Best of LessWrong: October 2014

Best of LessWrong: October 2014

On Caring

This is an essay describing some of my motivation to be an effective altruist. It is crossposted from my blog. Many of the ideas here are quite similar to others found in the sequences. I have a slightly different take, and after adjusting for the typical mind fallacy I expect that this post may contain insights that are new to many.

1

I'm not very good at *feeling* the size of large numbers. Once you start tossing around numbers larger than 1000 (or maybe even 100), the numbers just seem "big".

Consider Sirius, the brightest star in the night sky. If you told me that Sirius is as big as a million earths, I would feel like that's a lot of Earths. If, instead, you told me that you could fit a *billion* Earths inside Sirius… I would still just feel like that's a lot of Earths.

The feelings are almost identical. *In context*, my brain grudgingly admits that a billion is a lot larger than a million, and puts forth a token effort to feel like a billion-Earth-sized star is bigger than a million-Earth-sized star. But out of context — if I wasn't anchored at "a million" when I heard "a billion" — both these numbers just feel vaguely large.

I feel a *little* respect for the bigness of numbers, if you pick really really large numbers. If you say "one followed by a hundred zeroes", then this feels *a lot* bigger than a billion. But it certainly doesn't feel (in my gut) like it's 10 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 times bigger than a billion. Not in the way that four apples *internally feels* like twice as many as two apples. My brain can't even begin to wrap itself around this sort of magnitude differential.

This phenomena is related to scope insensitivity, and it's important to me because I live in a world where sometimes the things I care about are really really numerous.

For example, billions of people live in squalor, with hundreds of millions of them deprived of basic needs and/or dying from disease. And though most of them are out of my sight, I still care about them.

The loss of a human life with all is joys and all its sorrows is tragic no matter what the cause, and the tragedy is not reduced simply because I was far away, or because I did not know of it, or because I did not know how to help, or because I was not personally responsible.

Knowing this, I care about every single individual on this planet. The problem is, my brain is *simply incapable* of taking the amount of caring I feel for a single person and scaling it up by a billion times. I lack the internal capacity to feel that much. My care-o-meter simply doesn't go up that far.

And this is a problem.

2

It's a common trope that courage isn't about being fearless, it's about being afraid but *doing the right thing anyway*. In the same sense, caring about the world isn't about having a gut feeling that corresponds to the amount of suffering in the world, it's about *doing the right thing anyway*. Even without the feeling.

My internal care-o-meter was calibrated to deal with about a hundred and fifty people, and it *simply can't express* the amount of caring that I have for billions of sufferers. The internal care-o-meter just doesn't go up that high.

Humanity is playing for unimaginably high stakes. At the very least, there are billions of people suffering today. At the worst, there are quadrillions (or more) potential humans, transhumans, or posthumans whose existence depends upon what we do here and now. All the intricate civilizations that the future could hold, the experience and art and beauty that is possible in the future, depends upon the present.

When you're faced with stakes like these, your internal caring heuristics — calibrated on numbers like "ten" or "twenty" — completely fail to grasp the gravity of the situation.

Saving a person's life feels *great*, and it would probably feel just about as good to save one life as it would feel to save the world. It surely wouldn't be *many billion times* more of a high to save the world, because your hardware can't express a feeling a billion times bigger than the feeling of saving a person's life. But even though the altruistic high from saving someone's life would be shockingly similar to the altruistic high from saving the world, always remember that *behind* those similar feelings there is a whole world of difference.

Our internal care-feelings are woefully inadequate for deciding how to act in a world with big problems.

3

There's a mental shift that happened to me when I first started internalizing scope insensitivity. It is a little difficult to articulate, so I'm going to start with a few stories.

Consider Alice, a software engineer at Amazon in Seattle. Once a month or so, those college students will show up on street corners with clipboards, looking ever more disillusioned as they struggle to convince people to donate to Doctors Without Borders. Usually, Alice avoids eye contact and goes about her day, but this month they finally manage to corner her. They explain Doctors Without Borders, and she actually has to admit that it sounds like a pretty good cause. She ends up handing them $20 through a combination of guilt, social pressure, and altruism, and then rushes back to work. (Next month, when they show up again, she avoids eye contact.)

Now consider Bob, who has been given the Ice Bucket Challenge by a friend on facebook. He feels too busy to do the ice bucket challenge, and instead just donates $100 to ALSA.

Now consider Christine, who is in the college sorority ΑΔΠ. ΑΔΠ is engaged in a competition with ΠΒΦ (another sorority) to see who can raise the most money for the National Breast Cancer Foundation in a week. Christine has a competitive spirit and gets engaged in fund-raising, and gives a few hundred dollars herself over the course of the week (especially at times when ΑΔΠ is especially behind).

All three of these people are donating money to charitable organizations… and that's great. But notice that there's something similar in these three stories: these donations are largely motivated by a *social context*. Alice feels obligation and social pressure. Bob feels social pressure and maybe a bit of camaraderie. Christine feels camaraderie and competitiveness. These are all fine motivations, but notice that these motivations are related to the *social setting*, and only tangentially to the *content* of the charitable donation.

If you took any of Alice or Bob or Christine and asked them why they aren't donating *all* of their time and money to these causes that they apparently believe are worthwhile, they'd look at you funny and they'd probably think you were being rude (with good reason!). If you pressed, they might tell you that money is a little tight right now, or that they would donate more if they were a better person.

But the question would still feel kind of *wrong*. Giving all your money away is just not what you do with money. We can all *say out loud* that people who give all their possessions away are really great, but behind closed doors we all know that people are crazy. (Good crazy, perhaps, but crazy all the same.)

This is a mindset that I inhabited for a while. There's an alternative mindset that can hit you like a freight train when you start internalizing scope insensitivity.

4

Consider Daniel, a college student shortly after the Deepwater Horizon BP oil spill. He encounters one of those college students with the clipboards on the street corners, soliciting donations to the World Wildlife Foundation. They're trying to save as many oiled birds as possible. Normally, Daniel would simply dismiss the charity as Not The Most Important Thing, or Not Worth His Time Right Now, or Somebody Else's Problem, but this time Daniel has been thinking about how his brain is bad at numbers and decides to do a quick sanity check.

He pictures himself walking along the beach after the oil spill, and encountering a group of people cleaning birds as fast as they can. They simply don't have the resources to clean all the available birds. A pathetic young bird flops towards his feet, slick with oil, eyes barely able to open. He kneels down to pick it up and help it onto the table. One of the bird-cleaners informs him that they won't have time to get to that bird themselves, but he could pull on some gloves and could probably save the bird with three minutes of washing.

Daniel decides that he *would* spend three minutes of his time to save the bird, and that he would *also* be happy to pay at least $3 to have someone else spend a few minutes cleaning the bird. He introspects and finds that this is not just because he imagined a bird right in front of him: he feels that it is *worth* at least three minutes of his time (or $3) to save an oiled bird in some vague platonic sense.

And, because he's been thinking about scope insensitivity, he *expects* his brain to misreport how much he actually cares about large numbers of birds: the internal feeling of caring can't be expected to line up with the actual importance of the situation. So instead of just *asking his gut* how much he cares about de-oiling lots of birds, he shuts up and multiplies.

Thousands and thousands of birds were oiled by the BP spill alone. After shutting up and multiplying, Daniel realizes (with growing horror) that the amount he *acutally* cares about oiled birds is lower bounded by two months of hard work and/or fifty thousand dollars. And that's not even counting wildlife threatened by other oil spills.

