Best of LessWrong: January 2017

1. Double Crux — A Strategy for Mutual Understanding
2. Project Hufflepuff
3. 80,000 Hours: EA and Highly Political Causes
4. Most empirical questions are unresolveable; The good, the bad, and the appropriately under-powered

Best of LessWrong: January 2017

1. Double Crux — A Strategy for Mutual Understanding
2. Project Hufflepuff
3. 80,000 Hours: EA and Highly Political Causes
4. Most empirical questions are unresolveable; The good, the bad, and the appropriately under-powered

Double Crux — A Strategy for Mutual Understanding

Preamble

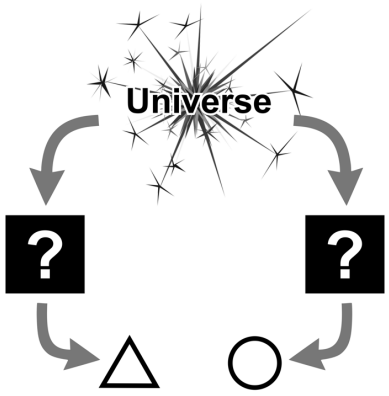
Double crux is one of CFAR's newer concepts, and one that's forced a re-examination and refactoring of a lot of our curriculum (in the same way that the introduction of TAPs and Inner Simulator did previously). It rapidly became a part of our organizational social fabric, and is one of our highest-EV threads for outreach and dissemination, so it's long overdue for a public, formal explanation.

Note that while the core concept is fairly settled, the execution remains somewhat in flux, with notable experimentation coming from Julia Galef, Kenzi Amodei, Andrew Critch, Eli Tyre, Anna Salamon, myself, and others. Because of that, this post will be less of a cake and more of a folk recipe—this is long and meandering on purpose, because the priority is to transmit the *generators* of the thing over the thing itself. Accordingly, if you think you see stuff that's wrong or missing, you're probably onto something, and we'd appreciate having them added here as commentary.

Casus belli

To a first approximation, a human can be thought of as a black box that takes in data from its environment, and outputs beliefs and behaviors (that black box isn't really "opaque" given that we *do* have access to a lot of what's going on inside of it, but our understanding of our own cognition seems uncontroversially incomplete).

When two humans disagree—when their black boxes output different answers, as below—there are often a handful of unproductive things that can occur.



The most obvious (and tiresome) is that they'll simply repeatedly bash those outputs together without making any progress (think most disagreements over sports or politics; the people above just shouting "triangle!" and "circle!" louder and louder). On the second level, people can (and often do) take the difference in output as evidence that *the other person's black box is broken* (i.e. they're bad, dumb, crazy) or that the other person *doesn't see the universe clearly* (i.e. they're biased, oblivious,

unobservant). On the third level, people will often *agree to disagree,* a move which preserves the social fabric at the cost of truth-seeking and actual progress.

Double crux in the ideal solves all of these problems, and in practice even fumbling and inexpert steps toward that ideal seem to produce a lot of marginal value, both in increasing understanding and in decreasing conflict-due-to-disagreement.

Prerequisites

This post will occasionally delineate two versions of double crux: a *strong* version, in which both parties have a shared understanding of double crux and have explicitly agreed to work within that framework, and a *weak* version, in which only one party has access to the concept, and is attempting to improve the conversational dynamic unilaterally.

In either case, the following things seem to be required:

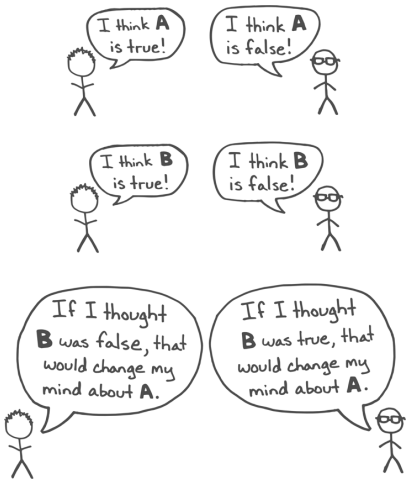
- **Epistemic humility.** The number one foundational backbone of rationality seems, to me, to be how readily one is able to think "It's possible that *I* might be the one who's wrong, here." Viewed another way, this is the ability to take one's beliefs as *object*, rather than being *subject to them* and unable to set them aside (and then try on some other belief and productively imagine "what would the world be like if *this* were true, instead of *that*?").
- **Good faith.** An assumption that people believe things for *causal reasons*; a recognition that having been exposed to the same set of stimuli would have caused one to hold approximately the same beliefs; a default stance of holding-with-skepticism what seems to be evidence that the other party is bad or wants the world to be bad (because as monkeys it's not hard for us to *convince* ourselves that we have such evidence when we really don't).1
- **Confidence in the existence of objective truth.** I was tempted to call this "objectivity," "empiricism," or "the Mulder principle," but in the end none of those quite fit. In essence: a conviction that for almost any well-defined question, there *really truly is* a clear-cut answer. That answer may be impractically or even impossibly difficult to find, such that we can't actually go looking for it and have to fall back on heuristics (e.g. how many grasshoppers are alive on Earth at this exact moment, is the color orange superior to the color green, why isn't there an audio book of Fight Club narrated by Edward Norton), but it nevertheless *exists.*
- **Curiosity and/or a desire to uncover truth.** Originally, I had this listed as truth-seeking alone, but my colleagues pointed out that one can move in the right direction simply by being curious about the other person and the contents of their map, without focusing directly on the territory.

At CFAR workshops, we hit on the first and second through specific lectures, the third through osmosis, and the fourth through osmosis and a lot of relational dynamics work that gets people curious and comfortable with one another. Other qualities (such as the ability to regulate and transcend one's emotions in the heat of the moment, or the ability to commit to a thought experiment and really wrestle with it) are also helpful, but not as critical as the above.

How to play

Let's say you have a belief, which we can label A (for instance, "middle school students should wear uniforms"), and that you're in disagreement with someone who believes some form of ¬A. *Double cruxing* with that person means that you're both in search of a second statement B, with the following properties:

- You and your partner both disagree about B as well (you think B, your partner thinks ¬B).
- The belief B is *crucial* for your belief in A; it is one of the *cruxes* of the argument. If it turned out that B was *not* true, that would be sufficient to make you think A was false, too.
- The belief ¬B is crucial for your partner's belief in ¬A, in a similar fashion.



In the example about school uniforms, B might be a statement like "uniforms help smooth out unhelpful class distinctions by making it harder for rich and poor students to judge one another through clothing," which your partner might sum up as "optimistic bullshit." Ideally, B is a statement that is somewhat *closer to reality* than A—it's more concrete, grounded, well-defined, discoverable, etc. It's less about principles and summed-up, induced conclusions, and more of a glimpse into the structure that led to those conclusions.

(It doesn't have to *be* concrete and discoverable, though—often after finding B it's productive to start over in search of a C, and then a D, and then an E, and so forth, until you end up with something you can research or run an experiment on).

At first glance, it might not be clear why simply *finding* B counts as victory—shouldn't you *settle* B, so that you can conclusively choose between A and ¬A? But it's important to recognize that arriving at B means you've *already* dissolved a significant chunk of your disagreement, in that you and your partner now *share a belief about the causal nature of the universe.*

If B, then A. Furthermore, if ¬B, then ¬A. You've both agreed that the states of B are *crucial* for the states of A, and in this way your continuing "agreement to disagree" isn't just "well, you take your truth and I'll take mine," but rather "okay, well, let's see what the evidence shows." Progress! And (more importantly) collaboration!

Methods

This is where CFAR's versions of the double crux unit are currently weakest—there's some form of magic in the search for cruxes that we haven't quite locked down. In general, the method is "search through your cruxes for ones that your partner is likely to disagree with, and then compare lists." For some people and some topics, clearly identifying your own cruxes is easy; for others, it very quickly starts to *feel like* one's position is fundamental/objective/un-break-downable.

Tips:

- **Increase noticing of subtle tastes, judgments, and "karma scores."** Often, people suppress a lot of their opinions and judgments due to social mores and so forth. Generally loosening up one's inner censors can make it easier to notice *why* we think X, Y, or Z.
- **Look forward rather than backward.** In places where the question "why?" fails to produce meaningful answers, it's often more productive to try making predictions about the future. For example, I might not know why I think school uniforms are a good idea, but if I turn on my narrative engine and start describing the better world I think will result, I can often sort of feel my way toward the underlying causal models.
- **Narrow the scope.** A specific test case of "Steve should've said hello to us when he got off the elevator yesterday" is easier to wrestle with than "Steve should be more sociable." Similarly, it's often easier to answer questions like "How much of our next $10,000 should we spend on research, as opposed to advertising?" than to answer "Which is more important right now, research or advertising?"
- **Do "Focusing" and other resonance checks.** It's often useful to try on a perspective, hypothetically, and then pay attention to your intuition and bodily responses to refine your actual stance. For instance: (*wildly asserts)* "I bet if everyone wore uniforms there would be a fifty percent reduction in bullying." (*pauses, listens to inner doubts)* "Actually, scratch that—that doesn't seem true, now that I say it out loud, but there *is* something in the vein of reducing *overt* bullying, maybe?"
- **Seek cruxes independently before anchoring on your partner's thoughts.** This one is fairly straightforward. It's also worth noting that if you're attempting to find disagreements in the first place (e.g. in order to practice double cruxing with friends) this is an excellent way to start—give everyone the same ten or fifteen open-ended questions, and have everyone write down their own answers based on their own thinking, crystallizing opinions *before* opening the discussion.

Overall, it helps to keep the *ideal* of a perfect double crux in the front of your mind, while holding the realities of your actual conversation somewhat separate. We've found that, at any given moment, increasing the "double cruxiness" of a conversation tends to be useful, but worrying about how far from the ideal you are in absolute terms doesn't. It's all about doing what's useful and productive in the moment, and that often means making sane compromises—if one of you has clear cruxes and the other is floundering, it's fine to focus on one side. If neither of you can find a single crux, but instead each of you has something like eight co-cruxes of which any five are sufficient, just say so and then move forward in whatever way seems best.

(Variant: a "trio" double crux conversation in which, at any given moment, if you're the least-active participant, your job is to squint at your two partners and try to model what each of them is saying, and where/why/how they're talking past one another and failing to see each other's points. Once you have a rough "translation" to offer, do so—

at that point, you'll likely become more central to the conversation and someone else will rotate out into the squinter/translator role.)

Ultimately, each move should be in service of reversing the usual antagonistic, warlike, "win at all costs" dynamic of most disagreements. Usually, we spend a significant chunk of our mental resources guessing at the shape of our opponent's belief structure, forming hypotheses about what things are crucial and lobbing arguments at them in the hopes of knocking the whole edifice over. Meanwhile, we're incentivized to obfuscate our own belief structure, so that our opponent's attacks will be ineffective.

(This is also terrible because it means that we often fail to even *find* the crux of the argument, and waste time in the weeds. If you've ever had the experience of awkwardly fidgeting while someone spends ten minutes assembling a conclusive proof of some tangential sub-point that never even had the *potential* of changing your mind, then you know the value of someone being willing to say "Nope, this isn't going to be relevant for me; try speaking to *that* instead.")

If we can move the debate to a place where, instead of *fighting* over the truth, we're *collaborating* on a search for understanding, then we can recoup a lot of wasted resources. You have a tremendous comparative advantage at knowing the shape of your own belief structure—if we can switch to a mode where we're each looking inward and candidly sharing insights, we'll move forward *much* more efficiently than if we're each engaged in guesswork about the other person. This requires that we want to know the *actual truth* (such that we're incentivized to seek out flaws and falsify wrong beliefs in ourselves just as much as in others) and that we feel emotionally and socially safe with our partner, but there's a doubly-causal dynamic where a tiny bit of double crux spirit up front can *produce* safety and truth-seeking, which allows for more double crux, which produces more safety and truth-seeking, etc.

Pitfalls

First and foremost, it matters whether you're in the strong version of double crux (cooperative, consent-based) or the weak version (you, as an agent, trying to improve the conversational dynamic, possibly in the face of direct opposition). In particular, if someone is currently riled up and conceives of you as rude/hostile/the enemy, then saying something like "I just think we'd make better progress if we talked about the *underlying reasons* for our beliefs" doesn't sound like a plea for cooperation—it sounds like a trap.

So, if you're in the weak version, the primary strategy is to embody the question **"What do you see that I don't?"** In other words, approach from a place of explicit humility and good faith, drawing out their belief structure *for its own sake,* to see and appreciate it rather than to undermine or attack it. In my experience, people can "smell it" if you're just playing at good faith to get them to expose themselves; if you're having trouble really getting into the spirit, I recommend meditating on times in your past when you were embarrassingly wrong, and how you felt *prior* to realizing it compared to *after* realizing it.

