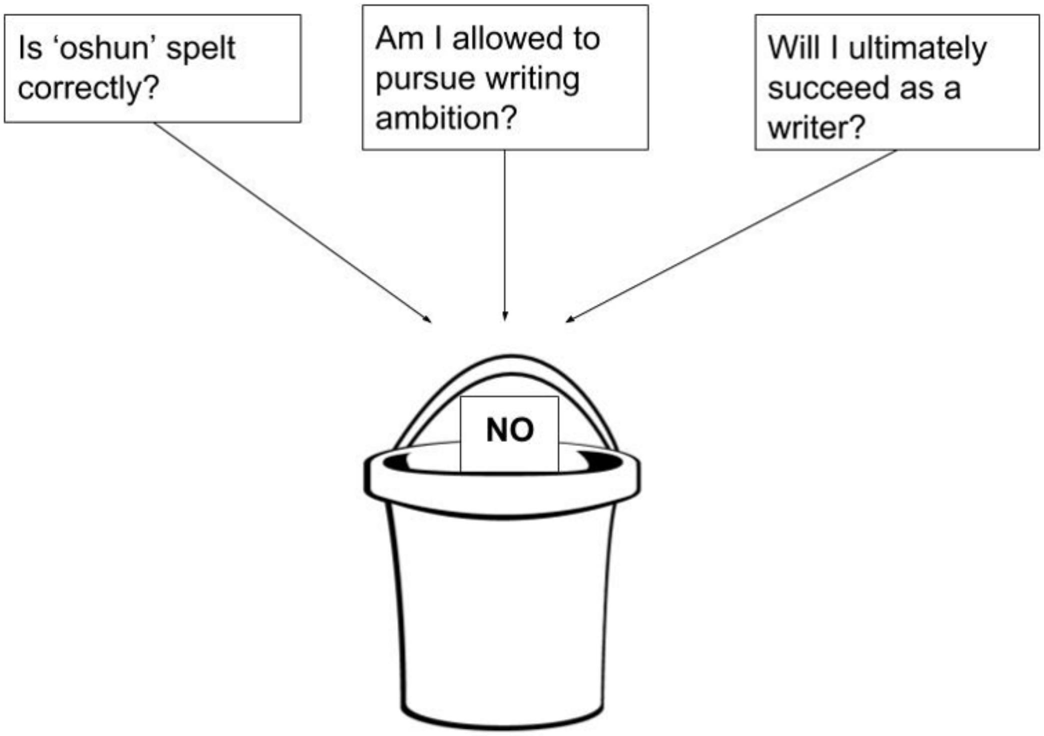
The goal, if one is seeking actual true beliefs, is to separate out each of these variables into its own separate bucket, so that the “is ‘oshun’ spelt correctly?” variable can update to the accurate state of "no", without simultaneously forcing the "Am I allowed to pursue my writing ambition?" variable to update to the *in*accurate state of "no".

Desirable state (requires somehow acquiring more buckets):


Is ‘oshun’ spelt correctly?
Am I allowed to pursue writing ambition?
Will I ultimately succeed as a writer?
NO
YES
?

The trouble is, the kid won’t necessarily acquire enough buckets by trying to “grit her teeth and look at the painful thing”. A naive attempt to "just refrain from flinching away, and form true beliefs, however painful" risks introducing a more important error than her current spelling error: mistakenly believing she must stop working toward being a writer, since the bitter truth is that she spelled 'oshun' incorrectly.

State the kid might accidentally land in, if she naively tries to "face the truth":

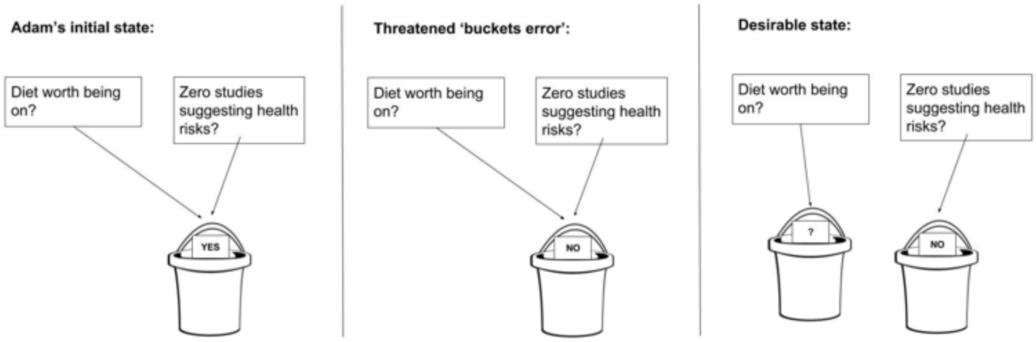

Is ‘oshun’ spelt correctly?
Am I allowed to pursue writing ambition?
Will I ultimately succeed as a writer?
NO

(You might take a moment, right now, to name the cognitive ritual the kid in the story *should* do (if only she knew the ritual). Or to name what you think you'd do if you found yourself in the kid's situation -- and how you would *notice* that you were at risk of a "buckets error".)

More examples:

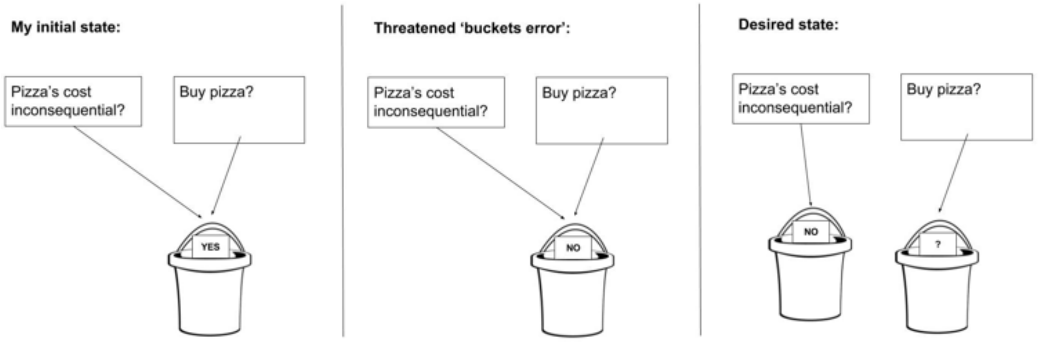
It seems to me that bucket errors are actually pretty common, and that many (most?) mental flinches are in some sense attempts to avoid bucket errors. The following examples are slightly-fictionalized composites of things I suspect happen a lot (except the "me" ones; those are just literally real):

Diet: Adam is on a diet with the intent to lose weight. Betty starts to tell him about some studies suggesting that the diet he is on may cause health problems. Adam complains: “Don’t tell me this! I need to stay motivated!”


Adam’s initial state:
Diet worth being on?
Zero studies suggesting health risks?
YES
Threatened ‘buckets error’:
Diet worth being on?
Zero studies suggesting health risks?
NO
Desirable state:
Diet worth being on?
Zero studies suggesting health risks?
?
NO

One interpretation, as diagramed above: Adam is at risk of accidentally equating the two variables, and accidentally *assuming* that the studies imply that the diet must stop being viscerally motivating. He semi-consciously perceives that this risks error, and so objects to having the information come in and potentially force the error.

Pizza purchase: I was trying to save money. But I also wanted pizza. So I found myself tempted to buy the pizza *really quickly* so that I wouldn't be able to notice that it would cost money (and, thus, so I would be able to buy the pizza):


My initial state:
Pizza's cost inconsequential?
Buy pizza?
YES
Threatened 'buckets error':
Pizza's cost inconsequential?
Buy pizza?
NO
Desired state:
Pizza's cost inconsequential?
Buy pizza?
NO
?

On this narration: It wasn't *necessarily* a mistake to buy pizza today. Part of me correctly perceived this "not necessarily a mistake to buy pizza" state. Part of me also expected that the rest of me _wouldn't_ perceive this, and that, if I started thinking it through, I might get locked into the no-pizza state _even if pizza was better_. So it tried to 'help' by buying the pizza *really quickly, before I could think and get it wrong*. [1]

On the particular occasion about the pizza (which happened in 2008, around the time I began reading Eliezer's LW Sequences), I actually managed to notice that the "rush to buy the pizza before I could think" process was going on. So I tried promising myself that, if I still wanted the pizza after thinking it through, I would get the pizza. My resistance to thinking it through vanished immediately. [2]

To briefly give several more examples, without diagrams (you might see if you can visualize how a buckets diagram might go in these):

- Carol is afraid to notice a potential flaw in her startup, lest she lose the ability to try full force on it.
- Don finds himself reluctant to question his belief in God, lest he be forced to conclude that there's no point to morality.
- As a child, I was afraid to allow myself to actually consider giving some of my allowance to poor people, even though part of me wanted to do so. My fear I was that if I allowed the "maybe you should give away your money, because maybe everyone matters evenly and you should be consequentialist" theory to fully boot up in my head, I would end up having to give away *all* my money, which seemed bad.
- Eleanore believes there is no important existential risk, and is reluctant to think through whether that might not be true, in case it ends up hijacking her whole life.
- Fred does not want to notice how much smarter he is than most of his classmates, lest he stop respecting them and treating them well.
- Gina has mixed feelings about pursuing money -- she mostly avoids it -- because she wants to remain a "caring person", and she has a feeling that becoming strategic about money would somehow involve giving up on that.