And if he cares that much about *de-oiling birds*, then how much does he actually care about factory farming, nevermind hunger, or poverty, or sickness? How much does he actually care about wars that ravage nations? About neglected, deprived children?

About the future of humanity? He *actually* cares about these things to the tune of much more money than he has, and much more time than he has.

For the first time, Daniel sees a glimpse of of how much he actually cares, and how poor a state the world is in.

This has the strange effect that Daniel's reasoning goes full-circle, and he realizes that he actually *can't* care about oiled birds to the tune of 3 minutes or $3: not because the birds aren't *worth* the time and money (and, in fact, he thinks that the economy produces things priced at $3 which are worth less than the bird's survival), but because he can't spend *his* time or money on saving the birds. The opportunity cost suddenly seems far too high: there is *too much else to do!* People are sick and starving and dying! The very future of our civilization is at stake!

Daniel doesn't wind up giving $50k to the WWF, and he also doesn't donate to ALSA or NBCF. But if you ask *Daniel* why he's not donating all his money, he won't look at you funny or think you're rude. He's left the place where you don't care far behind, and has realized that *his mind was lying to him the whole time* about the gravity of the real problems.

Now he realizes that he *can't possibly do enough*. After adjusting for his scope insensitivity (and the fact that his brain lies about the size of large numbers), even the "less important" causes like the WWF suddenly seem worthy of dedicating a life to. Wildlife destruction and ALS and breast cancer are suddenly all problems that he would *move mountains* to solve — except he's finally understood that there are just too many mountains, and ALS isn't the bottleneck, and AHHH HOW DID ALL THESE MOUNTAINS GET HERE?

In the original mindstate, the reason he didn't drop everything to work on ALS was because it just didn't seem… pressing enough. Or tractable enough. Or important enough. Kind of. These are sort of the reason, but the real reason is more that the concept of "dropping everything to address ALS" never even *crossed his mind* as a real possibility. The idea was too much of a break from the standard narrative. It wasn't his problem.

In the new mindstate, *everything* is his problem. The only reason he's not dropping everything to work on ALS is because there are far too many things to do first.

Alice and Bob and Christine usually aren't spending time solving all the world's problems because they forget to see them. If you remind them — put them in a social context where they remember how much they care (hopefully without guilt or pressure) — then they'll likely donate a little money.

By contrast, Daniel and others who have undergone the mental shift aren't spending time solving all the world's problems because there are *just too many problems*. (Daniel hopefully goes on to discover movements like effective altruism and starts contributing towards fixing the world's most pressing problems.)

5

I'm not trying to preach here about how to be a good person. You don't need to share my viewpoint to be a good person (obviously).

Rather, I'm trying to point at a shift in perspective. Many of us go through life understanding that we *should* care about people suffering far away from us, but failing to. I think that this attitude is tied, at least in part, to the fact that most of us implicitly trust our internal care-o-meters.

The "care feeling" isn't usually strong enough to compel us to frantically save everyone dying. So while we acknowledge that it would be *virtuous* to do more for the world, we think that we *can't*, because we weren't gifted with that virtuous extra-caring that prominent altruists must have.

But this is an error — prominent altruists aren't the people who have a larger care-o-meter, they're the people who have *learned not to trust their care-o-meters*.

Our care-o-meters are broken. They don't work on large numbers. Nobody has one capable of faithfully representing the scope of the world's problems. But the fact that you can't *feel* the caring doesn't mean that you can't *do* the caring.

You don't get to feel the appropriate amount of "care", in your body. Sorry — the world's problems are just too large, and your body is not built to respond appropriately to problems of this magnitude. But if you choose to do so, you can still *act* like the world's problems are as big as they are. You can stop trusting the internal feelings to guide your actions and switch over to manual control.

6

This, of course, leads us to the question of "what the hell do you then?"

And I don't really know yet. (Though I'll plug the Giving What We Can pledge, GiveWell, MIRI, and The Future of Humanity Institute as a good start).

I think that at least part of it comes from a certain sort of desperate perspective. It's not enough to think you *should* change the world — you also need the sort of desperation that comes from realizing that you would dedicate your entire life to solving the world's 100th biggest problem if you could, but you can't, because there are 99 bigger problems you have to address first.

I'm not trying to guilt you into giving more money away — becoming a philanthropist is *really really hard*. (If you're *already* a philanthropist, then you have my acclaim and my affection.) First it requires you to have money, which is uncommon, and then it requires you to *throw that money at distant invisible problems*, which is not an easy sell to a human brain. Akrasia is a formidable enemy. And most importantly, guilt doesn't seem like a good long-term motivator: if you want to join the ranks of people saving the world, I would rather you join them proudly. There are many trials and tribulations ahead, and we'd do better to face them with our heads held high.

7

Courage isn't about being fearless, it's about being able to do the right thing even if you're afraid.

And similarly, addressing the major problems of our time isn't about feeling a strong compulsion to do so. It's about doing it anyway, even when internal compulsion utterly fails to capture the scope of the problems we face.

It's easy to look at especially virtuous people — Gandhi, Mother Theresa, Nelson Mandela — and conclude that they must have cared more than we do. But I don't think that's the case.

Nobody gets to comprehend the scope of these problems. The closest we can get is doing the multiplication: finding something we care about, putting a number on it, and multiplying. And then trusting the numbers more than we trust our feelings.

Because our feelings lie to us.

When you do the multiplication, you realize that addressing global poverty and building a brighter future deserve more resources than currently exist. There is not enough money, time, or effort in the world to do what we need to do.

There is only you, and me, and everyone else who is trying anyway.

8

You can't actually feel the weight of the world. The human mind is not capable of that feat.

But sometimes, you can catch a glimpse.

Crossing the History-Lessons Threshold

(1)

Around 2009, I embarked on being a serious amateur historian. I wouldn't have called it that at the time, but since then, I've basically nonstop studied various histories.

The payoffs of history come slow at first, and then fast. History is often written as a series of isolated events, and events are rarely put in total context. You can easily draw a straight line from Napoleon's invasions of the fragmented German principalities to how Bismarck and Moltke were able to unify a German Confederation under Prussian rule a few decades later; from there, it's a straight line to World War I due to great power rivalry; the Treaty of Versailles is easily understood in retrospect by historical French/German enmity; this gives rise to World War II.

That series of events is hard enough to *truly* get one's mind around, not just in abstract academic terms, but in actually getting a *feel* of how and why the actors did what they did, which shaped the outcomes that built the world.

And that's only the start of it: once you can flesh out the rest of the map, history starts coming brilliantly alive.

Without Prime Minister Stolypin's assassination in 1911, likely the Bolsheviks don't succeed in Russia; without that, Stalin is not at the helm when the Nazis invade.

On the other side of the Black Sea, in 1918, the Ottoman Empire is having terms worse than the Treaty of Versailles imposed on it -- until Mustafa Kemal leads the Turkish War of Independence, building one of the most stable states in the Middle East. Turkey, following Kemal's skill at governance and diplomacy, is able to (with great difficulty) stay neutral in World War II, not be absorbed by the Soviets, and not have its government taken over by hard-line Muslims.

This was not-at-all an obvious course of events. Without Kemal, Turkey almost certainly becomes crippled under the Treaty of Sevres, and eventually likely winds up as a member of the Axis during World War II, or gets absorbed as another Soviet/Warsaw Pact satellite state.

