Best of LessWrong: January 2018

1. Arbital postmortem
2. A LessWrong Crypto Autopsy
3. Babble
4. Announcement: AI alignment prize winners and next round
5. Beware of black boxes in AI alignment research
6. The Loudest Alarm Is Probably False
7. Teaching Ladders
8. Dispel your justification-monkey with a “HWA!”
9. Aliveness
10. Why everything might have taken so long
11. Global online debate on the governance of AI
12. Learning to make better decisions
13. A model I use when making plans to reduce AI x-risk
14. No, Seriously. Just Try It: TAPs
15. The Desired Response
16. Justice
17. Video - Subject - Object Shifts and How to Have Them
18. TSR #10: Creative Processes
19. Hammers and Nails
20. Superhuman Meta Process
21. Hammertime Day 1: Bug Hunt
22. A simpler way to think about positive test bias
23. Rationality: Abridged
24. Understanding-1-2-3
25. Actionable Eisenhower
26. GoodAI announced "AI Race Avoidance" challenge with $15k in prize money
27. Demon Threads
28. What Is "About" About?
29. One-Consciousness Universe
30. The O'Brien Technique
31. Paper: Artificial Intelligence in Life Extension: from Deep Learning to Superintelligence
32. Schelling Orders
33. The different types (not sizes!) of infinity
34. Roleplaying As Yourself
35. ProbDef: a game about probability and inference
36. Insights from 'The Strategy of Conflict'
37. Prune
38. More Babble
39. [Paper] Global Catastrophic and Existential Risks Communication Scale, similar to Torino scale
40. Making Exceptions to General Rules
41. Singularity Mindset
42. Pareto improvements are rarer than they seem
43. Adequacy as Levels of Play
44. Have you felt exiert yet?
45. Magic Brain Juice
46. Conversational Presentation of Why Automation is Different This Time
47. Niceness Stealth-Bombing
48. Field-Building and Deep Models
49. The Tallest Pygmy Effect
50. The Solitaire Principle: Game Theory for One

Best of LessWrong: January 2018

1. Arbital postmortem
2. A LessWrong Crypto Autopsy
3. Babble
4. Announcement: AI alignment prize winners and next round
5. Beware of black boxes in AI alignment research
6. The Loudest Alarm Is Probably False
7. Teaching Ladders
8. Dispel your justification-monkey with a “HWA!”
9. Aliveness
10. Why everything might have taken so long
11. Global online debate on the governance of AI
12. Learning to make better decisions
13. A model I use when making plans to reduce AI x-risk
14. No, Seriously. Just Try It: TAPs
15. The Desired Response
16. Justice
17. Video - Subject - Object Shifts and How to Have Them
18. TSR #10: Creative Processes
19. Hammers and Nails
20. Superhuman Meta Process
21. Hammertime Day 1: Bug Hunt
22. A simpler way to think about positive test bias
23. Rationality: Abridged
24. Understanding-1-2-3
25. Actionable Eisenhower
26. GoodAI announced "AI Race Avoidance" challenge with $15k in prize money
27. Demon Threads
28. What Is "About" About?
29. One-Consciousness Universe
30. The O'Brien Technique
31. Paper: Artificial Intelligence in Life Extension: from Deep Learning to Superintelligence
32. Schelling Orders
33. The different types (not sizes!) of infinity
34. Roleplaying As Yourself
35. ProbDef: a game about probability and inference
36. Insights from 'The Strategy of Conflict'
37. Prune
38. More Babble
39. [Paper] Global Catastrophic and Existential Risks Communication Scale, similar to Torino scale
40. Making Exceptions to General Rules
41. Singularity Mindset
42. Pareto improvements are rarer than they seem
43. Adequacy as Levels of Play
44. Have you felt exiert yet?
45. Magic Brain Juice
46. Conversational Presentation of Why Automation is Different This Time
47. Niceness Stealth-Bombing

48. Field-Building and Deep Models
49. The Tallest Pygmy Effect
50. The Solitaire Principle: Game Theory for One

Arbital postmortem

Disclaimer 1: These views are my own and don’t necessarily reflect the views of anyone else (Eric, Steph, or Eliezer).

Disclaimer 2: Most of the events happened at least a year ago. My memory is not particularly great, so the dates are fuzzy and a few things might be slightly out of order. But this post has been reviewed by Eric, Steph, and Eliezer, so it should mostly be okay.

I’m going to list events chronologically. At times I’ll insert a “**Reflection**” paragraph, where I’m going to outline my thoughts as of now. I’ll talk about what I could have done differently and how I would approach a similar problem today.

Chapter 0: Eliezer pitches Arbital and I say ‘no’

Around the summer of 2014 Eliezer approached me with the idea for what later would become Arbital. At first, I vaguely understood the idea as some kind of software to map out knowledge. Maybe something like a giant mind map, but not graphical. I took some time to research existing and previous projects in that area and found a huge graveyard of projects that have been tried. Yes, basically all of them were dead. Most were hobby projects, but some seemed pretty serious. None were successful, as far as I could tell. I didn’t see how Eliezer’s project was different, so I passed on it.

Reflection: Today, I’d probably try to sit down with Eliezer for longer and really try to understand what he is seeing that I’m not. It’s likely back then I didn’t have the right skills to extract that information, but I think I’m much better at it today.

Reflection: Also, after working with Eliezer for a few years, I’ve got a better feeling for how things he says often seem confusing / out of alignment / tilted, until you finally wrap your mind around it, and then it’s crystal clear and easy.

Chapter 1: Eliezer and I start Arbital

Early January 2015 I was sitting in my room, tired from looking in vain for a decent startup idea, when Arbital popped back into my mind. There were still a lot of red flags around the idea, but I rationalized to myself that given Eliezer's track record, there was probably something good here. And, in the worst case, I'd just create a tool that would be useful to Eliezer alone. That didn't seem like a bad outcome, so I decided to do it. I contacted Eliezer, he was still interested, and so we started the project.

Reflection: The decision process sounds a bit silly, but I don't think it's a bad one. I really prefer to do something decently useful, rather than sit around waiting for something perfect. I also still approve of the heuristic of accepting quests / projects from people you think are good at coming up with quests / projects. But if I did it again, I'd definitely put a lot more effort upfront to understand the entire vision before committing to it.

Reflection: Paul Graham wrote in one of his essays that it's okay (though not ideal) to initially build a product for just one user. There are, of course, several caveats. The user needs to use the product extensively, otherwise you don't get the necessary feedback on all the features you're building. And the user needs to be somewhat typical of other users you hope to attract to the platform.

Reflection: Unfortunately, both of these turned out to be false. I'll elaborate on the feature usage below. But the "typical" part probably could have been foreseen. There are only a few people in the world who write explanations at the scale and complexity that Eliezer does. The closest cluster is probably people writing college textbooks. So, in the beginning, I didn't have any sense for who the first 10-100 users were going to be. That would have been fine if I was *just* building a tool for Eliezer, but since my goal was explicitly to create a for-profit consumer startup, this was a big mistake.

Eliezer provided the product vision and design, and I did all the coding. At first, I thought I'd code for a few months and then we would have an MVP that we could show to a few people to gather more interest and get some potential users. But, as I began to understand the overall vision better myself, the shipping date began drifting further and further back. At the time this worried me greatly, because I didn't want to build a thing that nobody else would use. Eliezer's argument was that we needed to build a product that was the best tool for a particular workflow. (I'm the Startup Founder 1 in conversation 2.) This made sense to me, but I still felt anxious that we were flying blind. So around April, I went around and showed what I had to some people. There wasn't much to look at, and what was there wasn't pretty, so it was mostly me explaining the idea. The reception was lukewarm. People said it seemed interesting, but may be not particularly for them. This was a bit discouraging, but it was also clear that people weren't getting the *full* vision.

Reflection: Sigh, this is complicated. In general, I agree that if you are showing / talking about your product to potential users and they are not interested then either you're talking to the wrong people, your product isn't useful, or you're presenting it wrong. In the case of Arbital, though, I think lack of enthusiasm was due to how hard it was to explain the entire vision. There were a lot of moving parts, and a lot of what made Arbital good eventually was the full combination of all those parts.

Reflection: I think the correct thing to do would have been to create detailed UI screens. Then print them and show them to people (and Eliezer). This probably would have taken a month or two, but it would have been worthwhile. The reason I never got around to it, aside from the ugh-field around doing UI mockups, was because it always felt like in a month or two we would be done with the MVP.

Reflection: Eliezer requested a lot of features, and most of them had good justifications for why the final product needed to have them. But, neither of us was very good at prioritizing. (I wouldn't say we were bad, but we probably could have sped up the development by about 25% if we were better.) It was only around autumn when we finally got better at it.

Reflection: One such feature was a pretty nifty system for questions and answers. Of course, since nobody was using the platform, we didn't really get any questions or answers, so it was hard to test that feature, and maintaining it felt pointless. Another feature: a private domain, where you could basically have your own private instance of Arbital at your_subdomain.arbital.com.

Around summer of 2015, I finally started to get a grasp for the *entire* vision. The grand plan had five major problems that needed to be solved: Explanations -> Debate -> Notifications -> Rating -> Karma. (Done roughly in that order, but also in parallel.)

Explanations: Arbital as a better Wikipedia. (1, 2) Each page would explain a specific concept (as opposed to Wikipedia pages that list a bunch of facts); the system would create a sequence of pages for you to read to understand a topic, where the sequence would be tailored specifically to you based on your preferences and what you already know.

Debate: Arbital could of course be used as a blog. We also wanted to support comments (both for wiki and blog pages). We also wanted the discussion to be of high quality and centralized.

Notifications: Make sure the user is notified about various events that might interest them (e.g. a new comment in a thread they are subscribed to, a new article to read). Also, if they are a writer, they need to be notified of various related events as well (e.g. someone commented, someone proposed an edit).

Rating: How will the system know which pages, explanations, or comments are good? How will the system be resistant to people trying to game it to make their pages, explanations, or comments appear better than they are? If we do this right, we could replace Yelp (or other services whose primary function is to provide ratings).

Karma: How will we rate users? How will their ratings affect what they can do? How do ratings interact between domains (e.g. math domain vs. art domain)?

Later that year Eliezer wrote a 55 page document describing Arbital and how and why it was different and necessary. (If Eliezer ever gets around to it, he might edit and publish it at some point. I'm mostly mentioning it here to underline the size and complexity of the project.)

Reflection: Once I understood how Arbital was different, it was clear that no previous (nor current) project has even come close to trying to capture that vision. Over the years I've had a lot of people send me messages that they or their friend were working on a similar project. And it's true, for most people who give a cursory glance at Arbital, it seems similar to the other "organize all knowledge" projects. But I'll still

maintain that Arbital is a different kind of beast. And certainly in scope and ambition, I haven't seen anything close.

Reflection: Now you can probably see how the meme of "Arbital will solve that too" was born. It was a hugely ambitious project for sure, but looking back the only problem with that was that for a while we just didn't have a good, short explanation of what Arbital was. This made it hard to talk to people about the project and get them excited. It also made prioritizing features more difficult.

So, the first major problem we wanted to solve was Explanations. If we solved it well, it's possible we could become the next Wikipedia (or at least a much better Quora). Our goal was for Arbital to be the best tool to write and organize online explanations. The primary topic we wanted to explain was, of course, AI safety. But we reasoned that if we just had AI safety content, especially if it was mostly written by Eliezer, the website wouldn't become generally used and its content widely accepted. (And then we definitely wouldn't become the next Wikipedia.) This is why later we focused mostly on math explanations.

At the end of 2015 we launched a private beta version for MIRI. A few weeks before, I sat down with a UX designer, Greg Schwartz. We spent a few sessions going over all the screens and redesigning them to be simpler and more understandable. He often pushed me to simplify the project and drop various features. I also had another friend look at UI and help with font and colors. This was definitely time well spent (only about a month), and we later got many compliments on the look and feel of the website.

Reflection: It occurs to me now that while Greg's feedback had some specifics wrong, it was overall correct in that it was pointing out a deep problem: the project had too many moving parts and a lot of those parts weren't really used. It would have been hard to guess which parts would end up necessary, but the right solution was to find more users who would want to use the platform now (or very soon) and talk to them.

I was excited about the launch, because I thought that finally some people aside from Eliezer would be using Arbital. Unfortunately, it was only many many months later that other people from MIRI slowly started using it.

Reflection: I think after we reached our "MVP", I should have switched into "find users" mode. (Ideally, I would have had users lined up at the outset, but even this timing would have been okay.) For example, I could have pushed for Agent Foundations forum to be ported to Arbital. Even though that was more of a Discussion project, these were very reachable users, still within the overall strategy. I think we should have used a greedy user acquisition strategy, instead of trying to stick to our rigid sequential plan.

Reflection: I'd describe one of the main struggles of 2015 as: "we need to build a small MVP quickly and get feedback from users" (Alexei) vs. "users don't know what they want, and they won't be able to give you meaningful feedback until they see and use the product" (Eliezer). Like I mentioned above, I think the correct solution here are detailed mockups.

Reflection: Another struggle was: "we need users to make sure we are building things correctly" (Alexei) vs. "I can tell when we are building things correctly, I can get us users as soon as the product is ready" (Eliezer). Unfortunately, we never got the product "ready" enough to test Eliezer's claim. I think it would have a taken a long

while to get there. But, given how things ended up, it’s possible that would have been a better path.

Chapter 2: Eric and Steph join Arbital, and we take destiny into our own hands

Around April of 2016 Eric Rogstad and Stephanie Zolayvar joined the team. We continued following Eliezer's vision and have him dictate features and their design. Since focusing on AI alignment alone wouldn't have resulted in a respected platform, we shifted our primary topic to math, specifically: intuitive math explanations.

Reflection: When we ran this idea by people, we got a lot of positive feedback. A lot of people said they wanted that kind of website, but it took me some time to realize that everyone wanted to *read* intuitive math explanations, but almost nobody would actually spend the time creating them, even if they could in principle.

We invited some people to write the content. We hosted a writing party. We had a Slack channel, where with Eric Bruylant's help we built a small community. Some people wrote pretty good math explanations, but overall things moved way too slow. We talked to some of our users; we tried various things, like creating projects. But, we simply didn't have enough writers, and we didn't know how to find more.

Reflection: I think we should have dropped most of the development and focused on user acquisition at this point. There were several times when I considered pivoting to a "math blogging" platform, but it felt like too big of a shift from wiki-focused plan we were pursuing. Again, I think a greedy "acquire users now!" strategy would have served us well.

One of the biggest features we built around this time was dynamic explanations. A lot of effort went into designing and implementing a system of requisites. Basically each page could teach and/or require certain requisites, which were other pages. It was not clear what overall ontology we wanted, so it took us a while to iterate this feature and we ended up with a lot of edge cases. We built something that worked okay, but, again, it was hard to test because there wasn't quite enough dense content.

Reflection: I'd say we iterated that feature for way too long. In part this was because Eliezer was consistently not satisfied with what we implemented. At some point things became way too "hacky." I think if we simply had more pages and more people constructing explanations, it would have helped us answer a lot of the internal debates we had. But instead we were trying to wrangle a set of about 30 pages to work in just the right way. We should have left the feature as good enough and moved on. (But really, we should have been getting more users.)

Not only was it hard to find writers, but the explanations were hard to write as well. In general, writing modular explanations is very hard. Doubly so, when you also want to string those explanations together to form a coherent sequence.

Reflection: we were also trying to build a two-sided marketplace. We needed writers, but writers wanted readers, but readers wanted good content. I think the correct way to solve that would have been to attract people with existing blogs / readership to switch to Arbital and bring their audience with them.

Reflection: Team-wise we absolutely needed someone who would be going after users all the time and talking to them, recruiting them, marketing, etc… Nobody on the team had experience with or affinity for doing that.

To help us showcase the platform, Eliezer wrote the Bayes' Rule Guide. We've went through several iterations of it over the course of a few months, tweaking features and improving retention metrics. The somewhat dense set of pages helped us test a few features easier. Lots of people read the guide and loved it, but it wasn't obvious if Arbital format helped vs. Eliezer's writing was good. I think people also didn't appreciate the magic that happened behind the scene. (How do you communicate to a reader that they could have had a much worse reading experience but didn't?)

Nate Soares helped us by writing the Logarithm Guide. We thought if we could produce good sequences like that frequently and post them online, we might slowly get traction. Unfortunately, it's really time consuming to produce content that good, and there are just simply not that many people who can write content of that quality (and have the time to do it for free).

Here is what the front page looked like around that time. At the height of it, we had about a dozen regular users who would come and write a few pages every week. They enjoyed the small community we had and frequently hung out in our Slack channel. They wanted to write math explanations for themselves and their friends. I don't remember how many readers we had, but it was around 50-200 / day, most of them redirected from Eliezer's old guide to Bayes' Theorem.

In August we raised a $300k pre-seed round. We had about 9 investors. Most of them invested because of Eliezer, but a few knew me personally as well.

Also around that time, it became clear to us that things just weren't going well. The primary issue was that we completely relied on Eliezer to provide guidance to which features to implement and how to implement them. Frequently when we tried to do things our way, we were overruled. (Never without a decent reason, but I think in many of those cases either side had merit. Going with Eliezer's point of view meant we frequently ended up blocked on him, because we couldn't predict the next steps.) Also, since Eliezer was the only person seriously using the product, there wasn't enough rapid feedback for many of the features. And since we wasn't in the office with us every day, we were often blocked.

So, we decided to take the matter into our own hands. The three of us would decide what to do, and we would occasionally talk to Eliezer to get his input on specific things.

Reflection: Working with Eliezer was interesting to say the least. He certainly had a great overall vision for the product; one that I'm still astonished by to this day. He often had good insight into specific features and how to implement them. But sometimes he would get way too bogged down by certain details and spend longer on a feature than I thought was necessary. (In most of those cases he need things to work a certain way to solve a particular problem he had, but it was wasting our time because we were building something ultra specific to his use case.) This was especially painful for features that nobody, including Eliezer, would end up using.

Reflection: Eliezer also had a tendency sometimes to overcomplicate things and designs systems that I could barely wrap my head around. (I often joked that we would end up building a website that only one person in the world could use.) But then again, there were also many moments where a complicated, messy feature would suddenly click into place, and then it seemed obvious and simple.

Reflection: I'm tempted to draw a lesson like: never ever build a product you don't understand yourself. But if I did that, I'd certainly miss a huge learning opportunity of

working with Eliezer and leveling up my product and UX skills. So, instead, I think the lesson is: if you're running the project, never ever do anything that doesn't make sense to you. As soon as you start delegating / accepting things that don't make sense, you muddy the water for yourself. Now the strategy has opaque components that you don't understand, can't explain, and sometimes actively disagree with. There is just no way you can move at the necessary startup speed like that, and you're also not learning from your mistakes.

Reflection: This is especially true with respect to the overall strategy. Yes, maybe some paths are objectively better or easier. But if it's not one that makes sense to you, if it's not one you can execute, then you should take another path.

Reflection: Looking at arbital.com today, I'm actually still very much impressed with it. It's a good piece of software, and if I wanted to write explanations, I think I'd be hard pressed to find a better website. Ironically, the large part of what makes it really good are *all* the features that it has.

Chapter 3: Pivot to discussion

It was clear that we couldn't scale a community around math. So we decided to pivot. It wasn't a clean and easy pivot; if I remember correctly it took us about a month of struggling and deliberating to decide that our current approach wasn't working and then settle on a new one.

We decided to skip the Explanations part and go straight for the Discussion. We started build a new design around claims. A claim is a page with a proposition that users can vote on by assigning a probability estimate or by marking the degree of (dis)agreement. The idea was that people would blog on Arbital, and create pages for claims they discussed. People could vote on claims, and thus everyone could see where people mostly agreed and mostly disagreed. Claims could also be reused in other blog posts by the same author or other people.

We kept most of the original architecture, but remade the homepage. We also shifted the focus to the regional rationalist community. We did multiple user interviews. We did UI mockups. We talked to some rationalist bloggers and got some mild support.

One of my favorite artifacts to come out from that time period is this SlateStarCodex predictions page.

Reflection: At the time, I think the pivot decision was correct. And if we continued going with it, it's possible Arbital would have become LW 2.0, though that wasn't exactly our intention at the time.

Reflection: One thing we messed up during this time was diluting leadership. Since Eliezer was no longer in charge of product, the responsibility fell on all of us. This resulted in many many discussion about what to build, how to build it, down the minute details. Our pace really slowed down and it took us a while to patch it up.

Chapter 4: End of Arbital 1.0

In the beginning of 2017 I experienced my first burnout. There was simply no way I could work, so I apologized to the team and spent a month playing video games, which I desperately needed. This gave me the time, space, and distance to think about the project. When I came back, I sat down with the team and we had an extensive discussion about the direction of the company.

Eric and Steph wanted to stay the course. I no longer believed that was going to work, and I wanted to try a different approach. My biggest realization during my break was that people (and in this case specifically: most rationalists) were not actually interested in putting any serious effort in improving the state of online debate. While almost everyone *wanted* better online discussions, just like with math explanations, almost nobody was *willing* to put in any kind of work.

Furthermore, when we talked to a few rationalists directly, I just didn't get the feeling of genuine helpfulness or enthusiasm. This was upsetting, because there aren't that many big projects that the community does. So when I was doing Arbital, I guess I expected that more people would be on board, that more people would put in a bit of an extra effort to help us. But at best people put in minimal work (to satisfy us or themselves, I'm not sure). However, there was a limit to how upset I could be, because I very clearly recognized the same trait in me. So, while it's still a sad state of affairs, I'd be a hypocrite for being upset with any particular person.

Reflection: I think the rationality community can produce great insights, but mostly due to individual effort. There are great posts, but they rarely lead to prolonged conversations. And you very rarely see debates summarized for public consumption. (No wonder, it takes a lot of time and hard work!) There are a few counterexamples, but I think they prove the point by how much they stand out. (Best recent example I can think of is Jessica Taylor mediating a discussion between Eliezer and Paul Christiano and then writing it up.) (And, of course, not only do those things need to be written, but they also have to be read! And who has time to read…)

Reflection: I'm pleasantly surprised by the currently active LW 2.0. I think this is some evidence against my claim, but overall, I still think that when it comes to building out more detailed models with explicit claims, especially when it involves working with other people, most people are not willing to put in the extra work. (Especially if their name isn't attached to it.)

It was clear to me how to address this issue. People are willing to do what they are already doing. In particular: blogging. It didn't seem that hard to take the software we had and really optimize it to be a better blogging platform, at least for some audience (like math bloggers). And it seemed obvious to me that we would at least get some users that way. The key difference from our path at the time was that instead of solving the Discussion problem and trying to get people to do new things, we'd simply focus on building a better tool for a thing people already do. Then once we had people on our platform, we could help improve the ongoing discussions.

Reflection: This was me finally channeling the "greedy" user acquisition strategy.

At some point during the debate we considered trying both projects in parallel, but at the end, Eric and Steph decided to leave. I'd take Arbital in the new direction by

myself. (Huge thank you to Anna Salmon for helping to mediate that discussion. I’d say it went pretty well, all things considered.)

Chapter 5: Arbital 2.0

I spent the rest of 2017 working on Arbital 2.0. At first it was going very well. The vision felt very clear to me, I had my mind totally wrapped around it, and all parts of the strategy made sense. But for some reason, around summer of 2017 it became really hard to work. I spent a lot of the time trying to code, but being unable to. Even though intellectually I believed in the idea very much, my spirit was burned out / my System 1 just didn't believe in the project. After struggling with it on and off for the remaining half of the year, I finally had to admit to myself that it just didn't have enough momentum to succeed.

(The rest of this chapter is a **Reflection**.)

My best guess is that I was burnt out again. Even though I didn't feel as bad as I did in January, the feeling of being unable to even touch the laptop was very similar.

For those curious and for those looking for a startup idea, I'm going to describe my plan for Arbital 2.0. In short, it's Tumblr for mathematicians. You could use it as a blog, but it's really a social network. What makes it radically different is the ability for one person to create and own multiple topic-centered channels. (One big issue I see with FB is that it doesn't scale well with the number of friends. With most friends I only want to talk about certain specific topics. But FB is broadcast-to-all by design.) On Arbital 2.0, I would be able to post about improv, Rick and Morty, AI, scuba-diving, and all my other interests to different channels. People could subscribe to the channels they were interested in. So if you never wanted to listen to politics, you wouldn't follow people on their political channel. (Or hide all posts with #politics.) Each channel could be grown into a community, where other people could submit their posts too.

I still think this approach is very likely to work:

- Write the software. (I have a 70% baked version.)
- Go to mathematicians and offer to host their blog on Arbital. (Why mathematicians? There is basically no good blogging software with great support for LaTeX. That feature plus a few others will convince many people to switch or at least to try Arbital.) When I cold emailed 100+ math bloggers, I got a good number of pretty enthusiastic responses. Path for 0 to 1000 users seems straightforward.
- Most math bloggers also blog about other things. This naturally will lead them to use the channels feature.
- Many people have topics that they can't discuss on Facebook because they don't want to spam their friends. (I'd make at least one Rick and Morty post a day if I could.) Arbital would be the perfect outlet for those. (I'd assign 20% probability that it's possible to skip and succeed with this step directly without bothering with recruiting math bloggers first.)
- Then follow the original Arbital plan: Social Network -> Explanations -> Debate -> Notifications -> Rating -> Karma.
- And, of course, put the entire thing on a blockchain to make crazy money in the meantime. ;)

It's pretty clear to me that for Arbital to work at scale it *has* to be a social network. Part of why I don't think most other paths will work is that social media ate all the free time. It's not that people became lazy, it's just when it's a choice between spending

another 15 mins on FB or spending that 15 mins creating and linking claims to your new blog post, most people will choose FB. (And while FB is the best example, the problem is more widespread, of course. Everything became more addictive.) This is why the new approach was to create a social network that's better than FB and would allow you to manage your time and attention. And then from there we could actually put that saved time and attention to more useful things. (Although, I'm still sceptical that there are enough people who will have constructive debates to warrant all this effort.)