It seems to me that in each of these cases, the person has an arguably worthwhile goal that they might somehow lose track of (or might accidentally lose the ability to act on) if they think some *other* matter through -- arguably because of a deficiency of mental "buckets".

Moreover, "buckets errors" aren't just thingies that affect thinking in prospect -- they also get actually made in real life. It seems to me that one rather often runs into adults who decided they weren't allowed to like math after failing a quiz in 2nd grade; or who gave up on meaning for a couple years after losing their religion; or who otherwise make some sort of vital "buckets error" that distorts a good chunk of their lives. Although of course this is mostly guesswork, and it is hard to know actual causality.

How I try to avoid "buckets errors":

I basically just try to do the "obvious" thing: when I notice I'm averse to taking in "accurate" information, I ask myself what would be bad about taking in that information.[3] Usually, I get a concrete answer, like "If I noticed I could've saved all that time, I'll have to feel bad", or "if AI timelines are maybe-near, then I'd have to rethink all my plans", or what have you.

Then, I remember that I can consider each variable separately. For example, I can think about whether AI timelines are maybe-near; and if they are, I can always decide to not-rethink my plans anyhow, if that's actually better. I mentally list out all the decisions that *don't* need to be simultaneously forced by the info; and I promise myself that I can take the time to get these other decisions not-wrong, even after considering the new info.

Finally, I check to see if taking in the information is still aversive. If it is, I keep trying to disassemble the aversiveness into component lego blocks until it isn't. Once it isn't aversive, I go ahead and think it through bit by bit, like with the pizza.

This is a change from how I used to think about flinches: I used to be moralistic, and to feel disapproval when I noticed a flinch, and to assume the flinch had no positive purpose. I therefore used to try to just grit my teeth and think about the painful thing, without first "factoring" the "purposes" of the flinch, as I do now. But I think my new ritual is better, at least now that I have enough introspective skill that I can generally finish this procedure in finite time, and can still end up going forth and taking in the info a few minutes later.

(Eliezer once described what I take to be the a similar ritual for avoiding bucket errors, as follows: When deciding which apartment to rent (he said), one should first do out the math, and estimate the number of dollars each would cost, the number of minutes of commute time times the rate at which one values one's time, and so on. But at the end of the day, if the math says the wrong thing, one should do the right thing anyway.)

---

[1]: As an analogy: sometimes, while programming, I've had the experience of:

1. Writing a program I think is maybe-correct;
2. Inputting 0 as a test-case, and knowing ahead of time that the output should be, say, “7”;

3. Seeing instead that the output was “5”; and
4. Being really tempted to just add a “+2” into the program, so that this case will be right.

This edit is the wrong move, but not because of what it does to MyProgram(0) — MyProgram(0) really is right. It’s the wrong move because it maybe messes up the program’s *other* outputs.

Similarly, changing up my beliefs about how my finances should work in order to get a pizza on a day when I want one *might* help with getting the right answer today about the pizza — it isn’t clear — but it’d risk messing up other, future decisions.

The problem with rationalization and mental flinches, IMO, isn’t so much the “intended” action that the rationalization or flinch accomplishes in the moment, but the mess it leaves of the code afterward.

[2] To be a bit more nitpicky about this: the principle I go for in such cases isn’t actually “after thinking it through, do the best thing”. It’s more like “after thinking it through, do the thing that, if reliably allowed to be the decision-criterion, will allow information to flow freely within my head”.

The idea here is that my brain is sometimes motivated to achieve certain things; and if I don’t allow that attempted achievement to occur in plain sight, I incentivize my brain to sneak around behind my back and twist up my code base in an attempt to achieve those things. So, I try not to do that.

This is one reason it seems bad to me when people try to take “maximize all human well-being, added evenly across people, without taking myself or my loved ones as special” as their goal. (Or any other fake utility function.)

[3] To describe this "asking" process more concretely: I sometimes do this as follows: I concretely visualize a 'magic button' that will cause me to take in the information. I reach toward the button, and tell my brain I'm really going to press it when I finish counting down, unless there are any objections ("3... 2... no objections, right?... 1..."). Usually I then get a bit of an answer — a brief flash of worry, or a word or image or association.

Sometimes the thing I get is already clear, like “if I actually did the forms wrong, and I notice, I’ll have to redo them”. Then all I need to do is separate it into buckets (“How about if I figure out whether I did them wrong, and then, if I don’t want to redo them, I can always just not?”).

Other times, what I get is more like quick nonverbal flash, or a feeling of aversion without knowing why. In such cases, I try to keep “feeling near” the aversion. I might for example try thinking of different guesses (“Is it that I’d have to redo the forms?… no… Is it that it’d be embarrassing?… no…”). The idea here is to see if any of the guesses “resonate” a bit, or cause the feeling of aversiveness to become temporarily a bit more vivid-feeling.

For a more detailed version of these instructions, and more thoughts on how to avoid bucket errors in general (under different terminology), you might want to check out Eugene Gendlin’s audiobook “Focusing”.

Is Caviar a Risk Factor For Being a Millionaire?

Today, my paper "Is caviar a risk factor for being a millionaire?" was published in the Christmas Edition of the BMJ (formerly the British Medical Journal). The paper is available at http://www.bmj.com/content/355/bmj.i6536 but it is unfortunately behind a paywall. I am hoping to upload an open access version to a preprint server but this needs to be confirmed with the journal first.

In this paper, I argue that the term "risk factor" is ambiguous, and that this ambiguity causes pervasive methodological confusion in the epidemiological literature. I argue that many epidemiological papers essentially use an audio recorder to determine whether a tree falling in the forest makes a sound, without being clear about which definition of "sound" they are considering.

Even worse, I argue that epidemiologists often try to avoid claiming that their results say anything about causality, by hiding behind "prediction models". When they do this. they often still control extensively for "confounding", a term which only has a meaning in causal models. I argue that this is analogous to stating that you are interested in whether trees falling in the forest causes any human to perceive the qualia of hearing, and then spending your methods section discussing whether the audio recorder was working properly.

Due to space constraints and other considerations, I am unable to state these analogies explicitly in the paper, but it does include a call for a taboo on the word risk factor, and a reference to Rationality: AI to Zombies. To my knowledge, this is the first reference to the book in the medical literature.

I will give a short talk about this paper at the Less Wrong meetup at the MIRI/CFAR office in Berkeley at 6:30pm tonight.

(I apologize for this short, rushed announcement, I was planning to post a full writeup but I was not expecting this paper to be published for another week)

Further discussion of CFAR’s focus on AI safety, and the good things folks wanted from “cause neutrality”

Follow-up to:

- CFAR's new focus, and AI safety
- CFAR's new mission statement (link post; links to our website).

In the days since we published our previous post, a number of people have come up to me and expressed concerns about our new mission. Several of these had the form “I, too, think that AI safety is incredibly important — and that is why I think CFAR should remain cause-neutral, so it can bring in more varied participants who might be made wary by an explicit focus on AI.”

I would here like to reply to these people and others, and to clarify what is and isn’t entailed by our new focus on AI safety.

First: Where are CFAR’s activities affected by the cause(s) it chooses to prioritize?

The question of which causes CFAR aims to help (via its rationality training) plugs into our day-to-day activities in at least 4 ways:

1) **It affects which people we target.** If AI safety is our aim, we must then backchain from “Who is likely both to impact AI safety better if they have more rationality skills, and also to be able to train rationality skills with us?” to who to target with specialized workshops.

2) **It affects which rationality skills we prioritize**. AI safety work benefits from the ability to reason about abstract, philosophically confusing issues (notably: AI); which presumably benefits from various rationality skills. Competitive marathon running probably also benefits from certain rationality skills; but they are probably different ones. Designing an “art of rationality” that can support work on AI safety is different from designing an “art of rationality” for some other cause. (Although see point C, below.)

3) **It affects what metrics or feedback systems we make interim use of, and how we evaluate our work.** If “AI safety via rationality training” is the mission, then “person X produced work A that looks existential risk-reducing on our best guess, and X says they would’ve been less able to do A without us” is the obvious proxy measure of whether we’re having impact. If we have this measure, we can use our measurements of it to steer.

4) **It affects explicit curriculum at AI-related or EA-related events.** E.g., it affects whether we're allowed to run events at which participants double crux about AI safety, and whether we're allowed to present arguments from Bostrom's Superintelligence without also presenting a commensurate amount of analysis of global poverty interventions.

In addition to the above four effects, it has traditionally also affected: 5) what causes/opinions CFAR staff feel free to talk about when speaking informally to participants at workshops or otherwise representing CFAR. (We used to try not to bring up such subjects.)

One thing to notice, here, is that CFAR's mission doesn't just affect our external face; it affects the details of our day-to-day activities. (Or at minimum, it *should* affect these.) It is therefore very important that our mission be: (a) actually important; (b) simple, intelligible, and usable by our staff on a day-to-day basis; (c) corresponding to a detailed (and, ideally, accurate) model in the heads of at least a few CFARians doing strategy (or, better, in all CFARians), so that the details of what we're doing can in fact "cut through" to reducing existential risk.

