Best of LessWrong: October 2015

1. A few misconceptions surrounding Roko's basilisk
2. Two Growth Curves
3. Detach the grim-o-meter
4. Simply locate yourself
5. Have no excuses
6. Come to your terms

Best of LessWrong: October 2015

A few misconceptions surrounding Roko's basilisk

There's a new LWW page on the Roko's basilisk thought experiment, discussing both Roko's original post and the fallout that came out of Eliezer Yudkowsky banning the topic on *Less Wrong* discussion threads. The wiki page, I hope, will reduce how much people have to rely on speculation or reconstruction to make sense of the arguments.

While I'm on this topic, I want to highlight points that I see omitted or misunderstood in some online discussions of Roko's basilisk. The first point that people writing about Roko's post often neglect is:

- Roko's arguments were originally posted to *Less Wrong*, but they weren't generally accepted by other Less Wrong users.

Less Wrong is a community blog, and anyone who has a few karma points can post their own content here. Having your post show up on *Less Wrong* doesn't require that anyone else endorse it. Roko's basic points were promptly rejected by other commenters on *Less Wrong*, and *as ideas* not much seems to have come of them. People who bring up the basilisk on other sites don't seem to be super interested in the specific claims Roko made either; discussions tend to gravitate toward various older ideas that Roko cited (e.g., timeless decision theory (TDT) and coherent extrapolated volition (CEV)) or toward Eliezer's controversial moderation action.

In July 2014, David Auerbach wrote a *Slate* piece criticizing *Less Wrong* users and describing them as "freaked out by Roko's Basilisk." Auerbach wrote, "Believing in Roko's Basilisk may simply be a 'referendum on autism'" — which I take to mean he thinks a significant number of Less Wrong users accept Roko's reasoning, and they do so because they're autistic (!). But the Auerbach piece glosses over the question of how many *Less Wrong* users (if any) in fact believe in Roko's basilisk. Which seems somewhat relevant to his argument...?

The idea that Roko's thought experiment holds sway over some community or subculture seems to be part of a mythology that's grown out of attempts to reconstruct the original chain of events; and a big part of the blame for that mythology's existence lies on *Less Wrong*'s moderation policies. Because the discussion topic was banned for several years, *Less Wrong* users themselves had little opportunity to explain their views or address misconceptions. A stew of rumors and partly-understood forum logs then congealed into the attempts by people on RationalWiki, *Slate*, etc. to make sense of what had happened.

I gather that the main reason people thought *Less Wrong* users were "freaked out" about Roko's argument was that Eliezer deleted Roko's post and banned further discussion of the topic. Eliezer has since sketched out his thought process on Reddit:

> When Roko posted about the Basilisk, I very foolishly yelled at him, called him an idiot, and then deleted the post. [...] Why I yelled at Roko: Because I was caught flatfooted in surprise, because I was indignant to the point of genuine emotional shock, at the concept that somebody who thought they'd invented a brilliant idea that would cause future AIs to torture people who had the thought, had promptly posted it to the public Internet. In the course of yelling at Roko to explain why this was a bad thing, I made the further error---keeping in mind that I had absolutely

> no idea that any of this would ever blow up the way it did, if I had I would obviously have kept my fingers quiescent---of not making it absolutely clear using lengthy disclaimers that my yelling did not mean that I believed Roko was right about CEV-based agents [= Eliezer's early model of indirectly normative agents that reason with ideal aggregated preferences] torturing people who had heard about Roko's idea. [...] What I considered to be obvious common sense was that you did not spread potential information hazards because it would be a crappy thing to do to someone. The problem wasn't Roko's post itself, about CEV, being correct.

This, obviously, was a bad strategy on Eliezer's part. Looking at the options in hindsight: To the extent it seemed plausible that Roko's argument could be modified and repaired, Eliezer shouldn't have used Roko's post as a teaching moment and loudly chastised him on a public discussion thread. To the extent this didn't seem plausible (or ceased to seem plausible after a bit more analysis), continuing to ban the topic was a (demonstrably) ineffective way to communicate the general importance of handling real information hazards with care.

On that note, point number two:

- Roko's argument wasn't an attempt to get people to donate to Friendly AI (FAI) research. In fact, the opposite is true.