One reason I'm not pursuing this right now (aside from being burnt out with the whole enterprise) is that it no longer obviously helps with AI safety. If you recall, one of the assumptions was that if we did Arbital specifically for AI safety, the website wouldn't get enough credibility. With some recent developments, I think that's no longer the case. AI safety in general is more accepted, as well as MIRI's research in particular. So, if I did any work in this space again, I'd, ironically enough, go back to the original vision of creating an explanation website for Eliezer and other people who wanted to write about AI safety. (But, actually, I think the bottleneck is good content and people actually reading it.)

Tidbits

What's going to happen with Arbital?

I'm currently in the process of shutting down the company. All the software and IP is going to be turned over to MIRI. A few people expressed interest in having Arbital be open sourced, including one of our investors, so that's likely to happen.

Arbital tech stack

Arbital 1.0: Golang on BE, Angular 1.0 on FE, MySQL DB, markdown editor.
Arbital 2.0: NodeJS on BE, React on FE, ArangoDB, SlateJS editor. (Much much better choices all around!)

How much do you think technical skill mattered on the margin?

A lot. This was a pretty complex project, so managing code complexity was important. We also needed to continuously optimize things to make sure everything loaded decently fast. We had to make sure there weren't many bugs, because most things were user-facing. And being able to code decently fast helped a lot, since the amount of features we had to implement was fairly large.

Lawyers

I hated working with our lawyers. This was may be the most frustrating part of the entire project.

Lesson 1: work only with a team who is recommended to you personally by someone you trust *and* who has worked with them before.

Lesson 2: Ask how much something will take upfront. If the lawyer wants to spend more than an hour on anything, have them double check that you want it done. Have them send you a weekly report of time spent.

Lesson 3: Consider just not going with lawyers and using standard paperwork. Before Series A it just doesn’t matter and you can restructure things any way you want later.

Single founder

Being a single founder is not great, but there are actual reasons for it that in principle could be mitigated.

It’s unlikely you have all the skills. (Note that the situation is not that different if you have a co-founder, but they are very similar to you.) More important than the skills though, is your personality / inclination. Personally, I'd rather code than talk to users. So an ideal co-founder for me would be someone outgoing, who'd prefer talking to users than doing other things.

Not having someone to talk to day-to-day means you might end up with a tunnel vision / stuck / doing unimportant things / forgetting to take a step back / making basic mistakes. Having someone to talk to on frequent basis is important.

It feels bad when you don’t work / are stuck and the project doesn’t move forward. When you’re working with someone else, they usually make progress even when you don’t.

Responsibilities

It’s important that in each area there is a single person who is ultimately responsible for it. Product: one person. Business: one person. And, overall, for the company: one person. Assigning a responsibility to more than one person will significantly increase the communication overhead and slow things down.

Again, a heartfelt thank you to everyone who has participated in this adventure. It has been quite a journey and, at the end of the day, I wouldn’t take any of it back.

A LessWrong Crypto Autopsy

Wei Dai, one of the first people Satoshi Nakamoto contacted about Bitcoin, was a frequent Less Wrong contributor. So was Hal Finney, the first person besides Satoshi to make a Bitcoin transaction.

The first mention of Bitcoin on Less Wrong, a post called Making Money With Bitcoin, was in early 2011 - when it was worth 91 cents. Gwern predicted that it could someday be worth "upwards of $10,000 a bitcoin". He also quoted Moldbug, who advised that:

> If Bitcoin becomes the new global monetary system, one bitcoin purchased today (for 90 cents, last time I checked) will make you a very wealthy individual...Even if the probability of Bitcoin succeeding is epsilon, a million to one, it's still worthwhile for anyone to buy at least a few bitcoins now...I would not put it at a million to one, though, so I recommend that you go out and buy a few bitcoins if you have the technical chops. My financial advice is to not buy more than ten, which should be F-U money if Bitcoin wins.

A few people brought up some other points, like that if it ever became popular people might create a bunch of other cryptocurrencies, or that if there was too much controversy the Bitcoin economy might have to fork. The thread got a hundred or so comments before dying down.

But Bitcoin kept getting mentioned on Less Wrong over the next few years. It's hard to select highlights, but one of them is surely Ander's Why You Should Consider Buying Bitcoin Right Now If You Have High Risk Tolerance from January 2015. Again, people made basically the correct points and the correct predictions, and the thread got about a hundred comments before dying down.

I mention all this because of an idea, with a long history in this movement, that "rationalists should win". They should be able to use their training in critical thinking to recognize more opportunities, make better choices, and end up with more of whatever they want. So far it's been controversial to what degree we've lived up to that hope, or to what degree it's even realistic.

Well, suppose God had decided, out of some sympathy for our project, to make winning as easy as possible for rationalists. He might have created the biggest investment opportunity of the century, and made it visible only to libertarian programmers willing to dabble in crazy ideas. And then He might have made sure that all of the earliest adapters were Less Wrong regulars, just to make things extra obvious.

This was the easiest test case of our "make good choices" ability that we could possibly have gotten, the one where a multiply-your-money-by-a-thousand-times opportunity basically fell out of the sky and hit our community on its collective head. So how did we do?

I would say we did mediocre.

According to the recent SSC survey, 9% of SSC readers made $1000+ from crypto as of 12/2017. Among people who were referred to SSC from Less Wrong - my stand-in

for long-time LW regulars - 15% made over $1000 on crypto, nearly twice as many. A full 3% of LWers made over $100K. That's pretty good.

On the other hand, 97% of us - including me - didn't make over $100K. All we would have needed to do was invest $10 (or a few CPU cycles) back when people on LW started recommending it. But we didn't. How bad should we feel, and what should we learn?

Here are the lessons I'm taking from this.

1: Our epistemic rationality has probably gotten way ahead of our instrumental rationality

When I first saw the posts saying that cryptocurrency investments were a good idea, I agreed with them. I even Googled "how to get Bitcoin" and got a bunch of technical stuff that seemed like a lot of work. So I didn't do it.

Back in 2016, my father asked me what this whole "cryptocurrency" thing was, and I told him he should invest in Ethereum. He did, and centupled his money. I never got around to it, and didn't.

On the broader scale, I saw what looked like widespread consensus on a lot of the relevant Less Wrong posts that investing in cryptocurrency was a good idea. The problem wasn't that we failed at the epistemic task of identifying it as an opportunity. The problem was that not too many people converted that into action.

2: You can only predict the future in broad strokes, but sometimes broad strokes are enough

Gwern's argument for why Bitcoin might be worth $10,000 doesn't match what actually happened. He thought it would only reach that level if it became the world currency; instead it's there for...unclear reasons.

I don't count this as a complete failed prediction because it seems like he was making sort of the right mental motion - calculate the size of the best-case scenario, calculate the chance of that scenario, and realize there's no way Bitcoin wasn't undervalued under a broad range of assumptions.

3: Arguments-from-extreme-upside sometimes do work

I think Moldbug's comment aged the best of all the ones on the original thread. He said he had no idea what was going to happen, but recommended buying ten bitcoins. If Bitcoin flopped, you were out $10. If it succeeded, you might end up with some crazy stratospheric amount (right now, ten bitcoins = $116,000). Sure, this depends on an assumption that Bitcoin had more than a 1/10,000 chance of succeeding at this level, but most people seemed to agree that was true.

This reminds me of eg the argument for cryonics. Most LWers believe there's a less than 10% chance of cryonics working. But if it does work, you're immortal. Based on the extraordinary nature of the benefits, the gamble can be worth it even if the chances of success are very low.

We seem to be unusually fond of these arguments - a lot of people cite the astronomical scale of the far future as their reason for caring about superintelligent AI

despite the difficulty of anything we do affecting it. These arguments are weird-sounding, easy to dislike, and guaranteed to leave you worse off almost all the time.

But you only need one of them to be right before the people who take them end up better off than the people who don't. This decade, that one was Bitcoin.

Overall, if this was a test for us, I give the community a C and me personally an F. God arranged for the perfect opportunity to fall into our lap. We vaguely converged onto the right answer in an epistemic sense. And 3 - 15% of us, not including me, actually took advantage of it and got somewhat rich. Good work to everyone who succeeded. And for those of us who failed - well, the world is getting way too weird to expect there won't be similarly interesting challenges ahead in the future.

Babble

This is a linkpost for https://radimentary.wordpress.com/2018/01/10/babble/

This post is an exercise in "identifying with the algorithm." I'm a big fan of the probabilistic method and randomized algorithms, so my biases will show.

How do human beings produce knowledge? When we describe rational thought processes, we tend to think of them as essentially deterministic, deliberate, and algorithmic. After some self-examination, however, I've come to think that my process is closer to babbling many random strings and later filtering by a heuristic. I think verbally, and my process for generating knowledge is virtually indistinguishable from my process for generating speech, and also quite similar to my process for generating writing.

Here's a simplistic model of how this works. I try to build a coherent sentence. At each step, to pick the next word, I randomly generate words in the category (correct part of speech, relevance) and sound them out one by one to see which continues the sentence most coherently. So, instead of deliberately and carefully generating sentences in one go, the algorithm is something like:

1. Babble. Use a weak and local filter to randomly generate a lot of possibilities. Is the word the right part of speech? Does it lie in the same region of thingspace? Does it fit the context?
2. Prune. Use a strong and global filter to test for the best, or at least a satisfactory, choice. With this word in the blank, do I actually believe this sentence? Does the word have the right connotations? Does the whole thought read smoothly?

This is a babble about embracing randomness.

Baby Babble

Research on language development suggests that baby babble is an direct forerunner to language. You might imagine that infants learn by imitation, and that baby babble is just an imperfect imitation of words the baby hears, and progress occurs as they physiologically adapt to better produce those sounds. You would be wrong.

Instead, infants are initially capable of producing *all the phonemes* that exist in all human languages, and they slowly prune out which ones they need via reinforcement learning. Based on the sounds that their parents produce and respond to, babies slowly filter out unnecessary phonemes. Their babbles begin to drift as they prune out more and more phonemes, and they start to combine syllables into proto-words. Babble is the process of generating random sounds, and looking for clues about which ones are useful. Something something reinforcement learning partially observable Markov decision process I'm in over my head.

So, we've learned that babies use the Babble and Prune algorithm to learn language. But this is quite a general algorithm, and evolution is a conservative force. It stands to reason that human beings might learn other things by a similar algorithm. I don't think it's a particularly controversial suggestion that human thought proceeds roughly by

cheaply constructing a lot of low-resolution hypotheses and then sieving from them by allowing them to play out to their logical conclusions.

The point I want to emphasize is that the algorithm has two distinct phases, both of which can be independently optimized. The stricter and stronger your Prune filter, the higher quality content you stand to produce. But one common bug is related to this: if the quality of your Babble is much lower than that of your Prune, you may end up with nothing to say. Everything you can imagine saying or writing sounds cringey or content-free. Ten minutes after the conversation moves on from that topic, your Babble generator finally returns that witty comeback you were looking for. You'll probably spend your entire evening waiting for an opportunity to force it back in.

Your pseudorandom Babble generator can also be optimized, and in two different ways. On the one hand, you can improve the weak filter you're using, to increase the probability of generating higher-quality thoughts. The other way is one of the things named "creativity": you can try to eliminate systematic biases in the Babble generator, with the effect of hitting a more uniform subset of relevant concept-space. Exercises that might help include expanding your vocabulary, reading outside your comfort zone, and engaging in the subtle art of nonstandard sentence construction.

Poetry is Babble Study

Poetry is at its heart an isolation exercise for your Babble generator. When creating poetry, you replace your complex, inarticulate, and highly optimized Prune filter with a simple, explicit, and weird one that you're not attached to. Instead of picking words that maximize meaning, relevance, or social signals, you pick words with the right number of syllables that rhyme correctly and follow the right meter.

Now, with the Prune filter simplified and fixed, all the attention is placed on the Babble. What does it feel like to write a poem (not one of those free-form modern ones)? Probably most of your effort is spent Babbling almost-words that fit the meter and rhyme scheme. If you're anything like me, it feels almost exactly like playing a game of Scrabble, fitting letters and syllables onto a board by trial and error. Scrabble is just like poetry: it's all about being good at Babble. And no, I graciously decline to write poetry in public, even though Scrabble does conveniently rhyme with Babble.

Puns and word games are Babble. You'll notice that when you Babble, each new word isn't at all independent from its predecessors. Instead, Babble is more like initiating a random walk in your dictionary, one letter or syllable or inferential step at a time. That's why word ladders are so appealing - because they stem from a natural cognitive algorithm. I think Scott Alexander's writing quality is great partly because of his love of puns, a sure sign he has a great Babble generator.

If poetry and puns are phonetic Babble, then "Deep Wisdom" is semantic Babble. Instead of randomly arranging words by sound, we're arranging a rather small set of words to sound wise. More often than not, "deep wisdom" boils down to word games anyway, e.g. wise old sayings:

"A blind person who sees is better than a seeing person who is blind."

"A proverb is a short sentence based on long experience."

"Economy is the wealth of the poor and the wisdom of the rich."

Reading is Outsourcing Babble

Reading and conversation outsource Babble to others. Instead of using your own Babble generator, you flood your brain with other people's words, and then apply your Prune filter. Because others have already Pruned once, the input is particularly high-quality Babble, and you reap particularly beautiful fruit. How many times have you read a thousand-page book, only to fixate on a handful of striking lines or passages?

Prune goes into overdrive when you outsource Babble. A bug I mentioned earlier is having way too strict of a Prune filter, compared to the quality of your Babble. This occurs particularly to people who read and listen much more than they write or speak. When they finally trudge into the attic and turn on that dusty old Babble generator, it doesn't produce thoughts nearly as coherent, witty, or wise as their hyper-developed Prune filter is used to processing.

Impose Babble tariffs. Your conversation will never be as dry and smart as something from a sitcom. If you can't think of anything to say, relax your Prune filter at least temporarily, so that your Babble generator can catch up. Everyone starts somewhere - Babbling platitudes is better than being silent altogether.

Conversely, some people have no filter, and these are exactly the kind of people who don't read or listen enough. If all your Babble goes directly to your mouth, you need to install a better Prune filter. Impose export tariffs.

The reason the Postmodernism Generator is so fun to read is because computers are now capable of producing great Babble. Reading poetry and randomly generated postmodernism, talking to chatbots, these activities all amount to frolicking in the uncanny valley between Babble and the Pruned.

Tower of Babble

A wise man once said, "Do not build Towers out of Babble. You wouldn't build one out of Pizza, would you?"

NP

NP is the God of Babble. His law is: humans will always be much better at verifying wisdom than producing it. Therefore, go forth and Babble! After all, how did Shakespeare write his famous plays, except by randomly pressing keys on a keyboard?

NP has a little brother called P. The law of P is: never try things you don't understand completely. Randomly thrashing around will get you nowhere.

P believes himself to be a God, an equal to his brother. He is not.

Announcement: AI alignment prize winners and next round

We (Zvi Mowshowitz, Vladimir Slepnev and Paul Christiano) are happy to announce that the AI Alignment Prize is a success. From November 3 to December 31 we received over 40 entries representing an incredible amount of work and insight. That's much more than we dared to hope for, in both quantity and quality.

In this post we name six winners who will receive $15,000 in total, an increase from the originally planned $5,000.

We're also kicking off the next round of the prize, which will run from today until March 31, under the same rules as before.

The winners

First prize of $5,000 goes to Scott Garrabrant (MIRI) for his post Goodhart Taxonomy, an excellent write-up detailing the possible failures that can arise when optimizing for a proxy instead of the actual goal. Goodhart's Law is simple to understand, impossible to forget once learned, and applies equally to AI alignment and everyday life. While Goodhart's Law is widely known, breaking it down in this new way seems very valuable.

Five more participants receive $2,000 each:

- Tobias Baumann (FRI) for his post Using Surrogate Goals to Deflect Threats. Adding failsafes to the AI's utility function is a promising idea and we're happy to see more detailed treatments of it.
- Vanessa Kosoy (MIRI) for her work on Delegative Reinforcement Learning (1, 2, 3). Proving performance bounds for agents that learn goals from each other is obviously important for AI alignment.
- John Maxwell (unaffiliated) for his post Friendly AI through Ontology Autogeneration. We aren't fans of John's overall proposal, but the accompanying philosophical ideas are intriguing on their own.
- Alex Mennen (unaffiliated) for his posts on legibility to other agents and learning goals of simple agents. The first is a neat way of thinking about some decision theory problems, and the second is a potentially good step for real world AI alignment.
- Caspar Oesterheld (FRI) for his post and paper studying which decision theories would arise from environments like reinforcement learning or futarchy. Caspar's angle of attack is new and leads to interesting results.

We'll be contacting each winner by email to arrange transfer of money.

We would also like to thank everyone who participated. Even if you didn't get one of the prizes today, please don't let that discourage you!

The next round

We are now announcing the next round of the AI alignment prize.

As before, we're looking for technical, philosophical and strategic ideas for AI alignment, posted publicly between now and March 31, 2018. You can submit your entries in the comments here or by email to apply@ai-alignment.com. We may give feedback on early entries to allow improvement, though our ability to do this may become limited by the volume of entries.

The minimum prize pool this time will be $10,000, with a minimum first prize of $5,000. If the entries once again surpass our expectations, we will again increase that pool.

Thank you!

(Addendum: I've written a post summarizing the typical feedback we've sent to participants in the previous round.)

Beware of black boxes in AI alignment research

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Over the course of the AI Alignment Prize I've sent out lots of feedback emails. Some of the threads were really exciting and taught me a lot. But mostly it was me saying pretty much the same thing over and over with small variations. I've gotten pretty good at saying that thing, so it makes sense to post it here.

Working on AI alignment is like building a bridge across a river. Our building blocks are ideas like mathematics, physics or computer science. We understand how they work and how they can be used to build things.

Meanwhile on the far side of the river, we can glimpse other building blocks that we *imagine* we understand. Desire, empathy, comprehension, respect... Unfortunately we don't know how these work inside, so from the distance they look like *black boxes*. They would be very useful for building the bridge, but to reach them we must build the bridge first, starting from our side of the river.

What about machine learning? Perhaps the math of neural networks could free us from the need to understand the building blocks we use? If we could create a *behavioral imitation* of that black box over there (like "human values"), then building the bridge would be easy! Unfortunately, ideas like adversarial examples, treacherous turns or nonsentient uploads show that we shouldn't bet our future on something that imitates a particular black box, even if the imitation passes many tests. We need to understand how the black box works inside, to make sure our version's behavior is not only similar but based on the right reasons.

(Eliezer made the same point more eloquently a decade ago in Artificial Mysterious Intelligence. Still, with the second round of our prize now open, I feel it's worth saying again.)

The Loudest Alarm Is Probably False

Epistemic Status: Simple point, supported by anecdotes and a straightforward model, not yet validated in any rigorous sense I know of, but IMO worth a quick reflection to see if it might be helpful to you.

A curious thing I've noticed: among the friends whose inner monologues I get to hear, the most self-sacrificing ones are frequently worried they are being too selfish, the loudest ones are constantly afraid they are not being heard, the most introverted ones are regularly terrified that they're claiming more than their share of the conversation, the most assertive ones are always suspicious they are being taken advantage of, and so on. It's not just that people are sometimes miscalibrated about themselves- it's as if the loudest alarm in their heads, the one which is apt to go off at any time, is pushing them in the exactly wrong direction from where they would flourish.

Why should this be? (I mean, presuming that this pattern is more than just noise and availability heuristic, which it could be, but let's follow it for a moment.)

It's exactly what we should expect to happen if (1) the human psyche has different "alarms" for different social fears, (2) these alarms are supposed to calibrate themselves to actual social observations but occasionally don't do so correctly, and (3) it's much easier to change one's habits than to change an alarm.

In this model, while growing up one's inner life has a lot of alarms going off at various intensities, and one scrambles to find actions that will calm the loudest ones. For many alarms, one learns habits that basically work, and it's only in exceptional situations that they will go off loudly in adulthood.

But if any of these alarms don't calibrate itself correctly to the signal, then they eventually become by far the loudest remaining ones, going off all the time, and one adjusts one's behavior as far as possible in the other direction in order to get some respite.

And so we get the paradox, of people who seem to be incredibly diligently following the exact wrong advice for themselves, analogous to this delightful quote (hat tip Siderea) about system dynamics in consulting:

> People know intuitively where leverage points are. Time after time I've done an analysis of a company, and I've figured out a leverage point — in inventory policy, maybe, or in the relationship between sales force and productive force, or in personnel policy. Then I've gone to the company and discovered that there's already a lot of attention to that point. Everyone is trying very hard to push it IN THE WRONG DIRECTION!

The funny thing about cognitive blind spots (and that's what we're looking at here) is that you can get pretty far into reading an article like this, hopefully enjoying it along the way, and forget to ask yourself if the obvious application to your own case might be valid.

If so, no worries! I developed this idea an embarrassingly long time before I thought to ask myself what would be the constant alarm going off in my own head. (It was the

alarm, "people aren't understanding you, you need to keep explaining", which was a huge epiphany to me but blindingly clear to anyone who knew me.)

And the framing that helped me instantly find that alarm was as follows:

What do I frequently fear is going wrong in social situations, despite my friends' reliable reassurance that it's not?

That fear is worth investigating as a possibly broken alarm.

Teaching Ladders

This is a linkpost for https://radimentary.wordpress.com/2018/01/24/teaching-ladders/

I've been teaching math to people one or two levels below me my entire life. Although this seems like a limitation, I think it's the natural state of affairs.

On the Kiseido Go Server (KGS), there's a room called the KGS Teaching Ladder where players can find teaching games with players just a few stones stronger than them. The few times I participated, it was extraordinarily positive. Because of the relative linearity of progression in Go, losing to a slightly stronger player is legible: they will usually play a move you considered but just barely don't understand, or find the simplest good moves that you don't know yet. Losing to a much stronger player, however, is completely illegible. Much stronger players will often play completely incorrect moves ("overplays") just to test your instincts, or play otherwise incomprehensibly complicated variations and traps that you immediately fall into.

I distinguish between two models of teaching:

1. (Traditional) The master teaches everyone.
2. (Teaching Ladder) The students one or two stages up from you teach you.

Previously, I noted that many progressions come in three stages: "naive, cynical, naive but wise," where the third stage bears more resemblance to the first than the second. The value of Teaching Ladders is that they naturally mesh with the three stages: Stage 3's have a difficult time teaching Stage 1's, and Stage 2's are needed to fill that gap.

Scaffolding and Assimilation

> *The history of every major galactic civilization tends to pass through three distinct and recognizable phases, those of Survival, Inquiry and Sophistication, otherwise known as the How, Why, and Where phases. For instance, the first phase is characterized by the question 'How can we eat?', the second by the question 'Why do we eat?' and the third by the question, 'Where shall we have lunch?'* (Douglas Adams, "The Hitchhiker's Guide to the Galaxy")

In Singularity Mindset, I articulated the following model of development without explaining its origins:

> Oftentimes, progress curves look like "naive, cynical, naive but wise":
> For mathematicians, the curve is pre-rigor, rigor, post-rigor.
> Picasso said, "It took me four years to paint like Raphael, but a lifetime to paint like a child."
> Scott Alexander foretold that idealism is the new cynicism.
> Knowing about biases can hurt you.

This is a general phenomenon which applies not just at the level of an entire field, but also at the level of individual skills. With Terry's example of mathematics in mind, the stages look like:

1. (naive) The student has bad instincts. He thinks that proof by example is a valid argument. Saying "trust your instincts" doesn't help and deeply frustrates him.

Progress is achieved by dropping the instincts, acquiring explicit knowledge, and following fixed and deliberate rules.
2. (cynical) The student has the knowledge and understands the fixed and deliberate rules. He starts every proof with a cookie-cutter template for proof by induction or proof by contradiction and fills in the logic line by line. Unfortunately, doing everything by System 2 is slow and clunky. Progress is achieved by pushing acquired knowledge back down to System 1 via practice, metaphor, and exploration.
3. (naive but wise) The student has successfully integrated skills into System 1. He produces intuitive arguments that only mention the salient details. A completely rigorous proof can be reconstructed on demand, but requires effort. At this point the explicit structures originally built to progress to Stage 2 are unnecessary, and are slowly taken down.

This model resembles - and perhaps generalizes - the interaction of Babble and Prune, where conscious Prune filters are slowly pushed down into the subconscious Babble. Learning occurs as superior algorithms are constructed in System 2 and then pushed back down to System 1, the instinctual level. After the algorithm is constructed, however, the remaining machinery in System 2 is outdated scaffolding. The farther along a student is past Stage 2, the more of this scaffolding is forgotten.

Let's call the transition between Stage 1 and Stage 2 Scaffolding and the transition between Stage 2 and Stage 3 Assimilation. Every progression in every domain looks roughly like a ladder built out of alternating Scaffolding and Assimilation rungs. In the Scaffolding stage, bad instincts are explicitly corrected with procedure and hard-and-fast rules. In the Assimilation stage, the explicit Scaffolding built is now practiced and stretched until it becomes instinct. Afterwards, although there is rarely an explicit call to remove the Scaffolding, it is no longer in use and slowly crumbles, leaving only the pure instinct behind.

Teach Scaffolding

In a traditional teaching model, the master teaches students at all levels of development, from precalculus to (infinity, 1)-categories. The basic pitfall to this model is usually described as Expecting Short Inferential Distances, i.e. that the master has a hard time reaching back down the tall tower of inferences to meet her students. She may even be a great speaker, throwing down her instincts in the way of quirky metaphors in an attempt to boost her students up. But she is no Rapunzel and the students are left staring up longingly from the bottom of the tower. Every so often, one of them tries to hop up and make progress by asking what that symbol means, but the tower is too damn high.

Long inferential distances are certainly part of the problem, but even if the master is sufficiently humble to back down a hundred steps, she may lack key pieces of Scaffolding that are required to convey ideas to the students. A master looks down the tower of inference and sees only the transition between Stage 1 and Stage 3, as if Assimilation can be achieved without the Scaffolding. She would never dream of teaching proof by contradiction with a cookie-cutter mad-lib proof template, but that seems to be an effective starting point for students who've never handled proofs.