So, okay, we just looked concretely at how CFAR's mission (and, in particular, its prioritization of AI safety) can affect its day-to-day choices.

It's natural next to ask what upsides people were hoping for from a (previous or imagined) "cause neutral" CFAR, and to discuss which of those upsides we can access still, and which we can't. I'll start with the ones we can do.

Some components that people may be hoping for from "cause neutral", that we can do, and that we intend to do:

A. For students of all intellectual vantage points, we can make a serious effort to be "epistemically trustworthy relative to their starting point".

By this I mean:

- We can be careful to include all information that *they, from their vantage point, would want to know* -- even if on our judgment, some of the information is misleading or irrelevant, or might pull them to the "wrong" conclusions.

- Similarly, we can attempt to expose people to skilled thinkers they would want to talk with, regardless of those thinkers' viewpoints; and we can be careful to allow *their own thoughts, values, and arguments to develop*, regardless of which "side" this may lead to them supporting.

- More generally, we can and should attempt to cooperate with each student's extrapolated volition, and to treat the student as *they* (from their initial epistemic vantage point; and with their initial values) would wish to be treated. Which is to say that we should not do anything that would work less well if the algorithm behind it were known, and that we should attempt to run such workshops (and to have such conversations, and so on) as would cause good people of varied initial views to stably on reflection want to participate in them.

In asserting this commitment, I do not mean to assert that others should believe this of us; only that we will aim to do it. You are welcome to stare skeptically at us about potential biases; we will not take offense; it is probably prudent. Also, our execution will doubtless have flaws; still, we'll appreciate it if people point such flaws out to us.

B. We can deal forthrightly and honorably with potential allies who have different views about what is important.

That is: we can be clear and explicit about the values and beliefs we are basing CFAR's actions on, and we can attempt to negotiate clearly and explicitly with individuals who are interested in supporting particular initiatives, but who disagree with us about other parts of our priorities.[1]

C. We can create new "art of rationality" content at least partly via broad-based exploratory play — and thus reduce the odds that our "art of rationality" ends up in a local optimum around one specific application.
That is: we can follow Feynman's lead and notice and chase "spinning plates". We can bring in new material by bringing in folks with very different skillsets, and seeing what happens to our art and theirs when we attempt to translate things into one another's languages. We can play; and we can nourish an applied rationality community that can also play.

Some components that people may be hoping for from "cause neutral", that we can't or won't do:

i. Appear to have no viewpoints, in hopes of attracting people who don't trust those with our viewpoints.

We can't do this one. Both CFAR as an entity and individual CFAR staff, do in fact have viewpoints; there is no high-integrity way to mask that fact. Also, "integrity" isn't a window-dressing that one pastes onto a thing, or a nicety that one can compromise for the sake of results; "integrity" is a word for the basic agreements that make it possible for groups of people to work together while stably trusting one another. Integrity is thus structurally necessary if we are to get anything done at all.

All we can do is do our best to *be* trustworthy in our dealings with varied people, and assume that image will eventually track substance. (And if image doesn't, we can look harder at our substance, see if we may still be subtly acting in bad faith, and try again. Integrity happens from the inside out.)

ii. Leave our views or plans stalled or vague, in cases where having a particular viewpoint would expose us to possibly being wrong (or to possibly alienating those who disagree).

Again, we can’t do this one; organizations need a clear plan for their actions to have any chance at either: i) working; or ii) banging into data and allowing one to notice that the plan was wrong. Flinching from clearly and visibly held views is the mother of wasting time. (Retaining a willingness to say “Oops!” and change course is, however, key.)

iii. Emphasize all rationality use cases evenly. Cause all people to be evenly targeted by CFAR workshops.

We can’t do this one either; we are too small to pursue all opportunities without horrible dilution and failure to capitalize on the most useful opportunities.

We are presently targeting all workshops at either: (a) folks who are more likely than usual to directly impact existential risk; or (b) folks who will add to a robust rationality community, and/or (c) allow us to learn more about the art (e.g., by having a different mix of personalities, skills, or backgrounds than most folks here).

Coming soon:

- CFAR’s history around our mission: How did we come to change?

---

[1] In my opinion, I goofed this up historically in several instances, most notably with respect to Val and Julia, who joined CFAR in 2012 with the intention to create a cause-neutral rationality organization. Most integrity-gaps are caused by lack of planning rather than strategic deviousness; someone tells their friend they’ll have a project done by Tuesday and then just… doesn’t. My mistakes here seem to me to be mostly of this form. In any case, I expect the task to be much easier, and for me and CFAR to do better, now that we have a simpler and clearer mission.

Be secretly wrong

> "I feel like I'm not the sort of person who's allowed to have opinions about the important issues like AI risk."
>
> "What's the bad thing that might happen if you expressed your opinion?"
>
> "It would be wrong in some way I hadn't foreseen, and people would think less of me."
>
> "Do you think less of other people who have wrong opinions?"
>
> "Not if they change their minds when confronted with the evidence."
>
> "Would you do that?"
>
> "Yeah."
>
> "Do you think other people think less of those who do that?"
>
> "No."
>
> "Well, if it's alright for other people to make mistakes, what makes YOU so special?"

A lot of my otherwise very smart and thoughtful friends seem to have a mental block around thinking on certain topics, because they're the sort of topics Important People have Important Opinions around. There seem to be two very different reasons for this sort of block:

1. Being wrong feels bad.
2. They might lose the respect of others.

Be wrong

If you don't have an opinion, you can hold onto the fantasy that someday, once you figure the thing out, you'll end up having a right opinion. But if you put yourself out there with an opinion that's unmistakably your own, you don't have that excuse anymore.

This is related to the desire to pass tests. The smart kids go through school and are taught - explicitly or tacitly - that as long as they get good grades they're doing OK, and if they try at all they can get good grades. So when they bump up against a problem that might actually be hard, there's a strong impulse to look away, to redirect to something else. So they do.

You have to understand that this system is not real, it's just a game. In real life you have to be straight-up wrong sometimes. So you may as well get it over with.

If you expect to be wrong when you guess, then you're already wrong, and paying the price for it. As Eugene Gendlin said:

> What is true is already so. Owning up to it doesn't make it worse. Not being open about it doesn't make it go away. And because it's true, it is what is there to be interacted with. Anything untrue isn't there to be lived. People can stand what is true, for they are already enduring it.

What you would be mistaken about, you're already mistaken about. Owning up to it doesn't make you any more mistaken. Not being open about it doesn't make it go away.

"You're already "wrong" in the sense that your anticipations aren't perfectly aligned with reality. You just haven't put yourself in a situation where you've openly tried to guess the teacher's password. But if you want more power over the world, you need to focus your uncertainty - and this only reliably makes you righter if you repeatedly test your beliefs. Which means sometimes being wrong, and noticing. (And then, of course, changing your mind.)

Being wrong is how you learn - by testing hypotheses.

In secret

Getting used to being wrong - forming the boldest hypotheses your current beliefs can truly justify so that you can correct your model based on the data - is painful and I don't have a good solution to getting over it except to tough it out. But there's a part of the problem we can separate out, which is - the pain of being wrong publicly.

When I attended a Toastmasters club, one of the things I liked a lot about giving speeches there was that the stakes were low in terms of the content. If I were giving a presentation at work, I had to worry about my generic presentation skills, but also whether the way I was presenting it was a good match for my audience, and also whether the idea I was pitching was a good strategic move for the company or my career, and also whether the information I was presenting was accurate. At Toastmasters, all the content-related stakes were gone. No one with the power to promote or fire me was present. Everyone was on my side, and the group was all about helping each other get better. So all I had to think about was the form of my speech.

Once I'd learned some general presentations at Toastmasters, it became easier to give talks where I did care about the content and there were real-world consequences to the quality of the talk. I'd gotten practice on the form of public speaking separately - so now I could relax about that, and just focus on getting the content right.

Similarly, expressing opinions publicly can be stressful because of the work of generating likely hypotheses, and revealing to yourself that you are farther behind in understanding things than you thought - but also because of the perceived social consequences of sounding stupid. You can at least isolate the last factor, by starting out thinking things through in secret. This works by separating epistemic uncertainty from social confidence. (This is closely related to the dichotomy between social and objective respect.)

Of course, as soon as you can stand to do this in public, that's better - you'll learn faster, you'll get help. But if you're not there yet, this is a step along the way. If the choice is between having private opinions and having none, have private opinions. (Also related: If we can't lie to others, we will lie to ourselves.)

Read and discuss a book on a topic you want to have opinions about, with one trusted friend. Start a secret blog - or just take notes. Practice having opinions at all, that you can be wrong about, before you worry about being accountable for your opinions. One step at a time.

Before you're publicly right, consider being secretly wrong. Better to be secretly wrong, than secretly not even wrong.

(Cross-posted at my personal blog.)

How does personality vary across US cities?

In 2007, psychology researchers Michal Kosinski and David Stillwell released a personality testing app on Facebook app called myPersonality. The app ended up being used by 4 million Facebook users, most of whom consented to their personality question answers and some information from their Facebook profiles to be used for research purposes.

