Staying Sane While Taking Ideas Seriously

Adding Up To Normality

Related: Leave a Line of Retreat, Living In Many Worlds

> "It all adds up to normality." Greg Egan, *Quarantine*

You're on an airplane at 35,000 feet, and you strike up a conversation about aerodynamic lift with the passenger in your row. Things are going along just fine until they point out to you that your understanding of lift is wrong, and that planes couldn't fly from the effect you thought was responsible.

Should you immediately panic in fear that the plane will plummet out of the sky?

Obviously not; clearly the plane has been flying just fine up until now, and countless other planes have flown as well. There has to be *something* keeping the plane up, even if it's not what you thought, and even if you can't yet figure out what it actually is. Whatever is going on, it all adds up to normality.

Yet I claim that we often do this exact kind of panicked flailing when there's a challenge to our philosophical or psychological beliefs, and that this panic is entirely preventable.

I've experienced and/or seen this particular panic response when I, or others, encounter good arguments for propositions including

- My religion is not true. ("Oh no, then life and morality are meaningless and empty!")
- Many-worlds makes the most sense. ("Oh no, then there are always copies of me doing terrible things, and so none of my choices matter!")
- Many "altruistic" actions actually have hidden selfish motives. ("Oh no, then altruism doesn't exist and morality is pointless!")
- I don't have to be the best at something in order for it to be worth doing. ("Oh no, then others won't value me!") [Note: this one is from therapy; most people don't have the same core beliefs they're stuck on.]

(I promise these are not in fact strawmen. I'm sure you can think of your own examples. Also remember that panicking over an argument in this way is a mistake even if the proposition turns out to be false.)

To illustrate the way out, let's take the first example. It took me far too long to leave my religion, partly because I was so terrified about becoming a nihilist if I left that I kept flinching away from the evidence. (Of course, the religion proclaimed itself to be the origin of morality, and so it reinforced the notion that anyone else claiming to be moral was just too blind to see that their lack of faith implied nihilism.)

Eventually I did make myself face down, not just the object-level arguments, but the biases that had kept me from looking directly at them. And then I was an atheist, and still I was terrified of becoming a nihilist (especially about morality).

So I did one thing I still think was smart: I promised myself not to change all of my moral rules at once, but to change each one only when (under sober reflection) I decided it was wrong. And in the meantime, I read a lot of moral philosophy.

Over the next few months, I began relaxing the rules that were obviously pointless. And then I had a powerful insight: I was so cautious about changing my rules *because I wanted to help people and not slide into hurting them*. Regardless of what morality was, in fact, based on, the plane was still flying just fine. And that helped me sort out the good from the bad among the remaining rules, and to stop being so afraid of what arguments I might later encounter.

So in retrospect, the main thing I'd recommend is to **promise yourself to keep steering the plane mostly as normal while you think about lift** (to stretch the analogy). If you decide that something major is false, it doesn't mean that everything that follows from it has to be discarded immediately. (False things imply both true and false things!)

You'll generally find that many important things stand on their own without support from the old belief. (Doing this for the other examples I gave, as well as your own, is left to you.) Other things will collapse, and that's fine; that which can be destroyed by the truth should be. Just don't make all of these judgments in one fell swoop.

One last caution: **I recommend against changing meta-level rules as a result of changing object-level beliefs**. The meta level is how you correct bad decisions on the object level, and it should only be updated by very clear reasoning in a state of equilibrium. Changing your flight destination is perfectly fine, but don't take apart the wing mid-flight.

Good luck out there, and remember:

It all adds up to normality.

[EDIT 2020-03-25: khafra and Isnasene make good points about not applying this in cases where the plane shows signs of *actually dropping* and you're updating on *that*. (Maybe there's a new crisis in the external world that contradicts one of your beliefs, or maybe you update to believe that the thing you're about to do could actually cause a major catastrophe.)

In that case, you can try and land the plane safely- focus on getting to a safer state for yourself and the world, so that you have time to think things over. And if you can't do that, then you have no choice but to rethink your piloting on the fly, accepting the danger because you can't escape it. But these experiences will hopefully be very rare for you, current global crisis excepted.]

