Best of LessWrong: May 2015

1. We Should Introduce Ourselves Differently
2. The File Drawer Effect and Conformity Bias (Election Edition)
3. Leaving LessWrong for a more rational life
4. Is Scott Alexander bad at math?
5. How my social skills went from horrible to mediocre
6. "Should" considered harmful
7. You're allowed to fight for something
8. Caring about something larger than yourself
9. You don't get to know what you're fighting for
10. Not because you "should"

Best of LessWrong: May 2015

1. We Should Introduce Ourselves Differently
2. The File Drawer Effect and Conformity Bias (Election Edition)
3. Leaving LessWrong for a more rational life
4. Is Scott Alexander bad at math?
5. How my social skills went from horrible to mediocre
6. "Should" considered harmful
7. You're allowed to fight for something
8. Caring about something larger than yourself
9. You don't get to know what you're fighting for
10. Not because you "should"

We Should Introduce Ourselves Differently

I told an intelligent, well-educated friend about Less Wrong, so she googled, and got "Less Wrong is an online community for people who want to *apply* the discovery of biases like the conjunction fallacy, the affect heuristic, and scope insensitivity in order to fix their own thinking." and gave up immediately because she'd never heard of the biases.

While hers might not be the best possible attitude, I can't see that we win anything by driving people away with obscure language.

Possible improved introduction: "Less Wrong is a community for people who would like to think more clearly in order to improve their own and other people's lives, and to make major disasters less likely."

The File Drawer Effect and Conformity Bias (Election Edition)

As many of you may be aware, the UK general election took place yesterday, resulting in a surprising victory for the Conservative Party. The pre-election opinion polls predicted that the Conservatives and Labour would be roughly equal in terms of votes cast, with perhaps a small Conservative advantage leading to a hung parliament; instead the Conservatives got 36.9% of the vote to Labour's 30.4%, and won the election outright.

There has already been a lot of discussion about why the polls were wrong, from methodological problems to incorrect adjustments. But perhaps more interesting is the possibility that the polls were right! For example, Survation did a poll on the evening before the election, which predicted the correct result (Conservatives 37%, Labour 31%). However, that poll was never published because the results seemed "out of line." Survation didn't want to look silly by breaking with the herd, so they just kept quiet about their results. Naturally this makes me wonder about the existence of other unpublished polls with similar readings.

This seems to be a case of two well know problems colliding with devastating effect. Conformity bias caused Survation to ignore the data and go with what they "knew" to be the case (for which they have now paid dearly). And then the file drawer effect meant that the generally available data was skewed, misleading third parties. The scientific thing to do is to publish all data, including "outliers," both so that information can change over time rather than be anchored, and to avoid artificially compressing the variance. Interestingly, the exit poll, which had a methodology agreed beforehand and was previously committed to be published, was basically right.

This is now the third time in living memory that opinion polls have been embarrassingly wrong about the UK general election. Each time this has lead to big changes in the polling industry. I would suggest that one important scientific improvement is for polling companies to announce the methodology of a poll and any adjustments to be made *before* the poll takes place, and commit to publishing *all* polls they carry out. Once this became the norm, data from any polling company that didn't follow this practice would be rightly seen as unreliable by comparison.

Leaving LessWrong for a more rational life

You are unlikely to see me posting here again, after today. There is a saying here that politics is the mind-killer. My heretical realization lately is that philosophy, as generally practiced, can also be mind-killing.

As many of you know I am, or was running a twice-monthly Rationality: AI to Zombies reading group. One of the bits I desired to include in each reading group post was a collection of contrasting views. To research such views I've found myself listening during my commute to talks given by other thinkers in the field, e.g. Nick Bostrom, Anders Sandberg, and Ray Kurzweil, and people I feel are doing “ideologically aligned” work, like Aubrey de Grey, Christine Peterson, and Robert Freitas. Some of these were talks I had seen before, or generally views I had been exposed to in the past. But looking through the lens of learning and applying rationality, I came to a surprising (to me) conclusion: it was philosophical thinkers that demonstrated the largest and most costly mistakes. On the other hand, de Grey and others who are primarily working on the scientific and/or engineering challenges of singularity and transhumanist technologies were far less likely to subject themselves to epistematic mistakes of significant consequences.

Philosophy as the anti-science...

What sort of mistakes? Most often reasoning by analogy. To cite a specific example, one of the core underlying assumption of singularity interpretation of super-intelligence is that just as a chimpanzee would be unable to predict what a human intelligence would do or how we would make decisions (aside: how would we know? Were any chimps consulted?), we would be equally inept in the face of a super-intelligence. This argument is, however, nonsense. The human capacity for abstract reasoning over mathematical models is in principle a fully general intelligent behaviour, as the scientific revolution has shown: there is no aspect of the natural world which has remained beyond the reach of human understanding, once a sufficient amount of evidence is available. The wave-particle duality of quantum physics, or the 11-dimensional space of string theory may defy human intuition, i.e. our built-in intelligence. But we have proven ourselves perfectly capable of understanding the logical implications of models which employ them. We may not be able to build intuition for how a super-intelligence thinks. Maybe—that's not proven either. But even if that is so, we will be able to reason about its intelligent behaviour in advance, just like string theorists are able to reason about 11-dimensional space-time without using their evolutionarily derived intuitions at all.

This post is not about the singularity nature of super-intelligence—that was merely my choice of an illustrative example of a category of mistakes that are too often made by those with a philosophical background rather than the empirical sciences: the reasoning by analogy instead of the building and analyzing of predictive models. The fundamental mistake here is that reasoning by analogy is not in itself a sufficient explanation for a natural phenomenon, because it says nothing about the context sensitivity or insensitivity of the original example and under what conditions it may or may not hold true in a different situation.

A successful physicist or biologist or computer engineer would have approached the problem differently. A core part of being successful in these areas is knowing when it is that you have insufficient information to draw conclusions. If you don't know what you don't know, then you can't know when you might be wrong. To be an effective rationalist, it is often not important to answer “what is the calculated probability of that outcome?” The better first question is “what is the uncertainty in my calculated probability of that outcome?” If the uncertainty is too high, then the data supports no conclusions. And the way you reduce uncertainty is that you build models for the domain in question and empirically test them.

The lens that sees its own flaws...

Coming back to LessWrong and the sequences. In the preface to Rationality, Eliezer Yudkowsky says his biggest regret is that he did not make the material in the sequences more practical. The problem is in fact deeper than that. The art of rationality is the art of truth seeking, and empiricism is part and parcel essential to truth seeking. There's lip service done to empiricism throughout, but in all the “applied” sequences relating to quantum physics and artificial intelligence it appears to be forgotten. We get instead definitive conclusions drawn from thought experiments only. It is perhaps not surprising that these sequences seem the most controversial.

I have for a long time been concerned that those sequences in particular promote some ungrounded conclusions. I had thought that while annoying this was perhaps a one-off mistake that was fixable. Recently I have realized that the underlying cause runs much deeper: what is taught by the sequences is a form of flawed truth-seeking (thought experiments favored over real world experiments) which inevitably results in errors, and the errors I take issue with in the sequences are merely examples of this phenomenon.

And these errors have consequences. Every single day, 100,000 people die of preventable causes, and every day we continue to risk extinction of the human race at unacceptably high odds. There is work that could be done now to alleviate both of these issues. But within the LessWrong community there is actually outright hostility to work that has a reasonable chance of alleviating suffering (e.g. artificial general intelligence applied to molecular manufacturing and life-science research) due to concerns arrived at by flawed reasoning.

I now regard the sequences as a memetic hazard, one which may at the end of the day be doing more harm than good. One should work to develop one's own rationality, but I now fear that the approach taken by the LessWrong community as a continuation of the sequences may result in more harm than good. The anti-humanitarian behaviors I observe in this community are not the result of initial conditions but the process itself.

What next?

How do we fix this? I don't know. On a personal level, I am no longer sure engagement with such a community is a net benefit. I expect this to be my last post to LessWrong. It may happen that I check back in from time to time, but for the most part I intend to try not to. I wish you all the best.

A note about effective altruism…

One shining light of goodness in this community is the focus on effective altruism—doing the most good to the most people as measured by some objective means. This is a noble goal, and the correct goal for a rationalist who wants to contribute to charity. Unfortunately it too has been poisoned by incorrect modes of thought.

Existential risk reduction, the argument goes, trumps all forms of charitable work because reducing the chance of extinction by even a small amount has far more expected utility than would accomplishing all other charitable works combined. The problem lies in the likelihood of extinction, and the actions selected in reducing existential risk. There is so much uncertainty regarding what we know, and so much uncertainty regarding what we don't know that it is impossible to determine with any accuracy the expected risk of, say, unfriendly artificial intelligence creating perpetual suboptimal outcomes, or what effect charitable work in the area (e.g. MIRI) is have to reduce that risk, if any.

This is best explored by an example of existential risk done right. Asteroid and cometary impacts is perhaps the category of external (not-human-caused) existential risk which we know the most about, and have done the most to mitigate. When it was recognized that impactors were a risk to be taken seriously, we recognized what we did not know about the phenomenon: what were the orbits and masses of Earth-crossing asteroids? We built telescopes to find out. What is the material composition of these objects? We built space probes and collected meteorite samples to find out. How damaging an impact would there be for various material properties, speeds, and incidence angles? We built high-speed projectile test ranges to find out. What could be done to change the course of an asteroid found to be on collision course? We have executed at least one impact probe and will monitor the effect that had on the comet's orbit, and have on the drawing board probes that will use gravitational mechanisms to move their target. In short, we identified what it is that we don't know and sought to resolve those uncertainties.

