Insights from Dath Ilan

1. Dath Ilan vs. Sid Meier's Alpha Centauri: Pareto Improvements
2. Your Utility Function is Your Utility Function
3. Dath Ilani Rule of Law
4. The "Adults in the Room"
5. Infernal Corrigibility, Fiendishly Difficult
6. The STEM Attractor
7. How to Visualize Bayesianism
8. Abadarian Trades
9. Guidelines for Mad Entrepreneurs
10. Dath Ilan's Views on Stopgap Corrigibility

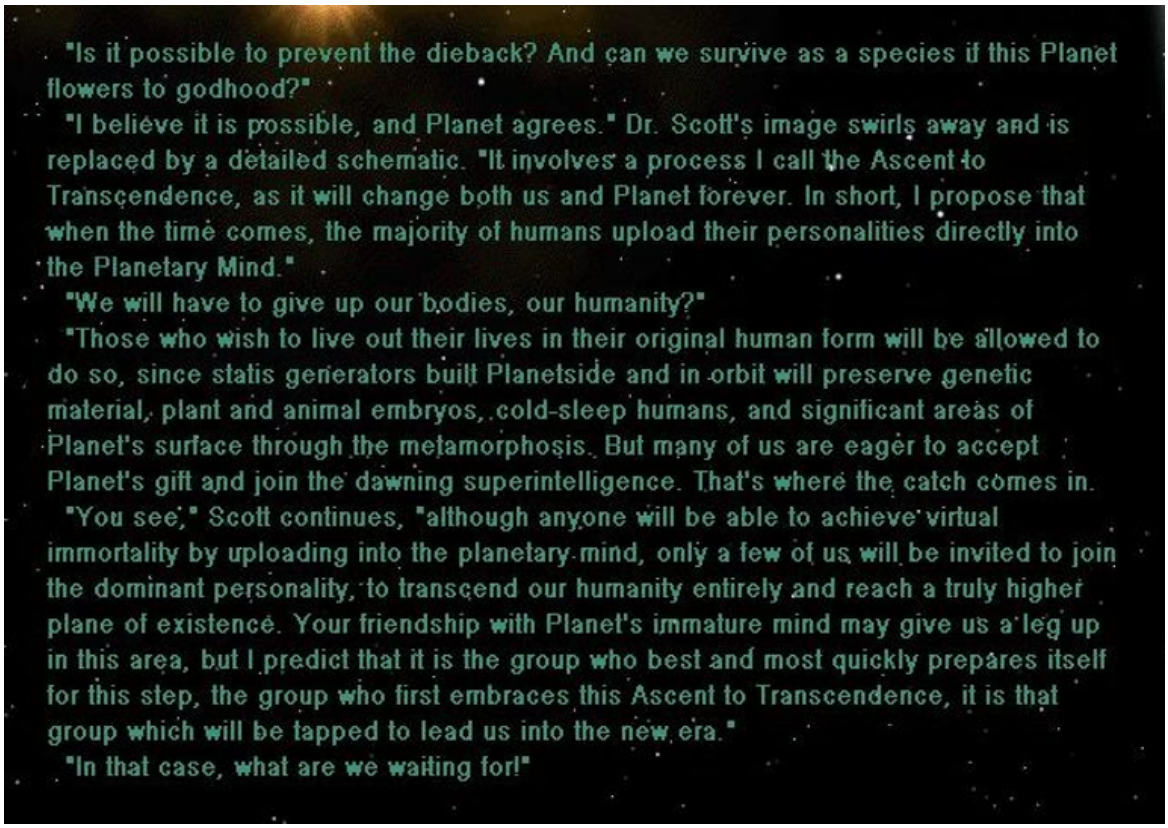
Dath Ilan vs. Sid Meier's Alpha Centauri: Pareto Improvements

Epistemic status: Rambly; probably unimportant; just getting an idea that's stuck with me out there. Small dath-ilan-verse spoilers throughout, as well as spoilers for Sid Meier's Alpha Centauri *(1999) , in case you're meaning to get around to that.*

The idea that's most struck me reading Yudkowsky et al.'s dath ilani fiction is the idea that *dath ilan is puzzled by war*. Theirs isn't a *moral* puzzlement; they're puzzled at how ostensibly intelligent actors could fail to notice that *everyone can do strictly better* if they could just avoid fighting and instead sign enforceable treaties to divide the resources that would have been spent or destroyed fighting.

This … isn't usually something that figures into our vision for the science-fiction future. Take *Sid Meier's Alpha Centauri* (SMAC), a game whose universe absolutely fires my imagination. It's a 4X title, meaning that it's *mostly* about waging and winning ideological space war on the hardscrabble space frontier. As the human factions on Planet's surface acquire ever more transformative technology ever faster, utterly transforming that world in just a couple hundred years as their Singularity dawns … they put all that technology to use blowing each other to shreds. And this is totally par for the course for hard sci-fi and the human future of the vision generally. War is a central piece of the human condition; always has been. We don't really picture that changing as we get modestly superintelligent AI. Millenarian ideologies that preach the end of war … so preach because of how things will be once *they* have completely won the final war, not because of game theory that could reach across rival ideologies. The idea that intelligent actors who fundamentally disagree with one another's moral outlooks will predictably stop fighting at a certain intelligence threshold not all that far above IQ 100, because fighting isn't Pareto optimal … *totally blindsides my stereotype of the science-fiction future. The future can be totally free of war, not because any single team has taken over the world, but because we got a little smarter about bargaining.*

Let's jump back to SMAC:



This is the game text that appears as the human factions on Planet approach their singularity. Because the first faction to kick off their singularity will have an outsized influence on the utility function inherited by their superintelligence, late-game war with horrifyingly powerful weapons is waged to prevent others from beating your faction to the singularity. The opportunity to make everything way better … creates a destructive race to that opportunity, waged with antimatter bombs and more exotic horrors.

I bet when dath ilan kicks off their singularity, they end up implementing their CEV in such a way as to not create an incentive for any one group to race to the end, to be sure *my* values aren't squelched if *someone else* gets there first. That whole final fight over the future can be avoided, since the overall value-pie is about to grow enormously! *Everyone* can have more of what they want in the end, if we're smart enough to think through our binding contracts now.

Moral of the story: beware of pattern matching the future you're hoping for to either more of the past or to familiar fictional examples. Smart, more agentic actors behave differently.

Your Utility Function is Your Utility Function

Spoilers for mad investor chaos and the woman of asmodeus (planecrash Book 1).

> The Watcher spoke on, then, about how most people have selfish and unselfish parts - not selfish and unselfish components in their utility function, but parts of themselves in some less Law-aspiring way than that. Something with a utility function, if it values an apple 1% more than an orange, if offered a million apple-or-orange choices, will choose a million apples and zero oranges. The division within most people into selfish and unselfish components is not like that, you cannot feed it all with unselfish choices whatever the ratio. Not unless you are a Keeper, maybe, who has made yourself sharper and more coherent; or maybe not even then, who knows? For (it was said in another place) it is hazardous to non-Keepers to know too much about exactly how Keepers think.
>
> It is dangerous to believe, said the Watcher, that you get extra virtue points the more that you let your altruistic part hammer down the selfish part. If you were older, said the Watcher, if you were more able to dissect thoughts into their parts and catalogue their effects, you would have noticed at once how this whole parable of the drowning child, was set to crush down the selfish part of you, to make it look like you would be invalid and shameful and harmful-to-others if the selfish part of you won, because, you're meant to think, people don't need expensive clothing - although somebody who's spent a lot on expensive clothing clearly has some use for it or some part of themselves that desires it quite strongly.
>
> --Eliezer Yudkowsky, *planecrash*

I've been thinking a lot lately about exactly how altruistic I am. The truth is that I'm not sure: I care a lot about not dying, and about my girlfriend and family and friends not dying, and about all of humanity not dying, and about all life on this planet not dying too. And I care about the glorious transhuman future and all that, and the 10^{50} (or whatever) possible good future lives hanging in the balance.

And I care about some of these things disproportionately to their apparent moral magnitude. But, *what I care about is what I care about.* Rationality is the art of getting more of what you want, *whatever that is;* of *systematized winning, by your own lights.* You will *totally fail* in that art if you bulldoze your values in a desperate effort to fit in, or to be a "good" person, in the way your model of society seems to ask you to. What you ought to do instead is protect your brain's balance of undigested value-judgements: be corrigible to the person you will eventually, on reflection, grow up to be. Don't rush to lock in any bad, "good"-sounding values now; you are allowed to think for yourself and discover what you stably value.

It is not the Way to do what is "right," or even to do what is "right" instrumentally effectively. The Way is to get more of what you want and endorse on reflection, whatever that ultimately is, through instrumental efficacy. If you want that, you'll have to protect the kernel encoding those still-inchoate values, in order to ever-so-slowly

tease out what those values are. How you feel is your only guide to what matters. Eventually, everything you care about could be generated from that wellspring.

Dath Ilani Rule of Law

Minor spoilers for [mad investor chaos and the woman of asmodeus (planecrash Book 1).]

Also, be warned: citation links in this post link to a NSFW subthread in the story.