Negotiating With Yourself

(Talk given on Sunday 21st June, over a zoom call with 40 attendees. orthonormal is responsible for the talk, jacobjacob is responsible for the transcription)

Talk

orthonormal: So, I'm doing a generalisation of the post that was curated and this post is sort of an elaboration of what Vaniver talked about. If you notice that there are differences between your private intuitions, and what you can publicly acknowledge, this is a system fast versus system slow thing-

orthonormal: This is a question of negotiating with yourself. I'm going to present a model and talk about some consequences, but I'll start with a question: why do people in our sphere tend to burn out or go nuts?

orthonormal: This is a pretty important question. I'll use an analogy many of us have heard before — the elephant and the rider. The conscious mind is the rider and the elephant is the unconscious mind. The rider wants to get somewhere, but the elephant has its own preferences about what happens.

orthonormal: Some features of this analogy I think are true and useful for minds, for humans, is that the elephant has these immediate preferences and some longer term needs, just like we have subconscious desires and subconscious needs. The rider has their own preferences and some carrots and sticks, but the real advantage is having a map. The elephant can just completely ignore the rider if its preferences are strong enough. So how this connects to being human is that our subconscious has these desires and needs and fears, and our consciousness may have a little bit of willpower but what it really has is strategy and planning, the ability to pick out a path so that the elephant won't want to deviate from it too much. If you go right by a river, the elephant is going to want to drink. So, if you don't want to stop for that, don't go by the river right now.

orthonormal: Finally, the subconscious, the elephant, can just overwhelm you in two ways: one of which is it controls your motivation, so it can burn you out or get you depressed if you're trying to defy it too much. And the second is, it can induce bias.

orthonormal: This is the subject of *The Elephant in the Brain* by Robin Hanson and Kevin Simler, claiming that a lot of motivated reasoning comes when the elephant wants something, the rider doesn't, and the elephant changes the rider's cognition to make the rider feel like it wants that thing (for noble reasons, of course).

orthonormal: This would be very bad. Depression is its own thing, but it doesn't change your way of thinking about the world. It doesn't make you go crazy. Going crazy is really bad. Citation — don't need it in this community.

orthonormal: What can you do about this? There are a couple of things. First: you can keep the elephant happy. You can choose a path along the map so that the elephant will be reasonably well fed, have enough to drink, not get tired going up and down mountains, etc. And you can still get to a place you'd like to go. Maybe not the place that's absolute best, but good enough.

orthonormal: This is analogous to a lot of things in Effective Altruism where I'm telling people, “give yourself permission to be happy”. Don't take a job that's going to make you miserable just because you think it is the best thing to do. Find something that meets you in the middle. I don't recommend living on minimum wage and giving away everything else to charity because you're going to burn out from that, or you're going to come up with some crazy reason why doing something else is better. So just let yourself be happy. 80/20 things.

orthonormal: The second thing is about positive versus negative reinforcement. I mentioned carrots versus sticks earlier, and this is really good for also keeping the elephant happy and keeping the elephant liking the rider. There's a wonderful book called Don't Shoot the Dog, which is primarily about animal training, but also about interacting with people — and even about interacting with yourself. It talks about achieve things in animal training by rewarding the animal or by punishing the animal. Rewarding the animal, you can get them to do great things. Punishing, you can get them to do some things... but they'll also just want to avoid the trainer. You don't want your subconscious mind to want to avoid your thoughts. It'll make it even harder to find out what's going on with your desires.

orthonormal: Finally, real quick, treat the elephant with respect, even if you disagree with it. It's really important for you to be able to say, not "Your desires are wrong", but "I understand why you want that, I want this other thing, let's find common ground." And I think those are some of the really important lessons about the elephant and the rider.

orthonormal: Thank you.

Q&A

Ben Pace: Cool. Thank you very much, orthonormal.

Ben Pace: I like the emphasis you made on having a respectful dialog with the elephant. You spoke about making the elephant happy. I understand the point you're making. But often my relationship with my elephant, when I try to have an internal dialog, is more about asking what it wants and making a commitment to getting it that thing. And those things are not necessarily happiness. They're sometimes respect, or status, or just commitments to find time for the elephant to do the things it wants, whilst also making agreements to work what the rider wants. Some of those motives are not directly about immediately pleasurable experiences. So I always make that distinction.