How then might one approach an existential risk like unfriendly artificial intelligence? By identifying what it is we don't know about the phenomenon, and seeking to experimentally resolve that uncertainty. What relevant facts do we not know about (unfriendly) artificial intelligence? Well, much of our uncertainty about the actions of an unfriendly AI could be resolved if we were to know more about how such agents construct their thought models, and relatedly what language were used to construct their goal systems. We could also stand to benefit from knowing more practical information (experimental data) about in what ways AI boxing works and in what ways it does not, and how much that is dependent on the structure of the AI itself. Thankfully there is an institution that is doing that kind of work: the Future of Life institute (not MIRI).

Where should I send my charitable donations?

Aubrey de Grey's SENS Research Foundation.

100% of my charitable donations are going to SENS. Why they do not get more play in the effective altruism community is beyond me.

If you feel you want to spread your money around, here are some non-profits which have I have vetted for doing reliable, evidence-based work on singularity technologies and existential risk:

- Robert Freitas and Ralph Merkle's Institute for Molecular Manufacturing does research on molecular nanotechnology. They are the only group that work on the

long-term Drexlarian vision of molecular machines, and publish their research online.
- Future of Life Institute is the only existential-risk AI organization which is actually doing meaningful evidence-based research into artificial intelligence.
- B612 Foundation is a non-profit seeking to launch a spacecraft with the capability to detect, to the extent possible, ALL Earth-crossing asteroids.

I wish I could recommend a skepticism, empiricism, and rationality promoting institute. Unfortunately I am not aware of an organization which does not suffer from the flaws I identified above.

Addendum regarding unfinished business

I will no longer be running the Rationality: From AI to Zombies reading group as I am no longer in good conscience able or willing to host it, or participate in this site, even from my typically contrarian point of view. Nevertheless, I am enough of a libertarian that I feel it is not my role to put up roadblocks to others who wish to delve into the material as it is presented. So if someone wants to take over the role of organizing these reading groups, I would be happy to hand over the reigns to that person. If you think that person should be you, please leave a reply in another thread, not here.

EDIT: Obviously I'll stick around long enough to answer questions below :)

Is Scott Alexander bad at math?

This post is a third installment to the sequence that I started with The Truth About Mathematical Ability and Innate Mathematical Ability. I begin to discuss the role of aesthetics in math.

There was strong interest in the first two posts in my sequence, and I apologize for the long delay. The reason for it is that I've accumulated hundreds of pages of relevant material in draft form, and have struggled with how to organize such a large body of material. I still don't know what's best, but since people have been asking, I decided to continue posting on the subject, even if I don't have my thoughts as organized as I'd like. I'd greatly welcome and appreciate any comments, but I won't have time to respond to them individually, because I already have my hands full with putting my hundreds of pages of writing in public form.

Where I come from

My father is a remarkable creature, and I'm grateful for the opportunity to have grown up around him. Amongst other things, we share a love of music. There's a fair amount of overlap in our musical tastes. But there's an important difference between us.

When a piece of music is complex, like a piano sonata or a symphony, I often need to listen to it repeatedly before I figure out what I like about it. When I share the piece with him that he's never heard before, he'll often highlight the parts that I like most in real time, on first listening, without my having said anything.

In the past, people would have attributed this to magic, or other supernatural constructs like telepathy. We now know that these explanations don't suffice.

You might hypothesize that the difference comes from him having greater abstract pattern recognition ability than my own. In fact, this is the case, but it doesn't suffice to account for the phenomenon. Some people with greater pattern recognition ability than me don't appreciate music at all. More significantly, my father doesn't figure out what I like by *thinking about it* – his reactions are instead emotionally rooted, for example, he broke into tears upon hearing the repetition of the original theme in the final movement of Beethoven's piano sonata Op. 109.

For whatever reason, my father's *initial* emotional responses are surprisingly often closely aligned with my *eventual* emotional responses than *my own* initial emotional responses are. They also seem to be more closely aligned with the *average* person's eventual emotional responses than my own initial emotional responses are. The phenomenon extends beyond music, into the visual arts and even math. It plays a role in his work as Art Director for the Wells Fargo website.

People are often surprised to learn that my IQ is about average for the Less Wrong community: they think that it you need to be a lot smarter to be as good at math as I am. They're not the only ones: a leading researcher in the field of exceptional intellectual talent expressed surprise that I was able to become a mathematician given that I have a nonverbal learning disability.

When I hear people say these things I smile inwardly.

Math is an art

You see, there are broad misconceptions that math is about *intelligence*. No, math is an *art*. This isn't just true of pure math, it's also true of applied math, statistics, physics and computer science. Sufficiently high quality mathematical thinking of any kind has a large aesthetic component. My unusually high mathematical ability doesn't come me having *higher intelligence* than my conversation partners. It comes from me having *unusually high aesthetic discernment*, something that I acquired from my father, both out of virtue of inheriting his genes, and out of virtue of having him as a strong environmental influence in my life.

That's how I was able to go from failing geometry in 9th grade to being the best calculus student in my high school class of ~650 people. I was far from being the sharpest of my classmates, but my aesthetic sense drove me in the direction of rediscovering how to do mathematical research, and at that point it became easy for me to reconstruct any part of the high school math curriculum. I transcended the paradigm of "memorizing without understanding very well" to gain a deep conceptual understanding of the material, without needing outside assistance.

Just as levels of innate intelligence vary greatly, levels of innate aesthetic discernment vary greatly, and this has profound ramifications. Even if I were as smart as John von Neumann, I *still* wouldn't be able to discover the fast Fourier transform in the early 1800's like Gauss did: I don't have enough aesthetic discernment. This shouldn't be surprising – even though I have some musical talent, there's no way that I could write music as great as Beethoven's late string quartets.

But if you're reading this post with interest, you've already distinguished yourself as somebody who can probably understand and appreciate math much more deeply than you would have imagined possible.

I understand that you may doubt me. The great mathematician Alexander Grothendieck understood too. He wrote to people in your position:

> *It's to that being inside of you who knows how to be alone, it is to this infant that I wish to speak, and no-one else. I'm well aware that this infant has been considerably estranged. It's been through some hard times, and more than once over a long period. It's been dropped off Lord knows where, and it can be very difficult to reach. One swears that it died ages ago, or that it never existed - and yet I am certain it's always there, and very much alive.*

Is Scott Alexander "bad at math"?

In The Parable of The Talents Scott Alexander discusses his mathematical ability:

> In Math, I just barely by the skin of my teeth scraped together a pass in Calculus with a C-. [...] Meanwhile, there were some students who did better than I did in Math with seemingly zero effort. I didn't begrudge those students. But if they'd started trying to say they had exactly the same level of innate ability as I did, and the only difference was they were trying while I was slacking off, then I sure as hell would have begrudged them. Especially if I knew they were lazing around on the beach while I was poring over a textbook.

I don't doubt that Scott Alexander struggled to get a C- in calculus, and worked much harder than some other students. But **almost surely, what he was seeing wasn't math in a meaningful sense**. What he was seeing was more akin a course that teaches scales and chords to piano students. It's just not true that if someone has substantially more trouble learning scales and chords than his or her classmates, he or she is "worse than them at music."

The signals of Scott's mathematical ability coming *outside* of formal math classes are **much stronger**. Some of these are fairly obvious — as Ilya Shpitser wrote:

> Scott's complaints about his math abilities often go like this: "Man, I wish I wasn't so terrible at math. Now if you will excuse me, I am going to tear the statistical methodology in this paper to pieces."

But these don't even constitute the *main* evidence that Scott Alexander is good at math.

When a friend pointed out a couple of his blog posts back in early 2010, I did a double take, and thought "wow, this guy has something really special." I'm not alone: there's a broad consensus that he's a great writer, both within and outside of the Less Wrong community. Ezra Klein has been named one of the 50 most powerful people in Washington DC and he responded to one of Scott's blog posts.

A large part of what makes Scott's posts a pleasure to read is his storytelling ability, which overlaps strongly with the ability to write narrative fiction. There are hints that come across in the cultural references that he makes that he has a strong appreciation for art in general.

When I mentioned the unsolvability of quintic to Scott in passing, it grabbed his attention, and he was visibly very curious as to how it could be possible to show that a general quintic polynomial has no solutions in terms of radicals. *It's the exact same reaction that my father has had to some of the deep math that I've showed him*. There aren't very many mathematicians who have such a strong level of interest in the unsolvability of the quintic when they first encounter it.

What accounts for the difference? Like my father, Scott has exceptional aesthetic discernment. If most mathematicians had as much as he did, they would rightly find what I mentioned as striking as Scott did: the problem of showing that the quintic isn't solvable in radicals is what led to Galois Theory, one of the pinnacles of mathematical achievement, and the backdrop for the study of the Absolute Galois Group, one of the deepest areas of contemporary mathematical research.

People don't believe me when I tell them they're good at math!

When I try to convince people like Scott that they're actually very good at math, they often say "No, you don't understand, I'm *really* bad at math, you're overestimating my mathematical ability because of my writing ability." To which my response is "I know you think that, I've seen many people in your rough direction who think that they're *really* bad at math, and say that I don't understand how bad they are, and they're almost always wrong: they *almost never know that what they were having trouble with wasn't representative of math*."

