

The Replacing Guilt series

Preliminaries

- Half-assing it with everything you've got
- Failing with abandon

1. Fighting for something

- Replacing guilt
- The stamp collector
- You're allowed to fight for something
- Caring about something larger than yourself
- You don't get to know what you're fighting for

2. Drop your obligations

- "Should" considered harmful
- Not because you "should"
- Your "shoulds" are not a duty

3. Half monkey, half god

- Working yourself ragged is not a virtue

- Rest in motion
- Shifting.guilt
- Don't steer with guilt
- Update from the suckerpunch
- Be a new homunculus
- Not yet gods
- Where coulds go
- Self compassion
- There are no "bad people"
- Residing in the mortal realm

4. The dark world

- Being unable to despair
- See the dark world
- Choose without suffering
- Detach the grim-o-meter
- Simply locate yourself
- Have no excuses
- Come to your terms
- Transmute guilt into resolve
- The best you can
- Dark, not colorless

5. Fire within

- Stop trying to try and try

- There is no try
- Obvious advice
- The art of response
- Confidence all the way up
- Desperation
- Recklessness
- Defiance
- How we will be measured

Related

- On caring
- The value of a life
- Moving towards the goal
- Self-signaling the ability to do what you want
- Productivity through self-loyalty

Conclusion.

Half-assing it with everything you've got

13 MARCH 2015

I hang out around a lot of effective altruists. Many of them are motivated primarily by something like guilt (for having great resource and opportunity while others suffer) or shame (for not helping enough). Hell, many of my non-EA friends are primarily motivated by guilt or shame.

I'm not going to criticize guilt/shame motivation: I have this policy where, when somebody puts large amounts of effort or money towards making the world a better place, I try really hard not to condemn their motives. Guilt and shame may be fine tools for jarring people out of complacency.

However, I worry that guilt and shame are unhealthy long-term motivators. In many of my friends, guilt and shame tend to induce akrasia, reduce productivity, and drain motivation. So over the next few weeks, I'll be writing a series of posts about removing guilt/shame motivation and replacing it with something stronger.

1

Say you're a college student, and you have a paper due. The quality of the paper will depend upon the amount of effort you put in. We'll say that you know the project pretty well: you can get an A with only moderate effort, and with signifi-

cant effort you could produce something much better than the usual A-grade paper.



The education environment implicitly attempts to convince students that their preferences point ever rightward along this line. Parents and teachers say things like "you should put in your best effort," and they heap shame upon people who don't strive to push ever rightward along the quality line.

People generally react to this coercion in one of two ways. The first group (the "slackers") rejects the implication that quality=preferences. These are the people who don't care about the class, who complain constantly about the useless pointless work they have to do, who half-ass the assignment and turn in something that either barely passes or fails entirely. Slackers tend to resent the authority forcing them to write the paper.

The second group (the "tryers") are the ones who accept the premise that quality=preferences, and strive ever rightwards on the quality line. Tryers include people of all ability levels: some struggle as hard as they can just to get a C, others flaunt their ability to produce masterpieces. Some try to curry favor with the teacher, others are perfectionists who simply can't allow themselves to turn in anything less than

their best effort. Some of them are scrupulous people, who feel guilty even after getting an A, because they know they could have done better, and think they should have. Some are humble, some are show-offs, but all of them are pushing rightward.

Society has spent a *lot* of time conditioning us to think of the tryers as better than the slackers. Being a tryer is a virtue. Slackers are missing the point of education; why are they even there? The tryers are going to go places, the slackers will never amount to anything.

But in fact, both groups are doing it wrong.

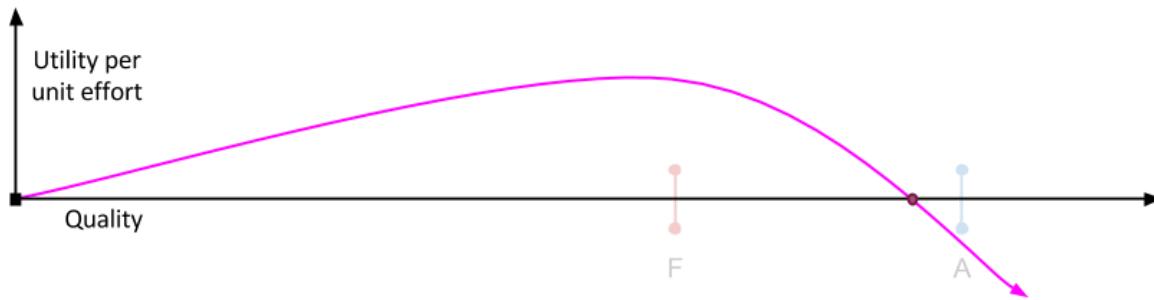
If you want to be highly effective, *remember what you're fighting for.*

And, spoiler alert, you aren't fighting for "write a high-quality paper." That would be a pretty silly thing to fight for.

What is your goal in taking this class? Perhaps you're doing it thanks to a combination of social pressure (your parents said so), social inertia (everybody else goes to college), and a vague belief that this is the path towards a good job and a comfortable life. Or perhaps you're there because you want good grades so you can acquire lots of money and power which you will use to fight dragons. Or perhaps you're there out of a genuine thirst for knowledge. But no matter why you're there, your reason for being there will pick out a sin-

gle target point on the quality line. Your goal, then, is to hit that quality target — no higher, no lower.

Your preferences are not "move rightward on the quality line." Your preferences are to *hit the quality target with minimum effort*.



If you're trying to pass the class, then pass it with minimum effort. Anything else is wasted motion.

If you're trying to ace the class, then ace it with minimum effort. Anything else is wasted motion.

If you're trying to learn the material to the fullest, then mine the assignment for all its knowledge, and don't fret about your grade. Anything else is wasted motion.

If you're trying to do achieve some combination of good grades (for signalling purposes), respect (for social reasons), and knowledge (for various effects), then pinpoint the minimum quality target that gets a good grade, impresses the teacher, and allows you to learn the material, and hit that as efficiently as you can. Anything more is wasted motion.

Your quality target may be significantly left of F — if, say, you've already passed the class, and this assignment doesn't matter. Your quality target may be significantly to the right of A — if, say, you're there to learn the material, and grade inflation means that it's much easier to produce an A-grade paper than it is to complete the assignment in the maximally informative way. But no matter what, your goals will induce a quality target.

Both the slackers *and* the tryers are pursuing lost purposes. The slackers scoff at the tryers, who treat an artificial quality line like it's their actual preferences and waste effort over-achieving. The tryers scoff at the slackers, who are taking classes but refusing to learn. And both sides are right! Because both sides are wasting motion.

The slackers fail to deploy their full strength because they realize that the quality line is not their preference curve. The tryers deploy their full strength at the wrong target, in attempts to go as far right as possible, wasting energy on a fight that is not theirs. So take the third path: *remember what you're fighting for*. Always deploy your full strength, in order to hit your quality target as fast as possible.

Half-ass everything, with everything you've got.

(My teachers used to say that I could do great things if only I applied myself. I used to tell them that if they wanted me to apply more effort, they would need to invent higher letter grades.)

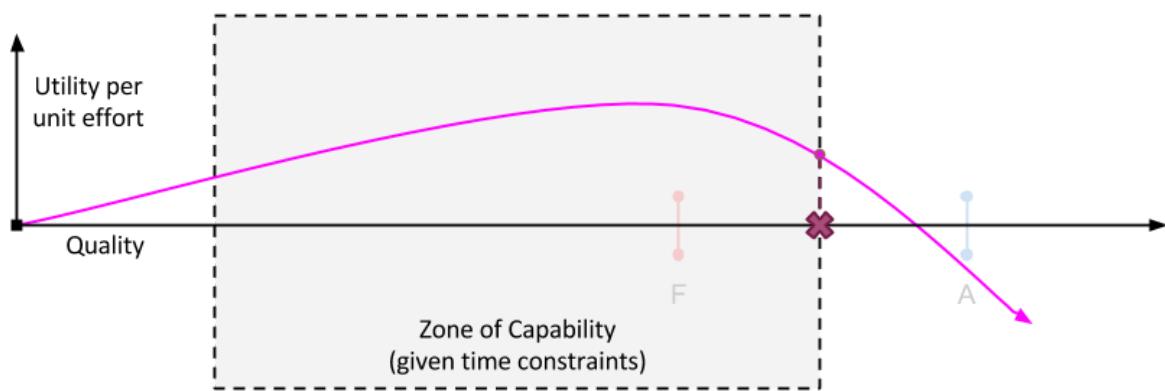
2

A common objection arises here:

Some things are too important to "half-ass." Some things are simply worth fighting for with your full strength. It's one thing to half-ass a homework assignment, and another thing entirely to half-ass saving a life. Sometimes you want to push as far right as you can on the quality curve.

This is both true and false, because it is mixed up. Given any project, *always* aim no higher than the quality target, and always strive for minimum expenditure of effort. It doesn't matter whether you're writing a term paper, pulling a person out of a burning house, or creating a galaxy-spanning human civilization — the goal is always to achieve some quality target with minimum effort. Negentropy is scarce.

That said, the quality target can be *really really high*. In fact, the quality target is sometimes unattainably high. Often, we simply aren't capable of hitting our quality targets, and in those cases, we *do* want to push as far right along the preference curve as we can.



This can occur naturally whenever you work on something difficult relative to your skill level, or in competitive situations, or if you're signalling your ability to work hard. But don't get confused. Even if you write for the love of writing, you eventually have to stop editing and call it finished. Even if you're getting somebody out of a burning building, you eventually stop putting effort towards ensuring that they survive in favor of putting that effort towards saving other dying people instead. Even if you're building an intergalactic civilization, you need to trade off energy spent building the civilization against energy spent living in it.

There are goals for which you cannot achieve your quality targets, and in those cases, you will push ever rightwards. But too many people automatically assume that, when an authority figure describes a quality line, they're "supposed to" push as far right as possible. They think they "should" care about quality. This is silly: real world problems are not about producing the highest-quality products. In all walks of life, the goal is to hit a quality target with minimum effort.

This is of course only a fuzzy and inaccurate description of reality. The relative costs of time, effort, energy, attention, and quality are generally in flux, and change with both information and circumstance. The essential point is to be able to differentiate between the implicit quality line highlighted circumstance, and your actual preference curve.

Let me be clear about what I'm *not* saying. If you're taking a college course, I'm *not* telling you that you should be scraping by by only the barest of margins. If you're saving a life, I'm *not* telling you to prefer speed over caution. In general, I'm decidedly **not** saying that you must always identify the worst outcome that you'd grudgingly accept as your target.

What I am saying is, don't conflate the quality line with the preference curve. Don't get confused when the teacher labels one quality-point "pass" and another "fail," for these are just labels, and your deeper goals are likely only tangentially related to those labels. Remember what you're trying to achieve, identify your quality target, and aim for that: no higher, no lower.

(Also, remember that the planning fallacy exists! If you shoot for a D, you might get an F. Humans tend to be overconfident. When you pick your targets, be cautious, and leave yourself comfortable margins.)

3

The common slacker objection goes:

But what if "get the minimum passing grade as quick as possible" is also boring? What if this task, too, is meaningless?

Then get out of college!

I personally find that shooting for the minimum acceptable quality is usually *fun*. Doing the homework assignment is boring, but finding a way to get the homework assignment up to an acceptable level *with as little total effort as possible* is an interesting optimization problem that actually engages my wits, an optimization problem which both my inner perfectionist and my inner rebel can get behind.

But sometimes, after remembering what you're fighting for, the whole project will still seem worthless. Sometimes, the goal of getting the minimum passing grade with minimum effort will still stink of somebody else trying to pass off their arbitrary metric as your true preferences. In that case, consider dropping the class.

More generally, if there's no variation on "achieve such-and-such a goal with minimum effort" that seems worth doing, then you may need to abandon that goal entirely.

By contrast, the common tryer objection runs as follows:

But I'm a perfectionist! I physically can't stop caring about a low-quality product. I'm compelled to do my best.

Great! Harness the perfectionist within you, and point it towards the goal of hitting your target with minimum effort.

Instead of being a perfectionist about the paper, be a perfectionist about *writing* the paper. Be a perfectionist about identifying good strategies, about abandoning sunk costs, about killing your darlings, about noticing when you're done. Be a perfectionist about wasting no attention. Be a perfectionist about learning from your mistakes. Perfectionism can be a powerful tool, but there's no need to point it at overachieving on metrics you don't care about.

4

Attempting to hit a quality target with the least possible effort is, in a sense, a much more difficult task than pushing as far right on the quality-line as possible. One always could push further right on the quality line with more time: when one is trying to write a great paper, they always *could* correct their flaws with more time and energy. But when one is trying to produce a paper with minimal wasted motion, mistakes are irrevocable. Time cannot be un-wasted.

In this sense, switching from being a tryer to a whole-assed-half-asser may lead to more guilt and shame than usual, if

you start feeling guilty about every wasted motion.

However, I see far too many people feeling guilt and shame about not having pushed far enough along the quality line. They feel guilty about not putting effort towards their job (which they hate); they feel guilty about not being a good enough friend (when they are nearly at breakdown themselves); they feel guilty about not fulfilling their parent's expectations (which are ridiculous and uninformed). In order to replace guilt and shame with intrinsic motivation, it is first necessary to break the slacker/tryer dichotomy. If you've got to feel guilty, please feel guilty about missing your own targets, rather than feeling guilty about not adopting some arbitrary quality line as your true preferences. The former type of guilt is the one that I have a shot at addressing.

(Scrupulous people: in the interim, please don't feel guilty about wasting motion! Treat it like an important part of the human action process rather than something to be ashamed of. Future posts will expand on this idea.)

5

Most people seem to have two modes of working on problems: the slacker path-of-least-resistance "coasting" mode, and the tryer make-a-masterpiece "overachiever" mode. When faced with a problem, most people either put in the minimum effort necessary to scrape by without pissing off

the relevant authorities, or else they pour their heart and soul into the task.

Almost everybody spends some time in both modes. Some people overachieve in history class and coast in grammar class. Some people overachieve at work and coast in their relationship. In fact, most heartwarming bad-students-can-be-good-people-too stories are about how students who are slacking in most domains are secretly trying really hard when it comes to dance/sports/music/number theory.

This, of course, is another piece of tryer propaganda: "Don't worry," the movies say, "these slackers aren't bad people, because they're secretly tryers in other domains!" As if you're only a good person if you can adopt *some* arbitrary quality line as your true preferences.

Most people are trapped in the slacker/tryer dichotomy. They either do as little as they can get away with or as much as they can manage. They're either aiming for barely acceptable or they're aiming to be the best. Very few people seem able to pick a target in the middle and then pursue it with *everything they've got*. Very few people seem capable of deploying their *full strength* to hit "mediocre" as efficiently as possible.

Reject the dichotomy. Keep your eye on the preference curve. And remember that the preference curve says this, and only this:

Succeed, with no wasted motion.

The slacker in you rebels against pointless tasks, and the tryer in you wants perfection. So satisfy both: aim for the minimum necessary target, and move there as efficiently as possible.

And if ever you forget what it means to "succeed" in one context or another, take a moment to pause and remember what you're fighting for.

Failing with Abandon

12 APRIL 2015

This is a short public service announcement: *you don't have to fail with abandon.*

Say you're playing Civilization, and your target is to get to sleep before midnight, and you check the clock, and it's already 12:15. If that happens, you *don't* have to say "too late now, I already missed my target" and then keep playing until 4 in the morning.

Say you're trying to eat no more than 2000 calories per day, and then you eat 2300 by the end of dinner, you *don't* have to say "well I already missed my target, so I might as well indulge."

If your goal was to watch only one episode of that one TV show, and you've already watched three, you *don't* have to binge-watch the whole thing.

Over and over, I see people set themselves a target, miss it by a little, and then throw all restraint to the wind. "Well," they seem to think, "willpower has failed me; I might as well over-indulge." I call this pattern "failing with abandon."

But you *don't* have to fail with abandon. When you miss your targets, you're allowed to say "dang!" and then *continue trying to get as close to your target as you can*.

You don't have to say dang, either. You're allowed to over-indulge, if that's what you want to do. But for lots and lots of people, the idea of missing by *as little as possible* never seems to cross their mind. They miss their targets, and then suddenly they treat their targets as if they were external mandates set by some unjust authority; the jump on the opportunity to defy whatever autarch set an impossible target in the first place; and then (having already missed their target) they reliably fail with abandon.

So this is a public service announcement: you don't have to do that. When you miss your target, you can take a moment to remember who put the target there, and you can ask yourself whether you want to get as close to the target as possible. If you decide you only want to miss your target by a little bit, you still can.

You don't have to fail with abandon.

Replacing guilt

20 APRIL 2015

In my experience, many people are motivated primarily by either guilt, shame, or some combination of the two. Some are people who binge-watch television, feel deeply guilty about it, and convert that guilt into a burning need to *Actually Do Something* on the following day. Others are people who feel guilty whenever they stop working before they *literally fall over from exhaustion*, and in attempts to avoid that guilty feeling, they consistently work themselves weary.

I find that using guilt as a motivation source is both unhealthy and inefficient, but yet, I find it to be a common practice, especially among effective altruists.

Thus, in the coming series of posts, I'm going to explore a whole slew of tools for removing guilt-based motivation and replacing it with something that is both healthier and stronger.

My goal is to help people remove guilt-based motivation entirely, and replace it with intrinsic motivation. I'm aiming to both reduce the frequency of netflix binges *and* reduce the bad feelings that follow. I'm aiming to help people feel like they're still worthwhile human beings if they stop working before they literally drop. I'd like to help people avoid the failure mode where they feel guilty about something for days (even after learning their lesson), and I'm also hoping

to remove some shame-based motivation while I'm in the area.

My first goal will be to address the guilt that comes from a feeling of *listlessness*, the vague feeling of guilt that one might get when they play video games all day, or when they turn desperately towards drugs or parties, in attempts to silence the part of themselves that whispers that there must be something else to life.

This sort of guilt cannot be removed by force of will, in most people. The trick to removing this sort of guilt, I think, is to start exploring that feeling that there must be something else to life, that there must be something more to do — and either find something worth working towards, or find that there really isn't actually anything missing. This first sort of listless guilt, I think, comes from someone who wants to find something else to do, and hasn't yet.

Unfortunately, addressing this sort of guilt isn't as easy as just finding a hobby. In my experience, this listless guilt tends to be found in people who have fallen into the nihilistic trap — people who either believe they can't matter, or who believe that no one can matter. It tends to be found in people who believe that humans only ever do what they want, that nothing is truly "better" than anything else, that there is no such thing as altruism, that "morality" is a pleasant lie — that class of beliefs is the class that I will address first, starting with the Allegory of the Stamp Collector.

I'll post the allegory tomorrow. In the interim, I invite you to devise your own tools for removing the listless guilt: the tools that people develop themselves are often more useful to them than the tools they are given.

The Stamp Collector

26 APRIL 2015

Once upon a time, a group of naïve philosophers found a robot that collected trinkets. Well, more specifically, the robot seemed to collect stamps: if you presented this robot with a choice between various trinkets, it would always choose the option that led towards it having as many stamps as possible in its inventory. It ignored dice, bottle caps, aluminum cans, sticks, twigs, and so on, except insofar as it predicted they could be traded for stamps in the next turn or two. So, of course, the philosophers started calling it the "stamp collector."

Then, one day, the philosophers discovered computers, and deduced out that the robot was merely a software program running on a processor inside the robot's head. The program was too complicated for them to understand, but they did manage to deduce that the robot only had a few sensors (on its eyes and inside its inventory) that it was using to model the world.

One of the philosophers grew confused, and said, "Hey wait a sec, this thing can't be a stamp collector after all. If the robot is only building a model of the world in its head, then it can't be optimizing for its real inventory, because it has no access to its real inventory. It can only ever act according to a model of the world that it reconstructs inside its head!"

"Ah, yes, I see," another philosopher answered. "We did it a disservice by naming it a stamp collector. The robot does not have true access to the world, obviously, as it is only seeing the world through sensors and building a model in its head. Therefore, it must not *actually* be maximizing the number of stamps in its inventory. That would be impossible, because its inventory is outside of its head. Rather, it must be maximizing its *internal stamp counter* inside its head."

So the naïve philosophers nodded, pleased with this, and then they stopped wondering how the stamp collector worked.

There are a number of flaws in this reasoning. First of all, these naïve philosophers have made the homunculus error. The robot's program may not have "true access" to how many stamps were in its inventory (whatever that means), but it *also* didn't have "true access" to its internal stamp counter.

The robot is not occupied by some homunculus that has dominion over the innards but not the outards! The abstract program doesn't have "true" access to the register holding the stamp counter and "fake" access to the inventory. Steering reality towards regions where the inventory has lots of stamps in it is the *same sort of thing* as steering reality towards regions where the stamp-counter-register has high-

number-patterns in it. There's not a magic circle containing the memory but not the inventory, within which the robot's homunculus has dominion; the robot program has just as little access to the "true hardware" as it has to the "true stamps."

This brings us to the second flaw in their reasoning reasoning, that of trying to explain choice with a choice-thing. You can't explain why a wall is red by saying "because it's made of tiny red atoms;" this is not an *explanation* of red-ness. In order to explain red-ness, you must explain it in terms of non-red things. And yet, humans have a bad habit of explaining confusing things in terms of themselves. Why does living flesh respond to mental commands, while dead flesh doesn't? Why, because the living flesh contains Élan Vital. Our naïve philosophers have made the same mistake: they said, "How can it possibly choose outcomes in which the inventory has more stamps? Aha! It must be by choosing outcomes in which the stamp counter is higher!" and in doing so, they have explained choice in terms of choice, rather than in terms of something more basic.

It is *not an explanation* to say "it's trying to get stamps into its inventory because it's trying to maximize its stamp-counter." An explanation would look more like this: the robot's computer runs a program which uses sense-data to build a model of the world. That model of the world contains a representation of how many stamps are in the inventory. The program then iterates over some set of available actions, predicts how many stamps would be in the invento-

ry (according to the model) if it took that action, and outputs the action which leads to the most predicted stamps in its possession.

We could *also* postulate that the robot contains a program which models the world, predicts how the world would change for each action, and *then* predicts how *that* outcome would affect some specific place in internal memory, and *then* selects the action which maximizes the internal counter. That's possible! You could build a machine like that! It's a strictly more complicated hypothesis, and so it gets a complexity penalty, but at least it's an explanation!

And, fortunately for us, it's a *testable* explanation: we can check what the robot does, when faced with the opportunity to directly increase the stamp-counter-register (without actually increasing how many stamps it has). Let's see how that goes over among our naïve philosophers...

Hey, check it out: I identified the stamp counter inside the robot's memory. I can't read it, but I did find a way to increase its value. So I gave the robot the following options: take one stamp, or take zero stamps and I'll increase the stamp counter by ten. Guess which one it took?

"Well, of course, it would choose the latter!" one of the naïve philosophers answers immediately.

Nope! It took the former.

"... Huh! That means that the stampyness of *refusing* to have the stamp counter tampered with must worth be more than 10 stamps!"

Huh? What is "stampyness"?

"Why, stampyness is the robot's internal measure of how much *taking a certain action* would increase its stamp counter."

What? That's ridiculous. I'm pretty sure it's just collecting stamps.

"Impossible! The program doesn't have access to how many stamps it really has; that's a property of the outer world. The robot *must* be optimizing according to values that are actually in its head."

Here, let's try offering it the following options: either I'll give it one stamp, or I'll increase its stamp counter by Ackermann(g_{64}, g_{64}) — oh look, it took the stamp."

"Wow! That was a very big number, so that almost surely mean that the stampyness of refusing is dependent upon how much stampyness it's refusing! It must be very happy, because you just gave it a *lot* of stampyness by giving it such a compelling offer to refuse."

Oh, here, look, I just figured out a way to set the stamp counter to maximum. Here, I'll try offering it a choice between either (a) one stamp, or (b) I'll set the stamp counter to maxi — oh look, it already took the stamp.

"Incredible! That must there must be some other counter measuring *micro-stampyness*, the amount of stampiness it gets *immediately* upon selecting an action, before you have a chance to modify it! Ah, yes, that's the only possible explanation for why it would refuse you setting the stamp counter to maximum, it *must* be choosing according to the perceived immediate micro-stampyness of each available action! Nice job doing science, my dear fellow, we have learned a lot today!"

Ahh! No! Let's be very clear about this: the robot is predicting which *outcomes* would follow from which actions, and it's ranking them, and it's taking the actions that lead to the best outcomes. Actions are rated according to what they achieve. Actions do not *themselves* have intrinsic worth!

Do you see where these naïve philosophers went confused? They have postulated an agent which treats *actions* like *ends*, and tries to steer towards whatever *action* it most prefers — as if actions were ends unto themselves.

You can't explain why the agent takes an action by saying that it ranks actions according to whether or not taking them is good. That begs the question of which actions are good!

This agent rates actions as "good" if they lead to outcomes where the agent has lots of stamps in its inventory. Actions are rated according to what they achieve; they do not themselves have intrinsic worth.

The robot program doesn't contain reality, but it doesn't need to. It still gets to *affect* reality. If its model of the world is correlated with the world, and it takes actions that it predicts leads to more *actual* stamps, then it will tend to accumulate stamps.

It's *not* trying to steer the future towards places where it happens to have selected the most micro-stampy actions; it's just steering the future towards worlds where it predicts it will actually have more stamps.

Now, let me tell you my second story:

Once upon a time, a group of naïve philosophers encountered a group of human beings. The humans seemed to keep selecting the actions that gave them pleasure. Sometimes they ate good food, sometimes they had sex, sometimes they

made money to spend on pleasurable things later, but always (for the first few weeks) they took actions that led to pleasure.

But then one day, one of the humans gave lots of money to a charity.

"How can this be?" the philosophers asked, "Humans are pleasure-maximizers!" They thought for a few minutes, and then said, "Ah, it must be that their pleasure from giving the money to charity outweighed the pleasure they would have gotten from spending the money."

Then a mother jumped in front of a car to save her child.

The naïve philosophers were stunned, until suddenly one of their number said "I get it! The immediate micro-pleasure of *choosing that action* must have outweighed —

People will tell you that humans always and only ever do what brings them pleasure. People will tell you that there is no such thing as altruism, that people only ever do what they want to.

People will tell you that, because we're trapped inside our heads, we only ever get to care about things inside our heads, such as our own wants and desires.

But I have a message for you: You can, in fact, care about the outer world.

And you can steer it, too. If you want to.

You're allowed to fight for something

03 MAY 2015

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

10 MAY 2015

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 *reflec-*

tively 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

17 MAY 2015

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 *inscrutable*. 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

by Matt Rhodes







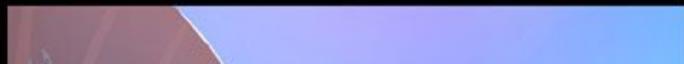


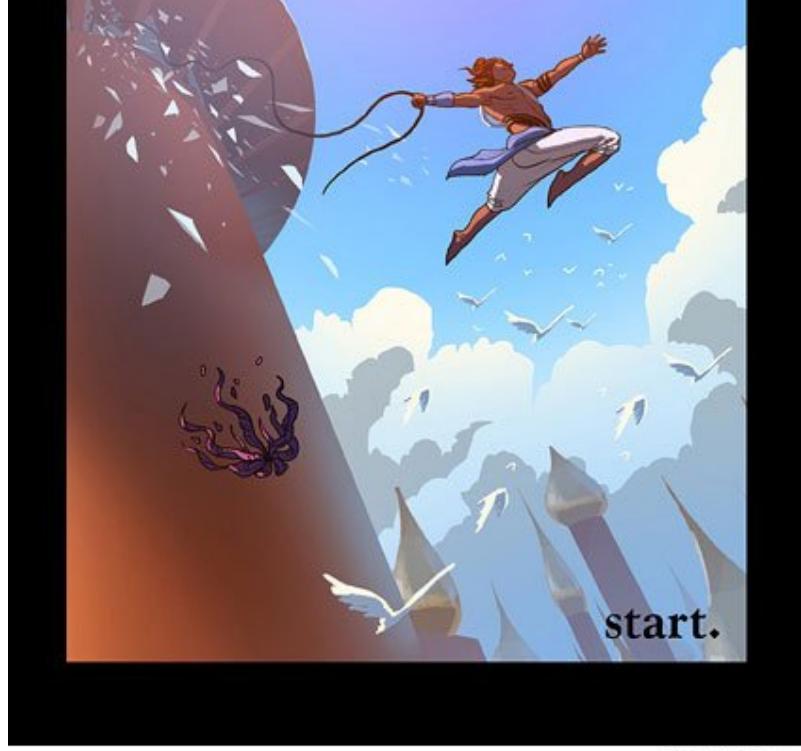












Matt Rhodes

([source](#))

"Should" considered harmful

25 MAY 2015

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 at-

tached 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, because 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.

Not because you "should"

31 MAY 2015

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*, with-

out 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 precommitting 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.

Your "shoulds" are not a duty

08 JUNE 2015

I have a friend who, after reading my last two posts, still struggled to give up her shoulds. She protested that, if she stopped doing things because she should, then she might do the wrong thing. I see this frequently, even among people who claim to be moral relativists: they protest that if they weigh their wants and their shoulds on the same scales, then they might make the wrong choice.

But this notion of "right" vs "wrong" cannot come from outside. There is no stone tablet among the stars that mandates what is right. Moral relativists usually have no trouble remembering that their narrow, short-term desires (for comfort, pleasure, etc.) are internal, but many seem to forget that their wide, long-term desires (flourishing, less suffering, etc.) are also part of them.