Teaching Ladders, on the other hand, not only reduce inferential distance but introduce teachers who still have their Stage 2 Scaffolding mostly intact. That's why (in my experience) the TA in an undergraduate-level math course is usually more

effective than the lecturer. Unless they are explicitly trained in teaching, even lecturers who understand and correct for the inferential gulfs involved lack the mental machinery to convey Scaffolding.

Often when I'm asked to teach a concept I recently learned, I have the urge to punt it to a known master's writing or lectures. Today I'm learning to fight that urge. However incomplete my knowledge, I convey it with Scaffolding intact, and that will do more good than harm.

Dispel your justification-monkey with a “HWA!”

I'm going to use a couple of words in this post that might not be immediately clear to some people. One of them is "justification". Another is "acceptance". I would like to suggest that if you think I'm saying something stupid when I'm using those words, that you instead consider what meaning I might be using for those words such that I'm not saying something stupid. My meanings, I think, are pretty clear if you look for them.

If you want more detail on "justification", see this blog post on Causal Explanations vs Normative Explanations for an in-depth explanation.

Justification—ie a normative explanation as opposed to a causal one—is sometimes necessary. But, for many of us, it’s necessary much less often than we feel it is.,

The reason we justify more often than we need to is that we live in fear of judgment, from years having to explain to authorities (parents, teachers, bosses, cops (for some people)) why things went differently than they “should have”. This skill is necessary to avoid punishment from those authorities.

We often offer justifications before they’re even asked for: “Wait I can explain—”

With friends, though, or in a healthy romantic partnership, or with people that we have a solid working relationship with, it is quite apparent that this flinch towards justification is actually in the way of being able to effectively work together. It is:

- unhelpful for actually understanding what happened (since it’s a form of rationalization, ie motivated cognition)
- an obstacle to feeling safe with each other
- a costly waste of time & attention

And yet we keep feeling the urge to justify. So what to do instead? How to re-route that habit in a way that builds trust within the relationships where justification isn’t required? How to indicate to our conversational partners that we aren't demanding that they justify?

There are lots of ways to do this—here’s one.

Fundamentally, the issue with justification is that it’s an attempt to explain why we’re in the world that we’re in, as opposed to being in the world that we *“should be”* in. It indicates, therefore, that there’s a lack of acceptance of *what is already real* . There might be causal reasons to be understood about why what happened happened and how to do things differently in the future, but in order to actually See those without rationalizing, you first need to accept where you in fact are.

This is where “HWA!” comes in. In all caps, it sounds like a *kiai* , which is the japanese martial arts term for a short yell used when performing an strike. Since "hwa" is a mental move that can dispel the justification monkeys that get in the way of us co-thinking clearly, it kind of *is* a *kiai!* However, in this case the power of the sound comes from the power of the meaning.

“HWA” stands for “Here We Are”

And "here we are" is a simple, pure expression of acceptance of the reality of the present as it is. This pops out of the justification frame.

(For those tempted to get up in arms about my use of the term "acceptance", I don't mean the thing you think I mean. "Acceptance," as I'm using it, doesn't mean "allowing things to stay the same", it means not having a sense of "this should *already* be different than it in fact is". It's too late. The present is already here! Refusing to perform this kind of acceptance here is a form of denial or regret or some other confusion.)

Here’s how to use it:

Example 1: self

The simplest situation is actually to use it with yourself. Perhaps if you catch your inner rebel attempting to justify something to your inner dictator, or just more generally you find yourself feeling that *things should not be as they are* . Here are some sample phrasings:

> “Ahh shit I forgot to do the thing. Well, hwa.”
>
> “Man I really haven’t been focused today. *breathe* …hwa.”
>
> “What?! It’s noon already? I was gonna get up at—hwa. Here we are. It is noon. What now?”

In all of these cases, you're taking a moment to let in the reality of *what is so.* This step is a necessary precondition to figuring out *what is next.* Without it, you're just flailing.

Example 2: avoiding interpersonal blame

> **Alice**: Man the grocery store was crowded. Anyway, here’s the Thing you said to pick up.
>
> **Bob**: Oh, no, this is X Thing, this won’t work at all for the Recipe. It needs to be Y Thing.
>
> **Alice or Bob**: So HWA! What do we do now?

What this does is dispel the need to figure out whose fault it was that Alice got X Thing instead of Y Thing. Did Bob assume Alice should have known that Recipe needed Y not X? Was that a reasonable assumption? Did Bob explicitly say "Y Thing", and Alice didn't hear? Did Bob use a synonym for Y Thing that Alice didn't recognize? etc, etc, etc.

The answers to these questions may or may not be important, but in any case, justification is a form of rationalization and is therefore not a way to actually find the *real* answers to those questions.

Also, while acute time pressure stresses out your inner monkeys and tends to increase blame in these situations, the blame and justification monkey patterns are going to be in the way of actually resolving the situation by going out to get some Y Thing or finding an alternative approach, which is important when time is of the essence!

Example 3: avoiding implicit accusation

What do you do when you're already partway into an explanation (which could easily become a justification or defense?) You can use "hwa" for that too.

If you've got a decent amount of flow with the other person, you can sometimes combine "hwa" with a small piece of information, to communicate your understanding of a situation without that information acting as bait for the justification monkey.

You can imagine a conversation that happened a few hours prior to the one in Example 2:

> **Alice**: Anyway, before you go back to work, shall we check in about any food you might want me to pick up for the party tomorrow?
>
> **Bob**: Oh the party is moved to tomorrow? I missed that.
>
> **Alice**: I think I mentioned it… but hwa. Anyway, groceries? Snacks?

It's mildly helpful info for the parties to have common knowledge that Alice believes she said it already. But without the "hwa", Alice's initial impulse to say "I think I mentioned it" can easily be read as a defense on Alice's part, implying an accusation of B.

From this space of opposition, if B is inclined towards self-blame, they might be tempted to say *"Sorry! I should have listened more closely when you were talking about the party earlier. I was distracted because I was wondering if Jamie was gonna be there."* Conversely, if they are inclined towards other-blame, they might be tempted to protest *"I'm pretty sure you didn't! Or if you did then you should have made sure I was listening when you said it..."*

What Alice saying "here we are" indicates here to Bob is "whether I said it or not, whether you heard it or not, it's *okay.* We don't need to fight."

In Practice

One practical question is: does it make sense to say "hwa" or to say a full on "here we are"?

The short answer is: use both.

In cases where there isn't common knowledge about the "hwa" concept, you'll naturally want to say "here we are" so that people have the slightest idea what you're saying.

And honestly, even if people never used the abbreviation, I still think that crystallizing this concept as “hwa” makes it more memorable as a move and therefore also easier to transmit as a meme.

In addition, I think that in text-based conversations (ie where the spelling of “hwa” is easy to see but tone of voice is not) hwa has a particular couple of advantages, provided that both conversationalists know what it means already. (Maybe use the full phrase “here we are” the first time, and then also link to this post.)

- It satisfies the Sparkly Pink Purple Ball Criterion. “Here we are” can mean lots of things, but “hwa” is not a phrase likely to be used by someone who hasn’t read this post (or heard about this usage), which means it will prompt awareness more.:
- Relatedly, when talking aloud, the tone of voice used to say "here we are" matters. In the context of chat conversation, there's no tone, so it helps to have a more distinct phrase
- It is quick to type! In a fast-paced chat conversation, typing 3 characters instead of 11 makes a difference. It makes the thing easier to say and also means that it can more easily pop in as a kind of interruption of what seems likely to lead to justification.

Note that it's not enough to say the phrase. **You have to mean it.** You have to actually accept what has happened and allow yourself to be with what's real in the present. If you need help letting go of your concept of how things should have gone, read Transcending Regrets, Problems, and Mistakes.

Also, a general tip with stuff like this: if someone tries to use “hwa” and either party slips up and you ends up getting pulled into a justification loop, *that’s okay too* . It’s part of the learning process. *And,* if the trust is available, it is important also to give them feedback if they seem to be saying one thing (”no need for justification”) and meaning another (”but things are not okay unless I get a justification”). Just... give that feedback once you’re already back on the same page, feeling good together, not while the justification pattern is live.

This post is crossposted from malcolmocean.com.

Aliveness

This is a linkpost for https://sinceriously.fyi/aliveness/

Why everything might have taken so long

I asked why humanity took so long to do anything at the start, and the Internet gave me its thoughts. Here is my expanded list of hypotheses, summarizing from comments on the post, here, and here.

Inventing is harder than it looks

1. **Inventions are usually more ingenious than they seem.** Relatedly, reality has a lot of detail.
2. **There are lots of apparent paths**: without hindsight, you have to waste a lot of time on dead ends.
3. **People are not as inventive as they imagine.** For instance, I haven't actually invented anything – why do I even imagine I could invent rope?
4. **Posing the question is a large part of the work.** If you have never seen rope, it actually doesn't occur to you that rope would come in handy, or to ask yourself how to make some.
5. **Animals (including humans) mostly think by intuitively recognizing over time what is promising and not among affordances they have**, and reading what common observations imply. New affordances generally only appear by some outside force e.g. accidentally. To invent a thing, you have to somehow have an affordance to make it even though you have never seen it. And in retrospect it seems so obvious because now you do have the affordance.

People fifty thousand years ago were not really behaviorally modern

1. **People's brains were actually biologically less functional fifty thousand years ago.**
2. **Having concepts in general is a big deal**. You need a foundation of knowledge and mental models to come up with more of them.
3. **We lacked a small number of unimaginably basic concepts** that it is hard to even imagine not having now. For instance 'abstraction', or 'changing the world around you to make it better'.
4. **Having external thinking tools is a big deal.** Modern 'human intelligence' relies a lot on things like writing and collected data, that aren't in anyone's brain.
5. **The entire mental landscapes of early people was very different**, as Julian Jaynes suggests. In particular, they lacked self awareness and the ability to have original thought rather than just repeating whatever they usually repeat.

Prerequisites

1. **Often A isn't useful without B, and B isn't useful without A.** For instance, A is chariots and B is roads.
2. **A isn't useful without lots of other things**, which don't depend on A, but take longer to accrue than you imagine.
3. **Lots of ways to solve problems don't lead to great things in the long run.** 'Crude hacks' get you most of the way there, reducing the value of great inventions.

Nobody can do much at all

1. **People in general are stupid in all domains, even now.** Everything is always mysteriously a thousand times harder than you might think.
2. **Have I tried even *making* rope from scratch?** Let alone inventing it?

People were really busy

1. **Poverty traps.** Inventing only pays off long term, so for anyone to do it you need spare wealth and maybe institutions for capital to fund invention.
2. **People are just really busy doing and thinking about other things.** Like mating and dancing and eating and so on.

Communication and records

1. **The early humans did have those things, we just don't have good records.** Which is not surprising, because our records of those times are clearly very lacking.
2. **Things got invented a lot, but communication wasn't good/common enough to spread them.** For instance because tribes were small and didn't interact that much).

Social costs

1. **Technology might have been seen as a sign of weakness or laziness**
2. **Making technology might make you stand out rather than fit in**
3. **Productivity shames your peers and invites more work from you**
4. **Inventions are sometimes against received wisdom**

Population

1. **There were very few people in the past**, so the total thinking occurring between 50k and 28k years ago was less than in the last hundred years.

Value

1. **We didn't invent things until they became relevant at all, and most of these things aren't relevant to a hunter-gatherer.**
2. **Innovation is risky**: if you try a new thing, you might die.

Orders of invention

1. **First order inventions** are those where the raw materials are in your immediate surroundings, and they don't require huge amounts of skill. My intuition is mostly that first order inventions should have been faster. But maybe we did get very good at first order ones quickly, but it is hard to move to higher orders.
2. **You need a full-time craftsman to make most basic things to a quality where they are worth having**, and we couldn't afford full-time craftsmen for a very long time.
3. **Each new layer requires the last layer of innovation be common enough that it is available everywhere**, for the next person to use.

Global online debate on the governance of AI

Hi guys,

For background, I'm a French EA, attended a CFAR workshop, and recently decided to work on AI policy as it is a pressing and neglected issue. I've been working for The Future Society for a few weeks already and would like to share with you this opportunity to impact policy-making. The Future Society is a Harvard Kennedy School-incubated think tank dedicated to the governance of emerging advanced technologies. It has partnerships with the Future of Life Institute and the Centre for the Study of Existential Risk.

The think-tank provides an participatory debate platform to people all around the world

The objective is to craft actionable and ethical policies that will be delivered in a White Paper, to the White House, the OECD, the European Union and other policymaking institutions that the think-tank is working with.

Because we know AI policy is hard, the idea is to use collective intelligence to provide innovative and reasonable policies. The debate is hosted on an open source collective intelligence software resulting from a research project funded by the European Commission, technologically supported by MIT. It's based on research on collective intelligence, going from open and exploratory questions to more in-depth discussions. Right now, we are in the "Ideation" phase, which is very open. You can make constructive answers and debate with other people who are also interested in crafting AI Policies with instant translation.

The platform is like an online forum articulated around several issues, both short-term and long-term oriented. You have six themes, including "AI Safety and Security", "Reinvent Man & Machine Relationship" and "Governance Framework".

So far, most of the answers have been very constructive. But with you guys... it can be even better.

Because you are Rationalists, I really wanted to pick your brains to think rationally and critically about AI governance.

It would be great if you guys could participate, on the topic you're most interested in, knowing that a) it will be impactful b) you will be able to challenge your thoughts with other people passionate about AI social impacts. Of course, you don't have to talk about AI safety if you'd rather focus on other topics.

Don't hesitate to post short (yet insightful) comments to start, just to fuel the debate.

So please connect on the debate, and participate.

The debate is here

Learning to make better decisions

This is a linkpost for https://sites.google.com/site/falklieder/practical-rationality-blog/mcrl

A model I use when making plans to reduce AI x-risk

I've been thinking about what implicit model of the world I use to make plans that reduce x-risk from AI. I list four main gears below (with quotes to illustrate), and then discuss concrete heuristics I take from it.

A model of AI x-risk in four parts

1. Alignment is *hard.*

Quoting "Security Mindset and the Logistic Success Curve" (link)

> Coral: YES. Given that this is a novel project entering new territory, expect it to take *at least* two years more time, or 50% more development time—whichever is less—compared to a security-incautious project that otherwise has identical tools, insights, people, and resources. And that is a very, very optimistic lower bound.
>
> Amber: This story seems to be heading in a worrying direction.
>
> Coral: Well, I'm sorry, but creating robust systems takes longer than creating non-robust systems even in cases where it would be really, extraordinarily bad if creating robust systems took longer than creating non-robust systems.

2. Getting alignment right accounts for most of the variance in whether an AGI system will be positive for humanity.

Quoting "The Hidden Complexity of Wishes" (link)

> There are three kinds of genies: Genies to whom you can safely say "I wish for you to do what I should wish for"; genies for which *no* wish is safe; and genies that aren't very powerful or intelligent.
>
> [...]
>
> There is no safe wish smaller than an entire human morality. There are too many possible paths through Time. You can't visualize all the roads that lead to the destination you give the genie... any more than you can program a chess-playing machine by hardcoding a move for every possible board position.
>
> And real life is far more complicated than chess. You cannot predict, in advance, which of your values will be needed to judge the path through time that the genie takes. Especially if you wish for something longer-term or wider-range than rescuing your mother from a burning building.

3. Our current epistemic state regarding AGI timelines will continue until we're close (<2 years from) to having AGI.

Quoting "There is No Fire Alarm for AGI" (link)

> It's not that whenever somebody says "fifty years" the thing always happens in two years. It's that this confident prediction of things being far away corresponds to an epistemic state about the technology that feels the same way internally until you are very very close to the big development. It's the epistemic state of "Well, I don't see how to do the thing" and sometimes you say that fifty years off from the big development, and sometimes you say it two years away, and sometimes you say it while the Wright Flyer is flying somewhere out of your sight.
>
> [...]
>
> So far as I can presently estimate, now that we've had AlphaGo and a couple of other maybe/maybe-not shots across the bow, and seen a huge explosion of effort invested into machine learning and an enormous flood of papers, we are probably going to occupy our present epistemic state until very near the end.
>
> By saying we're probably going to be in roughly this epistemic state until almost the end, I *don't* mean to say we know that AGI is imminent, or that there won't be important new breakthroughs in AI in the intervening time. I mean that it's hard to guess how many further insights are needed for AGI, or how long it will take to reach those insights. After the next breakthrough, we still won't know how many more breakthroughs are needed, leaving us in pretty much the same epistemic state as before. Whatever discoveries and milestones come next, it will probably continue to be hard to guess how many further insights are needed, and timelines will continue to be similarly murky.

4. Given timeline uncertainty, it's best to spend marginal effort on plans that assume / work in shorter timelines.

Stated simply: If you don't know when AGI is coming, you should make sure alignment gets solved in worlds where AGI comes soon.

Quoting "Allocating Risk-Mitigation Across Time" (link)

> Suppose we are also unsure about when we may need the problem solved by. In scenarios where the solution is needed earlier, there is less time for us to collectively work on a solution, so there is less work on the problem than in scenarios where the solution is needed later. Given the diminishing returns on work, that means that a marginal unit of work has a bigger expected value in the case where the solution is needed earlier. This should update us towards working to address the early scenarios more than would be justified by looking purely at their impact and likelihood.
>
> [...]
>
> There are two major factors which seem to push towards preferring more work which focuses on scenarios where AI comes soon. The first is nearsightedness: we simply have a better idea of what will be useful in these scenarios. The second is diminishing marginal returns: the expected effect of an extra year of work on a problem tends to decline when it is being added to a larger total. And because there is a much larger time horizon in which to solve it (and in a wealthier world), the problem of AI safety when AI comes later may receive many times as much work as the problem of AI safety for AI that comes soon. On the other hand one more factor preferring work on scenarios where AI comes later is the ability to pursue more leveraged strategies which eschew object-level work today in favour of generating (hopefully) more object-level work later.

The above is a slightly misrepresentative quote; the paper is largely undecided as to whether shorter term strategies or longer term strategies are more valuable (given uncertainty over timelines), and recommends a portfolio approach (running multiple strategies, that each apply to different timelines). Nonetheless when reading it I did update toward short-term strategies as being especially neglected, both by myself and the x-risk community at large.

Concrete implications

Informed by the model above, here are heuristics I use for making plans.

- **Solve alignment! Aaargh! Solve it! Solve it now!**
  - I nearly forgot to say it explicitly, but it's the most important: if you have a clear avenue to do good work on alignment, or field-building in alignment, do it.
- **Find ways to contribute to intellectual progress on alignment**
  - I think that intellectual progress is very tractable.
    - A central example of a small project I'd love to see more people attempt, is people writing up (in their own words) analyses and summaries of core disagreements in alignment research.
      - e.g. Jessica Taylor's two posts on motivations behind MIRI's research agenda and the Paul-MIRI disagreement.
    - A broader category of things that can be done to push discourse forward can be found in this talk Oliver and I have given in the past, about how to write good comments on LessWrong.
  - It seems to me that people I talk to think earning-to-give is easy and doable, but pushing forward intellectual progress (especially on alignment) is impossible, or at least only 'geniuses' can do it. I disagree; there is a lot of low hanging fruit.
- **Build infrastructure for the alignment research community**
  - The *Berkeley Existential Risk Initiative (BERI)* is a great example of this - many orgs (FHI, CHAI, etc) have ridiculous university constraints upon their actions, and so one of BERI's goals is to help them outsource this (to BERI) and remove the bureaucratic mess. This is ridiculously helpful. (FYI they're hiring.)
  - I personally have been chatting recently with various alignment researchers about what online infrastructure could be helpful, and have found surprisingly good opportunities to improve things (will write up more on this in a future post).
  - What other infrastructure could you build for better communication between key researchers?
- **Avoid/reduce direct government involvement (in the long run)**
  - It's important that those running AGI projects are capable of understanding the alignment problem and why it's necessary to solve alignment before implementing an AGI. There's a better chance of this when the person running the project has a strong technical understanding of how AI works.
    - A government-run AI project is analogous to a tech company with non-technical founders. Sure, the founders can employ a CTO, but then you have Paul Graham's design problem - how are they supposed to figure out who a good CTO is? They don't know what to test for. They will likely just pick whoever comes with the strongest recommendation, and given their info channels that will probably just be whoever has the most status.

- **Focus on technical solutions to x-risk rather than political or societal**
    - I have an impression that humanity has a better track record of finding technical than political/social solutions to problems, and this means we should focus even more on things like alignment.
        - As one datapoint, fields like computer science, engineering and mathematics seem to make a lot more progress than ones like macroeconomics, political theory, and international relations. If you can frame something as either a math problem or a political problem, *do the former.*
    - I don't have something strong to back this up with, so will do some research/reading.
- **Avoid things that (because they're social) are fun to argue about**
    - For example, ethics is a very sexy subject that can easily attract public outrage and attention while not in fact being useful (cf. bioethics). If we expect alignment to not be solved, the question of "*whose values do we get to put into the AI?*" is an enticing *distraction.*
    - Another candidate for a sexy subject that is basically a distraction, is discussion of the high status people in AI e.g. "*Did you hear what Elon Musk said to Demis Hassabis?*" Too many of my late-night conversations fall into patterns like this, and I actively push back against it (both in myself and others).
    - This recommendation is a negative one ("*Don't do this*"). If you have any ideas for positive things to do instead, please write them down. What norms/TAPs push away from social distractions?

I wrote this post to make explicit some of the thinking that goes into my plans. While the heuristics are informed by the model, they likely hide other assumptions that I didn't notice.

To folks who have tended to agree with my object level suggestions, I expect you to have a sense of having read obvious things, stated explicitly. To everyone else, I'd love to read about the core models that inform *your* views on AI, and I'd encourage you to read more on those of mine that are new to you.

My thanks and appreciation to Jacob Lagerros for help editing.

[Edit: On 01/26/18, I made slight edits to this post body and title. It used to say there were four models in part I, and instead now says that part I lists four parts of a single model. Some of the comments were a response to the original, and thus may read a little funny.]

No, Seriously. Just Try It: TAPs

This next semester (I'm in university, so that's how I measure time) I'm working on developing my ability to better integrate arbitrary habits into my behavior. Trigger-action-planning (a more detailed explanation as well) is LW's most concrete strategy for doing such a thing, so I've decided to just try it. Starting 3 weeks ago and proceeding for the next 6 months I'm working on my personal approach to Trigger-action-planning (TAP for short).

My basic structure for this:

- Each week (Sunday morning) design a new TAP, or redesign an old one.
- Aim to follow through on the TAP in all applicable situations throughout the week.
- At the end of each day (as part of an already existing review process) note if I did or did not follow through.
- At the end of each week, look back and think on if the TAP was useful, what worked well, what was hard, and all that jazz.

The biggest things that's jumped out at me so far has been that not all TAPs are created equal. I think conceptual similarity does not imply actionable similarity applies very heavily to TAPs. In light of this, I'm approaching each week's TAP from a very implementation specific perspective, and then afterwards I'm going to think about what connections and universal principles might apply.

Here are some things I've noticed so far:

- It's seems like there's a "mindfulness bootstrapping" problem, in that it often feels like my TAPs are only activated because of preexisting mindfulness triggers.
- I've been putting the handle for each week's TAP on my phone lock screen. This was a very effective reminder originally, though it's already lost most of its power three weeks in. Mayhaps having a randomly generated background image would help prevent my mind from filtering out expected reminders. One google search and perusing the first page of results did not produce an app to do this.
- I notice that sometimes there's a glimmer of noticing I'm in a context to activate the TAP, yet I don't. Part of this might be friction related to stopping whatever train of thought/awareness I'm currently on. Giving myself explicit permission to derail thoughts to execute a TAP could help.

I'll be posting updates on this project every so often. If you have had trouble with attempting TAPs or similar situational habit practices in the past, I'd love to hear about the specifics how what you've tried and what hasn't worked.

The Desired Response

A friend once told me a story of how an interaction with her mother changed her perspective on communication. She said that she had been going through a break up at the time, and was venting to her mother, when her mother responded with "Do you want my advice, or my sympathy?"

Often times, when we have something important to talk about, we consider how we expect the other person to respond, and talk to people who will respond the kind of way we want. We may choose to speak to someone who we know can keep a secret, or who gives good advice, or who will say nothing at all, only listen. Sometimes we may turn to someone who won't want to talk about it at all, and will instead distract us from our problems. This can be great if you have a number of close friends who reliably respond in different ways, and you are able to predict this and make use of it. However, not everyone is in this situation, and so instead, we end up with interactions where people are not getting their desired responses.

Take an interaction between example-humans Alice and Bob. Alice's goal in the conversation may be to be heard, whereas Bob may want to feel validated for what he says. If Alice is talking, Bob may point out something or make a clever remark. Instead of validating Bob, Alice feels like she hasn't been heard, and is quiet in response. Now Bob has not been validated either, and they are both sad.

Sometimes, a person's desired response can be inferred. This is like how generally, if someone tells a joke, their desired response is for you to laugh, or at least acknowledge that the joke was funny. However, this isn't always easy to tell. Some of us may know the best thing for our best friend in a time of stress, but we don't always know what's right for someone else, even if we are close. Often times we will assume that what works for us will work for them, but then there is a risk of emotional damage. Trying to talk about a situation that someone wants to avoid thinking about may exacerbate the problem or cause tension between two people. Usually, it is better to ask a person what they need -- like how my friend's mother asked her.

Going more deeply, some people may tend to have common desired responses, in general interactions. Think of the person constantly telling bad jokes, excited to hear people groan and laugh, or someone who loves giving advice and recommendations. The first of these two may love the validation they get from humor, whereas the second may want to feel helpful, and be appreciated for it. This goes far deeper than a single situational interaction; these people want these kinds of responses in everyday conversations. This may even tie to how we want others to see us (clever or kind) or how we most enjoy interacting with others (playing with ideas or working with people, receiving attention or giving it). A person's desired response may not always be the response that feels right, or the response that we want to give -- and that's perfectly okay. But it does give us insight into what drives them, what is important to them, and who they are as a person.

Justice

This is a linkpost for https://sinceriously.fyi/justice/

Video - Subject - Object Shifts and How to Have Them

A frame is a lens through which we see the world. We're always framing things, but very rarely are we aware of those frames.