I *taught* myself how to do mathematical research in order to understand calculus deeply. I've been thinking deeply about mathematical education for 12 years. I spent hundreds of hours tutoring students in calculus in high school and college. I taught calculus for 6 semesters at University of Illinois. I completed a PhD in math. Scott's exposure to calculus seems to consist of a single year in calculus. Your Bayesian prior should be that I know more about Scott's mathematical potential than Scott does. :-)

But so often I've seen people in Scott's position not believe me. By the time people have reached their mid-20's, they generally have *such strong negative perceptions of their mathematical ability* that I can't get through to them: their confirmation bias is *too strong*, there's nothing that I can do about the situation. So it may be that Scott will incorrectly think that he's bad at math forever, and that there's nothing that I can do about it. But maybe this article will influence at least someone's thinking.

I'll substantiate my claim that aesthetic sense drives a large fraction of mathematical accomplishment in future posts.

How my social skills went from horrible to mediocre

Over the past few months, I've become aware that my understanding of social reality had been distorted to an extreme degree. It took 29 years for me to figure out what was going on, but I finally now understand.

The situation is very simple: The amount of time that I put into interacting within typical social contexts was very small, so I didn't get enough feedback to realize that I had a major blindspot as I otherwise would have.

Now that I've identified the blindspot, I can work on it, and my social awareness has been increasing at very rapid clip. I had no idea that I had so much potential for social awareness. I had been in a fixed mindset as rather than a growth mindset, I had thought "social skills will never be my strong point, so I shouldn't spend time trying to improve them, instead I should focus on what I'm best at." I'm astonished by how much my relationships have improved over a span of mere weeks.

I give details below.

How I spent my time growing up

I've been extremely metacognitive and reflective since early childhood, and have spent most of my time optimizing for my intellectual growth. Even as a child, the things that I thought about where very unusual: at age 7, upon reflection, I realized that there's no free will in the sense that people usually think of it: that brain chemistry drives our decisions in a very strong sense.

As I grew up, my interests became more and more remote from those of my peers, and the pool of conversation topics of mutual interest diminished rapidly as I got older. For this reason, I generally found my interactions with others to be very unfulfilling: other people were rarely interested in talking about what I wanted to talk about, and I struggled to find points of mutual interest.

Because I was much more unusual than most of my conversation partners, there was an implicit assumption that the responsibility of finding common ground fell exclusively on me, rather than being shared by me and my conversation partners. Even when I tried really hard to connect with my conversation partners, it often came across to my conversation partners though I wasn't trying, because our interests were *so different* that even if I bridged 95% of the gap, the remaining 5% was uncomfortably large for them, so that they would feel resentful toward me.

There were almost no people who shared my interests. So my choices seemed to be

1. Socialize and talk about things that I have no interest in.
2. Socialize and try to talk about what I'm interested in, at risk of alienating my conversation partners.
3. Keep to myself

Each of (1) and (2) tended to be very unpleasant, which pushed strongly in the direction of (3). I essentially **never** engaged in normal social activities. In college,

every day I would see my classmates sitting with their friends in the cafeteria, whereas I would almost always be sitting alone.

In the subsequent intervening years I developed further and further in the direction of having deep insights about the world.a strong focus on abstraction and generality. I essentially **never** engaged in usual social activities. In college, almost all of my classmates would sit at tables in the cafeteria with their friends, and I would almost always sit alone. I **almost** never went to Less Wrong meetups, because I had already thought about most of what people discussed, so that it was more efficient for me to learn on my own.

I found these reminders of my isolation to be depressing, but didn't think much about it. In hindsight I see that I erred in not thinking about the situation more deeply.

I've found that Malcolm Gladwell's view that developing mastery of a field takes ~10,000 hours is largely true.

When people tell me that they were bad at calculus, my internal response had become "When I was learning calculus I spent **~20 hours a week** on it. it's not at all surprising to me that you wouldn't become good at calculus without having done so, independently of whether or not you had the ability to."

How many many hours had I spent socializing by age 29? Lots, but almost exclusively with a small handful of people who are very unusual in the same ways that I am. When I was in a group conversation, I would usually find the conversation uninteresting, and let my mind wander, without attempting to participate myself. Thinking it over, I probably spent less than 5% as much time participating in usual social contexts as other 29 year olds had by the same age.

It didn't occur to me how significant this was. The number of hours that I had is perhaps as small as the number of hours that most people have by age 10. In hindsight it's obvious: of **course** I didn't have good social skills relative to other adults, in the same way that a 10 year old doesn't have good social skills for an adult. I just hadn't put nearly enough time in!

What went horribly wrong

Throughout my life, I've yearned for companionship, and have had a strong desire to contribute to global welfare. Up until the past year, I was extremely socially isolated, and my positive social impact was utterly negligible. This gave rise to a huge amount of cognitive dissonance. As Eliezer wrote:

> I keep trying to say that rationality is the winning-Way [...] Be careful [...] any time you find yourself defining the "winner" as someone other than the agent who is currently smiling from on top of a giant heap of utility.

If I cared so much about connecting people and about contributing to global welfare, then *why wasn't I getting anything done*?

My theory of mind was based on my knowledge of my own mind (c.f. Yvain's post Generalizing From One Example). I engage almost exclusively in metacognition and deep reflection. I therefore had *no reference frame* for what other people are like: I projected my own style of thinking on the people who I interacted with. The effect of

this was that I became a figurative space alien, almost totally out of touch with the rest of the human race.

Concretely, how was I socially oblivious?

My implicit model of other people's minds was along the lines "everyone always has access to a transcript of all conversations that we've ever had at his or her disposal." This probably seems loony, and rightly so. I was very focused on carefully organizing my interactions with everyone in my mind. It just didn't occur to me that my conversations partners weren't doing the same thing! My *subjective sense* of what was going on in my conversation partners' mind turns out to have usually been *completely* different from what was *actually* going on in my conversation partners' mind.

Some of my common self destructive patterns of behavior were:

(a) "Person X expresses insecurity over Y. I spend several dozen hours contemplating how to reassure person X. I then broach the subject with person X without offering any background context, assuming that person X knows that I'm following up on a specific conversation thread from several weeks ago, and wants to continue the conversation about the subject. In person X's eyes, it looks like I'm bringing up a triggering subject for no reason, and person X develops an Ugh Field around me, of the type "when I talk to Jonah, he says things that make me feel bad, so I don't want to talk with him anymore."

My reaction to this was "this is so weird, these people are *really touchy,* such that they're unable to have conversations about topics that they themselves bring up. How is it even possible for people to have conversations given how touchy they are?"

I didn't know that when someone brings up a sensitive subject, *that's not necessarily an invitation to talk about it*, and that *they didn't realize that I was responding to something that they had said weeks ago*.

(b) A woman sends signals of romantic interest, either accidentally, or whimsically. I mistakenly assume that she's carefully deliberating over the possibility of dating me, as I would be in her position. I decide to express interest in her.

She hasn't been thinking about whether or not she'd like to date me *at all*, she was instead engaging in casual preliminary flirting and/or wasn't carefully guarding against accidentally sending signals of romantic interest. So from her point of view it looks like "This guy expressed romantic interest in me *without paying attention to how I'm feeling*." She reactively reprimands me, or cuts contact with me, usually with connotations (even if slight) that I might not respect her boundaries.

I mistakenly think that she had *carefully deliberated* on how to respond to my expression of romantic interest. So I mistakenly perceive the false dichotomy:

1. I'm a delusional potential rapist, and she sees this.
2. I'm not a delusional potential rapist, she knows that she's made me feel like I might be one. The woman who I loved has turned out to have so little empathy that she doesn't mind the fact that she's done this.

Both of these possibilities are *extremely* upsetting, and I fall into severe depression, totally oblivious to the fact that she was behaving in a reactive way and that her reaction is neither evidence that I'm a potential rapist, nor evidence that she doesn't mind me feeling like a potential rapist.

(c) A lot of things that people find offensive I don't find at all offensive. For example, if a student tells me that I'm the worst teacher he or she has ever had, it makes me feel bad because I feel like I'm not contributing value, but I'm not at all upset with *the student*: my attitude is that the student is conveying valuable information to me, and that I should be appreciative.

I always knew that it's best to soften such things, but I didn't know how triggering unexpected criticism is – I didn't know that **far** more gentle remarks can be triggering for most people.

So I might tell a friend:

> "There's strong evidence that there are only a few people in the world who have a chance of solving the math research problem that you've been working on for the past few years. It's very unlikely that you have the innate ability to solve it regardless of how hard you work on it. You're a good mathematician, you could make a lot of progress on easier problems, and that would probably make you happier."

In **my mind**, what's salient is **the facts that I want him to be happy, and that he'd plausibly be happier doing something else**. But that's not what's most salient about the situation **him:** instead it just sounds like I'm just saying "you're too bad at math to be able to meet your goals."

(d) Often when I talk, I'm trying to illustrate a general principle, and give an example to illustrate it. Sometimes I'm drawing an analogy between two general principles, and give an example to illustrate one of them. Sometimes I'll state a general principle as a special case of a still more general principle, and give an example to illustrate one of the two.

What's salient to my conversation is the example that I'm giving, not the general principle that I'm trying to illustrate. So my conversation partner and I are talking past each other: I don't know that the person doesn't know that I'm talking about a general principle, or an analogy between general principles, or a general principle being a special case of a general principle, my conversation partner doesn't know that I don't know.

For example, I might say:

> "Sometimes our perceptions of social reality are very distorted. For example, **I used to be confused and mistakenly believed** that I'm the only good person in the world."