Ben Pace: Abram has a question.

Abram Demski: Yeah, sometimes I've heard this advice that you should identify with the elephant instead of the rider. It's also a diversity question, you're speaking to people who identify with the rider rather than the elephant but some people identify more with the elephant — or so I've heard. One part of it is normative. Like, maybe we should identify with the elephant instead of the rider? So the question is: what do you have against that, if anything?

orthonormal: It would be nice to be unified, but one thing I think is true is that the rider is good at language and the elephant is not. So the part of you that just asked

me that question is the rider.

Abram Demski: I guess I have this drug experience where I was high and I completely separated my consciousness and my audio loop. So, my inner dialog did not feel conscious and instead, I felt like I was the consciousness that the inner dialog is talking about. Which doesn't change my day-to-day thinking that much but makes me able to take that framework where it's like words are coming out of my mouth from this thing that's looking at my actual conscious experience. But my conscious experience is not this thing. I don't have conscious access to... Compare how people know grammar without having explicit knowledge of grammar *[Editor's note: source].* So it's like there's this grammar thing here that somehow knows grammar and it's looking at my conscious experience and producing words that try to describe my conscious experience but that doesn't mean that my conscious experience is the thing that's... You're sort of not talking to my words, my words are like a special case module. You're interacting with my conscious experience indirectly through my words, but my words are kind of dumb.

orthonormal: This is just a hard thing to talk about. I very much believe your experiences and I very much believe that there is something to, through meditation or drugs or whatever, getting more in touch with the non-verbal part of you and having more compassion and connection to that. It's just very complicated to describe in words what that looks and feels like, for obvious reasons.

Abram Demski: Yeah.

Ben Pace: Thanks, Abram, I appreciate that way of thinking about yourself. I think I will probably meditate on that some more afterwards.

Ben Pace: Kamil, do you want to ask a question?

Kamil: Yeah I think that this concept looks like internal double-crux, and if so, my question is, maybe there would be some more sub-personalities, more than just elephant and rider — maybe, some other decision makers in our mind?

orthonormal: Absolutely. The elephant/rider is an extremely simplified version of things. Personally I like the internal family systems approach to understanding myself. Again, all of these are metaphor, but metaphors can be very useful. The internal family systems metaphors treat different desires and feelings as different agents, more or less, that can talk to each other. So, whatever metaphor works well for people, I encourage them to use that while being aware that it's a metaphor, and also to experiment with other ways of thinking about themselves.

Ben Pace: Cool. Thanks, Kamil, does that sound good to you or do you want to follow up?

Kamil: Yeah, thanks. If so, what are the constraints of this model? Of the model of the elephant and the rider?

orthonormal: Right. The fundamental constraint for my metaphor, at first, is that the conscious part of the mind, which for me includes the verbal part of the mind, is just less strong than everything else that happens, whether that everything else is unified or an aggregate of other parts.

Kamil: Thanks.

Map Errors: The Good, The Bad, and The Territory

What happens when your map doesn't match the territory?

There's one aspect of this that's potentially very helpful to becoming a rationalist, and one aspect that's very dangerous. The good outcome is that you could understand map errors more deeply; the dangerous outcome is that you could wind up stuck somewhere awful, with no easy way out.

The first version, where you notice that the map is wrong, comes when the map is undeniably *locally* wrong. The map says the path continues here, but instead there's a cliff. (Your beliefs strongly predict something, and the opposite happens.)

The ordinary result is that you scratch out and redraw that part of the map – or discard it and pick up an entirely different map – and continue along the new path that looks best. (You decide you were wrong on that one point without questioning any related beliefs, or you convert to a completely different belief system which was correct on that point.)

The really valuable possibility is that you realize that there are probably *other* errors besides the one you've seen, and probably unseen errors on the other available maps as well; you start to become more careful about trusting your maps so completely, and you pay a bit more attention to the territory around you.

This is a really important formative experience for many rationalists:

- Take ideas seriously enough to notice and care if they fail
- Get smacked in the face with an Obvious But False Belief: your past self couldn't have imagined you were wrong about this, and yet here we are.
- Deeply internalize that one's sense of obviousness *cannot be trusted*, and that one has to find ways of being way more reliable where it matters.