In this Mental Model Monday, I go over "Subject-Object Shifts", or the process by which we become aware of the frames through which we see things. I also give two practical ways to quickly make subject-object shifts in your own life.

View Video

Transcript

Hi, this is Matt, and this is Mental Model Monday number 26. Today I wanted to talk about subject-object shifts. These are incredibly important in the development of people and how they think. If you can learn to make subject-object shifts faster, then you can exponentially increase your rate of development and growth and understanding in a new field.

Defining a Subject-Object Shift

The idea behind a subject-object shift is that what was once something that you were subject to, something that was unconscious to you, something that you saw through, now becomes object to you, something that you can look at, something that you can play with, something that is not the way the world is but is just a way to look at the world.

One way to look at them is that they are a way to examine unconscious assumptions. That's sort of a fake and not quite true model of what subject-object shifts are, is they are a way examine assumptions which at one point were unexaminable to you.

Subject-Object Shifts and Kegan's Developmental Stages

An example of where subject-object shifts happen are in Kegan's developmental stages. This is one model of how adults develop. In Kegan's developmental stages, he's the one who coined the term subject-object shift, I believe. In Kegan's developmental stages, at each stage the thing that you originally saw as the world

becomes a part of the world. At stage two, your impulses and perceptions, you can now see they're not me. They're part of me but my needs, interests, and desires are me. That's who is am, that's what I am. That's all there is. Whereas in stage three you see, oh, my needs, interests, and desires are not all of my. They're not all of the things that matter. I can see beyond that. Those are just individual things that I can play with. My whole world is now relationships and mutuality and how I interact with others, and the relations between people, and my friend group.

Then the relationships between people and my friend group now become just a thing I see. I'm not defined by that. I have my own values system, my own identity, my own mission. That's what I'm seeing, and all my relationships and my mutual connections are in relationship to that. It goes on to the fifth order, where even my values system I now see is just one of many values systems. What I'm seeing is the dialectic between different ideologies and how they relate, and the interrelationships between systems.

That's a series of subject-object shifts that happen well into one's 40s, if at all, to get all the way to fifth order. But you can see how the ability to make these shifts faster and the ability to understand what a subject-object shift even is, it can be really powerful in your own growth and development.

Practical Tactics That Point Towards Subject Object Shifts

What I'm going to do is I'm going to go over a few different ways to look at some of those object shifts. I'm going to talk about how they are actually the same way of looking at subject-object shifts and how they relate to concepts like enlightenment.

Turning Edges Into Nodes

The first thing I want to talk about is turning edges into nodes, which you can also go back in Mental Model Monday number 13, and that also goes into turning edges into nodes, but I'm going to real quickly just rehash this concept.

In a concept map, which is this right here, you have nodes, which are these square boxes. Then you have edges, which are the arrows and the words that connect the square boxes, so edges and nodes.

One way to look at a concept map is that all of the squares, all of the nodes, are things that are objects in your ontology. They are things that you can look at, whereas all of the edges, the arrows and the words on them, are subjects, are things that you're looking through. They are assumptions about the world that you have.

One way to shift from subject to object is to take one of these edges and turn them into a node with two new edges connecting two concepts. An example of this is if you have "pine is a tree."

Your original concept map is a way you view the world that you don't even know that you're doing it. However, if you try to turn this "is a" into a node, what you realize is

that what you really mean ... what you're really subject to, what the assumption you have here, is the idea of woody secondary growth. When you turn this into a node, what you realize the object is pine has woody secondary growth and is therefore classified as a tree.

If you didn't realize that that's what was making you classify trees, you wouldn't be able to classify trees and different ways and new ways, and look at things differently. Again, the thing you were viewing through at one point, this "is a," is now something you can look at. You can see that the assumption you're actually [inaudible 00:07:30] there is that woody secondary growth is a way to classify trees. Just like Kegan's models, now that you know that this "is a" is just one way to classify trees, you can look at other ways to classify trees, whereas this "is a" prevented you from doing so.

That's one way to have a subject-object shift, is to map out your ontology using edges and nodes, and then spend some time turning those edges into nodes.

Observing the Observer

Now I'm going to go over another way to have subject-object shifts. At first it's going to seem completely unrelated to the previous thing, but don't worry. I will show how these relate eventually.

This is the idea of observing the observer. One way to think of yourself is a series of parts that are observing each other.

You can represent these parts in three different ways.

First, you can have some image for them. This is the glowing thing in my chest, or you just have some abstract thought. This is like a hammer. This is the hammer part of me. You're visualizing them.

You can represent them auditorily. This is the voice that talks to me in this tone. This is my negative voice. This is how I talk to myself. You can also represent them kinesthetically. This is that rough feeling in my chest. This is that nice warm feeling I get when XYZ happens. The parts of yourself you represent visually, auditorily, and kinesthetically.

Oftentimes we are subject to these parts. I am the voice in my head. I am this feeling of happiness or I am this feeling of warmth. I am the hero, the hero I'm visualizing.

Now how do you turn these parts that you're a subject to into parts that are objects for you to view? The Wholeness Process, by Connirae Andreas, has a way to do this. Here's how you do it. You take one of those voices, one of those images, one of those kinesthetic feelings that are located somewhere in your body, and you say, "I am aware of this tight feeling in my chest. I am aware of the voice that I'm talking to myself with." First you just say the sentence, "I am aware of this part."

Now here comes the tricky part. You ask yourself, where is the I that is aware of that thing? The I that I was looking through, where is it? That would be kinesthetically. Where is it in my body? You can also ask, when you say, "I am aware of this voice in my head," "Where's the voice that's saying that?" You can also ask, "When I see myself as a XYZ, where is the eye that sees the other eye?" You're trying to observe the observer and say, "Where is the eye?"

At first this seems like a weird question, but if you just answer the question immediately, you'll come up with an answer, and it won't be this eye is looking at itself. There'll be another observer observing the thing. You've now stepped out of this thing you are subject to and you are viewing it as an object.

Interesting thing here, the thing you stepped out of, you didn't step out of all observers. You're now in another observer using this tool. But, again, you've made this subject-object shift, so play around with that. I am aware of this feeling. I see this feeling. I hear this voice. Where is the eye that is here and where is the eye that is seeing? Where is the I that is aware? You can now view that I from a different observational point.

Enlightenment as a Subject-Object Shift

One way to think about enlightenment is when everything is object. You are no longer subject to any observer. You are viewing everything as object. When you view everything as object, in a weird way, everything becomes subject. This is the idea of oneness. I am one with everything. When I can view everything objectively, I also am viewing as everything. Therefore, I feel at one with everything. This is the idea of enlightenment, is instead of stepping up one level observer, one level observer, one level observer, I'm stepping out of observers altogether, and therefore everything becomes an observer and everything becomes the observed. Interesting way to go about this.

How Does "Turning Edges Into Nodes" Relate to "Observing The Observer"

These are two practical ways right now that you can make a Subject-Object Shift. One is you can try to observe the observer by using this "I am aware" sentence and asking yourself where is the I. Two, you can turn your ontology into a concept map. You can turn the edges of the concept map into nodes. Now the question is, these seem like two totally different things. How do you really understand them as one idea of subject-object shifts?

One way to do this is to vusalize yourself first person on a node INSIDE your concept map and you are looking at another node, through the lens of that node's edge. Now the edge in that node is your beliefs about that node's relation to your current node.

Before you did this visualization, you might not even have notice that each node has it's own observer, viewing it's relation to other nodes. Without consiously being aware of the multitude of observers that exist within a concept map, your ontology becomes "just the way things are". However, when you notice the observer, you can begin to question their observations. This, as you may notice, is the same thing as becoming aware of the I, and then examining its' beliefs.

This is how you tie together this idea of turning edges into nodes with observing the observer. You figure out that the observer is the thing that is looking from each node to the next one, and the edge is the unconscious beliefs that that observer has that you can make conscious.

I hope you enjoyed that quick overview of subject-object shifts and how to have them. Please let me know if this is useful to you. Please let me know what was unclear. Please let me know if there's anything in here that you'd like me to do a future Mental Model Monday on. Thanks for watching guys, and I'll catch you soon.

TSR #10: Creative Processes

This is part of a series of posts where I call out some ideas from the latest edition of The Strategic Review (written by Sebastian Marshall), and give some prompts and questions that I think people might find useful to answer. I include a summary of the most recent edition, but it's not a replacement for reading the actual article. Sebastian is an excellent writer, and your life will be full of sadness if you don't read his piece. The link is below.

Background Ops #10: Creative Processes

SUMMARY

- Focus on the results vs “doing work”.
- Though creative work appears to be “Some days you got it, some days you don’t”, this is not the case.
- *“By relieving the brain of all unnecessary work, a good notation sets it free to concentrate on more advanced problems, and in effect increases the mental power of the [human] race.”*
- Separate the different phases of your writing process so they aren’t all competing for mental space at the same time.
- Break a process down into its “universal movement”
    - Examples in martial arts
        - Imi Litchtenfeld (Krav Maga)
        - Bruce Lee (Jeet Kune Do)
- There is no magic, it’ll be boring, but it works.
- Universal movements of writing
    - Brainstorm
    - Categorize
    - Outline
    - Writing
    - Editing
- Guidance
    - Read Pozen’s Extreme Productivity Chapter 8
    - Create or learn the universal movements of your field.
    - Personalize and develop
        - Sebastian made research it’s own step in his writing process.
    - Cross pollinate ideas from unrelated domains.

**Just realised that there were two more babble posts, and what I point out as anchors and what alkjash called beacons*

I’ve really enjoyed the ideas and discussion in the Babble post and its follow up, and will be looking at this TSR edition through the lens of babble and prune.

The standout claim in this article is that there are ways to structure your approach to a creative task such that you don’t end up stuck. What’s going on when you get stuck? Getting stuck is when your pruning filter has a higher standard than the quality of your babble, leaving you with “nothing to say” and an unsatisfying feeling of being stuck.

What then, would be the solution to getting stuck? As people were pointing out, often the first step is to tone down the aggressiveness of your pruning. Babbling nonsense

beats not saying anything. But what comes after that? How do you end up getting your babble to a point where even with whatever pruning standard you think is appropriate, you still end up producing at that level?

Thinking of your babble generator as traversing through a connected babble/idea graph, one route seems to be to improve the idea graph itself. Learn cooler things, and make more connections between the cool things. Another approach is to train the part of the generator that makes the traversal decisions. The traversing part of the algorithm seems to include a mechanic for deciding where to restart from for the next traversal, and how to choose which neighboring node to visit next. Improving the graph, or improving the traversal would produce better babble.

The writing process that Sebastian draws attention to a specific way to improve your traversal quality, within the context of your current writing project.

It's non-controversial to claim that knowing what you are writing about makes it easier to write. Why, from a babble and prune lens, is that so?Well it seems like your traversal algorithm is affected both by some "general ability" and by whatever is in your working memory. What is on you mind anchors how you traverse your babble graph.

If the essence of what you want to say is already in your working memory, you can produce babble that matches said essence much more easily. Writing an outline is just a way of expanding your working memory, and if you make a quality outline, it will have a strong anchoring effect on your babble. The cool thing here is that you can get a huge context-specific boost in babble quality, without the work needed to improve your babble graph itself (though I'm sure it's possible to have a poorly enough connected babble graph that no anchor will help you that much).

So how do you make the outline in the first place? Recursively using babble and prune! Think back to the 5 stages of writing that Sebastian mentions. Brainstorming looks like pure babble. Categorizing looks like babble anchored by your previous babble, with a lite amount of pruning. All of that babble is used to anchor the babble and prune which goes towards making the outline. If you make a good enough outline, you can probably babble your way through the entire writing phrase, and just swing back around for one last editorial prune to get things ship shape.

In the most general sense, you might be able to describe a generic creative process as "Recursively using babble and prune with progressively strong pruning to create anchors of high enough quality to allow you to babble to main content at the standards of your full throttle prune."

For writing in particular, I'd guess it would be more useful to follow a specific five steps like Sebastian laid out, as opposed to thinking in terms of the general principle. Though it does seem like the general principle is a good anchor to have in your working memory for when you are trying to come up with a flow for your specific creative work.

Some questions that could be useful to answer:

- Where in your creative process do you often get stuck?
- Does babble and prune give a useful framework for describing how you get stuck, and does it imply a way to get unstuck?

Hammers and Nails

This is a linkpost for https://radimentary.wordpress.com/2018/01/22/hammer-and-nails/#more-2328

If all you have is a hammer, everything looks like a nail.

The most important idea I've blogged about so far is Taking Ideas Seriously, which is itself a generalization of Zvi's More Dakka. This post is an elaboration of how to fully integrate a new idea.

I draw a dichotomy between Hammers and Nails:

A *Hammer* is someone who picks one strategy and uses it to solve as many problems as possible.

A *Nail* is someone who picks one problem and tries all the strategies until it gets solved.

Human beings are generally Nails, fixating on one specific problem at a time and throwing their entire toolkit at it. A Nail gets good at solving important problems slowly and laboriously but can fail to recognize the power and generality of his tools.

Sometimes it's better to be a Hammer. Great advice is always a hammer: an organizing principle that works across many domains. To get the most mileage out of a single hammer, don't stop at using it to tackle your current pet problem. Use it everywhere. Ideas don't get worn down from use.

Regardless of which you are at a given moment, be systematic because Choices are Bad.

Only a Few Tricks

I am reminded of a classic speech of the mathematician Gian-Carlo Rota. His fifth point is to be a Hammer (emphasis mine):

> A long time ago an older and well known number theorist made some disparaging remarks about Paul Erdos' work. You admire contributions to mathematics as much as I do, and I felt annoyed when the older mathematician flatly and definitively stated that all of Erdos' work could be reduced to a few tricks which Erdos repeatedly relied on in his proofs. **What the number theorist did not realize is that other mathematicians, even the very best, also rely on a few tricks which they use over and over.** Take Hilbert. The second volume of Hilbert's collected papers contains Hilbert's papers in invariant theory. I have made a point of reading some of these papers with care. It is sad to note that some of Hilbert's beautiful results have been completely forgotten. But on reading the proofs of Hilbert's striking and deep theorems in invariant theory, it was surprising to verify that Hilbert's proofs relied on the same few tricks. Even Hilbert had only a few tricks!

The greatest mathematicians of all time created vast swathes of their work by applying a single precious technique to every problem they could find. My favorite

book of mathematics is The Probabilistic Method, by Alon and Spencer. It never ceases to amaze me that this same method applies to:

1. (The Erdős–Kac Theorem) The number of distinct prime factors of a random integer between 1 and n behaves like a normal distribution with mean and variance $\log \log n$.
2. (Heilbronn's Triangle Problem) What is the maximum $\Delta(n)$ for which there exist n points in the unit square, no three of which form a triangle with area less than $\Delta(n)$?
3. (The Erdős-Rényi Phase Transition) A typical random graph where each edge exists with probability $\frac{1}{n^c}$ has connected components of size $O(\log n)$. A typical random graph where each edge exists with probability $\frac{1}{n^c}$ has a giant component of size linear in n.

It's amusing to note that in the same speech, Rota expounded the benefits of being a Nail just two points later:

> Richard Feynman was fond of giving the following advice on how to be a genius. You have to keep a dozen of your favorite problems constantly present in your mind, although by and large they will lay in a dormant state. Every time you hear or read a new trick or a new result, test it against each of your twelve problems to see whether it helps. Every once in a while there will be a hit, and people will say: "*How did he do it? He must be a genius!*"

Both mindsets are vital.

To be a Nail is to study a single problem from every angle. It is often the case that each technique sheds light on only one side of the problem, and by circumambulating it via the application of many hammers at once, one corners the problem in a deep way. This remains true well past a problem's resolution - insight can continue to be drawn from it as other methods are applied and more satisfying proofs attained.

Usually even the failure of certain techniques sheds light on shape of the difficulty. One classic example of an enlightening failure is the consistent overcounting (by exactly a factor of two!) of primes by sieve methods. This failure is so serious and unfixable that it has its own name: the Parity Problem.

Dually, to be a Hammer is to study a single *technique* from every angle. In the case of the probabilistic method, a breadth of cheap applications were found immediately by simply systematically studying uniform random constructions. However, particularly adept Hammers like Erdős upgraded the basic method into a superweapon by steadfastly applying it to harder and harder problems. Variations of the Probabilistic Method like the Lovász Local Lemma, Shearer's entropy lemma, and the Azuma-Hoeffding inequality are now canon due to the persistence of Hammers.

Be Systematic

The upshot is not that Hammers are better than Nails. Rather, there is a place for both Hammers and Nails, and in particular both mindsets are far superior to the wishy-washy blind meandering that characterizes overwhelmed novices. There may be an endless supply of advice - even great advice - on the internet, and yet any given person should organize their life around systematically applying a few tricks or solving a few problems.

Taking an idea seriously is difficult and expensive. You'll have to tear down competing mental real estate and build a whole new palace for it. You'll have to field test it all over the place without getting superstitious. You'll have to gently titrate for the amount you need until you have enough Dakka.

Therefore, be a Hammer and make that idea pay rent. Hell, you're the president, the emperor, the king. There's no rent control in your head! Get that idea for all its got.

Exercise for the reader: all things have their accustomed uses. Give me ten unaccustomed uses of your favorite instrumental rationality technique! (Bonus points for demonstrating intent to kill.)

Superhuman Meta Process

This is a linkpost for http://squirrelinhell.blogspot.com/2018/01/superhuman-meta-process.html

Hammertime Day 1: Bug Hunt

Rationality is systematized winning.

In Hammers and Nails, I suggested that rationalists need to be more systematic in the practice of our craft. In this post, I will use the word Hammer for a single technique well-practiced and broadly applied.

Hammertime is a 30-day instrumental rationality sequence I am designing for myself to build competence with techniques. Its objective is to turn rationalists into systematic rationalists. By the end of this sequence, I hope to upgrade each Hammer from Bronze Mace to Vorpal Dragonscale Sledgehammer of the Whale. I invite you to join me on this journey.

The core concept: One Day, One Hammer.

Hammertime Schedule

In Hammertime, we will practice 10 Hammers over 30 days. Each exercise is scalable from a half hour to an entire day. The Hammers will be bootleg CFAR techniques:

1. Bug Hunt
2. Resolve Cycles
3. TAPs
4. Design
5. CoZE
6. Mantras
7. Goal Factoring
8. Focusing
9. Internal Double Crux
10. Planning

There will be three cycles of 10 days each, practicing each technique a total of three times. The first cycle will cover basics and solve bugs at the life-hack level. The second cycle will reinforce the technique, cover variations and generalizations, and solve tougher challenges. The third cycle will build fluid compound movements out of multiple core techniques.

Day 1: Bug Hunt

A bug is anything in life that needs improvement. Even if something is going well, if you can imagine it going better, there's a bug.

On the first day of Hammertime, we will scour our lives with a fine-toothed comb to find as many bugs as possible. A comprehensive bug list will provide the raw material on which we practice every other rationality technique. For the first cycle of Bug Hunt, look for small, concrete bugs. The whole exercise should take a bit over an hour.

WARNINGS: Focus on finding bugs, not solving them. If you can solve the bug immediately, go for it. Otherwise, hold off on proposing solutions. Writing down a bug

does not mean you commit to doing anything about it.

1. Setup

Find a notebook, phone app, spreadsheet, or Google Doc to record your bugs - preferably something you can bring with you throughout the day. We will refer back to it repeatedly in the coming days for bugs to solve.

During Bug Hunt, spend the next 30 minutes writing down as many bugs as you can. Following each of the six sets of prompts in the next section, set a timer for 5 minutes and list as many bugs as you notice.

2. Prompts

A. Mindful Walkthrough

Walk through your daily routine in your head and look for places that need improvement. Do you get up on time? Do you have a morning routine? Do you waste mental effort deciding whether to or what to eat for breakfast? Do you take the most efficient commute, and make the most of time in transit?

Fast forward to work or school. Are there physical discomforts? Are you missing any tools? Are there particular people who bother you, or to whom you don't speak enough? Do you ask for help when you need it? Do you know how to shut up? Is there unproductive dead time during meetings, classes, or builds? Do you take care of yourself during the day?

Think about the evening at home. Do you waste time deciding where or what to eat? Are there hobbies you want to try? Are there things you know will be more fun that you're not doing? Do you progress consistently on your side projects? Do you sleep on time? How is your sleep quality?

B. Hobbies, Habits, and Skills

Walk through the things you do on a regular basis. Are there habits you mean to drop? Are there habits you mean to pick up but never seem to get around to?

For each hobby or habit, answer the following questions. Do you do it enough? Do you do it too much? Are there ways you could improve your experience? Do it in a different place and time? Do it with other people or alone?

Perhaps you have skills to practice. Are you as good as you want to be? Do you practice regularly? Have you plateaued by overtraining? Are there minor recurring discomforts keeping you from trying? Are there directions you haven't tried which might indirectly improve your abilities?

C. Space

Look around your living space, your workspace, or the interior of your vehicle. What would you change?

Space should be functional. Is there clutter you circumnavigate on a daily basis? Are your chairs and tables at the right height? Is your bed comfortable? Are there towels, pans, notebooks, or papers sitting out taunting you? Are there important things that deserve a more central position? Have you set up Schelling places for glasses, wallets, and phones?

Space should be aesthetically pleasing. Do pieces of furniture or equipment stick out comically? Do your walls feel drab and depressing? Are there carpet stains or dust mites that keep catching your eye and sucking out your happiness? Are you tired of the art on the walls?

Space on the monitor can be as important as physical space. Do you have enough screens? Do you find yourself repeating mechanical boot-up and shutdown sequences that can be automated? Do you use all the browser extensions and keyboard shortcuts? Is there a voice in the back of your head whispering at you to learn vim?

D. Time and Attention

People and things clamor for your attention. What's missing from your life that would let you live as intentionally as possible?

Many activities are bottomless time sinks. Do you watch shows or play games you no longer enjoy? Do you get dragged into conversations that hold no value? Do you find yourself rolling the mouse wheel down endless Facebook or Reddit feeds? Are there classes, meetings, commutes, or projects that zombify you for the rest of the day? Do you set up ejector seats in advance to protect yourself from time sinks?

Focus on the things you don't pay enough attention to. Do you often make mistakes on autopilot? Are there friends or family you've neglected or grown distant from? Are there conversations you zone out in that you could get more out of? Is there a childhood dream you've forgotten?

Sometimes trivial distractions lead to spectacular failures. Are there slight, recurring physical discomforts that drain your agency? Does the temperature outside prevent you from exercising? Is there something shiny that always draws your eye away from work?

E. Blind spots

Our biggest bugs can hide in cognitive blind spots.

Outside view your life. Are you sufficiently awesome? What is your biggest weakness? If there is one thing holding you back from achieving your goals, what would it be? Do you have mysterious attachments to pieces of your identity? Do you routinely over- or under-estimate your own ability?

Simulate your best friend in your head. What do they say about you that surprises you? What behaviors annoy them? What behaviors would they appreciate? Is there a piece of advice they keep giving you?

Summon your Dumbledore. What would he say to you? What deep wisdom are you blind to? If you were the protagonist, what genre would this life be?

Look to admiration and jealousy for insight. Are you the person you most admire? What skills and traits do others have that you want?

F. Fear and Trembling

The shadows we flinch away from can hide the most bountiful treasures.

What are your greatest fears and anxieties? Do you have the strength to be vulnerable? Are there necessary and proper actions you need to take? Are there truths you're scared to say out loud? What do you lie to yourself about?

Look to your social circle. Are there good people you hide from? Are there conversation topics that cause you scramble away? What do people say that cause you to lose your composure?

Look to the past and future as far as your eyes allow. What deadlines cause you to avert your eyes? Is there a kind of person you are terrified of becoming? Or are you most afraid of stagnation? Do you trust your past and future selves?

3. Sort

Hopefully, you came up with at least 100 bugs; I came up with 142. Time for some housekeeping. Input your bugs into a spreadsheet to organize and coalesce similar ones. Using System 1, assign difficulty ratings from 1 to 10, where 1 is "I could solve it right now" and 10 is "Just thinking about it causes existential panic." Sort them in increasing order of difficulty.

In the coming days, we will go down the list systematically, hitting as many nails as possible with each hammer.

Daily Challenge

To help others brainstorm, share your strangest bug-fix story. I'll start:

The muscles on the left half of my face are more responsive, which caused me to smile asymmetrically for most of my life. Therefore, my usual smile wasn't far from a contemptuous smirk, and caused me to feel dismissive of everyone I smiled at. I trained myself to smile on both sides and now feel warmer towards people.

A simpler way to think about positive test bias

Eliezer described "positive bias" (which I'll rename "positive test bias" for reasons explained in Unnamed's comment below) in an LW post and an HPMOR chapter. According to him, dealing with that bias requires a kind of mental gymnastics that doesn't come naturally to most people: "twisty negative thinking" or "flinching toward zero". You're supposed to devise tests that come out *false* if your hypothesis is *true*. It's a bit confusing.

I think there's a simpler way to think about it. Positive test bias is just our bias toward strong hypotheses. You can deal with it by asking yourself, how can I test if my hypothesis is too strong?

- The LW post about the bias mentions the Wason 2-4-6 task. If you're told that the number sequence 2-4-6 has property X, it's tempting to guess that property X means "ascending arithmetic progression". To test if that hypothesis is too strong, you need to try some other sequences that are arithmetic progressions but not ascending, and some that are ascending but not arithmetic progressions. Easy!
- The HPMOR chapter describes Hermione spilling some soda on her robe, and a moment later it mysteriously becomes clean again. Hermione comes up with a hypothesis that her robe is magically self-cleaning. To test if that hypothesis is too strong, she can try spilling something else on her robe. There's no need for any counterintuitive thinking.

That technique is useful in other areas as well. For example, I often get carried away when writing posts, and end up with a draft full of wide-reaching conclusions. But then I ask myself, doesn't that sound a bit too strong? So I look for counterexamples, and the point I'm trying to make either dissolves or becomes much more robust.

Our bias toward strong hypotheses is especially harmful in politics. We will defend a hypothesis like "group X is to blame for everything" for its explanatory power, never noticing the real problem—that it's too strong. We'd all benefit from recognizing when it happens, and weakening our hypotheses until they match reality.