My conversation partner might respond to this "Look Jonah, you're very confused, you're not the only good person in the world!", because what's salient is "I'm the only good person in the world", not "I used to be confused and mistakenly believed..." I **mistakenly** interpret the situation as "people are so obsessed with status that they're totally blind to anything that's not a status grab," when the person doesn't *actually have any way of knowing what I was trying to say*, because my strong focus on general principles is so unusual.

(e) I say something that someone doesn't understand. I think "maybe the person needs more context," and follow up by giving more context. The person *still* doesn't understand, so I think "ok, I guess I have to give even more context" and so continue in the same direction. In fact, the amount of context that I would need to give for my point to be clear would take ~100 hours to convey, so that what I'm doing is actually not at all productive. The person perceives the situation as

> Jonah is totally ignoring the fact that I'm not understanding what I'm saying, and keeps going on and on about the same thing, oblivious to my feelings

because he or she has no way of knowing that *I'm explicitly trying to address the fact that the person is uncomfortable about not understanding*.

(f) I mistakenly believe that when people are unhappy with me, they'll tell me, because I know that I wouldn't be offended, and because I'm so verbal that I relate a very large fraction of my thoughts when I talk with someone.

So people will smile and show superficial indications of good will while being unhappy with me, and I have no idea what's going on.

If you've followed what I've said so far, it's probably not hard to understand how my misunderstandings would almost totally nullify my ability to contribute to global welfare :-).

How did I escape?

(a) Learning data science resulted in a *huge* boost to my intellectual caliber – the ways of thinking about the world that I developed are **very powerful**, and confer an advantage of the same magnitude as learning about selection effects and regression to the mean.

After this, even my closest friends could no longer understand what I was talking about, and told me as much, and I realized "Ok, I have some sort of serious blindspot, my intuitive sense is that people are understanding me when they're not, I need to figure out what's going on.

(b) A relative who's a salesman gave me very helpful advice after I had been rejected from a large number of jobs that I interviewed for explaining "When somebody asks you a question, you're giving answers that are way too long and you're not gauging where your interviewers are coming from. When they ask for you to describe your project, they're looking for a 1-2 minute response, not a ~6 minute response – from their perspective you're hijacking the conversation and talking about something that they're not interested in.

After this, I paid closer attention to my interviewers' body language and how they were directing the conversation, and I saw that he had been right.

I recently got very helpful explicit feedback from students that made me realize that I was on a *totally* different wavelength from the students, when I had no idea that that was the case.

(c) I started socializing more with people who are similar to me on dimensions other than the one that this post is about, and this resulted in me getting more useful feedback, because they could understand me more deeply than most people who had gotten upset with me for no apparent reason

(d) Learning data science made me realize that I could use the Wisdom of the Crowds to tease out what the common problem was in all of my interactions with people. It wasn't easy: the different instances were superficially totally different. It's not at all *a priori* clear what the two things

- "I expressed interest in a woman when she appears to have sent signals of romantic interest, and she rejects me in harsh terms."
- "I told a friend that I think that the math problem that he's working on is really hard and probably not feasible for him to solve, and he's mortified and cuts contact with me."

have in common. But learning data science gave me new ways of thinking about the world that enabled me to see the underlying pattern.

Figuring out what was going on has enabled me to improve my relationships with my family members, patch up relationships with friends who I alienated earlier in life, and interact more productively with people who I've just met.

Generalizable takeaways

(a) Focusing on understanding how one is similar to others and how one is different from others can be a better way to become socially aware than usual efforts to "develop social skills."

I knew that people thought I had bad social skills, but they weren't able to explain the situation to me in a way that I could understand, because they were totally misinterpreting me, on account of not knowing what was going on in my mind. So almost everything that they said about my social skills seemed wrong – they would claim that I didn't care about people's feelings, to which my response was along the lines "What are you talking about? I spend dozens of hours thinking about my friends' feelings."

They didn't have the information that they would have needed to help me: they didn't know that they needed to say "I know that you're thinking a lot about people's feelings *as they appear to you from the outside*, but you're not thinking about people's *actual* feelings: you can't assume that you know what's going on in their minds, you have to carefully feel out the situation."

(b) Finally figuring out what was going on corresponds to a huge boost in potential productivity: I finally have nontrivial prospects of transforming from

> "The guy who has deep insights but who doesn't get anything done, because he he's socially dysfunctional so nobody listens to him"

to

> "The guy who has deep insights and can use them to change the world"

(c) I now have realistic prospects for having a romantic relationship, which was not the case before.

My **past** attitude had been "The emotional cost of going through yet another traumatic experience of a woman getting angry at me for telling her that I love her isn't worth it. Even if I *were* able to make a favorable impression, I wouldn't want to date a woman who would hurt me so much *just because I approached her in the wrong way*."

Now I see that the women in question *had no idea what was going on*, so I can work on improving my communication skills. Once I get to the point of being able to communicate clearly, I can plausibly have a happy relationship.

(d) The experience made clear to me the extent to which the people who had appeared to be hostile toward me weren't hostile toward *me*, they were instead hostile toward *their construal of me*. They wouldn't have been at all hostile if they and known what was going on in my mind.

I had always known *on some level* that this was true, but I didn't *feel* it. I now have a deep understanding that there are many instances in which people *appeared* to be hostile toward me when their feelings weren't directed at *me*, instead they *didn't know enough about what was going on in my mind to be able to see that I wasn't the person who they thought that I was*.

I've developed the capacity to feel universal love and compassion the way Martin Luther King was able to. If somebody is angry at me and insults me, I know that it's not *me* who the person is insulting, it's instead *the person's perception of me*. So people can't hurt me anymore. Instead my response is "let me try to understand where the person is coming from, and help the person understand where I'm coming from."

This has made my life *so much better* than it had been before. I understand *intuitively* that Martin Luther King wasn't some sort of god, that he was human like you and me, and that the human race has the capacity to shift in his direction, and be much happier than we are now.

"Should" considered harmful

This is a linkpost for https://mindingourway.com/should-considered-harmful/

My last few posts have been aimed at addressing what I call the "listless guilt," the vague sense of guilt that stems from not doing anything in particular. I said:

> *The listless guilt is a guilt about not doing anything. To remove it, we must first turn it into a guilt about not doing something in particular.*

If you didn't have a listless guilt, or if you did and the last few posts worked for you, then you may now find yourself wrestling with a very *pointed* sort of guilt that stems from not doing *particular* things. These next few posts will address the pointed guilts.

One of the most common sources of pointed guilt that I encounter stems from neglected obligations. Imagine someone who thinks they should stop watching Netflix (because they care about something important, and watching Netflix isn't helping), but who can't seem to stop. Or imagine someone who thinks they should be spending more time working on their thesis, but can't make themselves do it. Or imagine someone who thinks they should be smarter, and that their homework shouldn't be taking them this long, and who feels worse and worse as they work. In each case, the pattern is the same: the subject thinks there's something they should be doing (or some way they should be), and they're not doing it (or aren't being it), and so they feel really guilty.

I claim that the word "should" is causing damage here.

In fact, as far as I can tell, the way that most people use the word "should," most of the time, is harmful. People seem to use it to put themselves in direct and unnecessary conflict with themselves.

For example, imagine the person who wakes up feeling a bit sick. They may well say to themselves, "ugh, I should go to the pharmacy and pick up medication before work." Now picking up meds feels like an obligation: if they don't get meds, then that's a little bit of evidence that they're incompetent, or akrasiatic, or bad. Now they *must* go get meds, if they want to be a competent person. In the lingo of CFAR, this "should" is the exact opposite of an urge-propagation: it disconnects the reason from the task, it abolishes the "why". The person feeling sick now feels like they have an obligation to pick up medication, and so if they do it, they do it grudgingly, resenting the situation. (And if they don't, then they've failed, and they're at risk of failing with abandon.)

Now imagine they say this, instead: "ugh, if I went to the pharmacy to pick up medication, I'd feel better at work today." Notice the difference? Now the reason remains attached to the task. Now neither option makes them "bad," and both options are tradeoffs.

I see lots of guilt-motivated people use "shoulds" as ultimatums: "either I get the meds, or I am a bad person." They leave themselves only two choices: go out of their way on the way to work and suffer through awkward human interaction at the pharmacy, or be bad. Either way, they lose: the should has set them up for failure.

But the actual options aren't "suffer" or "be bad." The actual options are "incur the social/time costs of buying meds" or "incur the physical/mental costs of feeling ill." It's just a choice: you weigh the branches, and then you pick. Neither branch makes you "bad." It's ok to decide that the social/time costs outweigh the physical/mental costs. It's ok to decide the opposite. Neither side is a "should." Both sides are an option.

Don't say "I really should finish this paper." Say "if I don't finish this paper, I'll get a worse grade than I was planning to, and my teacher will frown at me, and my parents will frown at me." Then weigh your options. Then choose.

This is not necessarily easy! Breaking a "should" into its component goals, tasks, and desires may be particularly difficult for people who are still confusing the quality line with the preference curve and forgetting that it's possible for their preferences to diverge from the expectations of others. I've often seen people confuse "an authority figure expects me to try hard to do X" or "my friends expect me to do X" with "I should do X," and many people find it very hard to tease these apart. (Future posts will touch on this a little.)

Unpacking a "should" can also be very difficult for a reason that's a little harder to articulate. Have you ever seen a person who can't even *imagine* the thought of failure start to fail? They start to panic, their actions get rushed, their hands start to shake (which is particularly fatal if their task is one requiring dexterity), they put on blinders to the fact that they're about to fail as they frantically repeat an action they wish would succeed over and over.