(If you're unable or unwilling to swallow your pride or set aside your sense of justice or fairness hard enough to really do this, that's **actually fine**; not every disagreement benefits from the double-crux-nature. But if your *actual goal* is improving the conversational dynamic, then this is a cost you want to be prepared to pay—going the

extra mile, because a) going what *feels like* an appropriate distance is more often an undershoot, and b) going an *actually appropriate* distance may not be enough to overturn their entrenched model in which you are The Enemy. Patience- and sanity-inducing rituals recommended.)

As a further tip that's good for either version but particularly important for the weak one, *model* the behavior you'd like your partner to exhibit. Expose your *own* belief structure, show how your *own* beliefs might be falsified, highlight points where you're uncertain and visibly integrate their perspective and information, etc. In particular, if you don't want people running amok with wrong models of what's going on in *your* head, make sure you're not acting like you're the authority on what's going on in *their* head.

Speaking of non-sequiturs, beware of getting lost in the fog. The very first step in double crux should *always* be to operationalize and clarify terms. Try attaching numbers to things rather than using misinterpretable qualifiers; try to talk about what would be observable in the world rather than how things feel or what's good or bad. In the school uniforms example, saying "uniforms make students feel better about themselves" is a *start,* but it's not enough, and going further into quantifiability (if you think you could actually get numbers someday) would be even better. **Often, disagreements will "dissolve" as soon as you remove ambiguity—this is success, not failure!**

Finally, *use paper and pencil,* or whiteboards, or get people to treat specific predictions and conclusions as immutable objects (if you or they want to change or update the wording, that's *encouraged*, but make sure that at any given moment, you're working with a clear, unambiguous statement). Part of the value of double crux is that it's the *opposite* of the weaselly, score-points, hide-in-ambiguity-and-look-clever dynamic of, say, a public political debate. The goal is to have everyone understand, at all times and as much as possible, what the other person is *actually* trying to say—not to try to get a straw version of their argument to stick to them and make them look silly. Recognize that *you yourself* may be tempted or incentivized to fall back to that familiar, fun dynamic, and take steps to keep yourself in "scout mindset" rather than "soldier mindset."

Algorithm

This is the double crux algorithm as it currently exists in our handbook. It's not strictly connected to all of the discussion above; it was designed to be read in context with an hour-long lecture and several practice activities (so it has some holes and weirdnesses) and is presented here more for completeness and as food for thought than as an actual conclusion to the above.

1. Find a disagreement with another person

- A case where you believe one thing and they believe the other
- A case where you and the other person have different confidences (e.g. you think X is 60% likely to be true, and they think it’s 90%)

2. Operationalize the disagreement

- Define terms to avoid getting lost in semantic confusions that miss the real point
- Find specific test cases—instead of (e.g.) discussing whether you should be more outgoing, instead evaluate whether you should have said hello to Steve in the

office yesterday morning
- Wherever possible, try to think in terms of actions rather than beliefs—it's easier to evaluate arguments like "we should do X before Y" than it is to converge on "X is better than Y."

3. Seek double cruxes

- Seek your own cruxes independently, and compare with those of the other person to find overlap
- Seek cruxes collaboratively, by making claims ("I believe that X will happen because Y") and focusing on falsifiability ("It would take A, B, or C to make me stop believing X")

4. Resonate

- Spend time "inhabiting" both sides of the double crux, to confirm that you've found the core of the disagreement (as opposed to something that will ultimately fail to produce an update)
- Imagine the resolution as an if-then statement, and use your inner sim and other checks to see if there are any unspoken hesitations about the truth of that statement

5. Repeat!

Conclusion

We think double crux is super sweet. To the extent that you see flaws in it, we want to find them and repair them, and we're currently betting that *repairing and refining double crux* is going to pay off better than *try something totally different.* In particular, we believe that embracing the spirit of this mental move has huge potential for unlocking people's abilities to wrestle with all sorts of complex and heavy hard-to-parse topics (like existential risk, for instance), because it provides a format for holding a bunch of partly-wrong models at the same time while you distill the value out of each.

Comments appreciated; critiques highly appreciated; anecdotal data from experimental attempts to teach yourself double crux, or teach it to others, or use it on the down-low without telling other people what you're doing **extremely** appreciated.