The very large sample size and matching data from Facebook profiles make it possible to investigate many questions about personality differences that were previously inaccessible. Koskinski and Stillwell have used it in a number of interesting publications, which I highly recommend (e.g. [1], [2] [3]).

In this post, I focus on what the dataset tells us about **how big five personality traits vary by geographic region in the United States**.

The Five Factor Model of Personality

The *Five Factor Model (FFM)* or *Big Five personality trait model* is currently the dominant paradigm in personality research. The model is founded on the lexical hypothesis:

> The lexical hypothesis is generally defined by two postulates. The first states that those personality characteristics that are most important in peoples' lives will eventually become a part of their language. The second follows from the first, stating that more important personality characteristics are more likely to be encoded into language as a single word.

When people are asked questions about whether various adjectives describe them (or describe someone who they know), their answers are pairwise correlated with one another. Applying factor analysis to the responses yields a small number of underlying factors that explain a large fraction of the variance common to the answers.

Empirically, it's been found that a model with 5 factors often fits the data well (though some researchers claim that one gets 6 or 7 factors if one uses a question battery that fully exhausts descriptive adjectives, see e.g. the HEXACO model of personality and The Big Seven Model of Personality and Its Relevance to Personality Pathology for more information).

The five factors referred to as the "Big Five" are labelled extraversion, neuroticism, agreeableness, conscientiousness and openness. I will describe these more below.

It's likely that the Big Five personality model falls far short of carving reality at its joints, and I'm in broad agreement with the stance that Jack Block expresses in A contrarian view of the five-factor approach to personality description. Nevertheless, the five factors in the model satisfy some desirable criteria, such as

- **Longitudinal stability**, with correlations between self-report and self-report 10 years later being ~0.7.

- **Self-other agreement**, with correlation of ~0.4 between self-report and a friend's perceptions, and ~0.5 between self-report and spouse perceptions).
- **External validity**, with correlations found between self-reported traits and objective behaviors.
- **Heritability**, with twin studies yielding estimates that 40%-60% of the variance in the underlying traits is explained by genetics.
- **Cross-cultural validity**, c.f. The Geographic Distribution of Big Five Personality Traits: Patterns and Profiles of Human Self-Description Across 56 Nations.

and much of the data that's available uses Big 5 personality questionnaires, so it's often what we have to work with.

Data, methodology and high level results

There were ~680k Americans who both answered 20+ questions on a Big Five Personality Test, and who made their hometown available to researchers. After excluding hometowns with <= 30 users, about ~3500 hometowns were represented. Questions were answered on a scale from 1 (strongly disagree) to 5 (strongly agree).

I estimated personality trait averages for each city using Bayesian hierarchical modeling in order to account for regression to the mean when sample sizes are small. This results in relatively large cities being more prominently represented at the extremes of the estimates, on account of the larger sample sizes making it possible to have greater confidence in city averages deviating substantially from the mean. A CSV file with all estimates of city averages is available on Dropbox.

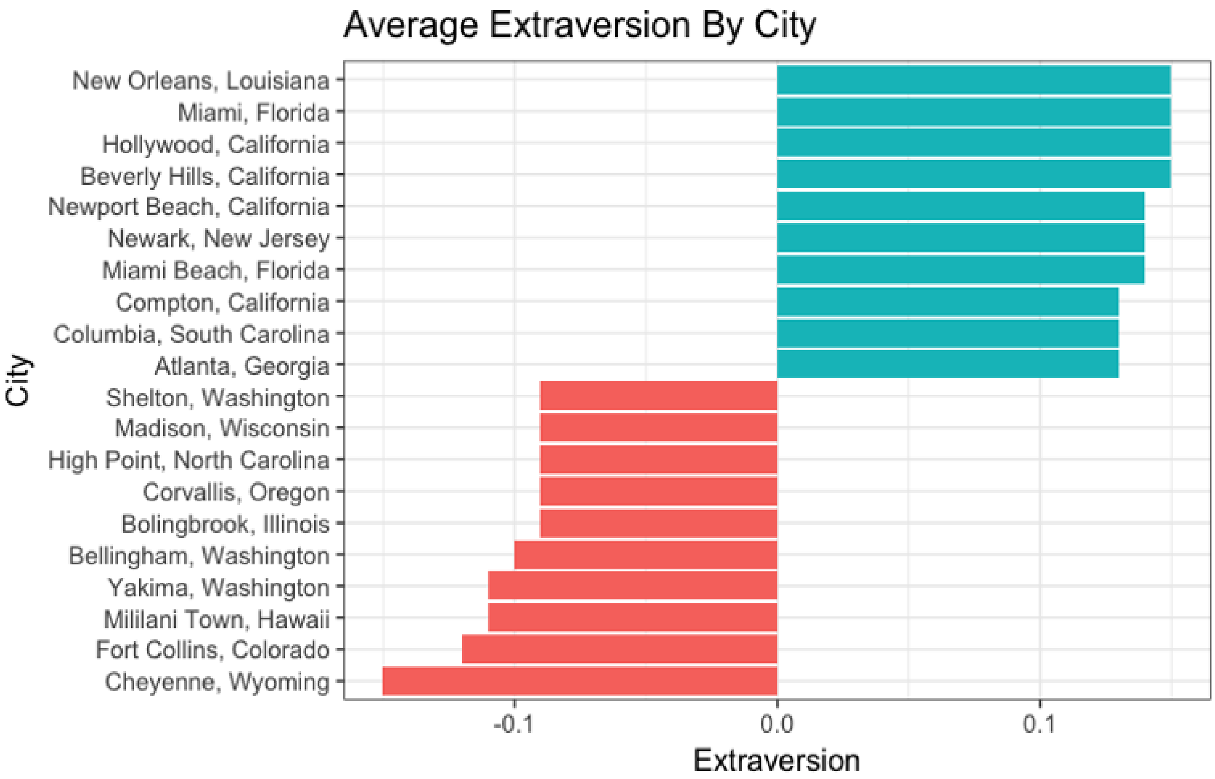
The units in the graphs below are *standard deviations away from the mean of the entire sample*. Roughly speaking, average self-reported personality by city varies from -0.2 to 0.2 standard deviations from the mean. However, this likely understates the magnitudes of differences in underlying traits across cities, owing to people anchoring on the people who they know when answering the questions rather than anchoring on the national population, as described in Birds of a feather do flock together: behavior and language-based personality assessment reveal personality homophily among couples and friends:

> Friends and spouses tend to be similar in a broad range of characteristics (i.e. *homophily*), such as age, educational level, attitudes, values, and general intelligence. Surprisingly, little evidence has been found for similarity in personality—one of the most fundamental psychological constructs. We argue that the lack of evidence for personality homophily derives from the tendency of individuals to make personality judgments in relation to a salient comparison group rather than in absolute terms when responding to the self-report and peer-report questionnaires commonly used in personality research (i.e. *reference-group effect*)

Extraversion

Representative questions:

- Do not mind being the centre of attention
- Make friends easily
- Keep in the background (reversed)
- Avoid contact with others (reversed)



Party cities have high average extraversion

The appearance of New Orleans, Miami, Hollywood, Beverly Hills and Newport Beach as amongst the highest on average extraversion is consistent with the the cities' reputations as having high prevalence of partying & socialization. New Orleans and Miami are both highest average extraversion in the data, and 2 of the 3 American cities on this list flist of top 20 party cities in the world.

The Seattle Freeze

Andrew J. Ho comments that the high frequency of cities in Washington state reminds him of the Seattle Freeze:

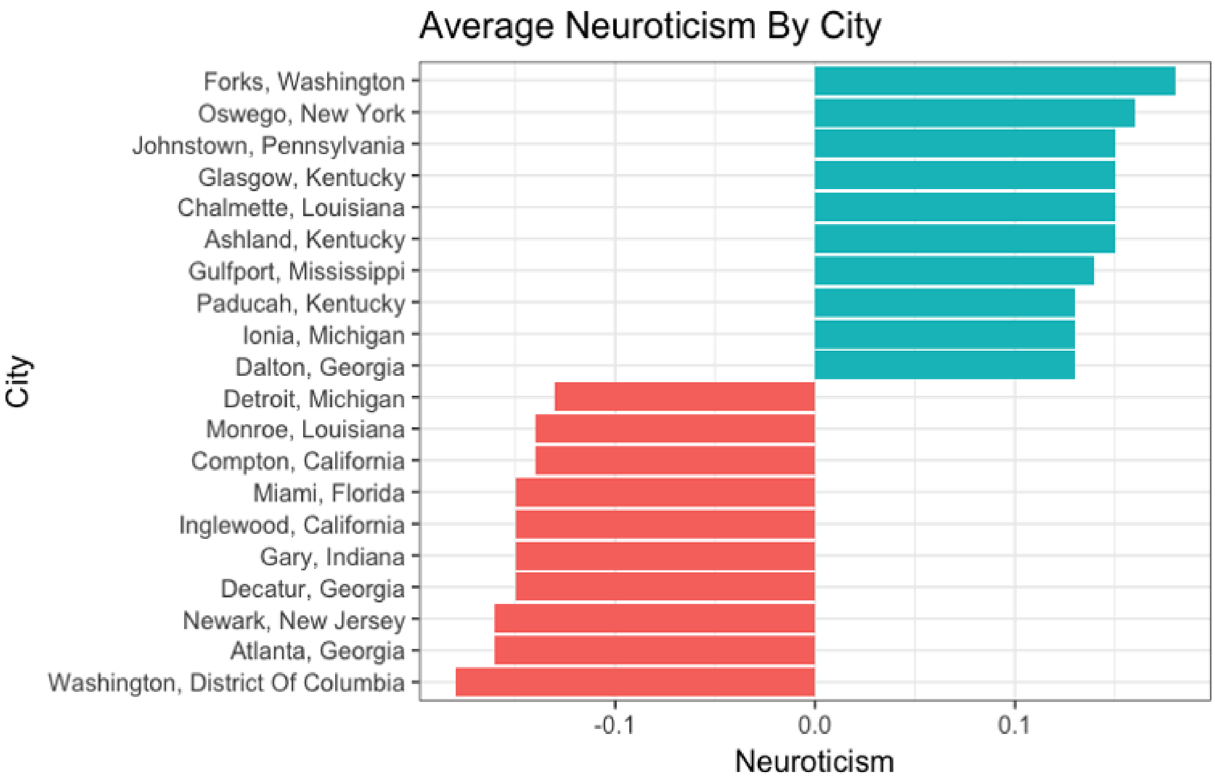
> Newcomers to the area have described Seattleites as being standoffish, cold, distant, and not trusting.[3] While in settings such as bars and parties, people from Seattle tend to mainly interact with their particular clique.[4] One author described the aversion to strangers as: "people are very polite but not particularly friendly." [5] In 2008 a peer-reviewed study published in Perspectives on Psychological Science found that among all states, Washington residents ranked 48th in the personality trait extroverted.

Neuroticism

Representative questions:

- Often feel blue
- Get stressed out easily
- Feel comfortable with myself (reversed)

- Am not easily bothered by things (reversed)

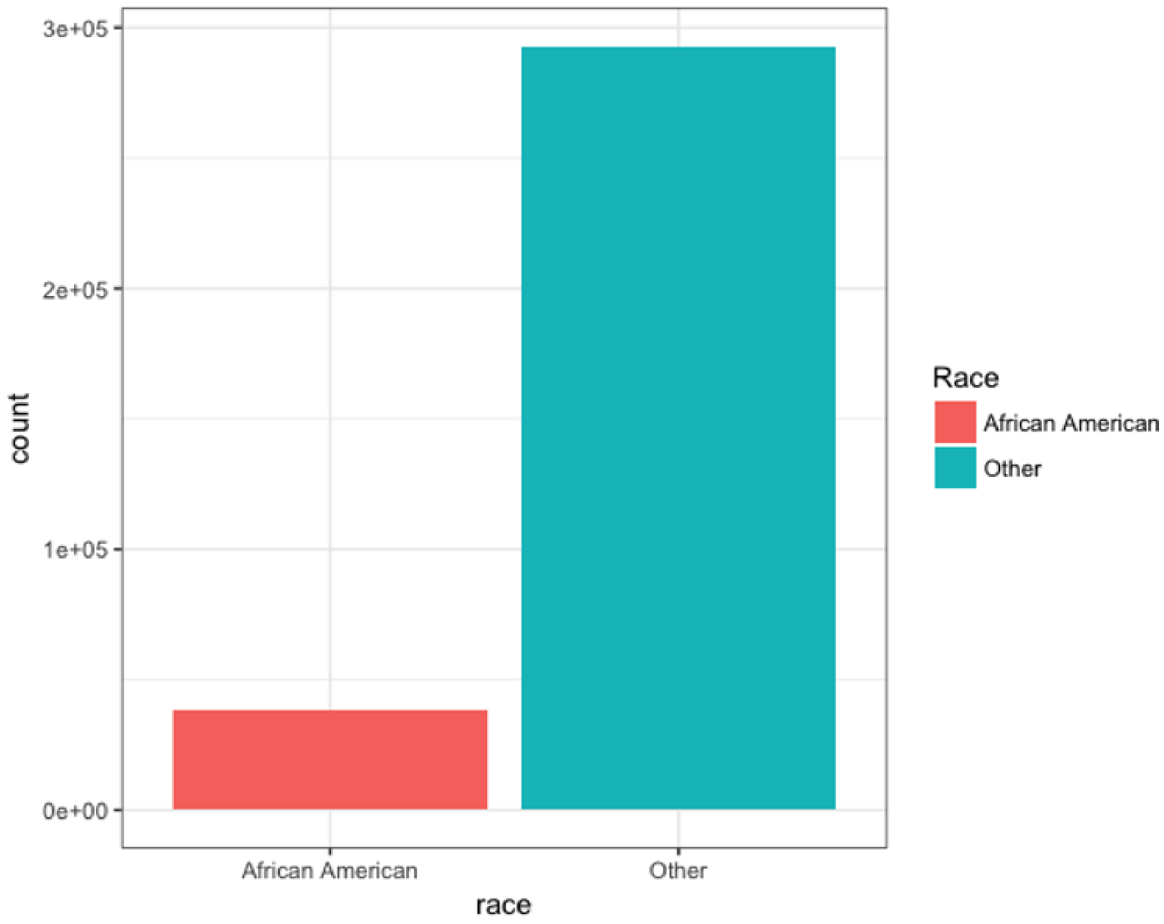


Ethnicity as an underlying factor

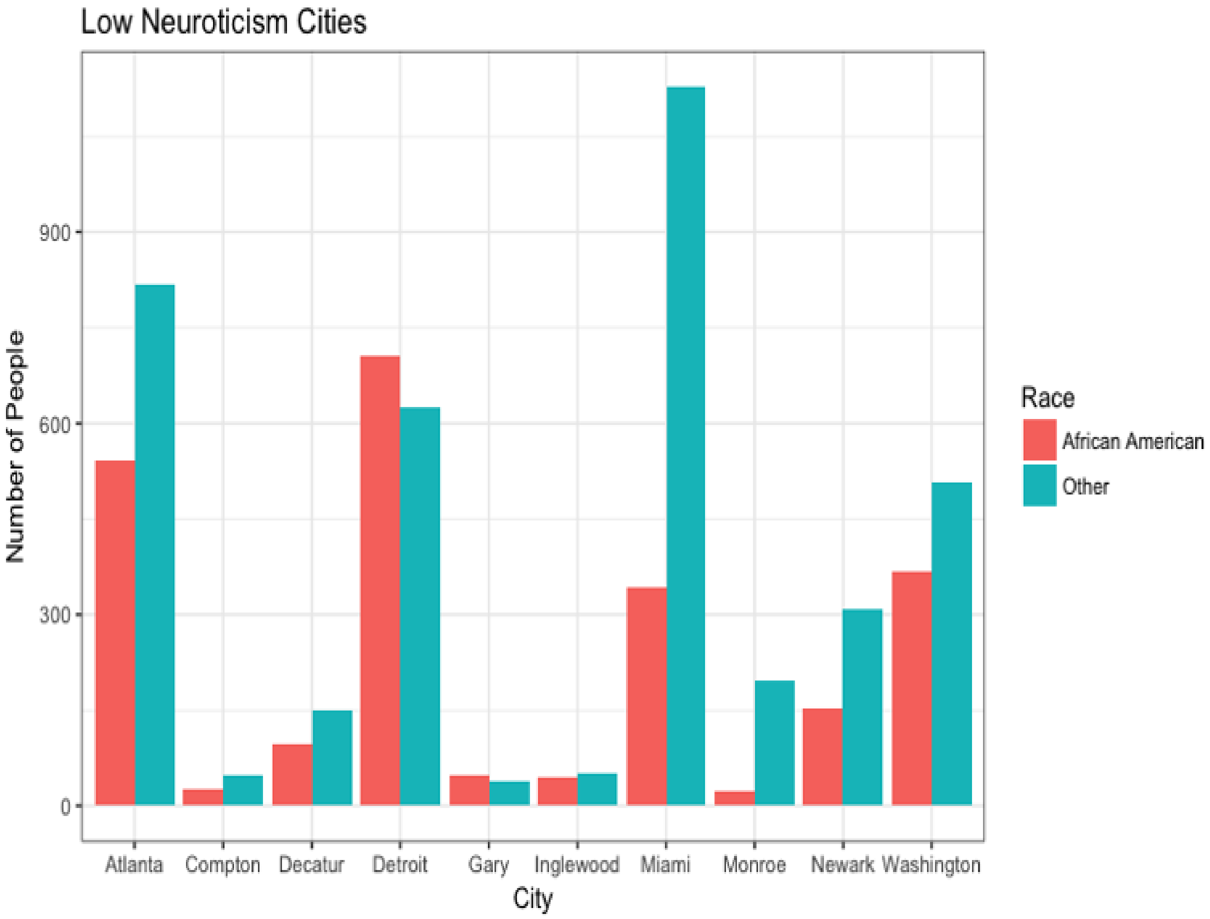
Washington DC and Atlanta stand out as having unusually large African American populations, constituting roughly 50% of the population. From Wikipedia:

> Atlanta has long been known as a center of black wealth, political power and culture; a cradle of the Civil Rights Movement[1] and home to Dr. Martin Luther King, Jr. It has often been called a "black mecca".