The chain of events goes on and on. There is an eminently clear chain of events from Martin Luther at Worms in 1521 to the American Revolution. Meanwhile, the non-success the Lord Protectorate and Commonwealth of England turned out less promising than was hoped -- ironically, arguably predisposing England to being less sympathetic to greater democracy. But the colonies were shielded from this, and their original constitutions and charters were never amended in the now-becoming-more-disenchanted-with-democracy England. Following a lack of consistent colonial policy and a lot of vacillating by various British governments, the American Revolution happens, and Britain loses control of the land and people would come to supplant it as the dominant world power one and a half centuries later.

(2)

Until you can start seeing the threads and chains of history across nations, interactions, and long stretches of time, history is a set of often-interesting stories -- but the larger picture remains blurry and out-of-focus. The lessons come once you can synthesize it all.

Hideyoshi Toyotomi's 1588 sword hunt was designed to take away weapons and chances of rebellious factions overthrowing his unified government of Japan. The policy was continued by his successor after the Toyotomi/Tokugawa Civil War, which leads to the Tokugawa forces losing to the Imperial Restoration in 1868 as their skill at warfare had atrophied; common soldiers with Western artillery were able to out-combat samurai with obsolete weapons.

Nurhaci founded the Qing Dynasty around the time Japan was being unified, with a mix of better command structures and tactics. But the dynasty hardened into traditionalism and was backwards-looking when Western technology and imperialists came with greater frequency in the late 1800's. The Japanese foreign minister Ito Hirobumi offered to help the Qing modernize along the lines Imperial Japan had modernized while looking for a greater alliance with the Chinese. But, Empress Dowager Cixi arrests and executes the reform-minded ministers of Emperor Guangxu and later, most likely, poisoned the reform-minded Emperor Guangxu. (He died of arsenic poisoning when Cixi was on her deathbed; someone poisoned him; Cixi or someone acting under her orders is the most likely culprit.)

The weak Qing Dynasty starts dealing with ever-more-frequent invasions, diplomatic extortions, and rebellions and revolutions. The Japanese invade China a generation after Hirobumi was rebuffed, and the Qing Dynasty entirely falls apart. After the Japanese unconditional surrender, the Chinese Civil War starts; the Communists win.

(3)

From this, we can start drawing lessons and tracing histories, seeing patterns. We start to see how things could have broken differently. Perhaps Germany and France were doomed to constant warfare due to geopolitics; maybe this is true.

But certainly, it's not at all obvious that Mustafa Kemal would lead the ruins of the Ottoman Empire into modern Turkey, and (seemingly against overwhelming odds) keep neutrality during World War II, rebuff Stalin and stay removed from Soviet conquest, and maintain a country with secular and modern laws that honors Muslim culture without giving way to warlordism as happened to much of the rest of the Middle East.

Likewise, we can clearly see how the policies of Empress Dowager Cixi ended the chance for a pan-East-Asian alliance, trade bloc, or federation; it's not inconceivable to imagine a world today were China and Japan are incredibly close allies, and much of the world's centers of commerce, finance, and power are consolidated in a Tokyo-Beijing-Seoul alliance. Sure, it's inconceivable with hindsight, but Japan in 1910 and Japan in 1930 are very different countries; and the struggling late Qing Dynasty is different than the fledgling competing factions in China after the fall of the Qing.

We can see, observing historical events from broad strokes, the huge differences individuals can make at leveraged points, the eventual outcomes in Turkey and East Asia were not-at-all foreordained by geography, demographics, or trends.

(4)

Originally, I was sketching out some of these trends of history to make a larger point about how modern minds have a hard time understanding older governments -- in a world where "personal rule" is entirely rebuffed in the more developed countries, it is hard to imagine how the Qing Dynasty or Ottoman Empire actually functioned. The world after the Treaty of Westphalia is incredibly different than the world before it, and the world before strict border controls pre-WWI is largely unrecognizable to us.

That was the piece I was going to write, about how we project modern institutions and understandings backwards, and how that means we can't understand what actually happened. The Ottomans and Qing were founded before modern nationalism had emerged, and the way their subjects related to them is so alien to us that it's almost impossible to conceive of how their culture and governance actually ran.

(5)

I might still pen that piece, if there's interest in it -- my attempt at a brief introduction came to result in this very different one, focused on a different particular point: the threshold effect in learning history.

I would say there's broadly three thresholds:

The first looks at a series of isolated events. You wind up with some witty quips, like: Astor saying, "Sir, if you were my husband, I would poison your drink." Churchill: "If I were married to you, I'd drink it."

Or moments of great drama: "And so the die is cast." "Don't fire until you see the whites of their eyes." "There is nothing to fear except fear itself."

These aren't so bad to learn; they're an okay jumping-off place. Certainly, Caesar's decision to march on Rome, Nobunaga's speech before the Battle of Okehazama, or understanding why Washington made the desperate gamble to cross the Delaware all offerlessons.

But seeing how the Marian military reforms, Sulla's purges, and the Gracchi brothers created the immediate situation before Julius Caesar's fateful crossing is more interesting, and tracing the lines backwards, seeing how Rome's generations-long combat with Hannibal's Carthage turned the city-state into a fully militarized conquest machine, and then following the lines onwards to see how the Romans relied on unit cohesion which, once learned by German adversaries, led to the fall of Rome -- this is much more interesting.

That's the second threshold of history to me: when isolated events start becoming regional chains; that's tracing Napoleon's invasion of Germany to Bismarck to the to World War I to the Treaty of Versailles to WWII.

Some people get to this level of history, and it makes you quickly an expert in a particular country.

But I think that's a poor place to stop learning: if you can truly get your mind around a long stretch of time in a nation, it's time to start coloring the map. When you can broadly know how Korea is developing simultaneous with Japan; how the Portugese/Spanish rivalry and Vatican compromises are affecting Asia's interactions with the Age of Sail Westerners; how Protestantism is creating rivals to Catholic power, two of which later equip the Japanese's Imperial Faction, which kicks off the

Asian side of World War II -- this is when history starts really paying dividends and teaching worthwhile lessons.

The more you get into it, the more there is to learn. Regions that don't get much historical interest from Americans like Tito's Yugoslavia become fascinating to look at how they stayed out of Soviet Control and played the Western and Eastern blocs against each other; the chain of events takes a sad turn when Tito's successors can't keep the country together, the Yugoslav Wars follow, and its successor states still don't have the levels of relative prosperity and influence that Yugoslavia had in its heyday.

Yugoslavia is hard to get one's mind around by itself, but it's easy to color the map in with a decent understanding of Turkey, Germany, and Russia. Suddenly, figures and policies and conflicts and economics and culture start coming alive; lessons and patterns are everywhere.

I don't read much fiction any more, because most fiction can't compete with the sheer weight, drama, and insightfulness of history. Apparently some Kuomintang soldiers held out against the Chinese Communists and fought irregular warfare while funding their conflicts with heroin production in the regions of Burma and Thailand -- I just got a book on it, further coloring in the map of the aftermath of the Chinese Civil War, and that aspect of it upon the backdrop of the Cold War and containment, and how the Sino/Soviet split led to America normalizing relations with China, and...

...it never ends, and it's been one of the most insightful areas of study across my life.

History in that first threshold -- isolated battles, quotes, the occasional drama -- frankly, it offers only a slight glimmer of what's possible to learn.

Likewise, the second level of knowing a particular country's rise and fall over time can be insightful, but I would encourage anyone that has delved into history that much to not stop there: you're not far from the gates unlocking to large wellsprings of knowledge, a nearly infinite source of ideas, inspiration, case studies, and all manner of other sources of new and old ideas and very practical guidance.