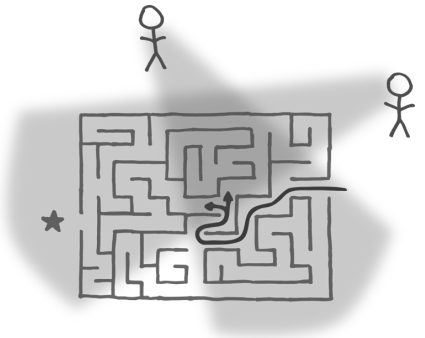
- Duncan Sabien

---

[1]One reason good faith is important is that even when people are "wrong," they are usually *partially* right—there are flecks of gold mixed in with their false belief that can be productively mined by an agent who's interested in getting the whole picture. Normal disagreement-navigation methods have some tendency to throw out that gold, either by allowing everyone to protect their original belief set or by replacing everyone's view with whichever view is shown to be "best," thereby throwing out data, causing information cascades, disincentivizing "noticing your confusion," etc.

The central assumption is that the universe is like a large and complex maze that each of us can only see parts of. To the extent that language and communication allow

us to gather info about parts of the maze without having to investigate them ourselves, that's great. But when we disagree on what to do *because* we each see a different slice of reality, it's nice to adopt methods that allow us to integrate and synthesize, rather than methods that force us to pick and pare down. It's like the parable of the three blind men and the elephant—whenever possible, avoid generating a bottom-line conclusion until you've accounted for *all* of the available data.



The agent at the top mistakenly believes that the correct move is to head to the left, since that seems to be the most direct path toward the goal. The agent on the right can see that this is a mistake, but it would never have been able to navigate to that particular node of the maze on its own.

Project Hufflepuff

(This is a crossposted FB post, so it might read a bit weird)

My goal this year (in particular, my main focus once I arrive in the Bay, but also my focus in NY and online in the meanwhile), is to join and champion the growing cause of people trying to fix some systemic problems in EA and Rationalsphere relating to "lack of Hufflepuff virtue".

I want Hufflepuff Virtue to feel exciting and important, because it is, and I want it to be something that flows naturally into our pursuit of both epistemic integrity, intellectual creativity, and concrete action.

Some concrete examples:

- on the 5 second reflex level, notice when people need help or when things need doing, and do those things.

- have an integrated understanding that being kind to people is *part* of helping them (and you!) to learn more, and have better ideas.

(There are a bunch of ways to be kind to people that do NOT do this, i.e. politely agreeing to disagree. That's not what I'm talking about. We need to hold each other to higher standards but not talk down to people in a fashion that gets in the way of understanding. There are tradeoffs and I'm not sure of the best approach but there's a lot of room for improvement)

- be excited and willing to be the person doing the grunt work to make something happen

- foster a sense that the community encourages people to try new events, actively take personal responsibility to notice and fix community-wide problems that aren't necessarily sexy.

- when starting new projects, try to have mentorship and teamwork built into their ethos from the get-go, rather than hastily tacked on later

I want these sorts of things to come easily to mind when the future people of 2019 think about the rationality community, and have them feel like central examples of the community rather than things that we talk about wanting-more-of.

80,000 Hours: EA and Highly Political Causes

this post is now crossposted to the EA forum

80,000 hours is a well known Effective Altruism organisation which does "in-depth research alongside academics at Oxford into how graduates can make the biggest difference possible with their careers".

They recently posted a guide to donating which aims, in their words, to (my emphasis)

> use evidence and careful reasoning to work out how to best promote the ***wellbeing of all***. To find the highest-impact charities this giving season ... We ... summed up the main recommendations by area below

Looking below, we find a section on the problem area of criminal justice (US-focused). An area where the aim is outlined as follows: (quoting from the Open Philanthropy "problem area" page)

> investing in criminal justice policy and practice reforms to substantially reduce incarceration while maintaining public safety.

Reducing incarceration whilst maintaining public safety seems like a reasonable EA cause, if we interpret "pubic safety" in a broad sense - that is, keep fewer people in prison whilst still getting almost all of the benefits of incarceration such as deterrent effects, prevention of crime, etc.

So what are the recommended charities? (my emphasis below)

1. Alliance for Safety and Justice

> "The Alliance for Safety and Justice is a US organization that aims to **reduce incarceration and racial disparities in incarceration** in states across the country, and replace mass incarceration with new safety priorities that prioritize prevention and protect low-income **communities of color.**"

They promote an article on their site called "black wounds matter", as well as how you can "Apply for VOCA Funding: A Toolkit for Organizations Working With Crime Survivors in **Communities of Color** and Other Underserved Communities"

2. Cosecha - (note that their url is www.lahuelga.com, which means "the strike" in Spanish) (my emphasis below)

> "Cosecha is a group **organizing undocumented immigrants** in 50-60 cities around the country. Its goal is to **build mass popular support for undocumented immigrants**, **in resistance to** incarceration/detention, **deportation**, denigration of rights, and discrimination. The group has become especially active since the Presidential election, given the immediate threat of mass incarceration and deportation of millions of people."