The ironic thing is, especially in timed, dexterity-based tasks, if the person *didn't* panic, they would have a better chance of succeeding. It seems to me that, more often than not, it's the fact that they can't even *consider* their failure that is harming them most. If only they had come to terms with failure beforehand, then they could keep a level head as failure looms, and this would buy them one or two more shots at success.

This is related to leaving yourself a line of retreat: If you find yourself *unable to think* about a certain outcome, it can be very useful to think all the way through the painful outcome — not to convince yourself that everything would actually be fine, but just so that you can *actually think about it.* It's the thoughts you can't think that really screw you.

Similarly, it's the options you can't weigh that really cost you. People often seem to use the word "should" to assign a value of "negative infinity" to all alternative actions. They should do X, so if they don't do X, they're *bad*, end of story. Some people have trouble unpacking a should for the same reason they have trouble staring at a failure: they have a mental geis against seriously considering alternatives, against weighing them on the scales. One common symptom of this behavior is a tendency to do a fake unpacking of the should, e.g. by translating "I should finish this paper" to "I *need* to finish this paper": notice how this trades one negative-infinity analysis for another, without ever reconnecting the task to the goal or acknowledging the alternatives.

I'm not saying that the alternatives are always good: perhaps the should unpacks into "I want to finish this paper, becaues if I don't, then I will very likely fail my course, lose my scholarship, get kicked out of college, disappoint my parents, and destroy my job prospects." The alternative options might be *really bad.* Yet, I claim that there is power in laying them all out, no matter how bad they are. Make the values finite, so you can actually weigh them on your scales. When you should yourself without looking at the

alternatives, you run a high risk of making yourself feel obligated and resentful. When you lay out all the options you can think of and choose the best, then it's much easier to work with yourself rather than against yourself — sometimes you have to settle for the best of a bad lot, but this is much easier once you've actually *looked at the whole lot.*

If you often suffer from guilt, then I strongly suggest cashing out your shoulds. Get a tally counter and start training yourself to notice when you say the word "should," and then once you're noticing it, start training yourself to unpack the sentence. "I should call my father this week" might cash out to "if I don't call my father this week, he'll feel disappointed and lonely." "I shouldn't play that video game" might cash out to "if I play that video game, I'll lose lots of time that I was planning to use for studying." "I should work on my homework right now" might cash out to "I want to have my paper finished by tomorrow, and I also want to go socialize right now, and these goals are mutually exclusive."

You can almost always re-state a should-sentence without the should. It may seem like a trivial transformation on sentences, but it might also really help remove the burden of an obligation.

Of course, cashing out your shoulds isn't all it takes to stop feeling guilty — not by a long shot. Once you've cashed out a should, you're often left with conflicting interests (remember that it's quite possible to disagree with yourself! I've seen people should themselves simply because they refuse to acknowledge that they might be under internal conflict). Frequently, after unpacking a should you're still left with a really hard choice. Furthermore, it's also quite common to cash a should out, weigh both options, decide that one option is better, and then *still find yourself doing the worse thing.* (This last problem is a doozy, and I'll discuss it more in future posts.) I'm not handing you a silver bullet, here.

But it's still a bullet. Don't use shoulds as an ultimatum! Your options are not divided up into "choices which make you good" and "choices which make you bad": your options are stratified by how much they move you towards the goal. So pick your shoulds apart into their component tasks and desires, and keep the tasks connected to the goal: don't say "I should get meds," say "I need to get meds if I want to feel good."

I've found it very helpful to treat almost all shoulds as a toxic attempt to blind me to the alternatives. Be careful: the thoughts you can't think do you harm, and the options you can't weigh cost you dearly.

So cash out your shoulds, and weigh *all* your options on the scales — and then choose what is best, free of obligation.

You're allowed to fight for something

This is a linkpost for https://mindingourway.com/youre-allowed-to-fight-for-something/

The first sort of guilt I want to address is the listless guilt, that vague feeling one gets after playing video games for twelve hours straight, a guilty feeling that you should be doing something else. Many people in my local friend group don't suffer from the listless guilt, because many people in my sphere are effective altruists who feel a *very acute and specific* sense of guilt when they think they've spent their time poorly. Specific guilt tends to be as bad or worse than the listless guilt, but before I address specific guilt, I need to confront the listless guilt.

It seems to me that the listless guilt usually stems from not doing anything in particular. I'm not sure how to remove that feeling of guilt in people who aren't doing anything in particular. But if they shift the guilt to being guilty about not doing *one thing in particular*, then I have some tools that might help.

Warning: in this post, I'm going to encourage people with listless guilt to find something to care about, and to shift their guilt away from a vague sense of not doing anything towards a specific sense of not doing one thing in particular. If you already have strong specific guilts, consider skipping this post.

The message of the allegory of the stamp collector is this: *you can care about things in the world.* There is no difference in kind between steering reality towards futures where there are more happy-chemicals in your head, and steering reality towards futures where there are lots of happy humans outside your head. Your decision process is implemented by the lump of meat between your ears, but it builds a map of the entire universe, and you can act (according to the map) towards whatever end you please.

You only ever see the map, but you walk the territory.

Many people will say that humans only ever do what they want. They wheel out phrases such as "revealed preference" and say that no matter what people do, they do it because they wanted to. But here's the thing:

If you use the word "want" to mean "whatever humans actually do," then I need new words to differentiate activities-I-do-for-personal-enjoyment (stargazing, studying physics, tinkering, cavorting) from activities-I-do-for-the-sake-of-others-I-care-about (attempting to reduce existential risk, donating to charities, community service). These are very different clusters of behavior that feel very different, and I need words to distinguish between them.

If a word describes everything, then it distinguishes nothing, and is useless. If you use the word "want" to mean "whatever people do," then it can't be used for talking about actions. In order for "wants" to be *about* goals humans are trying to achieve for various purposes, it must apply to some goals and not others.

I'm happy to *split* the word "want," because it's a pretty loaded word. Sometimes I use it to distinguish between the stargazing/cavorting cluster and the charity/altruism cluster, and other times I use it to distinguish between tasks-I-reflectively-approve-of-doing (such as studying an interesting topic) and tasks-I-reflectively-disapprove-of-

doing (such as procrastinating by reading boring web pages), which is a different way of cutting up things-I-do that I also find useful.

Distinguishing between clusters of things is what words are *for.* If anything, we need to make the word "want" *more* specific, not less specific.

Nihilists may tell you that nothing matters, that there is no altruism, that people only do what they want to, and these are all traps that lead to the listless guilt. They help people half-convince themselves that nothing matters, and then the other half of them, which fails to be fooled, goes on yearning for something more.

So if you're experiencing nihilism along with a vague sense of discomfort or guilt, consider taking a moment to remind yourself that it *is* possible for you to care about things beyond yourself, for non-selfish reasons.

I've been surprised, in the past, by how many people vehemently resist the idea that they might not actually be selfish, deep down. I've seen some people do some incredible contortions in attempts to convince themselves that their ability to care about others is actually completely selfish. (Because iterated game theory says that if you're in a repeated game it pays to be nice, you see!) These people seem to resist the idea that they could have selfless values on general principles, and consistently struggle to come up with selfish explanations for their altruistic behavior.

Don't get me wrong, selfishness is fine. Yet, true selfishness doesn't lead to the listless guilt. If you think you *must* be selfish, and you also feel vaguely guilty about life, then perhaps you care about what goes on beyond your head.

In case you're skeptical, here's a little thought experiment:

> *Imagine you live alone in the woods, having forsaken civilization when the Unethical Psychologist Authoritarians came to power a few years back.*
>
> *Your only companion is your dog, twelve years old, who you raised from a puppy. (If you have a pet or have ever had a pet, use them instead.)*
>
> *You're aware of the fact that humans have figured out how to do some pretty impressive perception modification (which is part of what allowed the Unethical Psychologist Authoritarians to come to power).*
>
> *One day, a psychologist comes to you and offers you a deal. They'd like to take your dog out back and shoot it. If you let them do so, they'll clean things up, erase your the memories of this conversation, and then alter your perceptions such that you perceive exactly what you would have if they hadn't shot your dog. (Don't worry, they'll also have people track you and alter the perceptions of anyone else who would see the dog, so that they also see the dog, so that you won't seem crazy. And they'll remove that fact from your mind, so you don't worry about being tracked.)*
>
> *In return, they'll give you a dollar.*

Under the assumption that you will in fact believe and perceive the same things you would have if they hadn't shot the dog, *and* have an extra dollar for your trouble, would you take the offer?

Most people reject it. You're allowed to reject it! You're allowed to reject *arbitrarily good* amounts of faked pleasure-experience in order to avoid bad real-world outcomes. You're allowed to care about whether your beliefs are actually hooked up to reality. You're allowed to care about things outside of you!

One friend of mine, after probing around in thought experiments such as this one, said "Huh. Well, so I definitely care about myself experiencing pleasure, and also I seem to care about other people actually existing and experiencing pleasure, though I don't know why."

She seemed surprised and confused to notice that she cared about others, as though this fact demanded explanation.

You don't need an excuse. You can just care about things outside yourself.

If you have the listless guilt, if something seems like it's missing in life, if it seems like there's something else you should be doing with your time, then probe the feeling. Figure out what's missing.