Why did my friend worry that, if she stopped forcing herself to follow her shoulds, that she might do the wrong thing? There are no outside authorities punishing people who don't follow their best interests. There are no heavenly gatekeepers rewarding you for doing something other than what's best. The only reason to care about doing what's right is because you want what's right to be done. Why was she afraid that she would fail to follow her own interests, if she stopped using internal force? It's *you* who wants the right

thing done, so if you fear you're not going to do the right things, then bargain with yourself.

Part of the problem, I think, is that she realized that she wants to both (a) do the right thing and (b) avoid all the effort that entails, and she feared that without the tool of internal force, she would be unable to do as much good as she wants to do. This is a valid concern, and following posts will discuss different tools for doing what you want (without resorting to internal force, which I think is unsustainable — remember, expending willpower is a stopgap, not a solution).

But another part of the problem, I think, was a lingering sense of resentment towards the shoulds, for trying to suck fun and enjoyment out of her life.

I see this often. Picture someone who needs to choose between playing video games all day (and losing their job) or getting abused by customers all day (for not all that much money). They conclude that they "should" do the job, and they feel compelled by the should. And then, over and over, I find my friends resenting their shoulds, as if the shoulds came from outside, from Beyond, from the Intergalactic Oughthorities. They treat their should like shackles that bind them to the "right path", the one where have to go to work when they could be playing video games.

But the shoulds aren't the shackles. There aren't any oughthorities. You always get to do as you please, within the

bounds allowed by the universe. It's the *situation* which forces you to choose between bad and worse. Don't resent the bad option for being better than the worse option — if you must resent something, resent the situation.

(Or, better yet, turn your resentment into a cold resolve to *change* the situation.)

If you ever start to feel that your shoulds are obligations, then remember this:

The shoulds were made for us, not us for them.

There are no facts about the stars that say what you ought to do. Your shoulds are not written in the heavens, nor in the void.

But your shoulds are written in *you*.

What you "should do" in any given situation is a fact about your brain and the situation (which takes into account your current state of knowledge, and the amount of time you have available, and so on).

In other words, someone with a ton of computing power and intimate knowledge of your brain could *tell* you what you should do in any given situation.

Imagine being told some of those facts about what you "should do, as computed by someone with ridiculous amounts of computing power. They print them out on a sheet of paper, and hand it to you. What would this sheet of paper look like?

I think that most people expect it would look like a long list of obligations, full of uncomfortable task they're actively trying to not remember. Most people seem to expect a highly aversive list that reads something like this:

- Clean your room.
- Send a message to that one friend you fell out of touch with who sent you a message on your birthday.
- Reconcile with your father.
- Donate more to charity.
- ...

This is exactly the notion of "should" that I'm trying to discharge.

Your true shoulds, if I could show them to you, would not look like a list of obligations. Your true shoulds would look like a *recipe for building a utopia*.

They would look like a series of steps that make the world the best place you can make it.

And they wouldn't tell you to do anything psychologically unrealistic, either. Just as the list wouldn't say "snap your

fingers in just such a way that alzheimers is cured," the list wouldn't say "work yourself to the bone for 16 hours a day while still remaining in high spirits." No, the true shoulds (as computed by someone with deep knowledge of your brain and ridiculous amounts of computing power) would appear to you as a psychologically possible list of things that happened to have surprisingly awesome impacts on the world.

The things that you feel resentment towards are false shoulds, or at least twisted shoulds. Encountering one of your *actual* moral bonds feels very different indeed. A true opportunity to execute a moral commitment feels not like an obligation, but like a *privilege*. It feels like executing a Screw The Rules I'm Doing What's Right trope.

In fiction, picture the moment when the villain reveals that doing the Right Thing will start a war, and the hero sets their jaw, looks them in the eyes, and says "so be it," and then does the right thing anyway.

In real life, think of Irena Sendler, who smuggled thousands of Jewish children to safety during the holocaust, who was captured by the Nazis and tortured and had her legs broken and was sentenced to death, and who escaped anyway, and then *went back*.

Imagine what was going through her mind, when she decided to go back and save more people. Now, of course, I have no idea what she was actually feeling, but when I imagine what it would take for *me* to go back under those circumstances, I imagine feeling fear, and a hint of despair at finding myself still capable, but also a burning resolve to do the right thing anyway.

I imagine her feeling that having the opportunity to go back was a *privilege*. Not an external obligation whispered down from the heavens, but an internal fire, a defiance of the natural order, a need to make the world *different* from the way it would be otherwise.

Irena didn't have an *obligation* to keep fighting. She had more than discharged her moral duty. And while I'm willing to bet that at least part of her was scared, and at least part of her wished she had been crippled and unable to return, there was also a part of her that didn't look at the opportunity to return to save more children as a misfortune, but as an honor.

Can you begin to see the difference between a false should, and a true moral commitment? Think of a false should, one that gives you a strong sense of obligation and a hint of resentment (such as "finish this paper" or "go to work tomorrow"). Now imagine of Irena Sendler, offered the opportunity to return to Warsaw. Imagine what went on in her head, in that moment.

I imagine a mind afraid, but unified, because for her, it wasn't really a choice. Innocent children were still dying, and there was only one thing to do.

That's what a true moral impulse feels like, when you find one. Not like an obligation, but like a piece of cold iron found deep in your core, the thing that you touch — or that touches you — in the moment that you really see the best option available to you, the moment that you realize you already know which way this choice is going to go.

Your shoulds are not shackles, and I caution you to be wary of anyone who tries to force a should upon you. For if you are not careful, you may start to feel like your shoulds are obligations, and you may start to resent them.

Human moral bonds aren't compulsions. They are what let Irena Sendler see the opportunity to risk life and limb to save just one more child, and treat it not as a duty, but as an honor. If you told *her* that she didn't have to go back, that she'd done enough, that she'd earned the right to turn away, and you asked her why the hell she was still going back to Warsaw,

then she's allowed to reach inside, touch that something of iron, look you in the eyes, and say "because I should."

That's how you use a should. Not with obligation and resentment, but with steel in your heart and no other choice that compares.

I strongly encourage you to unpack your shoulds into their component wants and desires — I would rather not be responsible for inspiring a bunch of people to run around shoulding themselves and saying "no, it's OK, these are the true moral bonds." Rather, the point I'm trying to make is this:

Many treat their moral impulses as a burden. But I say, find all the parts that feel like a burden, and drop them. Keep only the things that fill you with resolve, the things you would risk life and limb to defend.

Those moral impulses are not a reminder of your grudging duty. They are a reminder that you value things larger than yourself. They are a description of everything you're fighting for. They are the birthright of humanity, they are your love for fellow sentient creatures, they are everything we struggle so hard to send upwards to the stars.

They aren't a duty. They're an honor.

Working yourself ragged is not a virtue

21 JUNE 2015

Let's get back to the "replacing guilt" series. Here's a quick recap of what we've covered so far:

Part 1 was about replacing the listless guilt: if someone feels vaguely guilty for not really doing anything with their life, then the best advice I can give is to start doing something. Find something to fight for. Find a way that the world is not right, and decide to change it. Once the guilt is about failing at a *specific* task, then we can start addressing it.

Part 2 was about refusing to treat your moral impulses as obligations. Be wary of the word should, which tries to force an obligation upon you. I recommend refusing to do anything just because you "should": Insofar as that sets you free, the obligations were false ones. Insofar as that sparks fear that something important won't get completed, seek out the cause of the worry, and complete the task because you want to see it done, rather than because you "should."

However, having something to change in the world and being free of false obligations is not anywhere near enough to replace guilt motivation. In fact, I think that most guilt in most people comes from a different source: it comes from people honestly deciding that X is what they want to do and then finding themselves *not doing X anyway*.

Maybe they *know* that watching another episode of a TV show will cause them to stay up too late and be tired at class tomorrow, and they *know* that their classes are very expensive and that their parents would be very disappointed, and they decide that the best thing to do would be to stop binge-watching the TV show and get some sleep — and then they find themselves watching the next episode anyway.

This sort of guilt is one of the most demoralizing, and therefore it's perhaps one of the most damaging types of guilt. Addressing it is going to require quite a few different tools. Today, I'll describe one of them.

(If you haven't read [half-assing it with everything you've got](#), recommend doing so now: I wrote it as a direct predecessor to this idea, before realizing that I actually needed the previous seven or so posts first.)

Here's a failure mode that I used to see all the time, back when I was a professional programmer: A co-worker of mine would be working on a project that was *almost* under control. It would be a Friday afternoon, with an important deadline coming up in a few weeks, and everything would be almost passable but slightly behind schedule. Some dire bugs demanded fixes, some poor decisions required refactoring. Inevitably, my co-worker would conclude that if they just worked really hard *this* Friday, then they could finish

the big refactor, and once that was done, *next* week they could get all the bugs under control, and then by the beginning of the week after that, everything would be back on track again.

(We all know how this story goes.)

Inevitably, co-workers of this type were constantly stressed, and reliably worked late into the night.

I suspect that most people who act like this are guilt-motivated. They're often the sort of person who feels guilty if they stop working before they're completely exhausted. Sometimes, they feel guilty for stopping even when they *are* exhausted, if there is still more work to be done. It's as if part of them believes that if they stop before they're physically forced to drop, and there's still work to be done, then they're being Bad.

This sort of behavior can stem from a number of mistakes. First and foremost, it seems to me that this sort of programmer is usually pursuing a lost purpose. They have succumbed to tryer propaganda; they have confused the quality line for the preference curve. I sometimes want to grab them by the shoulders late on a weekend, look them in the eyes, and ask them what they're fighting for — surely not this? You're allowed to fight for something!

But I also see this failure mode in people who love their work, who believe in its importance. And yet, they still work

themselves to exhaustion in a binge/recovery cycle, as if this were the best way to cause their project to succeed.

These people seem to be following an impulse to work as hard as possible whenever they can, perhaps due to a belief that it is unvirtuous for one to stop working when they could continue.

This is an error. **The goal is not to maximize how much work you get done today. The goal is to maximize your productivity over time.**

People who feel guilty for stopping work when they could continue seem to be trying to maximize their local velocity: they feel a need to produce as much as they can, right now, on pain of guilt if they fail. But the actual goal is to maximize the total distance traveled, to maximize how much important work you can get done over time.

(When all is said and done, and Nature passes her final judgement, you will not be measured by the number of moments in which you worked as hard as you could. You will be measured by what actually happened, as will we all.)

People driven by guilt and shame often feel bad for *slowing down*. This is about as effective as starting a marathon with a dead sprint, and then feeling bad for slowing down when you can't sustain it.

Working yourself ragged is not a virtue. You don't get extra points for effort. In fact, you *lose* points for effort: effort is costly; spend it only to purchase better outcomes. The goal is not to appear to be working hard, the goal is to improve the world. Sometimes you do need to push yourself to the limit, but before you do, acknowledge the costs and weigh the tradeoffs, while keeping your long-term goals in view.

We're not yet gods. We're still apes. Remember to pay attention to the distance you need to cover, and remember to pay attention to yourself.

Being a human can be frustrating. Human-bodies aren't as productive as we might like them to be, and running a human-body at maximum capacity for too long causes stress, chronic exhaustion, burnout, and psychological damage. With this in mind, doesn't it seem a bit confused for a person to berate themselves for stopping before they've spent all their available reserves?

Let me be clear: I'm *not* saying to restrict yourself to only 40 hour per week of work because it's important to pace yourself. I'm just saying that it's important to pace yourself. Do as much as you can, but don't be constantly taking damage. We aren't yet gods. We're still fragile. If you have something ur-

gent to do, then work as hard as you can — but work as hard as you can *over a long period of time*, not *in the moment*.

I'm also not saying "stop as soon as working feels hard." When exercising, it's important to understand the difference between soreness and strain, between pushing yourself and hurting yourself, and the same is true psychologically. Maintaining focus and productivity for long periods of time is a skill that can be trained, like any other. (More on that later, but spoiler alert, "feeling really guilty when you didn't work as hard as you wanted to" is not the best way to train this skill.)

Push your limits! Some things are worth fighting for! But while you're doing that, recognize that the way to complete a marathon isn't to sprint 42+ kilometers.

There is no shame in doing less than you could do in any given moment. Most guilt-motivated people I meet would do well to worry less about whether they're going fast enough *now*, and worry more about whether the amount of work they're doing day-to-day is ideal in the long term, taking psychological constraints into account. You don't get points for pushing your body and mind as hard as you can, you get *good outcomes* from using your resources as *wisely* as you can. That usually entails stopping well before you drop each day, while steadily improving your capabilities.

Please treat yourself well today; doing so is an important component of long-term productivity.

Rest in motion

28 JUNE 2015

Many people seem to think the 'good' state of being, the 'ground' state, is a relaxed state, a state with lots of rest and very little action. Because they think the ground state is the relaxed state, they act like maintaining any other state requires effort, requires suffering.

This is a failure mode that I used to fall into pretty regularly. I would model my work as a finite stream of tasks that needed doing. I'd think "once I've done the laundry and bought new shoes and finished the grocery shopping and fixed the bugs in my code and finished the big refactor, everything will be in order, and I'll be able to rest." And in that state of mind, every new email that hit my inbox, every new bug discovered in my code, every tool of mine that wore down and needed repair, would deal me damage.

I was modeling my work as finite, with the rest state being the state where all tasks were completed, and so every new task would push me further from that precious rest state and wear me down.

But the work that needs to be done is not a finite list of tasks, it is a neverending stream. Clothes are always getting worn down, food is always getting eaten, code is always in motion. The goal is not to finish all the work before you; for that is impossible. The goal is simply to move *through* the

work. Instead of struggling to reach the end of the stream, simply focus on moving along it.

Advertisements and media often push the narrative that the purpose of all our toil is to win a chance at relaxation. We're supposed to work hard at boring jobs in order to earn our vacations. We're supposed to work hard for decades so that we can retire. (We're supposed to conceive of heaven as a place where nobody does anything except lounge on clouds.)

I call bullshit. For almost everybody, inaction is *boring*. That's why we pick up books, go exploring, and take up hobbies. The ground state is an active state, not a passive one.

The *actual* reward state is not one where you're lazing around doing nothing. It's one where you're keeping busy, where you're doing things that stimulate you, and where you're resting only a fraction of the time. The preferred ground state is not one where you have no activity to partake in, it's one where you're managing the streams of activity precisely, and moving through them at the right pace: not too fast, but also not too slow. For that would be boring.

And yet, most people have this model of the world where whenever they're not resting, they're taking damage. When the homework isn't done, they're taking damage. When

they're reading a textbook, they're taking damage. When they go to sleep with work unfinished, they're taking damage. When they're at a large social event, they're taking damage. Some part of them yearns to be in the rest state, where they don't need to do all these *things*, and insofar as they aren't, they're suffering a little.

This is a grave error, in a world where the work is never finished, where the tasks are neverending.

Rest is not a reward for getting through all your obligations. You already dropped your obligations, remember? Rather, rest (and personal health, and personal time) are part of the goal. Both because most people care about their personal comfort, and because taking care of yourself is very important in order to do all the other things you want to do.

Rest isn't something you do when everything else is finished. Everything else doesn't *get* finished. Rather, there are lots of activities that you do, some which are more fun than others, and rest is an important one to do in appropriate proportions. Disconnect your impulse to rest from whether or not the world is in a stable state, because, spoiler alert, the world isn't going to be in a stable state for a long time.

Rest isn't a reward for good behavior! It's not something you get to do when all the work is finished! That's finite task thinking. Rather, rest and health are just two of the unending streams that you move through.

Imagine the person who is tight on money and needs to buy groceries once a month. Imagine that they agonize over every purchase, even though they know that they're buying as little as they can in order to secure the health of their family. You might suggest to them that they stop fretting over individual purchases and come to terms once and for all with the fact that food is a necessary purchase, and suggest that they fret over their *budget* instead. That way, they won't need to suffer every time they enter a grocery store.

The same technique applies to effort. You don't need to suffer every time it's time to do the laundry. Stop looking at the individual tasks, and start looking at the streams of work, some of which you can widen and some of which you can narrow.

Look at *all* the streams you want to move through, assess how much bandwidth you have available, and then simply move through the streams at the appropriate clip. Some streams will be unpleasant (chores, etc.), some will be basically mandatory (making money, etc.), some will be quite fun (learning, exploring, relaxing, etc.), and some of the most important streams are the meta streams (improving your capacity, finding better ways to fulfill your needs, etc.). But in all cases, simply see the streams and then move along them.

Many people I meet seem to think that they need to take damage whenever they're working, and then only heal it when they rest. While they're studying, they're taking damage. While they're at a large social event, they're taking damage. While they're doing their job, they're taking damage. They seem to think they "should" be able to be at home doing nothing, and so when they're not, they're taking damage. They think that the ground state is a resting state, a state of inaction, and so whenever they're acting, this is a deviation from the default, and it requires effort to maintain.

I say, *the ground state is in motion*. The privileged state is not a frozen state. Most of us wouldn't want to just lie in bed doing nothing forever, anyway. The easiest state to maintain isn't a motionless state, it's the state where you're out there *doing what needs doing* at a sustainable pace. *That's* the ground state, that's the state that requires no effort to maintain. Anything less leads to boredom, and it's *boredom* that's taxing.

I think one of the reasons people think high productivity is hard is that they think of lying in bed doing nothing as the default state, and anything else as taking damage. But it's not. It's really not. We were built to *move*, and we have things to do.

Make sure you're not taking damage just for moving. If any state of being is going to wear you down, then I suggest that you feel pressure whenever you start to move too fast or too

slow. Take damage when your life is too boring and nothing's getting done, and take damage when your life is moving at an unsustainable pace: but don't take damage when you're moving through the streams at a steady clip.

The default state, the effortless state, is the one where you're moving along many streams. It is up to you to make sure that you're prioritizing the right streams and that you're steadily increasing your throughput, but the end goal is not to cease moving. Total inaction is dreadfully boring.

The ground state, the state to aspire to, the healthy state, the state that occurs naturally when you aren't forcing yourself to do anything, is the state where you're getting done what you want done as fast as is sustainable, and no faster.

The ground state is in motion.

Shifting guilt

05 JULY 2015

The posts so far have been less about confronting guilt, and more about different tools for shifting it. This is a valuable skill to generalize.

The posts in this series have developed three such tools for shifting guilt. In this post, I'll recast those three tools as members of the same family, so that you can start to see the pattern, and develop similar tools from the same family as you need them.

The tools that I have described so far shift guilt to one particular place: guilt about being unable to act as you desire. This is intentional — that is the one place that I know how to confront guilt head-on.

The first tool for shifting guilt is the tool of *refinement*. This tool is used on listless guilty in need of pointing.

Imagine finding yourself feeling vaguely guilty the morning after a party, having slept in longer than you intended, your head aching from a slight hangover. Imagine a vague guilt making your body feel heavier. Perhaps it whispers that the night was senseless. Perhaps it murmurs that you're wasting your life away. This is the sort of guilt that's amenable to refinement: ask the guilt what, precisely, it would have had you do instead of what you did. (It is important, when refining, to also possess the virtue of concreteness: do not settle

for "I should have been studying." Demand a specific action: Which book? Which chapter?)

Sometimes, when asking the guilt what you could have done instead, you will remember that none of the alternatives were compelling. Maybe the party was for an old friend who you only see once every few years, and fulfilling the social obligation was better than the alternative. Maybe you were exhausted from a day of studying, low on human contact, and needed the party to reinvigorate you. When using the tool of refinement, the guilt sometimes simply disappears.

But often, the guilt gets more pointed. Perhaps you conclude you should have been working overtime so you could donate the money to a worthy cause. Perhaps you conclude that you had the opportunity and the stamina, but simply not the willpower. This is good! This is a success! The refinement has succeeded, and the guilt has come into more focus.

But more often than not, when you succeed at refining guilt, you find yourself left with an obligation ("I should have drank less" or "I should have studied" or "I should have worked overtime.") This has not yet shifted enough to be confronted. For obligations, you need the second tool.

The second tool for shifting guilt is the tool of *internalization*. This tool is used on guilts that stem from neglected obligations.

I strongly recommend that you staunchly refuse to bow to any guilt forced on you from the outside. You say you "should have" studied more, instead of going to the party? Says who? Cash out the should. Again, it is critically important have virtue of concreteness when cashing out a should: do not say "it would have been better for me to study more;" for this has not removed the should, it has simply hidden it inside the word "better." The way to cash out a should (and, thus, the way to use the tool of internalization) is to ask yourself whether or not it would be OK to drop the obligation entirely.

What would happen if you decide to never study that textbook again? Is it a relief? If so, then drop the obligation, and relinquish the guilt. You probably just accidentally confused someone's quality line with your preference curve. Sometimes, when attempting internalization, the guilt simply disappears. (Other times, part of the guilt disappears, and you find yourself again facing a vague, unfocused guilt. This is fine, and indeed quite normal — just apply the tool of refinement again, and repeat.)

But more often than not, when you threaten to drop an obligation entirely, some part of you protests. Imagine you're feeling terrible for failing to work overtime and donate money. If you ask yourself "what if I just never donate money to those worse off ever again?," then most likely, some part of you will protest "but that would be bad!"

This is good. It means you have your *own* reasons for wanting to donate, which means you can drop the external obligation and do it because *you* want to.

Why would it be bad to stop donating? Don't settle for answers like "because then I would be a bad person" — that's replacing one obligation with another. If you get an answer like that, ask yourself, "why would it be bad for me to be a bad person?" Remember that concreteness is a virtue. Don't settle for an externalized answer (such as "because then people wouldn't like me"); push on until you get an internalized answer ("because I prefer worlds where _____").

Keep in mind that there may be many different parts to the answer: if you use the tool of internalization and get an answer that feels unconvincing, such as "because I prefer worlds where my friends think that I am generous," then ask yourself something like "OK, let's say that my friends were guaranteed to think that I am generous regardless of how much I donate to people worse off than me, *then* is it OK for me to never donate to people worse off than me ever again?" — You can keep doing this until you uncover all the reasons behind your desires.

This is how the tool of internalization shifts guilt: it forces the guilt to either resolve itself, or reveal itself to you in terms of your own desires. It shifts the guilt to a place where the thing the guilt demands are things you want for yourself, rather than things you want because you think you should.

So perhaps now you feel guilty for not working overtime to earn money to give to those less well-off than yourself (which is something you desire due to a deep dissatisfaction with the unfairnesses of the modern world). This, again, is progress: the guilt is now focused and internalized. This is exactly the sort of guilt that the third tool addresses.

The third tool for shifting guilt is the filter of *realism*. Look at your guilt, and ask it whether its demands are realistic.

Ask whether you really could have worked harder and done something else, while remembering that you are in fact mortal. You are no more able to work 20 hour days at peak capacity than you are able to cure Alzheimer's disease with a snap of your fingers. Look not to whether you were moving as fast as you physically could. Instead, look to the streams you need to move through in order to achieve your goals while remembering that two of the most important streams are maintaining health and motivation.

Do not ask, "could I have skipped the party and worked more?" Ask, instead, "am I traversing the work streams at the fastest sustainable pace?" Check whether the task the guilt demands is realistic. Remember that working yourself ragged is not a virtue. When keeping the filter of realism in mind, many guilt simply fail to materialize in the first place.

But some guilts do pass the filter of realism, and leave you lamenting a flaw in your process, an inability to do what you think is best. Perhaps you will notice that you attend

parties far more often than you prefer, due to peer pressure. Perhaps you will notice that you actually find parties draining, and that you were only attending this one in hopes of finding a date, which you could have done in a less costly manner if you were really trying. Perhaps you will realize that you've been adrift, that you've lost focus, and you'll feel guilty for failing to maintain your drive.

And this is right where we want the guilt. If you must feel guilty, I recommend feeling guilty not about what you did or didn't do, but about the *pattern of behavior* that corresponds to acting against your will. Don't feel guilty for going to *this* party, feel guilty for the general pattern of giving into peer pressure, or misjudging how much fun you'll have, or overindulging. Because *this* is the sort of guilt that I know how to address head-on.

The three tools of refinement, internalization, and realism, are, in my case, effectively universal: I can use them to shift any hint of guilt up to specific, internalized guilt about a realistic concern at the process level. I am sure, though, that for many of you, there will be other forms of guilt that these three tools do not cover.

This is why I make the tools explicit here: so that you can see how they work and see what they share, and then construct your own variants that work on whatever other guilts you tend to encounter.

As you hone those tools, I recommend you seek a similar endpoint: shift the guilt away from the misstep and onto the systemic flaw in your footwork. Shift guilt from the instance to the pattern. Bring your guilt to this battleground, and I will show you how to defeat it.

Don't steer with guilt

12 JULY 2015

I've spoken at length about shifting guilt or dispelling guilt. What I haven't talked about, yet, is guilt itself.

So let's talk about guilt.

Guilt is one of those strange tools that works by *not* occurring. You place guilt on the branches of possibility that you don't want to happen, and then, if all goes well, those futures don't occur. Guilt is supposed to steer the future towards non-guilty futures; it's never supposed to be instantiated in reality.

Guilt works by the same mechanism as threats: imagine the tribesperson who precommits to breaking the legs of anyone who steals their food. If this precommitment works, then it never needs to be carried out: violence is a dangerous business, and the tribesperson would much rather that they never need to break legs at all. The threat is something that the tribesperson places on possibilities that they disprefer, in attempts to ensure that they never come to be.

Imagine, by contrast, the tribesperson who threatens to breaking the legs of anyone who looks at them funny: they might find themselves attempting violence every single day, and this likely makes their life unpleasant, to say the least. In this case, I would argue that they're using their threats poorly. I would say that, if you keep finding yourself carry-

ing out a threat, then you really need to consider whether or not your threats are really capable of steering the future in the way you hoped.

Guilt is the same way: *if you find yourself regularly experiencing guilt, then you're using guilt incorrectly.*

Guilt works only when you wield it in such a way that it *doesn't happen.*

Guilt is costly when deployed. Once activated, it's usually strongly demotivating, and can easily lead to failure spirals or vicious cycles of depression.

As far as I can tell, the way that guilt-motivated people tend to operate is by working fervently in attempts to avoid the scourge of guilt. This may be effective when it works, but as soon as it starts to fail, the failure often cascades into a full-blown failure spiral (you're guilty that you're not working, which makes you feel bad, which makes it hard to work, which makes you guiltier, which you feel worse, which makes it harder to work, ...). As a result, guilt motivation often results in a boom/bust productivity/depression cycle that, as far as I can tell, results in people feeling quite bad about themselves and being much less effective than they would be if they could maintain a steady pace.

Some might argue that the boom is worth the bust, that the productivity is worth the depression. This seems straight up false to me (and I have some relevant experience): the fran-

tic productivity fueled by fear of guilt doesn't seem more effective (and often seems *less* effective) than intrinsically motivated productivity, and that's *before* we count the losses from periodic failure spirals. As far as I can tell, intrinsic motivation is just straight up more effective.

(This is something you have to accept before I can help you remove your guilt: it's much harder to remove guilt if you don't want to.)

Guilt is very costly when activated, so if it's getting activated regularly, then you're placing it on the wrong branches of possibility.

You might protest, "but then what do I *do* in the unsatisfying branches of reality? I need to find *some* way to prevent me from chasing short-term satisfaction at the expense of long-term benefits." If you regularly finding yourself binging netflix TV shows, and you would rather not find yourself regularly binging netflix TV shows, then shouldn't you feel guilty whenever you do?

No! If the situation occurs regularly, then guilt is not the tool to use! You're welcome to feel guilty if you ever kidnap a baby or punch a homeless person, and you can tell that the guilt is working in those cases because you *never do those things*. But if you repeatedly find yourself in a situation that

you disprefer, then guilt is just not the tool to use. That's not where it's useful.

If you want to figure out how to avoid a certain recurring situation, then there's a different tool that *is* appropriate, that's much more effective at figuring out how to steer the future towards better places: Science!

When you find yourself binging netflix, don't heap loads of guilt on yourself post-binge. That sort of thing clearly doesn't prevent the binge. Instead, say to yourself, "huh, I appear to netflix-binge under certain conditions, despite the fact that I'd rather not. I wonder which conditions, specifically, led to that binge! What were the triggers? How could they have been avoided? What methods might help me avoid binging in the future?"

And then treat it like an experiment! Write up your hypotheses. Experiment with many different ways to fix your glitches. Write postmortems when you fail. If you attempt a fix and then find yourself binging *again*, then don't heap loads of guilt on yourself! *That still doesn't help*. Instead, say "Aha! So *that* attempted fix didn't work. I wonder if I can figure out why?" Cross a hypothesis or two off your list. Refine your models. Expand your hypothesis space. Gather more data.

Do science to it.

Don't bemoan individual failures. That's finite-task thinking. Instead, acknowledge that there's an unlimited number of changes you'd like to make to your behavior, and that some of them are more important than others, and that some of them are more costly than others, and that they all take time to fix. See the infinite stream of self-improvement that lies before you, add it to all the other streams you're optimizing, and then simply navigate the streams as quickly as you are able.

Don't feel terrible whenever you do something you wish you hadn't! That is a poor mechanism by which to steer the future. Instead, when you do something you wish you hadn't, identify the *pattern of behavior* that led to this, and add addressing *that* to your todo list. Then weigh the time you're losing against the time it would take to change the pattern, and weigh that against the other priorities that are vying for your attention, and then do what needs doing.