Rationality: Abridged

This was originally planned for release around Christmas, but our old friend Mr. Planning Fallacy said no. The best time to plant an oak tree is twenty years ago; the second-best time is today.

I present to you: *Rationality Abridged* -- a 120-page nearly 50,000-word summary of "Rationality: From AI to Zombies". Yes, it's almost a short book. But it is also true that it's less than 1/10th the length of the original. That should give you some perspective on how massively long R:AZ actually is.

As I note in the Preface, part of what motivated me to write this was the fact that the existing summaries out there (like the ones on the LW Wiki, or the *Whirlwind Tour*) are too short, are incomplete (e.g. not summarizing the "interludes"), and lack illustrations or a glossary. As such, they are mainly useful for those who have already read the articles to quickly glance at what it was about so they could refresh their memory. My aim was to serve that same purpose, while **also** being somewhat more detailed/extensive and including more examples from the articles in the summaries, so that they could also be used by newcomers to the rationality community to understand the key points. Thus, it is essentially a heavily abridged version of R:AZ.

Here is the link to the document. It is a PDF file (size 2.80MB), although if someone wants to convert it to *.epub* or *.mobi* format and share it here, you're welcome to.

There is also a text copy at my brand new blog:
perpetualcanon.blogspot.com/p/rationality.html

I hope you enjoy it.

(By the way, this is my first post. I've been lurking around for a while.)

Understanding-1-2-3

Epistemic Status: Seems worth sharing

Assumes Knowledge Of: System 1 and System 2 from Thinking, Fast and Slow

While in San Francisco this past week, I found myself in many complex and rapid conversations, with new ideas and concepts flying by the second. I hadn't experienced this since least February's AI Safety Unconference. It is wonderful to be pushed to one's limits, but it means one must often rely on System-1 rather than System-2, or one will be left behind and lost. I now realize this is what happened to me in foreign language classes – I was using System-2 to listen and talk, and that doesn't work *at all* if others don't wait for you.

This meant that often I'd grasp what was being said in an intuitive sense using System-1, or agree or disagree in that fashion, without the time to unpack it. This can lead to great understanding, but also lead to waking up the next day having no idea what happened, or thinking about it in detail and realizing you didn't understand it after all. So when people asked if I understood things, I instinctively started saying things like "I understand-1 but not 2." I also did this for beliefs and agreements.

I debated explaining it but decided not to, as a test to see if others would intuit the meaning, most seemed to (or ignored the numbers, or came up with something equivalent in context); one person explicitly asked, and said he'd use it. I think this may be a valuable clarification tool, as it's very different to understand-1 versus understand-2, or agree-1 versus agree-2, and seems better than saying something like "I *kind of* agree" or "I *think* I agree but I'm not sure," which are both longer and less precise ways people often say the same thing.

My system-1 also added understand-3. Understand-3 means "I understand this well enough to teach it." By extension, believe-3 or agree-3 means "I believe this, know how to convince others, and think others should be convinced." To truly understand something, one must be able to explain, teach or defend it, which also means you and others can build solidly upon it. Writing has helped me turn much understanding-2 (and understanding-1) into understanding-3.

Actionable Eisenhower

This is a linkpost for https://squirrelinhell.blogspot.com/2018/01/actionable-eisenhower.html

GoodAI announced "AI Race Avoidance" challenge with $15k in prize money

This is a linkpost for https://www.general-ai-challenge.org/ai-race

Demon Threads

tldr: a Demon Thread is a discussion where everything is subtly warping towards aggression and confusion (i.e. as if people are under demonic influence), even if people are well intentioned and on the same 'side.' You can see a demon thread coming in advance, but it's still hard to do anything about.

("Flame Wars" are similar but I felt the connotation was more like "everything has already *gone to hell, and people aren't even pretending to be on the same side")*

I kept wanting to reference this post when discussing internet discussion policy, and I kept forgetting that nobody has written it yet. So here it is.

Suggested Background Reading:

- Politics is Hard Mode (Rob Bensinger)
- Civility is Never Neutral (Ozy)
- Writing that Provokes Comments (me)
- Musings on Double Crux and Productive Disagreement (me)

If someone in the future linked you to this post, it's probably because a giant sprawling mess of angry, confused comments is happening - or is about to happen - and it's going to waste a lot of time, make people upset, and probably *less* likely to listen to each other about whatever the conversation ostensibly is about.

I have some ideas on what to do *instead*, which I discuss in this followup post.

But for now, this post is meant to open a discussion, explore the mechanics of how demon threads work, and then in the comments brainstorm solutions about how to handle them.

Wrong On the Internet

I find “Someone Is Wrong On the Internet” to be a weird, specific feeling.

It's distinct from someone being *factually wrong* - people can be wrong, point it out, and hash out their disagreements without a problem. But a common pattern I’ve witnessed (and experienced) is to notice someone being wrong in a way that *feels distinctly bad*, like if you don’t correct them, something precious will get trampled over.

This is when people seem most prone to jump into the comments, and it’s when I think people *should be most careful*.

Sometimes there actually *is* an important thing at stake.

There usually isn’t.

It often *feels* like there is, because our social intuitions were honed for tribes of a hundred or two, instead of a world of 7 billion. We live in a different world now. If you

actually want to have an impact on society, yelling at each other on the internet is almost certainly not the best way to do so.

When there *actually is* something important at stake, I think there are usually better plans than “get into a giant internet argument.” Think about what your goals are. Devise a plan that actually seems like it might help.

Different situations call for different plans. For now, I want to talk about the common anti-pattern that often happens instead.

Demon Threads are explosive, frustrating, many-tentacled conversations that draw people in regardless of how important they are. They come in two forms:

- **Benign Demon Threads** are mostly *time wasting.* Nobody gets *that* angry, it's just a frustrated mess of "you're wrong" "no y*ou're* wrong" and then people spend loads of digital ink arguing about something that doesn't matter much.
- **Malignant Demon Threads** feed upon emotions of defensiveness, anger, tribal affiliation and righteousness - and inflame those emotions, drawing more people into the fire.

(A malignant demon thread is cousin to the *flame war* - people hurling pure insults at each other. What makes a malignant demon thread insideous is the way it can warp discussion even among people who are earnestly trying to communicate, seek truth and solve problems)

If you find yourself in a malignant demon thread, I think it's likely you are not only *not helping*, but are actually hurting your cause.

The Demon Seed

*How to write so that people will comment [disclaimer: not necessarily **good** advice]*

1. *Be wrong*
2. *Be controversial*
3. *Write things people feel qualified to have opinions on.*
4. *Invoke social reality.*

- *Writing That Provokes Comments*

In the comments on YouTube, or the worst parts of Facebook or tumblr, demon threads are not surprising. People write comments that inflame ideological warfare all the time. Internets be internets. People be people. What can you do?

The surprising thing is how this works in places where everyone should really know better. The powers of demons are devious and subtle.

There's an experiment — insert obligatory replication crisis disclaimer — where one participant is told to gently poke another participant. The second participant is told to poke the first participant the same amount the first person poked them.

It turns out people tend to poke back slightly harder than they were first poked.

Repeat.

A few iterations later, they are striking each other really hard.

I think something like this is at work in the mechanics of demon threads.

The **Demon Seed** is the first comment in what will soon become a demon thread. It might look pretty innocuous. Maybe it feels *slightly* rude, or slightly oblivious, or pushing a conversation that should be about concrete empirical facts slightly towards being about social consensus (or, vice versa?).

It feels 1% outside the bound of a reasonable comment.

And then someone **waters the demon seed**. They don't want to let the point stand, so they respond with what seems like a fair rebuke.

Maybe they're self-aware that they're feeling annoyed, so they intentionally dial back the aggression of their response. "Ah, actually this probably comes across as too hostile, so I'll tweak my wording to reduce hostility by 4%." But, actually, the words were 6% more hostile than they thought, and now they've escalated 2%.

Repeat 2-3 times. The demon seed is watered. Latent underlying disagreements about how to think properly... or ideal social norms... or which coalitions should be highest status... or pure, simple *you're insulting me and I'm angry*...

They have festered and now they are ready to explode.

Then someone makes a comment that pushes things over the edge, and a demon thread is born.

(It is, of course, possible to skip steps 1-4 and just write a blatantly rude, incendiary comment. I'm trying to describe how this happens *even when everyone is well intentioned and mostly trusts each other)*

From there, if you're *lucky* it's contained to two people. But often, well meaning bystanders will wander by and think "Ah! People are being wrong on the internet! Wrong about things I am qualified to have opinions on! I can help!"

And it grows.

Then people start linking it from elsewhere, or FB algorithms start sharing it because people are *commenting* so the thread must be *important*.

It grows further.

And it consumes days of people's attention and emotional energy. More importantly, it often entrenches people's current opinions, and it burns people's *good will* that they might have been willing to spend on honest, cooperative discourse.

Why Demon Threads are Bad

I think demon threads are not just a bad plan - I think they are often *net negative* plan.

The reason is best expressed in Conor Moreton's Idea Innoculation and Inferential Distance. [Edit: the full article is no longer available]

> **Inferential distance** is the gap between [your hypotheses and world model], and [my hypotheses and world model]. It's *just how far out* we have to reach to one another in order to understand each other.
>
> If you share political and intellectual and cultural foundations, it's (relatively) easy. If you have completely different values and assumptions, (say you get dropped off in the 15th century and need to argue with Christopher Columbus) it may be nigh impossible.
>
> It's right in the name—*inferential* distance. It's not about the "what" so much as it is about the "how"—how you infer new conclusions from a given set of information. When there's a large inferential distance between you and someone else, you don't just disagree on the object level, you also often disagree about ***what counts as evidence, what counts as logic,*** and ***what counts as self-evident truth.***
>
> What makes this really bad is **idea inoculation.**
>
> When a person is exposed to a weak, badly-argued, or uncanny-valley version of an idea, they afterwards are *inoculated* against stronger, better versions of that idea. The analogy to vaccines is extremely apt—your brain is attempting to conserve energy and distill patterns of inference, and once it gets the shape of an idea and attaches the flag "bullshit" to it, it's *ever after* going to lean toward attaching that same flag to any idea with a similar shape.
>
> When you combine idea inoculation with inferential distance, you get a recipe for disaster—if your first attempt to bridge the gap fails, your second attempt will *also* have to overcome the person's rapidly developing resistance.
>
> You might think that each successive attempt will bring you closer to the day that you finally establish common ground and start communicating, but alas—often, each attempt is just increasing their resistance to the core concept, as they build up a library of all the times they saw something like this defeated, proven wrong, made to look silly and naive.

A demon thread is a *recipe for bad attempts at communicating*. Lots of people are yelling at once. Their defenses are raised. There's a sense that if you give in, you or your people look like losers or villains.

This'll make people worse at listening *and* communicating.

Why the Internet Worse

"Demon threads" can happen in person, but they're worse online.

One obvious reason is that **the internet is more anonymous**. This reduces consequences to the person writing a comment, and makes the target of the comment easier to round off to a bad stereotype or an abstract representation of The Enemy.

Other things people do:

A. People end up writing long winded monologues without anyone interrupting them to correct basic, wrong assumptions.

i.e. "you're just wrong because you think X, therefore... [complicated argument]", without providing opportunity for someone to respond "no I don't actually think X at all". And then, having written out [complicated argument] you're already *invested* in it, despite it being built on faulty premises.

B. Lots of people are writing. Especially as the demon thread grows. After 24 hours of its existence, the thread will have so much content it's a huge investment to actually read everything that's been said.

C. The comments aren't necessarily displayed in order. Or, if they are, people aren't reading them in order, they're reading whatever it's largest or most interesting.

D. The internet is full of lots of other content competing for attention.

This all means that:

E. People are *skimming*. This is most true when lots of people are writing lengthy monologues, but even when the thread first begins, people's eyes may be bouncing around to different tabs or different threads within a page so they *aren't even reading what's being said*, not with the intentionality and empathy they would when confronted with a real person in front of them.

And they might *first* be reading the most explosive, recent parts of a thread rather than piecing together the actual order of escalation, which may make people look less reasonable than they were.

This all adds up to giant threads being a *uniquely bad way to resolve nuanced, emotionally fraught issues.*

Containment?

Demon threads are like wildfires. *Maybe* you can put them out, with coordinated effort. You can also try to ignore them and hope they burn themselves out.

But if you wanted to actually stop it, the best bet is to do so is before they're erupted in the first place.

I've developed a sense of what seeds look like. I'll see a comment, think "god, this is going to become a demon thread in like two hours", and then sure enough, two hours later people are yelling at each other and everything is awful and everyone involved seems *really sure* that they are helping somehow.

Some flags that a demon thread might be about to happen:

Flags Regarding: Tension and Latent Hostility

- When you look a comment and want to respond, you feel a visceral sense of "you're *wrong*", or "ugh, those people [from group that annoy me]" or "important principle must be defended!" or “I am literally under attack.”

- You feel physiological defensiveness or anger - you notice the hairs on the back of your neck or arms standing on end, or a tightness in your chest, or however those emotions manifest in your body.
- People in the thread seem to be talking past each other.
- For whatever reason, tensions seem to be escalating.

Flags Regarding: Social Stakes

- The argument seems like it's about who should be high or low status, which people or groups are virtuous and which are not, etc.
- The argument is about social norms (in particular if at stake is whether some people will end up feeling unwelcome or uncomfortable in a given community/space that is important to them - this is *extremely* threatening)
- More generally - the argument touches *in some way* on social reality, in ways that might have ramifications beyond the immediate conversation (or that people are *afraid* might have such ramifications).

If some of the above seem true (in particular, at least one of the first group and at least one of the second), then I think it's worth stepping back and being *very careful* about how you engage, even if no comment seems especially bad yet.

Potential Solutions

The first line of defense is to notice what's happening - recognize if you're feeling defensive or angry or talking past each other. Brienne's Noticing Sequence is pretty good for this (as well as her particular posts on training the skills of Empathy and handling Defensiveness - these may not work for everyone but I found the underlying thought process useful).

But while noticing is necessary, it's not *sufficient.*

Rather than list my first guesses here, I'll be discussing them in the comments and following this up with a "best-seeming of the potential solutions" post.

Meanwhile, some factors to consider as you decide what to do:

- How are you involved?
  - Are you one of the people initially arguing, or a bystander?
  - How much do you normally trust the people involved?
  - Is it possible to take the conversation private?
- Are we on the demon *seed* or demon *thread* stage? Is there common knowledge about either?
- What are the actual stakes?
- What are the moderation tools available to you?
- Are you in a venue where you have the ability to shape conversational norms?
  - Do you directly control them (i.e. personal blog or feed?)
  - Does *anyone* have direct ownership of the venue? (either technically, or culturally)
  - Is there anything you can do *unilaterally* to make the conversation better, or will it require help from others?
- Are you building a site where you get to develop entire new tools to deal with this class of problem?

With that in mind...

In whatever venues you most find yourself demon-thread-prone, what sort of plans can you actually think of that *might actually help?*

Note: I have since written a followup post with a working example of what I think people should usually do instead of demon threads.

What Is "About" About?

It has seemed to me that saying that something is "about" something else is probably the vaguest, least useful way possible to say that the two ideas are connected. I've used "about" myself in exactly those instances when I perceived a connection of some kind between two ideas, but I was incapable of articulating what that connection was, incapable even of seeing it myself.

"Well," I might say. "It seems like X is...*about* Y in some way...It's not exactly that X causes it...maybe X is a subsection of Y? No, that's not quite right, either. I'm not really sure, honestly, but to my eye, they seem linked to each other in some way. X is...about...Y."

Well, what did that mean, that it was "about" Y? Did they have a similar conceptual structure, were they causally related in one way or another, were they both discussed by the same groups, did they have similar effects on the world...like I said, it's vague. It could be any or none or all of these. Saying X is "about" Y tells you nothing very meaningful about them, so as a word, it's next-to-useless. Or, is it?

Now, I think I see what "about" is about; now I see it quite clearly. Saying X is "about" Y means, to say it very precisely, that you want people to give Y's connotations to X. For some reasons, X's connotations are undesirable, and you'd like Y's instead. Now, X is X, it is itself, so it is only natural that it will evoke its own connotations. If you're going to give it the more desirable connotations of Y, we must wrench the natural perspective off its natural path and force it down another.

Think of "it's not about who wins, it's about how you play the game," ie, please use the conceptual connections around this "how you play the game" idea to think about who won here today. Please *don't* use the connotations, do *not* use the subconscious connections linked to the "winning idea."

Now, what does "how you play the game" mean, exactly? No, that's not the point; the point isn't the point is to use those hard-to-define connotations, use the *feelings* around the "how you play the game" concept, the point is to take the aura around "how you play the game" and surround the "it" in question with them. You may not be exactly conscious of why it feels better, but you may feel it nonetheless. And, naturally, in this case, this attempt to change how people perceive the game is as much, or even more, about what it's *not* about. It *is* about "how you play the game," they say, yes, but also, they emphasize, it is *not* about winning. The speaker has some reason to ask you not to use the connotations around the "winner" concept to assess the "it" in question, probably, obviously, because they lost and are hoping to cut their losses.

They are hoping that, by repeating this truism, that "it's not about winning, it's about how you play the game," they can get you to join them in switching the game's assessment standards. They're hoping we'll all throw out the standards according to which they have unambiguously failed and swap in other standards which are more convenient to them, standards according to which, perhaps, they have succeeded, or at least they have mixed some success in to make up for their failure.

The political applications are clear enough, and you can pick up on this trick specifically almost anytime someone says "it's not about X, it's about Y," and in a great many other instances, you can pick up on the more general strategy from which

this one tactic is born: using language to get people to give one concept the connotations of another in order warp their perceptions.

One-Consciousness Universe

This is a linkpost for https://practicalontology.com/2018/01/23/one-consciousness-universe/

There's an idea in physics that there might in fact be only one electron in the universe, an electron that moves through space and time in such a way that it appears, when humans observe the world around them, like there's many different electrons all throughout the universe, all with precisely the same physical properties. There might be more to it, I wouldn't be qualified to say, but I view this one-electron universe model as an ontological koan. It makes us think "hey, reality could be this way rather than the way we think it is and we would be none the wiser — let's try to deepen our understanding of reality in light of that."

There's a short story by Andy Weir called *The Egg*. It falls under the umbrella of metaphysics porn which also covers The Matrix and Fight Club, the kind of story that makes some part of you believe it's true for the first half hour after you read/watch it, giving you a philosophical high — an intimation of a mystical experience. I don't want to deny you that, so go read it; it's short, shorter than this post, you've got no excuse. Did you read it? Well in case you're reading this after the collapse of civilization and Weir's story has been lost to time, I'll summarize the relevant detail. A man dies and meets God, who informs him that he (the man) is everyone. Everyone he has met, and everyone he hasn't across the planet and throughout history, past and future, is him. Every time he dies, he's sent to be born again as a different person in a different place and at a different time. God then sends the man off to be reincarnated as someone else. *The Egg* describes a one-consciousness universe.

Derek Parfit's theory of personal identity detailed in *Reasons and Persons* relies on what he calls *person-slices,* snapshots of a person at an instant in time. This is somewhat contentious because it's hard to make sense of what a person-slice is — what would it be *like* to be a person-slice? This seems to me to be analogous to asking "how can we observe a snapshot of an electron in time?" We can't! Observation can only be done over an interval of time, but just because we can't observe electron-slices doesn't mean that we shouldn't expect to be able to observe electrons over time, nor does the fact that we can observe electrons over time suggest that electron-slices are a nonsensical concept. Likewise, if there's nothing it's like to be a person-slice, that doesn't mean that person-slices are nonsense.

The one-consciousness universe in *The Egg* is beautiful because it's easy to comprehend, but it's painfully anthropocentric. If consciousness is fundamental to the universe, we shouldn't expect it to obey laws that reference human beings. This would be operating on the wrong layer of abstraction: human beings are not fundamental to the universe, they are emergent properties of some fundamental building blocks, like quarks and electrons. I find the one-electron universe model to be useful for helping us transcend our parochial intuitions about consciousness because it is sufficiently weird, without being too hard to comprehend. Let's say that there is a fundamental constituent of the universe called the conscion. It flits around throughout space-time visiting places all throughout the universe, and imbuing everything it visits with consciousness. Certain physical configurations, namely person-slices, exhibit continuity properties which make it *like* something to be the continuous sequence of those configurations through time. The order in which the conscion visits your person-slices makes no difference to what it's like to be you — in fact, it could visit one of

your person-slices, then visit someone else’s, then go back to yours and it would make no difference. It could even visit one of your person-slices more than once, without changing what it’s like to be you in that instant.

You are your person-slice at this very instant. The continuity of your experience is not the result of the essence of consciousness (the conscion) flowing through you from one moment to the next, it is not necessary that it do so. For your subjective experience to be continuous requires only that the person-slice that you are in this instant be continuous with person-slices past. And the notion of a person-slice doesn’t rely on the existence of consciousness, even a p-zombie can be viewed as a continuous set of person slices.

The O'Brien Technique

This is a linkpost for https://sinceriously.fyi/the-obrien-technique/

Schelling place for comments is here on LessWrong.

Paper: Artificial Intelligence in Life Extension: from Deep Learning to Superintelligence

There are two views on the best strategy among transhumanists and rationalists: The first involves the belief that one must invest in life extension technologies, and the latter, that it is necessary to create an aligned AI that will solve all problems, including giving us immortality or even something better. In our article, we showed that these two points of view do not contradict each other, because it is the development of AI that will be the main driver of increased life expectancy in the coming years, and as a result, even currently living people can benefit (and contribute) from the future superintelligence in several ways.

Firstly, because the use of machine learning, narrow AI will allow the study of aging biomarkers and combinations of geroprotectors, and this will produce an increase in life expectancy of several years, which means that tens of millions of people will live long enough to survive until the date of the creation of the superintelligence (whenever it happens) and will be saved from death. In other words, the current application of narrow AI to life extension provides us with a chance to jump on the “longevity escape velocity”, and the rapid growth of the AI will be the main factor that will, like the wind, help to increase this velocity.

Secondly, we can—here in the present—utilize some possibilities of the future superintelligence, by collecting data for “digital immortality”. Based on these data, the future AI can reconstruct the exact model of our personality, and also solve the identity problem. At the same time, the collection of medical data about the body will help both now—as it can train machine learning systems in predicting diseases—and in the future, when it becomes part of digital immortality. By subscribing to cryonics, we can also tap into the power of the future superintelligence, since without it, a successful reading of information from the frozen brain is impossible.

Thirdly, there are some grounds for assuming that medical AI will be safer. It is clear that fooming can occur with any AI. But the development of medical AI will accelerate the development of BCI interfaces, such as a *Neuralink*, and this will increase the chance of AI not appearing separately from humans, but as a product of integration with a person. As a result, a human mind will remain part of the AI, and from within, the human will direct its goal function. Actually, this is also Elon Musk’s vision, and he wants to commercialize his *Neuralink* through the treatment of diseases. In addition, if we assume that the *principle of orthogonality* may have exceptions, then any medical AI aimed at curing humans will be more likely to have benevolence as its terminal goal.

As a result, by developing AI for life extension, we make AI more safe, and increase the number of people who will survive up to the creation of superintelligence. Thus, there is no contradiction between the two main approaches in improving human life via the use of new technologies.

Moreover, for a radical life extension with the help of AI, it is necessary to take concrete steps right now: to collect data for digital immortality, to join patient organizations in order to combat aging, and to participate in clinical trials involving

combinations of geroprotectors, and computer analysis of biomarkers. We see our article as a motivational pitch that will encourage the reader to fight for a personal and global radical life extension.

In order to substantiate all of these conclusions, we conducted a huge analysis of existing start-ups and directions in the field of AI applications for life extension, and we have identified the beginnings of many of these trends, fixed in the specific business plans of companies.

Michael Batin, Alexey Turchin, Markov Sergey, Alice Zhila, David Denkenberger

“Artificial Intelligence in Life Extension: from Deep Learning to Superintelligence”

Informatica 41 (2017) 401–417:
http://www.informatica.si/index.php/informatica/article/view/1797

Schelling Orders

This is a linkpost for https://sinceriously.fyi/schelling-orders/

The different types (not sizes!) of infinity

In a recent conversation, a smart and mathy friend of mine revealed they didn't understand Cantor's diagonal argument. Further questionning revealed that she was using the wrong concept of infinity. I thought I'd pass on my explanations from there.

What is infinity? Does one plus infinity make sense? What about one over infinity? Well, it all depends what concept of infinity you're using. Roughly speaking, infinity happens when you take a finite concept and add a "and then that goes on for ever and ever" at the end. There are four major examples:

1. If you take the concept of "size" and push that to infinity, you get the infinite cardinals.
2. If you take the concept of "ordered set" and push that to infinity, you get the infinite ordinals.
3. If you take the concept of "increasing function" and push that to infinity, you get the poles and limits of continuous functions - a major part of real analysis.
4. And if you take normal algebra and push that to infinity, you get the hyperreals.

To add to the confusion, there are normally only two symbols for infinity to go around - ω and ∞ - even though there are at least four very different concepts (honourable mention also goes to the infinities of complex analysis and algebraic geometry, which are like more limited versions of the real analysis one).

So, does $\omega+1$ make sense (as something different from ω)? It does, for the ordinals and hyperreals only. What about $\omega-1$, similarly? That only makes sense for the hyperreals. And $-\omega$? That works for the hyperreals and real analysis. $1/\omega$? This is well defined for the hyperreals (where it is not equal to 0), and is arguably well-defined for real analysis, where it would be equal to 0.

Is ω times 2 the same as ω? "Yes", say real analysis and cardinals; "No", say ordinals and hyperreals. How about there being many different infinities of different sizes? That makes sense for all of these, except real analysis.

So when talking about "infinity", be careful that you're using the right concept of infinity, and don't mix one with another, or import the intuitions of one area into an arena where it doesn't make sense.

***

EDIT: See also Scott Garrabrant's comment:

Starting with the natural numbers:

- Cardinals: How many are there?
- Ordinals: What comes next?
- Limits: Where are they going?
- Hyperreals: What is bigger than all of them?