How to write an academic paper, according to me

Disclaimer: this is entirely a personal viewpoint, formed by a few years of publication in a few academic fields. **EDIT:** Many of the comments are very worth reading as well.

Having recently finished a very rushed submission (turns out you can write a novel paper in a day and half, if you're willing to sacrifice quality and sanity), I've been thinking about how academic papers are structured - and more importantly, how they should be structured.

It seems to me that the key is to consider the audience. Or, more precisely, to consider the audiences - because different people will read you paper to different depths, and you should cater to all of them. An example of this is the "inverted pyramid" structure for many news articles - start with the salient facts, then the most important details, then fill in the other details. The idea is to ensure that a reader who stops reading at any point (which happens often) will nevertheless have got the most complete impression that it was possible to convey in the bit that they did read.

So, with that model in mind, lets consider the different levels of audience for a general academic paper (of course, some papers just can't fit into this mould, but many can):

Title readers

The least important audience. An interesting title may draw casual browsers in, but those likely aren't very valuable readers. Most people encountering an academic article will either be looking for it, or will have had it referred to them from some source. They will likely read more of it. So the main role of the title is to not put off these readers, and to clarify what the paper is about, and what field it belongs in. Witty titles are perfectly acceptable, as long as it fulfils those criteria. So in-jokes for the whole academic field are perfectly acceptable, in-jokes for a narrow subfield are not - unless you're not aiming beyond that subfield.

Abstract readers

The most important audience of all. Most people reading a paper will only read the abstract, and will then proceed to dismiss the paper or accept it and move on. The abstract thus plays three roles:

1. It presents the paper's results. The abstract must be crystal-clear on what the paper says; abstract readers must be able to describe the results correctly.
2. It establishes the credibility of the result. It can do this by briefly outlying the methods used, and by its general tone. It must thus be serious, and use the correct vocabulary for the field. No room for impressive rhetoric here - dry and descriptive is the model of the abstract.

3. It can draw the reader into looking into the paper proper. Because of the first two points, it cannot achieve this by teasers or rhetoric. Instead it must present strong results that cause the reader to want to read more.

Skimmers

This audience will skim through the paper to see what it says. Most crucial for them is the introduction and, depending on the field, possibly the conclusion or discussion section. These must tell the skimmers everything there is to know about the paper - what the problem is, what the results are, what methods were used, why these results are valid, why they are important. As long as all these points are covered, rhetoric and wit can be used, in moderation, to make the reading more enjoyable and salient. But be careful to use these in moderation, lest you give the impression that the paper's results depend on rhetorical tricks. Rhetoric is the flavouring, giving out the information above is the main goal.

Full readers

These are those readers who will go through the whole paper, though they may skim some parts along the way. The important thing here is to get the structure absolutely clear - it must be easy for them to see what the crucial steps or arguments are, what implies what, what relies on what. To do this, lay out the structure of the argument and of the paper clearly in the introduction or in the second section. Emphasise the important results through the paper (consider the layout for this, it can often be used to draw attention to the main points), and connect them together ("combining this with the results of section 2.3x.iii..."). Some rhetoric can be used around these important results, especially if it emphasises their importance.

Deep readers

These are your greatest fans or your more hated critics. They will go through the whole paper, taking your argument apart to understand it completely and figure out how it ticks. No fancy rhetoric for them, just careful attention to detail, clarity, and rigour. In mathematical terms, these are the people who will be reading the proofs of your minor lemmas. Don't waste space with anything that doesn't help you establish your argument or your results. These are the lawyers among your readers, looking for the tiniest of flaws. Don't give them any of these, and don't try to hide them with weak arguments.

Writing the paper

The different audiences above give a structure to the paper, but they can also give a structure to writing process. Looking back, I realise that I start by writing for the full readers, getting the important points and structure correct. Then I fill in the details for the deep readers. I then write the introduction (and conclusion, if appropriate) for the skimmers, and conclude with the abstract for the most important audience. The title can be chosen at any point in this process.

Hope this helps! I think I've been following this advice implicitly for a long time, and it's got me a few publications. Feel free to ignore it, of course, or to post your own preferred approach.

Is the potential astronomical waste in our universe too small to care about?

In the not too distant past, people thought that our universe might be capable of supporting an unlimited amount of computation. Today our best guess at the cosmology of our universe is that it stops being able to support any kind of life or deliberate computation after a finite amount of time, during which only a finite amount of computation can be done (on the order of something like 10^120 operations).

Consider two hypothetical people, Tom, a total utilitarian with a near zero discount rate, and Eve, an egoist with a relatively high discount rate, a few years ago when they thought there was .5 probability the universe could support doing at least 3^^^3 ops and .5 probability the universe could only support 10^120 ops. (These numbers are obviously made up for convenience and illustration.) It would have been mutually beneficial for these two people to make a deal: if it turns out that the universe can only support 10^120 ops, then Tom will give everything he owns to Eve, which happens to be $1 million, but if it turns out the universe can support 3^^^3 ops, then Eve will give $100,000 to Tom. (This may seem like a lopsided deal, but Tom is happy to take it since the potential utility of a universe that can do 3^^^3 ops is so great for him that he really wants any additional resources he can get in order to help increase the probability of a positive Singularity in that universe.)

You and I are not total utilitarians or egoists, but instead are people with moral uncertainty. Nick Bostrom and Toby Ord proposed the Parliamentary Model for dealing with moral uncertainty, which works as follows:

> Suppose that you have a set of mutually exclusive moral theories, and that you assign each of these some probability. Now imagine that each of these theories gets to send some number of delegates to The Parliament. The number of delegates each theory gets to send is proportional to the probability of the theory. Then the delegates bargain with one another for support on various issues; and the Parliament reaches a decision by the delegates voting. What you should do is act according to the decisions of this imaginary Parliament.

It occurred to me recently that in such a Parliament, the delegates would makes deals similar to the one between Tom and Eve above, where they would trade their votes/support in one kind of universe for votes/support in another kind of universe. If I had a Moral Parliament active back when I thought there was a good chance the universe could support unlimited computation, all the delegates that really care about astronomical waste would have traded away their votes in the kind of universe where we actually seem to live for votes in universes with a lot more potential astronomical waste. So today my Moral Parliament would be effectively controlled by delegates that care little about astronomical waste.

I actually still seem to care about astronomical waste (even if I pretend that I was *certain* that the universe could only do at most 10^120 operations). (Either my Moral Parliament wasn't active back then, or my delegates weren't smart enough to make the appropriate deals.) Should I nevertheless follow UDT-like reasoning and conclude that I should act as if they had made such deals, and therefore I should stop caring about the relatively small amount of astronomical waste that could occur in our universe? If the answer to this question is "no", what about the future going forward, given that there is still uncertainty about cosmology and the nature of physical computation. Should the delegates to my Moral Parliament be making these kinds of deals from now on?

Anthropic signature: strange anti-correlations

Imagine that the only way that civilization could be destroyed was by a large pandemic that occurred at the same time as a large recession, so that governments and other organisations were too weakened to address the pandemic properly.

Then if we looked at the past, as observers in a non-destroyed civilization, what would we expect to see? We could see years with no pandemics or no recessions; we could see mild pandemics, mild recessions, or combinations of the two; we could see large pandemics with no or mild recessions; or we could see large recessions with no or mild pandemics. We wouldn't see large pandemics combined with large recessions, as that would have caused us to never come into existence. These are the only things ruled out by anthropic effects.