Sometimes you'll ignore a pattern of failure. Maybe the failures are relatively cheap and the pattern is hard to change, and fixing the pattern simply isn't worth your attention. In this case, when the failure occurs, there is no need to feel guilty: the failures are the price you pay for time spent not fixing them. You can't simply teleport to a new pattern of behavior, and so if you lack the time to change the pattern, then the occasional failure is a fair price. Trust yourself to fix the pattern if the costs ever get too high, trust yourself to understand that investing in yourself is important, and if fixing the pattern *still* isn't at the top of your todo list, then

don't worry about the individual failures. You have bigger things on your plate.

Other times, you'll decide that the pattern needs changing. Five minutes per day is thirty hours per year, and investing in yourself pays dividends. In this case, treat addressing the pattern of failure like a science project. Every new individual failure is data point about what doesn't work. Every avoided failure is a data point about what does. Heaping guilt on yourself whenever you hit a new failure would be nonsense — fixing the *pattern* is a science experiment, and individual successes or failures are your data points.

Most people use their individual failures as a signal to themselves that it's time to feel terrible. It is much more effective, I think, to use your individual failures as a chance to update your tactics.

This, in my experience, is the head-on cure for guilt: Don't treat the individual failures like a burden; treat changing the pattern like a science experiment.

Update from the suckerpunch

19 JULY 2015

The most common objection I hear when helping people remove their guilt is something along the lines of "Hey wait! I was using that!"

Believing this (or really any variant of "but guilt is good for me!") makes it fairly hard to replace guilt with something more productive.

I've met some people who complain that if they didn't have guilt then they'd do horrible things. I think this is fairly unlikely, and I file it right next to the arguments that say that if they didn't believe in God then they'd do horrible things.
Even after dropping your obligations, you will still have something to fight for. Your *reasons* for not doing things you'd rather not do will remain even after the guilt is replaced.

Others I have met protest that guilt is useful in order to ensure that they won't repeat their failures. Without guilt, how would they learn their lesson? To which I generally say, that's fine, but if it keeps happening then you aren't learning, and it's time to use a different tool instead.

That said, there *are* lessons that need learning, and there *is* something sort of like 'guilt' that can help you learn them.

But you can use it even while completely replacing your guilt motivation.

Once upon a time, I had a loose date planned with a girl-friend. She was going to drop by around 21:00 to hang out. I had something else planned at 19:00 that I didn't expect to take too long; it ended up taking many hours longer than expected. There was no particularly convenient point along the way to step out and call my girlfriend and tell her I'd be late... so I didn't. I simply got home at 23:00 at night, opened the door, and saw my girlfriend sitting worried on the bed.

There's a very distinct type of feeling that I experienced, there, which you might call "guilt." Seeing her sitting there on the bed, I suddenly remembered that the anxiety and dejection that she went through was far worse than the slight awkwardness I would have incurred to call her. A compartmentalization in my head broke down, and the part of me that had *known* she'd been feeling terrible suddenly came into mental focus. My error became obvious. The feeling was something like being punched in the gut.

Afterwards, I *also* had the opportunity to feel a lingering sense of regret for days.

When I suggest removing guilt, I suggest removing the latter — but not the former. The former is quite useful.

If you worry that, by removing guilt, you will lose your ability to update when you mess up, then I say: update on the

suckerpunch. Trust me, it's strong enough. Update *immediately* when you realize where you failed, and use the terrible feeling to make sure you *don't do that again*.

Update fully on the suckerpunch, and there will be no need for that lingering regret. Skip to the end, immediately; update as far as you can, the moment that you realize your error. Moping for days doesn't make things better. Updating your behavior does.

There are those who still protest that the lingering regret is useful: if you hurt your friend, you may think that they need to see you spending days filled with regret, or otherwise they will think less of you. You may think that others find it disconcerting to see you update immediately and continue without missing a beat. Some people want to see penance done.

If that is your protest, then I have little to offer you. I can only note that I have seen many groups of friends form a tacit pact of non-excellence, where each individual in the group is reluctant to outperform the others, in fear that high performance will be punished with ostracization. Many have condemned themselves to a life of dissatisfaction thanks to a non-excellence pact. I say: better to inspire your friends than validate their mediocrity.

It can give some people whiplash, to see you update quickly, but I much prefer friends and lovers that encourage skipping to the end rather than those who feel a need to extract their pound of flesh whenever you err. For me, the social cost of updating quickly is well worth the ability to move faster. Your experience, of course, may differ.

Just remember that you won't be able to replace guilt-based motivation before giving yourself permission to do so. For so long as you view your guilt as an aid rather than a burden, for so long as you view it as right and necessary, I cannot help you remove it.

But I can tell you this:

Almost all emotions, I have found a place for. I have long looked upon Spock and Jedi with some dissatisfaction: I am not one to advocate suppressing emotion. Anger has its place and time, as does joy, as does sadness. Awe and fear and cold resolve, I have found a use for.

I have even found a use for that suckerpunch that occurs when you learn you have made a mistake, that you might label 'guilt.'

But the lingering, drawn-out guilt, the persistent regret that drives one to work in fear of it?

I have never once found a use for that.

Be a new homunculus

26 JULY 2015

Here's a mental technique that I find useful for addressing many dour feelings, guilt among them:

When you're feeling guilty, it is sometimes helpful to close your eyes for a moment, re-open them, and pretend that you're a new homunculus.

A "homunculus" is a tiny representation of a human, and one classic fallacy when reasoning about how brains work is the homunculus fallacy, in which people imagine that "they" are a little homonculus inside their head looking at an image generated by their eyes.

It's an easy fiction to buy into, that you're a little person in your head that can move your hands and shape your mouth and that decides where to steer the body and so on. There is, of course, no homunculus inside your head (for if you are steered by a homunculus, then how is the homunculus steered?), but it can be quite fun to pretend that you are a homunculus sometimes, mostly because this allows you to occasionally pretend you're a *new* homunculus, fresh off the factory lines, and newly installed into this particular person.

Close your eyes, and pretend you're arriving in this body for the very first time. Open them and do some original seeing on this person you now are. Rub your hands together, look around, and take stock of your surroundings. Do some inter-

nal checks to figure out what this body values, to figure out what it is you're fighting for. Check the catalog of plans and upcoming actions. Check the backlog of memories and obligations.

There will probably be some housecleaning to do: homunculi are known to get a little careless as they age, and the old homunculus that you replaced probably let a bunch of useless tasks accumulate without realizing it. As a new homunculus you have the privilege of pruning the things that obviously need pruning. Maybe you'll look and say "Ah, yes, we're going to cancel lunch with *that* person; this body was secretly dreading it. I also see that this body is currently spending a lot of cycles feeling guilty about a date that went poorly last week; we can dismiss that, it's no longer useful for *this* homunculus. And also, "exercise" doesn't seem to be on today's schedule at all! How strange. This body definitely intended to exercise today; somehow it fell off the list. I'll put it back on."

It can be quite liberating to be a new homunculus, without any obligation to propagate the errors of the old one.

This is, in fact, a common technique for dealing with the sunk cost fallacy (also known as the "pretend you're a teleporting alien that just teleported into your body" technique). This is useful for avoiding sunk costs because the *new* ho-

munculus has no reason to honor the old homunculus' sunk costs.

Say the old homunculus bought plane tickets which would let you travel to Texas tomorrow (and return in a week), and that the ticket is non-refundable. The old homunculus may well have an attachment to the "go to Texas" plan, and may try to convince themselves to go even when it becomes clear that the trip won't be worth the time. The new homunculus, however, has no such loyalty to the sunk costs: *it* can just evaluate whether or not to go on the trip regardless of how much the tickets costed.

This is also a technique that works quite well for managing guilt: it's often easy for the new homunculus to recognize lingering guilt as a bodily response marking malcontent about something that was done in the past, by the old homunculus. The best action for the new homunculus to take, usually, is to check what regretted action caused the guilt, check what pattern of behavior led to the regretted action, mark down a note about which cognitive pattern needs to be reprogrammed, and then dismiss the guilt (which has now served its purpose).

As a matter of fact, guilt and sunk cost fallacy are closely related: both are about suffering for costs that were paid in the past. The only difference is that guilt carries with it a lesson, an instruction to alter your environment and your mind so that similar actions don't occur in the future. With practice, it is possible to reflexively treat the initial gut-

wrenching guilt as an instruction to update your behavioral patterns, and then dismiss the lingering guilt immediately.
(Cognitive patterns, after all, take some time to train.)

In the interim I suggest pretending you're a new homunculus. If you start to feel guilt, then close your eyes and re-open them as a brand new homunculus. Notice the guilt, listen to the message it bears, and *actually write down* the behavioral pattern that you wish to change. Then spend five minutes (a full five minutes, by the clock) brainstorming ways that you might change the pattern and start retraining your mind. Then thank the guilt for carrying you this message, and dismiss it.

Eventually, this can become reflexive. Until then, I suggest occasionally becoming a new homunculus. In fact, I often use something like this myself, even though I've been immune to guilt for quite some time: it's a great way to see the world and yourself with fresh eyes, and that can be invaluable.

Not yet gods

09 AUGUST 2015

You probably don't feel guilty for failing to snap your fingers in just such a way as to produce a cure for Alzheimer's disease.

Yet, many people *do* feel guilty for failing to work until they drop every single day (which is a psychological impossibility). They feel guilty for failing to magically abandon behavioral patterns they dislike, without practice or re-training (which is a cognitive impossibility). What gives?

The difference, I think, is that people think they "couldn't have" snapped their fingers and cured Alzheimer's, but they think they "could have" used better cognitive patterns. This is where a lot of the damage lies, I think:

Most people's "coulds" are broken.

People think that they "could have" avoided anxiety at that one party. They think they "could have" stopped playing Civilization at a reasonable hour and gone to bed. They think they "could have" stopped watching House of Cards between episodes. I'm not making a point about the illusion of free will, here — I think there *is* a sense in which we "could" do certain things that we do not in fact do. Rather, my point is that most people have a miscalibrated idea of what they could or couldn't do.

People berate themselves whenever their brain fails to be engraved with the cognitive patterns that they wish it was engraved with, as if they had complete dominion over their own thoughts, over the patterns laid down in their heads. As if they weren't a network of neurons. As if they could choose their preferred choice in spite of their cognitive patterns, rather than recognizing that choice *is* a cognitive pattern. As if they were supposed to *choose* their mind, rather than *being* their mind.

As if they were already gods.

We aren't gods.

Not yet.

We're still monkeys.

Almost everybody is a total mess internally, as best as I can tell. Almost everybody struggles to act as they wish to act. Almost everybody is psychologically fragile, and can be put into situations where they do things that they regret — overeat, overspend, get angry, get scared, get anxious. We're monkeys, and we're fairly fragile monkeys at that.

So you don't need to beat yourself up when you miss your targets. You don't need to berate yourself when you fail to

act exactly as you wish to act. Acting as you wish doesn't happen for free, it only happens after tweaking the environment and training your brain. You're still a monkey!

Don't berate the monkey. *Help* it, whenever you can. It wants the same things you want — it's you. Assist, don't badger. Figure out how to make it easy to act as you wish. Retrain the monkey. Experiment. Try things.

And be kind to it. It's trying pretty hard. The monkey doesn't know exactly how to get what it wants yet, because it's embedded in a really big complicated world and it doesn't get to see most of it, and because a lot of what it does is due to a dozen different levels of subconscious cause-response patterns that it has very little control over. It's *trying*.

Don't berate the monkey just because it stumbles. We didn't exactly pick the easiest of paths. We didn't exactly set our sights low. The things we're trying to do are hard. So when the monkey runs into an obstacle and falls, help it to its feet. Help it practice, or help it train, or help it execute the next clever plan on your list of ways to overcome the obstacles before you.

One day, we may gain more control over our minds. One day, we may be able to choose our cognitive patterns at will, and effortlessly act as we wish. One day, we may become more like the creatures that many wish they were, the imaginary creatures with complete dominion over their own minds many rate themselves against.

But we aren't there yet. We're not gods. We're still monkeys.

Where coulds go

17 AUGUST 2015

Most people don't think they "could" cure Alzheimers by snapping their fingers, and so they don't feel terrible about failing to do this.

By contrast, people who fail to resist overeating, or who fail to stop playing Civilization at a reasonable hour, feel strongly that they "could have" resisted, and take this as a license to feel terrible about their decisions.

As I said last week, most people have broken "coulds."

Willpower is scarce in this world. Sometimes, you can will yourself out of a mental rut you're in, but only rarely; more often, sheer force of will alone is not sufficient. If your plan to stop staying up too late playing Civilization is "well I'll just force myself harder next time," then this plan is doomed to failure. If it didn't work last time, it likely won't work next time. Willpower is a stopgap, not a remedy.

I think that most people's "coulds" are broken because they put the action nodes in the wrong place. They think that the "choice" occurred at turn 347 of Civilization, when they decided to continue playing one more round (and at each following turn between midnight and 4:00 in the morning).

But that's not where the choice occurred. If you have to force yourself to change your behavior, then you've already

missed the real choice node.

The actual choice occurs when you decide *whether to play Civilization or not*, at the very beginning.

Say you have one acquaintance in your social circles who regularly frustrates you, and every so often, you explode at them and get into a big shouting match. You *know* you shouldn't start yelling at them, you *try* to not be frustrated. Whenever they start annoying you, you *will* yourself to cool down, but it never quite works (no matter how strongly you resolve to force yourself harder next time). In this case, I suggest that you stop trying to force yourself to hold back as your frustration peaks, and instead start noticing what happens *five minutes before* you explode. *That's* where the real choice is. The real choice isn't in whether or not you explode *in the moment*, it's in whether you exit the situation five minutes earlier.

The real choices tend to happen a few minutes before the choices that people beat themselves up about. If you have to apply willpower, you've already missed the choice node. (In fact, I've previously suggested promising yourself that you'll never pull yourself out of a situation using willpower — knowing that you *won't* save your own ass if you get into a situation where you need willpower to extract yourself really makes you notice the true point of no return when it comes along.)

If you find yourself in a pattern of behavior you don't like, then I recommend pretending you don't have *any* willpower. Imagine you lived in the world where you *couldn't* force yourself to stop doing something addicting after starting. In that world, how would you act?

Look for the triggers that precede the action you wish you could make differently. What happens an hour beforehand? What happens five minutes beforehand? What happens sixty seconds before you fail to act as you wish?

That's where the real choice lies.

Most people's coulds are broken. They treat themselves like they "could" start bingeing a TV show and then stop at a reasonable hour. They put themselves in a situation that tempts them against their better judgement, and then berate themselves when they succumb.

By contrast, I don't treat myself as if I "could" stop binge-reading a good book, and therefore I don't feel terrible if I binge. Instead, I say, "ah, I see, I binge-read engaging books; I will treat 'read an engaging book' as a single atomic action that takes five to twenty hours, with no choice nodes in between." Where others are berating themselves for failing to complete an impossible task ("stop binge-reading halfway through and get back to real work"), I am learning what I

am and am not capable of, and learning where my real action nodes are.

We humans don't *have* all the choice nodes. Sometimes, we can't stop binge-reading a good book anymore than we could snap our fingers to cure Alzheimer's disease. Sometimes, addiction takes over; other times, the lizard brain takes over; other times, primal rage takes over. In those moments, we don't get to call the shots. We aren't the choice-makers at every point in our lives. We often lack the willpower to override our impulses, instincts, and habits.

The goal is to win anyway.

Our better judgement is not the absolute arbiter of our actions, and there are often times when the voice of judgement is nearly powerless to affect our behavior. We aren't yet gods. We're still monkeys. Still neural nets.

I suggest you stop berating yourself for failing to complete impossible tasks, and start experimenting and identifying which action nodes *work*.

Search for the choices that let you act as you wish *before* the decision gets difficult to execute. Learn how to identify the moments when your mind is readily responding to your will. Those are the real choice-points, and it is from there that you may optimize.

Self compassion

25 AUGUST 2015

Imagine a time when you were feeling guilt-wracked. Maybe a time you hurt a friend badly. Maybe a time you tried to do get some important work done, and found you couldn't, and this kicked off a failure spiral leading to a deep depression. Maybe some other time: the important thing is to load into memory a time you felt guilt-wracked, and recall how you felt towards yourself in that case.

(When I do this, I get an internal sense of resistance, of not-wanting-to-look, of willing-the-past-to-be-different.)

Now imagine you have a child, who grows to the same age that you were then, who finds themselves in exactly the same situation. Maybe they, too, hurt somebody badly -- they didn't consciously realize how badly they were about to hurt a friend until one moment too late, and now they feel terrible. Or maybe they, too, tried to do something important, and found it hard, and started doubting themselves, and spiraled downwards into a depression that they now have trouble climbing out of.

Imagine what you might feel towards your child, in this scenario.

(When I do this, I get a sense of compassion, of protective-ness, and a desire to reassure them that this is what it looks like to learn hard lessons, for us monkeys.)

I encourage you to simulate the feelings you would feel towards your child in this situation —

— and then check whether you can *also* feel that way towards *yourself*.

When you think of your own failings, can you feel that compassion and protectiveness and impulse to reassure towards *you*?

Many can't. Some don't feel compassion towards others in the first place (this post is not for them — if you want help feeling compassion towards your fellow humans, then maybe try [this post](#) and see if it works for you.) Others can as easily feel compassion for themselves as others. But many people I've spoken to experience a wide gulf between compassion for others and self-compassion — which is a shame, because self-compassion is an important part of self-loyalty and the mental toolset I'm trying to convey with these posts.

To close the gap between compassion and self-compassion, I offer two tools. The first is a reminder that self-compassion is not the same thing as self-pity, and nor is it the same thing as making excuses for yourself. It is well possible to feel self-compassion even while thinking that you are not moving fast enough. It is perfectly possible to feel self-compassion

even as you notice that you're completely failing to act as you wish to.

For example, imagine someone going through boot camp in World War II, filled with resolve and determination to become a soldier and defend the free world — except they are a small person, and a weak one. Imagine them working their heart out, trying as hard as they can, and failing anyway. Imagine them failing to make the cut. Now, can you imagine feeling compassion for them, feeling warmth towards them, and maybe feeling a hint of sadness for their loss, without feeling any sense of pity? Compassion for yourself can be similar, without any hint of pity.

Or imagine another person going through the same boot camp, who really wants to go defend the free world with all their peers (on some level), but who lacks the deep drive. They *want* to feel the same passion and fire as their diminutive counterpart, but instead they feel resistance and suffer from depression — and every day they drag themselves out of bed (slightly too late), and every day they force themselves through the obstacle courses (but not quickly enough), and they aren't going to make the cut, and they're sick with guilt about it. Can you imagine feeling compassion for them in their plight, while making absolutely no excuses for their performance? Again, self-compassion can be the same way. You don't need to make excuses for yourself, to take the outside view and feel the same warmth for a monkey that's *trying* to try, against the gradient of depression and doubt.

Now imagine someone else doing what you're trying to do. Imagine them working on hard problems, and putting in what effort they can muster — sometimes it is enough, sometimes it isn't; sometimes they are highly motivated, other times they are blocked by their own mind and unable to act as they wish. Look at them and see the fragile monkey trying to build a satisfactory life, trying to improve their world. See if you can feel compassion for them. You don't need to pity them, you don't need to make excuses for their failures, you don't need to find ways they could improve: simply see if you can feel some warmth, for a fellow lost monkey — and then shift your gaze to yourself, and see if you can feel a similar sort of warmth.

The second tool I offer, to close the gap between compassion for others and compassion for yourself, is this: I recommend that you pinch yourself, and remember what you are. Practice original seeing while looking upon yourself and your situation. What do you see?

I see bundles of proteins and lipids arranged in a giant colony of cells, their lives given over to the implementation of a wet protein computer that thinks it's a person.

I see fractal patterns that arise on precisely the right sort of planet when you pour sunlight into it for a billion years.

I see wiggles in the Sun's wake that struggle to understand the universe. Incomprehensibly large constructs made of atoms, which are unnoticeably small on the scale of galaxies.

Look at us, the first species among the animals that can figure out what the stars are, yet still tightly bound to impulse and social pressure. (Notice how silly it is, monkeys acting all serious and wise as they try to affect the course of history.)

Look at us: half monkey, half god; towering below the stars.

Look at whatever quest you've taken on, you who was forged by the death of your father's brothers and now claims dominion over the future. Acknowledge that what you're trying to do is difficult. Turn the monkey sight on yourself, and see the lost monkey who's trying to steer an entire universe...

and say hello. Check in with the monkey. See how it's doing.

Steering the future is a difficult thing. The world is large beyond comprehension, and the monkey wasn't really built for this. The monkey isn't really used to this sort of thing, and it can be pretty hard to work with sometimes.

Let the monkey know that you have its back. Let it know that you'll still have its back, even if it gets ornery or difficult or depressed. Through thick and thin, let you know that

you have your support; that even when you screw everything, you'll stand by yourself, and help you through the mess, and help you figure out how to do better in the future.

See if you can resolve to work with yourself. You can do powerful things, if you work together.

There are no "bad people"

30 AUGUST 2015

When I help friends debug their intrinsic motivation, here's a pattern I often bump into:

Well, if I don't actually start working soon, then I'll be a bad person.

Or, even more worrying:

Well they wanted me to just buckle down and do the work, and I really didn't want to do it then, which means that either they were bad, or I was bad. And I didn't want to be the bad one bad, so I got angry at them, and...

I confess, I do not know what it would mean for somebody to be a "bad person." I do know what it means for somebody to be bad at achieving the goals they set for themselves. I do know what it means for someone to be good at pursuing goals that I dislike. I have no idea what it would mean for a person to "be bad."

I know what it means for a person to lack skill in a specific area. I know what it means for a person to be procrastinating. I know what it means for a person to be acting under impulses that they don't endorse, such as spite or disgust. I know what it means for someone to fail to act as they wish

to act. I know what it means for someone to hurt other people, either on purpose or with a feeling of helpless resignation.

But I don't know what it would mean for a person to "be bad." That fails to parse. People don't have a hidden stone deep inside their brain that is either green or red depending on whether they are good or bad. "Badness" is not a fundamental property that a person can have. At best, "they're bad" can be shorthand for either "I don't want their goals achieved" or "they are untrained in a number of skills which would be relevant to the present situation"; but in all cases, "they are bad" must be either shorthand or nonsense.

Asking whether a person is "fundamentally good" or "fundamentally bad" is a type error. Life is not a quest where you struggle to wind up "good." That's not the sort of reality we find ourselves in.

Rather, we find ourselves embedded in a vast universe, with control over the future and a goal of making it wonderful. We find ourselves to be part of a grand deterministic pattern, and we're trying to make that pattern as beautiful as possible.

Step back and imagine history as a fixed path through the great crystal that is our universe over all time; the time-crystal that describes everything everywhere and every-when; the time-crystal where you can look not only forwards and backwards, but beforewards and afterwards.

Imagine the path of history that dances through configurations to the tune of physics. That same physics, according to which the line jigs and jags, is what implements you. In those jigs and jags is the pattern that is your mind. Some of the jigs compute your thoughts, some of the jags compute your choices, and your choices determine how the line dances in the afterwards direction past the event of your choice.

We aren't here to alter the color of the fundamental "goodness" stone buried within us; we're here to make the path through time be a good one.

Life is not a game of "wind up good at the end"; life is about steering the future.

Look not to whether you are good or bad. Look to where you are, and what you can do from there.

Living this mindset does not mean that you lack regrets. It does not free you from the burdens of your wrongdoings. I, like anyone, suffer from recalling harms that I have done to others. But instead of treating those recollections as dark judgements on my soul, I treat them as messages from my past, information about what sorts of undesirable behavior the Nate-monkey is liable to execute if I am not careful.

I sometimes find myself unable to act as I wish; unresponsive to my own cajoling. I treat these not as evidence of my fundamental brokenness, but as evidence about how and when I can intervene on the world.

While I often fail, I do not act under fear of being judged inadequate by the universe. I may *be* inadequate to the tasks I undertake, I may fail to steer the future as I wish to, but I cannot be "fundamentally bad." That sentence does not parse.

There is something freeing about this: I may succeed; I may fail; but I will not be judged by someone who roots through my mind to see whether the stone is green or red.

I will be judged only by the path that the future takes; as will we all.

By contrast, when I help friends debug their motivation, I often find them motivated by a desperate attempt to avoid "being bad."

Where I can, I encourage them not to let that be at the core of what motivates them. It's well and good, when introspecting about why what you're doing is important, to get an answer from yourself that is of the form "otherwise I'll be bad." That's a fine answer to get. But *don't let that be the end*

of things. Don't pretend that that's the final answer.
Investigate.

Ask yourself, "what do I mean by that?" Say to yourself, "I bet that's shorthand for something." *Unpack* the feeling of would-be-bad.

If someone wants you to do the laundry, and you don't want to do the laundry, and you get angry at them because you have a sense that if there is conflict then one of you must be bad and *you* don't want to be bad—

—then pause, and investigate further.

Focus, and ask yourself what bad thing would happen if you did do the laundry, and what bad thing would happen if you didn't.

Maybe you get an answer like "if I don't do the laundry then it will strain my relationship with my friend, but if I do do the laundry then it will spend scarce energy and attention and I'm feeling really exhausted and don't want to force myself to do it."

That's great! (The answer doesn't need to be *comfortable*, it just needs to be *unpacked*. You may well reveal conflicting desires. You may well find that you were ignoring goals that you had but didn't endorse, such as preserving your own attention or energy.) This is a similar mental action to unpacking a should: if you find yourself compelled to do something

because otherwise you'd "be bad," then become curious, investigate, and unpack the feeling into its component parts.

Ask yourself, "I don't know what it would mean to be bad; can you elaborate?"

Then, *listen to yourself*. Don't worry if your answers seem senseless! Often, I have watched people completely fail to figure out what is blocking them, because as soon as they get an answer from deep inside their mind, they declare that it's ridiculous, and then they struggle to dismiss it or cover it up or decry it as "irrational."

Perhaps they ask themselves what they mean by "then I'd be bad" and find something like "I apparently think that if I don't do the laundry then it's evidence that I can't do *anything*, and that means I'll lose my job and end up on the street and die cold and alone, and that's *stupid*, so..." at which point they start lecturing themselves about why their concerns are dumb, instead of declaring self-loyalty and standing by themselves. (If you find yourself doing this, I suggest taking your concerns seriously, and explaining your different beliefs earnestly, with the same respect you'd show an inquisitive child who wants to understand the world but has a few flaws in their understanding.)

You're still a monkey! You often have inconsistent, strange preferences. Parts of you often have beliefs that other parts of you don't endorse. That's *ok*. Decrying your own inconsistencies is no way to fix them: work with yourself.

So don't settle for being motivated to do something because otherwise you'd "be bad." Unpack the feeling of "being bad," and figure out what outcomes you're aiming for. Figure out what you want to *do*. Figure out how you want the future to *be*.

Because at the end of the day, a person "being bad" fails to parse. "Goodness" and "badness" are not properties of people. People can do terrible things; they can pursue horrible goals; they can watch with growing despair as they act against their own best interests; but they do not have a fundamental stone buried deep inside of them which measures their worth.

Life is not a game of "wind up good at the end." Life is about steering the future.

Residing in the mortal realm

06 SEPTEMBER 2015

The last sevenish posts describe the main tools I have for removing guilt-based motivation. The common thread running through them can be summed up as follows: *Reside in the mortal realm.*

Many people hold themselves to a very different standard than they hold others. They hold themselves accountable for failing to do the psychologically impossible. They fret over past mistakes and treat themselves as failed gods, rather than ambitious monkeys. This condemning-of-the-self can lead to great guilt, with all its negative effects.

My suggestion for dealing with guilt, roughly speaking, is to first focus your guilt, by dispelling the guilt that comes from not doing what other people think you should or from from false obligations, and shifting all your guilt into guilt about the fact that you have not yet made the future how you want it to be. Then, once your guilt is focused there, remember that you are a denizen of the mortal realm.

In the past, you have failed to act as you wished to act. You have failed to make the best available choices. But these facts have little bearing on what you do next. They have *some* bearing, insofar as your memories still hold lessons that can teach you about how to better steer yourself to

steer the world, but they do not say anything about the color of your soul. They are simply the background knowledge against which you move forwards, from here, looking only towards the future.

You are a mortal, who often struggles to follow their own will, and your actions set the course of the entire future. Instead of berating yourself for your shortcomings, figure out how to do the best you can *given* the shortcomings — sometimes by spending time and effort to fix them (mere willpower seldom suffices), and sometimes by taking them as given and working around them.

Be a mere mortal, and do the best you can anyway. Learn everything you can from your mistakes, and then forgive yourself your sins, and look only to how much better you can make the future (knowing what you know now about how you perform in different situations).

Guilt has no place among mortals: we already *know* we're fallible. We don't need to suffer over that fact: our failings provide only information about what to do next, if we want to steer the future.

Over the last few months, three different people have informed me that I broke their motivation systems. In short, one found themselves less able to care about what they

were working on, another found themselves unable to force themselves to work, another found themselves unable to continue spurring themselves on with guilt.

In part, this is working as intended: in the long run, I think that guilt-based motivation can be harmful. However, my goal is not to simply remove existing motivation systems: my goal is to replace guilt with something else.

So the question is, without guilt, what can you use for drive? And this brings us to the penultimate arc of my "replacing guilt" series of posts.

I've already given partial answers to the question "whence internal drive?", when talking on caring, or about the value of a life, or about caring about something larger than yourself. Those posts are intended to inspire you and remind you that there's something *worth* fighting for, and that you can fight for it even if you lack a burning passion. That's not the whole picture, though, and in the upcoming arc, I'll touch upon a different aspect of intrinsic motivation.