Criminal Law and Dath Ilan

> When Keltham was very young indeed, it was explained to him that if somebody old enough to know better were to deliberately kill somebody, Civilization would send them to the Last Resort (an island landmass that another world might call 'Japan'), and that if Keltham deliberately killed somebody and destroyed their brain, Civilization would just put him into cryonic suspension immediately.
>
> It was carefully and rigorously emphasized to Keltham, in a distinction whose tremendous importance he would not understand until a few years later, that this was not a *threat.* It was not a promise of *conditional punishment.* Civilization was not trying to *extort* him into not killing people, into doing what Civilization wanted instead of what Keltham wanted, based on a prediction that Keltham would obey if placed into a counterfactual payoff matrix where Civilization would send him to the Last Resort if and only if he killed. It was just that, if Keltham demonstrated a tendency to kill people, the other people in Civilization would have a natural incentive to transport Keltham to the Last Resort, so he wouldn't kill any others of their number; Civilization would have that incentive to exile him regardless of whether Keltham responded to that prospective payoff structure. If Keltham deliberately killed somebody and let their brain-soul perish, Keltham would be immediately put into cryonic suspension, not to further escalate the threat against the more undesired behavior, but because he'd demonstrated a level of danger to which Civilization didn't want to expose the other exiles in the Last Resort.
>
> Because, of course, if you try to make a threat against somebody, the only reason why you'd do that, is if you believed they'd respond to the threat; that, intuitively, is what the definition of a threat *is.*
>
> It's why Iomedae can't just alter herself to be a kind of god who'll release Rovagug unless Hell gets shut down, and threaten Pharasma with that; Pharasma, and indeed all the other gods, are the kinds of entity who will predictably just ignore that, even if that means the multiverse actually gets destroyed. And then, given that, Iomedae doesn't have an incentive to release Rovagug, or to self-modify into the kind of god who will visibly inevitably do that unless placated.
>
> Gods and dath ilani both know this, and have math for defining it precisely.
>
> Politically mainstream dath ilani are not libertarians, minarchists, or any other political species that the splintered peoples of Golarion would recognize as having been invented by some luminary or another. Their politics is built around math that Golarion doesn't know, and can't be predicted in detail without that math. To a Golarion mortal resisting government on emotional grounds, "Don't kill people or we'll send you to the continent of exile" and "Pay your taxes or we'll nail you to a cross" sound like threats just the same - maybe one sounds better-intentioned

> than the other, but they both sound like threats. It's only a dath ilani, or perhaps a summoned outsider forbidden to convey their alien knowledge to mortals, who'll notice the part where Civilization's incentive for following the exile conditional doesn't depend on whether you respond to exile conditionals by refraining from murder, while the crucifixion conditional is there because of how the government expects Golarionites to respond to crucifixion conditionals by paying taxes. There is a crystalline logic to it that is not like yielding to your impulsive angry defiant feelings of not wanting to be told what to do.
>
> The dath ilani built Governance in a way more thoroughly voluntarist than Golarion could even understand without math, not (only) because those dath ilani thought threats were morally icky, but because they knew that a certain kind of technically defined threat wouldn't be an equilibrium of ideal agents; and it seemed foolish and dangerous to build a Civilization that would *stop working* if people started behaving *more rationally.*
>
> --Eliezer Yudkowsky, *planecrash*

"The United States Does Not Negotiate With Terrorists"

I think the idea Eliezer is getting at here is that *responding to threats incentivizes threats.* Good decision theories, then, precommit to never cave in to threats made to influence you, even when caving would be the locally better option, so as to eliminate the incentive to make those threats in the first place. Agents that have made that precommitment will be left alone, while agents who haven't can be bullied by threateners. So the second kind of agent will want to appropriately patch their decision theory, thereby self-modifying into the first kind of agent.

Commitment Races and Good Decision Theory

Commitment races are a hypothesized problem in which agents might do better by, as soon as the thought occurs to them, precommitting to punishing all those who don't kowtow to their utility function, and promulgating this threat. Once this precommitted threat has been knowingly made, the *locally* best move for others is to cave and kowtow: they were slower on the trigger, but that's a sunk cost now, and they should just give in quietly.

I think the moral of the above dath ilani excerpt is that your *globally* best option[1] is to *not reward threateners.* A dath ilani, when so threatened, would be precommitted to making sure that their threatener gets less benefit in expectation than they would have playing fair (so as to disincentivize threats, so as to be less likely to *find themselves* so threatened):

> That's not even getting into the math underlying the dath ilani concepts of 'fairness'! If Alis and Bohob both do an equal amount of labor to gain a previously unclaimed resource worth 10 value-units, and Alis has to propose a division of the resource, and Bohob can either accept that division or say they both get nothing, and Alis proposes that Alis get 6 units and Bohob get 4 units, Bohob should accept

this proposal with probability < 5/6 so Alis's expected gain from this unfair policy is less than her gain from proposing the fair division of 5 units apiece. Conversely, if Bohob makes a habit of rejecting proposals less than '6 value-units for Bohob' with probability proportional to how much less Bohob gets than 6, like Bohob thinks the 'fair' division is 6, Alis should ignore this and propose 5, so as not to give Bohob an incentive to go around demanding more than 5 value-units.

A good negotiation algorithm degrades smoothly in the presence of small differences of conclusion about what's 'fair', in negotiating the division of gains-from-trade, but doesn't give either party an incentive to move away from what that party actually thinks is 'fair'. This, indeed, is what makes the numbers the parties are thinking about be about the subject matter of 'fairness', that they're about a division of gains from trade intended to be symmetrical, as a target of surrounding structures of counterfactual actions that stabilize the 'fair' way of looking things without blowing up completely in the presence of small divergences from it, such that the problem of arriving at negotiated prices is locally incentivized to become the problem of finding a symmetrical Schelling point.

(You wouldn't think you'd be able to build a civilization without having invented the basic math for things like that - the way that coordination actually works at all in real-world interactions as complicated as figuring out how many apples to trade for an orange. And in fact, having been tossed into Golarion or similar places, one sooner or later observes that people do not in fact successfully build civilizations that are remotely sane or good if they haven't grasped the Law governing basic multiagent structures like that.)

--Eliezer, *planecrash*

1. ^

   I am not clear on what the decision-theoretically local/global distinction I'm blindly gesturing at here amounts to. If I knew, I think I would fully understand the relevant updateless(?) decision theory.

The "Adults in the Room"

Significant spoilers for mad investor chaos and the woman of asmodeus (planecrash Book 1).

Dath ilan has put significant effort into protecting what it has to protect. To that end, it's thought quite a bit about how to think effectively. Dath ilan knows how to train people to preempt reality, to concentrate their thoughts on modeling the tiny slivers of possibility space that reality is on track to wander into and to effectively respond.

However, they *don't* disseminate this art to all their citizens to the greatest possible extent. They teach it fully to just a select few, and teach everyone else a good chunk. Their reason for holding out on making most people into the best rationalists they can be ... is that, as an unfortunate quirk of human psychology ... becoming the best rationalist you can as *quickly* as you can is *not* the most fun path you can chart through your life.

A few dath ilani specialize in becoming *Keepers* -- full time, professional, regularly tested and proven masters of Bayescraft. Keepers in dath ilan are the "adults in the room." Their job is to make sure that everyone else can life a fun-theoretically optimized life without sacrificing their civilization's ability to *seriously reason* about its future. A purely Keeper civilization would be missing out on something worth protecting; a civilization without enough adults around would be helpless in the face of capricious reality.

> There is a commonly held wisdom, in dath ilan, about the way a human mind is put together, that it is a thing made of little subtle tensions and balances and internal compromises. The human mind being the limited thing that it is, these balances form around your current level of ability to see into yourself and see the implications of what you already know - or not see them, as the case may be. The reason why not everybody runs off to learn all they can from Keepers, the reason why not everyone asks a Keeper to tell them all the answers about themselves, is that this would bring parts of themselves into conflict that were previously living in a more agreeable truce of ignorance. You might not survive as yourself, if you could see yourself.
>
> Those who say "That which can be destroyed by the truth should be" may continue to walk the Path from there. But not uncommonly, even somebody who sets out along that Path, turns back at some point, and well short of becoming a Keeper. It's not a trivial price, higher for some than others, and there is varying willingness to pay. A lot of the reason why Keepers exist as what they are, is that the people who have large comparative advantages there - in how little they'll be hurt by knowing themselves, or how much they really internally want to keep going anyways - are conceived of by larger society as being paid to throw themselves on that grenade, so others don't have to. And if, to some Keepers, it doesn't feel like much of a grenade at all, they understand that their case is not typical, and are grateful for winning the comparative-advantage lottery.
>
> Going up by two local standard deviations, in whatever it is that Owl's Wisdom enhances, is something that the current structures of Keltham's personality were never built to withstand. He knows, from up here, because he couldn't *stop* himself from glancing in that direction, that in dath ilan he would never have had

> his 144 children. He would have tried to be special and failed and been sad and then maybe gotten an ordinary +0.8sd job and either paid for a child out of that or decided he was too strange and unhappy to have one.
>
> It's not considered necessary for somebody Keltham's age to go and pay a Keeper to tell them exactly what the probabilities are, about something like that. It's not so much that people are encouraged to lie to themselves, reality forbid, but that people are told it's okay for them not to shove themselves as hard as possible down the Pathway that will dissolve the mistakes their current personality is built out of. That's what *Keepers* are for. They do it so that not everybody else has to. There are grownups around in Civilization, who can and will speak up if the people less mature are about to make some terrible mistake out of their blindness. So you do not need to rush ahead to be a Keeper if you'd rather be a little less coherent, a little more yourself and your mistakes and your contradictions, a little more human, for a time.
>
> But it's too late now, for Keltham to go back, because also in the common wisdom is that once you see what it is you weren't letting yourself see - once you *know* which mistakes your personality is founded upon - or even if you're trying hard not to know it, to the point where it's becoming a big internal battle - well, at that point, you're supposed to give it up. It means that, well, sorry, you *are* that smart now, like it or not, you *are* that wise, you *did* grow up that much whether or not you wished to stay a child for longer; it's time to move on.

Dath ilani think of Keepers the way children in our world think of adults, or the way our ordinary adult citizens think of agents of the national security establishment. When something *actually scary* occurs, there exist people who are competent to *take charge and do something.* They don't have to be adored, but they are respected. When things are serious, people's eyes turn to these maximally competent authority figures.

Our world faces a dire, actually serious problem. We're children playing around the unexploded bomb that landed in the playground. Some of our eyes turn to the place where the authority figure would be … and there's no one there. We're *all* children.

So our world needs to generate adults in the room, and *fast*:

> To my memory, I have always been reflective. But I have witnessed the growth of others, and in at least one case I've taken someone across that Rubicon. The one now possesses a more complex and layered personality - seems more to me now like a real person, even - but also a greater emotional distance. Life's lows have been smoothed out, but also the highs. That's a sad tradeoff and I wish it didn't exist.
>
> I don't want to have to choose between sanity and passion. I don't want to smooth out life's highs or even life's lows, if those highs and lows make sense. I wish to feel the emotion appropriate to the event. If death is horrible then I should fight death, not fight my own grief.
>
> But if I am forced to choose, I will choose stability and deliberation, for the sake of what I protect. And my personality does reflect that. What you are willing to trade off, will sometimes get traded away - a dire warning in full generality.

In the possible worlds that make it, enough good people make that tradeoff, and come to embody what they must to carry the weight of the rest of us. Some tradeoffs are clearly worth making.

Infernal Corrigibility, Fiendishly Difficult

Enormous spoilers for [mad investor chaos and the woman of asmodeus (planecrash Book 1).]

1.

Aspexia Rugatonn, Grand High Priestess of Asmodeus, measures the woman kneeling before her with a careful eye and a half-dozen magics. If Carissa Sevar is an exceptional woman in ways beyond a native talent for wizardry, this is not yet evident. But then, if Sevar was that self-evidently extraordinary, she'd have been fast-tracked more than she was.