Maybe start by saying, aloud, "I can care about how the world is," and "I want the world to be different than it currently is," if that helps unstick something. And then *listen* to that listless feeling saying there must be something more, and look at the world with fresh eyes, and ask yourself what is wrong. Ask yourself what you would like to see changed.

Is the world totally perfect? No? What would you change, if granted omnipotence? Do you want to acquire power, fame, or riches? Do you want to reduce inequality? Do you want to make it easier for humans to connect? Do you want to reduce loneliness and despair? Do you want to put an end to disease and suffering? Do you want to slay Moloch, the avatar of a runaway civilization that chews humans to pieces, twisted them into bitter shells of their former selves by forcing them to take degrading jobs in order to survive?

Don't just look for ideas that sound nice. Look for changes in the world that *compel* you, ideas such that thinking them makes something move in your chest. Look for places where the world is broken and in need of fixing. Look for things in the world that are *unacceptable*. Reject the natural order.

It doesn't have to be a grand and ambitious desire. Maybe you'll just want more personal gain. Maybe you'll find that there's one person in particular who you want to save, one person trapped in a hellhole that you want to shield them from. Or maybe you'll decide that you want to save the entire damn world. I don't know. But if you want to remove the listless guilt, then step one is finding something to fight for.

Step zero is *believing that you can.*

Lots of people seem to have these blinders on: the world is big and they are small, and they're just trying to scrape together a living or get by with skills that don't seem particularly relevant to their ambitions, and they don't have the time or ability or energy to make things better. And so they try not to think about it, and then they forget that they're allowed to have a way they want the future to be, that they're allowed to have a specific vision for what they want to achieve.

They forget that they're allowed to desperately want the future to be different from the present.

Finding something to fight for won't *eliminate* the listless guilt. In fact, it may do the opposite: it may refine the listless guilt into a more pointed thing, a guilt about not making the world better *right now.* It may make you feel guilty about there being so much wrongness and badness that you're not confronting, that you *can't* confront. That's OK: the goal of this exercise is not to eliminate the listless guilt, but to *shift* it. The pointed guilt is more painful, but easier to replace with intrinsic motivation.

The listless guilt is a guilt about not doing anything. To remove it, we must first turn it into a guilt about not doing *something in particular.*

If, instead of feeling vaguely guilty for binging netflix due to the feeling that there must be more to life, you feel *specifically* guilty because you could have been pursuing some concrete end, then we've made progress. The latter guilt, though often much more painful, is easier to address.

Caring about something larger than yourself

This is a linkpost for https://mindingourway.com/caring-about-some/

In my last post, I said that in order to address the listless guilt, step zero is believing that you can care about something, and step one is finding something to care about. This post is about step one.

There are many different ways to care passionately about one thing or another. Parents in particular are usually good at step one, and often care strongly about the welfare of their children. Others care strongly about their family, or the environment, or what-have-you. Many others claim to care about all humanity or about all sentient life.

On the other hand, some people have significantly more trouble caring about big things. They don't have any children to die for and they don't see the point in caring about everyone, and yet many of them still possess the listless guilt. When I suggest to such a person that they address their guilt by searching their motivations and finding something to care about, the response, more often than not, is simply "Why?"

This post is for them.

In order to answer, I'm going to talk about *my* answer to this "why?". Before we continue, I stress that my answer is not the only one, that my cause is not the only one, and that I endorse anyone's desires to pursue whatever it is they care deeply about, regardless of their cause. As with previous posts, don't treat this as a sermon about why you *should* care about things that are larger than yourself; treat it as a reminder that you *can,* if you want to.

I often encounter people who don't care much about humanity at large (or the future of sentient life), but seem vaguely curious as to why somebody would. When I suggest that it is possible for them, too, to care about things greater than themselves, the most common response by far, is "sure, but why would I want to do that?"

Why fight for humans? Why care about the fate of the Earth, or the fate of people we will never meet? Why care about the callous species that invented war and torture? Why care for people at large, when most of them are stupid or annoying or members of the wrong political party or possessing of incorrect beliefs? Most humans are *annoying*, so why would you possibly want to care about them?

I have encountered many people who claim that they only care for their immediate friend group.

Now, if you *actually* only care about your immediate friend group, then be it not upon me to change your preferences. Yet, in my experience, people who *think* they only care about their immediate friend group tend to be confused.

One friend of mine insisted that he only cares about the people he's close to, while simultaneously putting privacy concerns (e.g. privacy of communication over the internet) very high on his priority list. When I asked why, he claimed (after some

exploring) that it's because he cares about the autonomy and freedom of people in general. Noticing the inconsistency, he quickly added that he only cares about autonomy and freedom for the masses *because of the pleasurable feeling this creates within him;* it was of course a *selfish* desire, and he *still* only cared about the people close to him. (That was in fact the conversation where I first concocted the allegory of the stamp collector.)

What's going on, here? One thing, I think, is a tendency to confuse *feelings* with *caring.* Most people only have strong feelings of affection for their close friends, and they don't have feelings that are nearly so strong for nameless strangers, and so they conclude that they must not care about strangers. They forget that feelings and caring are separate things! I reassure you that I, too, have deeper feelings for people close to me than for strangers — but I still care about the strangers anyway. In fact, I suspect this is true of nearly everybody who claims to care about humanity at large. Courage isn't about not being afraid, it's about being afraid and doing the right thing anyway; and similarly, caring isn't about being overwhelmed by emotion, it's about not having the emotional compulsion and *doing the right thing anyway.* It's possible to both lack deep feelings of affection for strangers *and still care for them nearly as much as you care for friends.*

This is at least one reason why I think people tend to insist that they don't care about strangers, but it still doesn't answer the "why." Even once people admit it's *possible* to start acting like they care about humanity at large, they still tend to wonder why in the world they would ever want to do such a thing.

And I can't tell you whether or not you want to do this. But I can tell you why *I* wanted to do this, and at least help you understand why someone would.

We humans are reflective creatures: we get to examine what we feel and what we care about, and choose to change ourselves. As it happens, when I reflect upon myself and my desires, I find many that I approve of, and some that I don't.

I, like many, spend a large chunk of time frustrated by other human beings (especially when they fail to read my mind). I have unconscious biases against people who don't look sufficiently similar to the people I grew up near. I automatically bristle at members of my outgroup. I'm uncomfortable around vast segments of the population. And yet, *at the same time*, I care about all people, about all of Earth's children, about all sentient life.

Why? In large part, by choice. My default settings, roughly speaking, make it easy for me to feel for my friends and hate at my competitors. But my default settings *also* come with a sense of aesthetics that prefers fairness, that prefers compassion. My default feelings are strong for those who are close to me, and my default sensibilities are annoyed that it's not possible to feel strongly for people who *could have been* close to me. My default feelings are negative towards people antagonizing me, and my default sensibilities are sad that we didn't meet in a different context, sad that it's so hard for humans to communicate their point of view.

My point is, I surely don't lack the capacity to feel frustration with fools, but I *also* have a quiet sense of aesthetics and fairness which does not approve of this frustration. There is a tension there.

I choose to resolve the tension in favor of the people rather than the feelings.

Why? Because when I reflect upon the source of the feelings, I find arbitrary evolutionary settings that I don't endorse, but when I reflect upon the sense of aesthetics, I find something that goes straight to the core of what I value.

Because when I reflect, I see that I am an inconsistent mess of a brain born of a long and blind evolutionary process, full of desires and feelings and fears that capture everything I hold dear, and also a bunch of arbitrary junk that was kind of tacked on there. In making me, Time coughed up a *reflectively unstable* mind: the causal process of my past constructed me to value everything I value, and some things that I (upon reflection) don't.

So I look upon myself, and I see that I am constructed to both (a) care more about the people close to me, that I have deeper feelings for, and (b) care about fairness, impartiality, and aesthetics. I look upon myself and I see that I *both* care more about close friends, *and* disapprove of any state of affairs in which I care more for some people due to a trivial coincidence of time and space.

And I am constructed such that when I look upon myself and find inconsistencies, I care about resolving them.

So, why do I care about humanity? Because, for me, resolving this inconsistency is easy. My strong feelings are in conflict with my quiet aesthetics, but when push comes to shove, the quiet aesthetics win hands-down. To me, the feelings look like they are arbitrary remnants of the tribal days, while the aesthetics look like they are echoes of my deeper values. I know which one *I'm* more loyal to.

This is not a knock-down argument, by any means. One person's modus ponens is another person's modus tollens, and some people, looking upon themselves, would prefer to forgo a sense of fairness and impartiality instead of choosing to care about strangers. But I, and many others, don't *want* to care only about our friends. We feel more loyalty to our aesthetics than our default feelings — and so the choice is easy.

Caring about others may sound great in theory, but for jaded and cynical people (who can't stand interacting with idiots), the points above probably aren't enough.

And you know what? It can be *really hard* to muster any feeling of caring for humans, even if you've decided that you want to.

It's too easy to look at them and see the tarnished, ugly, greedy, stupid species.

It's too easy to look at individuals and see idiots.

(I have this feeling too, sometimes.)

But here's something strange:

Imagine you've had a pet dog that you've raised from a puppy, and grown close to over the course of a decade. Imagine somebody napped your dog and started harming it, for fun.

How would this make you feel? How much would you like to find this person, and bring them to justice?

Most people are able to feel a much larger burst of empathy and caring for suffering animals than for suffering humans.

Imagine you're being mugged by a homeless man in an alley. Someone notices, comes to help, push comes to shove, they scare the man off, and then ask if you're all right. Now imagine a stray dog growling at you in an alley. Someone notices, comes to help, kicks the dog when it won't back down, scares it off, and asks if you're all right.