(For me the Obvious But False Belief was about religion; for others it was politics, or an academic field, or even their own identity.)

Now, the dangerous outcome – getting trapped in a dismal swamp, with escape very difficult – comes when you've not seen an undeniable local map failure, so that you never notice (or never have to admit) that the map isn't matching up very well with the territory, until it's too late.

(I'm thinking of making major life decisions badly, where you don't notice or admit the problem until you're trapped in a situation where every option is a disaster of some sort.)

Sometimes you really do need to make bold plans based on your beliefs; how can you do so without taking a big risk of ending up in a swamp?

I suggest that you should ensure things look at least decent, according to a more "normal" map, while trying to do very well on yours. That is, make sure that your bold plan fails gracefully if the more normal worldview around you is correct. (Set up your

can't-miss startup such that you're back to the grind if it fails, not in debt to the Mob if it fails.)

And get advice. Always get advice from people you trust and respect, before doing something very uncommon. I could try and fit this into the map framework, but it's just common sense, and way too many good people fail to do it regardless.

Best of luck adventuring out there!

The Loudest Alarm Is Probably False

Epistemic Status: Simple point, supported by anecdotes and a straightforward model, not yet validated in any rigorous sense I know of, but IMO worth a quick reflection to see if it might be helpful to you.

A curious thing I've noticed: among the friends whose inner monologues I get to hear, the most self-sacrificing ones are frequently worried they are being too selfish, the loudest ones are constantly afraid they are not being heard, the most introverted ones are regularly terrified that they're claiming more than their share of the conversation, the most assertive ones are always suspicious they are being taken advantage of, and so on. It's not just that people are sometimes miscalibrated about themselves- it's as if the loudest alarm in their heads, the one which is apt to go off at any time, is pushing them in the exactly wrong direction from where they would flourish.

Why should this be? (I mean, presuming that this pattern is more than just noise and availability heuristic, which it could be, but let's follow it for a moment.)

It's exactly what we should expect to happen if (1) the human psyche has different "alarms" for different social fears, (2) these alarms are supposed to calibrate themselves to actual social observations but occasionally don't do so correctly, and (3) it's much easier to change one's habits than to change an alarm.

In this model, while growing up one's inner life has a lot of alarms going off at various intensities, and one scrambles to find actions that will calm the loudest ones. For many alarms, one learns habits that basically work, and it's only in exceptional situations that they will go off loudly in adulthood.

But if any of these alarms don't calibrate itself correctly to the signal, then they eventually become by far the loudest remaining ones, going off all the time, and one adjusts one's behavior as far as possible in the other direction in order to get some respite.

And so we get the paradox, of people who seem to be incredibly diligently following the exact wrong advice for themselves, analogous to this delightful quote (hat tip Siderea) about system dynamics in consulting:

> People know intuitively where leverage points are. Time after time I've done an analysis of a company, and I've figured out a leverage point — in inventory policy, maybe, or in the relationship between sales force and productive force, or in personnel policy. Then I've gone to the company and discovered that there's already a lot of attention to that point. Everyone is trying very hard to push it IN THE WRONG DIRECTION!

The funny thing about cognitive blind spots (and that's what we're looking at here) is that you can get pretty far into reading an article like this, hopefully enjoying it along the way, and forget to ask yourself if the obvious application to your own case might be valid.

If so, no worries! I developed this idea an embarrassingly long time before I thought to ask myself what would be the constant alarm going off in my own head. (It was the

alarm, "people aren't understanding you, you need to keep explaining", which was a huge epiphany to me but blindingly clear to anyone who knew me.)

And the framing that helped me instantly find that alarm was as follows:

What do I frequently fear is going wrong in social situations, despite my friends' reliable reassurance that it's not?

That fear is worth investigating as a possibly broken alarm.

Don't Make Your Problems Hide

I've seen a worrying trend in people who've learned introspection and self-improvement methods from CFAR, or analogous ones from CBT. They make better life decisions, they calm their emotions in the moment. But they still look just as stressed as ever. They stamp out every internal conflict they can see, but it seems like there are more of them beyond the horizon of their self-awareness.