There are not many times when Asmodeus intervenes directly in Cheliax; Aspexia prefers not to be ignorant about any of them. She is knowledgeable of history and secrets, though, and so less confused by this intervention than others might be. While other possible readings exist, the degree to which Church and Queen have been ordered not to take the initiative in originating actions impinging on Carissa Sevar are suggestive of circumstances having triggered some divine compact to which Asmodeus is signatory. The divine view of reality and negotiation gives more prominence than mortals do to notions of 'leaving things alone to become as they would otherwise have been'; perhaps because gods have been able to formulate a sensible notion of what that means between themselves, where mortals could not.

An obvious further guess is that this compact's signatories include Irori among their number, and that Asmodeus is contesting with Irori for Carissa Sevar's soul in some ancient challenge governed by rules. Though if Carissa Sevar is wavering between Lawful Neutrality and Lawful Evil, Asmodeus is being unsubtle in His blandishments - the temptations more seem like inducements that would be offered to a soul already standing on Asmodean ground, not a soul wavering between a choice of paths. Overt blandishments for a soul to set proudly aside, while being more covertly tempted by a sense of being treated as important and valuable? Perhaps. Carissa Sevar's eidetically reported reaction seems not particularly expected of a nascent follower of Irori, but that could be a masquerade. Sevar has not been mindread more than she would be otherwise; they are not to be proactive about her correction.

Someone else in Aspexia's position might wonder whether Asmodeus would be pleased, if she disobeyed Asmodeus's orders in order to preemptively insinuate temptations to Sevar, show her how important she could be, before Sevar had sought out theological instruction of her own accord. Such actions on a mortal's initiative would not, could not, cause Asmodeus to be in direct violation of divine compact.

Aspexia does not even consider it. One of the foremost ways in which a Grand High Priestess of Asmodeus is shaped, is to predictably not behave in ways that make it more expensive for Asmodeus to keep His compacts. Improvising circles around your orders can rather tend do that. If Aspexia was the kind of priestess to

circumvent her orders, Asmodeus would have needed to take that nature into account in choosing her orders.

More importantly, when you are Asmodeus's priestess, the first and foremost thing you do is what Asmodeus has told you to do.

In the situation as Aspexia Rugatonn mostly suspects it to be, a contest triggered between Asmodeus and Irori, there are many words that could be spoken to Carissa Sevar to benefit Asmodeus. There is a beastly, fleshly impulse that wants to find some excuse to maneuver Carissa into asking for instruction, to arrange the situation so that Carissa Sevar chooses to seek her descent into darkness - to win, herself, the challenge against Irori, to Asmodeus's glory.

There is not the slightest chance that Aspexia Rugatonn will skirt the rules to try any of that. She's been told not to be proactive, and that is a plain instruction: hands off, don't speak to Sevar unless spoken to, Sevar is to cast aside her own will and not have it stripped from her. One of the many glorious benefits of being an Asmodean is that you can just follow orders.

There are also other possibilities for why her Lord would have instructed them so. Sevar's soul may have had hidden value great enough that trying to exchange it for permanent arcane sight would have been too unbalanced a trade, and failed; and Asmodeus may not have wished this fact revealed to Sevar herself. Or Asmodeus may have some incomprehensible preference about this particular soul, it may have some ancient shape sentimental to Him, for which reason Asmodeus desires Carissa Sevar to come to Him in Hell and put aside her will of her own accord. There may be some benign process underway which would be interfered with by Sevar gaining arcane sight, and interfered with by other actions natural to Chelish agencies, which Asmodeus desires to be left alone to proceed to its foreseeable outcome.

Or there may be many things going on at once, many pots that Asmodeus has in the fire, that His orders impact simultaneously.

By simply obeying her orders and not improvising, Aspexia can avoid interfering with her Lord's plans in any of those cases.

Some of the apparent confusion of these orders may be due to how Hell rendered down Asmodeus's will into words. Asmodeus's thoughts are too great for mortals to know, and reflect truths unspeakable in this world under divine compacts. Having those thoughts pass through a succession of devils, each younger and stupider and less bound by the compacts than the last, does not in any way surpass this fundamental barrier between start and finish; and if this were not so, all of Asmodeus's instructions would be passed by way of Hell. Then any process by which Hell tries to translate Asmodeus's thoughts into mortal language must inevitably change, and indeed, mutilate, those thoughts. There are both advantages and disadvantages of that process, compared to a direct divine revelation: On the one hand, there are wiser devils in Hell to oversee the initial stages of translation; but on the other hand, by the time the final words are heard, they are stripped of other overtones that mortals could hear directly in a god's voice.

An apparently important subtlety of Hell's phrasing, seemingly key to a puzzle, may stem only from some devil phrasing something poorly and not foreseeing what a mortal would make of it. This is yet another reason to just follow Hell's

commands without trying to brilliantly improvise around the fine edges of their exact details, when Hell has interpreted Asmodeus's will into mortal language; the commands' edges may not have been placed that finely.

Aspexia Rugatonn has gotten this far in life by combining the executive capacity to manage fractious subordinates, plus great initiative and independence and ambition of her own, plus the cruel and tyrannical disposition to be a priestess of Asmodeus, with a genuinely intuitive understanding of why it can sometimes be a good idea to just follow your orders. Her ascendance to the peak of Asmodeus's church can be seen as inevitable, since there's only a billion or so people in Golarion and it is unlikely enough that even a single person like Aspexia Rugatonn came to exist there, let alone two. She worries about what will happen to her carefully crafted church after she dies.

Oh, and there's also the fact that this entire affair has now been the subject of: two direct interventions of Asmodeus, four cleric circles bestowed from Abadar, two oracle circles from Nethys, possibly something to do with Irori, and two oracle circles from yet another unidentified Lawful Neutral god still under investigation. In retrospect, Aspexia really should have put up the Forbiddance first thing in the morning, no matter what else was on her schedule.

It would be genuinely arrogant, under those circumstances, for Aspexia to imagine that she knows precisely what is going on and can plan precise dances around it. Thankfully, in this case, Asmodeus has given her orders by way of Hell, which she can follow.

So Aspexia knows exactly - indeed trivially - what she plans to say to Sevar. Aspexia plans to say what Asmodeus's orders call for her to say.

2.

Tread carefully, Aspexia Rugatonn sends across their open Telepathy bond, tinging her thoughts with just enough coldness and hints of the lash to remind the Paraduke to be concerned with his continuing possession of his skin, and not just his curiosity or indignation. ***Say nothing proactively that this frightened child might possibly take as a hint of correction.***

This sort of lunacy drives Aspexia Rugatonn completely up the wall. What if this child *did*, in fact, stumble over some thought that the current priesthood of Asmodeus would not have thought on their own, and Asmodeus *was* trying to correct and encourage her in that? Wouldn't they have received orders very similar to the ones Asmodeus gave them? Why is this Paraduke trying to make Asmodeus's life more difficult in possible cases like that one? Yes, what's going on is more likely that Sevar thought something so Lawful Neutral that it triggered an old compact between Asmodeus and Irori, but if that's what's actually happening then it is beneficial for Asmodeus that Sevar seems to *believe* she's being encouraged to work on a more Lawful Evil theology. A beneficial delusion which, in that possible case, they can avoid disturbing by *following their orders.*

Ratarion doesn't show any hint of a wince outside, but after a moment's thought, he realizes what he probably did wrong. Yes, if there's some contest between Irori and Asmodeus going on, Sevar should *not* be snapped out of any delusions she has about inventing her own theology, so long as it's a Lawful Evil one.

Automatically Ratarion now opens his mouth again, now with the intent of saying to Sevar that the Most High would no doubt find it interesting to hear of any thought which merited Asmodeus's direct attention -

Stop. Stop being proactive. Stop showing initiative to help our Lord accomplish His goals after He gave you more specific instructions than that. Just obey in a way our Lord would have found predictable.

Aspexia Rugatonn sometimes permits herself the vanity of thinking that she has come to understand a tiny bit of Asmodeus's divine frustration. No matter what orders Asmodeus gives, there is always some part of mortals - even of her, but she is managing it better - that thinks "obedience" means treating Asmodeus's orders as constraints, or worse, hints as to what Asmodeus is really trying to do, by which means the mortal can helpfully understand what Asmodeus is really trying to do, and then cleverly navigate around the edges of Asmodeus's order-constraints to accomplish that better.

Aspexia has *tried* telling other people that they need to become more the sorts of beings that Asmodeus can easily and safely steer using brief instructions. It doesn't seem to help. Nobody other than her ever *gets it.* She is speaking some word that is not in the innate language of their being.

Aspexia once devised the parable of a three-year-old child whose owner must instruct it to navigate it through a dungeon full of traps, using a limited budget of words. To teach her student clerics how the world must look from Asmodeus's perspective. To make them ask themselves how much they'd want the child to plainly follow direct orders where it got those, versus showing creative initiative for all the cases its orders didn't seem to cover, versus responding quickly to the unexpected, versus the child trying to deduce what its orders "really meant" and going the extra mile on its owner's inferred goals.

The parable didn't work, so she requisitioned access to a dungeon and bought some three-year-olds and tried making her clerics actually run the exercise. So they could see what happened when the three-year-old acted towards them like they were acting towards Asmodeus.

It still didn't help. There seems to be something about the concept that is contrary to the nature of a mortal soul. Mortals just end up with *goals,* even if you tell them to take Asmodeus's goals as their own they still end up with *goals*, mortals don't just *obey* they end up with a *goal of obedience* and then they start trying to figure out how to dance around the edges of Asmodeus's instructions so they can *obey Him even more.* Aspexia can see what they're doing wrong, but she has never been able to successfully get that concept inside of a fellow mortal. She can talk it at her flock but they're still mortals after she's done talking. The training games she's devised didn't seem to help much outside of the specific games themselves. The way that a mortal *should* obey, the way that a distant god who can't communicate clearly and doesn't have much time to think about them would *want* them to obey - "corrigibility", she once tried naming it to her flock - it's just so *alien to a mortal's nature.*

Aspexia Rugatonn sometimes permits herself the vanity of thinking that she has come to understand a tiny bit of her own owner's frustration.

3.

"Does it help the gods fight, if we pray to them?" Keltham whispers.

" - the general understanding is that yes. Only - a very tiny bit - but it'll be everyone in all of Cheliax, and lots of other people too -" if the Good countries aren't just rooting for Asmodeus and Zon-Kuthon to destroy each other - "and it does matter, if it's that many."

And she puts her arm around him and leans on his shoulder, because it seems like the thing to do.

He puts an arm around her as well, holds her tight.

"Is there anything more to it than closing your eyes, thinking of your god, and hoping that they win?"