Does it feel inconsistent to you, the difference between the way you feel for mistreated animals, versus the way you feel towards mistreated humans? Does it seem strange, how easy it is to like dogs, how difficult it is to like men?

You may, of course, conclude that you actually don't like men. But you don't *have* to. You can, as before, listen to the quiet sense of aesthetics that is in conflict with your default feelings. Why are our default feelings hooked up how they are? I can't say for sure, but here's a theory:

> *An influential version of social theory is the 'Machiavellian Intelligence' hypothesis (Byrne and Whiten 1988; Whiten and Byrne 1997). Social interactions and relationships are not only complex but also constantly changing and therefore require fast parallel processing (Barton and Dunbar 1997). The similarity with Niccolò Machiavelli (1469—1527), the devious adviser of sixteenth-century Italian princes, is that much of social life is a question of outwitting others, plotting and scheming, entering into alliances and breaking them again. All this requires a lot of brain power to remember who is who, and who has done what to whom, as well as to think up ever more crafty wiles, and to double bluff the crafty wiles of your rivals — leading to a spiralling arms race. 'Arms races' are common in biology, as when predators evolve to run ever faster to catch their faster prey, or parasites evolve to outwit the immune systems of their hosts. The notion that some kind of spiralling or self-catalytic process is involved certainly suits what Christopher Wills (1993) calls 'the runaway brain', and this idea is common among theories that relate language evolution to brain size.*

(Sue Blackmore, *The Meme Machine*)

I mean, look at us. Humans are the sort of creature that sees lightning and postulates an angry sky-god, because angry sky-gods seem much more plausible to us than Maxwell's equations — this despite the fact that Maxwell's equations are *far* simpler to describe (by a mathematical standard) than a generally intelligent sky-god. Think about it: we can write down Maxwell's equations in four lines, and we can't yet describe how a general intelligence works. Thor *feels* easier for us to understand, but only because we have so much built-in hardware for psychologically modeling humans.

Our brains are hard-wired to see human-like agents *everywhere.* Cartoons work: we see them as people (and attribute feelings to them) despite their simplicity. We see intentionality everywhere — religious folks have no trouble finding apparent affirmation that their mundane actions are part of some grand plan, superstition runs rampant, and many different types of mental disorders (schizophrenia, mania, etc.) are characterized by delusions that either everybody is against you or that your entire life has been carefully engineered — symptoms of a brain over-eager to see things in terms of human plots and schemes.

When we look at humans, we see them as plotters or schemers or competition. But when we look at puppies, or kittens, or other animals, none of that social machinery

kicks in. We're able to see them as just creatures, pure and innocent things, exploring an environment they will never fully understand, just following the flow of their lives.

If you back a puppy into a corner and frighten it, and it snaps at you, it's easy to feel a wave of compassion rather than hatred.

But when a *human* snaps at you, the social machinery engages. It's easy to get stuck inside the interaction. When a human is backed into a corner and lashes out, we tend to lash back.

Which is why, every so often, I take a mental step back and try to *see* the other humans around me, not as humans, but as innocent animals full of wonder, exploring an environment they can never fully understand, following the flows of their lives.

I try to see people in the same way I would see a puppy, reacting to pains and pleasures, snapping only when afraid or threatened. I try to see the tragedies in humans who have been conditioned by time and circumstance to be suspicious and harmful, and feel the same compassion for them that I would feel for an abused child.

I look at my fellow humans and strive to remember that they, too, are innocent creatures.

Someone told me once that, in order to feel compassion for others, it's useful to visualize them as having angel's wings. I think there's something to this. There's something powerful about looking at people and seeing the angels that never had a shot at heaven — though I prefer to see not angels, but monkeys who struggle to convince themselves that they're comfortable in a strange civilization, so different from the ancestral savanna where their minds were forged.

Some use 'animal' as a derogatory, and may think that it's demeaning to try to see humans as animals instead of people. For me, the opposite is true, for the same reason that it's easier to feel compassion for a homeless dog than a homeless man — it helps me, to detach my automatic impulses to see other humans as competitors or allies or enemies, and just look at them the same way I would look at a kitten, as a pure creature possessing of the same wonder and innocence.

Why do I care about humans and humanity, about Earth and all its children, about all sentient life? How can I say I do given that I, too, often feel more strongly for friends than strangers, and more compassion for dogs than men?

When I look upon myself, I see a tension between what I feel and a sense that my feelings are ill-calibrated. When I look closer, I find that the feelings are calibrated in ways I don't endorse, in a tribal setting, where it was important to love the ingroup and hate the outgroup. But when I look at the sense that those feelings are ill-calibrated, I find *good* reasons, and a sense that this is *actually* what matters, that it is not arbitrary but valuable.

And so for me, "why care?" has an easy answer.

Let me stress again that you don't have to resolve your internal tensions in the same way I do. Your answer to "why care?" might be "I don't." You might side more with your current feelings over your deeper sense of aesthetics, or you might have very different feelings and aesthetics. Either way, if you *listen* to that internal sense of friction, if you use your feelings as a guide rather than an answer, if you figure out

why you feel and care as you do, and reflect upon your reasons, and separate feeling from caring, and choose to care about what seems right and good to care about —

then you may find that "why care?" has an easy answer for you, too.

You don't get to know what you're fighting for

This is a linkpost for https://mindingourway.com/you-dont-get-t/

A number of my recent posts may have given you the impression that I know exactly what I'm fighting for. If someone were to ask you, "hey, what's that Nate guy trying so hard to do," you might answer something like "increase the chance of human survival," or "put an end to unwanted death" or "reduce suffering" or something.

This isn't the case. I mean, I am doing those things, but those are all negative motivations: I am *against* Alzheimer's, I am *against* human extinction, but what am I *for?*

The truth is, I don't quite know. I'm for *something*, that's for damn sure, and I have lots of feelings about the things that I'm fighting for, but I find them rather hard to express.

And in fact, I highly doubt that *anyone* knows quite what they're fighting towards — though it seems that many people think they do, and that is in part why I'm writing this post.

When I wrote on rationality, one commenter replied:

> *I would just note upfront that*
>
> *> Reasoning well has little to do with what you're reasoning towards.*
>
> *and*
>
> *> Rationality of this kind is not about changing where you're going, it's about changing how far you can go.*
>
> *are white lies, as you well know. It's not unusual in the process of reasoning of how to best achieve your goal to find that the goal itself shifts or evaporates.*
>
> "*How to best serve God*" *may result in deconversion.*
>
> "*How to make my relationship with partner a happy one*" *may result in discovering that they are a narcissistic little shit I should run away from. Or that both of us should find other partners.*
>
> "*How to help my neighborhood out of poverty*" *might become* "*How to make the most money*" *in order to donate as much as possible.*

This is a fine point. Humans are well-known for their ability to start out pursuing one goal, only to find that goal shift drastically beneath them as their knowledge of the world increases. In fact, this is a major plot point in many stories (such as, say, The Foundation Trilogy, The Dresden Files, and The Neverending Story). The goal you think you're pursuing may well not survive a close examination.

I claim this is true even if you think your goals are simple, objective, obvious, high-minded, or sophisticated. Just as the theist setting out to do the most good might

deconvert after deciding that they would still want humanity to flourish even without a divine mandate, so may the utilitarian setting out to do the most good discover that their philosophy is incoherent.

In fact, I suspect this is *inevitable*, at least at humanity's current stage of philosophical development.

It's nice and clean and *easy* to say "I'm a total hedonic utilitarian," and feel like you know exactly what you value. But what does it mean, to be a utilitarian? What counts as a mind? What counts as a preference? Under whose interpretation, under whose process, are preferences extracted? Do you feel an obligation to create people who don't exist? Does a mind matter more if you run two copies of it side by side? I doubt these questions will have objective answers, but subjective resolutions will be complex and will depend on what we value, in which case "total hedonic utility" isn't really an answer. You can *say* you're fighting for maximum utility, but for now, that's still a small label on a complex thing that we don't quite know how to express.

And even if we could express it, I doubt that most humans are in fact total hedonic utilitarians. Imagine that an old friend of yours eats a sandwich which (unexpectedly) alters their preferences so that all they want to do all day is stare at a white wall and not be disturbed. Do you feel a moral obligation to help them find a white wall and prevent others from disturbing them? If there was a button that resets them to as they were just before they ate the sandwich, would you press it? I sure as hell would — because I feel loyalty not only to the mind in front of me, but to the *person*, the *history*, the *friend*. But again, we have departed the objective utilitarian framework, and entered the domain where I don't quite know what I'm fighting for.

If I am loyal to my old friend over the person who sits in front of the white wall, then am I also obligated to "save" people who naturally want to wirehead? Am I obligated to the values they had as a teenager? Am I obligated to maximize the utilities of babies, before they grow up?

I'm not saying you can't answer these questions. I'm sure that many people have. In fact, I'm sure that some people have picked simple-enough arbitrary definitions and then bitten all the associated bullets. ("Yes, I care about the preferences of rocks a little bit!" "Yes, I maximize the utility of babies!", and so on.) And I'm picking on the utilitarians here, but the same goes for the deontologists, the theists, and everybody else who thinks they know what they're fighting for.