(I may have experienced this myself.)

One reason for this is that there's a danger with learning how to consciously notice and interact with one's subconscious thoughts/feelings/desires/fears: the conscious mind may not like what it sees, and try to edit the subconscious mind into one that pleases it.

The conscious mind might *try*, that is, but the subconscious is stronger. So, what actually happens?

The subconscious develops defense mechanisms.

Suppressed desires disguise themselves as being about other things, or they just overwhelm the conscious mind's willpower every now and then (and maybe fulfill themselves in a less healthy way than could otherwise be managed).

Suppressed thoughts become stealthy biases; certain conscious ideas or narratives get reinforced until they are practically unquestionable. So too with fears; a suppressed social fear is a good way to get a loud alarm that never stops.

Suppressed feelings hide themselves more thoroughly from the searchlight, so that one never consciously notices their meaning anymore, one just feels sad or angry or scared "for no reason" in certain situations.

At its worst, the conscious mind tries ever-harder to push back against these, further burning its rapport with the subconscious. I think of pastors who suppress their gay desires so hard that they vigorously denounce homosexuality and then sneak out for gay sex. They'd have been living such a happier life if they'd given up and acknowledged who they are, and what they want, years ago.

Now, sometimes people do have a strong desire that can't be satisfied in any healthy way. And that's just a brutal kind of life to life. But they would still do better by acknowledging that desire openly to themselves, than by trying to quash it and only hiding it.

How can we become more integrated between conscious and unconscious parts, and undo any damage we've already caused?

In my talk about the elephant and rider, I suggested (or gestured at) a few relevant things:

- Pursue basic happiness alongside your conscious goals (and make sure that's happiness *for you*, not just e.g. keeping your friends happy by doing the things

they like)
- Use positive reinforcement on yourself rather than punishment - it's especially important not to punish yourself for noticing the "wrong" thoughts/feelings/desires/fears. Reward the noticing, even with just an internal "thank you for surfacing this".
- Treat the content of these thoughts/feelings/desires/fears with respect. You might think of them as a friend opening up to you, and imagine the compassion you'd have when trying to figure out a way forward where both of you can flourish.

It's important to be gentle, to be curious, and to be patient. You don't have to resolve the whole thing; just acknowledging it respectfully can help the relationship grow.

There are other approaches too. Many people believe in using meditation to better integrate their thoughts and feelings and desires, for instance.

When you do something that you thought you didn't want to do, or when you're noticing an unexpected feeling, it's an opportunity for you. Don't push it away.

Roleplaying As Yourself

(*This is a basic intuition pump I've found helpful in making decisions, and maybe you'll like it too.)*

For all its shortcomings, I think there was something quite useful about the "What Would Jesus Do?" meme within the Christian framework. Of course it's not a very sophisticated ethical guide, and it comes with all kinds of biases; but asking it does put the believer into a frame of mind that emphasizes things like compassion and duty, and it sometimes helps the believer generate options that weren't in their default solution space.

Is there a version of this handy tool for the consequentialist, with our muddled mixture of selfish and altruistic goals and impulses, and the added difficulty that we're looking to actually optimize rather hard?

The one that works best for me is a double roleplay.

Jernau Gurgeh, champion of strategic games across the galaxy, sits down to a nice futuristic immersive roleplaying game: The Orthonormal Experience. Gurgeh will be controlling a denizen of the early 21st century on Earth, someone with the online name of Orthonormal. Getting Orthonormal to do well by Orthonormal's own standards is Gurgeh's objective.

Just as our roleplaying games have game masters who can call out uncharacteristic plans, so too does Gurgeh's game. He can't simply use his superior vantage point to calculate the right stocks for Orthonormal to buy today and sell tomorrow, because Orthonormal couldn't do that except by luck. He can't even have Orthonormal think at peak performance on some days- there are character attributes (penalties like Anxiety Disorder) he has to play around.

But Gurgeh is able to think patiently, and strategically, about the various obstacles blocking Orthonormal's progress, and to guide Orthonormal's thoughts in plausible ways to work on these. There's a lot of points out there to be scored: better states to reach in Orthonormal's relationships, career, inner life, and more.