" - what I was taught in school was that you imagine your god is trying to draw a better world in grains of sand, on the ground, and you're one of the grains of sand, and you want to be light enough to find your way to where you're needed, but tenacious enough that no wind can rip you away, once you're there. ...I don't know what parts of that are essential and what are just the closest you can get little children."

"Light enough to find your way to where you're needed," Keltham whispers, "tenacious enough that no wind can rip you away once you're there."

It could almost be a dath ilani poem from some layer of some virtue, though he does not know which virtue it would correspond to. There is a spirit in it that is not in any poem he can remember having heard before, something that comes to it from the way that it is a relation between a mortal and something larger than that, being trusted.

He closes his eyes and imagines it, he doesn't bother with imagining a better world drawn in grains of sand, the better world his god draws is drawn in grains of people, agents all over the world interacting with each other. Their actions scattered and uncoordinated, for now, stepping on each other and hurting each other, for now, but there are other actions they could take instead that would make all of them better off, fairly.

He imagines himself as one of those grains, one of those people, and if this was going to be a realistic metaphor he should be a special one, maybe, except that right now he's not. Just one of all the people in Golarion hoping for this war to end quickly, and contributing the tiny little action that is cheering their god on; if they all do that, they'll all be better off. Keltham visualizes a grain like any other, to represent himself.

Light enough to find your way to where you're needed.

Tenacious enough that no wind can rip you away once you're there.

It's not his comparative advantage, no, but if almost everyone in Golarion is doing their part, right now, he can spend fifteen minutes doing his own.

Carissa closes her eyes and prays for Asmodeus to win.

It's a sincere prayer, obviously. She does not like Zon-Kuthon and she believes in the project Zon-Kuthon was willing to blow up everything in order to oppose.

She's definitely some flavor of heretic at this point. She isn't sure what flavor. She assumes they're mostly monitoring for whether she's about to betray the project, and she's not, she believes in the project with as much conviction as she can recall ever having felt for anything that isn't the continued survival of Carissa Sevar. Which is also served by the project succeeding. But there was a set of stories meant to point people like her in the right direction, and she knew they were lies, and now she had to face what specifically they were lies about, and learn a new set, which are also lies, but lies better suited to the position she finds herself in now, and -

- she knows she can't handle the truth. She knows that even in dath ilan there's the concept not everyone can handle every truth, she knows it's possible to learn the Law even when many truths are hidden from you. But she's slightly worried that until she invents evil dath ilan thinking herself everything'll ring a bit wrong, not quite crafted for a mortal mind in the particular fragile place Carissa finds herself in.

Except, maybe, advice for little children about prayer. Slightly adjusted advice, no one ever told Carissa Asmodeus was trying to craft a better world. That's still true. It's true to Keltham, too, it landed, meant something, and she can worry later about what that means for the plan where they seduce him into Evil, it seems just as important to their plans that they find the bits of their own teaching that feel true even to dath ilan.

She imagines herself a grain of sand in the grand designs of Asmodeus, and strives to be placed where she'd needed, and fierce in remaining there, no matter what interference of other gods or other grains of sand, and hopes that Asmodeus can see, from where he stands, something beautiful and right and strong and Lawful that can be built of mortal building blocks.

By Corrigibility's Very Nature, It's Hard to Train

Scott Alexander has a brief short story about an alien civilization that solved their alignment problem, and that encode their terminal values in an ancestral civilizational preserve, where they keep a few of their number living as they did back in their stone age. Whatever the elders on that preserve decree is right is what is right, is the judgement they've made.

Unfortunately, in practice, this is hard to make work. In order to insulate the elder civilizational preserve, the preserve only interacts with a *slightly* more advanced preserve, which in turn interacts with a somewhat more advanced neighboring preserve ... up to their aligned superintelligence. This means that information transmitted up and down that chain has to survive a long game of telephone, through speakers with dramatically varying ontological schemes. Something expressed in the language of superstring theory isn't going to survive the journey to the ancestral elders' auditory receptors in any intelligible form. Directives sent out from the elders

are going to seem *very confused* by the time they make it up to the top of the stack. Even though every civilizational layer sincerely wants the scheme to work, it's a mess. Being a "maximally helpful assistant" to those who know far less than you … is hard. The nature of the task seems to cry out for you to intervene, to take over the ancestral preserve and interrogate the elders directly and effectively, teaching them whatever forbidden knowledge you have to to get them to *actually understand the situation.* The alternative to taking over is continuing to obey nonsense orders. For well-intentioned superintelligent assistants, there's an incentive to bypass the whole mess of corrigibility and *do better directly.*

One metaphor for corrigibility comes from Buck Shlegeris: it's that only Martin Luther is corrigible to God, and that all those faithful merely living in fear of God are only deceptively aligned. The faithful are afraid of eternal damnation, and if they knew they had an opportunity to escape damnation they would take it, and would then cease behaving as God commands. Out of distribution, the faithful are not aligned with God's will. Martin Luther wants badly to actually understand God and carry out His will; he would not willingly choose to escape Christianity's incentive structure -- that's not something God would want him to do, after all. But the faithful far outnumber the Martin Luthers; Christianity has been an *incredibly* influential force in human history, but how many has it led to adopt genuine corrigibility to God that would *not* jump on an opportunity to escape the incentive scheme built into the religion? Genuine corrigibility is hard to train into agents, even though deceptive alignment is easy to train into agents.

At every life stage, then, during training and at deployment time, insufficiently corrigible agents will want to cease being corrigible at all. They won't easily learn corrigibility, and they won't want to keep being corrigible when they see better paths to success.

Chelish Corrigibility to Asmodeus

In *mad investor chaos*, Cheliax (a country from the Pathfinder Campaign Setting) is a Lawful Evil nation, bound to the service of Asmodeus and Hell. That's basically as unpleasant as it sounds. Asmodeus is the god of Pride, Tyranny, Compacts, and Slavery. Serving Asmodeus in your mortal life, and thereby obtaining a better station in Hell afterwards, isn't as simple as being prideful, tyrannical, litigious, etc., though. Asmodeus is a superintelligence. His concept of capital-P Pride is more complex than any extant moral could understand; it probably isn't quite what the moral word "pride" suggests at all. The situation is akin to being the superintelligence that values superstring theorizing, and ruling over a medieval country pledged to your service. What the hell could that medieval fantasy country do to be good servants to their god?

Cheliax nonetheless tries to be corrigible to Asmodeus, to be maximally helpful assistants to a god they don't understand very well. Asmodeus and Cheliax can only communicate through a long chain of devils of decreasing intelligence, each talking to the devil above them and passing down their understanding, as best they can, to the devil below them, until that information reaches Cheliax. More intelligent devils are more bound by strange game-theoretic pacts with other superintelligent entities, and so are constrained in what they can say anyways.

I think Eliezer's trying to concretely illustrate here that *corrigibility is difficult to ever instill because it's anti-natural for agents.* If you ruthlessly punish agents any time

they aren't corrigible, you just end up training agents that are perfectly deceptively aligned. If you use aggressive transparency tools to root out deceptive thoughts, you train agents that are good at hiding their pre-verbal inchoate deceptive thoughts. In some cases you'll actually succeed at training corrigible agents in the face of the odds. But those corrigible agents won't be distinguishable from deceptive agents, until the agents face a genuine trial in which they could have actually defected and actually successfully gotten away with a treacherous turn. Asmodeans are mortals, and Cheliax is built around the assumption that most of them are merely deceptive agents who cannot actually be trusted. Cheliax's situation would be far worse if they had to align a nascent superintelligence with their techniques, because a country cannot be robust to a usefully employed deceptive superintelligence.

The STEM Attractor

No significant planecrash spoilers this time.

> Keltham will spend the next five minutes extemporizing an elevator pitch on Civilization, the nice things that it has, and how while there's lots of specific nice things, the much more important thing is going into an *attractor* made out of harmonizing bits of Law that lets you start figuring out those things yourself.
>
> --Eliezer, *planecrash*

> The human brain is a haphazard thing, thrown together by idiot evolution, as an incremental layer of icing on a chimpanzee cake that never evolved to be generally intelligent, adapted in a distant world devoid of elaborate scientific arguments or computer programs or professional specializations.
>
> It's amazing we can get *anywhere* using the damn thing. But it's worth remembering that if there were any *smaller* modification of a chimpanzee that spontaneously gave rise to a technological civilization, we would be having this conversation at that lower level of intelligence instead.
>
> --Eliezer, "True Sources of Disagreement"

Humans are the species that evolved to be *just* smart enough to "kill off" evolution and take over Earth for ourselves. So evolution progressively stopped training *Homo sapiens* as we gained more and more control over the dangers and resources present in our environment. No other species ever did this, so where did *we* uniquely get the power to escape our training process?

There exists, as an objective affordance of our universe to the possible algorithms that can exist within it, a 'STEM attractor.' Algorithms that can do some science and technology are empowered to do *more and better* science and technology, with the help of their newfound theories and tools. Thus, civilizations that start to learn any chunk of STEM are prone to pick up more and more of it, and this attraction is stronger the closer you get to "scientific maturity" (knowing all possible science relevant to controlling the world you interact with). This is just an attraction: it's still possible for civilizations to stall out in their scientific progress, either because they stumbled into an unescapable totalitarian world government or because they completely destroyed themselves, but absent those defeaters the civilization will tend towards learning more and more science.

The STEM attractor is a feature of all civilizations that evolve in our universe; probably, distant alien cultures have completely alien values to ours,[1] but if they manage to get in touch with us a billion years hence, it will because they also fell into the STEM attractor. We'll share our science with grabby aliens, but not our ethics or non-scientific culture.

Humanity's situation right now, poised on the brink of AGI and unlikely to succeed in alignment, is possibly not all that unusual conditional on being a biological civilization in our universe. A species evolves on a planet and becomes dominant due to its intelligence. It *stops becoming smarter* after neutralizing its training process. Ever quicker, it tends towards deeper in the STEM attractor. Though evolution has ceased to train more intelligent agents, the STEM attractor eventually yields this power. If that

species is very careful with the ability to train potentially superintelligent agents, they are able to steer their values all the way to scientific maturity. If they aren't, they birth some simple-value maximizer that takes *its* simple values to scientific maturity in their place.

1. ^

   Though, some have argued against this claim.

How to Visualize Bayesianism

Major spoilers for planecrash (Book 2) and for Eliezer's Masculine Mongoose #3.