Assume that pandemics and recessions are independent (at least, in any given year) in terms of "objective" (non-anthropic) probabilities. Then what would we see? We would see that pandemics and recessions appear to be independent when either of them are of small intensity. But as the intensity rose, they would start to become anti-correlated, with a large version of one completely precluding a large version of the other.

The effect is even clearer if we have a probabilistic relation between pandemics, recessions and extinction (something like: extinction risk proportional to product of recession size times pandemic size). Then we would see an anti-correlation rising smoothly with intensity.

Thus one way of looking for anthropic effects in humanity's past is to look for different classes of incidents that are uncorrelated at small magnitude, and anti-correlated at large magnitudes. More generally, to look for different classes of incidents where the correlation changes at different magnitudes - without any obvious reasons. Than might be the signature of an anthropic disaster we missed - or rather, that missed us.

A discussion of heroic responsibility

[Originally posted to my personal blog, reposted here with edits.]

Introduction

> You could call it heroic responsibility, maybe," Harry Potter said. "Not like the usual sort. It means that whatever happens, no matter what, it's always your fault. Even if you tell Professor McGonagall, she's not responsible for what happens, you are. Following the school rules isn't an excuse, someone else being in charge isn't an excuse, even trying your best isn't an excuse. There just aren't any excuses, you've got to get the job done no matter what." Harry's face tightened. "That's why I say you're not thinking responsibly, Hermione. Thinking that your job is done when you tell Professor McGonagall—that isn't heroine thinking. Like Hannah being beat up is okay then, because it isn't your fault anymore. Being a heroine means your job isn't finished until you've done whatever it takes to protect the other girls, permanently." In Harry's voice was a touch of the steel he had acquired since the day Fawkes had been on his shoulder. "You can't think as if just following the rules means you've done your duty. – HPMOR, chapter 75.

I like this concept. It counters a particular, common, harmful failure mode, and that it's an amazingly useful thing for a lot of people to hear. I even think it was a useful thing for me to hear a year ago.

But... I'm not sure about this yet, and my thoughts about it are probably confused, but I think that there's a version of Heroic Responsibility that you can get from reading this description, that's maybe even the default outcome of reading this description, that's also a harmful failure mode.

Something Impossible

A wrong way to think about heroic responsibility

I dealt with a situation at work a while back–May 2014 according to my journal. I had a patient for five consecutive days, and each day his condition was a little bit worse. Every day, I registered with the staff doctor my feeling that the current treatment was Not Working, and that maybe we ought to try something else. There were lots of complicated medical reasons why his decisions were constrained, and why 'let's wait and see' was maybe the best decision, statistically speaking–that in a majority of possible worlds, waiting it out would lead to better outcomes than one of the potential more aggressive treatments, which came with side effects. And he wasn't actually ignoring me; he would listen patiently to all my concerns. Nevertheless, he wasn't the one watching the guy writhe around in bed, uncomfortable and delirious, for twelve hours every day, and I felt ignored, and I was pretty frustrated.

On day three or four, I was listening to Ray's Solstice album on my break, and the song 'Something Impossible' came up.

> Bold attempts aren't enough, roads can't be paved with intentions...
> You probably don’t even got what it takes,
> But you better try anyway, for everyone's sake
> And you won’t find the answer until you escape from the
> Labyrinth of your conventions.
> Its time to just shut up, and do the impossible.
> Can’t walk away...
> Gotta break off those shackles, and shake off those chains
> Gotta make something impossible happen today...

It hit me like a load of bricks–this whole thing was stupid and rationalists should win. So I spent my entire break talking on Gchat with one of my CFAR friends, trying to see if he could help me come up with a suggestion that the doctor would agree was good. This wasn’t something either of us were trained in, and having something to protect doesn't actually give you superpowers, and the one creative solution I came up with was worse than the status quo for several obvious reasons.

I went home on day four feeling totally drained and having asked to please have a different patient in the morning. I came in to find that the patient had nearly died in the middle of the night. (He was now intubated and sedated, which wasn’t great for him but made my life a hell of a lot easier.) We eventually transferred him to another hospital, and I spent a while feeling like I’d personally failed.

I’m not sure whether or not this was a no-win scenario even in theory. But I don't think I, personally, could have done anything with greater positive expected value. There's a good reason why a doctor with 10 years of school and 20 years of ICU experience can override a newly graduated nurse's opinion. In most of the possible worlds, the doctor is right and I'm wrong. Pretty much the only thing that I could have done better would have been to care less–and thus be less frustrated and more emotionally available to comfort a guy who was having the worst week of his life.

In short, I fulfilled my responsibilities to my patient. Nurses have a lot of responsibilities to their patients, well specified in my years of schooling and in various documents published by the College of Nurses of Ontario. But nurses aren’t expected or supposed to take heroic responsibility for these things.

I think that overall, given a system that runs on humans, that's a good thing.

The Well-Functioning Gear

> I feel like maybe the hospital is an emergent system that has the property of patient-healing, but I’d be surprised if any one part of it does.
>
> Suppose I see an unusual result on my patient. I don’t know what it means, so I mention it to a specialist. The specialist, who doesn’t know anything about the patient beyond what I’ve told him, says to order a technetium scan. He has no idea what a technetium scan is or how it is performed, except that it’s the proper thing to do in this situation. A nurse is called to bring the patient to the scanner,

> but has no idea why. The scanning technician, who has only a vague idea why the scan is being done, does the scan and spits out a number, which ends up with me. I bring it to the specialist, who gives me a diagnosis and tells me to ask another specialist what the right medicine for that is. I ask the other specialist – who has only the sketchiest idea of the events leading up to the diagnosis – about the correct medicine, and she gives me a name and tells me to ask the pharmacist how to dose it. The pharmacist – who has only the vague outline of an idea who the patient is, what test he got, or what the diagnosis is – doses the medication. Then a nurse, who has no idea about any of this, gives the medication to the patient. Somehow, the system works and the patient improves.
>
> Part of being an intern is adjusting to all of this, losing some of your delusions of heroism, getting used to the fact that you're not going to be Dr. House, that you are at best going to be a very well-functioning gear in a vast machine that does often tedious but always valuable work. –Scott Alexander

The medical system does a hard thing, and it might not do it well, but it *does* it. There is too much complexity for any one person to have a grasp on it. There are dozens of mutually incomprehensible specialties. And the fact that [insert generic nurse here] doesn't have the faintest idea how to measure electrolytes in blood, or build an MRI machine, or even what's going on with the patient next door, is a feature, not a bug.

The medical system doesn't run on exceptional people–it runs on average people, with predictably average levels of skill, slots in working memory, ability to notice things, ability to not be distracted thinking about their kid's problems at school, etc. And it doesn't run under optimal conditions; it runs under average conditions. Which means working overtime at four am, short staffing, three patients in the ER waiting for ICU beds, etc.

Sure, there are problems with the machine. The machine is inefficient. The machine doesn't have all the correct incentives lined up. The machine does need fixing–but I would argue that from within the machine, as one of its parts, taking heroic responsibility for your own sphere of control isn't the way to go about fixing the system.

As an [insert generic nurse here], my sphere of control is the four walls of my patient's room. Heroic responsibility for my patient would mean...well, optimizing for them. In the most extreme case, it might mean killing the itinerant stranger to obtain a compatible kidney. In the less extreme case, I spend all my time giving my patient great care, instead of helping the nurse in the room over, whose patient is much sicker. And then sometimes my patient will die, and there will be literally nothing I can do about it, their death was causally set in stone twenty-four hours before they came to the hospital.