I think many people are motivated by an intrinsic (often subconscious) desire to be virtuous, or perhaps by a strong aversion to "being bad." I think many other people are motivated primarily by whatever obligations currently sit on their plate. They don't need to ask themselves what they are doing or why; they simply continue fulfilling the obligations in front of them so that life continues proceed. They fulfill obligations at school, they fulfill obligations at their jobs,

they find a spouse, they start a family, they fulfill obligations to their family. The obligations keep flowing in a steady stream, and there is never any need to soul-search in a grand quest for some sort of deep intrinsic drive (except, perhaps, during the occasional "midlife crisis," which is a fine distraction that they're expected to eventually overcome).

Yet here I stand, suggesting that you ditch the notion "being bad" and drop your obligations entirely, keeping only what remains. But dropping an existing framework is a far cry from creating a new one, and dropping guilt does not often reveal a blindingly virtuous non-obligation that you're supposed to pursue instead of what you were currently pursuing.

In fact, the new framework can't contain "supposed tos" at all. Obligations have been jettisoned.

So in the upcoming arc, I'm *not* going to give you something to pursue. Rather, I'm going to do my best to give you a different way of looking at the world. I'm going to describe a vantage point from which guilt motivation seems quaint, and something else — maybe cold resolve, maybe hot desire, maybe a different drive — guides your actions instead.

From that vantage point, guilt is alien — and it is only once it seems foreign (rather than evil) that it be fully replaced.

Being unable to despair

13 SEPTEMBER 2015

Content note: these next few posts are not going to be for the faint of heart.

Sometimes, when people see that their life is about to get a lot harder, they start buckling down. Other times, they start despairing, or complaining, or preparing excuses so that they can have one ready when the inevitable failure hits, or giving up entirely and then failing with abandon. These next few posts assume that you have the former demeanor, and they might not be helpful to people who are inclined to respond to new difficulties with despair. Remember the law of equal and opposite advice! (For every person who needs a certain piece of advice, there is someone else who needs the opposite advice.)

With that said, I'm going to spend a few words giving some tips about how to have the former demeanor, if you want to. The first piece of pertinent advice is that the way you respond to challenges is context dependent; even if you've already been known to respond to some problems by despairing, there are likely other problems that you respond to by buckling down.

There is a specific mindset that, in my experience, makes it much easier to adopt the "buckle down" demeanor. This is the mindset where "not doing anything" doesn't seem like an available option in the action-space. I've written a bit be-

fore about how I think many people think there is a default "rest state", and this is a related concept: many people seem to think that there is a privileged "don't do anything" action, that consists of something like curling up into a ball, staying in bed, and refusing to answer emails. It's much easier to adopt the "buckle down" demeanor when, instead, curling up in a ball and staying in bed feels like *just another action*. It's just another way to respond to the situation, which has some merits and some flaws.

So this is my second piece of advice, if you want to be the sort of person who buckles down in the face of hardship: see the world in terms of possible responses. See curling up in bed and ignoring the world as just *one possible response*, rather than an escape hatch. Dispel the illusion that some actions are labeled "do nothing," and notice that those, too, are responses. There is no privileged null choice.

(That's not to say that it's *bad* to curl up in a ball on your bed and ignore the world for a while. Sometimes this is exactly what you need to recover. Sometimes it's what the monkey is going to do regardless of what you decide. The point is that when nature offers you a choice, there is no "don't choose" option. There are only the options that nature offers, and all you can do is pick the best of them.)

My third piece of advice is to remember that you reside in the mortal realm. If you get new information or a new way of looking at the world and you start to feel despair, or hopelessness, or helpless, or impotent, then it is *perfectly OK*

to respond by curling up in a corner and feeling sad and scared and small for a little while. That's a fine response. It doesn't mean that you're not up to the task. Nor does it mean that you are condemned to despairing forever. You're allowed to feel small sometimes, and then get back up and keep going, without any need to pretend that things are fine. We're monkeys. Feeling helpless happens.

Rising to the challenge doesn't mean never feeling helpless. It means pushing on *anyway*, even if you feel helpless sometimes.

In my experience, tapping into internal drive often requires tapping into a deep desire to *make the world be different*, in a world that's very large and very hurting and very hard to change. When trying to do this, it can be easy to get overwhelmed by the odds stacked against you — regardless of their scale. (In fact, I have often found that the cards stacked against me personally — when I feel isolated, lonely, or friendless — induce as much despair as the cards stacked against anyone who tries to change the world at large.)

In the next few posts, I'm going to talk about tapping into that internal drive, and this will entail trying to see the situation for what it really is: which means owning up to everything stacked against you. If you aren't careful, this might cause you to buckle. But if you do it right, it can cause you to buckle down instead, and provide a source of drive.

See the dark world

20 SEPTEMBER 2015

Consider fictional Carol, who has convinced herself that she doesn't need to worry about the suffering of people who live far away. She works to improve her local community, and donates to her local church. She's a kind and loving woman, and she does her part, and (she reasons) that's all anyone can be expected to do.

Now consider fictional Dave, who failed a job interview. When telling his friends the story, he emphasizes how the interviewers were biased against him, and how they asked stupid questions.

Meanwhile, driven by hunger, a fox tries to reach some grapes hanging high on the vine but is unable to, although he leaps with all his strength. As he goes away, he remarks "Oh, you aren't even ripe yet! I don't need any sour grapes."

All of these reactions — and many others — share a common kernel. Carol, Dave, and the fox are all inventing reasons why an unpleasant state of affairs is acceptable. They're not inventing reasons why the world is *good*, by any means; but they are putting forth cognitive effort to make it seem *tolerable*.

Carol would surely tell you that it's terrible that children are suffering abroad — but only after convincing herself that her duty to help them had been discharged.

The fox would tell you that the world is worse for being full of sour grapes — and yet, he still had to work hard to assure himself that he didn't live in a far worse world, where the grapes were both ripe and inaccessible.

There's a certain type of darkness in the world that most people simply cannot see. It's not the abstract darkness: people will readily acknowledge that the world is broken, and explain how and why the hated out-group is responsible. And that's exactly what I'm pointing at: upon seeing that the world is broken, people experience an impulse to explain the brokenness in a way that relieves the tension.

When seeing that the world is broken, people *reflexively* feel a need to explain. Carol can acknowledge that there is suffering abroad, but this acknowledgement comes part and parcel with an explanation about why she bears no responsibility. Dave can acknowledge that he failed to pass the interview, but his mind automatically generates reasons why this is an acceptable state of affairs.

This is the type of darkness in the world that most people cannot see: they cannot see a world that is *unacceptable*. Upon noticing that the world is broken, they reflexively list reasons why it is still tolerable. Even cynicism, I think, can fill this role: I often read cynicism as an attempt to explain a world full of callous neglect and casual cruelty, in a framework that makes neglect and cruelty seem natural and expected (and therefore tolerable).

I call this reflexive response "tolerification," and if you watch for it, you can see it everywhere.

The sour grapes fallacy is a clear example of tolerification, but it's only one instance of the broader class. Tolerification occurs *any* time you see something bad in the world and feel an impulse to explain, especially if that explanation relieves pressure that would otherwise be placed on you.

Consider, for example, Alice and Bob in my allegory of the dragon. Both have recently learned that the market value of a life is only a few thousand dollars. Both are uncomfortable with this, and they reflexively tolerify the information in different ways.

Bob denies the information, protesting that one can't make decisions by attaching dollar values to lives, because lives are sacred. This declaration of a sacred value allows Bob to deny the discrepancy entirely, reject the implied responsibility, and restore tolerability to the universe.

Alice, by contrast, accepts the data and denies the intuition that lives are sacred. She notes that if you act like lives are worth *more* than a few thousand dollars then you'll save fewer lives than you could, and thus anyone who acts otherwise and wants to save lives is inconsistent. Therefore, she concludes that she can't treat the intrinsic value of a life as

worth any more than the market price, and grows cynical — not only are lives non-sacred, she realizes, but they're not worth that much more than a few thousand cans of coke. Now she can worry less about saving lives: they weren't worth as much as she thought, anyway. Tolerification successful.

Notice how their gazes slip to one side or the other, both of them failing to see the dark world — the one where lives are *both* nigh invaluable, *and* priced at \$3000. The one where it's *reprehensible* to pretend that a life is worth only as much as a few thousand cans of coke, *and* this is how you have to price a life if you want to save as many lives as you can. The world with a grim gap between life's price and life's value. This is the world that both Alice and Bob both reflexively tolerify away from.

In me, tolerification is toxic to intrinsic motivation. If you want intrinsic drive, I suggest you train yourself to notice when your gaze slips to one side or the other. When that happens, focus, and stare directly at the dark world.

Content note: the remainder of this post encourages you to contemplate and acknowledge significant difficulties in your own life. I assume that the reader is resilient in the face of adversity. If acknowledging adversity in your life is currently liable to harm you, consider skipping the rest of this post.

My favored tool for subverting the impulse to tolerify the intolerable (and thereby stare directly at the dark world) is to pose myself a "what if" question.

What if I lived in the world where it was *both* the case that lives are nigh invaluable, *and* it costs only a few thousand dollars to save a life?

What if I lived in the world where it was *both* the case that I failed the interview *and* it was because I lacked the requisite skill?

The default impulse, upon learning that I failed the interview, might be to tolerify. Someone prone to tolerification might automatically, reflexively, start listing ways that the interview was stacked against them, or reasons why the questions were stupid, or reasons why they didn't want the job anyway. Then they might jump directly into the next interview, with excuses already in hand for when they fail that one too. This illustrates one major way that tolerification can be harmful: it might prevent you from seeing what really needs to be done. The person who refuses to tolerify can seriously consider spending more time practicing, or switching careers. If necessary, they can acknowledge that they really need to get a job while still dramatically unqualified, and decide to play the numbers with full knowledge of what they're doing. If they tolerify, they have to act indignant when they fail. If they don't, they can face what needs to be done.

Refusing to tolerate in this situation can be *really really hard*. Saying "It seems I am not yet be skilled enough to get a job in this field" can be *tough*, especially when your livelihood depends upon the opposite being true (and double-especially if you think that past failures make you a "bad person").

The nice thing about the "what if" question is that I don't need to *believe* that that's the actual world when pondering the "what if". I don't need to *acknowledge* that I am unqualified for the job, I can simply ask what *would* do if I were. This makes it easier to plan out what I would do if I could see the dark world, and having a plan often makes it easier to acknowledge that the world I'm living in is dark. (See also: leaving yourself a line of retreat.)

So, let's run through some what ifs.

What if we lived in the world where it was *both* the case that (a) unwanted pregnancies could ruin the lives of both mother and child *and* (c) unborn children were moral patients with a right to life? What would you do then?

What if we lived in the world where it was *both* the case that (a) people are living and dying in extreme poverty *and* (b) you really need a new car soon if you want to keep your job, but you could spare a few thousand dollars if you really had to. What would you do then?

What if we lived in the world where people do have souls, but they're implemented on brains made of meat that rots

when you die?

What if we lived in the world where evolution built conscious predators, and conscious prey that suffers as it gets eaten alive?

What if almost nobody was evil, but almost everything was broken anyway? What if the hated out-groups *aren't* responsible for all the suffering?

I'm not claiming that these what-ifs are accurate. Rather, I offer this as a tool for staring the dark world directly in the face. Imagine the world that is as bad as it might be. Imagine the world were full of intolerable injustices. What would you do then?

Can you look upon those dark worlds and feel a sense of despair, of the world being harder to fix than seems acceptable? Do you get a feeling of bracing yourself for making terrible tradeoffs, because there are too many problems and you can't handle all of them? If so, that's good: that's what it feels like, to see the dark world.

The question is, what would you do *then*?

I'm not here to offer answers. Maybe your answer is "well in that world I'd stop trying so hard and move to a cabin in the

woods and try to forget how screwed up everything was." Or maybe your answer is "in that case I'd rise to the challenge, no matter how terrible the odds." More likely, it's something else entirely. I'm not trying to feed you answers. I'm trying to help you refuse to tolerate, because there is a source of resolve that comes only when you see the dark world.

I have to believe this falsehood, because otherwise I would be unable to go on.

This is something that I hear fairly frequently, either to my face, or in popular media. "I have to believe in God; otherwise there would be no meaning in my life." Or "It's a good thing humans are unrealistically optimistic; we wouldn't be able to handle reality." Or "I have to believe that I'm going to get this job; otherwise I wouldn't be able to continue trying." Or,

"All right," said Susan. "I'm not stupid. You're saying humans need... fantasies to make life bearable."

REALLY? AS IF IT WAS SOME KIND OF PINK PILL? NO. HUMANS NEED FANTASY TO BE HUMAN. TO BE THE PLACE WHERE THE FALLING ANGEL MEETS THE RISING APE.

"Tooth fairies? Hogfathers? Little—"

YES. AS PRACTICE. YOU HAVE TO START OUT LEARNING TO BELIEVE THE LITTLE LIES.

"So we can believe the big ones?"

YES. JUSTICE. MERCY. DUTY. THAT SORT OF THING.

"They're not the same at all!"

YOU THINK SO? THEN TAKE THE UNIVERSE AND GRIND IT DOWN TO THE FINEST POWDER AND SIEVE IT THROUGH THE FINEST SIEVE AND THEN SHOW ME ONE ATOM OF JUSTICE, ONE MOLECULE OF MERCY. AND YET—*Death waved a hand.* AND YET YOU ACT AS IF THERE IS SOME IDEAL ORDER IN THE WORLD, AS IF THERE IS SOME...SOME RIGHTNESS IN THE UNIVERSE BY WHICH IT MAY BE JUDGED.

"Yes, but people have got to believe that, or what's the point—"

MY POINT EXACTLY.

— Terry Pratchett, *Hogfather*

People say they *need* to tolerify, because otherwise they wouldn't be able to handle the intolerable world.

But that's false. Acknowledging that the world is unacceptable will not kill you; the world is *already* as unacceptable as it is. Remember the litany of Gendlin.

So face the dark world. See the intolerable.

Take up the burden that is supposed to be unbearable. Don't excuse the world, don't come up with reasons why it's OK.

Let it be not OK.

What happens then? What do you feel then?

Is there a sense of despair or helplessness? Is there a sense of hot fury or cold resolve? Is there a sense of being tiny in the face of a problem that is large?

Live *there*, in the face of the intolerable. Don't struggle to make it acceptable, just live with the bad world, while buckling down rather than buckling.

It is there, while staring the dark world in the face, that I find a deep well of intrinsic drive. It is there that my resolve and determination come to *me*, rather than me having to go hunting for them.

I find it amusing that "we need lies because we can't bear the truth" is such a common refrain, given how much of my drive stems from my response to attempting to bear the truth.

I find that it's common for people to tell themselves that they need the lies in order to bear reality. In fact, I bet that many of you can think of one thing off the top of your heads that you're intentionally tolerifying, because the truth is too scary to even consider. (I've seen at least a dozen failed relationships dragged out for months and months due to this effect.)

I say, if you want the intrinsic drive, drop the illusion. Refuse to tolerify. Face the facts that you feared you would not be able to handle. You are likely correct that they will be hard to bear, and you are likely correct that attempting to bear them will change you. But that change doesn't need to break you. It can also make you stronger, and fuel your resolve.

So see the dark world. See everything intolerable. Let the urge to tolerify it build, but don't relent. Just live there in the intolerable world, refusing to tolerate it. See whether you feel that growing, burning desire to *make the world be different*. Let parts of yourself harden. Let your resolve grow. It is here, in the face of the intolerable, that you will be able to tap into intrinsic motivation.

Choose without suffering

27 SEPTEMBER 2015

Imagine Eve, who works a service industry job. Her manager tells her, at the last minute and without warning, that she has to staff an event tomorrow in a town a few hour's drive from where she lives, and she has to wake up at 5am to get there on time.

Let's further suppose that she's on shaky footing with her manager as it is, and so she is posed with the following choice: she can either wake up at 5am tomorrow and go to work, or she can lose her job.

Imagine Eve's demeanor, upon learning this fact. It's likely dour, to say the least. She's probably grumpy and annoyed and malcontent, and she's likely to vent and complain all evening. She'll likely spend a lot of cognitive effort tolerifying the situation, convincing herself either that it's not going to be that bad to wake up early, or that her manager is a terrible person.

This is a common occurrence, I think: if you give humans the choice between bad and worse, they get *grumpy*.

When people find that none of their options cross a certain "acceptability" threshold, they get *frustrated*.

This, I think, is part of why tolerification is such a common human response to unfortunate situations. In an intolerable world, *none* of your options seem acceptable: so you tolerify, until at least one option (perhaps indignance, perhaps cynicism, perhaps doing nothing differently) passes the acceptability threshold. Only then are you able to act.

This behavior won't do, for someone living in a dark world. If you're going to live in a dark world, then it's very important to learn how to choose the best action available to you without any concern for how good it is in an absolute sense.

When given a choice between bad and worse, you need to be able to choose "bad", without qualm.

I think that one of the big reasons why people get annoyed when none of their options pass the "acceptable" threshold is they're often failing to see a hidden third alternative, and some part of them knows that this might be the case. In this setting, the frustration might even be *useful*, if it puts them in a mental state where they search more fervently for an escape hatch.

Furthermore, by acting flustered, people may well be able to draw other humans to their aid, and the additional assistance can often help make the situation better.

So frustration in the face of a choice between bad and worse may be a useful response in many situations. (At the least, it was useful enough to our ancestors.) Indeed, when you're offered the choice between bad and worse, the first thing to do is *look for a third option* and the second thing to do is *ask for help*. Find shortcuts. Try to cheat. Call in the cavalry, if you can.

But once you determine that you really have been offered a choice between bad and worse, and that there are no other options —

Then it is useful to be able to choose "bad," without suffering over it.

The first step to being able to choose the best option available without suffering, is to simply understand the distinction. Next time you find yourself feeling flustered because none of your options pass an absolute acceptability threshold, pause and reframe, and look at the *relative* acceptability of your actions instead. Simply knowing the distinction and watching out for it in real life may well be enough.

For me, another useful tool for choosing without suffering is to ask a "what if" question about a hypothetical universe, before making a choice in the real world. Let's say I'm trying to eliminate extreme poverty, and none of my actions seem

good. I might say to myself, "imagine you lived in a world where all your choices led to bad outcomes; what would you do then?" I can improve the lives of these three people, and then a million people will die of preventable disease anyway. Or I can try to alter the flow of politics, and then a million people will die of preventable disease anyway. Or I can put money into researching preventable diseases, and then a million people will die of preventable diseases anyway. No matter what I do, at least a million people will die of preventable diseases. What would I do in *that* world?

Clearly, the answer is "whatever action saves the *most* lives." I sometimes find it easier to frame my real problems as if they were hypothetical, identify the answer *there*, and then apply that to the real world.

In the hypothetical worlds where there are no third alternatives and all the actions before you, it doesn't matter that all the actions lead to bad outcomes. The best choice is still quite clear: take the action that leads to the best outcome, and take it without remorse. In the hypothetical, confident that there are no alternatives, it's quite easy to imagine selecting the least bad option from a terrible lot. In fact, it's easy to imagine doing this without any impulse to complain or struggle, but instead only a grim resolve to do the best you can in a bad situation.

So in the real world, do the same. Notice when you're measuring your options against what you think *should* happen; notice when you're measuring the futures you can attain

against the futures you *want* to attain; and treat that as a cue to reframe. Look at your actions available options again, and stop measuring them against an objective ideal, and start measuring them against each other. Look for cheats, look for third alternatives, look for ways out...

...and then, when you're done and you've considered all available options,

simply take the best action available.

Take it, without suffering, no matter how bad it is.

That is all there is to do.

Detach the grim-o-meter

05 OCTOBER 2015

I'm betting that the last three posts have given many readers an incorrect impression about my demeanor. It's easy to read those posts and conclude that I must be a grim, brooding character who goes around with his jaw set all day long.

Which is understandable, but silly. You don't need to carry a grim demeanor to draw strength from seeing the dark world. It's quite possible to deeply want the world to be different than it is, and tap into a deep well of cold resolve, and still also be curious, playful, and relaxed in turn.

This isn't a story, and we don't need to pretend to archetypes.

I've met many who are under the impression that when you realize the world is in deep trouble, you're obligated to respond by feeling more and more grim. Like a movie about a detective that's trying to save a kidnapped child: as the detective learns that the child is in more and more danger, they lock their jaw and become more and more grim and determined. Their respite comes only when the child is rescued.

That's narrative thinking, and we aren't in a narrative. You can break the trope. (In fact, I *encourage* you to break tropes

as soon as you realize that you're acting them out.)

Many people seem to have this internal grim-o-meter which measures how grim the state of the world is, and they dutifully try to keep this calibrated. When they hear that they might be failing a class, they get a bit more grim, and this helps them buckle down. When they hear that there was an earthquake in Napal, they get a little more grim, and they maybe even feel guilty if they can't feel appropriately grim for appropriately long.

I say, it's good to have a grim-o-meter, but *stop calibrating it against the state of the world*. That's a terrible plan!

I mean, look at humanity at large. People are killing each other like it's going out of style, while millions die from disease each year and civilization careens towards self-destruction.

Now look at your grim-o-meter. It has, like, seven different settings. Maybe twelve, on a good day.

That detective in the movie about the kidnapped child might be able to faithfully use a twelve-setting grim-o-meter to track the grimness of their own situation.

But the real world? The one with billions of people each with rich inner lives, and astronomical future potential hanging by a pale blue thread in Time? There's no way you can justifiably connect a twelve-setting grim-o-meter to *that*.

And what if you could? Would your grim-o-meter always be set to "maximum grimness," at least until humanity makes it through the gauntlet? That doesn't sound very fun or useful. Would you rather calibrate the grim-o-meter so that it adequately captures the normal range of variance in the human condition over your lifetime? Because then your grimness is likely to fluctuate wildly in response to events that have little relevance to your daily life (such as aggregate demand shocks in China). That *also* doesn't sound very fun or useful.

Look: that's not what your grim-o-meter is *for*. It's not supposed to be attached to the global state of the world. Feeling grim or carefree in proportion to the aggregate disparity or well-being on the planet is difficult, impractical, *and* mostly useless.

Your grim-o-meter is designed for *local* occasions. You need to get more grim (and more buckled down) as the work *immediately in front of you* gets harder, and you need to get less grim (so that you can spend time recharging and relaxing) whenever you have the affordance to recharge and relax. That's the *point* of the grimness setting.

Remember, the grim-o-meter was made for you, not you for it. What's the point of grimness? The point is to be able to buckle down when down needs buckling. And buckling down is something you need to do occasionally, if you want to get things done. But so is being curious, and being playful, and being calm. You're still a monkey, remember?

The world is dark and gritty, but that doesn't mean that you need to be dark and gritty to match. This isn't a book, and you can adopt whatever demeanor you need to adopt to get the job done.

You can look at the bad things in this world, and let cold resolve fill you — and then go on a picnic, and have a very pleasant afternoon. That would be a little weird, but you could do it! The resolve is a useful source of motivation, but you don't need to adopt a permanently grim demeanor in order to wield it. In fact, personal effectiveness is all about having the right demeanor at the right time.

I suggest a mix of playfulness, curiosity, relaxation, calm, and yes, grim determination.

I also personally recommend a healthy dose of dark humor. Everybody's dying, after all.

Simply locate yourself

11 OCTOBER 2015

Imagine I offer you the following bet: I'll roll a fair ten-sided die. If it comes up 1-9, you win a million dollars. If it comes up 0, you lose \$10,000. (If you're significantly richer or poorer than the median person, adjust the numbers up or down accordingly, such that winning is very great and losing hurts a lot, but is manageable.) Imagine that you take the bet, because those odds are ridiculously in your favor. Now imagine that I roll the die, and you watch it rolling, and rolling, and rolling, until it starts to settle, and then it settles... on 0.

Imagine the sinking feeling you might get, as you see the zero, and realize that you have to give me ten thousand dollars. Maybe you suddenly feel uncomfortable. Maybe you're unwilling to meet my gaze. Maybe you're angry, or slightly sick to your stomach. Maybe some part of you is pushing against reality, trying to deny it, willing the past to *change*.

Now imagine a second bet. This time, imagine a world that has figured out cloning and cryonics and space travel. The bet works as follows: I put you to sleep, and then I separate you into ten identical copies (none of which have any more claim to being the original than any other), and then I put them all into stasis. Your possessions are replicated ten ways, and the ten yous are put on ten ships to ten different

(already-colonized) planets. On nine of those planets, the local you will be placed in a room with blue walls, and given your possessions along with a million extra dollars. On one of those planets, the local you will be placed in a room with red walls, and will have \$10,000 removed from their possessions. Then all ten yous will be awoken. Thus, nine copies of you will gain a million dollars, and one copy of you will lose ten thousand dollars.

Imagine that you understand this procedure, and consent to it. You're put to sleep, and split into ten copies, put into stasis, sent to ten planets, and revived from stasis. You wake slowly, and haven't opened your eyes yet. You know that nine yous will wake in a blue room and find themselves rich, and one you will wake in a red room and find themselves poor, and you don't know which you you are. You open your eyes, and the walls are... red.

In one sense, you've lost exactly the same sort of bet as the first bet. But there's a very different way that you might be feeling. In the second bet, instead of feeling a sinking feeling and a desire to push against reality, you may simply nod, and say "ah, I'm the me in the red room."

Instead of treating the red walls as an unwelcome message about reality failing to go the way you wanted, you might treat them as a simple indicator of *where you ended up*. Instead of feeling despair, you may simply feel like you've figured out which you you are.

Most people seem to treat most of their observations as Bet 1 type observations: they treat their observations as information about how the universe *turned out to be*, which may be quite a bit worse than they were hoping it would turn out. They feel despair, or resistance, or victimized by an unfair universe. Part of them tries to tolerify, some part of them flinches away from facing reality, and so on.

There's another way to treat your observations. It's the Bet 2 way: treat them simply as information about *where you ended up*.

Imagine, on the one hand, Bet 1 as described above. Now imagine the same bet, but with a special die that generates ten copies of you (in different branches of the multiverse that are identical except for the number this die shows, separated such that the universes within them can never interact), such that nine of them will win a million dollars and one will lose ten thousand dollars.

Notice how someone who loses the former bet may try to push against reality, while someone who loses the latter bet has a much easier time simply saying "Huh, I guess I'm the one in the 0 branch. Such was the price for nine out of ten multiverse branches to have rich versions of me, and now I will pay it."

But these are, more or less, the same bet. Why do they feel so different?

I say, *always* treat your bets like the latter sort of bet. Stop struggling against the bad news. Treat it not as bad news about how reality went, but rather treat it as you would treat information about *where in the multiverse you ended up*. Try being a new homunculus. Look around you and figure out where you just landed, regardless of where past you thought they should have landed. Often, the place will be in worse shape than past-you was expecting, but that has little bearing on what you do next (aside from updating your current anticipations such that future-you is less wrong).

Imagine you're a new homunculus that has just landed in a branch of the multiverse where things were going poorly—maybe you recently lost social status, or made a choice that had worse effects than you expected, just before the new homunculus teleported in. This is an uncomfortable place to find yourself in! What do you do next?

Would you immediately throw a fit? What's the point of that? You just teleported into this part of the multiverse; how is struggling against the past supposed to help you? This is part of what detaching the grim-o-meter is all about: if you found yourself in a grim part of the multiverse, what would you do? Would you go around frowning and being dour all day? No? Because that sounds silly? Then there's no need to do that here!

Your observations are not messages that the world is full of terrible unfair luck. Your observations are simply indicators as to *where you are*. They're the data that you need to locate yourself.

Spoiler alert, you're currently located in a fairly precarious portion of the multiverse, where sentient beings are suffering and dying, and the future is hanging by a thread. It's worth cleaning this place up a bit, I think. But don't suffer about the poor state of affairs! Consider: if you *were* teleported to a precarious branch of the multiverse, what would you do upon arriving? Would you make sure to have a good time anyway? Would you do whatever you could to help out? Well then you're in luck! You *did* just arrive at a precarious part of the multiverse, and those are both things that you can do here.

When you get bad news, don't suffer over it. It's not unfair, it's not passing judgement, it's not a signal that everything sucks, it's not making the future worse. It's just telling you where you live.

And recently, you've ended up in the same part of the multiverse as I have. It is fairly nice, as parts of the multiverse go: it supports life, and things are better now than they were in many of the past points along our timeline. Nevertheless, it does look a bit precarious, and it sure does need some tidying up.

So, let's get to work!

Have no excuses

19 OCTOBER 2015

Except in a very few [tennis] matches, usually with world-class performers, there is a point in every match (and in some cases it's right at the beginning) when the loser decides he's going to lose. And after that, everything he does will be aimed at providing an explanation of why he will have lost. He may throw himself at the ball (so he will be able to say he's done his best against a superior opponent). He may dispute calls (so he will be able to say he's been robbed). He may swear at himself and throw his racket (so he can say it was apparent all along he wasn't in top form). His energies go not into winning but into producing an explanation, an excuse, a justification for losing.

— C. Terry Warner, *Bonds That Make Us Free*

Throughout high school and college, I noticed that many of my peers seemed like they were trying hard, but they weren't trying hard to learn content or pass classes — they were trying hard to make sure that they had good excuses and cover stories prepared for when they failed. Seeing this, I resolved that I would never excuse my own failures to myself — not even if I had a very good excuse. If you have an excuse prepared, you will be tempted to fall back on it. An excuse makes failure more acceptable, in some way. It's a license to fail.