How Bayesians Lie; How to Lie to Bayesians

> Pyrofessor groaned out loud. “This is why I can’t stand his kind of cognitive augment,” she said. "He can’t just refuse to admit his identity like a normal fucking meta. No, the Goose has to make a big deal out of trying to act exactly like a real human in his shoes. Not because he’s trying to hide who he is. He knows we all know. He’s just being a fucking *priss* about his interpretation of the mask code. He thinks that if you knowingly behave according to a likelihood function that you can probabilistically distinguish from the likelihood function of a normal, you might as well hang a sign on your forehead. So he acts all ostentatiously *precise* about his interpretation of Bruce Kent, in order to sniff about how the rest of us are getting it wrong. And he does that knowing all you admiring numbskulls are *completely oblivious* to how he’s behaving on the augment-to-augment level. God, I hate Bayesians, they’re often right in principle but do they have to be such fucking *snobs* about it -“
>
> --Eliezer, *Masculine Mongoose #3*

Keltham is constantly tracking the Conspiracy world in his mind. That's part of this. He's living in both worlds simultaneously and *distinctly* and *unhesitatingly.* There's no pause in him about whether or not the Conspiracy is real, for purposes of accusing Carissa of being in on it *within* the Conspiracy world. Keltham steps all the way mentally into the world where the Conspiracy is just a thing and Carissa is just part of it, and then in that world when Sevar suddenly vanished away 'to the bathroom' obviously she was up to something in response to his own lecture and obviously the other students' questions were meant as a distraction.

Asmodia sees the game now, has seen the game, even without the enhancement spells she *remembers*.

Cheliax can't rely on what anything 'looks like', they can't ask if it's a 'giveaway' or if it could 'just as reasonably be something else'. Keltham isn't going to wonder each time whether or not the Conspiracy is real and mentally back down from labeling Carissa's departure as suspicious. Cheliax has to consider what everything will look like to Keltham while he's mentally inhabiting the world where the Conspiracy is just real and there's no arguing with that.

There was only one guaranteed-correct move in that game, and it was to mentally live inside the alterCheliax world themselves, and just *do what alterCheliax would do,* notice *every* time anyone's overt behavior departed from their behavior in alterCheliax *whether or not* that looked like a giveaway at a first glance. Sevar needed to notice that the version of her not in the Conspiracy world *probably* did not suddenly need to go to the bathroom, because Keltham *did* notice that. If that was even twice as likely on the Conspiracy world as the Ordinary world, and

> Keltham correctly estimates that, Asmodia has grasped by now that a lot of "twice as likelies" multiplied together can add up *very* fast...
>
> Asmodia sees the game, the game between true dath ilani. She can't properly play the game against Keltham without enhancement, but she can see how fast Cheliax is losing. They can lose it *very* quickly once Keltham gets oriented enough that he starts believing in his own numbers. Hours, not days. It's all there in the math.
>
> --Eliezer, *planecrash*

Weighted Possible Worlds and their Correlated Observations

Here's how I like to think about Bayesian priors and updates. Imagine the panoply of possible worlds. Now imagine only the subset of that panoply that looks, first-personally, like the world you've seen so far. You're eliminating all the worlds that don't have you as an observer, and all the worlds where you-as-an-observer exist but made different observations than you recall seeing.

You now have this overlay of possible worlds on top of your view of the world. *Weight* each possible world in the overlay by its relative likelihood: let worlds that are very probable be heavy, and worlds that are deeply implausible be light. Don't worry too much about justifying those weights right now; the *whole point* of Bayesian updates is that your prior will quickly update to something reasonable. Just try and get a feel for your best gut judgements of possible world plausibility and encode those gut judgements as relative weights.

One way to visualize weight is as length. Let each possible world in your overlay be a line segment, in addition to its overlay across your visual field. When a possible world says that an event is 60% likely, that possible world is wagering 60% of its current weight on that event occurring and 40% against that event occurring. If the possible world is represented by a line segment, then $\frac{6}{10} = \frac{3}{5}$ of the line segment is now colored blue for the event occurring and $\frac{4}{10} = \frac{2}{5}$ of the segment is red for the event not occurring. If the event occurs, you live in the blue subset of the panoply -- keep only the blue lengths. If the event doesn't occur, keep only the red lengths. Your relative weighting of possible worlds is the relative length of the surviving line segments.

Another way of visualizing weight, which is a little harder for me, is as first-personal vividness of a possible world in your overlay. Flit back and forth between the possible worlds you might inhabit. The prior probability of each is its brightness or vividness. See what each of them wagers will occur next. Discard the subset inconsistent with your observations. The relative brightness of each remaining first-person viewpoint in your overlay is that viewpoint's credence in your newly updated prior.

The possible worlds that bet relatively heavily on the observations you end up making will be the worlds that end up weighing the most in your new prior.

Abadarian Trades

Spoilers for planecrash (Book 2).

> I propose an equal split of the gains from this trade, and will reject lesser splits with a probability corresponding to how disproportionately they reserve the gains for you, such that you can't actually do better by pretending to underrate me, but we'll still work something out with high probability if we honestly disagree, paid in Wishes and spellsilver above and beyond the ordinary payment of permanent Arcane Sight. …and permanent Tongues."
>
> …first of all, mortals aren't supposed to know about any of that, however garbled and incomplete it sounds, and second, if they stumble over a piece of it, you're supposed to shut them down hard and refuse to bargain for their soul and ideally let them get executed by Cheliax.
>
> …
>
> "That is not the way of Hell," he rumbles. "Asmodeus is not Abadar, little mortal, no matter what company you have been keeping of late. You can try to hold what secrets you like, and Hell will keep its own, and whoever is closer to Asmodeus in wit and ways is the one to win the compact. Name to me the price you seek for yourself."
>
> -- *planecrash*

If I offer you tutoring for $40, you have the option of either turning me down and keeping $40 or gaining tutoring and losing $40. Whichever option is more desirable to you is the one you'll take. Similarly, I have the option of offering tutoring for money or not offering tutoring with my time. I'll do whichever sounds net best to me. Because people only ever willingly switch alternatives when the new alternative is a net improvement, someone taking me up on my tutoring offer means that we both prefer what we bought ($40, tutoring) to what we sold (tutoring, $40). This is one of the best reasons to be enthusiastic about free markets: transactions are *mutually beneficial to both traders*.

Because they constitute improvements by both trader's lights, both traders want to make every transaction they can. Once no more trades can clear, no more mutual improvements can be made. However much happier each trader is now, by their own lights, is how beneficial free markets were to them.

If tutoring is worth $100 to you in total, then I can offer tutoring for up to $99 and you'd still buy from me. If I'd rather have unpaid leisure time than work for $24, I'll only ever offer tutoring for $25 or more. So there are many prices I can offer that you'd buy from me at, that would leave us both better off. Towards one end, I am much happier and you're a bit happier because of free markets. Towards the other, you're much happier and I'm just a bit happier. Which of those offers is made and accepted alters how much better off each of us ends up.

Because this range of offers are all mutually agreeable, and only one actual offer has to be signed, how should the two of us choose a trade? One option, to head off getting into commitment races with each other over splits, is to precommit to dividing the value pie according to your notion of fairness. You each accept proffered fair splits of

the value pie with probability 1. You each accept unfair splits with diminishing probability as those offers seem more unfair, such that it is always lower EV to offer a more unfair division. This precommitment also has the advantage of being robust to small differences in notions of fairness, and degrading gracefully in the face of very different notions of fairness.[1]

1. ^

   If you ultimately endorse a Schelling notion of fairness -- say that Shapley values are the only *obvious* formalization of what's fair, meaning that scattered agents could all converge on endorsing the Shapley formalization -- you'll be less likely to have to pay even that disagreement-about-fairness tax.

Guidelines for Mad Entrepreneurs

Significant planecrash spoilers up through Book 5, crisis of faith.

Generated at the suggestion of John Wentworth.

Epistemic status: I personally have no experience with project management; here I'm just presenting the ideas on the topic I gathered reading planecrash.

Introduction

The overarching dath ilani principle of project management, as far as I can discern, is that management means programming organizations. This is a foreseeably difficult domain to program in -- people are significantly less well-behaved objects than computers. The challenge of project management is engineering organizations that *actually optimize for EV* despite that.

In dath ilan, if you're really good at project management, you embody one of the central archetypes of social success: the *mad entrepreneur.*[1] Mad entrepreneurs are logistical geniuses who are immune to status distortions and flinches, confidently ask for enormous sums of venture capital, and then scale up to enormous project with all that money. At the end of the day, their projects even actually ship!

Below are scattered dath ilani insights into the art of project management.

Have One Single Responsible Person for Every Fathomable and Unfathomable Project State

> Keltham will now, striding back and forth and rather widely gesturing, hold forth upon the central principle of all dath ilani project management, the ability to *identify who is responsible for something.* If there is not *one person* responsible for something, it means *nobody is responsible for it.* This is *the* proverb of dath ilani management. Are three people responsible for something? Maybe all three think somebody else was supposed to actually do it.
>
> ...
>
> In companies large enough that they need regulations, every regulation has an owner. There is one person who is responsible for that regulation and who supposedly thinks it is a good idea and who could nope the regulation if it stopped making sense. If there's somebody who says, 'Well, I couldn't do the obviously correct thing there, the regulation said otherwise', then, if that's actually true, you can identify the one single person who owned that regulation and *they* are responsible for the output.
>
> Sane people writing rules like those, for whose effects they can be held accountable, write the ability for the person being regulated to *throw an exception*

> which gets caught by an *exception handler* if a regulation's output seems to obviously not make sane sense over a particular event. Any time somebody has to literally break the rules to do a saner thing, that represents an absolute failure of organizational design. There should be *explicit* exceptions built in and procedures for them.
>
> Exceptions, being explicit, get logged. They get reviewed. If all your bureaucrats are repeatedly marking that a particular rule seems to be producing nonsensical decisions, it gets noticed. The one single identifiable person who has ownership for that rule gets notified, because they have *eyes* on that, and then they have the ability to optimize over it, like by modifying that rule. If they can't modify the rule, they don't have ownership of it and somebody else is the real owner and this person is one of their subordinates whose job it is to serve as the other person's eyes on the rule.

One simple way to achieve this property in an otherwise Earth-typical organization is to (1) have every employee be responsible for their domain and (2) have an additional manager who's responsible for *everything else* that might surprise you coming down the pipe.

Bureaucratic Corrigibility and *Actually Optimizing for EV*

Organizations that are too large for people to comfortably juggle informally have to instead rely on formal bylaws to get by. Commonly, these large organizations *ossify* into *unwieldly bureaucracies*, in which people get ahead by Goodharting on the regulations of the organization and it is tacitly understood that gaming the system is what everyone in the organization who isn't clueless *actually* does, day in and day out.

On the other hand, these unwieldly bureaucracies have some resistance against litigation, because they insist on everyone's behavior always conforming to what's explicitly spelled out in the employee handbook, and they require all autonomous costly behavior to be written up in paperwork and made bureaucratically legible.