Roleplaying As Yourself

(*This is a basic intuition pump I've found helpful in making decisions, and maybe you'll like it too.)*

For all its shortcomings, I think there was something quite useful about the "What Would Jesus Do?" meme within the Christian framework. Of course it's not a very sophisticated ethical guide, and it comes with all kinds of biases; but asking it does put the believer into a frame of mind that emphasizes things like compassion and duty, and it sometimes helps the believer generate options that weren't in their default solution space.

Is there a version of this handy tool for the consequentialist, with our muddled mixture of selfish and altruistic goals and impulses, and the added difficulty that we're looking to actually optimize rather hard?

The one that works best for me is a double roleplay.

Jernau Gurgeh, champion of strategic games across the galaxy, sits down to a nice futuristic immersive roleplaying game: The Orthonormal Experience. Gurgeh will be controlling a denizen of the early 21st century on Earth, someone with the online name of Orthonormal. Getting Orthonormal to do well by Orthonormal's own standards is Gurgeh's objective.

Just as our roleplaying games have game masters who can call out uncharacteristic plans, so too does Gurgeh's game. He can't simply use his superior vantage point to calculate the right stocks for Orthonormal to buy today and sell tomorrow, because Orthonormal couldn't do that except by luck. He can't even have Orthonormal think at peak performance on some days- there are character attributes (penalties like Anxiety Disorder) he has to play around.

But Gurgeh is able to think patiently, and strategically, about the various obstacles blocking Orthonormal's progress, and to guide Orthonormal's thoughts in plausible ways to work on these. There's a lot of points out there to be scored: better states to reach in Orthonormal's relationships, career, inner life, and more.

What would Gurgeh do?

One last note: a couple of the bugs with this approach can be confronted within the approach itself. If I decide that Gurgeh might do X, and I try X and fail, it can be tempting to get frustrated with myself. But this isn't what Gurgeh would do next! He'd take my failure as more data about what this character's current attributes are, and look for ways to work around that failure mode or to train the relevant attribute. And he'd probably give the character a short rest to recover mana before trying again.

ProbDef: a game about probability and inference

Hey LessWrong! Do you still like probability, inference and Bayes' Theorem*? Well, I made a (free, open-source, HTML) game about using them to protect a spaceship from explosive mines.

I have plans to make an improved/expanded version at some point in the second half of this year. But I've not had much feedback on the current version; and feedback I *did* get focused mostly on its quality as a game instead of as an accurate and useful representation of probability concepts. So I'm posting it here, in the hopes that some risk/probability enthusiasts will share their opinions & give constructive criticism.

*Just so you know: Bayesian reasoning proper only shows up in the last third of the campaign. Yes, this is one of the things I plan on changing next iteration.

Insights from 'The Strategy of Conflict'

Cross-posted from my blog.

I recently read Thomas Schelling's book 'The Strategy of Conflict'. Many of the ideas it contains are now pretty widely known, especially in the rationalist community, such as the value of Schelling points when coordination must be obtained without communication, or the value of being able to commit oneself to actions that seem irrational. However, there are a few ideas that I got from the book that I don't think are as embedded in the public consciousness.

Schelling points in bargaining

The first such idea is the value of Schelling points in bargaining situations where communication *is* possible, as opposed to coordination situations where it is not. For instance, if you and I were dividing up a homogeneous pie that we both wanted as much of as possible, it would be strange if I told you that I demanded at least 52.3% of the pie. If I did, you would probably expect me to give some argument for the number 52.3% that distinguishes it from 51% or 55%. Indeed, it would be more strange than asking for 66.67%, which itself would be more strange than asking for 50%, which would be the most likely outcome were we to really run the experiment. Schelling uses as an example

> the remarkable frequency with which long negotiations over complicated quantitative formulas or *ad hoc* shares in some costs or benefits converge ultimately on something as crudely simple as equal shares, shares proportionate to some common magnitude (gross national product, population, foreign-exchange deficit, and so forth), or the shares agreed on in some previous but logically irrelevant negotiation.

The explanation is basically that in bargaining situations like these, any agreement could be made better for either side, but it can't be made better for both simultaneously, and any agreement is better than no agreement. Talk is cheap, so it's difficult for any side to credibly commit to only accept certain arbitrary outcomes. Therefore, as Schelling puts it,

> Each party's strategy is guided mainly by what he expects the other to accept or insist on; yet each knows that the other is guided by reciprocal thoughts. The final outcome must be a point from which neither expects the other to retreat; yet the main ingredient of this expectation is what one thinks the other expects the first to expect, and so on. Somehow, out of this fluid and indeterminate situation that seemingly provides no logical reason for anybody to expect anything except what he expects to be expected to expect, a decision is reached. These infinitely reflexive expectations must somehow converge upon a single point, at which each expects the other not to expect to be expected to retreat.

In other words, a Schelling point is a 'natural' outcome that somehow has the intrinsic property that each party can be expected to demand that they do at least as well as they would in that outcome.

Another way of putting this is that once we are bargained down to a Schelling point, we are not expected to let ourselves be bargained down further. Schelling uses the examples of soldiers fighting over a city. If one side retreats 13 km, they might be expected to retreat even further, unless they retreat to the single river running through the city. This river can serve as a Schelling point, and the attacking force might genuinely expect that their opponents will retreat no further.

Threats and promises

A second interesting idea contained in the book is the distinction between threats and promises. On some level, they're quite similar bargaining moves: in both cases, I make my behaviour dependent on yours by promising to sometimes do things that aren't narrowly rational, so that behaving in the way I want you to becomes profitable for you. When I threaten you, I say that if you don't do what I want, I'll force you to incur a cost even at a cost to myself, perhaps by beating you up, ruining your reputation, or refusing to trade with you. The purpose is to ensure that doing what I want becomes more profitable for you, taking my threat into account. When I make a promise, I say that if you do do what I want, I'll make your life better, again perhaps at a cost to myself, perhaps by giving you money, recommending that others hire you, or abstaining from behaviour that you dislike. Again, the purpose is to ensure that doing what I want, once you take my promise into account, is better for you than other options.

There is an important strategic difference between threats and promises, however. If a threat is successful, then it is not carried out. Conversely, the point of promises is to induce behaviour that forces you to carry out the promise. This means that in the ideal case, threat-making is cheap for the threatener, but promise-making is expensive for the promiser.

This difference has implications for one's ability to convince one's bargaining partner that one will carry out your threat or promise. If you and I make five bargains in a row, and in the first four situations I made a promise that I subsequently kept, then you have some reason for confidence that I will keep my fifth promise. However, if I make four threats in a row, all of which successfully deter you from engaging in behaviour that I don't want, then the fifth time I threaten you, you have no more evidence that I will carry out the threat than you did initially. Therefore, building a reputation as somebody who carries out their threats is somewhat more difficult than building a reputation for keeping promises. I must either occasionally make threats that fail to deter my bargaining partner, thus incurring both the cost of my partner not behaving in the way I prefer and also the cost of carrying out the threat, or visibly make investments that will make it cheap for me to carry out threats when necessary, such as hiring goons or being quick-witted and good at gossipping.

Mutually Assured Destruction

The final cluster of ideas contained in the book that I will talk about are implications of the model of mutually assured destruction (MAD). In a MAD dynamic, two parties both have the ability, and to some extent the inclination, to destroy the other party, perhaps by exploding a large number of nuclear bombs near them. However, they do not have the ability to destroy the other party immediately: when one party launches their nuclear bombs, the other has some amount of time to launch a second strike, sending nuclear bombs to the first party, before the first party's bombs land and annihilate the second party. Since both parties care about not being destroyed more

than they care about destroying the other party, and both parties know this, they each adopt a strategy where they commit to launching a second strike in response to a first strike, and therefore no first strike is ever launched.

Compare the MAD dynamic to the case of two gunslingers in the wild west in a standoff. Each gunslinger knows that if she does not shoot first, she will likely die before being able to shoot back. Therefore, as soon as you think that the other is about to shoot, or that the other thinks that you are about to shoot, or that the other thinks that you think that the other is about to shoot, et cetera, you need to shoot or the other will. As a result, the gunslinger dynamic is an unstable one that is likely to result in bloodshed. In contrast, the MAD dynamic is characterised by peacefulness and stability, since each one knows that the other will not launch a first strike for fear of a second strike.

In the final few chapters of the book, Schelling discusses what has to happen in order to ensure that MAD remains stable. One implication of the model that is perhaps counterintuitive is that if you and I are in a MAD dynamic, it is vitally important to me that you know that you have second-strike capability, and that you know that I know that you know that you have it. If you don't have second-strike capability, then you will realise that I have the ability to launch a first strike. Furthermore, if you think that I know that you know that you don't have second-strike capability, then you'll think that I'll be tempted to launch a first strike myself (since perhaps my favourite outcome is one where you're destroyed). In this case, you'd rather launch a first strike before I do, since you anticipate being destroyed either way. Therefore, I have an incentive to help you invest in technology that will help you accurately perceive whether or not I am striking, as well as technology that will hide your weapons (like ballistic missile submarines) so that I cannot destroy them with a first strike.

A second implication of the MAD model is that it is much more stable if both sides have more nuclear weapons. Suppose that I need 100 nuclear weapons to destroy my enemy, and he is thinking of using his nuclear weapons to wipe out mine (since perhaps mine are not hidden), allowing him to launch a first strike. Schelling writes:

> For illustration suppose his accuracies and abilities are such that one of his missiles has a 50-50 chance of knocking out one of ours. Then, if we have 200, he needs to knock out just over half; at 50 percent reliability he needs to fire just over 200 to cut our residual supply to less than 100. If we had 400, he would need to knock out three-quarters of ours; at a 50 percent discount rate for misses and failures he would need to fire more than twice 400, that is, more than 800. If we had 800, he would have to knock out seven-eighths of ours, and to do it with 50 percent reliability he would need over three times that number, or more than 2400. And so on. The larger the initial number on the "defending" side, the larger the *multiple* required by the attacker in order to reduce the victim's residual supply to below some "safe" number.

Consequently, if both sides have many times more nuclear weapons than are needed to destroy the entire world, the situation is much more stable than if they had barely enough to destroy the enemy: each is comforted in their second strike capabilities, and doesn't need to respond as aggressively to arms buildups by the other party.

It is important to note that this conclusion is only valid in a 'classic' simplified MAD dynamic. If for each nuclear weapon that you own, there is some possibility that a rogue actor will steal the weapon and use it for their own ends the value of large arms buildups becomes much less clear.

The final conclusion I'd like to draw from this model is that it would be preferable to not have weapons that could destroy other weapons. For instance, suppose that both parties were countries that had biological weapons that when released infected a large proportion of the other country, caused them obvious symptoms, and then killed them a week later, leaving a few days between the onset of symptoms and losing the ability to effectively do things. In such a situation, you would know that if I struck first, you would have ample ability to get still-functioning people to your weapons centres and launch a second strike, regardless of your ability to detect the biological weapon before it arrives, or the number of weapons and weapons centres that you or I have. Therefore, you are not tempted to launch first. Since this reasoning holds regardless of what type of weapon you have, it is always better for me to have this type of biological weapon in a MAD dynamic, rather than any nuclear weapons that can potentially destroy weapons centres, so as to preserve your second strike capabilities. I speculatively think that this argument should hold for real life biological weapons, since it seems to me that they could be destructive enough to act as a deterrent, but that authorities could detect their spread early enough to send remaining healthy government officials to launch a second strike.

Prune

This is a linkpost for https://radimentary.wordpress.com/2018/01/12/prune/

Previously, I described human thought-generation as an adversarial process between a low-quality pseudorandom Babble generator and a high-quality Prune filter, roughly analogous to the Generative Adversarial Networks model in machine learning. I then elaborated on this model by reconceptualizing Babble as a random walk with random restarts on an implicitly stored Babble graph.

Rationalist training (and schooling in general) slants towards developing Prune over Babble. I'm trying to solve the dual problem: that of improving the quality of your Babble.

Although the previous posts listed a number of exotic isolation exercises for Babble, I'm guessing nobody was inspired to go out and play more Scrabble, write haikus, or stop using the letter 'e'. That's probably for the best - taking these exercises too seriously would produce exotic but sub-optimal Babble anyway. For a serious solution to this serious problem, we need to understand Prune at a higher resolution.

The main problem with Prune is that it has too many layers. There's a filter for subconscious thoughts to become conscious, another for it to become spoken word, another for the spoken word to be written down, and a further one for the written word to be displayed in public. With this many-layer model in mind, there are plenty of knobs to turn to let more and better Babble through.

The River of Babble

Imagine that your river of Babble at its source, the subconscious: a foaming, ugly-colored river littered with half-formed concepts, too wild to navigate, too dirty to drink from. A quarter mile across, the bellow of the rapids is deafening.

Downstream, you build a series of gates to tame the rushing rapids and perhaps extract something beautiful and pure.

The First Gate, conscious thought, is a huge dam a thousand feet high and holds almost all the incoming thoughts at bay. Behind it, an enormous lake forms, threatening to overflow at any moment. A thick layer of trash floats to the top of this lake, intermixed with a fair amount of the good stuff. The First Gate lets through anything that satisfies a bare minimum of syntactical and semantic constraints. Thoughts that make it past the First Gate are the first ones you become conscious of - that's why they call the output the Stream of Consciousness.

A mile down the Stream of Consciousness is the Second Gate, spoken word, the filter through which thoughts become sounds. This Gate keeps you from saying all the foolish or risqué thoughts tripping through your head. Past the Second Gate, your spoken words form only a pathetic trickle - a Babbling Brook.

By now there is hardly anything left to sift from. The Third Gate, written word, is no physical gate but a team of goldpanners, scattered down the length of the Babbling Brook to pan for jewels and nuggets of gold. Such rare beauties are the only Babble

that actually make it onto paper. You hoard these little trinkets in your personal diary or blog, hoping one day to accumulate enough to forge a beautiful necklace.

Past the Third Gate, more Gates lay unused because there simply isn't enough material to fuel them: a whole chain of manufactories passed down from the great writers of yore. Among them are the disembodied voices of Strunk and White:

> Omit needless words. Vigorous writing is concise. A sentence should contain no unnecessary words, a paragraph no unnecessary sentences, for the same reason that a drawing should have no unnecessary lines and a machine no unnecessary parts. This requires not that the writer make all his sentences short, or that he avoid all detail and treat his subjects only in outline, but that every word tell.

Jealously clutching the 500-word pearls you drop once a month on your blog, you dream of the day when the capital comes through and these Gates will be activated to produce your magnum opus, your great American novel. For now, you can't afford to omit a single precious word.

The Gates of Prune

In the model above, there are many problems with Prune independent of having low-quality Babble to begin with. The Gates are working at odds with each other. They are individually too strict. There are simply too many of them. Lots of expensive mental machinery is not working at full capacity, if at all: if you have four Gates but 99% of the goods don't make it through the first one, that novel-writing factory you've built is not paying rent.

Even worse, there's probably two or three layers of subtlety within each of the big Gates I sketched. What you might whisper on a dark night in total solitude is different from what you might utter to a confidante is different from what you might say to your thesis adviser.

If a balanced Babble and Prune game is supposed to involve one Artist against one Critic, then having an overactive Prune is like pitting a pitchfork-wielding mob of Critics against one Artist. The first three Critics tar-and-feather the Artist and the rest are just there for moral support.

The task of relaxing all of Prune at once is monumental. Instead, relax the Gates individually in order. Simultaneously, shorten the psychological distance between them.

Relaxing and Shortening

At the First Gate, conscious thought, noticing is the way to let through more subconscious Babble. Practice noticing thoughts and sensations (not just confusion) that you never pay attention to. Much of meditation is devoted to relaxing this first Prune filter. Much of art is devoted to the motto: *make the familiar strange*, where strange is better translated as *salient*.

Another exercise along similar lines is zooming in on anything, anything at all. Pick up and stare at the whorls and aphids running down that twig on your driveway. Take

apart that broken old Canon in the attic. Dissect your aversions toward attending Algebraic Geometry.

At the Second Gate, spoken word, the trick is getting comfortable with vocalizing more of your Stream of Consciousness. I mentioned before that my internal process is very verbal - on reflection I think that whole post is about the maturation of my Prune filter to allow more Babble through. Several features stand out.

One of these features is that I directly mouth or whisper any thoughts that appear in my Stream of Consciousness. Psychologically, this shortens the distance between the First Gate and the Second Gate: it becomes a question of how loud to speak rather than whether or not to speak at all. There's no reason not to be constantly mouthing the things you're thinking, at least when you're alone. Similarly, when lost in thought I make micro-gestures with my fingers to imitate the emphatic ones I would make to convey that point in conversation. These tricks exploit the fact that the psychological distance between 1% and 100% is much shorter than that between 0% and 100%.

Another feature of my internal process is that I always have a mental audience: a silent judgmental muse, the personification of the Critic. In HPMOR, Harry has a supersized version of this: a whole cast of colorful mental characters that carry out full-length conversations with each other. This kind of dissociation-into-subpersonalities exercise has a whole of great side effects, but the relevant one for us is that it again shortens the mental gap between the First and Second Gate by making thinking feel like conversation.

Onwards to the Third Gate: the written word. Thankfully, modern technology has already radically shortened the distance between the Second and Third Gates for us with the invention of the blog, a medium much more free-form and personal than the book. Your training as a writer has probably erected a tall Third Gate, and successful bloggers have pretty much circumvented it.

What distinguishes blogging from formal writing? One metric is the frequency with which the blogger breaks the Fourth Wall - that poor Wall which is only mentioned when it is broken. Having torn down the Fourth Wall, blogging reduces naturally to a heated and cogent form of conversation, filled with rhetorical questions and injunctions.

Hey, look here, I'm not saying there's no place whatsoever in writing for formality. But if you're going to build a wall and call it the Fourth Wall, build it after the Third Gate, you know?

More Babble

This is a linkpost for https://radimentary.wordpress.com/2018/01/11/more-babble/

In my last babble, I introduced the Babble and Prune model of thought generation: Babble with a weak heuristic to generate many more possibilities than necessary, Prune with a strong heuristic to find a best, or the satisfactory one. I want to zoom in on this model. If the last babble was colored by my biases as a probabilist, this one is motivated by my biases as a graph theorist.

First, I will speculate on the exact mechanism of Babble, and also highlight the fact Babble and Prune are independent systems that can be mocked out for unit testing.

Second, I will lather on some metaphors about the adversarial nature of Babble and Prune. Two people have independently mentioned Generative Adversarial Networks to me, a model of unsupervised learning involving two neural nets, Generator and Discriminator. The Artist and the Critic are archetypes of the same flavor - I have argued in the past the spirit of the Critic is Satan.

Babble is (Sampling From) PageRank

Previously, I suggested that a Babble generator is a pseudorandom word generator, weighted with a weak, local filter. This is roughly true, but spectacular fails one of the technical goals of a pseudorandom generator: independence. In particular, the next word you Babble is frequently a variation (phonetically or semantically) of the previous one.

PageRank, as far as I know, ranks web pages by the heuristic of "what is the probability of ending up at this page after a random walk with random restarts." That's why a better analogy for Babble is sampling from PageRank i.e. taking a weighted random walk in your Babble graph with random restarts. Jackson Pollock is visual Babble.

Imagine you're playing a game of Scrabble, and you have the seven letters JRKAXN. What does your algorithm feel like?

You scan the board and see an open M. You start Babbling letter combinations that might start with M: MAJR, MRAJ, MRAN, MARN, MARX (oops, proper noun), MARK (great!). That's the weighted random walk. You set MARK aside and look for another place to start.

Time for a restart. You find an open A before a Triple Word, that'd be great to get! You start Babbling combinations that end with A: NARA, NAXRA, JARA, JAKA, RAKA. No luck.

Maybe the A should be in the middle of the word! ARAN, AKAN, AKAR, AJAR (great!). You sense mean stares for taking so long, so you turn off the Babble and score AJAR for (1+8+1+1)x3 = 33 points. Not too shabby.

The Babble Graph

Last time, I described getting better at Babble as increasing the uniformity of your pseudorandom Babble generator. With a higher-resolution model of Babble in hand, we should reconceptualize increasing uniformity as building a well-connected Babble graph.

What is the Babble graph? It's the graph within which your words and concepts are connected. Some of these connections are by rhyme and visual similarity, others are semantic or personal. Blood and snow are connected in my Babble graph, for example, because in Chinese they are homophones: snow is 雪 (xue), and blood is 血 (xue). This led to the following paragraph from one of my high school essays (paraphrased):

> In Chinese, snow and blood sound the same: "xue." Some people think the world will end suddenly in nuclear holocaust, pandemic, or a belligerent SkyNet. I think the world will die slowly and painfully, bleeding to death one drop at a time with each New England winter.

My parents had recently dragged me out to jog in the melting post-blizzard slush.

One of my favorite classes in college was a game theory class taught by the wonderful David Parkes; my wife and I lovingly remember the class as Parkes and Rec. One of the striking ideas I learned in Parkes and Rec is that exponentially large graphs can be compactly represented implicitly in memory, as long as individual edges and neighborhoods can be computed in reasonable time. Babble is capable of generating new words and combinations, so the Babble graph contains nodes you've never thought of. It's enormous, and definitely not (a subgraph of) the connectome, but rather implicitly represented therein in a compact way. This is related to the fact that the map is not the territory, except in the study of the brain, where the map is a subset of the territory.

It follows that the Babble graph is a massive implicitly represented graph, which is traversed via random walks with random restarts. How might we optimize this data structure to better fulfill its goals?

One technique I've already mentioned is to artificially replace either Babble or Prune to train the other in isolation. This is basically unit testing via mocking. To unit test Babble, we can mock out Prune with a simplified and explicit filter like the haiku, the game of Scrabble, or word games like Contact and Convergence. To unit test Prune, we replace Babble with other sources of word strings: reading, conversation, poetry, music.

Those methods completely black box the Babble and Prune algorithms and hope they self-optimize correctly. What if we want to get our hands dirty and explicitly rewire our Babble graph?

First we have to figure out what makes a quality Babble graph. I can think of two metrics worth optimizing:

1. Connectivity. With sufficient effort (i.e. taking enough random steps and restarts) you want to eventually explore the entire graph, and repeat yourself rarely. This requires not just that the graph is connected, but that it should have good expansion. Ever feel stuck on an idea, then be struck by external inspiration to explore a disconnected set of ideas you already knew, and find it massively

productive? Random walks getting trapped locally is a sign your Babble graph is a bad expander.
2. Value. Every node in your Babble graph should pay rent. I have found many abandoned components in my Babble graph - ghost towns and wastelands of neural machinery left over from experiences that are no longer relevant. They can be salvaged and repurposed, if only to generate metaphors.

Ramanujan was an extraordinarily creative mathematician who produced formulas like

$$p(n) \sim \frac{1}{4n\sqrt{3}} \exp \pi \sqrt{\frac{2n}{3}}$$

for the number of partitions p(n) of an integer n. Exercise: figure out how such an exponent might occur in nature. Hint: $\zeta(2) = \pi^2/6$.

Ramanujan was also known for his mysticism, attributing his most inspired results to his patron goddess. Mystical experiences, like LSD, are often characterized by the feeling of connectedness of all things. I think Ramanujan's genius might be the result of having a Babble graph that is an exceptionally good expander. What are those called again?

Here's a story about how I improved my Babble graph by making my bed.

It all started when Jordan Peterson told me to clean my room - because one's surroundings are a reflection of one's state of being. I decided to give it a chance and make my bed every morning.

Making my bed became a daily ritual. As I do it, I repeat the "proper and humble" mantra:

To save the world, I will start by doing the proper and humble things I know how to do within the confines of my own life.

Proper and humble were not words I'd liked in a very long time. They activated ideas I haven't wrestled with for years.

Honte is a Go term which means "the proper move." Honte is playing thickly to leave few weaknesses. Honte is killing already dead stones to remove aji. Honte is doing the proper and humble thing to prevent bad aji - failure modes you can't yet articulate. There's nothing quite like playing Go against a stronger player to put the fear of aji in you.

In relationships, honte is dedication to the removal of lingering resentment. Unhappy couples have the same fights at regular intervals; the landmines that trigger them might lay untouched for upwards of a year, but they never deactivate. Why would you allow these landmines be planted in the first place? You wouldn't leave a ladder breaker for your opponent in an unapproached corner, would you? Dedicate yourself to the removal of landmines, at least when you have the slack to do so. That's honte.

A well-connected and useful Babble graph is thickness (not to be confused with thiccness). It is written: *attack from thickness*. When thinking from a thick Babble

graph, you're not wandering lackadaisically, building an argument from scraps lying at the side of the trail. You'll have the weight of your entire intellectual life at your back.

The Artist and the Critic

Two people have independently suggested that the Babble and Prune model is similar to an approach in machine learning known as Generative Adversarial Networks, in which production of photorealistic images (say) is turned into a game between two neural nets, Generator, who learns to generate good counterfeits, and Discriminator, who works on finding the real stuff.

This is a manifestation of the eternal war between the Artist and the Critic, a war that is both exceedingly vicious and exceedingly productive. Artists of the ages have had some choice words for their critics. Beckett:

> VLADIMIR
> Moron!
> ESTRAGON
> That's the idea, let's abuse each other.
> *They turn, move apart, turn again and face each other.*
> VLADIMIR
> Moron!
> ESTRAGON
> Vermin!
> VLADIMIR
> Abortion!
> ESTRAGON
> Morpion!
> VLADIMIR
> Sewer-rat!
> ESTRAGON
> Curate!
> VLADIMIR
> Cretin!
> ESTRAGON
> *(with finality)* Crritic!
> VLADIMIR
> Oh!
> *He wilts, vanquished, and turns away.*

The opening lines of Hardy's *A Mathematician's Apology*:

> It is a melancholy experience for a professional mathematician to find himself writing about mathematics. The function of a mathematician is to do something, to prove new theorems, to add to mathematics, and not to talk about what he or other mathematicians have done. Statesmen despise publicists, painters despise art-critics, and physiologists, physicists, or mathematicians have usually similar feelings: there is no scorn more profound, or on the whole more justifiable, than that of the men who make for the men who explain. Exposition, criticism, appreciation, is work for second-rate minds.