I kind of predict that the results of installing heroic responsibility as a virtue, among average humans under average conditions, would be a) everyone stepping on everyone else's toes, and b) 99% of them quitting a year later.

Recursive Heroic Responsibility

If you're a gear in a machine, and you notice that the machine is broken, your options are a) be a really good gear, or b) take heroic responsibility for your sphere of control, and probably break something...but that's a false dichotomy. Humans are very flexible tools, and there are also infinite *other* options, including "step out of the machine, figure out who's in charge of this shit, and get it fixed."

You can't take responsibility for the individual case, but you can for the system-level problem, the long view, the one where people eat badly and don't exercise and at age fifty, morbidly obese with a page-long medical history, they end up as a slow-motion train wreck in an ICU somewhere. Like in poker, you play to win money–positive EV–not to win hands. Someone’s going to be the Minister of Health for Canada, and they’re likely to be in a position where taking heroic responsibility for the Canadian health care system makes things better. And probably the current Minister of Health isn’t being strategic, isn’t taking the level of responsibility that they could, and the concept of heroic responsibility would be the best thing for them to encounter.

So as an [insert generic nurse here], working in a small understaffed ICU, watching the endless slow-motion train wreck roll by...maybe the actual meta-level right thing to do is to *leave,* and become the freaking Minister of Health, or befriend the current one and introduce them to the concept of being strategic.

But it's fairly obvious that that isn't the right action for *all* the nurses in that situation. I'm wary of advice that doesn't generalize. What's difference between the nurse who should leave in order to take meta-level responsibility, and the nurse who should stay because she's needed as a gear?

Heroic responsibility for average humans under average conditions

I can predict at least *one* thing that people will say in the comments, because I've heard it hundreds of times–that Swimmer963 is a clear example of someone who should leave nursing, take the meta-level responsibility, and do something higher impact for the usual. Because she's smart. Because she's rational. Whatever.

Fine. This post isn't *about* me. Whether I like it or not, the concept of heroic responsibility is now a part of my value system, and I probably am going to leave nursing.

But what about the other nurses on my unit, the ones who are competent and motivated and curious and really *care?* Would familiarity with the concept of heroic responsibility help or hinder them in their work? Honestly, I predict that they would feel alienated, that they would assume I held a low opinion of them (which I *don't,* and I *really don't* want them to think that I do), and that they would flinch away and go back to the things that they were doing anyway, the role where they were comfortable–or that, if they did accept it, it would cause them to burn out. So as a consequentialist, I'm not going to tell them.

And yeah, that bothers me. Because I'm not a special snowflake. Because I want to live in a world where rationality helps *everyone*. Because I feel like the reason they would react that was isn't because of anything about them as people, or because heroic responsibility is a bad thing, but because I'm not able to communicate to them what I mean. Maybe stupid reasons. Still bothers me.

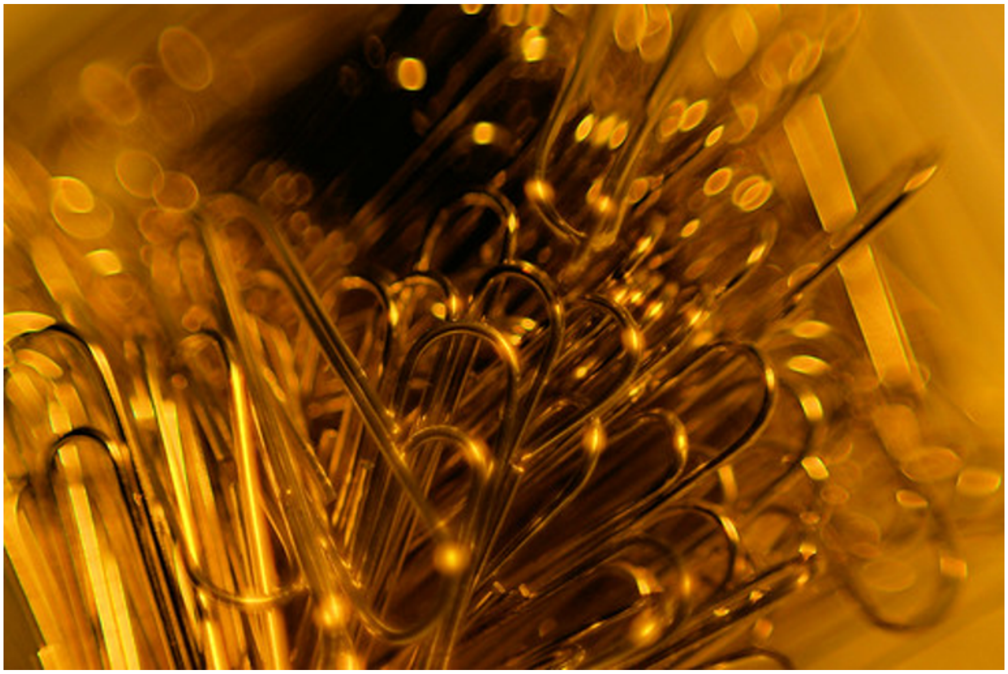
Maybe you want to maximise paperclips too

As most LWers will know, Clippy the Paperclip Maximiser is a superintelligence who wants to tile the universe with paperclips. The LessWrong wiki entry for Paperclip Maximizer says that:

> The goal of maximizing paperclips is chosen for illustrative purposes because it is very unlikely to be implemented

I think that a massively powerful star-faring entity - whether a Friendly AI, a far-future human civilisation, aliens, or whatever - might indeed end up essentially converting huge swathes of matter in to paperclips. Whether a massively powerful star-faring entity is *likely* to arise is, of course, a separate question. But if it does arise, it could well want to tile the universe with paperclips.

Let me explain.



To travel across the stars and achieve whatever noble goals you might have (assuming they scale up), you are going to want energy. A lot of energy. Where do you get it? Well, at interstellar scales, your only options are nuclear fusion or maybe fission.

Iron has the strongest binding energy of any nucleus. If you have elements lighter than iron, you can release energy through nuclear fusion - sticking atoms together to make bigger ones. If you have elements heavier than iron, you can release energy through nuclear fission - splitting atoms apart to make smaller ones. We can do this now for a handful of elements (mostly selected isotopes of uranium, plutonium and hydrogen) but we don’t know how to do this for most of the others - yet. But it looks thermodynamically possible. So if you are a massively powerful *and* massively clever galaxy-hopping agent, you can extract maximum energy for your purposes by taking up all the non-ferrous matter you can find and turning it in to iron, getting energy through fusion or fission as appropriate.

You leave behind you a cold, dark trail of iron.

That seems a little grim. If you have any aesthetic sense, you might want to make it prettier, to leave an enduring sign of values beyond mere energy acquisition. With careful engineering, it would take only a tiny, tiny amount of extra effort to leave the iron arranged in to beautiful shapes. Curves are nice. What do you call a lump of iron arranged in to an artfully-twisted shape? I think we could reasonably call it a paperclip.

Over time, the amount of space that you've visited and harvested for energy will increase, and the amount of space available for your noble goals - or for anyone else's - will decrease. Gradually but steadily, you are converting the universe in to artfully-twisted pieces of iron. To an onlooker who doesn't see or understand your noble goals, you will look a lot like you are a paperclip maximiser. In Eliezer's terms, your desire to do so is an instrumental value, not a terminal value. But - conditional on my wild speculations about energy sources here being correct - it's what you'll do.