Cosecha have a footprint in the news, for example this article:

> They have the ultimate goal of launching massive civil resistance and non-cooperation to show this country it depends on us ... if they wage a general strike of five to eight million workers for seven days, we think the economy of this country would not be able to sustain itself

The article quotes Carlos Saavedra, who is directly mentioned by Open Philanthropy's Chloe Cockburn:

> Carlos Saavedra, who leads Cosecha, stands out as an organizer who is devoted to testing and improving his methods, ... Cosecha can do a lot of good to prevent mass deportations and incarceration, I think his work is a good fit for likely readers of this post."

They mention other charities elsewhere on their site and in their writeup on the subject, such as the conservative Center for Criminal Justice Reform, but Cosecha and the Alliance for Safety and Justice are the ones that were chosen as "highest impact" and featured in the guide to donating.

Sometimes one has to be blunt: 80,000 hours is promoting the financial support of some *extremely* hot-button political causes, which may not be a good idea. Traditionalists/conservatives and those who are uninitiated to Social Justice ideology might look at The Alliance for Safety and Justice and Cosecha and label them as them racists and criminals, and thereby be turned off by Effective Altruism, or even by the rationality movement as a whole.

There are standard arguments, for example this by Robin Hanson from 10 years ago about why it is not smart or "effective" to get into these political tugs-of-war if one wants to make a genuine difference in the world.

One could also argue that the 80,000 hours' charities go beyond the usual folly of political tugs-of-war. In addition to supporting extremely political causes, 80,000 hours could be accused of being somewhat intellectually dishonest about what goal they are trying to further actually is.

Consider The Alliance for Safety and Justice. 80,000 Hours state that the goal of their work in the criminal justice problem area is to "substantially reduce incarceration while maintaining public safety". This is an abstract goal that has very broad appeal and one that I am sure almost everyone agrees to. But then their more concrete policy in this area is to fund a charity that wants to "reduce racial disparities in incarceration" and "protect low-income communities of color". The latter is *significantly* different to the former - it isn't even close to being the same thing - and the difference is highly political. One could object that reducing racial disparities in incarceration is merely a *means to the end* of substantially reducing incarceration while maintaining public safety, since many people in prison in the US are "of color". However this line of argument is a very politicized one and it might be wrong, or at least I don't see strong support for it. "Selectively release people of color and make society safer - endorsed by effective altruists!" struggles against known facts about redictivism rates across races, as well as an objection about the implicit conflation of equality of outcome and equality of opportunity. (and I do not want this to be interpreted as a claim of moral superiority of one race over others - merely a

necessary exercise in coming to terms with facts and debunking implicit assumptions). Males are incarcerated much more than women, so what about reducing gender disparities in incarceration, whilst also maintaining public safety? Again, this is all highly political, laden with politicized implicit assumptions and language.

Cosecha is worse! They are actively planning potentially illegal activities like helping illegal immigrants evade the law (though IANAL), as well as activities which potentially harm the majority of US citizens such as a seven day nationwide strike whose intent is to damage the economy. Their URL is "The Strike" in Spanish.

Again, the abstract goal is extremely attractive to almost anyone, but the concrete implementation is highly divisive. If some conservative altruist signed up to financially or morally support the abstract goal of "substantially reducing incarceration while maintaining public safety" and EA organisations that are pursuing that goal without reading the details, and then at a later point they saw the details of Cosecha and The Alliance for Safety and Justice, they would rightly feel cheated. And to the objection that conservative altruists should read the description rather than just the heading - what are we doing writing headings so misleading that you'd feel cheated if you relied on them as summaries of the activity they are mean to summarize?

One possibility would be for 80,000 hours to be much more upfront about what they are trying to achieve here - maybe they like left-wing social justice causes, and want to help like-minded people donate money to such causes and help the particular groups who are favored in those circles. There's almost a nod and a wink to this when Chloe Cockburn says (my paraphrase of Saavedra, and emphasis, below)

> I think his [A man who wants to lead a general strike of five to eight million workers for seven days so that the economy of the USA would not be able to sustain itself, in order to help illegal immigrants] work is a good fit **for likely readers of this post**.