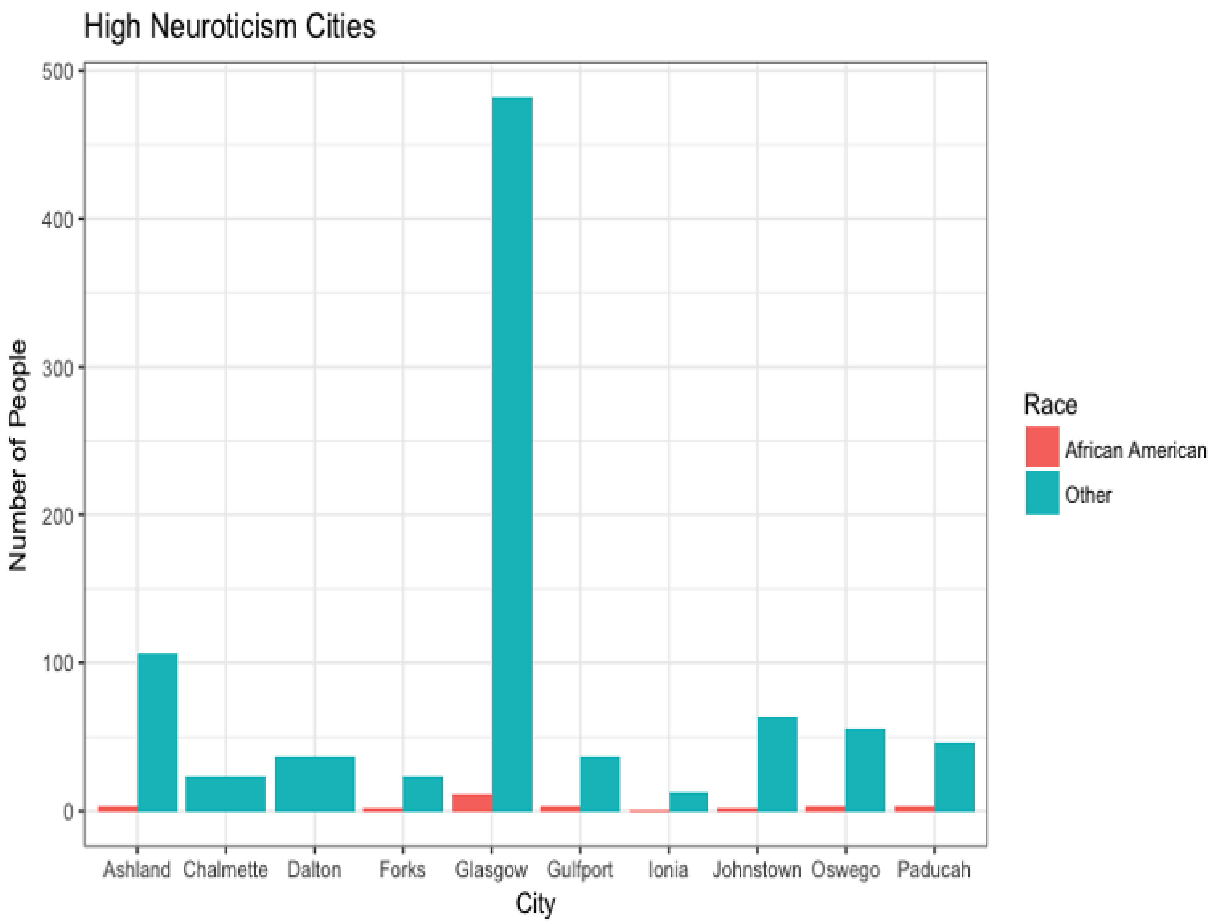
The researchers behind the myPersonality app labelled the Facebook profile photos of a subset of the users by their race, so we can stratify by race. The numbers of people for whom we have labelled photos are given below, by race.



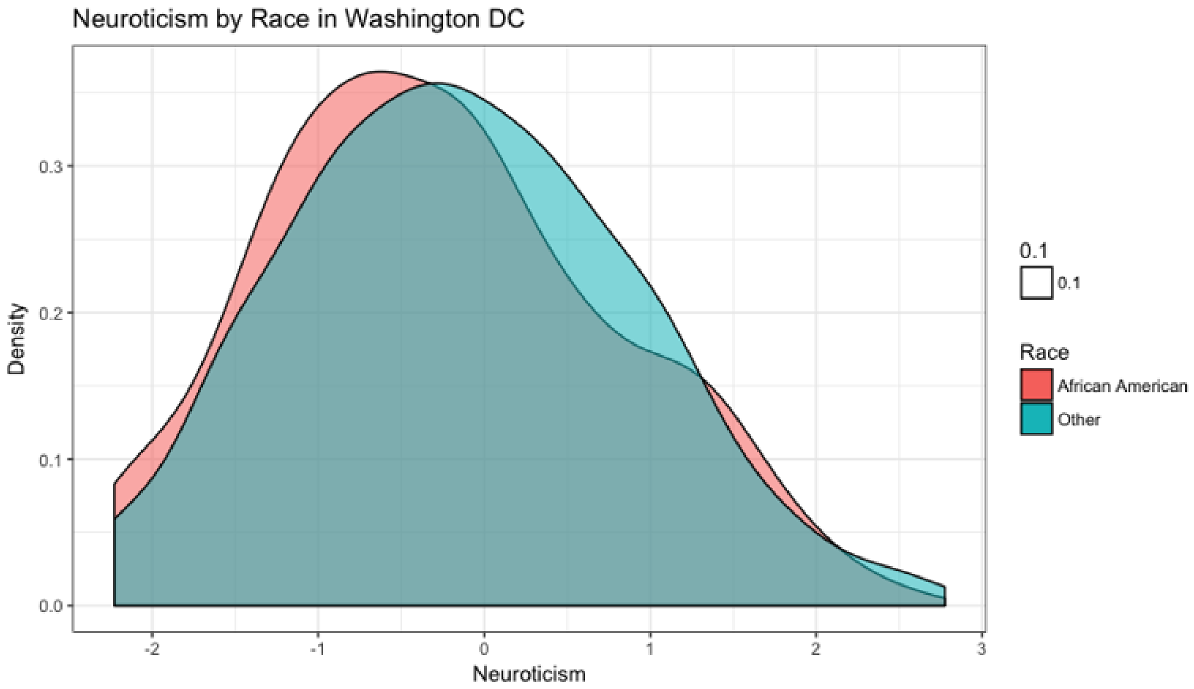
The people in cities with **low average neuroticism** are heavily disproportionately African-American:



whereas for **high neuroticism cities** the reverse is true:



This is not a coincidence. In fact, for the sample as a whole, African Americans' self-reported neuroticism is a full 0.2 standard deviations lower than the rest of the population. This remains true even if we restrict attention to a particular city, like Washington DC:



The finding of African Americans being relatively low on neuroticism is consistent with the literature on national differences in personality. The figure below is from The Geographic Distribution of Big Five Personality Traits: Patterns and Profiles of Human Self-Description Across 56 Nations. It depicts estimates of average neuroticism by continent, showing that Africans are as a group noticeably lower in neuroticism than people from other continents.