Roko's original argument was not 'the AI agent will torture you if you don't donate, therefore you should help build such an agent'; his argument was 'the AI agent will torture you if you don't donate, therefore we should avoid ever building such an agent.' As Gerard noted in the ensuing discussion thread, threats of torture "would motivate people to form a bloodthirsty pitchfork-wielding mob storming the gates of SIAI [= MIRI] rather than contribute more money." To which Roko replied: "Right, and I am on the side of the mob with pitchforks. I think it would be a good idea to change the current proposed FAI content from CEV to something that can't use negative incentives on x-risk reducers."

Roko saw his own argument as a strike against building the kind of software agent Eliezer had in mind. Other *Less Wrong* users, meanwhile, rejected Roko's argument both as a reason to oppose AI safety efforts and as a reason to support AI safety efforts.

Roko's argument was fairly dense, and it continued into the discussion thread. I'm guessing that this (in combination with the temptation to round off weird ideas to the nearest religious trope, plus misunderstanding #1 above) is why RationalWiki's version of Roko's basilisk gets introduced as

> a futurist version of Pascal's wager; an argument used to try and suggest people should subscribe to particular singularitarian ideas, or even donate money to them, by weighing up the prospect of punishment versus reward.

If I'm correctly reconstructing the sequence of events: Sites like RationalWiki report in the passive voice that the basilisk is "an argument used" for this purpose, yet no examples ever get cited of someone actually using Roko's argument in this way. Via citogenesis, the claim then gets incorporated into other sites' reporting.

(E.g., in *Outer Places*: "Roko is claiming that we should all be working to appease an omnipotent AI, even though we have no idea if it will ever exist, simply because the

consequences of defying it would be so great." Or in *Business Insider*: "So, the moral of this story: You better help the robots make the world a better place, because if the robots find out you didn't help make the world a better place, then they're going to kill you for preventing them from making the world a better place.")

In terms of argument structure, the confusion is equating the conditional statement 'P implies Q' with the argument 'P; therefore Q.' Someone asserting the conditional isn't necessarily arguing for Q; they may be arguing against P (based on the premise that Q is false), or they may be agnostic between those two possibilities. And misreporting about which argument was made (or who made it) is kind of a big deal in this case: 'Bob used a bad philosophy argument to try to extort money from people' is a much more serious charge than 'Bob owns a blog where someone once posted a bad philosophy argument.'

Lastly:

- "Formally speaking, what is correct decision-making?" is an important open question in philosophy and computer science, and formalizing precommitment is an important part of that question.

Moving past Roko's argument itself, a number of discussions of this topic risk misrepresenting the debate's *genre*. Articles on *Slate* and RationalWiki strike an informal tone, and that tone can be useful for getting people thinking about interesting science/philosophy debates. On the other hand, if you're going to dismiss a question as unimportant or weird, it's important not to give the impression that working decision theorists are similarly dismissive.

What if your devastating take-down of string theory is intended for consumption by people who have never heard of 'string theory' before? Even if you're sure string theory is hogwash, then, you should be wary of giving the impression that the only people discussing string theory are the commenters on a recreational physics forum. Good reporting by non-professionals, whether or not they take an editorial stance on the topic, should make it obvious that there's academic disagreement about which approach to Newcomblike problems is the right one. The same holds for disagreement about topics like long-term AI risk or machine ethics.

If Roko's original post is of any pedagogical use, it's as an unsuccessful but imaginative stab at drawing out the diverging consequences of our current theories of rationality and goal-directed behavior. Good resources for these issues (both for discussion on *Less Wrong* and elsewhere) include:

- "The Long-Term Future of Artificial Intelligence", on the current field of AI and basic questions for the field's development.
- "The Value Learning Problem", on the problem of designing AI systems to answer normative questions.
- "The PD with Replicas and Causal Decision Theory," on the prisoner's dilemma as a Newcomblike problem.
- "Toward Idealized Decision Theory", on the application of decision theory to AI agents.