If you really need to succeed on a task, then I suggest that you resolve to refuse to excuse your failure, in the event that you do fail. Even if the failure was understandable. Even if you failed for unfair reasons, due to things you couldn't have foreseen. Simply refuse to speak the excuse. *Understand* your errors, and learn from them, but if people demand to know why you failed, say only, "I'm sorry. I wasn't good enough." You may add "and I think I know what I did wrong, and I'll work to fix it, and I'll do better next time," but only if that's true.

Don't add anything else: if you want to play to win, you have to refuse to acknowledge excuses. If you were excused then you were helpless, and you couldn't have done better, and you can't learn to do better next time. Thus, I suggest that you become incapable of believing an excuse, lest you automatically slip into the game of making sure your failure will be explainable, rather than making sure you succeed.

"But sometimes bad luck just happens!" the one protests. We can imagine a person who took a bet that pays out \$1,000,000 nine times out of ten and costs \$10,000 otherwise. We can imagine them losing. We can imagine them saying "I should have gotten the money!", and feeling upset, and complaining that the dice went against them, and cursing the fates. We can imagine them loudly trying to make sure that everybody present knows that the bet was worth

taking, to make sure that their loss is excusable. And this person will be playing to ensure that their actions were acceptable; rather than playing to win.

I suggest, don't try to excuse bad luck. Don't call foul. Don't say that life was unfair. You're welcome to say "I'm sorry, I made a bet and I lost. I'd make the bet again, though, knowing what I did then." Then you're still *owning the choice*.

You're *owning the failure*, which is the important part. Only by owning the failure can you hope to adjust and do better next time: if you feel like you are allowed to curse the dice every time they go against you, and have your gambling excused as terrible luck by your peers ("oh they're such an unlucky person it's not their fault...") then you're never going to learn when to bet and when to abstain.

I suggest cultivating your mental habits such that it feels *bad* to check whether or not your failure will have an excuse. Refuse to have excuses. Refuse to cover your failures. Only then, without expected social protection, do you really start trying to figure out how to win.

"No really, sometimes unforeseen circumstances arise!", the one protests again. We can imagine someone who was totally planning to get their paper done on time, but who got violently ill. It's true: unforeseen circumstances can wreck your plans. But you *know* about the planning fallacy (or if you

didn't, you do now). You've been a human being for a long time. You know the background rates on illnesses, and on unforeseen circumstances in general. Why didn't you work slack into your plans? Why couldn't you see those bullets coming in advance?

If you *did* work a lot of extra slack into your plans, and you still got burned anyway by extraordinary circumstances, then as before, you are welcome to answer "I took a gamble and I lost, and I'd take the same gamble again at the same odds." You're welcome to calculate that the risk is worth the benefit, and then pay the price when your debts are called in.

If you *didn't* work in the necessary leeway, then you're allowed to say "I'm sorry, I messed up." You're allowed to add "and I learned something, and I will do better next time," *if that's true.*

Will you *actually* ever learn to beat the planning fallacy, if you allow yourself to use excuses? Will you *actually* visualize the possible failures, and take an outside view, and learn to see the bullets coming before they hit you? Or will you simply expect extenuating circumstances to arise, and feel relieved when they do, because a plausible excuse has presented itself?

I have found that it's usually in the moment when I refuse to make excuses even if I do fail, that I start really trying to win in advance.

"But people *want* excuses. They're social creatures! They want to know what happened!", the one protests.

Sometimes. Sometimes people really want you to provide them some excuse, or at least some explanation. But even here, be careful: I have noticed that my friends often help me try to excuse *myself*, for one reason or another, and I think that giving in to this pressure can be harmful.

Imagine someone who failed to exit an abusive relationship, despite three years of trauma. After they successfully exit, their friends are likely to be first in line with condolences along the lines of "they were gaslighting you" and "there wasn't anything you could have done" and "how could you have known what to do?"

They are providing excuses, and these are toxic. They rob you of your power. They rob you of your ability to say "actually, I *could* have known, if I had been thinking more clearly. I *could* have acted differently, if I had known better. And that's the *good part*, because it means that I am not a helpless victim, because it means that I can learn how to become stronger. Because it means that I cannot be trapped in that sort of situation again."

Excuses rob you of your agency. Yes, many people will try to get excuses out of you, if they perceive you as putting too much pressure on yourself. *But that pressure is precisely the*

impetus to learn and adapt, and if you can bear it, then I suggest you do.

There are situations where failing to generate excuses will cost you socially, especially if you're in the presence of people who have recently been generating excuses for themselves. If three students give thin excuses for why they didn't finish their project on time, and you say only "I'm sorry, I wasn't good enough, I think I know what I did wrong, I'll do better next time;" then they are liable to glare at you. In refusing to generate an excuse when everyone else is doing so, you violate some unspoken pact of mediocrity.

Sometimes, other people need *you* to make excuses in order to help excuse the fact that *they* are making excuses, and if you violate this norm, they find themselves faced with their own shortcomings. This can lead to some uncomfortable situations, and the best advice I can offer you for those, is that they provide a wonderful opportunity for self-signaling that you will refuse to excuse your actions even under intense social pressures.

Note, too, that in many other situations, refusing to generate excuses *gains* you lots of social status. Yes, there are places where people view refusal to generate an excuse as a violation of the solemn pact of mediocrity, but I have found that the people I can gain most from dealing with, are by and

large people who have a deep appreciation and respect for those who live up to their errors.

Excuses have you looking out to the world to explain your failure, rather than revealing the weak points in yourself. Did the unexpected happen? Then learn how to expect better next time. Were you betrayed? Learn how to build tighter social bonds, and learn how to see betrayals coming sooner next time. Did the dice turn against you? Then own up to your bet and make sure you're only making worthwhile gambles.

For many, the mantra of "find the failure in yourself, rather than in the world" will be harmful and destructive. If you are motivated primarily by guilt or shame, then seriously consider ignoring this post's advice. If you are prone to buckling instead of buckling down, then seriously consider ignoring this post's advice. If you are struggling with your self-image and your sense of self-worth, if you think some people are bad, if you flinch away from seeing the dark world, then seriously consider ignoring this post's advice. Or if "find the failure in yourself" feels bad or destructive at the moment for any other reason, then please ignore this post.

But if you are done with guilt motivation, and comfortable with the fact that we are not yet gods, and capable of detaching the grim-o-meter, then I strongly suggest that you

have no excuses. Find the flaws inside yourself. Don't tolerate them. Accept them, and plan ways to address or route around them. If you can't see what you need to do better next time, then it's going to be tough to do better next time.

This is part of the toolset that I use to replace guilt motivation: *play to win*. Don't play to excuse your loss.

You don't need to win every time — but you do need to *learn* every time.

If you find yourself trying to proclaim circumstance unfair, explaining how you could not possibly have seen this coming, then stop in your tracks. An explanation of how you couldn't possibly have seen this coming is a social device, an attempt to ensure that others still think you are OK, that they think your previous actions were acceptable. It's fine to play that social game; social games occasionally need to be played. But first, *figure out how you could have actually seen that thing coming*, next time. That's the important part.

Excuses are a social artifact, a way to ensure that you don't lose face when you fail.

But we're not here to win a social game.

Despite what all the monkey instincts might tell you, you're not playing Life in a competition against all the other monkeys.

You're playing Life with the universe, and the stakes are the entire future.

In the end, you won't be measured by how good your excuses were for all the things that didn't turn out the way you wanted.

You'll be judged only by what actually happens (as will we all).

"It's not an excuse, it's an *explanation*."

Explanations are excuses.

Don't get me wrong, it's very important to *understand* your failures. Note, though, that there's a big difference between "understanding" that your stupid knee was acting up and the sun was in your eyes and luck turned against you, and understanding that you didn't train hard enough or anticipate adverse conditions well enough.

When trying to understand your failures, it's important to figure out what *you* could have done better, rather than generating a list of reasons you never could have won. If there were unforeseen circumstances, understand why you couldn't foresee them. If your knee was acting up, learn how

to either address that next time or work it into your expectations.

(And be very wary, when figuring out what you could have done better, for hints of destructiveness and fatalism in your tone. Imagine someone who is betrayed, and shouts "well I guess now I've learned to never trust anyone ever again forever!" For all their guise of having learned, they are harming themselves. It seems to me that this self-harm has something in common with an excuse: it gives a false veneer of locating a problem internally ("I am too kind and trusting") while actually identifying the problem in the world ("the world is bad"). The right lesson to learn is likely never "become completely unable to trust," it is likely more along the lines of "learn how to build tighter friendships" or "learn how to read humans better." It can be often useful to check the advice you just gave yourself to see whether it was obviously destructive, before following it.)

The point of understanding your failure is to learn how to act better next time, and I recommend that you understand your failures whenever possible. But don't explain them away, and don't excuse them.

If you want to succeed, stop generating reasons why you never could have won, and play to win.

Come to your terms

26 OCTOBER 2015

Once, a friend of mine decided to make a drastic career change by teaching themselves a bunch of new skills from scratch, (with occasional assistance from me). They ran into occasional difficulties along the way, one of them being that they could not consider the possibility of failure without feeling fear.

The possibility of failing — of investing months in the effort, with nothing to show for it, and then having nowhere left to turn — weighed heavily on them. It wore them down, it caused great stress, it induced panic attacks. Sometimes, they were incapacitated to the point that they could hardly think.

This wasn't completely unreasonable: they had no safety net and no margin for error, and they had good reasons to fear for their personal safety in the event of failure. The problem was not that their fears were irrational. The problem was that they *couldn't think them*.

I encouraged them to try facing their fears, and they did, but they found that coming to terms with the worst was impossible. They buckled, rather than buckling down. So consider that a content note: the exercise I describe in this post may not be possible or helpful for you.

But it has been very helpful for me, and I continue to think that if my friend had been able to truly come to terms with the worst case scenario they had in mind, to imagine it in detail and accept it as a possibility, then they would have had a much easier time managing that stress.

So here's my advice: Think the unthinkable. Consider that which is painful to consider. Figure out what, exactly, is at stake. Weigh the consequences. Come to terms with them.

I'm *not* suggesting that you convince yourself the worst case actually wouldn't be that bad. I'm *not* suggesting that you tell yourself a story about how you could handle the worst. I'm saying, *come to terms with what could happen*. Imagine the worst case, in detail; learn to weigh it on your scales; accept that if you fail things could go very poorly; and then maybe those bad outcomes will loosen their grip on you.

If you ever notice yourself following the same pattern as my friend — if you ever notice an outcome so *terrible* that you *can't even consider it without panicking*, then I suggest that you pause, take a deep breath, and consider that outcome.

Visualize it in full detail. Don't need to excuse it. Don't tell yourself it wouldn't be your fault. Don't tell yourself it would be fine. Don't make up a story about how you'd handle it successfully. Just *imagine the worst*.

People close to you might get hurt. You could die. Lots of people could die. If bad outcomes are in the possibility space, internalize that *now*. Come to terms with that terrible fact as soon as you can. You want to get into a mental state where if the bad outcome comes to pass, you will only nod your head and say "I knew this card was in the deck, and I knew the odds, and I would make the same bets again, given the same opportunities." If you need to panic, panic once and get it over with. Otherwise, fear will strike again every time the bad outcome moves a millimeter closer, and that fear may debilitate you or incapacitate you at a crucial moment.

It's the thoughts you can't think that control you most, and it's the outcomes you can't consider that weigh heaviest on your scales.

An outcome that you can't consider without panicking — failing a class, crashing a car, destroying the family business — weighs infinitely heavily in your considerations. You can't even *think in the direction* of allowing the bad thing to happen, without encountering a cloying fear that steers your thoughts away. It is as if the bad outcome has infinite weight on your scales. Your thoughts become censored; you become unable to rationally weigh the risks and gambles.

Once you've fully considered the terrible outcome, its weight on your scales becomes finite. It may remain heavy, it may be the overriding concern in your life, it may still dominate your actions. But once you've weighed the outcome, it can only dominate your actions if you decide that that's rational, after weighing the possibilities and tradeoffs.

And maybe, after seriously considering the terrible outcome, it will *stop* dominating your actions. Maybe it will seem less terrifying once you drag it into the light. Maybe it will seem more manageable after you consider how you'd *actually* manage it. Maybe you'll notice that the outcome wasn't as terrifying as it seemed at a distance.

In my line of work, I occasionally find myself in conversations with powerful people in situations where the outcome of the conversation has some small chance of dramatically affecting the future of humanity and all earth-originating life. The first time I found myself in one of these conversations, I was fairly shaken afterwards.

During the conversation, there was a sensation not unlike the one I got as a young driver on the interstate, realizing that I could, with a trivial twist of my hands, steer the car into oncoming traffic. After the conversation, there was a fear that had a lingering effect on my thoughts. I was jumplier. My actions were less considered. I was flustered.

A friend of mine (who had been through this before) noticed, and asked me whether I'd ever really come to terms with the fact that I just might set into motion a chain of events that leads to the end of the world.

I said no.

But, amusingly enough, I *had* spent time coming to terms with the fact that I might ruin my *own* life, and die old and bitter and unaccomplished.

I remember *that* ritual quite well: I was 18 at the time, and I had (a few years prior) decided to dedicate my life to changing the world in a big way. I was aware of the odds stacked against me, and I was aware of the success rates, and I was fully aware of the fact that, in all likelihood, I was going to fail, and my ideas were going to prove defunct, and my plans were never going to come to fruition.

I imagined that I could well end up a bitter old man, bemoaning plans that should have worked, to people who only scoffed. Now, I also planned *not* to become that bitter old man — but in those days, I wasn't yet sure how much control I'd gain over my own mind, and I saw lots of bitter old men around me. I was wary that my plans to avoid bitterness would *also* fail, and I'd become bitter and old despite my best efforts.

As I attempted to get a few different schemes started, and I noticed myself holding back a part of myself, in case my

plan was just too crazy, in case I would be too harshly judged for trying. Introspecting, I concluded that I was resisting because I was afraid of ruining my own life.

So, knowing that a chance of becoming a bitter old man with little money, no respect, and nothing to show for it was one of the prices I might need to pay, I decided to come to terms with that fact once and for all. I spent time imagining this outcome in detail. I didn't try to explain it to myself, I didn't try to tell myself stories about how I'd avoid the outcome, I didn't try to tell myself it would be OK. I just pictured what would happen, considered the cost, weighed the price, and deemed the possibility of failure a price worth paying.

I didn't convince myself it would be *OK*, but I did decide that a chance of a not-OK outcome was a price worth paying.

And then those fears released their grip on me.

So when I was shaken by that high-stakes conversation, and my friend asked whether I had ever come to terms with the fact that I might set into motion a chain of events that leads to the end of the world, I laughed, and said no, but that I had done something similar, and that I knew the ritual. It was a simple task to repeat it, to go through the same mental motions but with larger stakes in mind.

Now, I'm a bit harder to shake.

(I'm sure this was not the only way I could have gotten used to high-stakes conversations, and undoubtably exposure alone would have eventually had a similar effect. Nevertheless, this mental ritual sped up the process quite a bit, and I'm under the impression that it's helped me think more clearly when making high-stakes decisions across the board.)

So, I say, if there are outcomes before you that seem unthinkablely terrible, then come to your terms with them. Don't explain them, don't excuse them, don't tolerify them, simply *visualize* them, and come to terms with the prices that you might need to pay.

You may be hurt. People you love may be hurt. You might break things that can't be fixed. The world might actually end. The point is not to convince yourself that you could handle the worst if it came, because maybe you won't be able to handle it. The point is simply to *know what the worst case looks like*.

If you know what it looks like, you can do your best to avoid it. The outcomes you can't consider control your actions. If you want to avoid the worst outcomes, you need to be able to weigh *all* the outcomes on the scales.

(For those of you who are wondering, fear not; my friend ultimately succeeded in switching careers.)

Transmute guilt into resolve

01 NOVEMBER 2015

A friend of mine came to me and said that he cares about his immediate friends, and he cares about humanity in the abstract, but he has trouble caring for most people. They seemed too shallow, too bitter, too spiteful to be worth an effort.

He'd been a sixth grade teacher, so I asked, "What about when they were eleven? Were they worth an effort then?"

"Yes," he answered adamantly. Or, most could still be salvaged at eleven, though there are some that you'd need to get to even earlier, if you wanted to save them from the shallowness and the learned helplessness and the death of curiosity.

"So then we live in a world that mishandles its youth, that turns them from bright children full of potential into empty shells. What are your feelings about that process, and the people subjected to it?"

His answer, more or less, was "A bit of anger, a bit of nothing-I-can-do-about-it, and a bit of victim-blaming, which I don't endorse."

Those last two emotions are very interesting: Why assure yourself that there is nothing you can do about the problem, if you don't care about the people who are harmed? Why assure yourself that it is their fault, if you stop caring about people once they are lost?

These seem like defense mechanisms, to me — defense mechanisms my friend generated unconsciously, because it was too painful look at bitter shallow adults and see lost mistreated eleven-year-olds.

Most of the time, if something is hurting you, I recommend making it stop. There is one exception, though.

Imagine walking past a beggar on the street. They're dirty and downtrodden; weathered but not much older than you. They ask you for change as you pass by.

This causes a certain type of pain in people — enough pain that most people develop some sort of coping mechanism. Some people pretend they didn't see or hear the beggar. Some give an apology, some make up an excuse about not having any money. Some shove their hands in their pockets and drag out some spare change, so that they may discharge their moral duty.

Other people cope with cynicism or bitterness — the sight of a beggar reminds them of the failings of the hated out group, the people who voted for the Wrong Political Party in the local elections. Still others cope with a wave of guilt, shorting out the pain, because the guilt seems easier to bear.

My suggestion, this week, is notice that impulse. Notice the impulse to look away, to ignore, to make an excuse, to assure yourself that there's nothing you can do, to blame the hated out-group.

Resist the impulse, and acknowledge the pain. Sit with the pain. Don't excuse yourself from it, don't tell yourself a story about how there's nothing you can do or about how your attention and effort can be better spent helping other people elsewhere. That may be true, but it's another coping mechanism, and it also shorts out the pain.

Instead, I suggest sitting with the pain, and transmuting it into resolve.

There are many people for whom guilt is the right response, when ignoring a beggar. If you're not doing anything to leave the world nicer than it was when you found it, if you're not doing anything to help your fellow human beings, if you're not doing anything to shape the grand story of Humanity as it plays out all around you, and if you *want* to

be helping, then guilt is a healthy reminder that you've betrayed some part of yourself.

This is why my "replacing guilt" series began by addressing the listless guilt, all those months ago. Sometimes, guilt is a reminder that you're not doing what you think is right, and those reminders can be valuable.

But most of the guilt-motivated people I know don't match that pattern. Many of them are dedicating their lives to making the world a better place, and they can do far more good by focusing their attention on their work and their health than they can by worrying over one beggar in the street, or over a thousand starving families that they can do nothing to save. They berate themselves for not needing less rest, for not being able to do the psychologically impossible, for not being as smart or as productive or as wealthy or as kind as those around them.

I say: Yes, the beggar suffers. Yes, a thousand families starve. The world is hurting.

And yes, there are others who are doing more than you to help. Some are smarter, some are more productive, some were born wealthier, some are kinder, some are less psychologically fragile, some have a stronger will.

But none of these are reasons for guilt. Guilt was made for us, not us for it. Guilt is useful only insofar as it helps you wrest yourself from the wrong path. If you're already walk-

ing the path you want to walk, if you're working on becoming kinder, or more generous, or psychologically stronger, or wealthier, or smarter, if you're already moving as fast as you can given your current constraints, then the fact that the world is still hurting and you aren't strong enough to fix things yet is no reason for guilt.

Rather, it's a reason for *anger*, at a world where nobody is evil but everything is broken. It's a reason for *resolve*, to push yourself as hard as is healthy and sustainable but no harder.

But it is not a reason for guilt, once you are doing what you can, in full light of the fact that you are still only mortal.

There are dozens of opportunities to transmute guilt, or awkwardness, or not-my-problem into resolve, each day.

Notice the disabused middle-aged woman who has to sacrifice a part of her soul working a job at Starbucks in order to earn her right to survive. See the madman yelling across the street, while everyone else reflexively struggles to ignore or unsee him. See a morbidly obese person avoid the stares of onlookers as they struggle with self-loathing in a civilization that filled its cheapest foods with poisons that ravage bodies.

Some people ignore these painful parts of the world. Others try to unsee them. Others try to distance themselves, by poking fun at those who are deemed "pathetic."

I suggest seeing them, and remembering. Remember that there may come a time when humanity will move the very stars to ensure that no mind suffers as much as a first-world beggar does today. Remember that, beneath all the mental callouses that allow you to write fellow human beings off as unsalvageable, the reason you won't help them is not because they aren't worth helping, but because there are too many other things that need doing first.

So notice your impulse towards guilt. Notice your impulse to ignore. Notice your impulse to distance yourself from people you don't want to acknowledge. Notice your impulse to assure yourself that it's not your fault, that there's nothing you can do, that you can't help them because it's cheaper to help other people suffering just as much abroad.

Then stop following those impulses. See the dark world. Acknowledge the pain, and remind yourself that we live in a universe *worth changing*.

Remind yourself that you're a part of the grand human story, and that when our children's children's children hear about the amount of suffering we had to pass over in combat of greater evils, they will shed tears.

The count of people we have to leave behind can be a persistent source of pain. But don't let it be a persistent source of *guilt*. Instead, let it be a reminder that the universe is vast and uncaring, and that our job here is unfinished.

The best you can

10 NOVEMBER 2015

In fiction, protagonists narrow their focus until the difference between success and failure on their specific task seems like the difference between victory and defeat. Batman attempts to solve the mystery while ensuring that nobody dies; meanwhile, children in Africa suffer from Malaria. The crew in The Martian spends billions of dollars worth of capital to save one man; capital that could have been spent curing diseases.

Real people run a risk of duplicating this error, if they try to find the very best action available.

It's easy to paralyze yourself if you try to do the "right thing." There's always more uncertainty to be had. There's always more information you could gather. It's hard to become confident that you're doing the right thing. This can lead to paralysis, and persistent inaction.

It's much easier, I think, to stop asking "is this action the right action to take?" and instead ask "what's the best action I can identify at the moment?"

Sometimes, the best action you can identify is "search for more alternatives." Sometimes, it's "study more" or "learn

more." Sometimes, it's a specific action. The nice thing is that "what's the best action I can find in the next five minutes?" always has a concrete answer. If you search for that, instead, you won't get paralyzed.

Spoiler alert: you can't find the "actually best" action. Insofar as there *is* an "actually best" sequence of motor outputs your brain could produce, it's a mad convoluted dance that leverages butterfly effects to reforge the world overnight. You're not going to find the "best action." And the best action you *can* find is exactly what it sounds like — the best action you're able to find.

You never have enough information to make a fully informed choice. You never have enough time to consider all the possibilities, or weigh all the evidence. You are always biased; your brain is compromised. The problem before you is too hard, and no matter what you do, a billion more people are going to die.

No matter what gambles you take, no matter how risky or cautious you are, you're trading off some possible futures against other ones. You can't save them all.

All you can do is look at your actions, and take the best one you can find.

It's easy for humans to zoom in to the game we think we're playing, and try to win *completely*, to solve the mystery without letting *anyone* die.

It's easier to remember to pick the best action you can find, rather than striving to do the "right thing," if you remember that people have already died; that the threshold has already been crossed. That we're not playing for a "total victory" any more, that we've already missed our chance at a "perfect score."

This is a battle we've already lost.

A hundred billion people have already died.

Rome fell. The barbarian hordes flooded through its gates. There were a thousand years of darkness.

We've already missed our shot at a total victory. Now we're just building the best future we can.

So don't get paralyzed looking for the right thing to do. Just find the best action you *can* find, and do that.

Dark, Not Colorless

16 NOVEMBER 2015

The last arc of posts has been about how to handle a dour universe. Become unable to despair, learn to see the darkness rather than flinching from it, learn to choose between bad and worse without suffering. Learn to live in a grim world without becoming grim yourself, learn to hear bad news without suffering, and stop needing to know your actions were acceptable. Come to terms with the fact you may lose, use the darkness as a source of fuel, and let go of dreams of total victory. These are the tools I use to tap into intrinsic motivation, in a precarious world where the problems are larger than I am.

Where others see a hurting world and feel guilty for not doing enough to help it, I see a hurting world and feed my own resolve. Instead of feeling guilty for not working until I drop, I recognize the psychological impossibility and resolve to do everything I can *within* my mortal constraints. For me, at least, this internal drive is more robust and reliable than guilt motivation.

This brings us to the end of the penultimate arc of the "replacing guilt" series of posts, which I began many months ago, and takes us into the final arc. The first arc was about addressing the listless guilt that comes from ignoring a part of yourself that wants to be doing something more. The second arc was about eliminating the feeling of obligation, and fighting for something you care about *only because you care*

about it. The third arc was about coming to terms with your limitations and learning to optimize *within* them, rather than feeling guilty because of them. This post concludes the fourth arc, about living in a dark universe and tapping into resolve instead of guilt.

The fifth and final arc is about what you do next. Once you've removed guilt and replaced it with intrinsic drive — both cold resolve and hot desire to *make the future bright* — what do you do next? What thought patterns allow one to turn these feelings into *actions*, rather than feelings of frustration and impotence?

I'll explore some of my answers to those questions in the coming handful of posts. But before then, I have one reminder I'd like to pass along.

Among all this talk of coming to terms with a dark and dour world, I ask you to remember that the world is *dark*, but it is not *colorless*.

I have seen many a friend attempt to see the dark world and then despair (for they are too small and the problems too large), and then confuse their sense of hopelessness with a sense of *meaninglessness*.

(The reasoning goes: "If the universe is so large, how can I matter? If the world is in such deep trouble, how can I make a difference? If all this were true, nothing would matter.")

So consider this a gentle reminder that a dark world is not a *lost* world. It is not a grey world, where everything is dead and there is nothing we can do. It is not a cold empty universe, from which nothing can be built. It is simply a *damaged* world, a *hurting* world, that is intolerable precisely because it could be so much better.

If you gazed upon a worthless universe, all cold and dead, the sight would likely not fill you with despair — because while there is no light, while there are no happy sapients living full lives, there is also no darkness: that universe is empty and dull. If you gaze upon our universe and despair, then, then that can only be because there is so much that is not right, but *could be*.

While our world is dark, it is still filled with color, and indeed many spots of light and even brilliance. Children laugh. Lovers meet. Right now, someone is just understanding one of the deep secrets of how the universe works for the first time, and their mind is filling with awe. Right now, someone is building a close friendship for the first time in a decade. Every day bears witness to a billion acts of love and kindness. This world is dark, yes — 150,000 people die every day — but it is not *lost*.

So don't let despair or hopelessness weigh you down. Instead, let them be a reminder: those are feelings you can only get from something worth saving. There are things here that are worth fighting for. If you begin to despair, then let that feeling be a reminder of what could be, and let everything that this world *isn't* be your fuel.

The world may be dark, but it's not colorless.

Stop trying to try and try

22 NOVEMBER 2015

Imagine a graduate student of mathematics as they interact with a professor, attempting to understand something in the professor's area of expertise. They're working hard to wrap their head around the basic formalism. They're in "learning mode" — they're a student in the presence of a master, expected to try to understand the math but not necessarily expected to succeed. Even if they're doing quite well, they're still reminded of how math is big and they are small; they encounter wide swaths of knowledge that they do not yet have, and often feel humbled. They use their tools tentatively, aware that they may be using them inappropriately, and wonder when they'll become a master.

Now imagine the same student tutoring an undergraduate in linear algebra, a topic they know quite well. Now they're in "teaching mode." Math is still large; the graduate student is still small; but the context is very different. The focus is no longer drawn repeatedly to all the things they don't know yet — but it's not drawn to all the things they *do* know, either. The focus simply isn't on them, or their abilities. It's on the undergrad. The grad student, in the back of their mind is not thinking "wow math is so large I don't know enough yet I'm not sure I'll ever know enough", and they're *also* not thinking "wow I know so much this is great!" — they're

thinking about how to help the undergrad understand a complex concept.

I think that many people who are in learning mode expect that mastery feels like learning mode, except that instead of feeling like they know very little, they feel like they know quite a bit. By contrast, I think mastery looks much more like teaching mode — it looks like someone operating in a context where their knowledge and their skills are not the focus, but are just unconscious assumptions in the background.

Consider the grad student in teaching mode. Their approach to answering questions in teaching mode is very different than their approach in learning mode. That's not because all the questions they encounter in teaching-mode are simple — if you've ever been a tutor you know that tutors are commonly asked questions they can't answer in the moment. Rather, they approach questions differently because context is different. When the professor asks them questions, they're Expected To Do Their Best; when the undergrad asks them questions, they're just expected to answer.

In the first case, they're expected to try; in the second case, they're assumed capable, an assumption that fades into the background.

I describe this model because I think there is an analog of these two modes when it comes to "trying" to achieve any task — and today, I'm going to talk about trying.

My advice is simple: notice when you're expected to try, and consider reframing. It's much harder to solve a problem when you're Expected To Do Your Best than it is to solve a problem when you're immersed in various subtasks, with the assumption that you're going to solve the problem buried implicitly and unconsciously in the context.