What would Gurgeh do?

One last note: a couple of the bugs with this approach can be confronted within the approach itself. If I decide that Gurgeh might do X, and I try X and fail, it can be tempting to get frustrated with myself. But this isn't what Gurgeh would do next! He'd take my failure as more data about what this character's current attributes are, and look for ways to work around that failure mode or to train the relevant attribute. And he'd probably give the character a short rest to recover mana before trying again.

The Real Standard

Long-delayed followup To: Roleplaying As Yourself

(Another simple intuition pump, this one especially useful for effective altruists who are struggling with wanting to do more or worrying they're not doing enough.)

Previously I wrote about a mental tool for prompting good consequentialist reasoning: ask yourself what a skilled alien roleplayer (here Gurgeh, from Player of Games) would do if they were controlling you, had to take only actions you could plausibly take, and scored points for achieving your goals.

This also serves as a standard for comparing your own actions, though as an aspiration rather than as an expectation.

The reason I mention this is that a good number of people in the rationality and effective altruism communities suffer from scrupulosity, the sense of guilt for not living up to an unattainable standard of conduct. And if we're going to speak to that sense, we need to start by getting the right standard of excellence to bargain with.

(If you're feeling scrupulous about altruism in particular, then you can imagine that Gurgeh gets points for achieving only your altruistic aims, though he's still constrained by your actual needs - he wouldn't steer you into a burnout, that wouldn't maximize his score.)

This standard is insanely daunting. Fortunately, it's not fair to ask you to meet it.

After all, you're not perfectly altruistic, and the other parts get to bargain too.

In Nobody Is Perfect, Everything Is Commensurable, Scott suggests that we deal with scrupulosity by letting ourselves be okay with the standard of giving 10% of our output to the most effective charitable causes. He runs into a bit of a problem when dealing with the fact that people are in different places in their careers (and that a tenth of one's income can be a large or small chunk of one's disposable income), and punts on the question a bit:

> If you make $30,000 and you accept 10% as a good standard you want to live up to, you can either donate $3000 to charity, or participate in political protests until your number of lives or dollars or DALYs saved is equivalent to that.

I think this is the right place to introduce the alien gamer roleplaying your character. Are you building intangible expertise or career capital? Gurgeh notices the high payoff in later rounds of the game from these resources, and would be happy to forgo a little more short-term impact if your time/money/attention can translate into those resources more effectively. Are you torn between multiple opportunities to do good? Gurgeh checks once to see if there's a synergy between them (a way to get a higher combined total than he would optimizing for either alone), and if not, he ruthlessly picks the one that translates more efficiently into points, and doesn't feel bad about leaving behind a less efficient path.

So here's my suggestion:

Figure out the expected score that you'd actually expect Gurgeh to get in "The Altruistic You Experience", then consider ways to *achieve at least one-tenth of that score*, and let *that* be your target for moral achievement.

This is still a really high standard, one that few achieve! It almost surely isn't enough to take your default path in life while giving even 50% of your income to the best charity. It may require you to change your career, your social circles, your everyday habits. It may ask you to do lots of self-experimentation, with the corresponding expectation of frequent failure.

But it at least leaves more slack for your own flourishing than attempting to achieve the altruistic high score. It lets you seek a way of achieving excellence that satisfies your other wants and needs well. Maybe you don't take your altruistic best option if your second best is much more personally fulfilling; maybe you go ahead and splurge on something big every now and then. But you don't lose sight of your aspiration.

I just want to emphasize:

It's okay to give yourself more happiness and more leisure than you need in order to be effective. It's okay to care about your own well-being, and that of your family and friends, than that of strangers in far-off lands or times.

It's okay to be mostly selfish. Just be *strategic* about the altruistic part.

Choosing the Zero Point

Summary: *You can decide what state of affairs counts as neutral, and what counts as positive or negative. Bad things happen if humans do that in our natural way. It's more motivating and less stressful if, when we learn something new, we update the neutral point to [what we think the world really is like now].*

A few years back, I read an essay by Rob Bensinger about vegetarianism/veganism, and it convinced me to at least eat much less meat. **This post is not about that topic.** It's about the way that essay differed, psychologically, from many others I've seen on the same topic, and the general importance of that difference.