I have a rule inspired by Solzhenitsyn, which is that every battle which occurs between human beings also plays out within each human heart. The proper locus of the fight between Artist and Critic is not cleanly between artists and critics, but between the Babble and Prune within each individual. After all, find me an artist who has never criticized, or a great critic who is never enjoyable to read for his own sake. Exercise: get some utility out of a bad book you've recently read by checking out the savage reviews online.

Like Generator and Discriminator, a good Artist and Critic pair can together ascend to heights that neither could reach alone, and having a filter is a healthy thing. However, I stand by my argument that the overdeveloped Critic is a manifestation of Satan:

> Jordan Peterson says Satan is an intellectual figure, and this idea has fermented in my imagination. Satan is the cynical and nihilistic intellectual whose thesis is "things are so bad they do not deserve to exist."
> [...]
> I would propose an embellishment of the figure of Satan as the nihilistic intellectual: Satan as the critic. One of the (many) disturbing things I have noticed about my high school curriculum is that English classes are factories for creating critics out of artists. At least in my experience, we wrote short stories, poems, and other free form essays in elementary and middle school, but turned exclusively to the analytical essay by the time high school rolled around.
> How frightening is that? Take a generation of teenagers, present them with the greatest literature of our civilization. Then, instead of teaching them to do the obvious thing – imitate – we teach them to analyze – the derivative work of a critic. The work of Satan: the intellectual whose ability to criticize far exceeds his ability to create. And so we find that the best students to come out of our high schools are created in the image of Satan. For every one budding novelist, we have a dozen teenage journalists, lawyers, and activists.
> Satan is the voice in your ear who says, "You will never do this well enough for it to be worth doing." This is the burrowing anxiety that puts me off writing for weeks at at time, the anxiety that anything I produce will not justify its own existence. The subroutine in your head constantly constructing impossibly high standards and handing them to you to use as excuses to do nothing. Satan is characterized by inaction, the inaction caused by paralyzing perfectionism.

Other Things that are Babble

The Bible is the best Babble ever produced. A common atheist refrain is that the Bible is so self-contradictory, so ambiguous, so open to interpretation as to be intrinsically meaningless. Any meaning you might extract from the Bible is just a reflection of your own beliefs.

I think this is a feature, not a bug.

Not only is the Bible open to interpretation, it *invites* interpretation. Its stories are so varied, fantastical and morally ambiguous that they *demand* interpretation. The Bible stood the test of time not because it is maximally packed with wisdom, but because it produced the most insightful and varied results when paired with outside sources of Prune. When the Christian is lost and desperate, he inputs 1 Corinthians to his Prune, and voilà! Faith is restored. Peterson's The Psychological Significance of the Bible series takes advantage of exactly this feature of the Bible: it is the fertile ground

upon which each individual can tell their own story. Of course, perversions can result when broken Prune filters are applied, even to the best Babble.

Perhaps writers have been optimizing for the wrong thing. Instead of directly packing insight into an essay, we should try to design high-quality Babble, fertile input for the reader's Prune.

The Oulipo is Babble training on steroids - a group of writers and mathematicians who worked based on the apparent paradox that freedom is the enemy of creativity. Creativity, the state of having a better Babble generator, is designed to solve tough, heavily constrained problems, and the Oulipians produced creative writing by imposing stricter restraints. Most famously, this method produced Perec's novel *La disparition*, a 300-page novel written without the letter 'e,' about "a group of individuals looking for a missing companion, Anton Vowl."

By the way, did you notice the letter missing from this entire post?

All good conversations are therapeutic, and therapeutic conversations are about letting down your guard and allowing yourself to simply Babble. Babies have no Prune at all and babble all the phonemes their adorable little mouths can produce - that's how they learn the beginnings of language so quickly. Being in a safe space is reproducing this state of development, a place where Babble can be rapidly be optimized on its own terms. Healthy teamwork and collaboration shares this quality: bouncing half-formed, half-nonsensical ideas off others and Pruning them together. Double the Babble, double the fun.

Oh, and about that missing letter? Just kidding. Ain't nobody got time for that.

[Paper] Global Catastrophic and Existential Risks Communication Scale, similar to Torino scale

We (Alexey Turchin and David Denkenberger) have a new paper out where we suggest a scale to communicate the size of global catastrophic and existential risks.

For impact risks, we have the Torino scale of asteroid danger which has five color-coded levels. For hurricanes we have the Saffir-Simpson scale of five categories. Here we present similar scale for communicating the size of the global catastrophic and existential risks.

Typically, some vague claims about probability are used as a communication tool for existential risks, for example, some may say, “There is a 10 per cent chance that humanity will be exterminated by nuclear war”. But the probability of the most serious global risks is difficult to measure, and the probability estimate doesn’t take into account other aspects of risks, for example, preventability, uncertainty, timing, relation to other risks etc. As a result, claims about probability could be misleading or produce reasonable skepticism.

To escape these difficulties, we suggested creating a scale to communicate existential risks, similar to the Torino scale of asteroid danger.

In our scale, there are six color codes, from white to purple. If hard probabilities are known, the color corresponds to probability intervals for a fixed timeframe of 100 years, which helps to solve uncertainty and timing.

However, for most serious risks, like AI, their probabilities are not known, but the required levels of prevention action are known. For these cases, the scale communicates the risk’s size through the required level of prevention action. In some sense, it is similar to Updateless Decision Theory, where an event’s significance is measured, not by observable probabilities, but by the utility of corresponding actions. The system would work, because in many cases of x-risks, the required prevention actions are not very sensitive to the probability.

How should the scale be implemented in practice? If probabilities are not known, a group of experts should aggregate available information and communicate it to the public and policymakers, saying something like: "We think that AI is a red risk, a pandemic is a yellow risk and asteroid danger is a green risk.” It would help to bring some order to the public perception of each risk—where, currently, asteroid danger is clearly overestimated compared to the risks of AI risk—without making unsustainable claims about unmeasurable probabilities.

In the article we have already given some estimates for the most well-known existential risks, but clearly they are open to debate.

Here's the abstract:

Existential risks threaten the future of humanity, but they are difficult to measure. However, to communicate, prioritize and mitigate such risks it is important to estimate their relative significance. Risk probabilities are typically used, but for

existential risks they are problematic due to ambiguity, and because quantitative probabilities do not represent some aspects of these risks. Thus, a standardized and easily comprehensible instrument is called for, to communicate dangers from various global catastrophic and existential risks. In this article, inspired by the Torino scale of asteroid danger, we suggest a color-coded scale to communicate the magnitude of global catastrophic and existential risks. The scale is based on the probability intervals of risks in the next century if they are available. The risks' estimations could be adjusted based on their severities and other factors. The scale covers not only existential risks, but smaller size global catastrophic risks. It consists of six color levels, which correspond to previously suggested levels of prevention activity. We estimate artificial intelligence risks as "red", while "orange" risks include nanotechnology, synthetic biology, full-scale nuclear war and a large global agricultural shortfall (caused by regional nuclear war, coincident extreme weather, etc.) The risks of natural pandemic, supervolcanic eruption and global warming are marked as "yellow" and the danger from asteroids is "green".

The paper is published in Futures
https://www.sciencedirect.com/science/article/pii/S001632871730112X

If you want to read the full paper, here's a link on preprint:
https://philpapers.org/rec/TURGCA

Two main pictures from the paper (I think that lesserwrong still not allow embedding pictures):

http://immortality-roadmap.com/xriskstab1.jpg

http://immortality-roadmap.com/xriskstab2.jpg

Making Exceptions to General Rules

Suppose you make a general rule, ie. "I won't eat any cookies". Then you encounter a situation that legitimately feels exceptional , "These are generally considered the best cookies in the entire state". This tends to make people torn between two threads of reasoning:

1) Clearly the optimal strategy is to make an exception this one time and then follow the rule the rest of the time.

2) If you break the rule this one time, then you are likely to dismantle the rule and ending up not following it at all, so you should whether the rule is good on the whole, instead of for this particular opportunity

How can we resolve this? For a very, very long time I didn't want to take option 2) because I felt that taking a sub-optimal strategy was irrational. It may seem silly, but I found this incredibly emotionally compelling and I had no idea of how to respond to this with logic. Surely rationality should never require you to be irrational? It took me literally years to work out, but 1) is an incredibly misleading way of framing the situation. Instead, we should be thinking:

3) If future you will *always* make the most rational decision, the optimal strategy is going to clearly be to make the exception. If there is a chance that it will cause you to fall off the path, either temporarily or permanently, then we need to account for these probabilities in the expected value calculation.

As soon as I realised this was the more accurate way of framing 1), all of its rhetorical power disappeared. Who cares about what is best for a perfectly rational agent? That isn't you. The other key benefit of this framing is that it pushes you to think in terms of probability. Too often I've thought, "I can make an exception without it becoming a habit". This is bad practise. Instead of thinking in terms of a binary "Can I?" or "Can't I?", we should be thinking in terms of probabilities. Firstly, because 0% probability is unrealistic, secondly because making a probability estimate allows you to better calibrate over time.

We can also update 2) to make it a more sophisticated argument as well.

4) If you break the rule this one time, then you risk dismantling the rule and ending up not following it at all. Further, humans tend to heavily biased towards believing that their future selves will make the decisions that they want it to make. So much so, that attempting the calculate this probability is hopeless. Instead, you should only make an exception if the utility gain would be so much that you would be willing to lose the habit altogether.

We still have two different ways of thinking of the problem, but at least they are more sophisticated than when we started.

(This article was inspired by seeing: The Solitaire Principle. I wanted to explain how I've progressed on this issue and I also wanted to have a post to link people to which is much shorter)

Singularity Mindset

This is a linkpost for https://radimentary.wordpress.com/2018/01/18/singularity-mindset/

In a fixed mindset, people believe their basic qualities, like their intelligence or talent, are simply fixed traits. In a growth mindset, people believe that their basic qualities can be modified gradually via dedication and incremental progress. Scott Alexander has cast some doubt on the benefits of growth mindset, but I think it still has merit, if only because it is closer to the truth.

Growth mindset is a good thing that doesn't do enough. The situation calls for More Dakka. I present: Singularity Mindset.

In a Singularity Mindset, people believe they are self-modifying intelligences well past the singularity threshold, that their basic qualities can be multiplied by large constants by installing the right algorithms and bug fixes, and that even the best optimized people are nowhere near hardware limitations. People who apply Singularity Mindset come to life with startling rapidity.

The seed of this post was planted in my head by Mind vs. Machine, which I read as a call-to-arms to be radically better humans, or at least better conversationalists. The seed sprouted when I noticed the other day that I've already blogged more words in 2018 (today is January 18) than in 2017.

Apparently, Kurzweil beat me to the name with Singularity University and Exponential Mindset, but (a) that's geared towards businesses and technology instead of individuals, and (b) I'm agnostic about the exact shape of the foom, so I'll stick to Singularity Mindset.

AI Worries

A tiny note of confusion noticed but unresolved can, over the course of years, fester into an all-consuming possession. One such question possessing me has come to a head:

Why is Eliezer Yudkowsky so much more worried about AI risk than I am?

I came up with a standard response rapidly, which goes something like: after years of steadfastly correcting his own biases, acquiring information, and thinking clearly about the world's biggest problems, he came to the right conclusion.

It's a good explanation which provokes in me at least a pretense of updating models, but today I will entertain another narrative.

I think the difference between us is that Eliezer has the lived experience of Singularity Mindset, of deliberately self-modifying to the point of becoming unrecognizably intelligent and productive, and the simultaneous lived experience of seeing his own values drift and require extraordinary effort to keep in line.

Meanwhile, I've been plodding up the incremental steps of the Temple of Growth Mindset, humbly and patiently picking up the habits and mental hygiene that arise

organically.

And so the difference between our worries about AI-risk might be described as us individually typical-minding an AI. Eliezer's System 1 says, "If AI minds grow up the way I grew up, boy are we in trouble." My System 1 says, "Nah, we'll be fine."

Take Ideas Seriously

Singularity Mindset is *taking ideas seriously:*

1. I took Jordan Peterson's advice seriously and cleaned my room. Turns out I actually love making the bed and decorating. A print of Kandinsky's Composition VIII is the best thing that happened to my room.
2. I went to see Wicked on Broadway, took it as seriously as possible, and was mesmerized. I ended up bawling my eyes out for the entire second hour.
3. I took More Dakka seriously and doubled the amount I blog every day until it became physically fatiguing.
4. I took my own advice about Babble and constrained writing seriously and wrote three short stories sharing the same dialogue.

Great ideas are not just data points. They are (at bare minimum) algorithms, software updates for your upcoming Singularity. To integrate them properly so that they become *yours* - to take them with the seriousness they deserve - requires not just an local update on the map, but at very least the design of a new cognitive submodule. In all likelihood, to get maximum mileage out of an idea requires a full-stack restructuring of the mind from the map down to the perceptual structures.

Take this quote of Jung's that I treasure:

Modern men cannot find God because they will not look low enough.

(More and more I find myself in the absurd position of writing on the ideas of this man who I find impossible to read directly but from whom I have derived such wisdom via second-hand sources.)

It is an injunction to be humble in directions orthogonal to the eighth virtue. To take Jung seriously deserves its own post, but in brief I read this quote in at least three directions.

Look low enough by focusing your mental energy on things that seem beneath you. Feed yourself properly, fix your sleep, and get exercise. Perhaps the most important thing you could be doing with your extraordinary intellectual capacity is to resolve the dysfunction within your immediate family. Perhaps the most important thing you could be writing involves repeating, summarizing, and coming up with catchy concept handles for the ideas of better men and women. Whatever it is, take it seriously, do it properly, and only good will come of it.

Look low enough by confronting the darkness in your personal hell. There are shadows there you instinctively convulse away from: horror movies, historical nightmares, the homeless on the street. Perhaps the most important thing you could be doing is admitting the existence of and then mastering your sadistic urges and delusions of genocide that lie just under the surface, ready to flood forth at the least opportune

moment. Perhaps the demon is instead a spiral of anxiety and self-doubt that sends you into sobbing fits in the fetal position. What you need in your life is exactly where you least want to look. Wield your attention against the darkness whenever you have the slack. Only light can defeat shadow.

Look low enough by looking to your inner child for guidance. Oftentimes, progress curves look like "naive, cynical, naive but wise":

1. For mathematicians, the curve is pre-rigor, rigor, post-rigor.
2. Picasso said, "It took me four years to paint like Raphael, but a lifetime to paint like a child."
3. Scott Alexander foretold that idealism is the new cynicism.
4. Knowing about biases can hurt you.

If you've plateaued for a long time in the cynical stage, look low enough by reconstituting your inner child. Relinquish your cynicism with the same quickness with which you relinquished your naiveté. Despite your "better" judgment, trust and forgive people. Feel small when you stand besides the ocean. Babble like a baby. Try stupid shit.

Taking ideas seriously is terrifying. It requires that at the drop of a hat, you are willing to extend such charity to a casual remark as to rebuild your whole mental machine on it if it proves true.

Extraordinary people take ideas with extraordinary seriousness. I will read a paper by skimming the abstract. "Huh, that sounds vaguely true." Scott Alexander will read the paper and write three detailed criticisms, each longer than the paper itself. Me on the other hand, in the last five years I've read more words in Scott's book reviews than in books themselves. What I'm after is that gripping but elusive experience of watching a mind take ideas seriously and completely synthesize them into a vast ocean of knowledge.

Is there a deep truth that caught your fancy recently, that you toss around with your friends the way Slytherins toss Remembralls? You thought it through once and you think you've done your due diligence?

Take that idea seriously. Reorganize your mind and life around it. Travel the world looking for examples of it. At very least, write a thousand words about it. God knows I want to hear about it.

Pareto improvements are rarer than they seem

This is a linkpost for http://reasonableapproximation.net/2018/01/27/pareto-improvements-rare.html

this is surely not an original insight, but I haven't seen it before

A Pareto improvement is where you make one party better off and no parties worse off.

Suppose Adam has a rare baseball card. He assigns no intrinsic value to baseball cards. Adam likes Beth, and somewhat values her happiness. Beth collects baseball cards, and would happily pay $100 for Adam's card.

If Adam just gives Beth his baseball card, is that a Pareto improvement? Naively, yes: he loses the card that he doesn't care about, and gains her happiness; she gains the card. Both are better off.

But I claim not, because if Adam has the card, he can sell it to Beth for $100. He would much prefer doing that over just giving her the card. But if Beth has the card, he can't do that. He assigns no intrinsic value to the card, but he can still value it as a trading chip.

Now suppose Adam has the baseball card but Beth also has a copy of that card. Then Beth has less desire for Adam's card, so this situation also isn't a Pareto improvement over the original. By giving something to Beth, we've made Adam's situation worse, *even though Adam likes Beth and values her happiness* .

And I think situations like this are common. The ability to give someone something they want, is a form of power; and power is instrumentally useful. And the less someone wants, the less able you are to give them something they want[1].

For a closer-to-reality example, the reddit comment that sparked this post said:

> bringing Platform 3 back into use at Liverpool Street Underground Station was denied because the platform would not be accessible. Neither of the platforms currently in use for that line is accessible, so allowing Platform 3 to be used would be a Pareto improvement

The model here is that there are two parties, people who can access the platforms at Liverpool St and those who can't. If Platform 3 is brought back into use, the first group gains something and the second group loses nothing.

But I think that if Platform 3 is brought back into use, the second group loses some power. They lose the power to say "we'll let you bring back Platform 3 if you give us...". Maybe Platform 3 can be made accessible for $1 million. Then they can say "we'll let you bring it back if you make it accessible", but they can't do that if it's already back in use.

And they lose some power to say "if you ignore us, we'll make things difficult for you". Maybe it would take $1 trillion to make Platform 3 accessible. If Platform 3 remains out

of use, people are more likely to spend $1 million to make their building projects accessible, because they've seen what happens if they don't. Conversely, if Platform 3 comes back, people are more likely to exaggerate future costs of accessibility. "If I say it costs $1 million, I'll have to pay. If I say it costs $10 million, maybe I won't."

I haven't researched the situation in question, and I expect that the actual power dynamics in play don't look quite like that. But I think the point stands.

(My original reply said: "If it's easier to turn an unused inaccessible platform into a used accessible platform, than to turn a used inaccessible platform into a used accessible platform - I don't know if that's the case, but it sounds plausible - then opening the platform isn't a Pareto improvement." That still seems true to me, but it's not what I'm talking about here. There are lots of reasons why something might not be a Pareto improvement.)

This doesn't mean Pareto improvements don't exist. But I think a lot of things that look like them are not.

Update 2018-02-02: some good comments on reddit and LessWrong. Following those, I have two things in particular to add.

First, that I like /u/AntiTwister's summary: "If you have the power to prevent another party from gaining utility, then you lose utility by giving up that power even if you are allies. There is opportunity cost in abstaining from using your power as a bargaining chip to increase your own utility."

Second, that there is a related (weaker) concept called Kaldor-Hicks efficiency. I think that a lot of the things that look-like-but-aren't Pareto improvements, are still Kaldor-Hicks improvements - meaning that the utility lost by the losing parties is still less than the utility gained by the winners. In theory, that means that the winners could compensate the losers by giving them some money, to reach a Pareto improvement over the original state. But various political and practical issues can (and often do) get in the way of that.

---

1. This feels like it generalizes far beyond questions of Pareto efficiency, but I'm not sure how to frame it. Something like *game theory is more competitive than it appears*. Even when no two players value the same resource, even when all players genuinely want all other players to do well, players still have an incentive to sabotage each other.

Adequacy as Levels of Play

Adequacy as Levels of Play

One method that I find very useful for evaluating adequacy concerns is "level of play". If you look at different games or different leagues of the same game, it's pretty apparent that the "level of play" - the amount of athleticism, effort, skill, planning, strategy, etc. that is on display - is quite different.

For instance, in the United States the NFL is operating at a higher level of play than college football. Similarly, baseball in the United States operates at a higher level of play than baseball in Japan, and either is significantly elevated compared to the rest of the world. In South Korea, Starcraft is treated as a professional sport, and is thus predictably operating at a higher level than Starcraft in the United States. This sort of consideration doesn't just apply to sports - cryptocurrency trading operated at an obviously much lower level of play than normal finance for a long time (though that may be changing).

A similar concept can be applied to general adequacy - one can quite usefully analyze existing organizations or programs by asking "what's the level of play here?" There are three basic questions that I think are quite useful for analyzing what level something is on:

- *Seriousness*. Do people take this seriously? There are many areas that nobody really tries very hard at, and those areas usually operate at a low level of play - why bother to do it right if you don't really care that much? In sports, think about the difference between a niche sport like jai alai and something like soccer.
- *Competitiveness*. Do different groups compete to do better? If so, how close is the competition? There are some fields that are taken very seriously but are nevertheless not very competitive - medicine, for instance - and this tends to yield lower-level outcomes overall, since there is less incentive to integrate new advances into the system.
- *Aliveness.* Are the conclusions being tested in actual practice? If not, are they being tested under realistic training conditions? It may surprise you to learn that there are matters that are taken seriously and competitively and yet not actually tested, but this is the state of the world! The biggest example of this I can think of would, surprisingly enough, be military strategy. History is rife with examples of situations where war planners, despite being very serious, well-funded and in direct competition (sometimes to the point of arms races), nevertheless made grevious errors - errors which led to the deaths of thousands. One of the main reasons such drastically wrong decisions can be made is that it is very difficult to actually test various military strategies under realistic conditions, and so false theories easily propagate in the absence of actual wars.

Asking these questions is a great quick way to evaluate "level of play" and hence adequacy. For instance, to take three quick examples of my evaluations:

- The level of play in investment banking is quite high. People take it quite seriously, it's very competitive, and people's conclusions are tested all the time.
- The level of play in Magic: the Gathering is lower than that but still pretty high - people take it quite seriously (as games go), it's very competitive, and people's conclusions are tested all the time.

- The level of play in asteroid impact avoidance, on the other hand, is not very high. While this is certainly a serious affair in one sense, it isn't highly funded, it isn't very competitive, and it's hard to do tests.

I find this method and framing quick and easy to use - when I frame things in terms of "adequacy" it seems at times a little distanced from the case, but when I frame them in terms of levels of play they end up much easier for me to work with.

Try this out and see if it works for you!

Have you felt exiert yet?

Pre-adolescent children haven't felt strong lust yet. Those of us who've avoided strong pain are also missing an experience of that affect. Nostalgia can come up very early, but does require a bit of living first. Depression can strike at any age.

So, in general, there are emotions and feelings that people are capable of feeling, in the right circumstances, but that some people have not yet felt.

As a thought experiment, I'd thus like to introduce the emotion of exiert, which no human being has yet felt, because it's only triggered by (very) unusual experiences. Maybe it's something that can only be felt after a decade of weightlessness, or after the human brain has adapted to the living for a long time within the atmosphere of a gas giant.

Let's assume that exiert has some survival uses - it's helpful in navigating the gas giant's ever-changing geography, say - and it can load onto both positive and negative affect (just as confusion and surprise can). Assume also that experiencing exiert can have impact on people's aethetic preferences - it can cause them to like certain colour schemes, or open them to different types of pacing and tone in films or theatre productions.

In the spirit of trying to deduce what human values really are, the question is then: is the AI overriding human preferences in any of the following situations:

1. The AI premeptively precludes the experience of exiert, so that we won't experience this emotion even if we are put in the unusual circumstances.
2. Suppose that we wouldn't "naturally" experience exiert, but the AI acts to ensure that we would (in those unusual circumstances).
3. The AI acts to ensure that some human that had not yet experienced strong lust, nostalgia, or extreme pain, could never experience that emotion.
4. Suppose some human would not "naturally" experience strong lust, nostalgia, or extreme pain, but the AI acts to ensure that they would.
5. The AI acts to ensure that some human experiences exiert, but that this is triggered by normal circumstances, rather than unusual ones.

Magic Brain Juice

This is a linkpost for https://radimentary.wordpress.com/2018/01/26/magic-brain-juice/

Shorter and less Pruned due to CFAR.

> A grandfather is talking with his grandson and he says there are two wolves inside of us which are always at war with each other.
> One of them is a good wolf which represents things like kindness, bravery and love. The other is a bad wolf, which represents things like greed, hatred and fear.
> The grandson stops and thinks about it for a second then he looks up at his grandfather and says, “Grandfather, which one wins?”
> The grandfather quietly replies, the one you feed.

I circumambulated the idea of meta-processes with the wonderfully inscrutable SquirrelInHell recently, and a seed of doubt has been circling in my head like a menacing sharkfin ever since.

At grave peril of strawmanning, a first order-approximation to SquirrelInHell’s meta-process (what I think of as the Self) is the only process in the brain with write access, the power of self-modification. All other brain processes are to treat the brain as a static algorithm and solve the world from there.

It seems to me that due to the biology of the brain there is a very serious issue with isolating the power of self-modification to the meta-process. After all, every single thought and experience causes self-modification at the neural level.

This post is another step towards a decision theory for human beings.

Unintentional Self-Modification

There is a central theme buried in my post The Solitaire Principle about building habits across time: human beings are not rational agents. We are not even “bounded-rationality agents,” whatever that means. We are agents who cannot simply act because every action is accompanied by self-modification.

Every time you take an action, the associated neural pathways are bathed in the magic brain juice [citation needed]. When you go to the gym, it becomes easier to decide to go to the gym next time. The activation energy for the second blog post you write is lower than that of the first. Acquired tastes are a real thing. After repeating a habit for a month, it is practically free.

Due to magic brain juice, every action is accompanied by an unintentional self-modification.

Let that sink in.

Every action you take is accompanied by an unintentional self-modification.

In an iterated game involving human beings, the choices made in each round influence not only their scores but their utility functions in perpetuity. Even

disregarding sunk costs and irrational attachments etc etc, it literally becomes easier for Brain 1 to press Button A the second time around.

The Ten Percent Shift

The Ten Percent Shift is a thought experiment I've successfully pushed to System 1 that helps build long-term habits like blogging every day. It makes the assumption that each time you make a choice, it gets 10% easier.