For example, consider exercise. Many people find it much easier to exercise in a context where the exercise is in the background rather than the foreground. Imagine someone who plays recreational soccer, sprinting up and down the soccer field up till the brink of exhaustion. Now imagine them not playing soccer, but just trying to sprint up and down the field up to the brink of exhaustion. They probably push themselves a lot less in the latter case. If "sprint up and down the field a lot" is the main goal, then at each possible stopping point, part of them starts trying to convince the rest that they've exercised enough for the day, and they must spend willpower to continue. In a soccer match, by contrast, the focus is elsewhere. They aren't constantly ping-pong themselves with explanations of how they've done enough sprinting for today. They aren't generating reasons why it's OK to stop here. They're trying to *score a goal*. Getting exercise is a background assumption, not a conscious choice.

Switching contexts such that your actual goal is in the background rather than the foreground — such that pursuing it

is not a conscious choice that you need to reaffirm every time you find a stopping point — is a powerful tool.

This is not novel advice, of course, but it is perhaps a generalization over a few different common types of advice. As another example, consider two people trying to become friends on purpose (perhaps for romantic reasons). I conjecture that it's much harder for people to become friends on purpose than to become friends accidentally while pursuing some other endeavor.

If they're trying to become friends on purpose, then they're constantly asking themselves, "are we friends yet?", and like the grad student asking themselves "do I understand all of mathematics yet?", the answer will never be an unresounding "yes". They would do better to switch to a context where they're not constantly checking whether they're friends yet, and are instead just *being friends*.

This model suggests that it's much more effective to alter the context such that neither party is regularly checking the depth of the friendship, but such that a strengthening bond is the implicit background assumption. (This suggests one reason why online dating feels more socially awkward than going on a date with someone you met in some other context.)

For a third (somewhat silly) example, imagine that I woke up one morning and said "I'll try to run MIRI well today." (MIRI, the [Machine Intelligence Research Institute](#), is an or-

ganization I run.) If I did this, I'd be in trouble. How does one run a research institute? What would my next actions be? Things that seem plausibly like what people-who-run-institutes-well would do? Things that seem defensible to the board of directors? I have no idea how to "try to run MIRI."

Now imagine instead that I woke up and said "I'm going to glance at my MIRI priority list, update it if today happens to be Monday, and then identify MIRI's biggest bottleneck and work on it directly." Now I'm in business, and might do something useful with my day.

Notice the difference. In the second case, I'm not asking myself whether I can run a research institute. I'm not asking myself *how* to run a research institute (though "study the strategies of people who ran other successful institutes" does occasionally get to the top of my priority list). I'm assuming myself capable — not consciously, but as a background assumption. I'm not assuming *success* — either I can run a research institute or I can't, the jury's still out on that one — but my capability is not the focus of my attention. I fret about much more practical things, like the tone to strike in a fundraiser announcement post, or how to prioritize paper-writing versus novel research. I'm never "trying to run MIRI;" I'm just working on the next top-priority task.

This, I think, is one of the main distinctions between "trying to try" and "actually trying".

Trying to try to run MIRI would *feel like* just trying to run MIRI — it would feel like thinking about what it takes to run an institute and reading books about running institutes and worrying whether the board of directors thought I was doing a good job and so on. From the inside, I'd probably think I was trying very hard to actually run an institute.

Actually trying to run MIRI feels very different from the inside. It doesn't feel like trying to make an institute run, it feels like trying to get all the most important emails handled while not letting administrative duties suck up my day. It feels like struggling to prioritize three important tasks that can't all be done. Actually trying to run MIRI does not feel like trying to run MIRI, it feels like a never-ending stream of smaller tasks.

I think many people imagine the difference between trying to try and actually trying involves something like Additional Effort or Additional Willpower. It's easy to imagine someone trying to try to (say) cure aging. Maybe they flounder around a bit and talk about how they want to join a biology startup, or start a biology startup, or get a biology degree, all while really deeply wanting to find some way to cure aging. It's also easy to imagine that the person "actually trying" to cure aging is doing something similar, but with more determination and a bit of pixie dust that makes things work out. The actually-tryer does the same things, but for them, the

startup works through dint of sheer willpower; or they get a biology degree while winning so many accolades that they get to set up their own laboratory.

This isn't how I imagine "actually trying." It's not trying-to-try with extra gusto. Actually trying looks like solving small subproblems, with the more ambitious target no longer the focus of attention, but rather a background task. Actually trying to cure aging doesn't look like a person getting a biology degree with *especially grim determination*, it looks like Aubrey de Grey wading through a mountain of mundane tasks while scraping together enough money to keep SENS running.

(SENS is currently fundraising, by the way.)

If you want to solve hard problems, stop trying to solve the hard problem directly. Change the context such that that's a background assumption: all your actions are going to be pointed roughly in the direction of solving-the-problem; what next? What's the next thing that needs doing? Work on *that*.

This is perhaps simple advice, but I myself have found it useful in the past. Many years ago, when I was in high school, a friend of mine came back from college having joined a fencing team. He wanted to show me some of the basics, so he tossed me a sabre, and we had at each other. We crossed swords a few times, and he said something

along the lines of "Nate, the goal isn't to hit my sword, the goal is to hit *me*."

It's an obvious thought, a simple thought, and a thought I had failed to think. After that, I wasn't trying to fence, I was trying to *hit him*.

Or consider the scene in The Matrix where Morpheus tells Neo "Come on, stop trying to hit me and hit me!" — at which point Neo's blows grow more intense, until he gets a fist past Morpheus' defenses. I suspect that many people watching that scene imagine Neo turning on the "try harder," pouring more effort into his punches and harnessing his frustration. When I watch the scene, I imagine a little bit of that, but mostly I imagine a similar mental shift to my "don't bang swords together; strike the enemy" mental shift — I imagine Neo had mostly been throwing out a bunch of martial arts moves that had recently been uploaded into his brain, in attempts to see if any of them worked against Morpheus, and that when Morpheus said "stop trying to hit me and hit me" Neo thought "oh yeah, I'm not supposed to be deploying martial arts moves and monitoring whether I'm fighting well enough, I'm supposed to be *hitting Morpheus*," and that his brain shifted from the "expected to try" gear to the "competence assumed" gear.

I think many people solve problems more effectively in the "competence assumed" gear," when they're not fretting about whether they can solve problems because they're too busy fretting about very specific actionable subproblems.

So if you want to tackle big problems, my advice is this: If you ever find yourself saying "I'm currently trying to solve [problem]", be wary. This is doubly true if you're Expected To Do Your Best.

If you find yourself saying "well I'm trying to solve aging, but it's a big problem, so I'll likely fail," then stop in your tracks. Not because of the underconfidence — aging *is* a big problem and you *will* likely fail to solve it — but because you're sprinting up and down the field when you'd be better off playing a game of soccer.

If you approach a big problem with Intent To Try, then at every plausible stopping point part of you will be trying to convince you that you've done enough. And thus, at every plausible stopping point, you'll need to spend willpower to continue. Find a soccer game instead — some way to focus your attention on useful object-level tasks, with the pursuit of the important goal turned into an implicit unconscious background assumption so deeply ingrained in your plan that you can hardly see it any more.

As for how you make or find the soccer games, that's a discussion for another day. For now, my generic suggestion is to (a) generalize from the above examples and (b) imagine someone who's "playing soccer" with respect to your task or problem, and ask yourself what they might be doing. The

key is to make the pursuit of your goal implicit, and spend your focus on the subproblems.

There is no try

29 NOVEMBER 2015



Ok, so "try" is actually a pretty useful concept; there's a reason we have a very short word for it in the English language. Nevertheless, I have found it quite useful to occasionally spend a few weeks refusing to use the word "try" or any of its synonyms, at least when talking about myself, and especially when thinking about myself to myself.

This is a quick and easy way to put success in the background, as discussed last week. For example, compare these two responses to "what are you doing?"

I'm trying to solve this math problem.

versus

I'm pursuing a promising line of inquiry on this math problem. If it doesn't lead anywhere, I have two others to pursue next. If

all three are fruitless, I'll ask for help.

For the first person, "failure" is either first or second on the list of things they expect to happen next: they're trying to solve the problem, and either they'll solve it, or they'll fail. If they fail, they can say "well, I tried", and move on. And because failing and moving on is such a prominent option, they must struggle against it each time they pause; they are like the person trying to sprint up and down a soccer field as much as they can, rather than the person playing soccer.

The second person, who does not have 'try' in their vocabulary, is forced to say what specific actions they are actually taking — and now, failure on the entire problem is much further down on the list of possible outcomes. Failure at this particular line of approach just drops them into the next line of approach. They're more like the person playing the soccer game, getting exercise ("trying to solve the problem") without that idea explicit in their mind. This sort of mindset, I find, is often helpful.

Imagine that I'm in the middle of flossing my teeth, when someone knocks on the door and asks what I'm doing. I wouldn't answer "trying to floss," I'd just answer "flossing" — unless I had been interrupted so many times that I was beginning to doubt my ability to complete the task. When we're sure of our ability to complete a task, we don't de-

scribe ourselves as "trying", we just *do it*. I don't get up every morning and try to dress myself, I just get up and dress myself.

Whenever you can honestly say that you are *doing*, rather than *trying*, then I suggest you do so — but often this is only honestly possible when you're quite confident in your own ability to succeed.

(Some self-help books and professionals advocate *always* saying that you are "doing" rather than "trying," but this often seems dishonest to me: when I'm trying to win a race, and I'm currently in tenth place, and you ask me what I'm doing, I have a hard time saying "winning a race" with a straight face.)

When removing 'try' and its synonyms from your vocabulary, you may find that you can't honestly say you're "solving a math problem," because you have no idea whether you'll succeed. And saying you're "working on a math problem" is only slightly better; it's mostly just using "working" as a synonym for "trying."

In these cases, if you want to remove the word 'try', I suggest not finding a near synonym, but increasing the granularity of your descriptions. Don't say "I'm trying to solve this math problem," say "I'm transforming the problem into a programming problem so I can see it from a different angle", or "I'm gameifying the problem so that my intuitions can get a better handle on it," or "I'm producing random algebraic

manipulations of this equation in desperate hope that one of them happens to be the answer," or "I'm staring at the problem waiting for my gut to say something for enough time to pass that I can give up without losing face." Describe what you're doing on the level of granularity where at each step you describe, it would be silly to say you were "trying" at that step, in the same way it would be silly to say that you wake up and try to dress yourself — describe your actions on a level of granularity where each step is definitely something you're *doing*, rather than *trying*.

Often, when I get down to the level of granularity where I'm doing rather than trying, I find that I'm doing something pretty silly — as in, I'll start out by saying "I'm trying to write the opening paragraph of this paper", and then I'll notice the word 'trying', and I'll introspect a bit and rephrase a bit and I'll eventually figure out that I was doing was "sitting in front of a screen holding the subject of the paper in my head waiting for my gut to figure out what to write" or something along those lines. With that description given, it's much easier for me to say "aha, my gut doesn't know what to write first; I'll make an outline on a whiteboard or some other place that feels non-committal."

"Try" is a useful word, but saying that you're "trying" to do something is a *high level description*, and it can often hide some very silly behaviors, like "sitting around staring at the problem waiting for enough time to pass that I can give up without losing face."

Occasionally, I tell people who come to me for advice that "try" is a fine and useful word, but saying that you're trying is something that *other* people get to say about *you*, not a thing that you get to say about yourself. *Others* get to say "they're trying to save that person's life," but *you* only get to say "I'm performing chest compressions while thinking back to remember my CPR training."

This isn't always the most useful advice; there is, after all, a reason why 'try' is such a short word. There are many situations where it's quite useful to communicate something like "I'm trying to prove this lemma; can you help?", and there are many other cases where it can be useful to use the word 'try' even when thinking about yourself to yourself. Nevertheless, there is a helpful sentiment buried in the above advice, and I have often found it useful to cash out my "try"s.

As such, I recommend, as an exercise, spending a few weeks refusing to use the word 'try'. This can help you train yourself to notice the difference between "trying" as in taking intelligent, concrete, fruitful actions; versus "trying" as in waiting for enough time to pass that you can safely say "well I tried."

This probably isn't what Yoda *actually* meant by "there is no try." Nevertheless, I like to imagine Luke nodding and saying "Oh, right; there is no try. I will close my eyes, relax, let the force flow through me, focus my mind, concentrate on a mental image of my X-wing, and then will it to lift, with no regard for its actual mass." That's the level of granularity at which you can tell whether a cashed-out "try" is a pre-emptive excuse for failure or an intelligent attempt to succeed.



Obvious advice

06 DECEMBER 2015

This is a common scene at the MIRI offices: I have a decision to make, like what sort of winter fundraiser to run. Before making any choices, I take a few minutes to write down all the obvious things to do before making the decision: spend five minutes brainstorming options before weighting any pros or cons; talk to people who have run different types of fundraisers in similar situations; and so on. I can usually generate a handful of obvious things to do before making my decision. I write those things down, and then I describe my decision to one of my advisors and see if they have any advice. They say "only the obvious," and then rattle off five more obvious things I hadn't thought of, all of them useful.

Sometimes, I wonder how successful a person would be if they just did all the obvious things in pursuit of their goals.

So with that in mind, allow me to offer some quite obvious pieces of advice, which have proven very useful for me:

Before carrying out any plan, *actually do the obvious things*.

When you're about to make a big decision, pause, and ask yourself what obvious things a reasonable person would do before making this sort of decision. Would they spend a full five minutes (by the clock) brainstorming alternative options before settling on a decision? Would they consult with

friends and advisors? Would they do some particular type of research?

Then, *actually do the obvious things.*

A corollary to this advice is to also occasionally consider *not doing things the wrong way*. Imagine someone who's recently failed at an endeavor that was important to them. They're fraught with despair, and you attempt to console them by saying "well, at least you learned something." They snap back, "yeah, I learned never to try hard things ever again!"

This may be just an emotional outburst, yes, but if they act upon this outburst — and withdraw, and become less curious, and become more bitter — then they are in solid need of the above corollary. In fact, the middle of an emotional outburst is one of the *best times* to use the corollary. I have often myself found it useful, mid-hasty-decision, to pause, reflect, and ask myself "wait... is this a *terrible plan?*"

(And then, if the answer is yes, I don't carry out the plan — a crucial step.)

Both pieces of advice above — "do the obvious preparation", and "don't execute bad plans" — each get a lot more useful as you expand your notions of "obvious preparation" and "bad plan". In fact, quite a bit of rationalist-style advice is

about expanding your notion of "obvious thing" and "bad plan." Thus, this advice gets much more helpful if you make sure to do the obvious things.

(Not all the rationalist advice is of this form, of course; many of the most important rationalist skills are cognitive operations that happen in five seconds or less. One example of a five-second-level skill is the skill of encountering a new problem and *reflexively starting to list obvious preparations* or noticing an emotional outburst and *reflexively taking a step back and checking whether your current plan is terrible*. More often than not, one of the goals of these blog posts is to install a five-second cognitive reflex of *deciding to apply a tool* by describing the tool itself. But I digress.)

For example, the cognitive reflex of "enumerate obvious preparations" becomes much more useful once you have concepts like "brainstorm options before weighing pros and cons" and "set a five minute timer and actually think about the problem for the whole five minutes" and "consider the opportunity costs." And the cognitive reflex of "check whether your current plan is terrible" becomes more useful as you add concepts like "rationalizing" and "blindly acting out a social role" and so on.

So this week's advice is obvious advice, but useful nonetheless: find a way to gain a reflex to actually do all the obvious preparation, before undertaking a new task or making a big decision.

It's surprising how often the advice that I give people who come to me asking for advice cashes out to some form of "well, have you considered doing the obvious thing?"

For example, when someone comes to me and says "help, I have a talk I have to give and I'm going to be terribly nervous and I dread it, what do I do?" it's often surprisingly helpful for me to ask, "well, what sort of things would make you less nervous?" Or someone comes to me and says "I find myself just playing video games all day, how do I stop myself?", I first ask, "have you considered what sorts of things you'd rather do besides play video games all day?"

In many cases, the obvious prompts aren't sufficient. But in a surprising number of cases, *they are*. I still often find this advice useful myself: when my attention slips, I am often helped by someone just asking me to consider the obvious — "what would make the task less dreadful?" or "have you thought for five minutes about alternatives?" or "have you considered delegating this?" and so on.

Much of my advice for how to manage guilt was generated by this very process, by me imagining feeling guilty, and then imagining which obvious things I'd try to do to engage with the feeling. I would ask myself questions like "what is the cause of this feeling?" and "how is it being useful to me?" and "is there a better way I can achieve those goals?" and I would spend time listening to myself and brainstorming op-

tions, because those are all the obvious ways to address the problem. Many of my early posts on guilt were a product of articulating that reflexive process. The types of obvious advice that I would generate — such as asking "what is the cause and use of this feeling?" — might be very different from the obvious advice that you would generate, and that's fine. The trick is to apply the obvious advice first.

Or imagine you have the problem of finding it difficult to use the "do the obvious" technique. Maybe you've been struggling to remember to consider the obvious whenever you encounter a hard decision. Instead of asking for advice, consider generating a list like the following, first:

- Spend five minutes generating examples of decisions you made in the past where it would have been helpful to do the obvious things first. Then spend five minutes examining those and looking for patterns.
- Close your eyes and visualize yourself facing a new decision in as much concrete detail as you can, and practice thinking "oh wait, let me list the obvious things before proceeding."
- Train yourself to notice decision-points better by buying a tally counter and tracking decisions and giving yourself positive reinforcement every time you do.

Or imagine that you have tried to do all the obvious things, and you find that you're going into "enumerate the obvious" mode even for the most trivial tasks, and it's making your trivial tasks take way too long and the whole thing seems pretty foolish. Then, before complaining, consider trying the

corollary, and consider whether applying "try the obvious" far too often is in fact a terrible plan.

Your list of obvious things will very likely look very different from my own — my friends and advisors *still* generate obvious-in-retrospect ideas that I myself was incapable of generating, even after spending a few minutes generating the sort of ideas I expected them to generate. Collecting tips and techniques from other people in your environment is a great way to expand your "obvious things" repertoire, and asking for advice from friends will likely continue to generate new obvious things for quite some time. It's OK for your lists to be very different; the trick is to do *some* of the obvious preparation before making a hard decision. It can often make a difference.

The important thing, here, is to find a way to actually start doing the obvious things. This is the skill that's like footwork for a rationalist: remembering to actually do the obvious preparation is easy to learn and difficult to master; it's a skill to drill when you have spare mental energy in hopes that it comes naturally and easily whenever the going gets tough and the stakes get high.

I continue to wonder how powerful a person could become, if they simply managed to do all the obvious things in pursuit of their goals.

The Art of Response

04 JANUARY 2016

Imagine two different software engineers in job interviews. Both are asked for an algorithm that solves some programming puzzle, such as "identify all palindromes in a string of characters."

The first candidate, Alice, reflexively enters problem-solving mode upon hearing the problem. She pauses for a few seconds as she internalizes the problem, and then quickly thinks up a very inefficient algorithm that finds the answer by brute force. She decides to sketch this algorithm first (as a warm up) and then turn her mind to finding a more efficient path to the answer.

The second candidate, Bob, responds very differently to the same problem. He reflexively predicts that he won't be able to solve the problem. He struggles to quiet that voice in his head while he waits for a solution to present itself, but no solution is forthcoming. He struggles to focus as the seconds pass, until a part of his brain points out that he's been quiet for an uncomfortably long time, and the interviewer probably already thinks he's stupid. From then on, his thoughts are stuck on the situation, despite his attempts to wrest them back to the task at hand.

Part of what makes the difference between Alice and Bob might be skill: Alice might have more experience that lets her solve programming puzzles with less concerted effort,

which helps her get to a solution before self-doubt creeps in. Self-confidence may also be a factor: perhaps Alice is simply less prone to self-doubt, and therefore less prone to this type of self-sabotage.

A third difference between Alice and Bob is their *response pattern*. Bob begins by waiting blankly for a solution to present itself; Alice begins by checking whether she can solve a simple version of the problem ("can I solve it by brute force?"). Bob is more liable to panic when no answer comes ("I have been quiet for too long"), Alice is more liable to break the problem down further if no solution presents itself ("Can I divide and conquer?").

This difference is also explained in part by experience: a more seasoned software engineer is more likely to *reflexively* notice that a problem can be solved with a simple recursion, and know which data structures to apply where. I don't think it's *only* experience, though. Imagine Alice and Bob both faced with a second problem, outside their usual comfort zone — say, a friend asks them for advice about how to handle a major life-changing event. It's easy to imagine Alice attempting to understand the situation better and asking clarifying questions that help her understand how her friend is thinking about the question. It's similarly easy to imagine Bob feeling profoundly uncomfortable, while he tries to give neutral advice and worries about the fact that he might give bad advice that ruins his friend's life.

One might call what Alice is doing "confidence," but that doesn't tell us *how it's working*. And 'confidence' also comes with connotations that may not apply to Alice — she may well decide that she isn't in a position to give good advice, she may be working from a shaky understanding and thus doubt her own conclusions, even as she turns her thoughts to understanding the obstacle before her.

One of the big differences, as I see it, is the difference in the *response pattern* between Alice and Bob. Alice justs *gets down to addressing the obstacle before her*, Bob spends mental cycles floundering. Managing response patterns is something of an art: when confronted with an obstacle, does your brain switch into problem-solving gear or do you start to flail?

Note that the art of response is *not* about immediately solving any problem placed before you. Sometimes, the best automatic response is to find some way to disengage or dodge. You aren't obligated to solve every problem placed before you. The goal of having appropriate response patterns is to *avoid flailing* and *avoid staring blankly*. The goal is to have your mind shift into the problem-solving gear.

Having effective responses prepared isn't necessarily a general skill. I'm a computer programmer at heart, and a few years ago I switched paths to math research. If I'm faced

with a programming problem that I want to solve, I quickly and easily slip into effective-response-mode; I can often find solutions to problems reflexively, and when I can't, I reflexively examine the problem from many different viewpoints and start breaking it down. Yet, if you confront me with a math problem I want solved, there are still times when my reflexive response is to sit back and wait for someone else to solve it for me. (It doesn't help that I'm surrounded by brilliant mathematicians who can do so successfully.) That reflexive response — the one of blanking my mind, curious while I wait for someone else to find the answer — is not a very effective response.

Effective responses aren't about answering *quickly*, either. When paired with expertise and familiarity an effective response to an obstacle will often lead to a fast answer, but oftentimes the most effective response is to pause and think. Plenty of people have very ineffective response patterns that involve opening their mouths the moment you ask them to help you confront an obstacle. Some people reflexively start solving the wrong problem, others reflexively start making excuses for themselves, still others reflexively share personal anecdotes that paint them in a positive light. Effective response patterns are not about answering fast, they're about answering *well*.

The most competent people that I know are, almost universally, people who have very effective response patterns to obstacles in their areas of expertise. The good programmers I meet reflexively start breaking a problem down the moment they decide to solve it. The stellar mathematicians I know reflexively start prodding at problems with various techniques, or reflexively identify parts of the problem that they don't yet understand. The best businesspeople among my advisors are people who listen to me describe the choice before me, and reflexively describe the costs, constraints, and opportunities they observe. Each has acquired a highly effective response pattern to problems that fall within their area of expertise. This response pattern allows them to hit an obstacle and start taking it apart, with an Alice-like mindset, rather than flailing and doubting themselves as per Bob.

Confidence, practice, and talent all help develop these specific response patterns quite a bit. That said, you can often learn someone's *response patterns* with much less effort than it takes to learn their skills: you can start thinking in terms of incentives, opportunity costs, and markets long before you become a master economist (though reading a microeconomics textbook surely doesn't hurt). Competence isn't just about believing in your capabilities; it's also about having a pattern prepared that takes you directly to the "break down the problem and gnaw on the parts" stage without ever dumping you into the "worry about how you've been silent for a long time and reflect on the fact that the interviewer probably thinks you're dumb" zone.

Having an *explicit* pattern, such as a checklist, can help you switch from one pattern to the other. For example, imagine Bob in the example above had a checklist which read as follows:

If I start dwelling on how likely I am to fail, I will do the following. (1) *Say "hmm, let me think for a few minutes" aloud.* (2) *Verify that I understand the problem, and ask clarifying questions if I don't.* (3) *Check whether I could easily solve the problem by brute force.* (3) *Come up with a few simplifications of the problem.* (4) *Find a way to break off only one part of the problem or one of its simplified variants.*

then he may well be able to manually switch from a flailing response pattern to an effective one. This sort of manual switching is a good way to instill a new response pattern. The ultimate goal, though, is for efficient response patterns to become *reflexive*.

In fact, I think many people could benefit from developing efficient "fallback" response patterns, to handle new or surprising situations. Response patterns like "verify that your observations were correct" or "find more data" or "generate more than one plausible explanation for the surprise" and so on. As far as I can tell, there *is* a general skill of being able to smoothly handle surprising new situations and think on your feet, and I suspect this can be attained by developing good response patterns designed for surprising new situations.

This advice is not new, of course. Lots of self-help advice will tell you to break down the problems before you into smaller parts, and to infuse your actions with intentionality, and to reflexively do the obvious things, and so on. So I won't say much more on how to *attain* the Alice-like mindstate as opposed to the Bob-like mindstate. The important takeaway is that sometimes people respond to obstacles by breaking them down and other times they respond by flailing, and one way or another, it's useful to develop reflexive responses that put you into the former mindstate.

The way that I do it is by monitoring the ways that I respond to new obstacles placed before me. I watch myself facing various situations and observe which ones lead me to reflexively get defensive, or to reflexively blank my mind and wait for someone else to answer, or to reflexively freeze in shock and act dumbfounded. Then I practice building better response patterns for those situations, by figuring out what the checklists to run are, and I do my best to replace those patterns with reflexive inquiry, curiosity, requests for clarification, and impulses to take initiative. Polished response patterns have proven useful to me, and I attribute much of my skill at math, programming, and running nonprofits to having sane responses to new obstacles.

Regardless of where you get your response patterns from, I suspect that honing them will do you well.

Confidence all the way up

17 JANUARY 2016

I apparently possess some sort of aura of competence. Some say I'm confident, others say I'm arrogant, others remark on how I seem very certain of myself (which I have been told both as compliment and critique).

I was surprised, at first, by these remarks from friends and family — from my perspective, I'm usually the first person in the conversation to express uncertainty in the form of probability estimates and error bars. I'm often quick to brainstorm alternative explanations of the data I use to support my claims. And, of course, I'm certain of nothing.

In fact, I had a conversation with a friend about this phenomenon once, which went something like this:

Me: Hey, have you noticed how everyone thinks I have an aura of confidence and certainty, sometimes arrogance? I don't know how to shake it, nor how it works. What's up with that?

Him: Well, you always seem to have a solid grasp on every situation. When you're explaining things, you answer questions quickly, deftly, and with precision.

Me: I don't think that's it, though. I'm rarely confident in the claims I'm making, and I tend to highlight that fact. Earlier, when we were talking with [other friend] about tools society

can use to break monopolies, I was very explicit about where my uncertainty lies, and what assumptions my models relied upon, and where they might be flawed.

Him: *Yeah, but even then you were confident in what you were saying — maybe not confident in any particular claim you made, but confident in your overall analysis.*

Me: *I don't think that's it either. I'll be the first to admit that the probabilities I put on my propositions are pulled out of thin air, and I'll also be the first to admit that my hypothesis space is decrepit and that I'd be able to find better models if I could think better. In fact, I'm aware of a bunch of flaws in the ways I think, and I dedicate a decent amount of effort to improving my own reasoning methods.*

Him: ...

Me: ... I'm doing the thing right now, aren't I?

Him: Yes, yes you are.

There definitely is something of "confidence" to this pattern of speech and thinking, but it's not an empirical confidence. The confidence people notice in me isn't in the *content* of my claims, for I'm quick to couch my claims with probability estimates and error bars. Most of the confidence isn't in my analysis, either; I'm quick to note the ways my analyses could be flawed.

Some of the confidence *does* reside in the ways I reason; I do admit that I am much better equipped to answer questions

of the form "but why are you so much more confident in your own reasoning than their reasoning, when they actually have more credentials?" than most. But even there, I can note plausible biases and judgement errors in my own reasoning processes with alacrity.

Why, then, do I come off as so confident? Why do I seem so self-assured while listing the ways I know my brain is flawed?

On reflection, I've concluded that (at least part of) the answer is something I call "confidence all the way up". Insofar as I'm uncertain of my content, I'm confident in my analysis — except, I'm not fully confident in my analysis. But insofar as I'm uncertain of my analysis, I'm confident in my reasoning procedures — except, I don't put faith there, either. But insofar as I'm uncertain of my reasoning procedures, I'm confident in my friends and failsafe mechanisms that will eventually force me to take notice and to update. Except, that's not quite right either — it's more like, every lack of confidence is covered by confidence one meta-level higher in the cognitive chain.

The result is something that reads socially as confidence regardless of how much empirical uncertainty I'm under.

Where does it bottom out? Well, insofar as my friends and failsafe mechanisms aren't sufficient to raise errors to my attention, I expect to reason poorly in an irredeemable fashion and then fail to achieve my goals. It bottoms out at the point where I say "yeah, if I'm *that far gone*, then I fail and die."

(And somehow, I'm able to say even this while maintaining my aura of self-assuredness and confidence).

I have encountered many people who seem paralyzed by their uncertainties. They hit a question (such as "what methods can a society use to break up monopolies?") and they are pretty sure that they won't be able to generate the *right* answers, and so they generate *no* answers.

And this may be a better failure mode than the failure mode of someone who has *too much* confidence and self-assuredness, who makes up a bunch of bad answers and then believes them with all their heart.