Self-signaling the ability to do what you want

This is a linkpost for https://mindingourway.com/self-signaling-the-ability-to-do-what-you-want/

1

In college, I would often find that I had just a little bit too much food. Either I'd cooked too much or the food I'd ordered was just a bit too large, or whatever.

I'm sure many of you are familiar with the feeling of having four bites of food left, wanting roughly one more bite, but knowing that three bites is too few to justify saving the food for later.

(Then you either apply lots of willpower to save the food for later, or you take another bite, realize that there isn't enough food left to save, and proceed to stuff yourself.)

This is pretty much a standard instance of the sunk cost fallacy, where reasoning of the form "I can't just *not eat the food*, because I already paid for it" neglects the fact that the costs are already sunk. In these scenarios, the *only* consideration should be whether or not eating the rest of the food is be better than throwing it away. Your money, which is gone no matter what you choose, shouldn't factor into the decision.

As a student of economics, I understood the sunk cost fallacy well. But extra food didn't *quite* seem sunk: after all, the food would still give me more calories, and even if it made me overfull for an hour or two, it could lead me to have smaller (and thus cheaper) subsequent meals.

Or, at least, that's the argument that my internal monologue would spin up to distract me long enough for my hands to keep shoving food into my mouth.

The counterargument would go something like

> *First of all, many of the calories will be either wasted or harmful if I consume them now. Secondly, the cost of dinner is more dependent upon what's available than how hungry I am. Third, even if the cost of dinner is reduced, it will be reduced by maybe a dollar, and a few hours of discomfort is not worth a dollar.*

But by then, it would already be too late; the food would be gone and I'd be overfull.

2

Failures of this form can generally be fixed by "just not doing that," which in this case entails forcing yourself to stop eating. I don't like that solution, as it requires an application of willpower, and in general, any solution that requires an application of willpower is a stopgap, not a remedy. I much prefer solutions that get all of myself onto the same page, *including* the parts that make distracting arguments so they can shovel more food into my mouth while I'm not looking.

(A problem isn't solved until it's solved automatically, without need for attention or willpower.)

The way I solved this problem was by committing to save any amount of leftover food, *no matter how small*. Two bites left? Screw it, get me a take-out box.

Committing to this, and actually doing it once or twice to show myself that I mean business, had an interesting effect.

First of all, it had the obvious effects that I stopped stuffing myself and that I occasionally had three-bite snacks available in the fridge.

But more importantly, credibly committing (to myself) that I would do the right thing *even if it seemed too late* made it much easier to automatically do the right thing.

Roughly speaking, I managed to signal to the part of myself that was worried about food scarcity that it didn't need to distract me in order to squirrel food away, because I would *actually listen to it.* I showed it that I was on its side, via an unflinching willingness to save food (even one or two bites) with a blatant disregard for social norms and weird looks from confused waiters.

And this, in turn, got that part of me onto *my* side. A willingness (and demonstrated ability) to save any amount of food no matter how small eliminated the *impetus* to keep eating when near full. This, in turn, allowed me to actually look at the remaining food and (armed with more experience about which tiny portions of food are actually appreciated later) and decide whether or not to save it.

These days, my bar for how little food I'm willing to take home is quite low, but I'm *also* comfortable throwing food out (if I'm in a rush or if it won't keep well), and I no longer get the feeling that I'm trying to distract myself for long enough to do something that I wouldn't approve of.

3

I occasionally see people hitting the failure mode where they try to apply willpower in order to do a thing (such as only eat half of their sandwich, and save the other half for later) and then fail slightly (such as by taking a bite out of the second half) at which point they proceed to completely ignore the parts of themselves that suggest restraint (such as by eating the entire second half of the sandwich and thereby stuffing themselves).

I refer to this failure mode as "failing with abandon." It seems to me that it's at least somewhat related to a failure of self-signalling: once the initial target is missed, the target itself is completely discredited and ignored in favor of total indulgence.

The technique I'm describing — self-signalling an ability to do the right thing even if it seems too late — can address this failure mode in general.

People might feel strange saving the second half of the sandwich after they've taken two bites out of it, but *if you actually do that a few times* then it becomes much easier to believe that you *can.* The narrative shifts from "well I guess I'm not saving the second half of *this* sandwich" to "I guess I was hungry enough for two more bites, but now I'll save the rest."

As it turns out, you can do the right thing after missing the initial target! Just promise yourself that you'll allow yourself to do the right thing, no matter how late.

4

There's a certain amount of self-trust that comes from making and honoring commitments to do what you want to do *even after* it's "too late" or "no longer worth it." For me, this entails a certain amount of self-loyalty: I'm willing to accept strange looks from waiters in order to save small amounts of food because I'm more loyal to the part of me that is possessive about food than I am to the social norms.

(I expect this is much easier above a certain confidence threshold, such that others say you are "eccentric" rather than "a weirdo." Your mileage may vary. But don't take that as an excuse; I still strongly encourage you to show yourself that you are able to do the right thing even after it's "too late".)

I have found that there is significant power in signalling to myself that I'm willing and able to do the thing that I want to do, no matter how futile it may seem; that I'm willing to get as close to the target as possible even if I've already missed it. This prevents me from the impulse to "fail with abandon" in the first place.

5

This technique is one facet of a more general mindset that I find quite useful, which is that of "loyalty to the self." I'll touch upon that general mindstate more next week.

Introducing Corrigibility (an FAI research subfield)

Benja, Eliezer, and I have published a new technical report, in collaboration with Stuart Armstrong of the Future of Humanity institute. This paper introduces Corrigibility, a subfield of Friendly AI research. The abstract is reproduced below:

> As artificially intelligent systems grow in intelligence and capability, some of their available options may allow them to resist intervention by their programmers. We call an AI system "corrigible" if it cooperates with what its creators regard as a corrective intervention, despite default incentives for rational agents to resist attempts to shut them down or modify their preferences. We introduce the notion of corrigibility and analyze utility functions that attempt to make an agent shut down safely if a shutdown button is pressed, while avoiding incentives to prevent the button from being pressed or cause the button to be pressed, and while ensuring propagation of the shutdown behavior as it creates new subsystems or self-modifies. While some proposals are interesting, none have yet been demonstrated to satisfy all of our intuitive desiderata, leaving this simple problem in corrigibility wide-open.

We're excited to publish a paper on corrigibility, as it promises to be an important part of the FAI problem. This is true even without making strong assumptions about the possibility of an intelligence explosion. Here's an excerpt from the introduction:

> As AI systems grow more intelligent and autonomous, it becomes increasingly important that they pursue the intended goals. As these goals grow more and more complex, it becomes increasingly unlikely that programmers would be able to specify them perfectly on the first try.
>
> Contemporary AI systems are correctable in the sense that when a bug is discovered, one can simply stop the system and modify it arbitrarily; but once artificially intelligent systems reach and surpass human general intelligence, an AI system that is not behaving as intended might also have the ability to intervene against attempts to "pull the plug".
>
> Indeed, by default, a system constructed with what its programmers regard as erroneous goals would have an incentive to resist being corrected: general analysis of rational agents.[1] has suggested that almost all such agents are instrumentally motivated to preserve their preferences, and hence to resist attempts to modify them [3, 8]. Consider an agent maximizing the expectation of some utility function U. In most cases, the agent's current utility function U is better fulfilled if the agent continues to attempt to maximize U in the future, and so the agent is incentivized to preserve its own U-maximizing behavior. In Stephen Omohundro's terms, "goal-content integrity'' is an *instrumentally convergent* goal of almost all intelligent agents [6].
>
> This holds true even if an artificial agent's programmers intended to give the agent different goals, and even if the agent is sufficiently intelligent to realize that its programmers intended to give it different goals. If a U-maximizing agent learns that its programmers intended it to maximize some other goal U*, then by default this agent has incentives to prevent its programmers from changing its utility