Alternatively, they could try to reinvigorate the idea that their "criminal justice" problem area is politically neutral and beneficial to everyone; the Open Philanthropy issue writeup talks about "conservative interest in what has traditionally been a solely liberal cause" after all. I would advise considering dropping The Alliance for Safety and Justice and Cosecha if they intend to do this. There may not be politically neutral charities in this area, or there may not be enough high quality conservative charities to present a politically balanced set of recommendations. Setting up a growing donor advised fund or a prize for nonpartisan progress that genuinely intends to benefit *everyone* including conservatives, people opposed to illegal immigration and people who are not "of color" might be an option to consider.

We could examine 80,000 hours' choice to back these organisations from a more overall-utilitarian/overall-effectiveness point of view, rather than limiting the analysis to the specific problem area. These two charities don't pass the smell test for altruistic consequentialism, pulling sideways on ropes, finding hidden levers that others are ignoring, etc. Is the best thing you can do with your smart EA money helping a charity that wants to get stuck into the culture war about which skin color is most over-represented in prisons? What about a second charity that wants to help people

illegally immigrate at a time when immigration is the most divisive political topic in the western world?

Furthermore, Cosecha's plans for a nationwide strike and potential civil disobedience/showdown with Trump & co could push an already volatile situation in the US into something extremely ugly. The vast majority of people in the world (present and future) are not the specific group that Cosecha aims to help, but the set of people who could be harmed by the uglier versions of a violent and calamitous showdown in the US is basically the whole world. That means that even if P(Cosecha persuades Trump to do a U-turn on illegals) is 10 or 100 times greater than P(Cosecha precipitates a violent crisis in the USA), they may still be net-negative from an expected utility point of view. EA doesn't usually fund causes whose outcome distribution is heavily left-skewed so this argument is a bit unusual to have to make, but there it is.

Not only is Cosecha a cause that is (a) mind-killing and culture war-ish (b) very tangentially related to the actual problem area it is advertised under by 80,000 hours, but it might also (c) be an anti-charity that produces net disutility (in expectation) in the form of a higher probability a US civil war with money that you donate to it.

Back on the topic of criminal justice and incarceration: opposition to reform often comes from conservative voters and politicians, so it might seem unlikely to a careful thinker that extra money on the left-wing side is going to be highly effective. Some intellectual judo is required; make conservatives think that it was their idea all along. So promoting the Center for Criminal Justice Reform sounds like the kind of smart, against-the-grain idea that might be highly effective! Well done, Open Philanthropy! Also in favor of this org: they don't copiously mention which races or person-categories they think are most important in their articles about criminal justice reform, the only culture war item I could find on them is the world "conservative" (and given the intellectual judo argument above, this counts as a plus), and they're not planning a national strike or other action with a heavy tail risk. But that's the one that *didn't make the cut* for the 80,000 hours guide to donating!

The fact that they let Cosecha (and to a lesser extent The Alliance for Safety and Justice) through reduces my confidence in 80,000 hours and the EA movement as a whole. Who thought it would be a good idea to get EA into the culture war with these causes, and also thought that they were plausibly among the most effective things you can do with money? Are they taking effectiveness seriously? What does the political diversity of meetings at 80,000 hours look like? Were there no conservative altruists present in discussions surrounding The Alliance for Safety and Justice and Cosecha, and the promotion of them as "beneficial for everyone" and "effective"?

Before we finish, I want to emphasize that this post is not intended to start an object-level discussion about which race, gender, political movement or sexual orientation is cooler, and I would encourage moderators to temp-ban people who try to have that kind of argument in the comments of this post.

I also want to emphasize that criticism of professional altruists is a necessary evil; in an ideal world the only thing I would ever want to say to people who dedicate their lives to helping others (Chloe Cockburn in particular, since I mentioned her name above) is "thank you, you're amazing". Other than that, comments and criticism are welcome, especially anything pointing out any inaccuracies or misunderstandings in this post. Comments from anyone involved in 80,000 hours or Open Philanthropy are welcome.

Most empirical questions are unresolveable; The good, the bad, and the appropriately under-powered

This is a linkpost for https://medium.com/@davidmanheim/the-good-the-bad-and-the-appropriately-under-powered-82c335652930