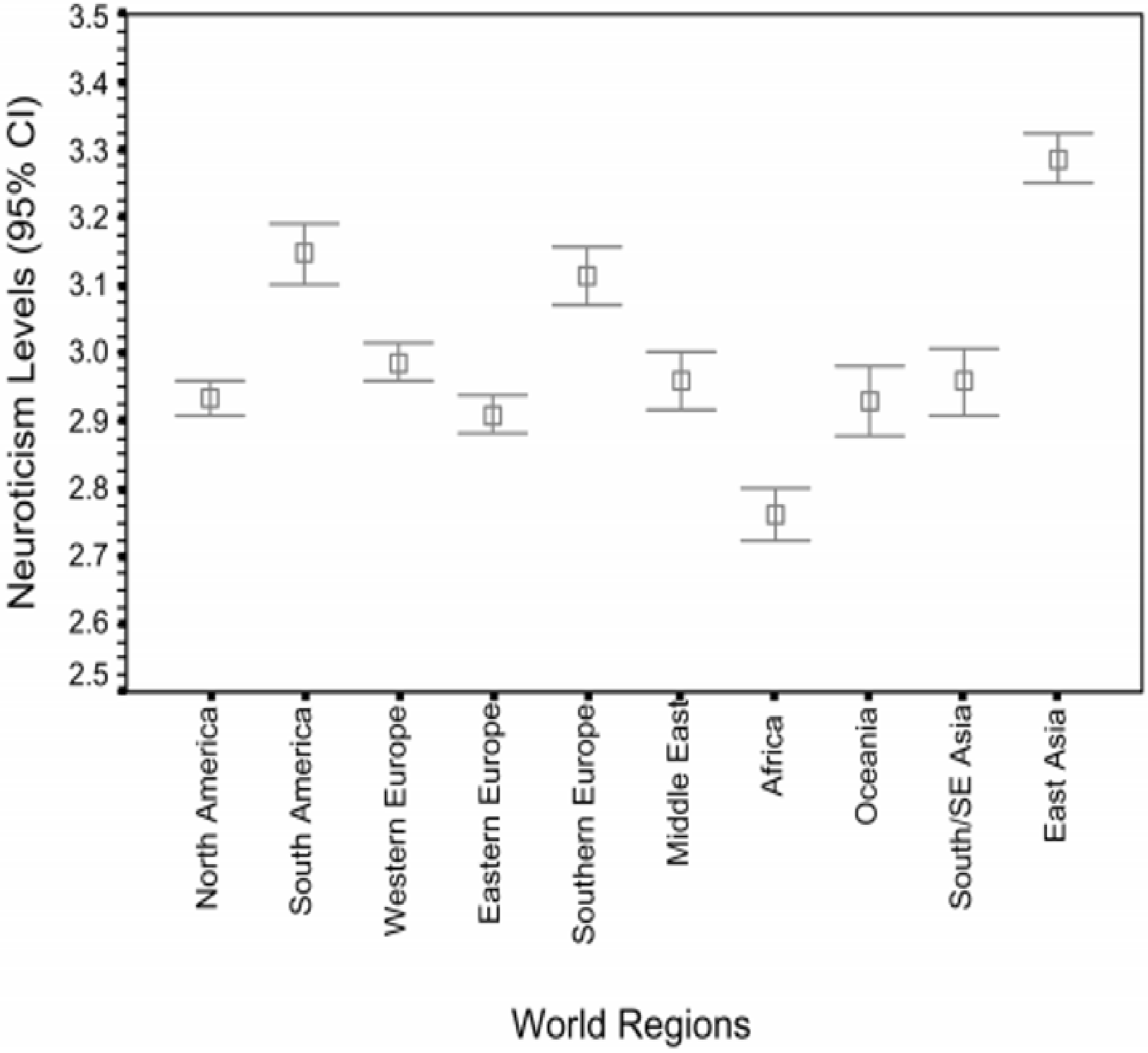
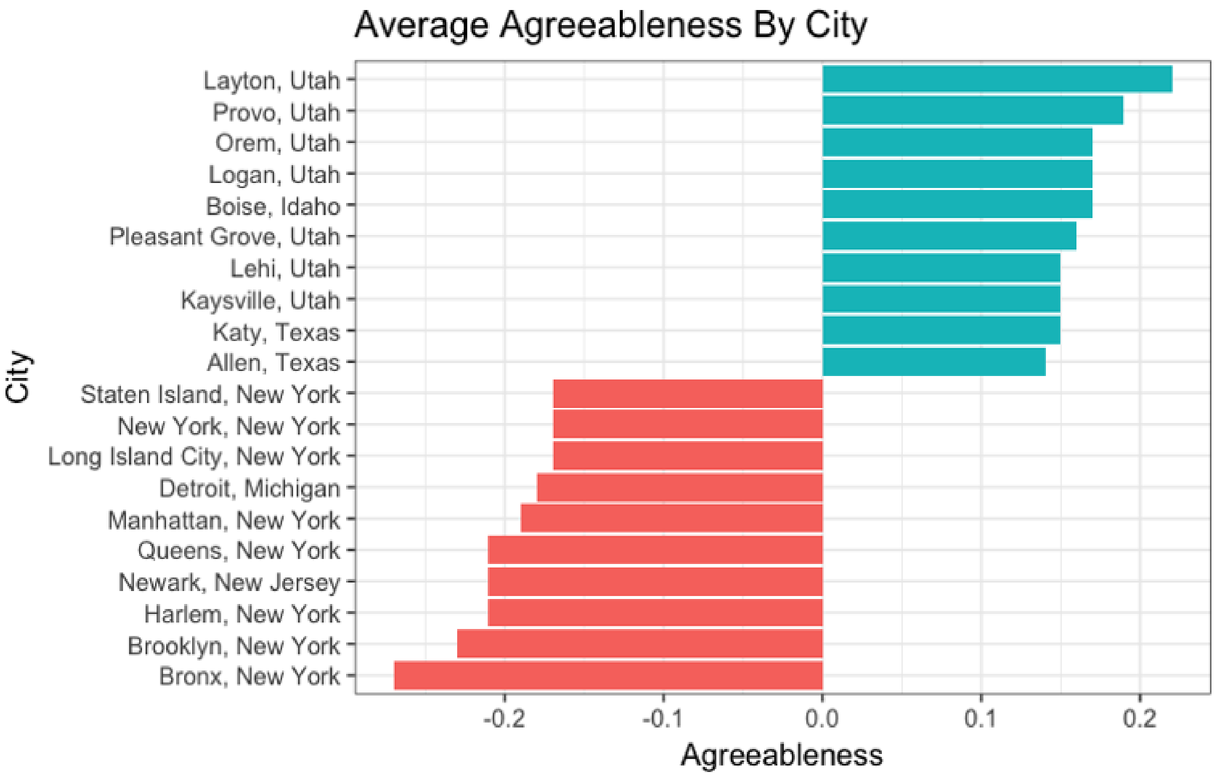


Figure 4: Neuroticism Levels (With 95% Confidence Interval [CI] Error Bars) Across the 10 World Regions of the International Sexuality Description Project

Agreeableness

Representative questions:

- Believe that others have good intentions
- Am easy to satisfy
- Hold a grudge(reversed)
- Cut others to pieces (reversed)



Agreeableness and Mormonism?

Seven of the 10 cities with highest average agreeableness are in Utah. This corresponds to Utah residents being almost 60% Mormon: as a group, Mormons have exceptionally high average agreeableness. One can do an analysis similar to the one that I did with race and neuroticism. I'll return to this later in the context of a more systematic discussion of agreeableness and religion.

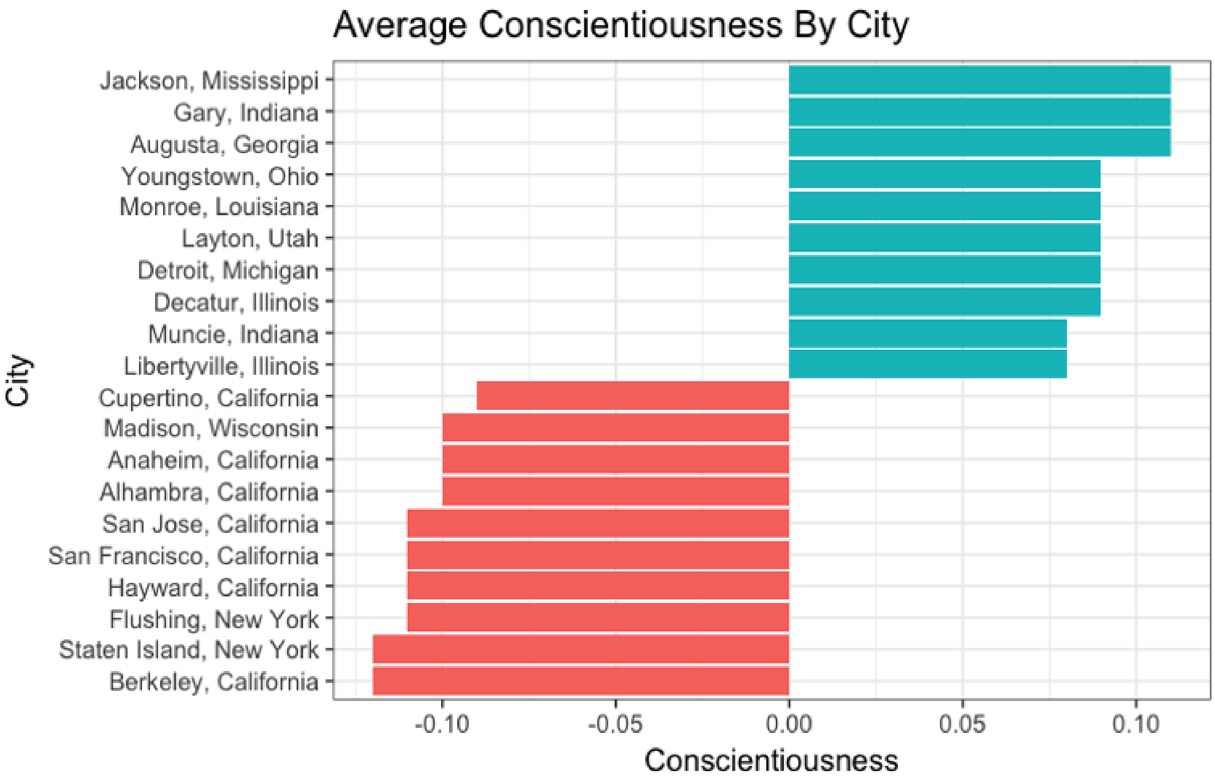
New Yorkers really are unusually disagreeable

The fact that 8 of the 10 cities listed correspond to some burrough of New York City is in accordance with stereotypes around New Yorkers being unfriendly / mean / aggressive / rude (c.f. New York City Ranked Sixth Most Unfriendly City in the World, Survey Finds).