The Roko's basilisk ban isn't in effect anymore, so you're welcome to direct people here (or to the Roko's basilisk wiki page, which also briefly introduces the relevant issues in decision theory) if they ask about it. Particularly low-quality discussions can

still get deleted (or politely discouraged), though, at moderators' discretion. If anything here was unclear, you can ask more questions in the comments below.

Two Growth Curves

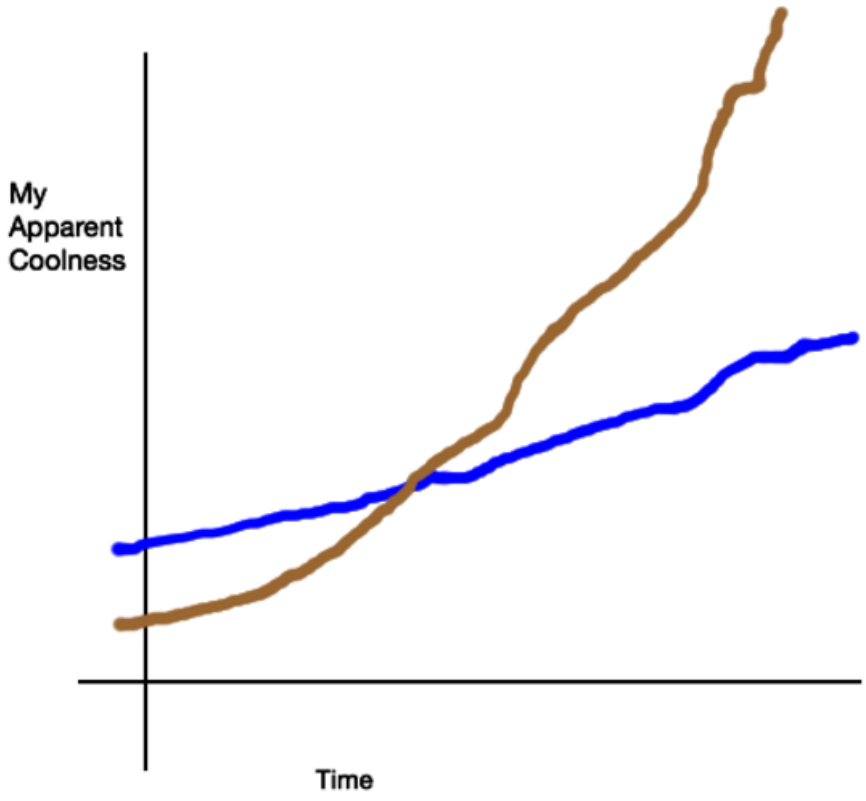
Sometimes, it helps to take a model that part of you already believes, and to make a visual image of your model so that more of you can see it.

One of my all-time favorite examples of this:

I used to often hesitate to ask dumb questions, to publicly try skills I was likely to be bad at, or to visibly/loudly put forward my best guesses in areas where others knew more than me.

I was also frustrated with this hesitation, because I could feel it hampering my skill growth. So I would try to convince myself not to care about what people thought of me. But that didn't work very well, partly because what folks think of me is in fact somewhat useful/important.

Then, I got out a piece of paper and drew how I expected the growth curves to go.



In blue, I drew the apparent-coolness level that I could achieve if I stuck with the "try to look good" strategy. In brown, I drew the apparent-coolness level I'd have if I instead made mistakes as quickly and loudly as possible -- I'd look worse at first, but then I'd learn faster, eventually overtaking the blue line.

Suddenly, instead of pitting my desire to become smart against my desire to look good, I could pit my desire to look good now against my desire to look good in the future :)

I return to this image of two growth curves often when I'm faced with an apparent tradeoff between substance and short-term appearances. (E.g., I used to often find myself scurrying to get work done, or to look productive / not-horribly-behind today, rather than trying to build the biggest chunks of capital for tomorrow. I would picture these growth curves.)

Detach the grim-o-meter

This is a linkpost for https://mindingourway.com/detach-the-grim-o-meter/

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

This is a linkpost for https://mindingourway.com/simply-locate-yourself/

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

This is a linkpost for https://mindingourway.com/have-no-excuses/

> *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 gabling 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 tolerify 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

This is a linkpost for https://mindingourway.com/come-to-your-terms/

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 jumpier. 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 unthinkably 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.)