Rob's essay referred to the same arguments I'd previously seen, but while other essays concluded with the implication "you're doing great evil by eating meat, and you need to realize what a monster you've been and immediately stop", Rob emphasized the following:

> Frame animal welfare activism as an astonishingly promising, efficient, and uncrowded opportunity to do good. Scale back moral condemnation and guilt. LessWrong types can be powerful allies, but the way to get them on board is to give them opportunities to feel like munchkins with rare secret insights, not like latecomers to a not-particularly-fun party who have to play catch-up to avoid getting yelled at. It's fine to frame helping animals as *challenging*, but the challenge should be to excel and do something astonishing, not to meet a bare standard for decency.

That shouldn't have had different effects on me than other essays, but damned if it didn't.

Consider a utilitarian Ursula with a utility function U. U is defined over all possible ways the world could be, and for each of those ways it gives you a number. Ursula's goal is to maximize the expected value of U.

Now consider the utility function V, where V always equals U + 1. If a utilitarian Vader with utility function V is facing the same choice (in another universe) as Ursula, then because that +1 applies to *every* option equally, the right choice for Vader is the same as the right choice for Ursula. The constant difference between U and V doesn't matter for any decision whatsoever!

We represent this by saying that a utility function is only defined up to positive affine transformations. (That means you can also multiply U by any positive number and it still won't change a utilitarian's choices.)

But humans aren't perfect utilitarians, in many interesting ways. One of these is that our brains have a natural notion of outcomes that are good and outcomes that are bad, and the neutral zero point is more or less "the world I interact with every day".

So if we're suddenly told about a nearby bottomless pit of suffering, what happens?

Our brains tend to hear, "Instead of the zero point where we thought we were, this claim means that we're really WAY DOWN IN THE NEGATIVE ZONE".

A few common reactions to this:

- *Denial.* "Nope nope that argument can't be true, I'm sure there's a flaw in it, we're definitely still in the normal zone"
- *Guilt.* "AAAAHHHH I need to drop everything and work super hard on this, I can't allow myself any distractions or any bit of happiness until this is completely fixed"
- *Despair.* "Oh no, there's no way I could get things back up to normal from here, I can't do anything, I'll just sit here and hate myself"

The thing about Rob's post is that it suggested an alternative. Instead of keeping the previous zero point and defining yourself as now being very far below it, **you can reset yourself to take the new way-the-world-is as the zero point**.

Again, this doesn't change any future choice a *utilitarian you* would make! But it does buy *human you* peace of mind. What is true is already so- the world was like this even when you didn't believe it.

The psychological benefits of this transformation:

- *Acceptance.* Is it too scary to consider the new hypothesis? No! If you accept it, you'll still start at zero, you'll just have an opportunity to do more kinds of good than you previously thought existed.
- *Relief.* Must you feel ashamed for not working your fingers to the bone? No! If you're pushing the world into the positive zone, it feels much more okay to 80-20 your efforts.
- *Hope.* Must you despair if you can't reach your old zero? No! Seen from here, this was always the world, but now you can help move it up from zero! It doesn't have to go higher than you can reach in order to be worthwhile.

A few last notes:

- I really recommend doing this for oneself first of all, and then extending it to one's efforts of persuasion.
- There are a few cases where a desperate effort is called for, but even then we can frame it as building something great that the world urgently needs.
- When it comes to personal virtue, the true neutral point for yourself shouldn't be "doing everything right", because you're consigning yourself to living in negative-land. A better neutral point is "a random person in my reference class". How are you doing relative to a typical [insert job title or credential or hobby here], in your effects on that community? Are you showing more discipline than the typical commenter on your Internet forum? That's a good starting point, and you can go a long way up from there.
- (Thanks to Isnasene for helping me realize this.) If many bad things are continuing to happen, then the zero point of "how things are right now" will inexorably lead to the world sliding into the deep negative zone. The zero point I've actually been using is "the trajectory the world would be on right now if I were replaced with a random person from my reference class". **That** is something that's within my power to make worse or better (in expectation).

Now go forth, and make the world better than the new zero!