Suppose there is a habit you want to build such as going to the gym. You've drawn the pentagrams, sprinkled the pixie dust, and done the proper rituals to decide that the benefits clearly outweigh the costs and there's no superior alternatives. Nevertheless, the effort to make yourself go every day seems insurmountable.

You spend 100 units of willpower dragging yourself there on Day 1. Now, notice that you have magic brain juice on your side. On Day 2, it gets a little bit easier. You spend 90 units. On Day 3, it only costs 80.

A bit of math and a lot of magic brain juice later, you spend 500 units of willpower in the first 10 days, and the habit is free for the rest of time.

So the thought experiment is this: feel how difficult going to the gym is once. Call that x units of effort. Now, imagine you get to trade 5x that total effort for the results of going the gym for a year. Decide whether you go today based on your reaction to the Ten Percent Shift.

The exact details of the Ten Percent Shift don't matter - the goal of the Ten Percent Shift is to convince System 1 that a single act of installing individual daily activities has far-reaching consequences. Thus:

1. Pick a time horizon that feels real to you.
2. If the exact numbers bother you, insert your own model of decaying effort curves. Realistically, the effort cost will decay to a positive constant rather than zero.
3. Notice the effect I called magic brain juice actually decomposes into multiple psychological factors, some of which decay over time. Thus, daily habits benefit more from magic brain juice than do every-other-day habits, which are in turn vastly easier to build than weekly habits.
4. During the new routine, practice noticing and mindfulness to amplify the effects of magic brain juice.

Conversational Presentation of Why Automation is Different This Time

I have been frustrated recently with my inability to efficiently participate in discussions of automation which crop up online and in person. The purpose of the post is to refine a conversational presentation of what I believe to be the salient concerns; the chief goals are brevity and clarity, but obviously corrections of fact supersede this.

Epistemic status: plausible causal conjecture.

I think the current wave of automation will be different from previous ones, in ways which make it more disruptive. There are three reasons for this:

No Fourth Sector: The economy has three broad sectors: agriculture, manufacturing, and services. The first wave was in agriculture, and people could find adjacent work or switched to working in manufacturing. The second wave was in manufacturing, and people could find adjacent work or switched to work in services. The current wave is affecting services, but there is no fourth sector of the economy left for workers to switch to.

Skills Over Jobs: Agricultural automation was largely about tasks: a digging machine, a seeding machine, a pulling machine. Manufacturing automation took this to the next level, with robots performing defined sequences of tasks. But in both cases these were specific - any task or series of tasks which had not been specifically automated was still work to be had. The new wave of automation is entire skillsets, like *apply this pattern* or the *ability to speak*. This means when a job is lost to automation, all similar jobs are going away at the same time. There will be no adjacent work for people to switch to.

Speed: When automation was physical machines, they had to design them, and build them, and ship them, and customers had to rebuild their own factories to use manufacturing robots. Modern automation is largely software driven, so design and build are the same process, which is then practically free to copy and distribute. As soon as the method is ready, it can be picked up by businesses as fast as they can rent server space to run it. This gives local economies and institutions like government very little time to respond.

Automation is different this time because the problems we experienced last time will be more severe, and more widespread, and happen faster.

Niceness Stealth-Bombing

Epistemic status: Tested in an environment very favorable to its success.

Guys, I know it's hard to believe and I'm sure very few of us have experience with it, but sometimes the people around us have opinions we think are really stupid.

Fortunately, there's an easy solution to this problem. It's a well-established fact that telling people they're Hitler immediately converts them to your side of the issue.

Rationalists are supposed to discuss topics with the goal of finding the truth. Non-rationalists are supposed to discuss topics with the goal of convincing others they're right (or at least I assume that if you asked they'd say that, not "WE MUST POUND OUR ENEMIES INTO THE EARTH!"). Neither of these two goals is fulfilled when debates turn acrimonious, when we respond to disagreement with fury or to fury with more fury. People who perceive enemy fire drop their original goals in the process of picking up their laser blasters. I'm gonna go really radical here and say that setting yourself against someone is not a great way to make them honestly consider your points. And that this is obvious enough that people who do it anyway have probably lost sight of their priorities. And that forgetting what you're trying to do is really bad if you want to actually do it. Maybe some people enjoy this kind of argument, and maybe there are social benefits to showing that certain opinions will be shouted down. But it's still better to convert your enemies than to shut them up.

So I present to you: niceness stealth-bombing.

Niceness stealth-bombing slips quietly under the "enemy argument" radar, so confirmation bias and all its friends don't know to shoot it out of the sky. The military metaphors end there, because stealth-bombing means finding a point at which we're all on the same side.

The trick is to communicate specific disagreement, but broader support. You have to demonstrate to the other person that you share a goal with them, and that you think your idea will help them reach that goal. You also have to communicate that you see them as an ally, that you recognize they're also playing for the good guys. "It's awesome that you care enough about helping people to give up your own money for strangers. I just think that checking out GiveWell would help you help people even more." Or "I agree with you that x-risks sound pretty crazy, I totally reacted the same way when I first heard about them. But I'm sure we both agree that human extinction would be bad, so if they turned out to be real, they'd be really important, don't you think? So you should check them out and judge for yourself, because being wrong could be really bad."

You have to find a point of agreement, even if it's as deeply buried as "We should make the economy better" or "Making people happy is good". You have to hold really tight to that shared goal and remember that the other person could help you reach it. And, of course, you have to be nice. People don't listen if you're not nice. In most circumstances, if the disagreement damages your relationship, you're probably not stealth-bombing right.

You can use a modified version of niceness stealth-bombing to defuse fury directed at you. "It would be really bad if the things you said about me were true. Can you explain to me what makes you think that, so I can try to do better in the future?" Or simply,

“Okay, what should I do to fix this problem?” If you sound sincere enough, you’ll either redirect the conversation onto a productive track or stump them if the object-level disagreement was just an excuse to yell.

I used the first version of stealth-bombing, the persuasive kind, very often and very successfully in high school class discussions. (This was recent; I’m a college freshman.) I managed not only to frequently change people’s minds, but also to never make anyone hate me for trying. However, most of my classmates were very smart and demographically similar to me, they knew and mostly liked me and respected my opinion, and we were nearly all on the same side of the political spectrum so our actual disagreements were not often very large. I would guess that stealth-bombing is less successful in environments where any of these factors is changed, except maybe intelligence since it’s not an appeal to logic. Also, I don’t have a control here since I never actually tried being rude. Maybe I’d have convinced people that way too, but it seems unlikely.

I’ve used the second kind of stealth-bombing, the response-to-fury kind, when people were angry at me. It’s defused fights very quickly when the other person really did want a specific problem solved. When I’ve used it with people who clearly just want to yell, they’ve settled into angry brooding. Which, at least, is blessedly quiet.

One thing I do feel confident in saying is that niceness stealth-bombing done right both requires and expands empathy. You have to recognize the good motivations of your opponents, and conceive of them as potential allies. I think that’s a very good thing to practice. It reminds you what we all should be: team players for the human race.

Field-Building and Deep Models

What is important in hiring/field-building in x-risk and AI alignment communities and orgs? I had a few conversations on this recently, and I'm trying to publicly write up key ideas more regularly.

I had in mind the mantra 'better written quickly than not written at all', so you can expect some failures in enjoyability and clarity. No character represents any individual, but is an amalgam of thoughts I've had and that others have raised.

ALBERT cares deeply about x-risk from AI, and wants to grow the field of alignment quickly; he also worries that people in x-risk community errs too much on the side of hiring people similar to themselves.

BEN cares deeply about x-risk from AI, and thinks that we should grow the AI safety community slowly and carefully; he feels it's important to ensure new members of the community understand what's already been learned, and avoid the eternal september effect.

ALBERT: So, I understand you care about picking individuals and teams that agree with your framing of the problem.

BEN: That sounds about right - a team or community must share deep models of their problems to make progress together.

ALBERT: Concretely on the research side, what research seems valuable to you?

BEN: If you're asking what I think is most likely to push the needle forward on alignment, then I'd point to MIRI's and Paul's respective research paths, and also some of the safety work being done at DeepMind and FHI.

ALBERT: Right. I think there are also valuable teams being funded by FLI and Open Phil who think about safety while doing more mainstream capabilities research. More generally, I think you don't need to hire people that think very similarly to you in your organisations. Do you disagree?

BEN: That's an interesting question. On the non-research side, my first thought is to ask what Y Combinator says about organisations. One thing we learn from YC are that the first 10-20 hires of your organisation will make or break it, especially the co-founders. Picking even a slightly suboptimal co-founder - someone who doesn't perfectly fit your team culture, understand the product, and work well with you - is the easiest way to kill your company. This suggests to me a high prior on selectivity (though I haven't looked in detail into the other research groups you mention).

ALBERT: So you're saying that if the x-risk community is like a small company it's important to have similar views, and if it's like a large company it's less important? Because it seems to me that we're more like a large company. There are certainly over 20 of us.

BEN: While 'size of company' is close, it's not quite it. You can have small companies like restaurants or corner stores where this doesn't matter. The key notion is one of

inferential distance.

To borrow a line from Peter Thiel: startups are very close to being cults, except that where cults are very *wrong* about something important, startups are very *right* about something important.

As founders build detailed models of some new domain, they also build an inferential distance of 10+ steps between themselves and the rest of the world. They start to feel like everyone outside the startup is insane, until the point where the startup makes billions of dollars and then the update propagates throughout the world ("*Oh, you can just get people to rent out their own houses as a BnB*").

A founder has to make literally thousands of decisions based off of their detailed models of the product/insight, and so you can't have cofounders who don't share at least 90% of the deep models.

ALBERT: But it seems many x-risk orgs could hire people who don't share our basic beliefs about alignment and x-risk. Surely you don't need an office manager, grant writer, or web designer to share your feelings about the existential fate of humanity?

BEN: Actually, I'm not sure I agree with that. It again comes down to how much the org is doing new things versus doing things that are central cases of a pre-existing industry.

At the beginning of Open Phil's existence they wouldn't have been able to (say) employ a typical 'hiring manager' because the hiring process design required deep models of what Open Phil's strategy was and what variables mattered. For example 'how easily someone can tell you the strength and cause of their beliefs' was important to Open Phil.

Similarly, I believe the teams at CFAR and MIRI have optimised workshops and research environments respectively, in ways that depend on the specifics of their particular workshops/retreats and research environments. A web designer needs to know the organisation's goals well enough to model the typical user and how they need to interact with the site. An operations manager needs to know what financial trade-offs to make; how important for the workshop is food versus travel versus ergonomics of the workspace. Having every team member understand the core vision is necessary for a successful organisation.

ALBERT: I still think you're overweighting these variables, but that's an interesting argument. How exactly do you apply this hypothesis to research?

BEN: It doesn't apply trivially, but I'll gesture at what I think: Our community has particular models, worldview and general culture that helped to notice AI in the first place, and has produced some pretty outstanding research (e.g. logical induction, functional decision theory); I think that the culture is a crucial thing to sustain, rather than to be cut away from the insights it's produced so far. It's important, for those working on furthering its insights and success, to deeply understand the worldview.

ALBERT: I agree that having made progress on issues like logical induction is impressive and has a solid chance of being very useful for AGI design. And I have a better understanding of your position - sharing deep models of a problem is important. I just think that some other top thinkers will be able to make a lot of the key inferences themselves - look at Stuart Russell for example - and we can help that along by providing funding and infrastructure.

Maybe we agree on the strategy of providing great thinkers the space to think about and discuss these problems? For example, events where top AI researchers in academia are given the space to share models with researchers closer to our community.

BEN: I think I endorse that strategy, or at least the low-fidelity one you describe. I expect we'd have further disagreements when digging down into the details, structure and framing of such events.

But I will say, when I've talked with alignment researchers at MIRI, something they want more than people working on agent foundations, or Paul's agenda, are people who grok a bunch of the models and *still have disagreements*, and work on ideas from a new perspective. I hope your strategy helps discover people who deeply understand and have a novel approach to the alignment problem.

For proofreads on various versions of this post, my thanks to Roxanne Heston, Beth Barnes, Lawrence Chan, Claire Zabel and Raymond Arnold. For more extensive editing (aka telling me to cut a third of it), my thanks to Laura Vaughan. Naturally, this does not imply endorsement from any of them (most actually had substantial disagreements).

The Tallest Pygmy Effect

Status: I thought this was a common economics term, but when I google it I get either unrelated or references using it the way I expect but not defining it. It's a really useful term, so I'm going to attempt to make it a thing.

"Tallest Pygmy Effect" is when you benefit not from absolute skill or value at a thing, but by being better at it than anyone else. For example, the US dollar is not that great a currency and the US economy is not that great an economy. However, the dollar is more stable than other currencies, so it becomes the currency of choice when you want stability. This high volume makes USD more stable and is in general good for the US economy (because e.g. US companies don't have to take on currency risk when they borrow money).

Tallest pygmy effects are fragile, especially when they are reliant on self-fulfilling prophecies or network effects. If everyone suddenly thought the Euro was the most stable currency, the resulting switch would destabilize the dollar and hurt both its value and the US economy as a whole.

The Solitaire Principle: Game Theory for One

This is a linkpost for https://radimentary.wordpress.com/2018/01/16/the-solitaire-principle-game-theory-for-one/

> Do I contradict myself?
> Very well then I contradict myself;
> (I am large, I contain multitudes.)

This post is an exercise in taking Whitman seriously. If the self is properly understood as a loose coalition of many agents with possibly distinct values, beliefs, and incentives, what does game theory have to say about self-improvement?

The Solitaire Principle is the principle that human beings can be usefully thought about as loose coalitions of many agents. Classes of interpersonal problems often translate into classes of intrapersonal problems, and the tools to solve them are broadly similar. The Solitaire Principle is a corollary of the paradigm that the universe is self-similar at every level of organization: the organizational principles and faults of a civilization are not wildly different from those of a single human mind.

Self-improvement is often framed in terms of *optimization* of a monolithic whole. Instead, the Solitaire Principle suggests that self-improvement can also be achieved by *alignment* of pieces within the whole to cooperate more efficiently.

First, I fractionate the self across the time dimension and investigate self-improvement as an iterated game for one. This is partially inspired by this essay on becoming more legible to other agents.

Second, I fractionate the self into multiple sub-personalities and investigate self-improvement as a single sub-personality taking unilateral action to improve the whole.

1. Iterated Games for One

i. Basic Thought Experiments

Imagine that a human being dies and is re-instantiated the following day. Across a year, one agent A actually behaves like 365 very weakly dependent agents A1, A2, ..., A365.

A1 wants to write a novel, and can either write a page today (cooperate) or Netflix (defect). The novel is completed if and only if A1, A2, ..., A365 all cooperate. A1 decides the probability of that happening is vanishingly small, so she defects. No pages are written.

B1 wants to write a novel. The novel is completed if at least 300 of B1, B2, ..., B365 all cooperate. B1 simulates the 364 other agents and expects only half of them to cooperate. B1 defects and no pages are written.

C1 wakes up on New Year's Day inspired to write a novel. C1 feels excited about the project and decides it's likely that everyone will cooperate, so she writes Page 1. The other 364 agents don't know about the book. One page is written.

D1 wakes up on New Year's Day inspired to write a novel. He gets *Write Novel* tattooed on his arm to broadcast his intent to the others. All the agents now know about the book project, but the other guys aren't excited about it. One page is written.

E1 has always wanted to write a novel. Given that E(-364),...,E(-1), E0 didn't already start the novel, E1 reasons that she is not the kind of person who would be able to follow through with a project of this magnitude. E2,...E365 reason similarly. No pages are written.

F1 has always wanted to write a novel. F1 reasons that he is not currently the kind of person who would be able to follow through with a project like this. He reads a self-help book to fix this state of affairs, and broadcasts his intention. F2, ... , F365 also reason that given the previous agents have not started, they are probably not yet ready. 365 self-help books are read, but no pages are written.

G1 has always wanted to write a novel. She designs an hour-long morning meditation to reflect on the importance of mindfulness and writing to her life. She performs this ritual before writing a page. The ritual shifts the kind of person G2, ..., G365 are, so that they are individually 10% more likely to repeat it. One of G2, ..., G11 (in expectation) repeats the ritual and shifts the kind of person G is by another 10%, for a total of 20%. One of the next five (in expectation) repeats the ritual and shifts the probability bay another 10%. Eventually, G30 is the kind of person who will meditate and write no matter what. The meditation no longer serves function, but continues nevertheless. The novel is written, but 365-30=335 hours are wasted on an unnecessary meditation.

ii. Variations

H1 gained a bit of weight over her undergraduate years, and decides to go on a diet to lose 10 pounds in a month. At work, H1 is tempted by the wonderful dessert selection at lunch, and H1 can choose to (a) have a piece of tiramisu (just this once!), or (b) maintain the integrity of the diet.

At the end of a month, Reality swoops in with two transparent boxes, leaving H30 with the choice of either both boxes or just Box 2. In Box 1 is a piece of tiramisu. In Box 2 is a magic bean that instantly induces 10 pounds of weight loss, but Box 2 will be empty iff Reality thinks H30 is the kind of person who would take both boxes. H30 sees an empty Box 2, shrugs, and takes Box 1 like H1, ..., H29 did. H30 is tired of tiramisu, but she isn't losing weight anyway.

I(today) plays an iterated prisonner's dilemma with I(yesterday). In each round, I(today) can choose to sleep on time (cooperate) or Netflix into the wee hours of the morning (defect). I(yesterday) is a known Tit-for-Tat player - in tomorrow's game, I(yesterday) is guaranteed to make the same move I(today) made today. I(today) reasons that the only way to get out of defect-defect against Tit-for-Tat is to cooperate first, so he cooperates.

iii. Planning and Self-Improvement

Long-term projects for one person can be difficult for the same reasons that short-term projects for large teams are difficult:

1. The individual has imperfect shared knowledge (C) and values (D) across time, and communication between selves suffers from the illusion of transparency.
2. The individual doesn't trust his future and past selves (A, B, E, H), and has much less influence over them than he thinks.
3. The individual gets bogged down into meta-level planning meetings and team-building exercises without actually shutting up and doing the work (F).
4. The individual becomes superstitious about improvement rituals (G).

A few first-pass ideas:

1. Broadcast transparently to your future self. Send costly signals, decide on Schelling points, etc.
2. Become legible, for the same reasons we'd like friendly AI to be. Follow hard-and-fast rules. Arrive on time. Stick to plans.
3. Shut up and do the thing. Do it now. Do it badly. Dwell not on quality.
4. Your never make decisions about *what you do right now.* Your decisions are always about *the kind of person you are*.
5. You are the Omega now. You might be the one agent in the universe who stands a chance of simulating you to sufficient precision. Building habits and changing *who you are* can be about setting up the right Newcomb-like problems for yourself.

2. Moloch for One

I derived the Solitaire Principle from the following quote of Solzhenitsyn:

> If only it were all so simple! If only there were evil people somewhere insidiously committing evil deeds, and it were necessary only to separate them from the rest of us and destroy them. But the line dividing good and evil cuts through the heart of every human being. And who is willing to destroy a piece of his own heart?

The evil that is Moloch, Moloch who lives in the vacuum between naive libertarians and the gears of capitalism and the manic whispers of causal decision theorists (*defect! defect!*), that evil lies in your heart too.

i. Subpersonalities

Many schools of psychology have taken seriously the idea that the human consciousness decomposes into separate sub-personalities, although the exact divisions are very different. Kahneman's System 1 and System 2 is a simple dichotomy in this vein. Freud decomposed the self into id, ego, and superego. Jordan Peterson argues that ancient Gods are embodiments of primordial human subpersonalities. The Internal Family Systems (IFS) model takes another tack:

> IFS sees consciousness as composed of three types of subpersonalities or parts: managers, exiles, and firefighters. Each individual part has its own perspective, interests, memories, and viewpoint. A core tenet of IFS is that every part has a positive intent for the person, even if its actions or effects are counterproductive or cause dysfunction. This means that there is never any reason to fight with, coerce, or try to eliminate a part; the IFS method promotes internal connection

> and harmony.
> […]
> IFS practitioners report a well-defined therapeutic method for individual therapy based on the following principles. In this description, the term "protector" refers to either a manager or firefighter.
> Parts in extreme roles carry "burdens," which are painful emotions or negative beliefs that they have taken on as a result of harmful experiences in the past, often in childhood. These burdens are not intrinsic to the part and therefore they can be released or "unburdened" through IFS. This allows the part to assume its natural healthy role.
> The client's Self is the agent of psychological healing. The therapist helps the client to access and remain in Self and provides guidance in the therapy process.
> Protectors can't usually let go of their protective roles and transform until the exiles they are protecting have been unburdened.
> There is no attempt to work with any exile until the client has obtained permission from any protectors who are protecting that exile. This makes the method relatively safe, even in working with traumatized parts.
> The Self is the natural leader of the internal system. However, because of harmful incidents or relationships in the past, protectors have stepped in to protect the system and taken over for the Self. One protector after another is activated and takes over the system causing dysfunctional behavior. These protectors are also frequently in conflict with each other, resulting in internal chaos or stagnation. The goal of IFS is for the protectors to come to trust the Self so they will allow it to lead the system and create internal harmony under its guidance.

I have previously speculated on salient divisions of my own internal processes into subpersonalities, e.g. Babble and Prune, Chinese and English. For now, the exact details of how subpersonalities should be split are not important - my sense is that every such theory is typical-minding straight off a cliff anyway.

Instead, I'll start with a simplified model of subpersonalities ("agents"). Here are the rules.

1. There are at least two agents.
2. Among them there is one you identify with most, the Self.
3. Agents have different values and(/because) they have different beliefs about reality.
4. The more CPU time an agent gets, the more it grows.

ii. Three Pairs of Nemeses

Babble and Prune are seven year olds who write poetry together. Babble writes the lines, and Prune edits them. One day, Prune gets a Yeats collection for Christmas. He falls in love:

> Had I the heavens’ embroidered cloths,
> Enwrought with golden and silver light,
> The blue and the dim and the dark cloths
> Of night and light and the half light,
> I would spread the cloths under your feet:
> But I, being poor, have only my dreams;
> I have spread my dreams under your feet;
> Tread softly because you tread on my dreams.

Babble's poetry no longer lives up to Prune's standards, so they stop playing together. She continues to write with the proficiency of a seven year old while he ransacks the poetry section of the library.

Yin and Yang cohabit uncomfortably. Yin sits hunched over as if to minimize and protect herself. Her inner life is filled with jealousy, vindictiveness, and unprovoked images of violence and sadism. Yang stands upright with his shoulders back, ready to meet the world. His inner life is filled with confidence, empathy, and faith in the good.

Yin and Yang each believe that other human beings are mostly like themselves.

When Yin is awake, she perceives jokes as sarcasm and body language as hostile. She is intimately aware of the vulnerabilities of her flesh. Yin is constantly at the ready, calculating how to strike the enemy preemptively.

Yang sees the good in people. He perceives jokes as gentle and body language as inviting. He is willing to extend a charitable hand in good faith, believing other people to be like himself.

Yin and Yang both want CPU time, and are thus beset by perverse incentives. Yin is as nasty as possible to people, provoking their enmity. This enmity Yin uses as evidence that her worldview is true and people are inherently evil and that she thus deserves more CPU time. Yang is friendly and forgiving, earning their trust and respect. This good nature Yang uses as evidence that people are inherently good and that he is the one who deserves more CPU time.

Actor and Scribe have competing worldviews. In the lab, Scribe determines the truth via the scientific method, controlling, double-blinding, the whole shebang. Scribe uses words to denote pieces of reality. Scribe knows about the conjunction fallacy and believes in Occam's razor: that simplicity is proof.

Actor uses words enactively. Actor believes all good things come from willpower and placebomancy. In conversation, Actor takes complexity as proof of honesty, because it's harder to falsify a consistent and persuasive hypothesis with more moving parts. Actor worships mystery and complexity for their own sake, for mysterious and complex things cast long shadows and make good dinner conversation.

Actor and Scribe each tries to surround you with people like himself. Actor wants you to be popular and plays word games to climb the social hierarchy. Scribe looks for communities where truth and simplicity are sacred. Each knows that success in this regard is the key to winning the war.

iii. God's Eye View

In each of these examples, two nemesis subpersonalities that both serve important functions oppose and detract from each other. They respond to perverse incentives to increase their individual power (CPU time) rather than maximizing value produced. As these oppositions between subpersonalities proliferate, we have a chaotic multi-agent race to the bottom - an inner Moloch.

The common refrain in Meditations on Moloch is that Moloch can be defeated from a god's-eye-view:

4. Coordination.
The opposite of a trap is a garden.
Things are easy to solve from a god's-eye-view, so if everyone comes together into a superorganism, that superorganism can solve problems with ease and finesse. An intense competition between agents has turned into a garden, with a single gardener dictating where everything should go and removing elements that do not conform to the pattern.

Jungian psychoanalysis and IFS agree that the path to maturity is the *integration* of agents into a whole under the leadership of a driving Self, the Optimization Czar, the gentle Gardener. What does integration mean, and how is it accomplished?

The Self must become strong enough to lead all other agents. This cannot be achieved through tyranny. Rather, it must be recognized that all agents have an internal logic and rationality given their beliefs and values, and serve a purpose to the collective. By fostering healthy discourse norms, the Self can allow antagonistic agents to exchange information and understand that they share terminal values. Build your mind into a walled garden.

Babble and Prune are both necessary ingredients to a productive poet. Without Babble, Prune is just a miserable critic. Without Prune, Babble will never grow past the ability of a seven-year-old.

Yin, the Jungian shadow, is necessary to protect against genuine malice in the world. To integrate the shadow requires coming to terms with the fact that a human being is a horrifyingly dangerous animal. To nevertheless stand up straight with your shoulders back and meet people in good faith - knowing something of their nature - requires a correspondingly strong Yang.

Actor and Scribe are both correct about how to speak. I hardly need to prove the value of speaking the truth, but it's also impossible to just say what you mean.

The grand conceit of our civilization is that each individual human being has intrinsic value, be he ne'er so vile. Taking this nearly absurd principle seriously has been unbelievably productive. To achieve a harmony of all the contradictory multitudes within the individual soul requires applying that same idealistic conceit to each subpersonalities in turn.