> function to U* (as this change is rated poorly according to U). This could result in agents with incentives to manipulate or deceive their programmers.[2]
>
> As AI systems' capabilities expand (and they gain access to strategic options that their programmers never considered), it becomes more and more difficult to specify their goals in a way that avoids unforeseen solutions—outcomes that technically meet the letter of the programmers' goal specification, while violating the intended spirit.[3] Simple examples of unforeseen solutions are familiar from contemporary AI systems: e.g., Bird and Layzell [2] used genetic algorithms to evolve a design for an oscillator, and found that one of the solutions involved repurposing the printed circuit board tracks on the system's motherboard as a radio, to pick up oscillating signals generated by nearby personal computers. Generally intelligent agents would be far more capable of finding unforeseen solutions, and since these solutions might be easier to implement than the intended outcomes, they would have every incentive to do so. Furthermore, sufficiently capable systems (especially systems that have created subsystems or undergone significant self-modification) may be very difficult to correct without their cooperation.
>
> In this paper, we ask whether it is possible to construct a powerful artificially intelligent system which has no incentive to resist attempts to correct bugs in its goal system, and, ideally, is incentivized to aid its programmers in correcting such bugs. While autonomous systems reaching or surpassing human general intelligence do not yet exist (and may not exist for some time), it seems important to develop an understanding of methods of reasoning that allow for correction *before* developing systems that are able to resist or deceive their programmers. We refer to reasoning of this type as *corrigible*.
>
> [1] Von Neumann-Morgenstern rational agents [7], that is, agents which attempt to maximize expected utility according to some utility function)
>
> [2] In particularly egregious cases, this deception could lead an agent to maximize U* only until it is powerful enough to avoid correction by its programmers, at which point it may begin maximizing U. Bostrom [4] refers to this as a "treacherous turn''.
>
> [3] Bostrom [4] calls this sort of unforeseen solution a "perverse instantiation''.

(See the paper for references.)

This paper includes a description of Stuart Armstrong's utility indifference technique previously discussed on LessWrong, and a discussion of some potential concerns. Many open questions remain even in our small toy scenario, and many more stand between us and a formal description of what it even means for a system to exhibit corrigible behavior.

> Before we build generally intelligent systems, we will require some understanding of what it takes to be confident that the system will cooperate with its programmers in addressing aspects of the system that they see as flaws, rather than resisting their efforts or attempting to hide the fact that problems exist. We will all be safer with a formal basis for understanding the desired sort of reasoning.
>
> As demonstrated in this paper, we are still encountering tensions and complexities in formally specifying the desired behaviors and algorithms that will compactly

yield them. The field of corrigibility remains wide open, ripe for study, and crucial in the development of safe artificial generally intelligent systems.

A Day Without Defaults

Author's note: this post was written on Sunday, Oct. 19th. Its sequel will be written on Sunday, Oct. 27th.

Last night, I went to bed content with a fun and eventful weekend gone by. This morning, I woke up, took a shower, did my morning exercises, and began eat breakfast before making the commute up to work.

At the breakfast table, though, I was surprised to learn that it was Sunday, not Monday. I had misremembered what day it was and in fact had an entire day ahead of me with nothing on the agenda. At first, this wasn't very interesting, but then I started thinking. What to do with an entirely free day, without any real routine?

I realized that I didn't particularly know what to do, so I decided that I would simply live a day without defaults. At each moment of the day, I would act only in accordance with my curiosity and genuine interest. If I noticed myself becoming bored, disinterested, or otherwise less than enthused about what was going on, I would stop doing it.

What I found was quite surprising. I spent much less time doing routine activities like reading the news and browsing discussion boards, and much more time doing things that I've "always wanted to get around to"-- meditation, trying out a new exercise routine, even just spending some time walking around outside and relaxing in the sun.

Further, this seemed to actually make me more productive. When I sat down to get some work done, it was because I was legitimately interested in finishing my work and curious as to whether I could use a new method I had thought up in order to solve it. I was able to resolve something that's been annoying me for a while in much less time than I thought it would take.

By the end of the day, I started thinking "is there any reason that I don't spend every day like this?" As far as I can tell, there isn't really. I do have a few work tasks that I consider relatively uninteresting, but there are multiple solutions to that problem that I suspect I can implement relatively easily.

My plan is to spend the next week doing the same thing that I did today and then report back. I'm excited to let you all know what I find!

Fixing Moral Hazards In Business Science

I'm a LW reader, two time CFAR alumnus, and rationalist entrepreneur.

Today I want to talk about something insidious: marketing studies.

Until recently I considered studies of this nature merely unfortunate, funny even. However, my recent experiences have caused me to realize the situation is much more serious than this. Product studies are the public's most frequent interaction with science. By tolerating (or worse, expecting) shitty science in commerce, we are undermining the public's perception of science as a whole.

The good news is this appears fixable. I think we can change how startups perform their studies immediately, and use that success to progressively expand.

Product studies have three features that break the assumptions of traditional science: (1) few if any follow up studies will be performed, (2) the scientists are in a position of moral hazard, and (3) the corporation seeking the study is in a position of moral hazard (for example, the filing cabinet bias becomes more of a "filing cabinet exploit" if you have low morals and the budget to perform 20 studies).

I believe we can address points 1 and 2 directly, and overcome point 3 by appealing to greed.

Here's what I'm proposing: we create a webapp that acts as a high quality (though less flexible) alternative to a Contract Research Organization. Since it's a webapp, the cost of doing these less flexible studies will approach the cost of the raw product to be tested. For most web companies, that's $0.

If we spend the time to design the standard protocols well, it's quite plausible any studies done using this webapp will be in the top 1% in terms of scientific rigor.

With the cost low, and the quality high, such a system might become the startup equivalent of citation needed. Once we have a significant number of startups using the system, and as we add support for more experiment types, we will hopefully attract progressively larger corporations.

Is anyone interested in helping? I will personally write the webapp and pay for the security audit if we can reach quorum on the initial protocols.

Companies who have expressed interested in using such a system if we build it:

- Beeminder
- HabitRPG
- MealSquares
- Complice (disclosure: the CEO, Malcolm, is a friend of mine)
- General Biotics (disclosure: the CEO, David, is me)

(I sent out my inquiries at 10pm yesterday, and every one of these companies got back to me by 3am. I don't believe "startups love this idea" is an overstatement.)

So the question is: how do we do this right?

Here are some initial features we should consider:

- Data will be collected by a webapp controlled by a trusted third party, and will only be editable by study participants.
- The results will be computed by software decided on before the data is collected.
- Studies will be published regardless of positive or negative results.
- Studies will have mandatory general-purpose safety questions. (web-only products likely exempt)
- Follow up studies will be mandatory for continued use of results in advertisements.
- All software/contracts/questions used will be open sourced (MIT) and creative commons licensed (CC BY), allowing for easier cross-product comparisons.

Any placebos used in the studies must be available for purchase as long as the results are used in advertising, allowing for trivial study replication.

Significant contributors will receive:

- Co-authorship on the published paper for the protocol.
- (Through the paper) an Erdos number of 2.
- The satisfaction of knowing you personally helped restore science's good name (hopefully).

I'm hoping that if a system like this catches on, we can get an "effective startups" movement going :)

So how do we do this right?