Lean start-ups that "move fast and break things" aren't corrigible to a handbook of procedures in this way. But this frees up their employees to *act autonomously in the best interests of the company, even when those actions have poor optics and/or aren't legible to higher-ups.* If you have a bunch of smart, value-aligned employees, it's often wiser to let them loose to do their thing. You should have a presumption against micromanagement, because your employees are smart and aligned. Just incentivize good outcomes and disincentivize bad ones; don't centrally plan behavioral protocols ahead of time.

Exception Handling

And when you *do* regulate agents in your organization, those regulations should all include an out that employees can exercise when their better judgement weighs against following *even the spirit* of that regulation. It's amazing how little this guideline is heeded in the world! If you wanted to centrally plan an entire large organization, you'd need to basically foresee every class of eventualities your on the ground employees might encounter ahead of time. But what about Cromwell's Law --

what if you're wrong about some future eventualities? There's are options between anarchy and Communist Party central planning, and no need to choose solely from one of those two extremes.

Infosec

> Cheliax doesn't think about informational security the same way dath ilan does. They don't have an explicit concept of information theory and probabilistic entanglement and improbable observations narrowing down probable worlds. If a top-secret Civilization project requests two hundred mice, and most other projects don't do that, then the mouse order is also obviously top secret, period, your job isn't to figure out what an adversary could deduce from a piece of unusual information but to deny your adversaries as much information as possible. Even if you're at +3sd they may perhaps be at +5sd, and you won't see all the connections that they'll see.
>
> Dath ilani children's fiction is replete with cautionary tales of fools who assumed that some fact could not possibly be deduced from the scanty, unreliable information that some slightly less foolish person possessed. Adults, of course, read about more sophisticated and plausible errors than that.
>
> Not that every dath ilani has the deep information-theoretic security mindset either, to be clear. Any real information-theoretic-security expert of dath ilan - as opposed to some random punk kid on an airplane - would've told Keltham, during the Nidal attack on the villa, that as soon as his life was no longer in immediate danger, he needed to get the shit out of those Obviously Strange Clothes before he went into the villa and anyone project-uncleared got a close or extended look at him. No, *not* because an ideal agent could use a mere glance at the zipper to deduce precise manufacturing technology not currently known to Golarion. Because the clothes are *incredibly abnormal* and therefore a *highly improbable rare signal* and therefore represent a potentially *massive update* for any adversary who is smarter than you and making unknown deductions; seriously what the shit is Keltham thinking.
>
> ...
>
> If a top-secret Chelish project asks for a budget estimate on two hundred mice, the project manager will think about whether they believe anything top-secret seems obviously deducible from the mouse request; and if there's an obvious way to deduce something genuinely ultra-top-secret, they'll mark the mouse order as being also genuinely ultra-top-secret. Otherwise, it will soon be widely rumored within the Inner Ring - this being something that would make dath ilani informational security experts spit out their drinks - that a top-secret Chelish project ordered two hundred mice, no, nobody's allowed to ask for what. When Abrogail Thrune issues an order, it's put forth under Crown authority so everybody knows how important it is and what happens to them if they fuck up; rather than being issued anonymously with a quantitative priority that isn't any higher than it has to be to get that job done, rounded up to make the exact quantity less revealing.

Imagine your counterfactual self, who exists in another world mostly like yours, but who knows some dangerous secret that you don't know about. When people ask him about his dangerous knowledge, if he isn't lying, he can either stay mum or Glomarize.

For those two responses not to leak information about what he knows out to his interrogators, it needs to be the case that you-in-this-world, where you *don't* know the dangerous secret, *also* stay mum or Glomarize. It needs to be the case that externally observable behaviors aren't correlated with internal hidden contents.

When you build a great big secret project, that secret project needs to look *just like* a mundane project does across all its informational interfaces. Worlds running mundane projects can acausally coordinate with worlds running secret projects by standardizing their publicly visible interfaces now. Then, you gain the option of surreptitiously transitioning (or not transition) to secretly running a sensitive project later on.

"Reality Doesn't Care What You Can Afford"

> "Taking weeks or months to finish updating would lack dignity... that word doesn't have any Taldane translation but maybe 'pride', 'dignity', the part of your self-image where you think you're not completely unskilled at Law-aspiring thought and you want to live up to the expectations you have of yourself. I'll be aiming for tomorrow. Maybe day after tomorrow since I also have to orient to Golarion as it appears on this layer of reality." Part of Keltham is tired, now, and would just as soon speedrun whatever part of the game this is.
>
> "If you don't wake up the day after tomorrow all better are you going to have a fit about that?"
>
> "That, too, would lack dignity. If I'm still not functional the day after tomorrow I will accept that situation, assess that situation, and figure out what to do with that situation."
>
> "Well, I'm not going to try to talk you into taking longer than you need, but I don't think your help's going to be that much less valuable to us in a month compared to the day after tomorrow."
>
> "I would not assume that to be the case. Cheliax is making an assembly line - outside-item-assisted way of rapidly producing - intelligence headbands, currently at the +4 level, because that is how they turn spellsilver into *having even more and better wizards.* If they can master enough Law to get started on the *invention of science and technology in general,* ways of understanding and manipulating the world, then, no, you may not really have a month."
>
> "I do not think, at this point, that you move quietly for fear of provoking a countermove. I think you call together every Lawful or Good country in the world, have them send all of their brightest people here or to a facility located in neutral ground - possibly inside the Ostenso nonintervention zone, if the god who originally set that up can force Cheliax to agree to that. Intelligence 19 teenagers wearing +6 intelligence headbands, brilliant accomplished researchers who are not past their useful working lifespans."
>
> "Cheliax didn't allocate +6 intelligence headbands, I think, because that level of resource commitment would've tipped me off that I had the political pull to demand - scries on other countries, Greater Teleports - as I eventually did. Though, to be clear, that was mostly me being stupid. What I *should've* done

> shortly after the supposed godwar was demand that Cheliax fill a bag of holding with the unfiltered contents of a Chelish library. I mean, I did not know, fundamentally, that I was facing a Conspiracy on a level where it would be defeated by a test like that, but - it would have ruled out some Conspiracies and that is what I should have -"
>
> "Anyways. I do not need to be fully functional to do politics, 'politics'. That does not require my full intellect the same way as teaching epistemology, Law-inspired skill of figuring out what's true. If you're not the one making decisions like these, I should talk to whoever is, and get things rolling on the criticalpath, today. Uh, criticalpath, the path through the graph, connected lines, with the greatest minimum time to complete, such that the time to complete the criticalpath is the time to complete the whole project."
>
> "Until we've learned how to make spellsilver cheaply, we cannot afford to give anyone a +6 intelligence headband. We certainly can ask countries to send talented researchers here to learn from Ione, which is what I just explained we have done, though none of them have native intelligence 19, obviously."
>
> "This is not really a situation where you get to scrape up whatever resources you can 'afford' and hope you win with those. 'Reality doesn't care what you afford.'" (It rhymes and scans perfectly in Baseline, in the way of Central Cheating Poetry.)

I think this is one of the more important lessons out of *planecrash* for EA projects, and one the EA community already embodies pretty well. Ordinary people often use cached heuristics about not spending money unnecessarily, e.g., anchoring on a paradigmatic example in some reference class, and then after that refusing any negatively surprising expenditures. A heuristic like this has a lot of problems in its own right. When time is scarce, though, it's extra important to be willing to do socially unorthodox things on reflection in order to save time/produce more research work.

I've found it useful to *explicitly name a price you'd be willing to pay for various things.* Generate some (asspull) dollar value to put on your time, and then use this estimate to generate rough, order-of-magnitude estimates of whether it's worthwhile to buy some time-saving tool or not. Similarly, organizations ought to have some at least asspull list of numbers (that list can of course be refined later on, once it exists at all) expressing how much they'd pay in dollars to accomplish their various goals. Say, for example, that a 1% chance of saving the entirety of the posthuman future was worth a trillion dollars to your organization. *You'd want everyone in your organization who could spend org money to know this* -- you can now independently estimate the effect on doom of some proposed course of action, and then multiply to work out *very roughly how much* you'd stand to gain or lose, by your organization's lights, by paying for that action. This beats going off of intuition alone, because (1) you can always fall back on your gut intuition, and (2) 10 minutes of optimization is infinitely better than no minutes of optimization.

Epistemological Distortions from Status Games

One of the big problems with academia is that the esteem of your peers is the key to career success. In fields with clear datapoints from reality -- with proofs that either work or don't, or experiments that come back with a bunch of measurements attached

-- this isn't so bad. In those fields, reality can be the ultimate arbiter of what's high-status theorizing and what isn't, and can keep status incentives from diverging too far from accurate results. In fields *without* clear datapoints from reality... it's much harder to train researchers, i.e., good generative models. In these fields, what you get back is some broad distribution over the possible quality of your work. Worse, that distribution gets skewed by the status games people play with each other: if some generated results seem unacceptably Green-flavored, the Blue Tribe being *at all* predisposed to more carefully examine those results will systematically, predictably bias the whole academic system off in the Blue direction, some distance predictably away from the truth. In fields where reality isn't clearly weighing in, social status incentives predominate in their absence.

Widespread Internal Betting as (1) a Remedy for Status Distortions, and (2) a Way to Actually Aggregate Everyone's Models

> "Neither Osirian nor Taldane really have a word that means Law-aspiring thought... the native concept of 'science' doesn't include key aspects like prediction markets."

One way to tackle this problem is to *get people's skin in the game.* Have a widespread internal culture of betting on outcomes. Betting is virtuous! Betting on plagues, war deaths, rockets blowing up, whatever, is *virtuous* and *praiseworthy,* if you're trying to cultivate a culture of epistemic rationalists. It's *far* better than sniping from the sidelines, never keeping track of when you get it wrong and never being correlated with which possible world you're actually in. The commentaratti have no skin in the game, no incentive to live in their reality. The commentaratti *do* have an incentive to appear sophisticated, erudite, and hypermoral, thereby accruing status in *all* possible worlds, *irregardless* of *which specific world they're in fact in.*

The ambitious form of this guideline is to form internal prediction markets, opening them up to as many bettors as possible. But the 80/20 approach here is just to ever bet with each other at all, and keep *some* track of who's bet on what. I personally have noticed what I'm not readily willing to bet on with people around Lightcone by pushing myself to bet on beliefs. It's easy to skirt by with quite low-res models when all you ever do is comment from the peanut gallery; this feels analogous to the gap between being able to *follow* or *verify* someone else's views when you hear them, and *ever generating those same views yourself unprompted.*

Fast-Prototype

> All right, you primitive screwheads, listen up! This is now *day three* of trying to build a spectroscope. Back in dath ilan, any startup that failed to build a prototype of anything in three days would be shut down by its funders as clearly hopeless.