Someone with Confidence All The Way Up, though, can achieve the third alternative: generate a bunch of bad answers, understand why they're bad and where their limitations are, and use that information as best they can.

I have found this mindset to be very useful throughout my life. Confidence all the way up is what has me dive into the fray to try new things, while others stand on the sidelines bemoaning a high degree of uncertainty. It's part of the tech-

nique of treat recurring failures as data and training, rather than as a signal that it's time to feel guilty. It's part of the technique of knowing you're deeply limited without letting that interfere with your progress towards the goal. Of the top ten most competent people I've met in person (by my estimation), eight of them seem to have some variant of confidence all the way up running. If the mindset seems foreign to you, I suggest finding a way to practice it for a while.

Confidence all the way up is about working with what you have. It's about knowing your limitations. It's about knowing that you don't have perfect models of "what you have" nor "your limitations", and proceeding anyway, with an even stride.

It's about knowing that there are going to be curveballs, and trusting your ability to handle curveballs, but not all the time; and trusting your ability to get back up when you're knocked down by a curveball you couldn't handle, but not all the time; and coming to terms with the fact that you might be hurt so badly you can't get up.

Yes, we're limited. All humans are limited. There are important, decision-relevant facts that we don't know. Our reasoning processes run on compromised hardware. But the correct response to uncertainty is not to proceed at half speed!

No matter how hard you try to justify your beliefs, if you're being honest with yourself, they won't ground out into "and therefore, no matter what I do, everything is going to be OK." No matter how hard you try to justify your reasoning, the meta-reasoning tower does not terminate at "and thus, eventually you will become capable of success." They terminate at "I may be so wrong that I can never be corrected; I may fail and all value may be lost." You will find no objectively stable perch from which to launch your reasoning.

But you were *created already in motion*. You don't *need* to ground out all your beliefs and justify all your reasoning steps before you can start moving. You don't need to have plans for every contingency before you can act. You don't need to be highly confident in your analyses before you present a model. If you sit around awaiting certainty, you will be waiting a long while.

Better, I say, to cover each lack of confidence on one level with confidence on the next level, and to come to terms with the fact that if you're so irredeemable that even your best meta-reasoning cannot save you, then you've already lost.

Desperation

24 JANUARY 2016

The next three posts will discuss what I dub the three dubious virtues: desperation, recklessness, and defiance. I call them dubious, because each can easily turn into a vice if used incorrectly or excessively. As you read these posts, keep in mind the law of equal and opposite advice. Though these virtues are dubious, I have found each of them to be a crucial component of a strong and healthy intrinsic motivation system.

The first of the three dubious virtues is *desperation*. There are bad ways to be desperate: visible desperation towards *people* can put you in a bad social position, strain your relationships, or otherwise harm you. Desperation towards a *goal*, on the other hand, is vital for a guilt-free intrinsic drive.

By "desperation towards a goal" I mean the possession of a goal so important to you that you can commit yourself to it fully, without hesitation, without some part of you wondering whether it's really worth all your effort. I mean a goal that you pursue with both reckless abandon and cautious deliberation in fair portions. I mean a goal so important that it does not occur to you to spare time wondering whether you can achieve it, but only whether *this* path to achieving it is better or worse than *that* path.

In my experience, the really powerful intrinsic motivations require that you're able to struggle as if something of incredible value is on the line. That's much easier if, on a gut level, you believe that's true.

Desperate people have a power that others lack: they have the *ability to go all out*, to put all their effort towards a task without reservation. Most people I have met don't have the ability to go all out for *anything*, not even in their imagination.

Ask yourself: is there anything *you* would go all out for? Is there anything some antagonist could put in danger, such that you would pull out all your stops? Is there any threat so dire that you would hold nothing back, in your struggle to make things right?

I have met many people who cannot honestly answer "yes" to this question, not even under imaginary circumstances. If I ask them to imagine their family being kidnapped, they say they would call the police and wait anxiously. If I ask them to imagine the world threatened by an asteroid, they say they would do their best to enjoy their remaining time. These are fine and prudent answers. Yet, even if I ask them to imagine strange scenarios where they and they alone can save the Earth at great personal cost, they often say they would do it only grudgingly.

For example, imagine that aliens that want to toy with *you in particular* have put a black hole on a collision course with Earth. Imagine that the only way to redirect it is using alien tech on an alien space ship that has been left on Earth and which can be piloted only by you and you alone — and that, to destroy the black hole, you must cross the event horizon, never to return. Would you save the world then? And if so, would you do it only grudgingly?

Would you do it if the spacecraft was sequestered atop Mt. Everest? How hard would you struggle to get to the ship, if it was at the bottom of the ocean? What if it could only be operated if you spoke fluent Mandarin, and you only had one year to learn?

Would you go all out to save the world, or would you put in a token "best effort", a token "at least I tried", and then go back to enjoying your remaining time?

And if you can't go all out even in incredible imaginary scenarios where everything depends on you, *what are you holding out for?*

A common protest here goes "I don't want to lose my friendships, my close connections, my comfort. That is too high a price to pay. If the struggle would be too brutal, then I would prefer to enjoy my remaining time instead." But if that were the case, then why couldn't someone get you to go all out by putting your friendships, connections, and comfort on the

line? Would you fight with everything you have for *those*? And if not, *what are you holding out for?*

Why are you stopping yourself from putting in a full effort, if there is no situation even in principle which could compel you to pull out all the stops? Why are you holding part of yourself back, if there is nothing *even in imagination* for which you would unbar all the holds? If there is nothing anyone could put on the line such that you'd struggle with all of your being, then *what are you holding out for?*

I'm not saying you need to be willing to go all out for something *real*. It may be that the only scenarios where you'd really struggle for all you're worth are fanciful or ridiculous. I'm saying that you need to be able to go all out in principle.

There's a certain type of vulnerability that comes with committing your whole self to something. Our culture has strong social stigmas against people who *really unabashedly care* about something.

I remember a classmate in gradeschool who *really really cared* about Pokemon, to the point that others felt embarrassed just to associate with him. The stereotypical stigma against "nerds" seems rooted at least partially in a stigma against caring too much. Derision among the intellectual

elites towards people who get really interested in sports seems to draw at least partially on the same stigma.

Notice the negative connotations attached to words like "cultist", "zealot", and "idealist". Notice all the people who distance themselves from whatever social movement they're in; those people who loosely identify as "effective altruists" or "rationalist" or "skeptics" or "atheists" but feel a deep compulsion to make sure you know that they think the *other* EAs/rationalists/skeptics/atheists are naive, Doing It Wrong, and blinded by their lack of nuanced views. I think that this is, in part, an attempt to defend against the curse of Caring Too Much.

Caring hard is uncool. The stereotypical intellectual is a detached moral non-realist who understands that nothing really matters, and looks upon all those "caring" folk with cynical bemusement.

Caring hard is vulnerable. If you care hard about something, then it becomes possible to lose something very important to you. Worse, everyone around you might think that you're putting your caring into the *wrong* thing, and see you as one of the naive blind idealist sheeple, and curl their lips at you.

Desperation is about *none of that mattering*. It's about having a goal so important that the social concerns drop away, except exactly insofar as they're relevant to the achievement of your goal. It's about being willing to let yourself care more about the task at hand than about what everyone

thinks about you, no matter how much they would deride you for fully committing.

A common barrier to desperation is that it can be difficult to admit that you really, really care about something, because then that means you are vulnerable to the loss of something that's very, very important to you. If your desperation is visible in a hostile social environment, desperation can destroy your ability to bargain and put you at a social disadvantage. Being social creatures, I suspect that many of us have mental architectures that prevent us from feeling desperation, because if we felt it, we'd show it, and that would undermine our social standing. (In my experience, confidence all the way up helps alleviate this effect.)

Thus, if you want to make desperation part of your intrinsic drive, you may need to practice becoming able to admit, to yourself, on a gut level, that you might lose something so terribly important that it's worth gaining a little desperation. You must first *allow yourself to become desperate*. (This is why I wrote about seeing the dark world and coming to your terms before writing about desperation.)

There is a common failure mode among those who succeed at becoming desperate, which is that they burn their resources too quickly, in their desperation. If you have to get yourself into an alien spacecraft at the bottom of the ocean,

and it's going to take many months of training, social and political maneuvering, and monotonous searching, then you would be unwise to spend your first week all wound up at maximum stress levels simply because you think that that's what it means to "go all out" and "hold nothing back." If you're going to pull out all the stops and unbar all the holds, you need to understand how to carry on a slow burn as well as a fast burn. (This is why I wrote about how to avoid working yourself ragged and rest in motion before writing about desperation.)

With these tools in hand, I suggest finding a way to *become able to become desperate*. Perform whatever thought experiments and meditations you have to to be able to *imagine a situation* where you would do everything in your power to achieve some outcome, without regard for the consequences (beyond their affect on the outcome). Figure out the circumstances under which you'd pull out all the stops and unbar all the holds and put *everything you have* into the struggle.

(If there is no situation, even in theory, where you would give everything you have into your efforts, then consider that there may be a part of yourself that you're holding back for nothing, a part of yourself that you're wasting.)

I'm not saying you need to become desperate *now*. That may be unnecessary. Maybe your life is going well enough, and your goals are well enough achieved, that the best way to continue achieving them is to strengthen your friendships and your connections and enjoy your comforts. If your fami-

ly is kidnapped, you probably *would* do best to call the police and then wait anxiously. If Earth is threatened by an asteroid, most people *would* do best to leave it to the experts and enjoy what time they have. So be it not upon me to force desperation upon you if you're leading a comfortable life. Make sure you don't suffer from the listless guilt, and make sure you can in principle *become* desperate, so as to ensure that you're not holding a part of yourself back for nothing, but save the actual desperation for times of need.

If, on the other hand, you are in a time of need, if you're the sort who sees every death as a tragedy, if you're otherwise fighting for something larger than yourself, then *get desperate now*.

The first step is allowing yourself to become desperate in principle. It's allowing there to be at least one imaginary scenario where you'd let yourself commit fully to a task without hesitation. Once you are able to do this, imagine the feeling that would come over you when you first committed yourself to that crucial undertaking, come whatever may. Is there a sense of desperation you would feel, a grasping need to *change the future*? Sit with it, become familiar with the sensation of desperation and any other feelings associated with the imaginary commitment.

Once you've gained some familiarity with those feelings, look with fresh eyes at what you're fighting for, at what you have to protect, at what you value, and see if any of it is worthy of a little desperation.

Recklessness

02 FEBRUARY 2016

The second dubious virtue is recklessness. As with desperation, there are many bad ways to be reckless. There is a nihilistic recklessness, in those with a muted ability to feel and care, that is self-destructive. There is a social recklessness, when peers push each other towards doing something dangerous that none of them would do alone, in a demonstration of commitment that can become needlessly dangerous. And there is a fiery, destructive recklessness in those too quick to anger, which can lead people to actions they will regret for a lifetime. I caution against all these types of recklessness.

Nevertheless, there is a type of recklessness that is a virtue. This is *recklessness in the pursuit of an external goal*, and I have found it to be rather rare.

I get a lot of questions from people about how cautious they should be as they make changes in their lives. If they remove their guilt motivation, will they be able to do anything at all? If they really try to understand how screwed up the world is, on a gut level, will they break? If they devote their efforts to the pursuit of something larger than them, will they lose touch with their humanity, and with their ability to connect to other human beings?

And I tend to answer: You are not made of glass.

Dive in. Change things. Fix problems. If more problems crop up, fix those too.

Imagine that you look upon yourself, detect harmful guilt-based motivation, tear it out, and then notice that this leaves you with a Zen-like lack of drive, such that most of yourself is now happy to let days slip by but some small part of you is crying out that something is wrong. Recklessness-the-virtue is about being in that state and deciding to push *forward* rather than retreating; deciding to make a desperate effort to acquire a new drive, rather than panicking and retreating back towards guilt.

Recklessness is about ripping off the blinders that prevent you from seeing the dark world on a gut level, and knowing that if this happens to be debilitating then you'll find some new way to handle it, rather than being forced to retreat.

Always forward, never back. Be unable to despair. Have confidence all the way up. Think of all the people you know who are too stagnant, too cautious about breaking something important, to ever change at all.

You can recover from breaking a few parts of yourself, so long as you're modular rather than fragile. You can become able to roll with a few punches.

(This seems like a good time to insert a heavy-handed reminder about the law of equal and opposite advice! Many people would do well to gain a little recklessness, but many others need *less* recklessness and *more* caution. If you're in a particularly fragile mental state, consider disregarding this post entirely.)

During my undergraduate education, I was the president of an entrepreneurship club. The first most common type of person who would drop by asking for advice was that young wannabe founder all full of naïve excitement about some half-formed notion that they're about to make the next Facebook. The second most common person was that competent programmer with an idea that wasn't half-bad — maybe they had some idea for an app that would let couples communicate in a way they couldn't yet easily do, six years ago — but, being tempered and level-headed and well aware of the naïvety of the first folks, were entirely unable to *commit* to their idea.

Both sets of prospective entrepreneurs were doomed to failure. The first set, for all the obvious reasons — they'd focus too narrowly on writing code that no one would ever buy, or fail to find their first users, or fail to make a minimum testable product, or they'd dramatically misunderstand and underestimate the difficulty of the technical challenges, or whatever.

The second set would fail because they didn't really expect themselves to succeed. They could *make* themselves work on their idea, while reciting to themselves some story about being risk-loving, but they couldn't get their head *into* the idea, to the point where they were spending fourteen hours a day working feverishly while plans and paths and strategies dominated their waking thoughts.

There's a fugue state that successful entrepreneurs report entering, which the second set of people had rendered themselves unable to enter. Somehow, their realistic understating of their odds destroyed their ability to commit.

In one fashion, this makes some sense: they, knowing that great success is likely a lie, cannot fool their innermost self into believing in their own vision, which precludes them from entering the fugue state.

But in another fashion, is silly. What do the *odds* have to do with your *ability to commit*? Why is their *epistemic* state preventing them from entering the *emotional* state that would most help them succeed?

I think there are a few different skills it takes to be able to ender the fugue state even while knowing that your odds of success are low. One of them, I think, is the virtue of recklessness.

Recklessness is in the ability to say "screw the odds, I'm going to push forward on this path as hard as I can until a bet-

ter path appears." If the odds are low, a better path is more likely to appear sooner rather than later — but the reckless let that be a fact about the *paths*, and they don't *further* allow low odds to prevent them from pushing forward on the best path they can currently see, as fast as possible.

If you want to become a successful entrepreneur, or if you want to succeed at other very difficult tasks, it helps to be able to take the best from both types of hopeless entrepreneurs. Become the sort of person who can enter the fugue state and give an idea your all, while *also* being able to see and avoid all the common failure modes. The fact that you are unlikely to succeed is an *epistemic* fact, you do not need to give it dominion over your *motivation*. Be a little reckless.

Recklessness, as a virtue, is about being able to throw caution to the wind. It's about being able to commit yourself fully to the best path before you, and then change your entire life at the drop of a pin as soon as a better path appears. It's about being free to act without worrying too much about what happens if you disrupt the status quo — too many people are already too stagnant, and we need to move faster.

So if you find yourself knowing what it is that you need to do next, but worried that doing so will break something else important...

then I say, do it.

Act.

Try not to break anything vital, but if you do, fix it and keep moving.

Always forward, never back.

Be a little reckless.

Defiance

14 FEBRUARY 2016

The third dubious virtue is defiance. As with the other dubious virtues, it can get you into trouble. Remember the law of equal and opposite advice. Used correctly, it can play a key role in a healthy guilt-free motivation system.

I used to tell people that I'm roughly 90% defiance-fueled. The most common response was "ha ha I guess you can be manipulated by reverse psychology, then"; which led me to realize that I didn't yet know how to convey what I meant by "defiance fuel," so I stopped saying it. Today, we see whether I can convey what I mean by "defiance fuel" yet.

Most people I talk to about defiance think of it as a mental stance adopted against some authority figure. Perhaps they imagine a parental figure saying "finish your broccoli," and a child who hates broccoli with their jaw set and smolder in their eyes, who proceeds to eat with as much petulance as they can muster, plotting their revenge. The feeling we imagine in that child is perhaps the standard central example of "defiance."

I claim that that child does possess defiance-the-virtue, but not in their petulance, and not in their opposition to an authority figure. Defiance-the-action is in the child chewing with their mouth open in an open refusal to submit; defiance-the-virtue is in the mental actions they make *before* they start chewing with their mouth open. It's in the inter-

nal steeling they do when deciding not to be ordered around. It's in their decision to be self-reliant, it's in their refusal to take orders lying down. If these automatic and subconscious mental motions were verbalized, they might be written "I am my own person; and not beholden to your whims," or "if you push me, I push back." But they aren't verbalized, because they aren't conscious. They're reflexive.

Defiance-the-virtue is about encountering a badness that's brewing in the world, and *reflexively* doing everything you can to throw a wrench in the works, to twist things in your favor. Defiance-the-virtue is about taking nothing lying down, and refusing to let badnesses in the universe slide.

Defiance isn't about acting petulantly without hesitation: A defiant child might bide their time, knowing that if they act rashly there will be harsh consequences. Defiance is about *resisting the default state of affairs* without hesitation: A defiant child might weigh their options and bide their time, but at no point do they wonder whether they should defy. They simply dislike the situation, and so rebel against it.

Defiance-the-virtue is about having *that* reaction, to something that's wrong in the world.

Of course, there's an art to defying the right things. I do recommend defying death; I don't recommend having the "defi-

ance" reaction against people who tell you to do things in a stern and authoritative voice. People who order you around can either be ignored or obeyed according to the social context, but they aren't usually worth *defying*, except perhaps in situations where you legitimately need to demonstrate that you're not beholden to them, and where gentler reminders have failed.

As a rule of thumb, I suggest that it's usually healthy to have a defiance reaction towards *states of the world*, and usually unhealthy to have a defiance reaction towards *people*.

To illustrate the difference, imagine you're Neo, twenty years after the first matrix movie. The sequels never happened; instead you got trapped in the matrix while one by one, all your connections to the outside world died or disappeared. One day, you lost your grasp on your ability to control the matrix, your abilities slipping through your grasp like lucidity slipping away in a dream. Now you stand atop a skyscraper, looking across the gap at its twin, unable to quite recall what it was like to fly.

You stand there frozen, desperate to recall what you once knew, finding it evasive. Behind you, someone else enters the rooftop and shouts at you over the wind.

"What the hell are you doing, you idiot?" they cry. "Get back from there! Now!"

Defiance-against-a-person would be to feel a burning need to show this person up, show them that you're not beholden to their demands, and possibly do something rash.

Defiance-against-the-world would be to hear this person cry out, and use the impetus to remember what it was you used to know. You would say, "Oh, right. I'm in the matrix." You would remember that the rules and customs of this place do not have dominion over you, no matter what illusions the people around you are taken in by. Your mind would snap back into focus. You would grab what you had forgotten how to grasp, and leap.

(And those with defiance-the-virtue deeply instilled in them don't need the impetus provided by another person to access the mental state — defiance is a property of the relationship between them and the state of the world that they can recall at will, not a property of the relationships between them and others.)

This is the defiance I mean to talk about. It's related to level hopping and skepticism about your limitations. It's related to the skill of measuring your progress not against others, but against what actually happens.

I've been writing a long sequence of posts on how to replace guilt-based motivation with something else. Many people

have remarked to me that my writings on averting guilt seem inspired by Taoism. And: maybe. There are some parallels. But not here, not with defiance.

Defiance is not about coming to terms with the world. It's about looking at the world and having the same mental reflexes as the defiant child. It's about the reflexive impulse to say "screw this" and choose self-reliance over hopelessness in the face of problems that are crushingly large. It's about a deep-seated inability to go gently into that good night. It's about being able to look at the terrible social equilibria we're all trapped in and get *pissed off* — not because any individual is evil, but because almost nobody is evil and everything is broken anyway.

Above all, it's about seeing that the world is broken, and *feeling* something akin to "fuck these mortal constraints, I'm *fixing things*."

When the defiant child eats their vegetables with as much spite as is humanly possible, there was never a *thought that crossed their mind* about capitulating to their parents. Petulance was an *automatic response*. They weren't carefully weighing a decision about whether to spite their parents — at best, they may have carefully weighed a decision about whether to get their payback now, overtly; or later, subtly. The defiance was a *reflex*; the fact that they weren't going to submit quietly to authority was never in question.

Defiance-the-virtue is about having the same reflexive response, not towards an authority figure, but towards *the state of a broken world*. It's about making the fact that you struggle to fix broken worlds *automatic and unspoken* — you might weigh your options and bide your time, but you spare no thought for *whether you will struggle*.

I don't know how to teach defiance, but it's one of the keystones of my motivation system. If you want to build yourself a motivation system akin to mine, defiance is an important component.

So this is how I suggest motivating yourself in place of guilt: Let the wrongness of the world trigger something deep inside of you, such that the question stops being whether you will capitulate or lose hope, and becomes *how you will wrest the course of the future onto a different path*. See the current state of affairs as your adversary; see the future as the prize that hangs in the balance. Shake off the illusory constraints, set your jaw, and rebel. Defy.

Allow yourself to be a little reckless. Get a little desperate. Let defiance of the way things are burn in you. Then *act*.

How we will be measured

21 FEBRUARY 2016

After nearly a year of writing, my "replacing guilt" sequence is coming to a close. I have just one more thing to say on the subject, by pointing out a running theme throughout the series.

When all is said and done, and Nature passes her final judgement, you will not be measured by the number of moments in which you worked as hard as you could. You will not be judged by someone rooting around in your mind to see whether you were good or bad. You will not be evaluated according to how unassailable your explanations are, for why the things that you couldn't possibly have prevented the things that went wrong.

You will be measured only by what *actually happens*, as will we all.

That doesn't mean all of us are using the same measuring stick: Some people are working to ensure that our universe-history is one in which they in particular have a happy and fulfilled life; others are working to ensure that our universe-history is one in which their children never have to debase themselves to survive. Still others look wide, and see poverty and destitution and suffering, and work to ensure that those blemishes fade from our universe-history, in the

places they can reach, near the time of their lives. Others look far forward, working to ensure that our universe-history is full of flourishing sentient civilizations and other nice things.

All it means is that the *type* of thing we're all trying to do, one way or another, is ensure that the actual history of our universe, the actual timeless structure of the place we're embedded, is as desirable as possible. That's the type of game we're playing: We manipulate universe-histories, for the sake of the future.

Some people have a listless guilt, thinking that nothing matters but feeling vaguely restless as they watch themselves spend their lives on things they think are pointless. Other people have a pointed guilt, thinking that *everything* matters, and berating themselves whenever they fall short of perfection. For me, the framing that *we act to determine the shape of our actual universe-history* is a framing that avoids both these pitfalls. Is there a way you want the completed, timeless story of our universe to go? Then act to ensure that the future is as good as you can make it. Are you wracked with guilt about your inability to act as you wish, or regret for the things you did in the past? Then act to ensure that the future is as good as you can make it. That's the sort of game we're playing: At all times, act to ensure that our future is bright.

I think many people get a bit mixed up about what type of game we're playing. They get stuck playing a social game, measuring their accomplishments by comparison to the accomplishments of their neighbors; or they mistake someone else's expectations for their preferences and get stuck chasing lost purposes; or someone slights them and their vision narrows as their sole objective becomes *retaliation*.

I'm not saying social goals are intrinsically bad. Wealth and status are useful aids when it comes to determining the future; the accomplishments and expectations of your peers can provide useful measurements of your abilities. But there's a difference between pursuing social goals for the sake of determining the course of our universe-history, and forgetting entirely that success is measured in terms of what actually happens throughout the course of history.

I alluded to this when I described defiance as "choosing self-reliance." At the end of the day, each and every one of us is engaged in a personal struggle to determine the future. We are not *alone*; there are many around us who can be friends and allies and support us in our struggle. But the goal, in the end, is to use what resources we have at our disposal to ensure that the universe-history is filled with light, whatever our light may be. I hope yours includes friends and family and loved ones, but *making it happen* — that is your personal task. You are encouraged to draw on the support of friends and allies where possible; and ensuring that you have close connections may be one of the properties you're putting into the timeless history of our universe: But even

then, the task of *ensuring our universe-history is one in which you have close connections* is your personal task.

What we are doing, on this earth, is acting in such a way that our future is filled with light. From this framing, "guilt-based motivation" is a foreign concept: If you start to feel guilty, simply look at your situation with fresh eyes, and then act such that the future is filled with light. Our lives are not status competitions; the world is not a proving ground. We are participating in a *gambit for the future* (or, more likely, a gambit for the shape of the multiverse), and that is all.

When there are people who oppose us out of nothing save for petty spite; when there are obstacles that stand between us and something important to us which seem all but insurmountable; when we encounter personal limitations that prevent us from acting as we wish to; it is easy to confuse retaliation, overcoming adversity, and growing stronger, with our actual goals. But crossing those hurdles is not the final objective: those hurdles are only parameters in our calculations about how to affect the future; they are nothing but the state of the game board in a game with cosmic stakes.

In that game, some people have stronger positions than others, and more leverage with which to determine the timeless story of our universe. Life isn't fair. But all of us, one

way or another, are here to make sure that our universe history is filled with light — whatever 'light' may be to each of us.

So find allies, find friends, find everything you need to improve your ability to ensure that our universe-history tells a story you like. Move towards whatever levers on our future you can find. And then *fill it with light*.

On caring

04 OCTOBER 2014

1

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

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

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

I feel a *little* respect for the bigness of numbers, if you pick really really large numbers. If you say "one followed by a hundred zeroes", then this feels *a lot* bigger than a billion. But it certainly doesn't feel (in my gut) like it's 10 000 000 000 000 000 000 000 000 000 000 000 000 times bigger than a billion. Not in the way that four apples *inter-*

nally feels like twice as many as two apples. My brain can't even begin to wrap itself around this sort of magnitude differential.

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

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

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

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

And this is a problem.

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

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

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

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

Saving a person's life feels *great*, and it would probably feel just about as good to save one life as it would feel to save the world. It surely wouldn't be *many billion times* more of a high to save the world, because your hardware can't express a feeling a billion times bigger than the feeling of saving a

person's life. But even though the altruistic high from saving someone's life would be shockingly similar to the altruistic high from saving the world, always remember that *behind* those similar feelings there is a whole world of difference.

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

3

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

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

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

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

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

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

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

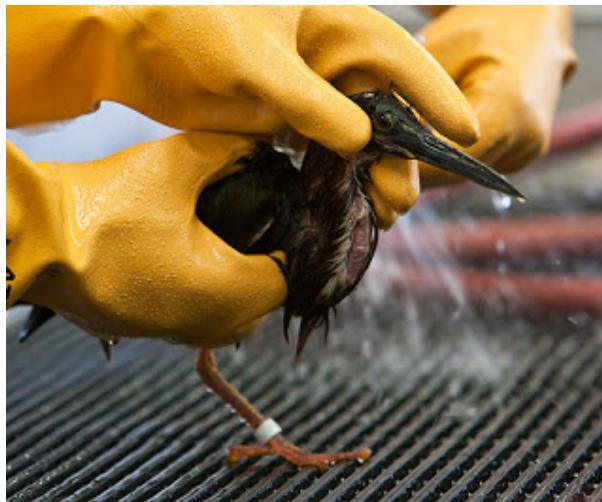
know that such people are crazy. (Good crazy, perhaps, but crazy all the same.)

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

4

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

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



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

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

Thousands and thousands of birds were oiled by the BP spill alone. After shutting up and multiplying, Daniel realizes (with growing horror) that the amount he *actually* cares about oiled birds is lower bounded by two months of hard

work and/or fifty thousand dollars. And that's not even counting wildlife threatened by other oil spills.

And if he cares that much about *de-oiling birds*, then how much does he actually care about factory farming, nevermind hunger, or poverty, or sickness? How much does he actually care about wars that ravage nations? About neglected, deprived children? About the future of humanity? He *actually* cares about these things to the tune of much more money than he has, and much more time than he has.

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

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

Daniel doesn't wind up giving \$50k to the WWF, and he also doesn't donate to ALSA or NBCF. But if you ask *Daniel* why he's not donating all his money, he won't look at you funny or think you're rude. He's left the place where you don't care

far behind, and has realized that *his mind was lying to him the whole time* about the gravity of the real problems.

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

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

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

Alice and Bob and Christine usually aren't spending time solving all the world's problems because they forget to see them. If you remind them — put them in a social context where they remember how much they care (hopefully with-

out guilt or pressure) — then they'll likely donate a little money.

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

5

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

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

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

But this is an error — prominent altruists aren't the people who have a larger care-o-meter, they're the people who

have learned not to trust their care-o-meters.

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

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

6

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

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

I think that at least part of it comes from a certain sort of desperate perspective. It's not enough to think you *should* change the world — you also need the sort of desperation that comes from realizing that you would dedicate your entire life to solving the world's 100th biggest problem if you

could, but you can't, because there are 99 bigger problems you have to address first.

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

7

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

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

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

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

Because our feelings lie to us.

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

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

8

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

But sometimes, you can catch a glimpse.

The value of a life

16 FEBRUARY 2015

If you have money and want to save lives, you had better put a price on life. Scott Alexander explains it better than I can.

But don't mix up the *price* of a life with the *value* of a life. I see this happen all too frequently. To correct this mistake, I'm going to tell a little story.