What I'm saying is, even if you *say* you know what you're fighting for, even if you *say* you accept the consequences and bite the bullets, *it's possible for you to be wrong about that.*

There is no *objective* morality writ on a tablet between the galaxies. There are no objective facts about what "actually matters." But that's because "mattering" isn't a property of the universe. It's a property of a *person.*

There *are* facts about what we care about, but they aren't facts about the stars. They are facts about *us.*

There is no objective morality, but also your morality is *not* just whatever you say it is. It is possible for a person to say they believe it is fine to kill people, and *be lying.* The mind is only part of the brain, and it is possible to have both (a) no objective morality, and (b) people who are wrong about what they care about.

There are facts about what you care about, but you don't get to know them all. Not by default. Not yet. Humans don't have that sort of introspective capabilities yet. They don't have that sort of philosophical sophistication yet. But they *do* have a massive and well-documented incentive to convince themselves that they care about simple things — which is why it's a bit suspicious when people go around claiming they know their true preferences.

From here, it looks very unlikely to me that anyone has the ability to pin down exactly what they really care about. Why? Because of where human values came from. Remember that one time that Time tried to build a mind that wanted to eat healthy, and accidentally built a mind that enjoys salt and fat? I jest, of course, and it's dangerous to anthropomorphize natural selection, but the point stands: our values come from a complex and intricate process tied closely to innumerable coincidences of history.

Now, I'm quite *glad* that Time failed to build a fitness maximizer. My values were built by dumb processes smashing time and a savannah into a bunch of monkey generations, and I don't entirely approve of all of the result, but the result is also where my approver comes from. My appreciation of beauty, my sense of wonder, and my capacity to love, all came from this process.

I'm not saying my values are dumb; I'm saying you shouldn't expect them to be simple.

We're a thousand shards of desire forged of coincidence and circumstance and death and time. It would be *really surprising* if there were some short, simple description of our values. Which is why I'm always a bit suspicious of someone who claims to know exactly what they're fighting for. They've either convinced themselves of a falsehood, or they're selling something.

Don't get me wrong, our values are not *inscruitable.* They are not *inherently unknowable.* If we survive long enough, it seems likely that we'll eventually map them out.

But we don't know them *yet.*

That doesn't mean we're lost in the dark, either. We have a hell of a lot of *evidence* about our values. I tend to prefer pleasure to pain and joy to sadness, most of the time. I just don't have an exact description of what I'm working towards.

And I don't *need* one, to figure out what to do next. Not yet, anyway. I can't tell you exactly where I'm going, but I can sure see which direction the arrow points.

It's easier, in a way, to talk about the negative motivations — ending disease, decreasing existential risk, that sort of thing — because those are the things that I'm pretty sure of, in light of uncertainty about what really matters to me. I don't know exactly what I want, but I'm pretty sure I want there to be humans (or post-humans) around to see it.

But don't confuse what I'm *doing* with what I'm *fighting for.* The latter is much harder to describe, and I have no delusions of understanding.

You don't get to know exactly what you're fighting for, but the world's in bad enough shape that you don't *need* to.

In order to overcome the listless guilt, I strongly recommend remembering that you have something to fight for, but I also caution you against believing you know exactly what that thing is. You probably don't, and as you learn more about the world, I expect your goals to shift.

I'll conclude with a comic by Matt Rhodes:


My Hero

Not because you "should"

This is a linkpost for https://mindingourway.com/not-because-you-should/

A few months ago, a friend of mine was describing her motivational issues to me. As an example, she explained she was having trouble making herself clean her room, despite her dissatisfaction with the constant messiness.

I asked: "Have you considered just not forcing yourself?"

She blinked, and cocked her head at me, and said "but then my room wouldn't get cleaned."

I called bullshit. Because look: either (a) you stop forcing yourself to clean the room, and you realize you don't actually care about having a clean room, and then your room stays messy *and that's fine because you don't care*; or (b) you stop forcing yourself to clean the room, and then you get a bit worried, because some part of you *actually wants the room cleaned*, so you listen to that part of yourself, and you work with it, and you find a time to clean the room *because you want to.*

Either way, you win. No need to use internal force.

This is a technique I've recommended before for motivational issues, and I recommend it again when dealing with shoulds. If you struggle with feelings of guilt, obligation, or inadequacy, then I strongly suggest the following remedy:

Just stop doing things because you "should".

As in, never let a "should" feel like a reason to do something. Only do things because they seem like the best thing to do after you've thought about it; never do things just because you "should."

A commenter to my last post said:

> *There's some meaning lost when you go from "I should X" to "If I X, I will achieve Y", which is "And I want to achieve Y enough to X, that's the best of the options."*

I think this is mostly correct. Only mostly, because as far as I can tell most people don't tend to use "I should X" to mean "X is my best option." More frequently, I see people use it to mean "I would conclude that X is my best option *if I knew more facts*," or "I would conclude that X is my best option *if I thought longer*", or "I would conclude that X is my best option *if I really cared about what I say I care about*."

Regardless, all these various interpretations of "I should X" share one property: It's extremely difficult to make these claims about X *while you're still deliberating.*

If you ever happen to figure out which option is best, then don't slap the label "should" on it and go back to thinking about your options! If you know what the best option is, then stop deliberating and *do it.*

After the fact, looking back, you are welcome to say "ah, knowing what I know now, I see that pressing the green button would have been better." But *in the moment*, all

you can do is evaluate all of your actions and see which one looks best given the information available. Shoulds are for retrospectives, not for deliberation.

What you *should* do is the option that actually seems best when you're done weighing your options, regardless of whether or not it has a "should" label attached. You can't figure out which action actually seems best by slapping "should" labels on options willy-nilly and then feeling bad when you ignore them.

Imagine you're trying to solve an algebra problem, with the following method:

1. Say to yourself, "The answer is going to be x=17. I know it."
2. Look at the problem. The problem is "2x = 12; solve for x."
3. Conclude the answer is x=6, and then feel really guilty because x wasn't 17.

This is not the best method ever for solving algebra problems. A better method might be to look at the problem *first*, without deciding what the answer is in advance or feeling guilty when it turns out you aren't prescient.

For the same reason, it's a bit silly to slap a "should" label on all your actions before you actually know which action seems best!

I've seen many people use the word "should" to highlight a conflict between what they perceive as desires and what they perceive as moral obligations. For example, they might say "well I *want* to buy this ice cream, but I *should* donate the money to the Against Malaria Foundation instead."

I say, this is a false conflict. Imagine this person precommiting to never doing anything just because they "should." How might they feel?

They might feel relieved, because they *actually didn't* care about helping others, not even a little bit. So they discharge their guilt, buy their ice cream, and go on their merry way.

But more likely (in someone who thought they "should" give to AMF), that would feel a little bad, and a little hollow. This person, when committing to never do things because they "should," might feel a bit of fear. They might worry that if they didn't keep themselves in check then they'd *never* do anything to help those less fortunate than themselves. That might seem bad, to them.

Which lets them actually see the true problem, for the first time: they *both* want to buy the ice cream *and* help those who are worse off than them. Now they can actually weigh both desires on the scales, or search for clever third options that fulfill both desires, and so on.

This is a big part of where guilt-free effective altruism comes from, I think: instead of forcing yourself to give to charities sporadically when the guilt overcomes you, promise yourself that you *won't* give sporadically due to guilt, and then *listen* to the part of you that says "but then when will I help others!?" Don't force yourself to be an altruist — instead, commit to *never* forcing yourself, and then work with the part of you that protests, and become an altruist *if and only if you want to help.*

Some people, when they stop forcing themselves to do things because they "should," will do a bit less to improve the world. They'll bow a bit less to social pressure, and insofar as the social pressure was pushing them to do what you think is good, you

might count that as a loss. Some people *don't* care about things larger than themselves, and that's *perfectly fine*, and making them more resilient to social pressure might lose the world some charity.

But I expect that far *more* charity is lost from people convincing themselves that their altruistic desires are external obligations and then resenting them. I expect that *most* people who feel obligated to improve the world and only do it because they "should" will become much more effective if they stop forcing themselves.

It might take them a while. There might be some backlash from years of using internal violence to fulfill a moral obligation that felt more like a bitter duty than a deep desire. Maybe when they first cut themselves free of the "shoulds" they'll go on a self-indulgent hedonistic spending spree. But most of them, I expect, will make their way back. Maybe they'll have to struggle through the listless guilt, maybe they'll have to do a lot of soul searching in order to figure out what they're really fighting for, but once they do, they'll be back stronger than ever.

A little while back, I said

> *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.*

And this is a big part of it. If you're going to struggle on the side of Earth and all its children, I expect you can pull harder if you're pulling because you want to, not just because you should.

Imagine promising yourself that you're never going to do something just because you "should," ever again. How does that make you feel?

Do you feel relieved? If so, then you were probably putting your "should" labels on the wrong things and forcing yourself to do things that weren't actually best.

Alternatively, do you feel anxious and worried? Is your mind saying "but wait, if I don't force myself to do what I should, then I'll never get anything done, and I'll lose my job, and I'll never help those less fortunate than myself, and that's bad!"? Because in that case, *listen to those concerns* when you're making your choices. Engage with that part of yourself. You may still decide to do a bunch of unpleasant work, but at least now you'll be doing it because it's better than the alternative, rather than because you're forcing yourself.

(There's still this one hitch where you decide A is best and find yourself doing B anyway; we'll get to this a few posts down the line.)

When you're making a decision, never let the force of action come from a "should." The "should" label is what you place on actions *after* you decide they're best. It's the label you place retrospectively on the answer, not something that can compel you towards the answer.

When you're deliberating, your only responsibility is to figure out which action seems best given the available time and information. Leave the "shoulds" to the historians.