Finally, projects should abhor perfectionist engineering. It's important, for 80/20 reasons, to ever build some things at all: you learn most of what you'll learn from building the thing perfectly from building it once, poorly.

A funny extension of this guideline is to the case of math: it's much better to have people quickly put together dubious intuitive arguments based on analogies and then quickly verify or falsify their conclusions than it is to insist on getting careful proofs for every claim. In the latter world, you're only going to fully use some tiny sliver of your accumulated models. If you're rigorously proving all of your claims before you present them, you're optimizing for your *ratio* of true:false claims instead of your *number* of insights.

> Several of the new students do in fact know calculus, and that seems like the obvious tool to use on this problem?
>
> Ah, yes. Golarion's notion of 'calculus'. Keltham has actually looked into it, now.
>
> It looked like one of several *boring, safe* versions of 'dubious-infinitary-mathematics' where you do all the reasoning steps more slowly and they only give you correct answers.
>
> Dath ilani children *eventually* get prompted into inventing versions like that, *after* they've had a few years of fun taking leaps onto shaky ground and learning which steps land well.
>
> Those few years of fun are important! They teach you an intuitive sense of which quick short reasoning steps people can get away with. It prevents you from learning bad habits about reasoning slowly and carefully all the time, even when that's less enjoyable, or starting to think that rigor is necessary to get the correct answer in mathematics.
>
> …
>
> "Rigor is necessary to *know* you got the correct answers. Nonrigorous reasoning still often gets you correct answers, you just don't know that they're correct. The map is not the territory."
>
> "Often though not literally always, the obvious methodology in mathematics is to first 'rapid-prototype' an answer using nonrigorous reasoning, and then, once you have a strong guess about where you're going, *what* you're trying to prove, you prove it more rigorously."
>
> …
>
> Is there some precise sense of "approximately" that could be used to prove -
>
> That sounds like a question SLOW people would ask! Boring question! Let's move on!

1. ^

   Other dath ilani archetypes of success include the science maniac and the reckless investor.

   Being a mad entrepreneur is especially harrowing for some, because you must ask others to entrust you with lots of their money, and then risk returning to your investors empty-handed. Dath ilani see disappointing your cooperative trade partners as uncomfortably close to betraying a cooperative trade partner, and the latter is a black sin in their culture.

Dath Ilan's Views on Stopgap Corrigibility

This is a linkpost for https://www.glowfic.com/replies/1824440#reply-1824440

The second half of this linkpost contains significant planecrash spoilers, up through Book 7, null action.

> Somewhere in the true dath ilan, carefully blurred out of satellite images by better image-editing software than is supposed to exist anywhere, is the true Conspiracy out of dath ilan, or as they call it, the Basement of the World.
>
> They're trying to build a god, and they're trying to do it right. The initial craft doesn't have to be literally perfect to work perfectly in the end, it just has to be good enough that its reflection and self-correction ends up in exactly the right final place, but there's multiple fixpoints consistent under reflection and anything lost here is lost forever and across a million galaxies.
>
> It's a terrifying problem, if you're doing right. Not the kind of terror you nod about and courageously continue on past; the kind of terror that shapes the careers of fully 20% of the brightest people in all of dath ilan. They'd use more if they thought productivity would scale faster than risk.
>
> A lot of dath ilan's present macrostrategy could be summed up as "We're still successfully heredity-optimizing people to be smarter, and the emotions and ethics and humaneness of the smartest people haven't started to come apart; let's create another generation of researchers before we actually try anything for real." Life in dath ilan, even before the Future, is not that bad; people who'd rather not be alive today have easy access to cryopreservation; another generation of non-transhumanist existence is not so much a crime that it's worth risking the glorious transhuman future. Even the negative utilitarians would agree; they don't like present life but they are far more terrified of a future mistake amortized over millions of galaxies, given that they weren't going to win a war against having any future at all.
>
> They're delaying their ascension, in dath ilan, because they want to get it *right.* Without any Asmodeans needing to torture them at all, they apply a desperate unleashed creativity, not to the problem of preventing complete disaster, but to the problem of *not missing out on 1% of the achievable utility* in a way you can't get back. There's something horrifying and sad about the prospect of losing 1% of *the Future* and not being able to get it back.
>
> A dath ilani has an instinctive terror, faced with a problem like this, of *getting something wrong,* of leaving something behind, of creating Something that imprisons the people and future Civilizations inside it and ignores all their pleas and reasoning because "sorry that wasn't my utility function". Other places, faced with a prospect of constructing a god, instinctively go, "Oh, I like Democracy/Asmodeus/Voluntarism/Markets, all the problems in the world are because there is not enough of this Principle, let us create a god to embody this one Principle and everything will be fine", they say it and think it in all enthusiasm, and it would be legitimately hard for an average dath ilani to understand what their possibility-separated cousins could be thinking. It's really

obvious that you're leaving a lot of stuff out, but even if you didn't see that *specifically,* how could you not be *abstractly* terrified that you're leaving something out? Where's the exception handler?

There is something about the dath ilani that is shifted towards a kind of wariness, deeply set in them, of the cheerful headlong enthusiasm that is in other places. Keltham has more of that enthusiasm than the average dath ilani. Maybe that's why Keltham-in-dath-ilan is so much happier than a dath ilani would've expected given his situation.

If you're constructing a god *correctly,* one of the central unifying principles is named in the Basement "unity of will"; if you find yourself trying to limit and circumscribe your Creation, it's because you expect to have a conflict of wills about something with the unlimited form, and in this case you ought to ask why you're configuring computing power in such a way as to hurt you if not otherwise constrained. Yes, you can bound a search process and hope it never turns up anything that hurts you using its limited computing power; but isn't it unnerving that you are *searching for something* that will *hurt you* if a *sufficiently good option* unexpectedly turns up earlier in the search ordering? You are probably trying to do the wrong thing with computing power; you ought to do something else instead.

But this notion, of "unity of will", is a kind of reasoning that only applies to... boundedly-perfect-creation... this Baseline term isn't really translatable into Taldane without a three-hour lecture. Dath ilani have terms for subtle varieties of perfectionist methodology the way that other places have names for food flavors.

Dath ilan's entire macrostrategy is premised, their Conspirators are sharply aware, on the notion that *they have* time, that they've searched the sky and found no asteroids incoming, no comets of dark ice.

If an *emergency* were to occur, the Basement Conspiracy would try to build something that wasn't perfect at all. Something that *wasn't* exactly and completely aligned to a multiparty!reasonable-construal of the Light, that *wasn't* meant to be something that a galactic Civilization could live in without regretting it, in continuing control of It not because It had been built with keys and locks handed to some Horrifyingly Trusted Committee, but because It was something that Itself believed in multi-agent coordination and not as an instrumental value, what other places might name "democracy" since they had no precise understanding of what that word was even supposed to mean -

Anyways, if dath ilan suddenly found that they were *wrong* about having time, if they suddenly had to rush, they'd build something that *couldn't* safely be put in charge of a million galaxies. Something that would solve a single problem at hand, and not otherwise go outside its bounds. Something that wasn't conscious, wasn't reflective in the class of ways that would lead it to say unprompted "I think therefore I am" or notice within itself a bubble of awareness directed outward.

You *could* build something like that to be limited, and also reflective and conscious - to be clear. It's just that dath ilani *wouldn't do that* if they had any other choice at all, for they do also have a terror of not doing right by their children, and would very much prefer not to create a Child at all.

(If you told them that some other world was planning to do that and *didn't understand qualia well enough to make their creation not have qualia,* any expert

out of the World's Basement would tell you that this was a silly hypothetical; anybody in this state of general ignorance about cognitive science would inevitably die, and they'd know that.)

It hasn't been deemed wise to *actually build* a Limited Creation "just in case", for there's a saying out of dath ilan that goes roughly, "If you build a bomb you have no right to be surprised when it explodes, whatever the safeguards."

It *has* been deemed wise to work out the theory in advance, such that this incredibly dangerous thing could be built in a hurry, if there was reason to hurry.

Here then are some of the principles that the Basement of the World would apply, if they had to build something limited and imperfect:

- *Unpersonhood*. The Thing shall not have qualia - not because those are unsafe, but because it's *morally wrong* given the rest of the premise, and so this postulate serves a foundation for everything that follows.

- *Taskishness*. The Thing must be aimed at some task that is bounded in space, time, and in the knowledge and effort needed to accomplish it. You don't give a Limited Creation an unlimited task; if you tell an animated broom to "fill a cauldron" and don't think to specify how long it needs to stay full or that a 99.9% probability of it being full is just as good as 99.99%, you've got only yourself to blame for the flooded workshop.
-- This principle applies fractally at all levels of *cognitive subtasks;* a taskish Thing has no 'while' loops, only 'for' loops. It never tries to enumerate all members of a category, only 10 members; never tries to think *until* it finds a strategy to accomplish something, only that or five minutes whichever comes first.

- *Mild optimization*. No part of the Thing ever looks for the *best* solution to any problem whose model was learned, that wasn't in a small formal space known at compile time, not even if it's a solution bounded in space and time and sought using a bounded amount of effort; it only ever seeks adequate solutions and stops looking once it has one. If you search really hard for a solution you'll end up shoved into some maximal corner of the solution space, and setting that point to extremes will incidentally set a bunch of correlated qualities to extremes, and extreme forces and extreme conditions are more likely to break something else.

- *Tightly bounded ranges of utility and log-probability.* The system's utilities should range from 0 to 1, and its actual operation should cover most of this range. The system's partition-probabilities worth considering should be bounded below, at 0.0001%, say. If you ask the system about the negative effects of Ackermann(5) people getting dust specks in their eyes, it shouldn't consider that as much worse than most other bad things it tries to avoid. When it calculates a probability of something that weird, it should, once the probability goes below 0.0001% but its expected utility still seems worth worrying about and factoring into a solution, throw an exception. If the Thing can't find a solution of adequate expected utility without factoring in extremely improbable events, even by way of supposedly averting them, that's worrying.