Conscientiousness

Representative questions:

- Complete tasks successfully
- Am always prepared
- Need a push to get started (reversed)
- Shirk my duties (reversed)



Low conscientiousness in the Bay Area

It's striking that each of Berkeley, San Francisco, San Jose, Hayward and Cuptertino make the list of 10 cities with lowest average conscientiousness, while simultaneously all being in the Bay Area.

Connection with the in person rationalist community?

The finding that Bay Area residents skew toward unusually low conscientiousness should be of especially strong interest to the rationalist community in light of the fact that the Bay Area has become the central hub of community activity.

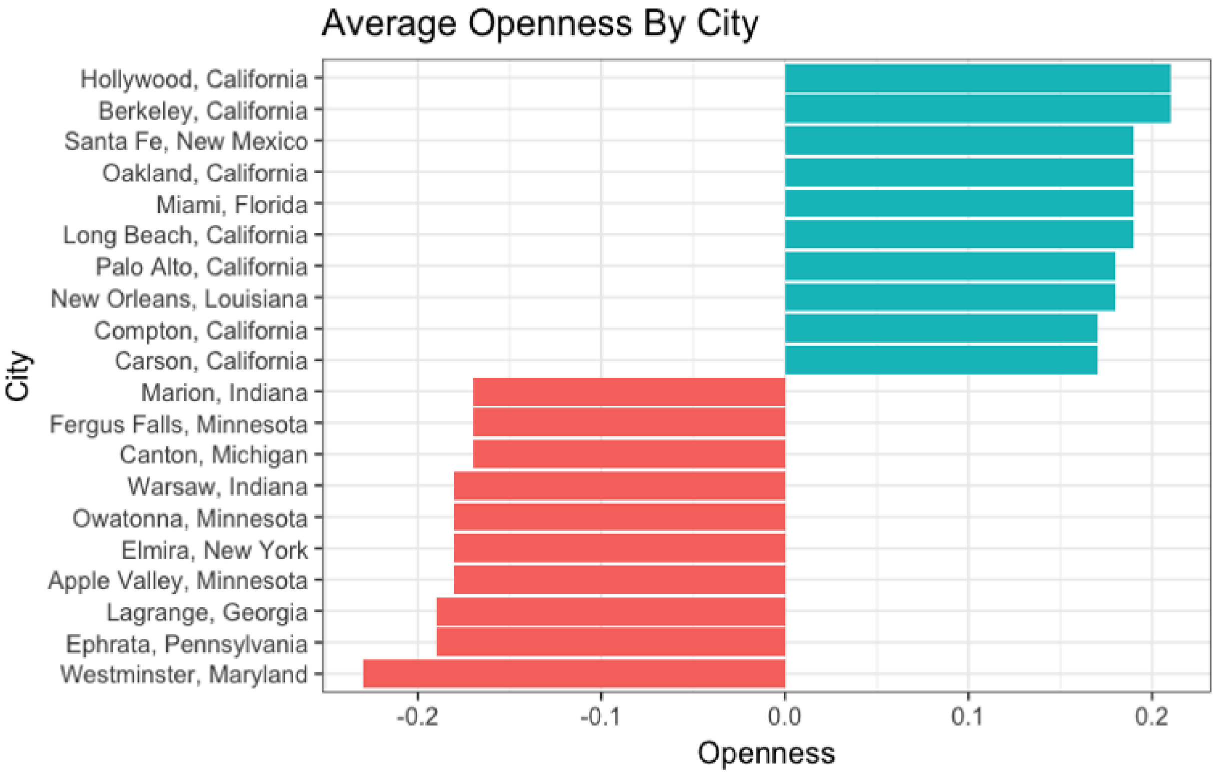
Slightly shifting the subject, in the 2016 Less Wrong Diaspora Survey, those respondents who reported to having involvement with the in-person community reported to being clincially diagnosed with ADHD with frequency ~20%, **roughly 2x more frequently** than those who reported to having no involvement with the in person community. Low conscientiousness is known to associate with ADHD, with people who have been diagnosed with ADHD scoring an average of 1 standard deviation below the population mean. In light of these things, it seems possible that there's some connection between high rates of clinical diagnosis of ADHD amongst people being involved with the in person community, and Bay Area residents being unusually low conscientiousness.

As with low extraversion, I'd welcome any ideas on what differentiates the cities with high average conscientiousness from others...

Openness

Representative questions.

- Have a vivid imagination
- Enjoy wild flights of fantasy
- Avoid philosophical discussions (reversed)
- Do not like poetry (reversed)



Artsy cities and openness

Openness is associated with artistic interests. Hollywood is the center of cinema in the United States. Sante Fe and New Orleans are considered two of the ten most artistic cities in America. So their appearance near the top of the list is in consonance with expectations.

Political Liberalism and Openness

Openness is known to be strongly predictive of liberal political affiliation (c.f. The Secret Lives of Liberals and Conservatives: Personality Profiles, Interaction Styles, and the Things They Leave Behind). So the appearance of many coastal California cities is also in consonance with expectations.

To Be Continued...

There's much more to say about personality and demographics, and I plan on writing more along these lines.

CFAR’s new focus, and AI Safety

A bit about our last few months:

- We’ve been working on getting a simple clear mission and an organization that actually works. We think of our goal as analogous to the transition that the old Singularity Institute underwent under Lukeprog (during which chaos was replaced by a simple, intelligible structure that made it easier to turn effort into forward motion).
- As part of that, we’ll need to find a way to be intelligible.
- This is the first of several blog posts aimed at causing our new form to be visible from outside. (If you're in the Bay Area, you can also come meet us at tonight's open house.) (We'll be talking more about the causes of this mission-change; the extent to which it is in fact a change, etc. in an upcoming post.)

Here's a short explanation of our new mission:

- We care a lot about AI Safety efforts in particular, and about otherwise increasing the odds that humanity reaches the stars.

- Also, we[1] believe such efforts are bottlenecked more by our collective epistemology, than by the number of people who verbally endorse or act on "AI Safety", or any other "spreadable viewpoint" disconnected from its derivation.

- Our aim is therefore to find ways of improving both individual thinking skill, and the modes of thinking and social fabric that allow people to think *together*. And to do this among the relatively small sets of people tackling existential risk.

To elaborate a little:

Existential wins and AI safety

By an “existential win”, we mean humanity creates a stable, positive future. We care a heck of a lot about this one.

Our working model here accords roughly with the model in Nick Bostrom’s book Superintelligence. In particular, we believe that if general artificial intelligence is at some point invented, it will be an enormously big deal.

(Lately, AI Safety is being discussed by everyone from The Economist to Newsweek to Obama to an open letter from eight thousand. But we’ve been thinking on this, and backchaining partly from it, since before that.)

Who we’re focusing on, why

Our preliminary investigations agree with The Onion’s; despite some looking, we have found no ultra-competent group of people behind the scenes who have fully got things covered.

What we have found are:

- AI and machine learning graduate students, researchers, project-managers, etc. who care; who can think; and who are interested in thinking better;
- Students and others affiliated with the “Effective Altruism” movement, who are looking to direct their careers in ways that can do the most good;
- Rationality geeks, who are interested in seriously working to understand how the heck thinking works when it works, and how to make it work even in domains as confusing as AI safety.

These folks, we suspect, are the ones who can give humanity the most boost in its survival-odds per dollar of CFAR’s present efforts (which is a statement partly about us, but so it goes). We’ve been focusing on them.

(For the sake of everyone. Would you rather: (a) have bad rationality skills yourself; or (b) be killed by a scientist or policy-maker who also had bad rationality skills?)

Brier-boosting, not Signal-boosting

Everyone thinks they’re right. We do, too. So we have some temptation to take our own favorite current models of AI Safety strategy and to try to get everyone else to shut up about their models and believe ours instead.

This understandably popular activity is often called “signal boosting”, “raising awareness”, or doing “outreach”.

At CFAR, though, we force ourselves not to do “signal boosting” in this way. Our strategy is to spread general-purpose thinking skills, not our current opinions. It is important that we get the truth-seeking skills themselves to snowball across relevant players, because ultimately, creating a safe AI (or otherwise securing an existential win) is a research problem. Nobody, today, has copyable opinions that will get us there.

We like to call this “Brier boosting”, because a “Brier score” is a measure of predictive accuracy.

We agree with physicists that "Moving the world" is probably not literally impossible. Still, we suspect it requires a rather crazy amount of focus.

(song by CFAR alum Ray Arnold)

---

[1] By "We believe X", we do not mean to assert that every CFAR staff member individually believes X. (Similarly for "We care about Y). We mean rather that CFAR as an organization is planning/acting as though X is true. (Much as if CFAR promises you a rationality T-shirt, that isn't an individual promise from each of the individuals at CFAR; it is rather a promise from the organization as such.)

If we're going to build an art of rationality, we'll need to figure out how to create an organization where people can individually believe whatever the heck they end up actually believing as they chase the evidence, while also having the organization qua organization be predictable/intelligible.

ETA:
You may also want to check out two documents we posted in the days since this post:

- Further discussion of CFAR’s focus on AI safety, and the good things folks wanted from “cause neutrality”
- CFAR's mission statement (link post, linking to our website).