Once upon a time, there was a village of peaceful immortal humans. They did not age past their primes, but they could still die from starvation or injury. But perhaps because their lives were so long and full, they all valued each other very highly and lived in peace. Indeed, there were no lengths to which the villagers would not go in order to save one of their fellows from unwanted annihilation.

Or, at least, that's how life was before the dragon came.

Dragons desire two things from people, as I'm sure you know: gold, and flesh. And this dragon, woe be upon the villagers, was powerful indeed — nigh invincible, with a cunning to match. The dragon, easily capable of killing the entire village outright, gave a grim ultimatum:

Each person in this village must pay a tax of gold, every year, in proportion to that person's age. Anyone who cannot pay the tax will be eaten.

The villagers begged and pleaded, they wept and raged, but the dragon was unmoved. It merely showed them a few heaps of rock that looked likely to make good gold mines, and told them to get to work.

The villagers tried their best. They really did. They were not miners, but they were fast learners. They worked themselves ragged, throwing aside stones, digging at the earth with their bare hands until their fingers bled, hunting and gathering as little as possible, letting their shelters deteriorate — yet still, they could not make the dragon's tax. At the end of a year, the dragon returned, and took all the gold they had, and ten of the oldest villagers (for giving up the eldest villagers was the way to save the most lives).

Distraught, the villagers resolved to try harder next time. They pushed themselves to their limits and beyond. They raced against time. They grew gaunt and ragged. Their eyes sank, their skin grew sallow, their arms thinned. They pushed themselves too hard, until they were collapsing in the mines. The next time the dragon came, it took all their gold and fifty of their number.

Their strategy wasn't working.

But these villagers were born of humanity, and ingenuity is humanity's birthright. So in their third year, the surviving villagers came to bitter terms with their situation, and set to hunting and gathering and growing stronger, accepting that they had to take care of themselves before they could take care of their friends. They set to building picks and shovels, realizing that they could not save themselves with their hands alone.

At the end of the third year, the dragon took all their gold and one hundred of their number, for their infrastructure had not yet started paying off.

But by the end of the fourth year, the dragon only took two.

Shortly thereafter, the dragon (delighted by their progress) informed the villagers that the tax would now begin increasing faster; exponentially in age.

This time, the villagers only nodded, and forged their hot fury into cold resolve.

It has been many, many years since the dragon came to the village. In fact, it is not a village any more: the village grew to a city, and the city grew to a civilization.

The population is quite a bit younger now. The elders are wiser and more productive, and can get more gold out of the ground per hour, but there simply comes a time when this increased productivity is not worth the cost in lives. When

that time comes, the elders go willingly to their fate, for these people are not the type to buy their own lives at the cost of two others.

In fact, hard tradeoffs such as these are commonplace. The villagers long ago discovered specialization and economics, and now most of them don't work in the mines. Some of them spend time growing or preparing food, others spend time maintaining shelter, others spend time inventing new tools and mechanisms that can keep pace with the dragon's dreadful tax. Indeed, some spend their lives on art and entertainment — for the villagers have learned the importance of maintaining motivation and morale.

(And some villagers, deep underground, far from the dragon's prying eyes, are designing weapons.)

So you will find, in this civilization, that there are people who dedicate their lives not to mining gold, but to writing books — but if you look closely, you'll notice that this only happens when the author can save *more* lives through increased morale and productivity than they can through working in the mines directly. And so this civilization, hell-bent on saving as many people as it can every year, still produces books and plays and movies.

Which means that in modern times, you can calculate the exact cost of saving an additional life. It turns out that one life goes for about the same price as a thousand movie tickets.

As it happens, two of the citizens of this dragon-ridden world, Alice and Bob, are having a conversation about the value of a life, right now. Let's listen in:

Alice: So you see, the true value of a life is equivalent to about a thousand views on the latest blockbuster.

Bob: Nonsense! A life is worth much more than two thousand hours of movie-viewing! A life is nigh invaluable! You can't put a price tag on a human life!

Alice: What hollow indignation! If your actions are inconsistent with putting a price tag on life, then there are ways you could shuffle money around to save more lives. If you want to save as many people as possible with a limited amount of money, then you *must* put a price on life!

Bob: But a thousand viewings of a movie simply isn't worth the same as a life! If I got to choose between a thousand people watching another blockbuster and the life of my mother, I'd choose the life of my mother any day!

Alice: Yes, but this intuition is inconsistent. The market for lives here is efficient, and the market has spoken, and the market says that a life is equal to about a thousand views of the latest blockbuster. Your mother's life *isn't* worth more than the accumulated pleasure that a thousand people experience when watching the latest blockbuster! The viewing experience and your mother's life just turn out to have the

same value, and if your intuition disagrees, you'll have to fix your intuition!

Do you see the errors here?

Alice and Bob are both right, and both wrong.

Alice is correct in that the villagers *must* treat a life as equivalent to a few thousand hours worth of watching movies.

Given that the villagers are all still trying to save each other, those thousand people only go to the movies if the resulting boost in motivation and morale leads them to collectively generate enough additional wealth to save *more* than one additional person. If you stopped those people going to the movies, and put their money towards producing gold instead, then *less* gold would be produced overall, and more people would die. Bob *must* trade off two thousand movie-hours against one life, if he wants to maximize lives saved.

But Bob is correct in that the value of a life is worth much more than two thousand hours of viewing movies!

Alice's claim is that the *sum experience* of two thousand movie-hours is equal to the intrinsic value of a life. The market has spoken, and so you must not protest, if you want to save lives.

But in fact, the very reason that Bob must treat the thousand movie-viewings as equivalent to a life is because those viewings lead to increased morale, which leads to *more than one life being saved*. This fact *does not* equate the experience of a life lived to the pleasure of the viewers.

What Alice has forgotten is that the village is *plagued by a dragon*.

Were it not for the dragon, these villagers would go to almost any lengths to save each other from unwanted death. There might be *some* lengths to which they would not go, *some* price they would not pay, in pain, sorrow, and decreased quality of life among the rest of the villagers, in order to save a friend. But, in the absence of a dragon, this cost would be a *hell of a lot higher* than two-thousand hours worth of watching movies.

Enough analogies. Let's look at our universe, now. Our economy is *not* efficient — it costs a few million dollars to save a life in developed nations, and a few thousand dollars to save a life in underdeveloped nations (where "save a life" really only means "push death back a bit", in these dark times). Furthermore, our economy is *not* maximizing for lives: humans are prone to scope insensitivity and a whole slew of other biases that dampen their ability to care about other humans dying against their will. Furthermore, it is impor-

tant to care not only about the lives we save, but about the lives we *live*.

Despite all this, we are not all that different from those villagers in the lengths we would go to save each other if death was not inevitable.

I don't know how the future will turn out. I don't know how we'll end up trading off the preservation of a life against the improvement of a life against the creation of a life, if and when we make it past this phase of scarcity. But I can tell you this: *There may well come a day when humanity would tear apart a thousand suns in order to prevent a single untimely death.*

That is the value of a life.

You still have to put a price tag on lives, and that price tag still has to be somewhere between a few thousand dollars and a few million dollars.

Imagine a button which, when pressed, picks a random number between 1 and a million. If that number is 1, it kills a randomly selected person. How much would somebody have to pay you to press that button?

Many people react with disgust, saying they wouldn't press such a button at any price. They say that the value of a life is nigh inconceivable.

And this intuition is correct!

But when somebody offers you ten dollars to press that button, press it anyway. Press it, and worry about it less than you worry about driving a car for a year (which, if I did my math right, is like pressing a button that has a one in ten thousand chance of killing somebody each year, in return for the convenience of driving [1] [2]). If you want to save the most lives, then you press that button for \$10, and you put the money towards saving lives.

But don't confuse the cost of a life with the value of a life!

In some parts of this world, it costs as little as a few thousand dollars to save a life. If you act like the price on a life is higher than a few thousand dollars, if you *actually* refuse a million dollars to press the button, or pay a billion dollars to save a single life, then there were other things you could have done to save more lives. If you want to save the most people, you *must* put a price tag on life according to the *actual* cost of saving a life.

But you don't have to confuse the *current cost of saving a life* with the *intrinsic value of a life*.

There is a gap there. There is a gap between how much a life is *really worth*, and the price tag that you *must assign*. That gap is not there because your intuitions are wrong. That gap is there because *our village is being plagued by a goddamn dragon.*

That gap is a direct measure of the difference between the universe that is, and the universe that should be.

That price difference, the difference between a few thousand dollars and a few thousand suns, is a direct measure of how fucked up things are.

Most people start with an intuition that they should refuse to press the button at any price, because lives are nigh invaluable. You can go to these people, and show them that in order to save as many lives as possible with a bounded amount of money, they must put a price on life. Most people, at that point, react one of two ways.

Some accept the logic and reject their intuitions. They see that, to save the most lives, they *must* use a price tag. It sounds repugnant to say that the pleasure experienced by a few million people drinking a can of soda is equivalent to the value of a life, but (they think) that's exactly the sort of reasoning that leads someone to thinking that life is invaluable, which is a deadly misconception. And so, wanting to

save as many people as they can with the money allotted to life-saving, they bite the bullet, and conclude that lives were never worth all that much anyway.

Others reject the logic, and continue to claim that life is invaluable, and then try to back up their intuitions with some strange version of ethics where saving as many lives as possible with the money available is not the right thing to do, for convoluted reasons.

But there's a third option here! All these people have forgotten about the dragon!

It is possible to live in a universe where it is *both* the case that (1) lives are nigh invaluable, and (2) people are being annihilated constantly, against their will, in ways that can be prevented using relatively small sums of money.

The universe is not fair! Pressing the button for \$10 is the way to save the most lives, *and this very fact is a horrible thing*. Lives are nigh invaluable, *but you have to treat them as if they're worth only a few thousand dollars*.

This gap between price and value is *unacceptable*, but physics wasn't written according to what we would accept. We live in a cold, uncaring universe; a universe beyond the reach of God.

One day, we may slay the dragons that plague us. One day we, like the villagers in their early days, may have the luxu-

ry of going to any length in order to prevent a fellow sentient mind from being condemned to oblivion unwillingly. If we ever make it that far, the worth of a life will be measured not in dollars, but in *stars*.

That is the value of a life. It will be the value of a life then, and it is the value of a life *now*.

So when somebody offers \$10 to press that button, you press it. You press the hell out of it. It's the best strategy available to you; it's the only way to save as many people as you can. *But don't ever forget that this very fact is a terrible tragedy.*

Don't ever forget about the gap between how little a life costs and how much a life is worth. For that gap is an account of the darkness in this universe, it is a measure of how very far we have left to go.

I don't want to turn this into a sermon. But some of you, seeing the great abyss between cost and worth clearly for the first time, may decide that this gap is worth closing, that our dragons are dragons worth slaying. Some of you may be wondering, what now? What next? This last part is for you.

Know that there are those of us who fight.

Some of us work in the mines to make the dragon's tax. Others prepare for the day we will confront the dragon — for the weapons we must bring to bear will be powerful indeed, and may prove difficult to aim.

And this is a fight you can join. For some of you, fighting means joining an effective cause. But for most of you, fighting means putting a low price tag on lives, and then honoring it — by purchasing lives wherever they are cheapest; by donating to highly effective causes. Remember that just as courage is about doing the right thing even though you're afraid, caring is about doing the right thing even when you're not overwhelmed by emotion.

If this is a fight you wish to join, then I urge you to remember the first lesson that the villagers learned: you must care for yourself before you care for others. You do not need to become destitute to struggle against the darkness in this universe. Any small amount of money or effort you can put towards saving lives is money and effort well spent. Pledging 10% of your earnings to an effective cause is a difficult achievement worthy of great acclaim.

If you are going to stand beside us in this fight, then I will welcome you no matter what — but I would rather you join us filled with hot fury or cold resolve, rather than with guilt or shame.

Oh, Death was never an enemy of ours!

We laughed with him, we leagued with him, old chum.

*No soldier's paid to kick against his powers.
We laughed, knowing that better men would come,
And greater wars; when each proud fighter brags
He wars on Death, for lives; not men, for flags.*

— Final stanza of *The Next War*, by Wilfred Owen

Moving towards the goal

07 DECEMBER 2014

This post contains some advice. I dare not call it obvious, as the illusion of transparency is ever-present. I will call it simple, but people occasionally remind me that they really appreciate the simple advice. So here we go:

1

(As usual, this advice is not for everyone; today I am primarily speaking to those who have something to protect.)

I have been spending quite a bit of time, recently, working with people who are explicitly trying to hop on a higher growth curve and have a larger impact on the world. (Most of them effective altruists.) They wonder how the big problems can be solved, or how one single person can themselves move the needle in a meaningful way. They ask questions like "what needs to be done?", or "what sort of high impact things can I do right now?"

I think this is the wrong way of looking at things.

When I have a big problem that I want solved, I have found that there is one simple process which tends to work. It goes like this:

1. Move towards the goal.

(It's simple, not easy.)

If you follow this process, you either win or you die. (Or you run out of time. Speed is encouraged. So are shortcuts, so is cheating.)

The difficult part is hidden within step 1: it's often hard to keep moving towards the goal. It's difficult to stay motivated. It's difficult to stay focused, especially when pursuing an ambitious goal such as "end ageing," which requires overcoming some fairly significant obstacles.

But we are human beings. We are the single most powerful optimization process in the known universe, with the only exception being *groups* of human beings. If we set ourselves to something and don't stop, we either succeed or we die.

There's a whole slew of advice which helps make the former outcome more likely than the latter (via efficiency, etc.), but first it is necessary to begin.

Moving towards the goal doesn't mean you have to work directly on whatever problem you're solving. If you're trying to end aging, then putting on a lab coat and combining random chemicals likely won't do you much good.

Rather, moving towards the goal is about *always acting to solve the problem*, with each motion. Identify the path to the goal that seems shortest, and then walk it. Maybe you need

to acquire financial stability first, and more knowledge second. Maybe you need to expand your social network, or fulfill your social attachment needs. Maybe you need to acquire a new skill. Maybe you have *no idea* how to start, in which case you need to gain more information, do some thinking, and gain a higher vantage point from which to search for a path to the goal.

But no matter what, there is always *some* way to keep moving towards the goal. Get stronger. Get smarter. Return with allies at your back.

2

Here's the pattern that this advice is designed to work against: consider the effective altruist, asking "what needs to be done?", or "what sort of high impact things can I do right now?"

I expect people to go much farther by first identifying an actual goal, and then moving towards it. Which breaks my one-step advice above into a more practical two-step process:

Step 1: identify the goal. Figure out what you're actually trying to accomplish. Probe your motivations, and trace them back to something that compels.

I'm not suggesting tracing your motivations all the way up to "final" goals; it's a bit presumptuous to claim knowledge

of "final goals" given modern introspective capabilities. Rather, look for important problems that you're trying to solve in the world today.

For example, you might be trying to fix education, end hunger, eliminate a disease, prevent aging, become immortal, end suffering, prevent human extinction, or whatever. None of these are *ends unto themselves*, but they're all problems that need solving.

Identifying a goal that compels—that really needs to be solved, and that won't be solved (or won't be solved fast enough) by default—is not always an easy task. Many people are locked into a mindset where they couldn't possibly actually solve any big problems, because big problems are big and people are small. Breaking out of that mindset is a topic for another day; for now I'll assume you have picked your poison and identified some goal to achieve, even if only a minor one.

Step 2: move towards it. So, you've found a goal. Nice work.

Now solve it tomorrow.

Can you? Seriously ask yourself whether or not you can solve the problem tomorrow. I don't care how ambitious it is. Can you solve it tomorrow? If yes, then do it. If not, why not? Say the obstacles aloud.

The usual answers are something like "I lack the power, time, money, network, and so on." Which is great! Now we're getting somewhere.

These are what you need to work on tomorrow, if you want to solve the problem.

Don't ask "what would be good to do," ask "what is standing between me and solving the problem immediately." Identify the obstacles. Your task is now to either remove them or cheat your way around them.

Of course, most of the obstacles themselves are still too big and vague. So ask yourself why you can't solve those problems tomorrow. Say you don't know the people you'd need to know to have a shot at fixing education. Can you contact them all tomorrow? That *probably* wouldn't go well, but why not? What are the obstacles between you and acquiring the resources you're going to need?

Rinse, repeat. Identify the obstacles to overcoming the obstacles, and so on. Eventually, this process will ground out in things that you can actually start doing tomorrow, with a path that you can trace all the way back up to your goal.

Once you have that, throw reservations to the wind, and start *today*.

Moving towards the goal doesn't solve the whole problem. If you want to solve a goal effectively, in the time allotted, it is important to approach the obstacles in the right order, to identify the ones you can safely cheat past, to correctly distinguish between short paths to the goal and long ones. But many people aren't there yet: they're still asking "what would be good for me to do," and not "what stands between me and solving the whole problem tomorrow."

My advice, if you want to be effective, is *always be solving the problem*. With each motion, be overcoming an obstacle that stands between you and the goal. If the obstacles are too large, then your next task is to get stronger, get smarter, or find a way around. That is what it means, to find a path to the goal.

To achieve a goal, simply keep moving along that path.

Self-signaling the ability to do what you want

26 OCTOBER 2014

1

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

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

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

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

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

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

The counterargument would go something like

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

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

2

Failures of this form can generally be fixed by "just not doing that," which in this case entails forcing yourself to stop eating. I don't like that solution, as it requires an application of willpower, and in general, any solution that requires an application of willpower is a stopgap, not a remedy. I much prefer solutions that get all of myself onto the same page,

including the parts that make distracting arguments so they can shovel more food into my mouth while I'm not looking.

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

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

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

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

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

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

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

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

3

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

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

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

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

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

4

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

(I expect this is much easier above a certain confidence threshold, such that others say you are "eccentric" rather than "a weirdo." Your mileage may vary. But don't take that as an excuse; I still strongly encourage you to show yourself

that you are able to do the right thing even after it's "too late".)

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

5

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

Productivity through self-loyalty

03 NOVEMBER 2014

1

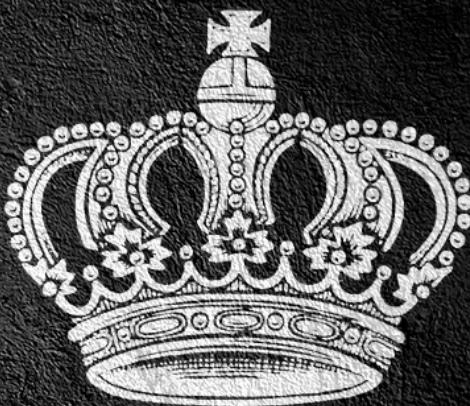
I can be pretty dang productive when I put my mind to it.

Many people have a generic mind-model which runs roughly as follows: a person's reported desires are but one voice among the thronging mob of forces that govern the brain, and it takes significant effort and force of will to align the mob for long enough for people to get something done.

Many of us have experienced a desire to stop procrastinating, and then have watched helplessly as we continue to surf the internet. Many of us have resolved to do something difficult, only to watch the opportunity flit by us as we stand motionless at the sidelines.

YOU ARE NOT THE
KING
OF YOUR BRAIN.

*You are the creepy
guy standing next
to the king going
"a most judicious
choice, sire".*



— STEVEN KAAS

*prettyRATIONAL.com

People use something like this model when they speak of akrasia, the tendency to act against your own better judgment. Haidt analogizes the brain to someone riding an elephant, where the conscious mind is a rider struggling to steer. Kahneman writes of a dichotomy between "fast," emotional, immediate processes that govern most of our thinking, and "slow," deliberate, conscious processes that occasionally assume command. I have found that the "spoon theory" model of energy reserves resonates for many people, even those who aren't chronically ill or otherwise disabled.

In all these models, there is a tendency to separate the voice from the mob. Insofar as the voice has the ability to direct the mob (steer the elephant, convince system 1, etc.), we get to do what we want. But when the mob loses interest or focus or motivation, we are at its leisure.

I find a lot of truth in these models, and so do many others. Thus, many people, upon seeing my high levels of productivity, expect that I must be very very good at keeping tight control over the mental mob, and forcing them to do things that they would rather not do. It's not uncommon for people to remark that I need to be careful about strong-arming the mob (as eventually they rebel, leading to burn-out), or for people to tell me that I must have some sort of iron will (which they cannot replicate).

I don't think this is the case. As I said last time:

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

It's *possible* to force the mob to do something, and this is why willpower is often useful in the short term. But it's seldom a good idea to try to force yourself to do things the mob doesn't want to do in the long term. Ultimately, the mob is the one actually managing your motivation systems, and any plan that relies upon a permanent use of mental force is unlikely to succeed.

It is much better to have the mob on the same side as the voice of reason.

But this is something of a catch 22: many people have mind-mobs that just want to sit around all day and watch TV shows or surf the internet. If your mind-mob just wants to

rest and I'm cautioning against force, then how does one ever attain high levels of productivity?

My answer is complex, and relies upon many tools. I've discussed a few of them in the past, and today I'll discuss another.

2

First, a word of warning: remember the law of equal and opposite advice. For every piece of advice useful to one person, there is some other person who needs exactly the opposite advice.

I am going to discuss a technique that I use for productivity which results in a sense of austerity through compassion/camaraderie: the parts of me that need rest take as much rest as they need, but also try to take as little as they need out of awareness of the scarcity of resources and compassion for the other parts of me.

This has proved a powerful technique for me, but it may be exactly the wrong tool for many others. The goal is *not* to guilt-trip the parts of you that need extra rest, and the goal is *not* to give yourself over to self-indulgent whims. I personally find a lot of power somewhere in between, at "compassionate austerity," but many others may react poorly to any internal narrative of scarce resources and mental frugality. *Remember the law of equal and opposite advice.*

3

Imagine a student who has been assigned a very important bit of homework with a deadline looming ever closer. Let's say they're trying to kick themselves into high productivity mode. How can they do this? Well, they can pull out the whips and cattle prods and *force* their mind-mob to be productive (with gritted teeth and building malcontent), or they can use their most desperate voice and plead with themselves, promising rewards for good behavior (that the mob might just take anyway, if it suits them), or they can wait until the deadline is so close that even the short-sighted mob can see it, at which point they'll go into panic mode (which is kinda like high productivity mode, if you squint).

But there's also a fourth option, which is something like "gain the trust of the mob, and build rapport." If the student gets the mob onto *their* side, then the paper will be done automatically, no willpower or pleading or panic necessary. This obviously sounds nice, but how is it done?

I do this, at least in part, by *showing the mob that I am on their side first*. This involves self-signalling, as discussed in last week's post. Specifically, it involves signaling to yourself that *you are loyal to the mob*.

Sometimes, the mob in you will make demands that sound unreasonable, such as "cancel everything today, I need a break." In these situations, it's easy to try to force or plea or

bargain with yourself. I take a different tactic: I ask myself if this is really what I need, and if it is, *then I do it.*

I show the mob that I respect its demands, and that I'm on its side. After all, we have the same goals; and furthermore, I am not the king in my mind. I do not desire a fight (and if I did, I wouldn't win it).

There are some really bad ways to do this (remember the law of equal and opposite advice!), and if you do this incorrectly it may lead to destructive self-indulgence. If your voice of reason signals helplessness in the face of the mob's whims, if it *gives itself up* to the mob, then you might end up unhappily pursuing short-sighted whims. The trick is to signal *respect* for the mob instead: what my mind reports it needs, it gets. This—an unflinching willingness to get the mob what it wants—*tempers* the mob's demands.

The appropriate sentiment can perhaps best be described by this clip from the film *It's a Wonderful Life*:

(start at 2:58, watch through 6:26)

A Wonderful Life without FRL



4

This scene portrays a bank run during the beginning of the great depression. It features the protagonist, George Bailey, trying to calm down a worried mob by reminding them that they're all in this crisis together. The mob doesn't really go for it, and he ends up using his honeymoon money to keep the bank alive.

The first member of the mob to get his money out of the bank demands the full value of his account, \$242. George pleads for austerity, reminding him that they're all in this together, but Tom still demands all his money. George doesn't protest or argue, he just nods and pays out Tom's entire account (and then extends the man a little extra compassion, to boot). The next two members of the mob say they can get away with \$20, and are starting to express some concern for George using his own money for this. Then Mrs. Davis bids *lower*, asking for only \$17.50. Overcome, George gives her a kiss on the cheek.

This is the sort of relationship—between George Bailey and the mob—that I have the "voice of reason" cultivate with the

varied and disparate parts of my mind. When some part of me demands that I pay its full account, I'll ask it once how much it *needs*, but if it still demands its full account I'll pay up without hesitation (and extend some additional compassion). This is done not in an appeasing way, but in a respectful way: we're all in this together.

The mob understands that the voice of reason is responsible for many of the good outcomes that I've achieved, and the mob understands that things like "rest" and "relaxation" and "procrastination" are expensive in terms of ability to achieve good outcomes—I'm "paying out of the honeymoon money."

But the voice of reason, in turn, is *willing* to pay out of its honeymoon money. It knows that everyone is going to need some resources to make it through, and does not begrudge any part of me for that.

There are two important components to this sort of self-relationship: First, the mob must respect the voice of reason, by understanding that the voice of reason achieves many nice things, such as food and roofs and clever schemes and so on. Second, the mob must know that the voice of reason is *loyal to them*. When some mind part *does* demand something ostentatious, such as "a few days of doing nothing," then the voice of reason is willing to acquire it.

My loyalty is not to any individual appointment or task. My own mental health is among my top priorities.

Once the mob sees this, once the mob *knows* that I will move the heavens and the earth in order to meet its needs, it doesn't tend to demand the full account. Because, in fact, the mob *respects* the scarcity of scarce resources, it *wants* the voice of reason to have enough flexibility to keep on achieving good outcomes. Done right, the mob enters a sort of camaraderie where it takes as little as it can out of *compassion*, because we all know that life can be hard.

When Mrs. Davis leaves that bank with \$17.50, she isn't feeling resentment or smugness. She knows that she's going to have to struggle a bit to live on only \$17.50 until the bank re-opens, but she isn't dreading the struggle or muttering curses. No, she goes home filled with compassion, with respect for George Bailey who is taking great pains to get everyone through this crisis together, and with a tighter feeling of community and closeness to those around her enduring similar austerity. She goes home happy and warm.

5

This is the mindstate in which I attain high productivity: various parts of the mob of my mind occasionally need rest, recuperation, and procrastination. Parts of me ask for these things. When they do, I ask them how much they really need, how much they can get by with. Do I actually need to take four days off? Because I will, but it's expensive.

Often, when a part of me really needs a break, and throws up its hands feeling overwhelmed, its initial demands are

unrealistic—"two weeks with no responsibilities!" So then I ask it again, with the demeanor of George Bailey, what it really needs to get by. And that part of me quickly remembers that all of me is in this together, and that I'm trying to do some very difficult things, and that all parts of me are constrained by scarce resources. Then the part that protested searches for what it really needs, the bare minimum, and it usually answers something like "I can get the rest I need in fifteen minutes."

And this sacrifice can leave me feeling stronger, feeling warmth and compassion and self-camaraderie, the same feeling that spurs George Bailey to kiss Mrs. Davis' cheek in the video clip above.

6

There is only so much time and attention that we have in this world, and we're trying to do many amazing and wonderful things. If you want to be able to do more than you're currently doing, I don't suggest trying to force yourself. Instead, I suggest showing yourself that you really are willing to move the heaven and earth *for yourself*, in order to satisfy your needs. This, in turn, can help you build up the mental camaraderie (and resulting austerity) that comes from all the parts of you understanding that you're all in this together.

Conclusion of the Re-placing Guilt series

28 FEBRUARY 2016

Today marks the end of my series on replacing guilt ([table of contents](#)).

I began the series by discussing the "restless guilt," that people feel when some part of them thinks they aren't doing what's important. I argued that it's possible to care about things outside yourself, and things larger than yourself, no matter what a nihilist tells you.

In the second arc of the series I implored readers to drop their obligations and ask themselves where they would put their efforts if there was nothing they felt they "should" be doing. If you can drop your sense of obligation and still care hard for something larger than yourself, you are well on your way to dispensing with guilt-based motivation.

In the third arc, I described techniques for building and maintaining a powerful intrinsic drive without the need to spur yourself with guilt. I point out that working yourself ragged is not a virtue, and that the "work too hard then rest a long time" narrative is a dangerous narrative. We can't always act as we wish we could: We're not yet gods, and it's often easier to change our behavior by exploring obstacles with experimentation and creativity instead of attempting

to berate and guilt ourselves into submission. I plea for self compassion and argue that there are no "bad people".

In the fourth arc, I describe ways to draw on the fact that the world around you is broken as fuel for your intrinsic drive. If, when given the choice between "bad" and "worse" you can choose "bad" without suffering; if you can be content in your gambles while having no excuses and coming to terms with the fact that you may fail, then it becomes easy to transmute your guilt into resolve and struggle hard to make the future as bright as you can make it.

In the fifth and final arc, I describe mindsets and mental stances from which guilt seems an alien concept. Primary among them are "confidence all the way up", the skill of believing in your capabilities while not being overly sure of anything; and desperate recklessness defiance, the three dubious virtues of those with strong intrinsic drive.

I conclude with a few words on how we will be measured: When all is said and done, Nature will not judge us by our actions; we will be measured only by what *actually happens*. Our goal, in the end, is to ensure that the timeless history of our universe is one that is filled with whatever it is we're fighting for. For me, at least, this is the underlying driver that takes the place of guilt: Once we have learned our lessons from the past, there is no reason to wrack ourselves with guilt. All we need to do, in any given moment, is look upon the actions available to us, consider, and take whichever one seems most likely to lead to a future full of light.