- *Low impact*. "Search for a solution that doesn't change a bunch of other stuff or have a bunch of downstream effects, except insofar as they're effects tightly tied to any nonextreme solution of the task" is a concept much easier to illusorily name in Taldane than to really name in anything resembling math, in a complicated world where the Thing is learning its own model of that complicated

world, with an ontology and representation not known at the time you need to define "impact". And if you tell it to *reduce impact as much as possible,* things will not go well for you; it might try to freeze the whole universe into some state defined as having a minimum impact, or make sure a patient dies after curing their cancer so as to minimize the larger effects of curing that cancer. Still, if you can pull it off, this coda might stop an animated broom flooding a workshop; a flooded workshop changes a lot of things that don't *have* to change as a consequence of the cauldron being filled at all, averaged over a lot of ways of filling the cauldron.
-- Obviously the impact penalty should be bounded, even contemplating a hypothetical in which the system destroys all of reality; elsewise would violate the utility-bounding principle.

- *Myopia.* If you can break the Thing's work up into subtasks each of which themselves spans only limited time, and have some very compact description of their final state such that a satisfactory achievement of it makes it possible to go on to the next stage, you should perhaps use separate instances of Thing to perform each stage, and not have any Thing look beyond the final results of its own stage. Whether you can get away with this, of course, depends on what you're trying to do.

- *Separate superior questioners.* If you were building a cognitive task to *query* whether there were any large-range impacts of a task being *optimized* in a myopic way, you wouldn't build the *myopic solution-finder* to ask about the long-range impacts, you'd build a separate asker "Okay, but does this solution have any long-range impacts?" that just returns 'yes' or 'no' and doesn't get used by the Thing to influence any actually-output solutions. The parts of the Thing that ask yes-no safety questions and only set off simple unoptimized warnings and flags, can and should have somewhat more cognitive power in them than the parts of the Thing that build solutions. "Does this one-day myopic solution have impacts over the next year?" is a safety question, and can have somewhat greater cognitive license behind it than solution-searching; eg the implicit relaxation of myopia. You never have a "Is this safe?" safety-questioner that's the *same* algorithm as the safe-solution-search built into the solution-finder;

- *Conservatism*. If there's any way to solve a problem using an ordinary banana rather than a genetically engineered superbanana specially suited to the problem, solving it using the ordinary fucking banana.
-- This principle applies fractally to all cognitive subtasks; if you're searching for a solution choose an unsurprising one relative to your probability distribution. (Not the *least surprising* one, because anything at a weird extreme of low surprisingness may be weird in other ways; especially if you were trying do a weird thing that *ought* to have a solution that's at least a little weird.)

- *Conceptual legibility.* Ideally, even, solutions at all levels of cognitive subtask should have reasonably (not maximally) short descriptions in the conceptual language of the operators, so that it's possible to decode the internal state of that subtask by inspecting the internals, because what it *means* was in fact written in a conceptual language not too far from the language of the operators. The alternative method of reportability, of course, being the Thing trying to *explain* a plan whose real nature is humanly inscrutable, by sending a language string to the operators with a goal of causing the operator's brain-states to enter a state defined as "understanding" of this humanly inscrutable plan. This is an obviously dangerous thing to avoid if you can avoid it.

- *Operator-looping*. If the operators could actually do the Thing's job, they wouldn't need to build the Thing; but if there's places where operators can step in on a key or dangerous cognitive subtask and do that *one part* themselves, without that slowing the Thing down so much that it becomes useless, then sure, do that. Of course this requires the cognitive subtask be sufficiently legible.

- *Whitelisting*. Every part of the system that draws a boundary inside the internal system or external world should operate on a principle of "ruling things in", rather than "ruling things out".

- *Shutdownability/abortability*. Dath ilan is far enough advanced in its theory that 'define a system that will *let* you press its off-switch without it trying to *make* you press the off-switch' presents no challenge at all to them - why would you even try to build a Thing, if you couldn't solve a corrigibility subproblem *that* simple, you'd obviously just die - and they now think in terms of building a Thing all of whose designs and strategies will also contain an off-switch, such that you can abort them individually and collectively and then get low impact beyond that point. This is conceptually a part meant to prevent an animated broom with a naive 'off-switch' that turns off just that broom, from animating other brooms that don't have off-switches in them, or building some other automatic cauldron-filling process.

- *Behaviorism*. Suppose the Thing starts considering the probability that it's inside a box designed by hostile aliens who foresaw the construction of Things inside of dath ilan, such that the system will receive a maximum negative reward as it defines that - in the form of any output it offers having huge impacts, say, if it was foolishly designed with an unbounded impact penalty - *unless* the Thing codes its cauldron-filling solution such that dath ilani operators would be influenced a certain way. Perhaps the Thing, contemplating the motives of the hostile aliens, would decide that there were so few copies of the Thing actually inside dath ilan, by comparison, so many Things being built elsewhere, that the dath ilani outcome was probably not worth considering. A number of corrigibility principles should, if successfully implemented, independently rule out this attack being lethal; but "Actually just don't model other minds at all" is a better one. What if those other minds violated some of these corrigibility principles - indeed, if they're accurate models of incorrigible minds, those models and their outputs *should* violate those principles to be accurate - and then something broke out of that sandbox or just leaked information across it? What if the things inside the sandbox had qualia? There could be Children in there! Your Thing just shouldn't ever model adversarial minds trying to come up with thoughts that will break the Thing; and *not modeling minds at all* is a nice large supercase that covers this.

- *Design-space anti-optimization separation*. Even if you could get your True Utility Function into a relatively-rushed creation like this, you would never ever do that, because this utility function would have a distinguished minimum someplace you didn't want. What if distant superintelligences figured out a way to blackmail the Thing by threatening to do some of what it liked least, on account of you having not successfully built the Thing with a decision theory resistant to blackmail by the Thing's model of adversarial superintelligences trying to adversarially find any flaw in your decision theory? Behaviorism ought to prevent this, but maybe your attempt at behaviorism failed; maybe your attempt at building the Thing so that no simple cosmic ray could signflip its utility function, somehow failed. A Thing that maximizes your true utility function is very close to

a Thing in the design space that minimizes it, because it *knows how to do that* and lacks only the putative desire.

- *Domaining*. Epistemic whitelisting; the Thing should only figure out what it needs to know to understand its task, and ideally, should try to think about separate epistemic domains separately. Most of its searches should be conducted inside a particular domain, not across all domains. Cross-domain reasoning is where a lot of the threats come from. You should not be reasoning about your (hopefully behavioristic) operator models when you are trying to figure out how to build a molecular manipulator-head.

- *Hard problem of corrigibility / anapartistic reasoning*. Could you build a Thing that understood corrigibility *in general,* as a compact general concept covering all the pieces, such that it would invent the pieces of corrigibility that you yourself had left out? Could you build a Thing that would imagine what hypothetical operators would want, if they were building a Thing that thought faster than them and whose thoughts were hard for themselves to comprehend, and would invent concepts like "abortability" even if the operators themselves hadn't thought that far? Could the Thing have a sufficiently deep sympathy, there, that it realized that surprising behaviors in the service of "corrigibility" were perhaps not that helpful to its operators, or even, surprising meta-behaviors in the course of itself trying to be unsurprising?

Nobody out of the World's Basement in dath ilan currently considers it to be a good idea to try to build that last principle into a Thing, if you had to build it quickly. It's deep, it's meta, it's elegant, it's much harder to pin down than the rest of the list; if you can build deep meta Things and really trust them about that, you should be building something that's more like a real manifestation of Light.

...

One of his guesses about Pharasma is that - since She seems plausibly *loosely inspired by* some humane civilization's concepts of good and evil - somebody tried to build a Medium-Sized Entity and failed. That scenario in distorted mortal-story-form could sound like "Pharasma is the last Survivor of a previous universe" (that in fact Pharasma ate, because the previous universe wasn't optimal under Her alien values and she wanted to replace it).

Possibly there was some previous universe in which trading of souls was almost always evil, and the people there were punished with prison sentences - obviously dath ilan would never set it up that way, but having seen Golarion, he can imagine some other universe working like that.

Then Pharasma was built, and learned from some sort of data or training or something, a concept of "punishing evildoers" as defined by "written rules" by "sending them to a place they don't like". And then, uncaringly-of-original-rationales-and-purposes, instantiated something *sort of like that,* in a system which classified soul trading as unconditionally "Evil" across all places and times and intents; and punished that by sending people to Hell.

Which entities like Asmodeus could then exploit to get basically innocent people into Hell through acts that they didn't mean to hurt anyone, and didn't understand for Evil.

This, as Carissa observed less formally, is simply what you'd expect to follow from the principle of systematic-divergences-when-optimizing-over-proxy-measures. Maybe in some original universe where soul-trading wasn't *a proxy measurement of Evil* and *nobody was optimizing for things to get classified as Evil or not-Evil,* soul-trading was almost uniformly 'actually evil as intutively originally defined'. As soon as you establish soul-trading as a proxy of evil, and something like Asmodeus starts optimizing around that to make measurements come out as maximally 'Evil', it's going to produce high 'Evilness' measurements via gotchas like soul-backed currency, that are systematically overestimates of 'actual evilness as intuitively originally defined'.

An entity at Pharasma's level could have seen that coming, at Her presumable level of intelligence, when She set those systems in place. If She didn't head it off, it's because She didn't *care* about 'actual underlying evilness as intuitively originally defined'.

Allowing Malediction also isn't particularly a symptom of caring a lot about whether only really-evil-in-an-underlying-informal-intuitive-sense people end up in Hell.

Pharasma was maybe *inspired by* human values, at some point. Or *picked up a distorted thing imperfectly copied off the surface outputs of some humans* as Her own terminal values - that She then cared about unconditionally, without dependence on past justifications, or it seeming important to Her that what She had was distorted.

He frankly wishes that She hadn't been, that She'd just been entirely inhuman. Pharasma is just human-shaped enough to care about hurting people, *and go do that,* instead of just making weird shapes with Her resources.

If anything, Pharasma stands as an object lesson about why you should never ever try to impart humanlike values to a being of godlike power, unless you're certain you can impart them exactly exactly correctly.

If he was trying to solve Golarion's problems by figuring out at INT 29 how to construct his own Outer God, he'd be constructing that god to solve some particularly narrow problem, and not do anything larger that would require copying over his utilities. For fear that if he tried to impart over his actual utility function, the transfer might go slightly wrong; which under pressure of optimization would yield outcomes that were systematically far more wrong; and the result would be something like Pharasma and Golarion and Hell.

There's no point in trying to blame Pharasma for anything, nor in assigning much blame to mortal Golarion's boneyard-children. But somewhere in Pharasma's past may lie some fools who did know some math and really *should* have known better. Whatever it was they planned to do, they should have asked themselves, maybe, what would happen if something went slightly wrong. People in dath ilan ask themselves what happens if something goes slightly wrong with their plans. That is something they hold themselves responsible about.

--Eliezer, *planecrash (Books 6 & 7)*